**ЗАТВЕРДЖЕНО**

Науково-методичною радою Державного університету «Житомирська політехніка»

протокол від \_\_ \_\_\_\_\_\_\_ 20\_\_ р. №\_\_

**КОНСПЕКТ ЛЕКЦІЙ**

**з навчальної дисципліни**

**«Цитфрове підприємництво в дії»**

для здобувачів вищої освіти освітнього ступеня «магістр»

спеціальності 051 «Економіка»

освітньо-професійна програма «Економіка»

факультет бізнесу та сфери обслуговування

(назва факультету)

кафедра цифрової економіки та міжнародних економічних відносин

 (назва кафедри)

Рекомендовано на засіданні кафедри цифрової економіки та міжнародних економічних відносин

 (назва кафедри)

\_\_\_ \_\_\_\_\_\_\_\_\_\_ 2021 р., протокол № \_\_\_

The Data Entrepreneurship in Action syllabus is developed in the framework of ERASMUS+ CBHE project “Digitalization of economic as an element of sustainable development of Ukraine and  Tajikistan” / DigEco 618270-EPP-1-2020-1-LT-EPPKA2-CBHE-JP

*This project has been funded with support from the European Commission. This document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained there in.*

*Цей проєкт фінансується за підтримки Європейської Комісії. Цей документ відображає лише погляди автора, і Комісія не несе відповідальності за будь-яке використання інформації, що міститься в документі.*

**Житомир  
 2021-2022 н.р.**

Інформаційне підприємництво в дії  
[Електронний ресурс] : Конспект лекцій з дисципліни «Цифрове підприємництво в дії» для здобувачів вищої освіти «Магістр» за спеціальністю: 051 «Економіка». – Житомир : Державний університет «Житомирська політехніка», 2021. – Режим доступу:

Конспект лекцій розроблено в рамках проєкту Erasmus+ “Діджиталізація економіки як елемент сталого розвитку України та Таджикистану (DigEco) 618270-EPP-1-2020-1-LT-EPPKA2-CBHE-JP”/ The proramm is developed in the framework of ERASMUS+ CBHE project “Digitalization of economic as an element of sustainable development of Ukraine and Tajikistan” / DigEco 618270-EPP-1-2020-1-LT-EPPKA2-CBHE-JP

Цей проект фінансується за підтримки Європейської Комісії. Цей документ відображає лише погляди автора, і Комісія не несе відповідальності за будь-яке використання інформації, що міститься в документі/This project has been funded with support from the European Commission. This document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained there in.

Розробник: аспірант кафедри інформаційних систем в управлінні та обліку, Дмитро Миколайович Захаров; доцент кафедри інформаційних систем в управлінні та обліку, кандидат економічних наук, доцент, Городиський Микола Петрович.

© Державний університет «Житомирська політехніка», 2021

© М. П. Городиський, 2021

**Змістовий модуль 1. Методи аналізу даних**

Тема 1. Статистичні методи аналізу даних

Основні поняття й завдання аналізу даних. Загальна методологія дослідження. Перевірка статистичних гіпотез. Дисперсійний аналіз. Кореляційний аналіз. Факторний аналіз. Завдання та методи класифікації даних. Побудови й дослідження регресійних моделей.

[5] c. 111-137; c. 147-176; c. 181-207; c. 213-235; c. 239-280; c. 288-315

[6] c. 63-120; с. 149-165; с. 191-233; с. 256-272

***Лекція 1.1. (2 год)***

**Тема 1. Статистичні методи аналізу даних**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Основні поняття й завдання аналізу даних
2. Загальна методологія дослідження
3. Завдання та методи класифікації даних
4. Перевірка статистичних гіпотез
5. Дисперсійний аналіз

# ***1. ОСНОВНІ ПОНЯТТЯ Й ЗАВДАННЯ АНАЛІЗУ ДАНИХ***

Успішність застосування будь-якого методу аналізу даних залежить від відповідності аналізованих даних його вихідним припущенням. Методи, придатні для одного типу даних, можуть призводити до серйозних помилок при їх використанні для даних інших типів.

*Класифікація ознак за шкалами вимірювання*

Першим етапом аналізу будь-яких даних є визначення їх типу. Основною є класифікація даних за шкалами їх вимірювання. Розрізняють такі типи ознак:

1. Номінальні ознаки **(**ознаки з невпорядкованими станами, кла**сифікаційні ознаки**) – це дані, що вимірюють в номінальній шкалі (класифікаційній, шкалі найменувань). Найменування класів можуть бути виражені за допомогою чисел, але ці числа можуть використовуватися лише для відповіді на питання: належать два об’єкти до одного класу чи ні. Прикладами номінальних ознак є назви біологічних видів, назви навчальних дисциплін, кольори тощо. З погляду автоматизації аналізу даних і застосування стандартних алгоритмів доцільно обирати такі позначення класів: 0, 1, 2, ... Але з цими числами не можливо виконувати будь-які дії, крім перевірки їх рівності або нерівності.

Порядкові ознаки **(**ознаки з упорядкованими станами, ординальні ознаки**)** – це дані, що вимірюють в порядкових шкалах. Ці дані можуть порівнюватися між собою у певному відношенні: «більше – менше», «легше – важче», «правіше – лівіше» тощо. Прикладами порядкових ознак є сила землетрусу, військові звання, оцінки студентів тощо. Якщо значення порядкової ознаки є числами, то вони можуть застосовуватися і для порівняння ступеня вияву класифікаційної ознаки, але відстані між класами при цьому будуть не визначені.

Кількісні **(**числові, варіаційні**)** ознаки – це ознаки, які вимірюють у кількісних (інтервальних, відносних, циклічних та абсолютних) шкалах вимірювань. Дії, що можуть виконуватися з числовими характеристиками даних, залежать від шкали вимірювань.

В узагальненому вигляді характеристики основних типів даних наведено в табл. 1.1.

Дані, отримані у шкалах вищих рангів, можуть приводитися до шкал нижчих рангів. Наприклад, дані, що виміряні у шкалі відношень, можна привести до інтервальної шкали. Такі перетворення називають зниженням шкали. Необхідність у них зазвичай виникає при обробці даних, що виміряні у шкалах різного типу. Зворотну операцію – перетворення даних, що виміряні у нижчих шкалах, до вищих – вважають некоректною. Зниження шкали призводить до втрати частини наявної інформації про досліджувані ознаки.

Таблиця 1.1

Характеристики основних типів даних

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Шкала  вимірювань | Визначальні  відношення | Еквівалентні  перетворення | Допустимі операції над даними | |
| Первинна  обробка | Вторинна  обробка |
| Номінальна | Еквівалентність | Перестановки  найменувань | Обчислення символу Кронекера | Обчислення відносних частот та операції над ними |
| Порядкова | Еквівалентність,  перевага | Монотонні (такі, що не змінюють порядку) | Обчислення та рангів | Обчислення відносних частот та квантилів, операції над ними |
| Інтервальна | Еквівалентність, перевага, збереження відношення інтервалів | Лінійне перетворення | Обчислення  рангів та інтервалів (різниць між даними) | Арифметичні дії над інтервалами |
| Циклічна | Еквівалентність, перевага, збереження відношення інтервалів, періодичність | Зсув | Обчислення  рангів та інтервалів (різниць між даними) | Арифметичні дії над інтервалами |
| Відношень | Еквівалентність, перевага, збереження відношення інтервалів, збереження відношення двох значень | Розтягання | Усі арифметичні операції | Будь-яка придатна обробка |
| Абсолютна | Еквівалентність, перевага, збереження відношення інтервалів, збереження відношення двох значень, абсолютна й безрозмірна одиниця, абсолютний нуль | Не існує (шкала є унікальною) | Усі арифметичні операції, використання як показника степеня, основи та аргументу логарифма | Будь-яка потрібна обробка |

Важливими типами класифікації є поділ ознак за дискретністю або неперервністю теоретичної функції розподілу, законом розподілу тощо.

Як характеристики вибірки можна використовувати точкові та інтервальні оцінки. Точковими оцінками параметрів вибірки називають такі оцінки, що визначаються одним числом. Прикладами таких оцінок є середні арифметичні й медіани вибірок. При малих обсягах вибірок, а також при їх значному відхиленні від нормального закону розподілу точкові оцінки можуть істотно відхилятися від істинних значень оцінюваних параметрів. Тому поряд з ними використовують інтервальні оцінки параметрів. Інтервальні оцінки визначаються двома числами – межами інтервалу, до якого із заданою ймовірністю потрапляє оцінюваний параметр.

# ***2. ЗАГАЛЬНА МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ***

Статистика – це точна наука, що вивчає методи збору, аналізу і обробки даних, які описують масові дії, явища і процеси. Дані, що вивчаються в статистиці, характеризують не окремі об’єкти, а їх сукупності. Статистика є одним з інструментів аналізу експериментальних даних і емпіричних спостережень, а також мовою, за допомогою якої повідомляються отримані математичні результати.

Статистичні методи включають як прості методи, які доступні навіть непідготовленим користувачам, так і складні математичні процедури, доступні лише кваліфікованим фахівцям високого класу. Методами статистичної обробки результатів експерименту називаються математичні прийоми, формули, способи кількісних розрахунків, за допомогою яких показники, що одержані в ході експерименту, можна узагальнювати, приводити в систему, виявляючи приховані в них закономірності.

Головна мета будь-якого статистичного методу – представити кількісні дані в систематизованій і стислій формі з тим, щоб полегшити їх розуміння.

Всі методи статистичного аналізу умовно діляться на первинні і вторинні.

Первинні – методи, за допомогою яких можна отримати показники, що безпосередньо відображають результати отриманих в експерименті вимірювань. Це: визначення середньої арифметичної; дисперсії; моди; медіани.

Вторинні – методи статистичної обробки, за допомогою яких на базі первинних даних виявляють приховані в них статистичні закономірності. Це: кореляційний аналіз; регресійний аналіз; факторний аналіз; методи порівняння первинних даних двох або декількох вибірок.

Сукупність, з якої відбирають певну частину її членів для спільного вивчення, називають генеральною. Відібрана тим чи іншим способом частина генеральної сукупності отримала назву вибіркової сукупності або вибірки. Загальну суму членів генеральної сукупності називають її обсягом і позначають буквою *N*.

Обсяг генеральної сукупності нічим не обмежений, тобто генеральну сукупність представляють як нескінченно велику множину відносно однорідних одиниць або членів, що складають її зміст. Обсяг вибірки, що позначається буквою n, може бути і великим, і малим, але він не може містити менше двох одиниць.

# ***3. ЗАВДАННЯ ТА МЕТОДИ КЛАСИФІКАЦІЇ ДАНИХ***

У загальному випадку **класифікацією (розпізнаванням образів)** називають поділ досліджуваної сукупності об’єктів на однорідні в певному розумінні групи (класи) або зарахування кожного із заданої множини об’єктів до деякого із заздалегідь відомих класів. При цьому вирізняють три групи завдань: дискримінацію, кластеризацію й групування. Останні дві групи є близькими за метою (поділ даних на класи або групи близьких у певному розумінні об’єктів), а також за алгоритмами. Але принципова різниця між ними полягає у тому, що у першому випадку межі класів є природними, а у другому – умовними й їх можна встановлювати суб’єктивно.

При побудові методів розв’язування цієї задачі зазвичай прагнуть мінімізувати ймовірність неправильної класифікації.

На результат класифікації істотно впливають типи класів. Найчастіше виокремлюють такі типи.

1. **Клас типу ядра, або згущення**. У цьому випадку всі відстані між об’єктами всередині класу є меншими, ніж їх відстані до будь-якого об’єкта, що не входить до цього класу.

2. **Кластер, або згущення у середньому**. Середня відстань між об’єктами всередині класу є меншою, ніж їх середня відстань до всіх інших об’єктів.

3. Для **класу типу стрічки** існує таке ε > 0, що в цьому класі є хоча б один об’єкт xi, для якого відстань до будь-якого іншого об’єкта цього класу xj dij < ε, а відстань до будь-якого об’єкта xk, що не належить до цього класу, dij < ε.

4. Характеристичною властивістю **класу із центром** є існування певних граничного значення R і точки x∗ у просторі ознак таких, що у багатовимірному шарі радіуса R з центром в точці x∗ містяться всі елементи цього класу й немає елементів, які не належать до нього.

Слід зазначити, що один й той самий клас може задовольняти визначення декількох типів. Тому вказана класифікація важлива не стільки з погляду зарахування конкретного класу до певного типу, а з погляду вибору методів розділення декількох класів. Класи, що перетинаються, можуть не задовольняти будь-які з наведених вимог. Але в окремих випадках їх також можна розділяти за допомогою формальних алгоритмів класифікації.

**Параметричні методи класифікації без навчання**

У методах класифікації без навчання програмна система на основі визначених нею самою критеріїв здійснює класифікацію певних об’єктів (образів). У деяких випадках можуть бути задані окремі параметри, але розподіл об’єктів за класами на основі цих параметрів виконується автоматично.

Параметричні методи класифікації без навчання використовують при класифікації об’єктів О1, О2, …, Оn, якщо апріорна інформація про класи може бути подана у вигляді суміші параметрично заданих одномодальних функцій щільності розподілу ймовірностей fj (X, Θj), j =1, ..., k з невідомими значеннями векторних параметрів Θі.

У задачах класифікації зазвичай розглядають дискретні суміші.

Розв’язання задачі розщеплення суміші розподілів передбачає побудову статистичних оцінок для кількості компонентів суміші (класів) k, їх питомих ваг (апріорних імовірностей) πj та функцій fj (X, Θj) для кожного із компонентів за наявною вибіркою спостережень X1, X2,..., Xn.

Основною ідеєю більшості методів розв’язування цієї задачі є прагнення зарахувати спостереження Хі до того класу, для якого функція правдоподібності буде максимальною. У найпростішому випадку із попередніх досліджень можуть бути відомі кількість класів, їх апріорні ймовірності та параметричний вигляд функцій щільності ймовірності fj (X, Θj), але невідомі значення параметрів Θj. Якщо при цьому є навчальні вибірки, то ми отримуємо задачу параметричного дискримінантного аналізу, яка більш докладно розглядається нижче. Якщо ж таких вибірок немає, то значення параметрів необхідно оцінити за наявною вибіркою спостережень за допомогою одного із статистичних методів оцінювання параметрів – максимальної правдоподібності, моментів тощо. Після отримання оцінок невідомих параметрів можна застосовувати схему параметричного дискримінантного аналізу. Аналогічний підхід використовують і в більш складних випадках, коли кількість класів та їх апріорні ймовірності є невідомими. У цьому разі їх також необхідно оцінити за наявною вибіркою.

Для розв’язання задачі розщеплення суміші розподілів часто використовують **EM (Expectation – Maximization)** алгоритм, вперше запропонований в 1977 р. американськими статистиками Артуром Демпстером, Неном Лайрдом і Дональдом Рубіном. Цей алгоритм дає змогу визначати методом найбільшої правдоподібності параметри статистичних моделей, що містять певні приховані змінні. Він передбачає здійснення двох кроків на кожній ітерації. Перший крок (Expectation) полягає в обчисленні значення функції правдоподібності за умови, що задані деякі значення прихованих змінних. На другому кроці (Maximization) обчислюють значення параметрів, що максимізують функцію правдоподібності. Обчислення виконують до виконання заданих умов збіжності. Недоліками алгоритму є залежність результату від вибору початкового наближення (якщо функція правдоподібності не є унімодальною). Крім того, цей алгоритм не дає змоги визначити кількість компонент суміші. Для усунення цих недоліків пізніше було запропоновано різноманітні модифікації EM алгоритму: медіанні, стохастичний (SEM), класифікаційний (CEM), узагальнений (GEM), з додаванням компонент тощо.

**Кластерний аналіз**

У непараметричному випадку ми не маємо інформації про загальний вигляд функцій fj (X, Θj). Ми можемо мати лише окремі загальні відомості про них: компактність або обмеженість діапазонів змінювання компонент класифікованих багатовимірних спостережень, неперервність або гладкість відповідних законів розподілу ймовірностей тощо. Вихідні дані зазвичай подають у вигляді матриці спостережень, яка містить значення всіх ознак для кожного із досліджуваних об’єктів, або матриці подібності, що містить попарні відстані між класифікованими спостереженнями.

Бажано, щоб компоненти вектора Х відповідали одному й тому самому типу даних. Для цього зазвичай використовують перехід від кількісних ознак до порядкових та від порядкових до номінальних. Але слід ураховувати, що при цьому втрачається частина корисної інформації.

Для формалізації задачі класифікації кожний об’єкт зручно інтерпретувати як точку в багатовимірному просторі ознак. Геометрична близькість точок у такому просторі відповідає близькості досліджуваних об’єктів з погляду досліджуваних властивостей (рис. 1.1). Наочне уявлення про зміст класифікації дає діаграма Герцшпрунга – Ресселла (рис. 1.2), яка є основою для однієї з найпоширеніших класифікацій зірок за поєднанням їх світності й кольору (температури або спектрального класу).

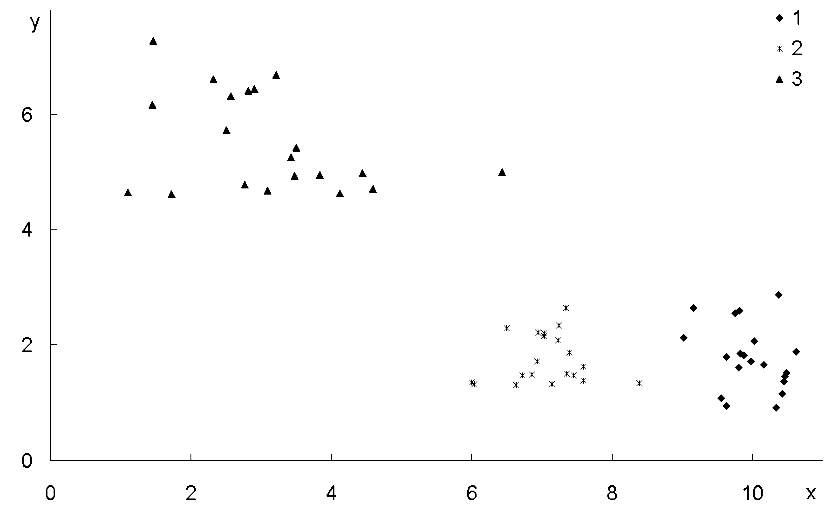


Рис. 1.1. Геометричне зображення сукупності об’єктів, що характеризуються двома ознаками й утворюють три кластери

Залежно від мети дослідження задачу класифікації можна сформулювати як розбиття аналізованих об’єктів на певну кількість груп, усередині яких вони розташовані на порівняно малій відстані один від одного, або як виявлення природного розшарування сукупності, що вивчається, на окремі кластери. Другу задачу можна також сформулювати як визначення областей підвищеної густини точок, що відповідають наявним спостереженням.

Перша задача завжди має розв’язок, а друга може не мати розв’язку. Це відповідає відсутності природного розшарування досліджуваних об’єктів (наприклад, вони утворюють один кластер або відповідні точки рівномірно заповнюють весь простір ознак).

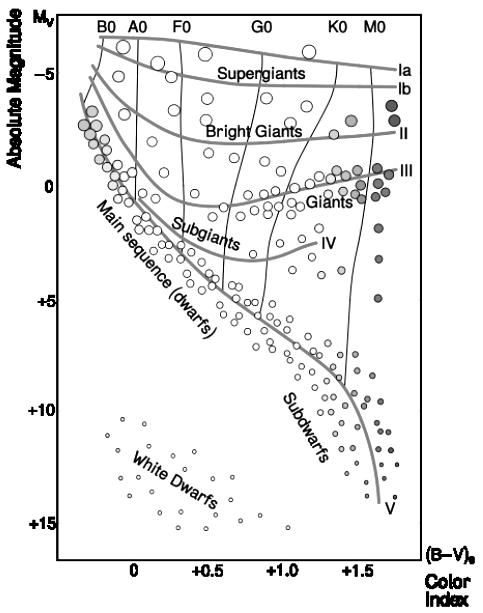


Рис. 1.2. Діаграма Е. Герцшпрунга – Г. Ресселла

Класичними непараметричними методами класифікації без навчання є методи **кластерного аналізу (таксономії)**. За їх допомогою вирішують проблему такого розбиття (класифікації, кластеризації) множини об’єктів, за якого всі об’єкти, що належать до одного класу, були б більш подібними один до одного, ніж до об’єктів інших класів. З формальної точки зору, основне завдання методів кластерного аналізу можна сформулювати, як визначення класів еквівалентності й рознесення за ними досліджуваних об’єктів. Під **класом**, як правило, розуміють генеральну сукупність, що описується одномодальною функцією щільності ймовірності f(X) або, у випадку дискретних ознак, – одномодальним полігоном імовірностей. Номери класів не мають змістового навантаження й використовуються лише для того, щоб відрізняти їх один від одного.

Для формування кластерів застосовують міри подібності та відмінності даних, які можуть бути поділені на три основних види:

– **міри подібності (відмінності) типу «відстань»** (при їх застосуванні об’єкти вважають тим більш подібними один до одного, чим меншою є відстань між ними);

– **міри подібності типу «зв’язок»** (у цьому випадку об’єкти вважають тим більш подібними, чим сильнішим є зв’язок між ними);

– **інформаційна статистика**.

Вибір міри відстані істотно впливає на результат класифікації. Тому для отримання надійних результатів необхідно враховувати мету дослідження, змістову й статистичну природу вектора спостережень і наявні відомості про характер розподілу досліджуваних ознак. Крім того, після закінчення розрахунків слід перевіряти адекватність отриманої класифікаційної моделі.

Найчастіше використовують евклідову та манхеттенську відстані, супремум-норму, а також відстань Махаланобиса. Вони відображають усе різноманіття підходів до цієї проблеми. Евклідову метрику традиційно застосовують як міру відстані. Манхеттенська відстань є найбільш відомою з класу метрик Мінковського. Відстань Махаланобиса, що не є метрикою, за допомогою дисперсійно-коваріаційної матриці пов’язана з кореляціями змінних і широко використовується у кластерному аналізі та інших методах аналізу даних. Вказані міри подібності можуть бути застосовані при реалізації методів ближнього зв’язку, середнього зв’язку Кінга, Уорда, k-середніх Мак-Куїна.

Як міри зв’язку для кількісних ознак можна обирати коефіцієнт кореляції Пірсона, кореляційне відношення і дисперсію-коваріацію. Застосування коефіцієнта кореляції Пірсона є обґрунтованим лише за умови, що зв’язок між ознаками є лінійним, але в окремих випадках його можна використовувати для нелінійного зв’язку після придатного перетворення вихідних ознак.

Для порядкових ознак призначені коефіцієнти рангової кореляції Спірмена й Кендалла. У цьому випадку їх називають, відповідно, **відстанями Спірмена й Кендалла**. Для дихотомічних ознак та ознак, що розміщуються в таблицях спряженості, використовують хеммінгову відстань, показник Жаккара, простий коефіцієнт зустрічальності, показник Рассела й Рао, коефіцієнт асоціації Юла, коефіцієнт спряженості Бравайса. Розглянуті показники (крім хеммінгової відстані) можна перетворити у відстані, віднімаючи обчислені значення від одиниці.

Для змішаних ознак користуються коефіцієнтом Гауера.

Перелічені міри зв’язку застосовують у методах ближнього зв’язку, кореляційних плеяд та максимального кореляційного шляху. Зазвичай за допомогою першого з цих методів класифікують об’єкти, а за допомогою двох інших – параметри. Але шляхом транспонування матриці вихідних даних можна легко змінити тип класифікації на протилежний. Результати класифікації різними методами, як правило, принципово не відрізняються.

Вибір метрики, навпаки, може істотно впливати на результати аналізу. Тому для кожної конкретної задачі його необхідно здійснювати окремо. При цьому треба враховувати головні цілі дослідження, фізичну й статистичну природу вихідних даних, повноту апріорних відомостей про тип функцій розподілу ймовірності. Зокрема, якщо кластери можна інтерпретувати як нормальні генеральні сукупності з однією і тією самою коваріаційною матрицею, то доцільно обирати відстані типу Махаланобиса, окремими випадками якої є евклідова, зважена евклідова та хеммінгова відстані.

Дивергенція між двома сукупностями і та j (повна середня інформаційна міра різниці двох класів, дивергенція Кульбака – Лейблера) запропонована американськими криптоаналітиками Соломоном Кульбаком та Річардом Лейблером у 1951 р.

**Міра Махаланобиса (відстань Махаланобиса, узагальнена евклідова відстань)** – це відстань від точки спостереження до центра ваги в багатовимірному просторі ознак.

**Евклідова відстань (евклідова метрика)** є відомою із загальноматематичних дисциплін. Вона збігається з відстанню Махаланобиса у випадку, коли незалежні змінні є некорельованими.

Поряд з евклідовою відстанню як міру близькості часто використовують її квадрат.

Евклідову відстань доцільно обирати, якщо:

– спостереження належать до генеральних сукупностей, які підпорядковуються багатовимірним нормальним законам, а компоненти вектора спостережень є незалежними і мають одну й ту саму дисперсію;

– компоненти вектора спостережень є однорідними з погляду змістової інтерпретації та однаково важливими для класифікації;

– простір ознак має розмірність 1, 2 або 3, і поняття близькості об’єктів у цьому просторі збігається із звичайною геометричною близькістю.

Недоліком евклідової метрики є те, що у випадках, коли ознаки виміряні у різних одиницях, зміна масштабу одиниць вимірювання може призвести до істотної зміни результатів класифікації. Для запобігання цьому використовують різні методи нормування даних.

Але слід зазначити, що нормування також впливає на результати класифікації. Зокрема, у випадках, коли кластери істотно розділяються за деякими ознаками й слабко за іншими, нормалізація може призвести до зменшення дискримінуючих можливостей першої групи ознак через збільшення шумового ефекту інших.

**Метрика Мінковського**, запропонована видатним німецьким математиком і фізиком Германом Мінковським у 1908 р., є узагальненням звичайної евклідової відстані.

У випадку r = 1 метрика Мінковського дає **манхеттенську відстань**.

При r → ∞ метрика Мінковського збігається із **супремум-нормою (відстанню Чебишева)**, яку було введено видатним російським математиком Пафнутієм Львовичем Чебишевим.

**Хеммінгову відстань**, яку ввів відомий американський математик Річард Хеммінг у 1950 р., використовують як міру відстані об’єктів, що характеризуються дихотомічними ознаками.

При конструюванні різноманітних процедур класифікації доцільно використовувати міри близькості кластерів один до одного. Найбільш поширеними з них є відстані, що вимірюють за принципами найближчого і далекого сусідів, середнього зв’язку та за центрами ваги. Вибір міри близькості кластерів є найбільш суттєвим для агломеративних ієрархічних методів кластерного аналізу.

Для розрахунку ступеня близькості класів використовують також розглянуті вище інформаційну відстань Каллбека (у випадку, коли класи можна розглядати як багатовимірні нормальні сукупності) та відстань Махаланобиса (якщо додатково відомо, що вони мають однакові коваріаційні матриці).

Порівняння різних способів розбиття досліджуваної сукупності об’єктів на класи здійснюють за допомогою **функціонала якості розбиття** Q(S). Найкращим вважають розбиття, при якому забезпечується екстремум цього функціонала. Не існує чітких методів обрання функціоналів якості.

У випадку, коли кількість класів є заданою, як функціонали якості найчастіше обирають такі.

**Узагальнену внутрішньокласову дисперсію** можна розраховувати як показник середньоарифметичної або середньогеометричної дисперсії.

Такі функціонали доцільно застосовувати, якщо допускають можливість зосередженості розбитих на класи спостережень у просторі меншої розмірності, ніж p. 169 За невідомої кількості класів функціонали якості розбиття зазвичай обирають у вигляді простої алгебраїчної комбінації (суми, різниці, добутку, частки) двох функціоналів, один з яких є незростаючою функцією кількості класів і характеризує внутрішньокласовий розкид спостережень, а другий – незгасаючою функцією кількості класів. Останній може характеризувати взаємну віддаленість (близькість) точок, втрати від надмірної деталізації вихідного масиву даних, концентрацію наявної структури точок тощо.

У **схемі О.М. Колмогорова** для побудови такого функціонала використовують **міру концентрації точок**.

Залежно від кількості вихідних спостережень виділяють задачі класифікації невеликих за обсягом (до декількох десятків точок) масивів спостережень і задачі класифікації великих масивів. Такий поділ зумовлений різницею процедур, які доцільно використовувати при класифікації відповідних даних.

З погляду апріорної інформації про кількість кластерів вирізняють такі типи задач:

– із заданою кількістю класів;

– з невідомою кількістю класів, яку треба оцінити;

– з невідомою кількістю класів, яку не потрібно оцінювати (таку задачу зазвичай формулюють як побудову ієрархічного дерева, або дендрограми вихідної сукупності.

Найпоширенішими методами кластерного аналізу є:

– ієрархічні методи (ближнього зв’язку, середнього зв’язку Кінга, Уорда, далекого зв’язку);

– ітеративні методи групування (метод k-середніх Мак-Куїна);

– алгоритми типу розрізування графа (кореляційних плеяд Терентьєва, вроцлавська таксономія).

**Ієрархічні (агломеративні та дивізимні)** методи призначені переважно для побудови ієрархічних дерев відносно невеликих за обсягом сукупностей. Іноді їх використовують також для задач класифікації перших двох типів. У цьому випадку реалізацію ієрархічного алгоритму продовжують до досягнення кількості класів, яка дорівнює заздалегідь заданому числу k, або до досягнення екстремуму одного з критеріїв якості розбиття.

Перевагами ієрархічних методів є можливість більш повного і тонкого аналізу структури досліджуваної сукупності порівняно з іншими методами, а також наочність подання результатів кластеризації. Їх основними недоліками є громіздкість обчислювальної процедури, яка пов’язана з перерахунком усієї матриці відстаней на кожному кроці, а також “скінченна неоптимальність” гранично оптимальних алгоритмів у багатьох випадках.

**Метод ближнього зв’язку** є найпростішим для розуміння з ієрархічних агломеративних методів кластерного аналізу. Процес класифікації в цьому випадку починають з пошуку та об’єднання двох найближчих один до одного об’єктів у матриці подібності.

На наступному етапі знаходять два наступні найближчі об’єкти й так само до повного вичерпання матриці подібності. Як правило, робота алгоритму закінчується, коли всі спостереження об’єднані в один клас. Для виокремлення кластерів після закінчення кластеризації задають пороговий рівень подібності, на якому можна виділити більше, ніж один кластер.

Описана процедура не завжди приводить до утворення одного великого кластера на останньому етапі. Часто вона закінчується явним розбиттям досліджуваних об’єктів на кластери.

У методі ближнього зв’язку два об’єкти потрапляють до одного й того самого кластера в тому випадку, коли існує ланцюжок близьких один до одного об’єктів, які їх з’єднують. Іноді це призводить до необґрунтованого зарахування об’єктів до одного й того самого кластера (**ланцюжковий ефект**). У процесі кластеризації можна явно простежити утворення таких ланцюжків. Для запобігання цьому ефекту можна задавати обмеження на максимальну відстань між елементами одного кластера.

Сучасний варіант цього методу, зважений алгоритм найближчих сусідів, запропоновали в 2005 р. іспанські математики Роберто Паредес та Енрікє Відал.

Кластери, одержувані за методом ближнього зв’язку, не обов’язково бувають опуклими. Залежно від обставин, це можна розглядати і як перевагу, і як недолік методу.

Після проведення кластеризації рекомендується візуалізувати результати шляхом побудови дендрограми, яка дає можливість отримати уявлення про загальну конфігурацію об’єктів.

Результати ієрархічних методів кластерного аналізу стають більш наочними, якщо їх подати у вигляді дендрограми (дендограми).

Пари об’єктів при побудові дендрограми з’єднують згідно з рівнем зв’язку, який відкладають по вісі ординат. Задаючи кількість кластерів, наприклад n = 3, знаходять, на якому рівні кількість перетинів горизонтальної лінії, яка відповідає рівню зв’язку, і вертикальних ліній, що відповідають об’єктам, дорівнює трьом.

**Метод середнього зв’язку Кінга** подібний до методу ближнього зв’язку. Його відмінність полягає в тому, що об’єднані до одного кластера об’єкти надалі вважають одним об’єктом з усередненими за кластером параметрами. При цьому новому об’єкту надають номер меншого з номерів об’єднуваних об’єктів, а об’єкти, що залишилися, перенумеровують. Таким чином їх загальна кількість зменшується на одиницю. Подальша процедура є подібною до попереднього методу. В іншому варіанті методу середнього зв’язку відстань між класами розраховують як середнє значення відстаней між усіма можливими парами представників цих класів. При використанні методу середнього зв’язку в процесі кластеризації також простежується формування ланцюжків об’єктів, що дає змогу задати пороговий рівень подібності, на якому можна виділити більше ніж один кластер. Часто процедура кластеризації закінчується явним розподілом об’єктів на кластери. Після закінчення кластеризації також доцільно здійснити візуалізацію результатів шляхом побудови дендрограми.

**К-узагальнена ієрархічна процедура** ґрунтується на тому, що перелічені відстані, а також відстань, що вимірюється за принципом далекого сусіда, яка застосовується в методі далекого зв’язку, є окремими випадками узагальненої відстані Колмогорова, яку й використовують у цьому випадку як міру близькості. Описані вище методи можна розглядати як окремі випадки К-узагальненої ієрархічної процедури.

**Порогові ієрархічні процедури** передбачають задання монотонної послідовності порогів с1, с2, …, сt. В агломеративних методах на першому кроці попарно об’єднують елементи, відстані між якими не перевищують с1. На другому кроці об’єднують елементи або групи елементів, відстані між якими не перевищують с2 тощо. При достатньо великих значеннях сt на останньому кроці всі елементи будуть об’єднані до одного загального кластера. Недоліком цих процедур є можливість перетину класів, тому вони бувають ефективними за умови, що ланцюжковий ефект слабко виражений, а вихідна сукупність природно поділяється на достатньо віддалені одне від одного скупчення точок у досліджуваному просторі ознак.

**Метод Уорда**, запропонований Дж. Уордом у 1963 р., є близьким до методу середнього зв’язку Кінга. Він відрізняється тим, що підставою для приєднання об’єкта до кластера є не близькість у значенні певної міри подібності, а мінімум дисперсії всередині кластера після поміщення до нього обраного об’єкта.

Паралельні ітеративні процедури передбачають одночасне використання всіх наявних спостережень, тому їх застосовують для розв’язання задач класифікації перших двох типів при порівняно малих обсягах досліджуваних сукупностей.

Послідовні ітераційні процедури на кожному кроці використовують лише невелику кількість спостережень, а також результат попереднього кроку кластеризації. Як правило, їх застосовують для розв’язання перших двох типів задач кластеризації при великих обсягах досліджуваних сукупностей.

Прикладом послідовної ітеративної процедури є метод k-середніх Мак-Куїна.

Розв’язується задача розбиття n об’єктів на k (k < n) однорідних у певному розумінні кластерів. На початковому етапі його реалізації вихідні точки впорядковують (можливо випадковим чином) і перші k точок у подальшому розглядають як окремі кластери, яким надають одиничні вагові коефіцієнти. Потім беруть точку Хk+1 і з’ясовують, до якого з наявних кластерів вона є найближчою. Цей кластер замінюють новим, розташованим у центрі ваги вихідного кластера й точки Хk+1. При цьому ваговий коефіцієнт отриманого кластера збільшують на одиницю порівняно із ваговим коефіцієнтом вихідного. Якщо точка Хk+1 є рівновіддаленою від декількох кластерів, то її вміщують до кластера з найменшим номером або з найбільшим ваговим коефіцієнтом. Потім почергово приєднують до наявних кластерів точки, що залишилися. При достатньо великих обсягах досліджуваних вибірок центри ваги отримуваних кластерів згодом перестають змінюватися, тобто ітераційна процедура збігається до певної границі. Якщо ж вона не збігається за задану кількість кроків, то використовують один із таких прийомів. Перший передбачає, що після розгляду останньої точки Хn повертаються до точок Х1, Х2 тощо. Другий підхід передбачає багаторазовий повторний вибір вихідних кластерів. При цьому на кожному етапі як вихідні обирають точки, що є найближчими до фінальних кластерів, що найчастіше отримували на попередніх етапах.

Особливістю методу є алгоритмічне гарантування того, що кожний із класифікованих об’єктів буде зарахований лише до одного з кластерів. При застосуванні цього методу немає особливої необхідності у візуалізації результатів. Але для наочності можна здійснити її за допомогою зображення просторових еліпсоїдів, що містять класифіковані об’єкти (якщо розмірність не перевищує трьох), або двовимірних зрізів простору. У багатьох випадках метод k-середніх дає змогу отримати розбиття, близьке до найкращого з погляду функціонала якості.

Якщо кількість класів є невідомою, необхідно задати дві константи: міру грубості φ та міру точності ψ. На нульовому кроці беруть довільне значення кількості класів k0, вихідні точки впорядковують і розглядають k0 перших точок як центри кластерів, яким надають одиничні вагові коефіцієнти. Потім здійснюють огрубіння вихідних кластерів. Для цього послідовно розраховують попарні відстані між ними і, якщо відстань між двома кластерами не перевищує φ, їх об’єднують до одного, який є їх зваженим середнім і має ваговий коефіцієнт, що дорівнює сумі вагових коефіцієнтів вихідних кластерів. Після закінчення цієї процедури ми отримуємо k ∗ 0 ≤ k0 кластерів.

Далі здійснюють послідовне рознесення точок, що залишилися, за кластерами. Для кожної точки визначають найближчий до неї кластер. Якщо відстань між ними не перевищує ψ, то відповідну точку приєднують до цього кластера за вищеописаною процедурою. У протилежному випадку її вважають центром нового кластера, якому надається одиничний ваговий коефіцієнт. Після рознесення усіх точок за кластерами повторюють процедуру огрубіння і переходять до чергового кроку ітерацій.

Обираючи різні значення констант φ та ψ, можна отримати різні розбиття вихідної сукупності. Вибір вважають задовільним, якщо результат класифікації є близьким до оптимального за оцінками експертів або з погляду функціонала якості. Можна довести, що алгоритм методу Мак-Куїна збігається до локального мінімуму суми внутрішньокласових дисперсій. Глобальний мінімум цього функціоналу може бути досягнутий за допомогою алгоритму Р. Дженсена, який базується на застосуванні динамічного програмування. Сутність методу кореляційних плеяд є такою. Візуально результати класифікації можна подати у вигляді кореляційного циліндра, розсіченого площинами, перпендикулярними його осі. Площини відповідають рівням від нуля до одиниці з кроком 0,1. На цих рівнях об’єднують класифіковані параметри або об’єкти. Метод наближається до методу ближнього зв’язку з фіксованими рівнями об’єднання. Графічно результати зображують у вигляді кіл-зрізів (плеяд) кореляційного циліндра. На них відмічають класифіковані об’єкти, зв’язки між якими вказують за допомогою хорд, що з’єднують відповідні точки кіл. Метод кореляційних плеяд є основою для багатьох порогових алгоритмів. 175

У методі **вроцлавської таксономії** визначають пари чисел, які вказують порядок з’єднання попарно найближчих один до одного об’єктів (параметрів), що підлягають класифікації. Одержуваний незамкнений найкоротший шлях можна відобразити графічно у вигляді певного оптимального дерева (дендрита). Цей метод є подібним до методу ближнього зв’язку, але його алгоритм належить до алгоритмів розрізання графів. Якщо мірою подібності обрати коефіцієнт кореляції, ми отримаємо метод найбільшого кореляційного шляху.

У методі вроцлавської таксономії результати розрахунків відображають у вигляді графа (**дендрита**).

# ***4. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ***

Існує велика кількість різноманітних методів перевірки статистичних гіпотез. При виборі методу для вирішення певного конкретного завдання необхідно виходити з відповідей на такі питання:

* якою є мета перевірки гіпотези;
* у яких шкалах виміряні аналізовані дані;
* чи є аналізовані вибірки незалежними або спряженими;
* скільки вибірок необхідно порівняти.

Розглянуті в даному пункті методи застосовують при порівнянні двох вибірок. При більшій кількості вибірок використовують методи дисперсійного аналізу.

Гіпотезу, що перевіряють, називають нульовою гіпотезою (Н0). Прикладами нульових гіпотез можуть бути такі твердження: “Середні значення двох вибірок суттєво не відрізняються одне від одного”; “Дисперсія першої вибірки суттєво перевищує дисперсію другої”; “Розподіл вибірки відповідає нормальному закону з певними параметрами” тощо. Гіпотезу, що суперечить нульовій, називають конкуруючою, або альтернативною гіпотезою (Н1). Для вказаних вище нульових гіпотез конкуруючими можуть бути такі твердження: “Середні значення двох вибірок суттєво розрізняються одне від одного”; “Дисперсія першої вибірки не перевищує істотно дисперсію другої”; “Розподіл вибірки не відповідає нормальному закону із вказаними параметрами”. Для однієї нульової гіпотези у загальному випадку можна сформулювати багато різних альтернативних гіпотез.

Розрізняють прості та складні гіпотези. **Простою** називають гіпотезу, що містить тільки одне твердження. **Складні гіпотези** складаються з декількох простих (при цьому кількість простих гіпотез може бути нескінченно великою).

Зазвичай при перевірці нульової гіпотези використовують певні модельні розподіли, що приблизно відповідають розподілу досліджуваного параметра. Їх називають **статистичними критеріями**. На практиці як критерії найчастіше використовують нормальний розподіл, χ2-розподіл, розподіли Стьюдента і Фішера. **Фактичним значенням критерію** називають його величину, яку розраховують за досліджуваними вибірками.

Для перевірки гіпотези весь вибірковий простір поділяють на дві області, що не перетинаються: критичну (***w***) та область прийняття (***W*** – ***w***). **Критичною областю** називають сукупність значень критерію, за яких нульову гіпотезу слід відхилити. **Областю прийняття гіпотези (областю допустимих значень)** називають сукупність значень критерію, за яких нульову гіпотезу приймають. Перевірка гіпотези передбачає розрахунок значення критерію і перевірку його потрапляння до області прийняття гіпотези.

Вирізняють **двобічні** й **однобічні** (лівобічні, правобічні) критичні області (рис. 1.3, 1.4). Їх використання залежить від вибору конкуруючої гіпотези.

|  |  |
| --- | --- |
| Рис. 1.3. Приклад двобічної критичної області | Рис. 1.4. Приклад правобічної критичної області |

Якщо розподіл імовірності спостережень, що відповідає нульовій гіпотезі Н0, є відомим, то критичну область визначають так, щоб при виконанні Н0 імовірність її відхилення була рівною заздалегідь заданій малій величині (**рівню значимості**) **α**.

, (1.1)

Замість рівня значимості можна використовувати також **довірчий рівень** .

Критерії, що базуються на використанні заздалегідь заданого рівня значимості, називають **критеріями значимості**. Рівень значимості визначає розмір критичної області: що більшим є рівень значимості, то ширшою буде критична область.

Розглядають два типи помилок, що можуть виникати при перевірці статистичних гіпотез:

– **помилкою першого роду** є відхилення правильної нульової гіпотези, рівень значимості а є ймовірністю такої помилки;

– **помилкою другого роду** є прийняття помилкової нульової гіпотези.

У деяких застосуваннях помилки першого та другого роду називають, відповідно, **ризиком виробника** та **ризиком споживача**.

Зменшення ймовірності помилки першого роду водночас призводить до підвищення ймовірності помилки другого роду **β**. З огляду на це додатково вводять поняття **потужності критерію** , яка є імовірністю відхилення помилкової нульової гіпотези, тобто ймовірністю потрапляння критерію до критичної області за умови, що правильною є конкуруюча гіпотеза:

, (1.2)

Потужність критерію можна підвищити, збільшуючи обсяг вибірки. При визначенні критичної області її зазвичай будують так, щоб максимізувати потужність обраного критерію. За наявності декількох критеріїв, що можуть використовуватися для перевірки досліджуваної гіпотези, рекомендується обирати більш потужні з них, якщо їх застосування є обґрунтованим.

Важливим завданням є визначення обсягу вибірки, який дає змогу гарантувати певне значення похибки першого роду за умови, що похибка другого роду не перевищує заданого значення. Для цього необхідно розв’язати таку систему:

(1.3)

Її аналітичне розв’язання можливо лише у найпростіших випадках. Але, в багатьох випадках істотне спрощення можна отримати, якщо замінити ймовірності α й β значеннями меж відповідних критичних інтервалів.

Загальна методика отримання висновків при перевірці гіпотез передбачає, що на першому етапі необхідно задати рівень значимості. Найчастіше його беруть рівним 0,01; 0,05 або 0,1. Обираючи рівень значимості, слід пам’ятати, що його зменшення знижує ймовірність помилки першого роду, але збільшує ймовірність помилки другого роду. Тому, виходячи з конкретних умов, потрібно знайти певний компроміс між ймовірностями припустити помилки різного типу.

На другому етапі за даними вибірки розраховують значення критерію та порівнюють його з обчисленими для заданого рівня значимості межами критичної області. Якщо розраховане значення критерію потрапляє до них, то нульову гіпотезу відхиляють. В іншому випадку вважають, що немає підстав для відхилення нульової гіпотези і або приймають її на заданому рівні значимості, або здійснюють додаткову перевірку. Для визначення меж критичної області застосовують спеціальні таблиці або розраховують їх на основі відомих законів розподілу використовуваних критеріїв.

Можливості сучасної комп’ютерної техніки та наявного програмного забезпечення дають змогу отримувати висновки іншим шляхом. Якщо за наявними емпіричними даними розрахувати значення критерію, то на наступному етапі можна визначити, для якого рівня значимості це значення буде критичним. Ураховуючи, що рівень значимості є ймовірністю відхилення правильної нульової гіпотези, ми можемо за його значенням зробити висновок про ймовірність правильності або помилковості нульової гіпотези. Залежно від того, задовольняє нас отримана ймовірність помилки чи ні, нульову гіпотезу приймають або відхиляють.

При перевірці гіпотез доцільно застосовувати різні методи, призначені для вирішення одних й тих самих завдань та однакових типів даних. Причинами розбіжності отримуваних при цьому результатів зазвичай є:

* помилки при введенні даних;
* непридатність окремих методик для типу даних, що розглядають;
* алгоритмічні помилки у програмах, що використовують для аналізу.

Залежно від наявності або відсутності можливості визначення напряму розбіжності порівнюваних вибірок, розрізняють **однобічні** та **двобічні критерії**. Перші застосовують, якщо наявні дані дають змогу вказати такий напрям, наприклад зробити висновок, що значення порівнюваної ознаки для одної вибірки є вищим, ніж в іншої. Двобічні критерії дають можливість зробити висновок лише про різницю вибірок за порівнюваною ознакою. Відповідно до цього говорять про однобічні й двобічні гіпотези. Для двобічних критеріїв рівень значимості є вдвічі більшим, ніж для відповідних однобічних. При використанні однобічних критеріїв рекомендується спочатку розраховувати двобічні. Якщо за двобічним критерієм різниці між вибірками немає, то наступне порівняння за однобічним є необґрунтованим.

Дані реальних експериментів можуть бути подані **незалежними** або **спряженими** вибірками. Для незалежних вибірок критерії допомагають виявити статистичну значущість різниці, що спостерігається. Прикладами незалежних вибірок є:

* мешканці двох різних населених пунктів (при демографічних дослідженнях);
* дві партії однотипної продукції, виготовлені різними працівниками на різному обладнанні (при розробці технології виробництва);
* випускники різних шкіл (при аналізі результатів зовнішнього незалежного оцінювання).

Критерії, що застосовують до вибірок з попарно спряженими даними, називають **парними**. Прикладами спряжених вибірок є:

* дані опитування громадської думки до й після певної суспільно значущої події;
* дві партії однотипної продукції, виготовлені одними й тими самими працівниками на одному й тому самому обладнанні до й після внесення певних змін до технології;
* одна й та сама партія виробів до і після певної технологічної обробки.

***Параметричні тести***

Критерії й тести, що застосовують для порівняння вибірок, поділяють на дві групи: параметричні й непараметричні. Особливістю параметричних критеріїв є припущення, що розподіл ознаки в генеральній сукупності підпорядковується певному відомому закону. Ця відповідність має бути доведена до застосування будь-якого з параметричних тестів. Переважна більшість параметричних тестів розроблена для нормально розподілених даних. Але для деяких типів гіпотез існують параметричні тести, призначені для вибірок, що підпорядковуються іншим законам розподілу.

Як правило, параметричні критерії є потужнішими за непараметричні. Застосування непараметричних критеріїв у випадках, коли можна використовувати параметричні, призводить до збільшення ймовірності прийняття помилкової нульової гіпотези, тобто помилки другого роду.

Якщо досліджувані вибірки підпорядковуються нормальному закону розподілу з відомими дисперсіями, то як критерій рівності їх середніх значень можна використовувати величину:

(1.4)

де , – кількість елементів у вибірках. Нульова гіпотеза полягає у рівності середніх.

**Z-критерій** є випадковою величиною, що підпорядковується стандартному нормальному розподілу. Якщо конкуруючою гіпотезою є , то праву критичну точку можна визначити з умови:

, (1.5)

де є функцією Лапласа (інтегралом ймовірностей), що пов’язана з функцією стандартного нормального розподілу співвідношенням . Ліва й права критичні точки для одного й того самого рівня значимості пов’язані умовою .

Якщо конкуруючою гіпотезою є: , то:

, (1.6)

і нульову гіпотезу приймають, якщо .

Якщо конкуруючою гіпотезою є: , то критичну точку визначають з умови (1.6), враховуючи, що . Нульову гіпотезу приймають, якщо .

Z-критерій можна застосовувати також для порівняння середніх значень довільно розподілених незалежних вибірок великого обсягу (), враховуючи, що в цьому разі вибіркові середні мають приблизно нормальний розподіл, а вибіркові дисперсії є достатньо точними оцінками генеральних дисперсій.

Для порівняння середніх значень вибірок застосовують **t-критерій Стьюдента**.

Розглядають дві незалежні нормальні вибірки з генеральних сукупностей, що мають рівні або нерівні, але відомі чи рівні невідомі дисперсії.

Значення критерію Стьюдента розраховують за формулою:

(1.7)

де , – відомі внутрішньогрупові дисперсії; та – чисельності груп. Для груп рівної чисельності статистика має t-розподіл з кількістю степенів вільності .

У випадку, коли обсяги вибірок є малими або істотно розрізняються, а їх дисперсії є рівними, останні замінюють вибірковим середнім квадратичним відхиленням, яке розраховують за формулою:

(1.8)

якщо стандартні відхилення вибірок оцінюють за самими вибірками, або:

(1.9)

якщо їх оцінюють незалежно. Формула для визначення розрахункового значення критерію у цьому разі набуває вигляду:

(1.10)

Відповідна статистика має розподіл Стьюдента з степенями вільності.

При застосуванні критерію Стьюдента у вигляді (1.10) необхідно спочатку перевірити гіпотезу про рівність дисперсій.

Критичні точки симетричні стосовно нуля. Нульову гіпотезу відхиляють: якщо при конкуруючій гіпотезі ; якщо конкуруючій гіпотезі ; якщо при конкуруючій гіпотезі .

При аналізі спряжених вибірок їх порівняння здійснюють з метою визначення наявності ефекту від певного фактора, наприклад, впливу змін у технології на якість виробленої продукції. Вимога щодо рівності дисперсій при цьому не висувається. Нульова гіпотеза полягає у відсутності різниці між середніми. Значення критерію розраховують за формулою:

(1.11)

де - кількість елементів у кожній із вибірок; , та – відповідні значення елементів першої та другої вибірок.

Іноді цей критерій називають **одновибірковим критерієм Стьюдента**. Відповідна статистика має розподіл Стьюдента з кількістю степенів вільності .

Якщо дисперсії або їх відношення невідомі й припущення про рівність дисперсій є необгрунтованим, то виникає так звана **проблема Беренса-Фішера**, що полягає у перевірці нульової гіпотези про рівність вибіркових середніх за таких умов. Одним з підходів до її вирішення є застосування **критерію Уелча (Крамера-Уелча)**. Його значення розраховують за формулою:

(1.12)

де , – розраховані за вибірками оцінки дисперсії.

Статистика цього критерію є приблизно такою самою, як для розподілу Стьюдента з кількістю степенів вільності:

(1.13)

Порівняння з (1.7) вказує, що основною відмінністю критерію Уелча з погляду прикладного аналізу є зміна кількості степенів вільності.

**F-критерій Фішера** запропоновано британським біологом і статистиком Рональдом Фішером в 1920 р. Його використовують для порівняння дисперсій двох вибірок. Гіпотезу про рівність дисперсій порівнюваних сукупностей відхиляють, якщо обчислене значення перевищує табличне при заданому довірчому рівні.

Для оцінювання значимості отриманого значення χ2 використовують **критерій В.І. Романовського** (запропонований радянським математиком Всеволодом Івановичем Романовським в 1928 р.):

Параметричні тести можна застосовувати також при **множинних порівняннях**, тобто при порівнянні двох груп вибірок одна з одною. Кожну групу задають подібно до того, як задають параметри масивів даних у методах двофакторного дисперсійного аналізу. При множинних порівняннях використовують багатовимірні узагальнення тестів, що були розглянуті вище.

**Непараметричні тести**

У багатьох випадках емпіричні дані не задовольняють нормальний розподіл. Тому для їх аналізу некоректно застосовувати параметричні тести. Серед непараметричних тестів важливе місце займають так звані **робастні методи**, що виявляють слабку чутливість до відхилень від стандартних умов і можуть використовуватися в широкому діапазоні реальних умов.

При перевірці нульової гіпотези про однорідність вибірок числових даних рекомендується використовувати омега-квадрат критерій або (за відсутності необхідних таблиць та програмного забезпечення) критерій Смірнова.

**Критерій омега-квадрат (критерій Крамера – фон Мізеса)** базується на розгляді відхилення між двома емпіричними функціями розподілу (або між емпіричною й теоретичною функціями розподілу при ідентифікації закону розподілу). Вперше його запропонували у 1928–1930 р. шведський математик Карл Харальд Крамер та американський математик і механік Ріхард фон Мізес, який народився у Львові.

Двохвибірковий варіант критерію запропоновано в 1951 р. американським статистиком Еріхом Лео Леманом та досліджено у 1952 р. американським математиком Мюреєм Розенблаттом. Тому його іноді називають також **критерієм Лемана – Розенблатта**. На сьогодні його вважають найпотужнішим з критеріїв, призначених для перевірки гіпотези про однорідність незалежних вибірок.

**Критерій Смірнова** запропонований в 1939 р. радянським математиком Миколою Васильовичем Смірновим. Він призначений для перевірки гіпотези про однорідність двох вибірок з неперервним законом розподілу.

Нульову гіпотезу відхиляють, якщо розрахункове значення критерію перевищить критичне для відповідного рівня значущості.

Для незалежних вибірок можна застосовувати **критерій рандомізації компонент**. Він розроблений Р. Фішером у 1920 р. для аналізу вибірок малого обсягу.

При порівнянні спряжених вибірок основою методу є перебирання можливих результатів, побудованих з різницевих оцінок. Нульова гіпотеза полягає у рівності вибіркових середніх.

**W-критерій Уїлкоксона (критерій рангових сум)** запропонований в 1945 р. американським хіміком і статистиком Френком Уїлкоксоном. Його застосовують для порівняння двох незалежних сукупностей за їх центральною тенденцією, тобто за центрами емпіричних функцій розподілу. Сукупності можуть мати як однакові, так і різні чисельності. Критерій оперує не числовими значеннями даних, а їх рангами – місцями у впорядкованих за згасанням або зростанням рядах даних. При його застосуванні передбачається, що розподіли вибірок є неперервними, а нульова гіпотеза полягає в тому, що функції розподілу вибірок збігаються одна з одною.

Процедура обчислення значення критерію є близькою до обчислення критерію рандомізації компонент. Різниця полягає в тому, що замість вихідних даних використовують їх ранги. Ранжирування порівнюваних вибірок здійснюють сумісно. Вихідні дані об’єднують до однієї вибірки, впорядковують, визначають ранги елементів об’єднаної вибірки. Потім формують дві нові вибірки, елементами яких є ранги відповідних елементів вихідних вибірок. Якщо деякі значення збігаються, то відповідним спостереженням призначають середній ранг.

При застосуванні W-критерію Уїлкоксона слід мати на увазі, що згідно з він належить до так званих критеріїв зсуву. Тобто найбільш потужним він є при виявленні різниці, пов’язаної з тим, що одну з вибірок отримано додаванням одного й того самого числа до всіх елементів іншої вибірки.

Він є нечутливим до різниці дисперсій порівнюваних вибірок, а також коефіцієнтів їх асиметрії та ексцесу. Зокрема, якщо дві вибірки мають симетричні функції розподілу з однаковими середніми значеннями, але різними стандартними відхиленнями, то в об’єднаної послідовності елементи однієї вибірки матимуть підвищену кількість елементів з високими та низькими рангами. Елементи іншої вибірки будуть мати підвищену кількість елементів із середніми значеннями рангів. Але суми рангів усіх елементів для цих двох вибірок можуть бути приблизно однаковими.

**U-критерій Манна – Уїтні** запропонований в 1947 р. американськими математиками Г.Б. Манном та Д.Р. Уїтні. Він призначений для перевірки нульової гіпотези про однаковість розподілу досліджуваних сукупностей або для перевірки рівності окремих параметрів цих розподілів, наприклад, середніх значень. Спостереження мають бути непарними. Цей критерій є найпотужнішим непараметричним аналогом t-критерію Стьюдента для незалежних вибірок. У деяких випадках його потужність може бути навіть більшою, ніж у t-критерію.

**Т-критерій Уїлкоксона (одновибірковий, знаковий ранговий критерій Уїлкоксона)** запропонований Ф. Уїлкоксоном в 1945 р. Його застосовують для порівняння вибірок з попарно спряженими значеннями. Він є непараметричним аналогом t-критерію Стьюдента для спряжених вибірок. Перевіряють нульову гіпотезу про симетричність розподілу різниць спряжених значень стосовно нуля. Методика розрахунку є близькою до розрахунку значення W-критерію Уїлкоксона, але в цьому випадку оперують модулями різниць відповідних значень.

Критерій χ2 (хі-квадрат) запропонований в 1900 р. видатним британським математиком, біологом та філософом Карлом Пірсоном. Його використовують для перевірки нульової гіпотези про однаковість розподілу досліджуваних випадкових величин. Його широко застосовують у дисперсійному аналізі та інших методах аналізу даних. Цей критерій оперує не первинними даними, а їх розподілом за класами. З огляду на це необхідно враховувати вимогу щодо мінімальних обсягів вибірок та кількостей 59 класів. За різними оцінками, мінімальна допустима кількість класів знаходиться у межах 4–7, а кількість елементів у вибірках – у межах 20–40.

Критерій χ2 можна застосовувати також і для порівняння вибірок значень номінальних ознак. У цьому випадку аналізують дані, подані у вигляді таблиці спряженості ознак. Елементами таблиці є числа, рівні кількостям елементів досліджуваних вибірок, для яких досліджувана ознака набуває значень, котрі відповідають певному класу. Кожний рядок таблиці характеризує розподіл елементів відповідної вибірки за класами, а кожний стовпець – наповненість певного класу в різних вибірках.

**Критерій серій Вальда – Волфовиця** розроблений в 1940 р. американськими математиками Абрахамом Вальдом і Джекобом Волфовицем. Його використовують для перевірки нульової гіпотези про те, що дві незалежні випадкові вибірки обсягами n1 та n2 не відрізняються одна від одної за досліджуваною ознакою.

Результати спостережень записують як варіаційний ряд об’єднаної вибірки, а їх належність до вихідних вибірок помічають за допомогою додаткової змінної, яка може набувати два значення, наприклад «0» та «1». Послідовність її значень називають **послідовністю кодів**.

**Серією** послідовності кодів називають будь-яку послідовність її однакових значень. Наприклад, у послідовності 00101111011 є такі серії: 00, 1, 0, 1111, 0, 11. Очевидно, що за умови справедливості нульової гіпотези кількість серій N має бути великою, а за умови її помилковості – відносно малою.

**Критерій знаків** запропонований Ф. Уїлкоксоном. Його використовують для перевірки нульової гіпотези про однорідність двох спряжених вибірок. Нехай хі і yі – значення відповідних елементів цих вибірок. Якщо вибірки є однорідними, то ймовірності появи додатних та від’ємних різниць xi − yі є рівними. Імовірність появи нульових значень цих різниць вважається нульовою, оскільки передбачається, що розподіл досліджуваної ознаки є неперервним. Якщо внаслідок випадкових похибок або округлення результатів такі різниці з’являються, то відповідні спостереження виключають з подальшого аналізу. За умови справедливості нульової гіпотези ймовірність р появи знаків певного типу (наприклад знаків «+») різниць xi − yі підпорядковується біноміальному розподілу з параметрами: p =1/ 2; m, де m – кількість різниць, що аналізують.

# ***5. ДИСПЕРСІЙНИЙ АНАЛІЗ***

Дисперсійний аналіз є сукупністю статистичних методів, призначених для перевірки гіпотез про зв’язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також для встановлення ступеня впливу факторів та їх взаємодії. У спеціальній літературі дисперсійний аналіз часто називають ANOVA (від англомовної назви Analysis of Variations). Вперше цей метод було розроблено Р. Фішером в 1925 р.

**Факторами** називають контрольовані чинники, що впливають на кінцевий результат. **Рівнем фактора**, або **способом обробки**, називають значення, що характеризують конкретний прояв цього фактора. Ці значення зазвичай подають у номінальній або порядковій шкалі вимірювань. Значення вимірюваної ознаки називають **відгуком**.

Часто вихідні значення факторів вимірюють у кількісних або порядкових шкалах. Тоді постає проблема групування вихідних даних у ряди спостережень, що відповідають приблизно однаковим значенням фактора. Якщо кількість груп взяти надмірно великою, то кількість спостережень у них може виявитися недостатньою для отримання надійних результатів. Якщо її взяти надмірно малою, це може призвести до втрати суттєвих особливостей впливу досліджуваного фактора на систему. Загальну методологію групування описано в розділі 1. Вибір конкретного способу групування даних залежить від їх обсягу і характеру варіювання значень фактора.

Кількість і розміри інтервалів при однофакторному аналізі найчастіше визначають за принципом рівних інтервалів або за принципом рівних частот. При багатофакторному аналізі застосовують три типи групування:

* групи з рівною кількістю спостережень;
* групи з різною кількістю спостережень;
* групи, кількості спостережень у яких відповідають певній пропорції.
* При цьому існують певні особливості обробки даних, залежно від
* типу групування, які не розглядаються у цьому посібнику.

***Однофакторний аналіз***

Основною метою однофакторного аналізу зазвичай є оцінка величини впливу конкретного фактора на досліджуваний відгук. Іншою метою може бути порівняння двох або декількох факторів один з одним з метою визначення різниці їх впливу на відгук, яку часто називають **контрастом факторів**. Попереднім етапом є перевірка нульової гіпотези про відсутність будь-якого впливу досліджуваного фактора (факторів), тобто гіпотези про те, що зміни значень ознаки в порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності.

Якщо нульову гіпотезу відкидають, то наступним етапом є кількісне оцінювання впливу досліджуваного фактора і побудова довірчих інтервалів для отриманих характеристик. У випадку, коли нульова гіпотеза не може бути відкинутою, зазвичай її приймають і роблять висновок про відсутність впливу. Але, якщо є підстави вважати, що такий вплив має бути присутнім (наприклад, це може випливати з теоретичних уявлень про об’єкт дослідження), то необхідно перевірити наявність інших факторів, що можуть його маскувати.

При **однофакторному дисперсійному аналізі** вихідні дані подають у вигляді таблиць, у яких кількість стовпчиків дорівнює кількості рівнів фактора, а кількість значень у кожному стовпчику - кількості спостережень при відповідному рівні фактора (табл. 1.2). Для різних рівнів фактора кількість спостережень може бути різною. При цьому виходять з припущення, що результати спостережень для різних рівнів є вибірками з нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими і не залежать від рівнів. Завданням аналізу є перевірка нульової гіпотези про рівність середніх значень сукупностей, що розглядаються.

Таблиця 1.2

Форма таблиці спостережень при проведенні однфакторного дисперсійного аналізу

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Результати вимірювань | Рівні фактора | | | |
| 1 | 2 | … | к |
| 1 |  |  | … |  |
| 2 |  |  | … |  |
| … | … | … | … | … |
|  |  |  | … |  |

Метод базується на основній тотожності дисперсійного аналізу, згідно з якою сума квадратів відхилень спостережень від загального середнього (**загальна варіація**) дорівнює:

(1.14)

де – загальне середнє; ; N– загальна чисельність; k – кількість вибірок; – кількість елементів у j-й вибірці; – середнє значення j-ї вибірки.

У правій частині (1.14) перший доданок (**факторна**, або **міжгрупова** **варіація**) є зваженою сумою квадратів відхилень групових середніх від загального середнього. Він характеризує коливання значень, зумовлені фактором, на основі якого здійснено групування даних. Другий доданок (**залишкова**, або **внутрішньогрупова варіація**) є сумою квадратів відхилень спостережень від відповідних групових середніх. Він характеризує коливання значень досліджуваної ознаки, зумовлені неврахованими факторами або випадковими чинниками.

Сутність методу полягає в тому, що за умови правильності нульової гіпотези величини

(1.15)

та

(1.16)

є незміщеними оцінками дисперсії похибок спостережень і мають бути приблизно рівними одна одній. Перша з них є мірою варіації всередині вибірок і не пов’язана з припущенням про рівність середніх значень, тому незалежно від справедливості нульової гіпотези. Друга оцінка характеризує варіацію між вибірками. При справедливості нульової гіпотези , а при її порушенні неї величина є тим більшою, чим більше відхилення від неї. Значення критерію розраховують за формулою:

(1.17)

Ця величина має F-розподіл Фішера з параметрами k −1 та N – k. Нульову гіпотезу відхиляють, якщо ймовірність P (F ≥ F\*), де F\* – значення, розраховане за емпіричними даними за формулою (1.16) є достатньо малою.

Непараметричним аналогом однофакторного дисперсійного аналізу є **ранговий однофакторний аналіз Краскела – Уолліса**. Він розроблений американськими математиком Вільямом Краскелом та економістом Вільсоном Уоллісом в 1952 р. Цей критерій призначено для перевірки нульової гіпотези про рівність ефектів впливу на досліджувані вибірки з невідомими, але рівними середніми. При цьому кількість вибірок має бути більшою ніж дві. Нульова гіпотеза полягає в тому, що k вибірок обсягами 1 2 , , ..., k nn n отримані з однієї і тієї самої генеральної сукупності. Критерій Краскела – Уолліса є узагальненням U-критерію Манна – Уїтні на випадок, коли кількість вибірок k > 2.

Рангові методи, у тому числі й метод Краскела – Уолліса, не передбачають нормальності розподілу результатів спостережень і можуть застосовуватися як для кількісних даних з невідомим законом розподілу, так і для порядкових ознак.

**Критерій Джонкхієра (Джонкхієра – Терпстра)** запропонований незалежно один від одного нідерландським математиком Т.Дж. Терпстрою в 1952 р. й британським психологом Е.Р. Джонкхієром в 1954 р. Його застосовують тоді, коли заздалегідь відомо, що наявні групи результатів упорядковані за зростанням впливу досліджуваного фактора, який вимірюють у порядковій шкалі. Таблиця даних має такий самий вигляд, як і в попередньому випадку. Будемо вважати, що її перший стовпчик відповідає найменшому рівню фактора, другий – наступному за величиною тощо, останній стовпчик відповідає найбільшому рівню. При виконанні таких припущень критерій Джонкхієра є більш потужним, ніж критерій Краскела – Уолліса, стосовно гіпотези про монотонний вплив фактора.

**М-критерій Бартлетта** запропонований британським статистиком Маурісом Стівенсоном Бартлеттом в 1937 р. Його застосовують для перевірки нульової гіпотези про рівність дисперсій кількох нормальних генеральних сукупностей, з яких взяті досліджувані вибірки, що у загальному випадку мають різні обсяги (обсяг кожної вибірки має бути не менше чотирьох).

**G-критерій Кокрена (Кочрена)** запропонований американським статистиком Вільмом Геммелом Кочреном в 1941 р. Його використовують для перевірки нульової гіпотези про рівність дисперсій k (k ≥ 2) нормальних генеральних сукупностей за незалежними вибірками рівного обсягу.

Непараметричний **критерій Левене**, запропонований американським математиком Ховардом Левене в 1960 р. є альтернативою критерію Бартлетта в умовах, коли немає впевненості у тому, що досліджувані вибірки підпорядковуються нормальному розподілу.

В 1974 р. американські статистики Мортон Б. Браун та Алан Б. Форсайт запропонували більш робастний тест (**критерій Брауна – Форсайта**).

Розглянуті вище критерії дають змогу встановити різницю дисперсій сукупностей, але не дають можливості дати кількісну оцінку впливу фактора на досліджувану ознаку, а також встановити, для яких саме сукупностей дисперсії є різними.

Для встановлення кількісного впливу досліджуваного фактора часто застосовують адитивну модель, яка передбачає, що значення відгуку є сумою впливу фактора і незалежної від нього випадкової величини.

Якщо гіпотезу про рівність середніх відхиляють, то наступним кроком може бути визначення вибірок, для яких ця різниця є суттєвою. Для цього використовують метод лінійних контрастів.

**Лінійним контрастом** у моделі адитивного впливу фактора на відгук називають лінійну функцію середніх значень k незалежних нормальних вибірок з невідомими рівними дисперсіями.

Для встановлення вибірок, що належать певній множині даних, дисперсії яких є різними, найчастіше застосовують **метод множинних порівнянь (Шеффе)**, запропонований американським статистиком Генрі Шеффе.

**Двофакторний аналіз**

Двофакторний дисперсійний аналіз застосовують для пов’язаних нормально розподілених вибірок. Дані подають у вигляді табл. 1.3, у стовпчиках якої наводять дані, що відповідають певному рівню першого фактора, а в рядках – дані, що відповідають рівням другого.

Таблиця 1.3

Таблиця даних двофакторного дисперсійного аналізу

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Рівні фактора В | Рівні фактора А | | | |
| 1 | 2 | … | k |
| 1 |  |  | … |  |
| 2 |  |  | … |  |
| ... | … | … | … | … |
| n |  |  | … |  |

Основною відмінністю від таблиці однофакторного дисперсійного аналізу є можлива неоднорідність даних у стовпцях, якщо вплив другого фактора є суттєвим. На практиці часто використовують і складніші таблиці двофакторного дисперсійного аналізу, зокрема такі, у яких кожна комірка містить набір даних (повторні вимірювання), що відповідають фіксованим значенням рівнів обох факторів.

У випадку, коли випадкова компонента εij підпорядковується нормальному розподілу з нульовим середнім і рівними для всіх i, j дисперсіями σ2 застосовують **двофакторний дисперсійний аналіз (дисперсійний аналіз за двома ознаками).**

Якщо припущення, необхідні для застосування двофакторного дисперсійного аналізу, не виконуються, то використовують непараметричний **ранговий критерій Фрідмана (Фрідмана, Кендалла та Сміта)**, розроблений американським економістом Мілтоном Фридманом наприкінці 1930р. Цей критерій не залежить від типу розподілу. Передбачається лише, що розподіл величин ij ε є однаковим і неперервним, а самі вони незалежні одна від одної.

У разі, коли в рядках вихідної таблиці є однакові значення, необхідно використовувати середні ранги. При цьому точність висновків буде тим гіршою, чим більшою є кількість таких збігів.

**Q-критерій Кокрена** запропонований В. Кочреном в 1937 р. Його використовують у випадках, коли групи однорідних суб’єктів піддаються впливам, кількість яких перевищує два, і для яких можливі два варіанти відгуків – умовно-негативний (0) та умовно-позитивний (1). Нульова гіпотеза полягає в рівності ефектів впливу.

Двофакторний дисперсійний аналіз дає можливість визначити існування ефектів обробки, проте не дає змоги встановити, для яких саме стовпців існує цей ефект.

При вирішенні цієї проблеми застосовують метод множинних порівнянь Шеффе для пов’язаних вибірок.

***Лекція 1.2. (2 год)***

Тема 1. Статистичні методи аналізу даних

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Кореляційний аналіз
2. Факторний аналіз
3. Побудови й дослідження регресійних моделей

# ***1. КОРЕЛЯЦІЙНИЙ АНАЛІЗ***

**Кореляцією (кореляційним зв’язком)** між випадковими величинами (ознаками) називають наявність статистичного або ймовірнісного зв’язку між ними. При цьому закономірна зміна певних ознак призводить до закономірної зміни середніх значень інших, пов’язаних з ними ознак. **Кореляційним аналізом** називають сукупність методів виявлення кореляційного зв’язку. Тому його можна застосовувати для формалізованого подання моделей зв’язків між окремими компонентами системи або між окремими процесами, що відбуваються в ній. Наявність кореляційного зв’язку не означає існування причинно-наслідкового зв’язку між досліджуваними ознаками. Вона може бути зумовлена тим, що обидві ознаки мають причинно-наслідковий зв’язок з певним іншим фактором. Наприклад, існує кореляція між цінами на нафту й на золото. Проте вона пояснюється тим, що обидві ціни виражаються у доларах США й залежать від динаміки його індексу. Кореляція також може бути випадковою.

Сучасну класифікацію мір подібності запропонували австрійський та американський біостатистик та антрополог Роберт Сокал та британський таксономіст Пітер Сніс у 1963 р. Згідно з нею виокремлюють такі типи мір подібності:

– міри асоціації, що відбивають різні співвідношення кількості ознак, що збігаються до загальної кількості ознак, а також близькі до них коефіцієнти спряженості (квантифіковані коефіцієнти зв’язку);

– вибіркові коефіцієнти зв’язку типу кореляції (нормовані косинусні міри);

– показники відстані у метричному просторі.

Перевірку зв’язку можна здійснювати лише для пов’язаних вибірок. Це означає, що між елементами обох досліджуваних вибірок існує взаємно однозначна відповідність, а кількість елементів у вибірках є однаковою.

Замість гіпотези про наявність кореляційного зв’язку часто розглядають протилежну гіпотезу про відсутність зв’язку між досліджуваними величинами. Нехай ознака А має r рівнів А1, А2, … Аr, а ознака В – s рівнів В1, В2, … Вs. Їх вважають незалежними, якщо події “ознака А набуває значення Аі” та “ознака В набуває значення Вj” є незалежними для всіх можливих пар і, j, тобто:

(1.18)

Це можна сформулювати в інший спосіб: ознаки є незалежними, якщо значення ознаки А не впливає на ймовірності реалізації можливих значень ознаки В:

(1.19)

Кореляційний аналіз здійснюють на початковому етапі вирішення всіх основних проблем статистичного аналізу даних. У проблемі статистичного аналізу залежностей і побудови регресійних моделей він дає змогу встановити сам факт існування зв’язку між змінними та оцінити ступінь його прояву. У проблемі класифікації даних за допомогою кореляційного аналізу отримують вихідну інформацію у вигляді коваріаційних і кореляційних матриць та інших характеристик парних порівнянь. Це дає змогу визначити подібні один до одного або до певних еталонів об’єкти, сформувати класи подібних об’єктів і здійснити класифікацію. У проблемі зменшення розмірності досліджуваного простору ознак також за допомогою коваріаційних і кореляційних матриць визначають ознаки, що можуть бути без втрати суттєвої інформації подані через інші наявні дані.

Загальна методика перевірки гіпотези про існування зв’язку між ознаками передбачає три основних етапи: визначення типу даних; перевірку гіпотези про відсутність зв’язку і, в разі її відхилення, оцінювання сили зв’язку. Тип вихідних даних суттєво впливає на вибір методів і критеріїв, які можна застосовувати на наступних етапах аналізу.

Для визначення сили зв’язку використовують різноманітні показники. Зазвичай їх прагнуть вибрати такими, щоб вони змінювалися від –1 до +1 або від 0 до 1. Значення, що є близькими за модулем до одиниці, свідчать про наявність сильного зв’язку. Близькі до нуля значення вказують або на відсутність будь-якого зв’язку, або на відсутність зв’язку того типу (найчастіше лінійного), для якого розроблено відповідний коефіцієнт. Знак коефіцієнта вказує на напрям зв’язку: прямий (для додатних значень) або зворотний (для від’ємних).

***Кореляційний аналіз кількісних ознак***

Методику кількісного оцінювання кореляції між ознаками вперше було запропоновано британським географом, антропологом та психологом Френсисом Гальтоном в 1888 р. Універсальною характеристикою ступеня тісноти зв’язку між кількісними ознаками є коефіцієнт детермінації.

Величина коефіцієнта детермінації може змінюватися в межах від нуля до одиниці й відображає частку загальної дисперсії досліджуваної ознаки, яка зумовлена зміною функції регресії f (X). При цьому нульове значення коефіцієнта детермінації відповідає відсутності будь-якого зв’язку, а його рівність одиниці – наявності строго функціонального зв’язку. Оскільки цей коефіцієнт є універсальним показником зв’язку, він має відбивати й такі зв’язки, що є немонотонними функціями. Тому питання напряму зв’язку у цьому випадку не має сенсу.

Слід зазначити, що для обмеженого набору даних часто можна побудувати декілька різних адекватних регресійних моделей. Групування даних також можна здійснювати різними способами. Тому існує певна невизначеність коефіцієнтів детермінації: при застосуванні різних регресійних моделей або різних способів групування ми будемо отримувати дещо різні значення коефіцієнта детермінації.

Інші поширені характеристики ступеня тісноти зв’язку між ознаками можна розглядати як окремі випадки коефіцієнта детермінації, отримані для конкретних математичних моделей зв’язку.

Розрізняють парні та частинні кореляційні характеристики. Парні характеристики розраховують за результатами вимірювань тільки досліджуваної пари ознак. Тому вони не враховують опосередкованого або спільного впливу інших ознак. Частинні характеристики є очищеними від впливу інших факторів, але для їх розрахунку необхідно мати вихідну інформацію не тільки про досліджувані ознаки, а й про всі інші, вплив яких необхідно усунути.

Для кількісних ознак найчастіше застосовують коефіцієнти кореляції Пірсона і Фехнера. **Коефіцієнт кореляції Пірсона (коефіцієнт кореляційного відношення Пірсона, парний коефіцієнт кореляції, вибірковий коефіцієнт кореляції, коефіцієнт Бравайса – Пірсона)** вимірює ступінь лінійного кореляційного зв’язку між кількісними скалярними ознаками. Він був запропонований К. Пірсоном у 1896 р. Часто, посилаючись на згадування К. Пірсона про ідеї математичного подання зв’язку, висловлені в 1846 р. відомим французьким фізиком та кристалографом Огюстом Браве, цей показник називають коефіцієнтом Бравайса – Пірсона (Бравайс – це викривлена транскрипція від французького Bravais, що закріпилася в літературі з кореляційного аналізу).

Застосування коефіцієнта Пірсона як міри зв’язку є обґрунтованим лише за умови, що спільний розподіл пари ознак є нормальним. Тому перед його розрахунком слід перевірити виконання цієї гіпотези. Якщо вона справедлива, то квадрат коефіцієнта кореляції Пірсона дорівнює коефіцієнту детермінації.

Значення коефіцієнта кореляції може змінюватися від −1 до +1. Значення −1 та +1 відповідають чіткій лінійній функціональній залежності, яка в першому випадку є спадною, а у другому – зростаючою. Для функціональної залежності y const = коефіцієнт кореляції, як видно з наведеної формули, є невизначеним, оскільки в цьому випадку знаменник дорівнює нулю. Що ближчим є значення коефіцієнта кореляції до −1 або +1, то більш обґрунтованим є припущення про наявність лінійного зв’язку. Наближення його значення до нуля свідчить про відсутність лінійного зв’язку, але не є доказом відсутності статистичного зв’язку взагалі.

На рис. 1.5. показано дві серії точок, координати яких відповідають двом парам спряжених вибірок.

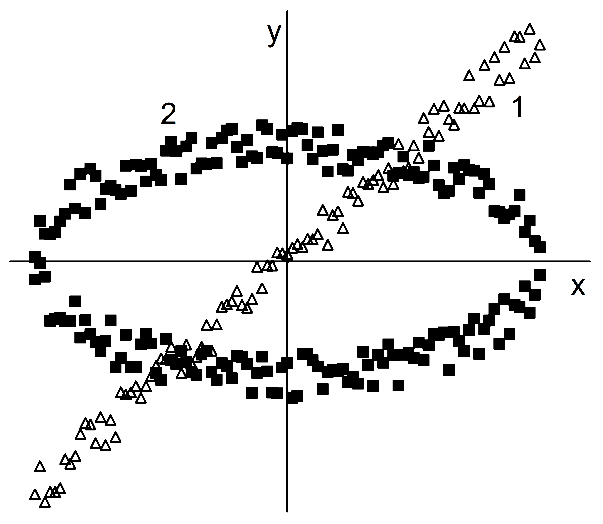


Рис. 1.5. Графічне зображення двох наборів тестових даних

Для обох пар вибірок є очевидним існування статистичного зв’язку між параметрами х та у. Але коефіцієнти кореляції для них дорівнюють, відповідно, r1 = 0,995 і r2 = 0,006 . Близькість коефіцієнта кореляції до нуля для другої пари вибірок пов’язана не з відсутністю зв’язку, а з його нелінійністю. Для порівняння, коефіцієнти детермінації для тих самих пар вибірок дорівнюють 0,98 та 1,00.

Показаний приклад свідчить, що в багатьох випадках для попереднього аналізу припущення про наявність і тип зв’язку між певними ознаками доцільно нанести наявні дані на графік.

Як видно, близькість коефіцієнта кореляції Пірсона до нуля в загальному випадку не є доказом незалежності ознак. Але можна довести, що у випадку, коли сумісний розподіл випадкових величин (x, y) є нормальним, рівність r = 0 свідчить про статистичну незалежність x і y.

Коефіцієнт кореляції Пірсона часто розглядають як універсальну міру кореляційного зв’язку. У багатьох пакетах загального призначення, зокрема в електронних таблицях MS Excel, не передбачено інших засобів його вимірювання. Але, як випливає з наведених вище даних, насправді сфера його обґрунтованого застосування є досить вузькою, оскільки лінійність залежності й нормальний розподіл даних навколо неї є скоріше винятком, ніж правилом.

При дослідженні багатовимірних сукупностей випадкових величин із коефіцієнтів кореляції, обчислених попарно між ними, можна побудувати квадратну симетричну кореляційну матрицю з одиницями на головній діагоналі. Вона є основним елементом при побудові багатьох алгоритмів багатовимірної статистики, наприклад у факторному аналізі.

Коефіцієнт кореляції Пірсона можна застосовувати для перевірки гіпотези про значущість зв’язку.

У випадку, коли між двома наборами ознак існує нелінійний зв’язок, для оцінювання ступеня його тісноти часто використовують кореляційне відношення, яке було запропоновано К. Пірсоном. Це можливо, якщо щільність розміщення емпіричних точок на координатній площині дає можливість їх групування за однією із змінних і підрахунку групових середніх значень другої змінної для кожного інтервалу

**Кореляційний аналіз порядкових ознак**

Під ранговою кореляцією розуміють статистичний зв’язок між порядковими ознаками. Вихідні дані зазвичай подають у вигляді табл. 1.4., де елемент xik є рангом і-го об’єкта за k-ю властивістю.

Таблиця 1.4.

Таблиця вихідних даних для рангового кореляційного аналізу

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Порядковий номер об’єкта | Порядковий номер досліджуваної ознаки | | | | | | |
| 0 | 1 | 2 | … | k | … | р |
| 1 |  |  |  | … |  | … |  |
| 2 |  |  |  | … |  | … |  |
| ... | … | … | … | … | … | … | … |
| і |  |  |  | … |  | … |  |
| … | … | … | … | … | … | … | … |
| п |  |  |  | … |  | … |  |

Завданнями аналізу в цьому випадку можуть бути: вивчення структури досліджуваних об’єктів; перевірка сукупної узгодженості ознак та умовне ранжирування об’єктів за ступенем тісноти зв’язку кожної з них з іншими ознаками; побудова єдиного групового впорядкування об’єктів (задача регресії на порядкових змінних).

У першому випадку кожну послідовність впорядкованих за k-ю ознакою n об’єктів подають як у n-вимірному просторі ознак. Найхарактернішими типами структури є такі.

1. Аналізовані точки рівномірно розкидані по всій області їх можливих значень. Це означає відсутність будь-якого зв’язку між досліджуваними ознаками.

2. Частина точок утворює ядро (кластер) із точок, що розташовані близько одна до одної, а інші випадково розкидані навколо цього ядра. Це відповідає існуванню підмножини узгоджених ознак.

3. Аналізовані точки утворюють декілька кластерів, розташованих відносно далеко один від одного. Це відповідає наявності декількох таких підмножин ознак, що існує істотний статистичний зв’язок між ознаками, які належать до однієї і тієї самої підмножини, і не існує значущого зв’язку між ознаками, які належать до різних підмножин.

Прикладом завдання другого типу є визначення узгодженості думок групи експертів з наступним впорядкуванням їх за рівнем компетентності. Для цього розраховують коефіцієнти конкордації для різних сукупностей досліджуваних змінних. Вирішення завдань третього типу зводиться до побудови такого впорядкування, яке б у певному значенні було б найближчим до кожного з наданих впорядкувань досліджуваних ознак. Для цього часто застосовують середнє арифметичне або медіану наявних базових рангів. Це можна розглядати як задачу найкращого у певному розумінні відновлення невідомого ранжирування за наявними емпіричними даними, що зумовлює можливість її розгляду як задачі регресії.

**Коефіцієнт рангової кореляції Спірмена (показник кореляції рангів Спірмена, коефіцієнт кореляції рангів)** запропонований британським психологом Чарльзом Едвардом Спірменом у 1904 р. Його використовують, якщо досліджується зв’язок між рядами даних, виміряними за порядковою шкалою. Його можна застосовувати також і для кількісних даних, але, як правило, це буває недоцільним. У найпростішому випадку досліджувані об’єкти класифікують за двома ознаками.

Коефіцієнт рангової кореляції Спірмена можна застосовувати як показник некорельованості вибірок.

Інший підхід використовує як міру подібності двох вибірок мінімальну кількість перестановок сусідніх об’єктів, потрібну для переведення послідовності рангів однієї вибірки до послідовності рангів іншої. Можна показати, що вона дорівнює кількості інверсій в однієї з цих послідовностей у випадку, коли інша послідовність впорядкована за зростанням.

Як і в попередньому випадку, кількість інверсій залежить від обсягу вибірки і є незручною для застосування як показника кореляції. Для цього використовують коефіцієнт рангової кореляції Кендалла (коефіцієнт кореляції рангів, ранговий коефіцієнт кореляції). Він був запропонований британським статистиком Маурисом Кендаллом у 1938 р.

Коефіцієнт рангової кореляції Кендалла призначений для визначення сили кореляційного зв’язку між двома рядами даних за тих самих умов, що і коефіцієнт рангової кореляції Спірмена. Як і для коефіцієнта Спірмена, його значення можуть змінюватися в межах від −1 до +1, при цьому −1 відповідає повній протилежності послідовностей рангів, а +1 – їх повному збігу. Слід зазначити, що обчислення коефіцієнта Кендалла є більш трудомістким, але з іншого боку, він має ряд переваг порівняно із коефіцієнтом Спірмена. Основними з них є такі:

– кращий рівень вивченості його статистичних властивостей, зокрема його вибіркового розподілу;

– можливість його застосування для визначення частинної кореляції;

– більша зручність перерахунку при додаванні нових даних.

**Кореляційний аналіз номінальних ознак**

Типовою ситуацією, коли необхідна перевірка зв’язку між номінальними ознаками, є обробка результатів соціологічних досліджень, що можуть містити такі комбінації ознак, як освіта, стать, професія, підтримка певної політичної партії, регіон проживання тощо.

При дослідженні зв’язків між категоризованими ознаками вихідні дані подають у вигляді таблиці спряженості (табл. 1.5.). До категоризованих зараховують номінальні ознаки, а також порядкові ознаки, для яких є відомим скінченний набір можливих градацій.

Величини fij показують, скільки разів зустрічалася комбінація ознак, за якої рівень першої має значення i, а рівень другої – j; mj є сумами стовпців, а ni – сумами рядків. За даними табл. 1.5. можна оцінити значення ймовірностей, що входять до формули (4.1):

Таблиця 1.5.

Таблиця спряженості категоризованих ознак

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Рівні ознаки 1 | Рівні ознаки 2 | | | | Разом |
| 1 | 2 | … | r |
| 1 |  |  | … |  |  |
| 2 |  |  | … |  |  |
| ... | … | … | … | … | … |
| c |  |  | … |  |  |
| Разом |  |  | … |  |  |

Нульову гіпотезу про відсутність зв’язку відхиляють, якщо різницю між ними й частотами, що спостерігаються, не можна пояснити випадковими чинниками. Як критерій можна використовувати величину .

На практиці частіше використовують **φ-коефіцієнт Пірсона**, або середньоквадратичну спряженість, яка може змінюватися від нуля до min {r-1, c-1 }.

Існує велика кількість показників ступеня тісноти статистичного зв’язку, призначених для категоризованих змінних, які не є універсальними, а відображають окремі властивості такого зв’язку.

**Коефіцієнт спряженості Крамера** був запропонований К.Х. Крамером у 1946 р. Він змінюється в межах від нуля до одиниці. При цьому значення С = 0 свідчить про статистичну незалежність аналізованих ознак, а значення С = 1 – про можливість однозначного відтворення значень однієї ознаки за відомими значеннями другої.

**Поліхоричний коефіцієнт спряженості Чупрова** призначений для дослідження кореляції номінальних ознак у таблиці спряженості c × r . Він був уведений російським статистиком О.О. Чупровим у 1926 р.

Існує велика кількість коефіцієнтів, що характеризують кореляцію між ознаками у випадку, коли кожна з двох ознак може мати лише два рівні, які найчастіше відповідають наявності та відсутності ознаки. У цьому випадку таблиця спряженості має розмір 2×2.

**Коефіцієнт (показник подібності) Жаккара**, уведений в 1901 р. французьким геоботаніком Полем Жаккаром. Значення цього коефіцієнта можуть змінюватися в межах від нуля до одиниці.

**Простий коефіцієнт зустрічальності (показник подібності Сокала й Міченера)** запропонований Р. Сокалом та американським ентомологом Чарльзом Дунканом Міченером у 1958 р. Як і в попередньому випадку, значення коефіцієнта можуть змінюватися в межах від нуля до одиниці.

**Показник подібності Рассела і Рао** запропонували в 1940 р. американський епідеміолог Поль Ф. Рассел та індийський і британський ентомолог Т. Рамакришна Рао. Його значення також можуть змінюватися в межах від нуля до одиниці.

**Коефіцієнт спряженості Бравайса – Пірсона** (показник подібності Чупрова) був уведений О.О. Чупровим у 1923 р. Значення цього коефіцієнта може змінюватися в межах від −1 до +1. Від’ємні значення коефіцієнта спряженості означають, що із збільшенням імовірності прояву одної ознаки, зменшується імовірність прояву іншою.

Легко показати, що цей показник є окремим випадком ϕ-коефіцієнта Пірсона для таблиць 2×2.

**Коефіцієнт асоціації Юла** був уведений відомим британським статистиком Джорджем Удні Юлом у 1900 р.

**Коефіцієнт колігації Юла**, що також був запропонований Дж.У. Юлом в 1912 р. Він не має переваг порівняно з коефіцієнтом асоціації. Значення обох коефіцієнтів змінюються в межах від −1 до +1.

**Хеммінгова відстань (метрика Хеммінга)** H = a + d також може застосовуватися для визначення кореляції. Проте, як і коваріація, вона не є безрозмірною величиною і може набувати будь-яких невід’ємних значень (верхньою межею є загальна кількість спостережень n). Цей показник був уведений відомим американським математиком Ричардом Веслі Хеммінгом у 1950 р.

**Кореляційний аналіз змішаних ознак**

**Коефіцієнт Гауера** був запропонований британським статистиком Джоном Кліффордом Гауером у 1971 р. Його застосовують у тому випадку, коли досліджувані ознаки виміряні в різних шкалах.

Для дихотомічних ознак алгоритм підрахунку внеску ознаки і визначення вагових коефіцієнтів збігається з коефіцієнтом Жаккара. Для порядкових ознак алгоритм підрахунку внеску ознаки збігається з хеммінговою відстанню, узагальненою на порядкові змінні, а вагові коефіцієнти беруть рівними одиниці для всіх ознак.

**Бісеріальний коефіцієнт кореляції** запропоновано К. Пірсоном. Його призначено для дослідження кореляції в таблицях розміром 2×n, які є дихотоміями за певною номінальною ознакою і класифікаціями за номінальною або порядковою ознакою, що класифікується за q класами і може бути впорядкованою або невпорядкованою. Вихідний розподіл має бути двовимірним нормальним.

Значення бісеріального коефіцієнта кореляції можуть змінюватися від −1 до +1.

**Бісеріальний коефіцієнт кореляції за таблицею Келлі – Вуда** запропонований американським психологом Луїсом Л. Терстоуном (Louis L. Thurstone) в 1928 р.

У випадку, коли одна із змінних дихотомізована, а інша – виміряна в кількісній шкалі, обчислюють **точково-бісеріальний коефіцієнт кореляції**.

При цьому передбачається, що дихотомічна змінна може набувати лише два значення: 1 (верхній рівень) та 0 (нижній рівень). З погляду теорії точково-бісеріальну кореляцію можна розглядати як окремий випадок коефіцієнта кореляції Пірсона

**Множинна кореляція**

Про множинну кореляцію мова йде в тому випадку, коли певна ознака може бути пов’язана не з однією, а із сукупністю декількох інших ознак.

У реальних дослідженнях можлива ситуація, коли на певну ознаку може впливати не одна, а декілька інших. В таких випадках парні показники кореляції будуть давати неправильну інформацію щодо наявності зв’язку між відповідними показниками, оскільки ці їх значення будуть викривлятися невраховуваними ознаками.

Для уникнення помилок використовують частинні показники кореляції, що усувають такий вплив. Ідея введення таких показників вперше була висунута Г.У. Юлом у 1896 р., а пізніше розвинена ним та К. Пірсоном.

Якщо досліджувані ознаки задовольняють багатовимірний нормальний розподіл, застосовують частинний коефіцієнт кореляції між двома ознаками і та j при фіксованих значеннях інших ознак. Цей показник запропоновано К. Пірсоном у 1897 р.

Частинні коефіцієнти кореляції мають всі властивості парних коефіцієнтів кореляції. Вони є показниками наявності лінійного зв’язку між двома незалежними ознаками, який не залежить від впливу інших ознак. Тісноту зв’язку між декількома змінними у випадку множинної регресії можна оцінити за допомогою **коефіцієнта множинної кореляції**.

Крім того, застосовують множинний коефіцієнт кореляції, якій є мірою лінійної кореляції між певною змінною y та сукупністю величин Х1, Х2, …, Хn і визначається як звичайний парний коефіцієнт кореляції між y та множинною лінійною регресією за Х1, …, Хn. При цьому припускають, що досліджувана сукупність підпорядковується багатовимірному нормальному закону.

**Множинний коефіцієнт кореляції**, запропонований у 1935 р. Г. Хотелінгом, є окремим випадком коефіцієнтів канонічної кореляції.

Його також можна визначити за частинними коефіцієнтами кореляції.

Множинний коефіцієнт кореляції мажорує будь-який парний або частинний коефіцієнт кореляції, що характеризує статистичні зв’язки досліджуваної ознаки.

Подання математичних об’єктів називають **канонічним**, якщо кожному об’єкту однієї множини відповідає один і тільки один об’єкт іншої множини й ця відповідність є взаємно однозначною. **Канонічний кореляційний аналіз** здійснюють між двома сукупностями (групами) вибірок. Він призначений для визначення лінійної функції від перших p компонент і лінійної функції від q компонент, що залишилися, таких, щоб коефіцієнт кореляції між цими лінійними функціями набув найбільшого можливого значення. Чисельності груп (кількість вибірок у першій та другій групах, р та q) можуть різнитися, але необхідною умовою є рівна кількість варіант у всіх вибірках, що становлять обидві групи.

Розв’язання зводиться до узагальненої проблеми власних значень.

Квадратні корені з власних значень називають **канонічними кореляціями**.

**Коефіцієнт конкордації** призначений для дослідження зв’язків між порядковими ознаками, кількість яких є більшою, ніж два. Як міру узгодженості беруть суму квадратів відхилень сум рангів спостережень (об’єктів) від їх спільного середнього рангу.

**Коефіцієнт конкордації Кендалла (W-коефіцієнт Кендалла)** було запропоновано М. Кендаллом у 1939 р. Його значення може змінюватися в межах від нуля до одиниці, при цьому він дорівнює одиниці лише за умови, що всі досліджувані ранжирування збігаються. Коефіцієнт конкордації дорівнює нулю, якщо k ≥ 3 і всі ранжирування є випадковими впорядкуваннями вихідної вибірки.

***2. ФАКТОРНИЙ АНАЛІЗ***

При дослідженні складних систем часто немає можливості безпосередньо вимірювати величини, що визначають їх властивості (фактори). Більше того, нерідко є невідомими кількість та зміст цих факторів. Але можуть вимірюватися інші величини, що залежать від них. Якщо невідомий фактор впливає на декілька вимірюваних ознак, останні виявляють певний зв’язок, наприклад корельованість, між собою. Тому загальна кількість факторів може бути значно меншою, ніж кількість вимірюваних ознак. Для виявлення таких факторів використовують факторний аналіз. Зменшення кількості факторів може бути необхідним також для забезпечення збіжності алгоритмів подальшого аналізу даних, скорочення ресурсів пам’яті ЕОМ та часу, потрібних для їх обробки, бажанням візуалізувати отримані результати тощо.

Основні ідеї факторного аналізу було сформульовано Ф. Гальтоном наприкінці XIX ст. Пізніше значний внесок у розвиток його методології зробили Р. Кеттелл, К. Пірсон, Ч. Спірмен, Л. Терстоун, Г. Хотеллінг та інші фахівці.

Першим етапом факторного аналізу зазвичай є вибір нових ознак (факторів), які є лінійними комбінаціями старих і відображають переваж- 138 ну частку загальної мінливості вихідних даних. Тому вони зберігають основну частину інформації, що містили ці дані. Другим етапом є обертання факторів з метою спрощення їх інтерпретації.

Об’єктом дослідження методами факторного аналізу, як правило, є кореляційна матриця, побудована із застосуванням коефіцієнта кореляції Пірсона для кількісних ознак. Основною вимогою до цієї матриці є її додатна напіввизначеність. Згідно з умовами Сильвестра для цього достатньо, щоб усі її головні мінори були невід’ємними. З додатної напіввизначеності кореляційної матриці випливає невід’ємність усіх її власних значень.

Методами факторного аналізу вирішують три основні групи проблем:

– пошук передбачуваних неявних закономірностей, що визначаються впливом зовнішніх або внутрішніх чинників на досліджуваний процес;

– виявлення та вивчення статистичного зв’язку ознак з факторами або головними компонентами;

– стискування інформації шляхом подання процесу за допомогою узагальнених факторів або головних компонент, кількість яких є меншою за кількість обраних спочатку ознак (параметрів), але достатньою для забезпечення відтворення кореляційної матриці з потрібною точністю.

Розрізняють **R-техніку** та **Q-техніку** факторного аналізу. Перша з них розроблена британським психологом Реймондом Б. Кеттеллом і передбачає розрахунок коефіцієнтів кореляції між параметрами (ознаками), що утворюють матрицю вихідних даних. Її використовують для зменшення кількості параметрів. Q-техніку запропонував британський психолог В. Стефенсон в 1935–1936 р. й докладно описав Р.Б. Кеттелл у 1946 р. За її допомогою вивчають кореляцію між об’єктами або станами об’єктів. Її застосовують для зменшення кількості об’єктів. З формального погляду в першому випадку шукають кореляцію між стовпчиками таблиці спостережень (табл. 1.6.), а у другому – між її рядками.

Таблиця 1.6.

Загальний вигляд таблиці спостережень для факторного аналізу

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Номери об’єктів (станів) | Параметри об’єктів (станів) | | | |
| 1 | 2 | … | р |
| 1 |  |  | … |  |
| 2 |  |  | … |  |
| ... | … | … | … | … |
| n |  |  | … |  |

Крім того, розроблено P-техніку, яку використовують при дослідженні одного об’єкта, якщо значення його ознак вимірюють у різні моменти часу, а також О-техніку, S-техніку та Т-техніку, які також докладно були описані Р. Кеттеллом в 1940 р. Понад 50% усіх завдань, що передбачають застосування факторного аналізу, вирішують за допомогою R-техніки.

Основними методами факторного аналізу є методи головних компонент, головних факторів, максимальної правдоподібності та центроїдний. Усі вони ґрунтуються на припущенні, що досліджувана залежність є лінійною. Вихідні дані мають підпорядковуватися багатовимірному нормальному розподілу, але центроїдний метод є досить стійким до відхилень від такого закону.

Метою факторного аналізу є зменшення кількості змінних та визначення структури взаємозв’язків між змінними (класифікація даних). З формального погляду його метою є одержання матриці факторного відображення. Її рядки є координатами кінців векторів, що відповідають n змінним у p’-вимірному факторному просторі. Близькість цих векторів свідчить про взаємну залежність змінних. Якщо кількість факторів перевищує одиницю, зазвичай здійснюють обертання матриці факторного відображення для одержання більш простої її структури.

Однією з проблем, що виникають при застосуванні факторного аналізу, є необхідність знаходження власних значень кореляційної матриці. Якщо вона є виродженою, ця задача може виявитися нерозв’язною. Для матриць високого порядку може відбуватися втрата значущості у процесі обчислень. У певних випадках проблему виродженості можна зняти виключенням лінійно залежних параметрів.

Метод Якобі дає змогу визначити власні значення і для вироджених кореляційних матриць. Але при цьому частина їх, яка дорівнює різниці між порядком та рангом матриці, буде мати значення, що не перевищують обчислювальної похибки. Завдяки цьому метод головних компонент виявляється стійкішим до аналізу відповідних даних, ніж метод максимуму правдоподібності. Водночас він є гіршим за останній з погляду можливості отримання точної оцінки загальності й досягнення повного відтворення кореляційної матриці.

Обов’язковими умовами факторного аналізу є такі:

– всі досліджувані ознаки мають бути кількісними;

– кількість ознак має бути принаймні вдвічі більшою, ніж кількість змінних;

– вибірка має бути однорідною;

– вихідні змінні повинні мати симетричний розподіл.

**Метод головних компонент**

**Метод головних компонент, або компонентний аналіз** вперше був запропонований К. Пірсоном у 1901 р., який розглядав задачу найкращої (з погляду мінімізації суми квадратів відхилень) апроксимації сукупності точок прямими та площинами. Потім він був докладно розроблений американським статистиком й економістом Гарольдом Хотеллінгом у 1933 р. Його важливою перевагою є те, що він є єдиним математично обґрунтованим методом факторного аналізу.

За своєю сутністю метод полягає у виборі нової ортогональної системи координат у просторі спостережень. Як першу головну компоненту обирають напрям, вздовж якого масив спостережень має найбільшу дисперсію. Кожну наступну компоненту обирають також з умови максимізації частки дисперсії, що залишилася, вздовж неї, доповненої умовою ортогональності всім раніше обраним компонентам. При цьому із зростанням номера компоненти буде зменшуватися пов’язана з нею частка загальної дисперсії.

Кількість компонент визначається значною мірою суб’єктивно, виходячи з розуміння того, яка величина загальної дисперсії відповідає випадковій мінливості, що відображає похибку вимірювань, вплив неконтрольованих випадкових чинників тощо.

Вибір критерію інформативності в методі головних компонент передбачає, що найбільш важливу інформацію про аналізовану систему можна 142 відобразити лінійною моделлю, яка відповідає такому вибору системи координат у тому самому просторі, що забезпечує максимальні дисперсії для проекцій досліджуваних об’єктів. Такий підхід є доцільним, якщо більшість вихідних ознак узгоджено впливає на властивість, що вивчається, і пригнічує вплив іррелевантних чинників на розподіл об’єктів. Адекватну модель можна отримати також у випадку, коли кількість пов’язаних інформативних ознак невелика, але вплив інших чинників є неузгодженим. У цьому разі не порушується однорідність еліпсоїда розсіювання, а лише зменшується його довгастість уздовж напрямку досліджуваної властивості.

У факторному аналізі використовують також інші міри інформативності, що дають змогу визначити кількість істотних факторів.

**Критерій Кайзера**, або критерій власних чисел, запропонований американським психологом Генрі Феліксом Кайзером, передбачає, що до моделі включають тільки фактори, для яких власні числа є не меншими, ніж одиниця. За змістом це означає, що таким факторам відповідає дисперсія, еквівалента принаймні дисперсії одної змінної. У противному випадку виокремлення фактора не має сенсу. Цей критерій іноді залишає в моделі занадто багато факторів.

**Критерій кам’янистого осипу (критерій відсіювання)** передбачає побудову графіка, де по осі абсцис відкладають порядковий номер власного числа, а по осі ординат – його значення. Згідно з Р. Кеттелом необхідно знайти точку найбільшого уповільнення спадання власних значень і враховувати лише фактори, яким відповідають власні числа, розташовані лівіше цієї точки. На відміну від попереднього цей критерій статистично необґрунтовані і часто залишає в моделі не всі істотні фактори. Втім у випадках, коли істотних факторів небагато, а кількість змінних є великою, обидва критерії є придатними для практичного застосування. На практиці часто здійснюють розрахунки, використовуючи різні критерії, а потім обирають модель, що містить найбільшу кількість факторів, яким можна надати змістову інтерпретацію.

Критерії, що ґрунтуються на аналізі визначників вихідної та відтвореної кореляційної матриць, часто виявляються нестійкими. Критерії, які базуються на величині власних значень кореляційної матриці, у підсумку призводять до аналізу відсотка дисперсії, виділеної факторами. Усі загальні фактори, кількість яких дорівнює кількості параметрів, пояснюють 100% дисперсії. Якщо сума відсотків за факторами перевищує 100%, це свідчить про отримання від’ємних власних значень і, відповідно, комплексних власних векторів, що може бути наслідком некоректної редукції вихідної кореляційної матриці. Доцільно здійснювати двохетапну процедуру аналізу. На першому етапі максимальну кількість факторів не задають. Після його проведення аналізують дисперсії, оцінюють приблизну кількість факторів і проводять повторний аналіз.

**Метод головних факторів**

Цей метод використовують для зменшення кількості змінних. У його основі лежить припущення, що не всі змінні, які вимірювали при дослідженні системи, є незалежними. Тому можливо формування нових змінних, що достатньо повно відображають наявну інформацію.

На відміну від методу головних компонент, метод головних факторів ґрунтується не на дисперсійному критерії інформативності множини ознак, а на поясненні кореляцій, що існують між цими ознаками. Він враховує, що вихідні дані можуть містити грубі помилки, які у багатовимірному аналізі призводять до помилок інтерпретації.

Тому метод головних факторів застосовують у більш складних випадках, зокрема за наявності сумісного прояву аналізованих й іррелевантних властивостей об’єктів, що є порівнянними за ступенем внутрішньої узгодженості, а також для виділення групи діагностичних показників із вихідної множини ознак.

У методі головних факторів припускають, що кожний елемент матриці X\* є результатом впливу m гіпотетичних загальних факторів та одного характерного фактора.

Характерні фактори вважають некорельованими один з одним, а також із загальними факторами. Загальні фактори пов’язані з істотними ваговими коефіцієнтами більше, ніж з одною ознакою. Ті з них, для яких істотними є всі вагові коефіцієнти, називають генеральними факторами.

Перший варіант методу, запропонований Ч. Спірменом на початку 1900 р. передбачав існування одного загального та одного характерного факторів. Пізніше у 1920 р. британський та американський психолог Раймонд Бернард Кеттел та американський психолог Карл Джон Хользінгер запропонували біфакторну модель, яка передбачала існування декількох, зазвичай двох, загальних факторів. Сучасний варіант методу головних факторів було запропоновано Г. Томсоном

У методах головних факторів, центроїдному та трикутної декомпозиції спочатку визначають загальності, а потім кількість факторів. У методах максимальної правдоподібності й максимальних залишків використовують іншу послідовність: спочатку задають кількість факторів m, а потім підбирають значення загальностей так, щоб ранг матриці Rh наблизився до m.

При розробці методів оцінювання загальностей виходять із визначення Д. Лоулі й К. Рао, згідно з яким загальності – це величини, які при статистично значущих факторах дають змогу найкращим чином відтворити кореляційну матрицю. Найчастіше використовують такі методи оцінювання загальностей. Перший метод ґрунтується на тому, що із зростанням кількості ознак при сталій кількості факторів нижня межа оцінки загальності збігається до її істинного значення.

Другий метод базується на припущенні, що за великої кількості ознак як значення загальності можна використовувати найбільший за модулем коефіцієнт кореляції даної ознаки (відповідних рядка або стовпчика) з іншими змінними, взятий зі знаком плюс. Цей метод немає теоретичного підґрунтя, але практика свідчить, що при кількості ознак понад 20 одержувані за його допомогою результати мало відрізняються від тих, що отримують за допомогою більш точних методів.

Згідно з третім методом як значення загальності беруть не максимальне, а середнє за відповідним рядком (стовпчиком) значення коефіцієнта кореляції.

Першим кроком алгоритму методу головних факторів є отримання матриці парних коефіцієнтів кореляції R, на головній діагоналі якої стоять одиниці. Наступним кроком є одержання редукованої матриці Rh із загальностями на головній діагоналі.

Далі визначають перший загальний фактор, виходячи з умови, що його внесок ν1 до дисперсії процесу має бути максимальним.

**Інші методи факторного аналізу**

У **методі максимуму правдоподібності**, який запропоновано Д. Лоулі, оцінювання загальностей до безпосереднього застосування алгоритму факторного аналізу не здійснюють. Їх визначають за результатами обчислень з умови повного відтворення кореляційної матриці. За будь-якої кількості факторів, що розглядаються, цей метод дає можливість відтворити її з точністю до похибки обчислень. Якщо кількість факторів дорівнює кількості параметрів, то оцінки загальностей будуть збігатися із загальностями нередукованої кореляційної матриці, тобто дорівнювати одиниці. Основним недоліком методу є його нестійкість при використанні окремих типів даних, зокрема даних, що містять однакові або лінійно залежні вектори. Це призводить до виродження матриці характерностей. У такому випадку можна спробувати зняти проблему шляхом виключення з розгляду лінійно залежних параметрів або застосування методу головних факторів.

У попередніх методах максимізується квадратичний критерій. На відміну від них, у **центроїдному методі**, розробленому американським психологом Луїсом Леоном Терстоуном у 1930 р., максимізують модульний критерій. З погляду змістової інтерпретації ці критерії є еквівалентними. Цікаво зазначити, що Л. Терстоун спочатку вивчав електроніку у Корнельському університеті. Потім працював асистентом видатного американського інженера й винахідника Томаса Едісона. Він винайшов метод озвучування кінофільмів, удосконалив конструкції кінокамери й кінопректора. Першу координатну вісь у центроїдному методі обирають так, щоб вона проходила через центр ваги сукупності точок (нульову точку вважають відомою). Другу вісь проводять перпендикулярно першій. На головній діагоналі матриці R ставлять найбільші коефіцієнти кореляції кожного рядка (кожного стовпчика).

Перевагою центроїдного методу, внаслідок його непараметричності, є відносна стійкість до відхилень від нормального розподілу.

Існує декілька схем обчислень, які відрізняються одна від одної операцією **відбиття**. Відбиттю підлягають змінні, які мають найбільшу кількість від’ємних значень (одна з таких змінних; спочатку змінна, що має найбільшу кількість від’ємних значень, потім наступна за кількістю тощо; змінна з номером стовпця, що має максимальну за модулем від’ємну суму значень, а потім інші змінні з від’ємними сумами за стовпцем). За Л. Терстоуном, максимальна кількість факторів, які можуть бути однозначно визначені за наявності n змінних.

У методі **контрастних груп** необхідно задати початкове наближення матриці факторного перетворення. При цьому вважають, що вона може містити зайві ознаки, які потрібно виключити при подальших розрахунках. Передбачається, що у просторі ознак, що включені до моделі, розподіл об’єктів вписується до певного еліпсоїда розсіювання, витягнутого вздовж напряму тенденції, що діагностується. Припускають також, що вплив зайвих факторів є мінімальним для об’єктів, що знаходяться поблизу полюсів головної діагоналі еліпсоїда розсіювання. Це дає змогу виокремити «контрастні групи» обсягом від 1/4 до 1/3 загального обсягу вибірки й перевірити на них вплив кожної ознаки на досліджувану характеристику. Наприклад, при конструюванні тестів використовують **ϕ-коефіцієнт Пірсона**.

Сучасні апроксимуючи методи виходять з припущення, що є певне початкове наближення, яке необхідно покращити. Крім розглянутих вище методів головних факторів, найбільшої правдоподібності й контрастних груп до них належать:

– **груповий метод** Л. Гуттмана й П. Хорста, що базується на попередньому відборі груп елементарних ознак;

– **метод мінімальних залишків** Г. Хартмана;

– **метод α-факторного аналізу**, запропонований Г. Кайзером й І. Кеффрі в 1965 р.;

– **метод канонічного факторного аналізу** К. Рао;

– **методи, що оптимізують**.

Всі ці методи дають змогу послідовно покращувати знайдені розв’язки на основі використання статистичних прийомів оцінювання випадкової величини або статистичних критеріїв й передбачають великий обсяг трудомістких обчислень.

***3. ПОБУДОВИ Й ДОСЛІДЖЕННЯ РЕГРЕСІЙНИХ МОДЕЛЕЙ***

Завданням дослідження складних систем і процесів часто є перевірка наявності й встановлення типу зв’язку між незалежними змінними хі (**предикторами, факторами**), значення яких можуть змінюватися дослідником і мають певну заздалегідь задану похибку, та залежною змінною (відгуком) z. Розв’язання таких завдань є предметом регресійного аналізу. Термін «Регресія» вперше був уведений Ф. Гальтоном наприкінці XIX ст. На практиці завдання регресійного аналізу зазвичай формулюють так: необхідно підібрати достатньо просту функцію, що в певному розумінні найкращим чином описує наявну сукупність емпіричних даних.

**Загальна характеристика методів і задач регресійного аналізу**

Класичний регресійний аналіз включає методи побудови математичних моделей досліджуваних систем, методи визначення параметрів цих моделей і перевірки їх адекватності. Він припускає, що регресія є лінійною комбінацією лінійно незалежних базисних функцій від факторів з невідомими коефіцієнтами (параметрами). Фактори й параметри є детермінованими, а відгуки – рівноточними (тобто мають однакові дисперсії) некорельованими випадковими величинами. Передбачається також, що всі змінні вимірюють у неперервних числових шкалах.

Звичайна процедура класичного регресійного аналізу є такою. Спочатку обирають гіпотетичну модель, тобто формулюють гіпотези про фактори, які суттєво впливають на досліджувану характеристику системи, і тип залежності відгуку від факторів. Потім за наявними емпіричними даними про залежність відгуку від факторів оцінюють параметри обраної моделі. Далі за статистичними критеріями перевіряють її адекватність.

При побудові регресійних моделей реальних систем і процесів вказані вище припущення виконуються не завжди. У більшості випадків їх невиконання призводить до некоректності застосування процедур класичного регресійного аналізу і потребує застосування більш складних методів аналізу емпіричних даних.

Постулат про рівноточність і некорельованість відгуків не є обов’язковим. У випадку його невиконання процедура побудови регресійної моделі певною мірою змінюється, але суттєво не ускладнюється. Більш складною проблемою є вибір моделі та її незалежних змінних. У класичному регресійному аналізі припускають, що набір факторів задається однозначно, всі суттєві змінні наявні в моделі й немає ніяких альтернативних способів обрання факторів. На практиці це припущення не виконується. Тому виникає необхідність розробки формальних та неформальних процедур перетворення й порівняння моделей. Для пошуку оптимальних формальних перетворень використовують методи факторного та дискримінантного аналізу. На сьогодні розроблено комп’ютеризовані технології послідовної побудови регресійних моделей.

Фактори в класичному регресійному аналізі вважають детермінованими, тобто вважається, що дослідник має про них всю необхідну інформацію з абсолютною точністю. На практиці це припущення часто не виконується. Відмова від детермінованості незалежних змінних зумовлює необхідність застосування моделей кореляційного аналізу. В окремих випадках можна використовувати компромісні методи **конфлюентного аналізу**, які передбачають можливість нормально розподіленого та усіченого розкиду значень факторів. Якщо ця умова виконується, побудову моделі можна звести до багаторазового розв’язування регресійної задачі.

Відмова від припущення про детермінованість параметрів моделей у регресійному аналізі призводить до суттєвих ускладнень, оскільки порушує його статистичні основи. Але на практиці це припущення виконується не завжди. У деяких випадках можна вважати параметри випадковими величинами із заданими законами розподілу. Тоді як оцінки параметрів можна брати їх умовні математичні сподівання для відгуків, що спостерігалися. Умовні розподіли та математичні сподівання розраховують за узагальненою формулою Байєса, тому відповідні методи називають **байєсівським регресійним аналізом**.

Регресійні моделі часто використовують для опису процесів, що розвиваються у часі. У певних випадках це зумовлює необхідність переходу від випадкових значень відгуків до випадкових послідовностей, випадкових процесів або випадкових полів. Однією з поширених і найпростіших моделей такого типу є **модель авторегресії**, згідно з якою відгук залежить не тільки від факторів, але також і від часу. Якщо останню залежність можна виявити, то проблема зводиться до стандартної задачі побудови регресії для модифікованого відгуку. В інших випадках необхідно використовувати більш складні прийоми.

Процедури класичного регресійного аналізу припускають, що закон розподілу відгуків є нормальним. Проте на практиці найчастішими є випадки, коли цей закон невідомий чи відомо, що він не є нормальним. Їх дослідження зумовило виникнення непараметричного регресійного аналізу, який не передбачає необхідності попереднього задання функції розподілу.

Важливою проблемою, яка виникає при оцінюванні параметрів регресійних моделей, є наявність грубих помилок серед набору аналізованих даних. Ці помилки можуть виникати внаслідок неправильних дій дослідника, збоїв у роботі апаратури, неконтрольованих короткотривалих сильних зовнішніх впливів на досліджувану систему тощо. У таких випадках 195 використовують два підходи, що дають змогу зменшити вплив грубих помилок на результати аналізу. У першому з них розробляють критерії та алгоритми пошуку помилкових даних. Потім ці дані відкидають. У другому підході розробляють алгоритми аналізу, які є нечутливими до наявних помилкових даних (алгоритми робастного оцінювання параметрів).

Одним з основних постулатів класичного регресійного аналізу є припущення, що найкращі оцінки параметрів можна одержати, використовуючи метод найменших квадратів. На практиці оцінки, одержані за допомогою цього методу, часто бувають недостатньо точними і містять великі похибки. Причиною цього може бути структура регресійної моделі. Якщо вона є лінійною комбінацією експонент або поліномом високого степеня, то це призводить до поганої зумовленості матриці системи нормальних рівнянь і нестійкості оцінок параметрів. Підвищення стійкості оцінок можна досягти шляхом відмови від вимоги щодо їх незміщеності. Розвиток цього напряму досліджень призвів до виникнення гребеневого, або рідж-регресійного аналізу.

Регресійні моделі називають **лінійними** або **нелінійними**, якщо вони є, відповідно, лінійними або нелінійними за параметрами. При цьому визначення «лінійна» часто опускають. Значення найвищого степеня предиктора в поліноміальних моделях називають **порядком моделі**.

Вибір типу регресійної моделі є нетривіальним завданням. Для моделей, що містять одну незалежну змінну, рекомендують спочатку нанести наявні емпіричні дані на графік. Це дає можливість визначити наявність чи відсутність залежності між досліджуваними величинами, а також зробити певні припущення про тип залежності.

Часто як попередній етап регресійного аналізу рекомендують за допомогою методів кореляційного аналізу перевіряти наявність значущого зв’язку між досліджуваними змінними. Але при цьому слід ураховувати, що звичайні методи кореляційного аналізу дають змогу перевіряти лише гіпотезу про наявність лінійного зв’язку. Якщо зв’язок є, але він нелінійний, висновки, отримані за допомогою кореляційного аналізу, можуть бути помилковими.

Важливою особливістю регресійних моделей є те, що їх не можна застосовувати поза межами тієї області значень вихідних параметрів, для якої вони були побудовані. При використанні регресійних моделей типу полінома, оберненого полінома, тригонометричного ряду та деяких інших слід враховувати, що, збільшуючи кількість членів ряду, можна одержати скільки завгодно близькі до нуля значення функціоналів. Проте це не завжди свідчить про якість апроксимації, оскільки ці функціонали не дають інформації про ступінь наближення моделі до емпіричної залежності у проміжках між наявними точками.

Іншою проблемою може бути наявність декількох локальних екстремумів функціоналів. У таких випадках необхідно враховувати, що більшість стандартних алгоритмів дає можливість знаходити лише локальні, а не глобальні екстремуми функціоналів, і результат мінімізації залежать від вибору початкових умов пошуку. Це часто зумовлює необхідність встановлення додаткових критеріїв вибору моделі, серед яких можуть бути як формальні критерії їх адекватності, так і неформальні критерії, що ґрунтуються на сукупності відомих даних про об’єкт дослідження.

Поліноміальні регресійні моделі, як правило, є формальними. Їх використовують для опису систем і процесів, теорію яких розроблено недостатньо. При цьому спираються на відомі властивості ряду Тейлора для аналітичних функцій. Більш цікавими для дослідників зазвичай є математичні моделі, які відображають структуру та зв’язки у системах, сутність і механізми процесів, що відбуваються у них. Якщо теоретичні основи досліджуваних систем і процесів достатньо розроблені, часто постає проблема визначення окремих параметрів моделі за наявними емпіричними даними. Для її вирішення у багатьох випадках можна використовувати формальні процедури регресійного аналізу.

На практиці часто доводиться користуватися нелінійними за параметрами та багатовимірними моделями. Під багатовимірними тут розуміють моделі, що розглядають декілька відгуків. Задачам, що розв’язуються у межах відповідних напрямів регресійного аналізу, властиві й інші ускладнення. Так у багатовимірних моделей окремі відгуки можуть бути пов’язані один з одним. Сама регресійна модель часто задається у неявному вигляді та є неаналітичним розв’язком певної системи алгебраїчних або диференціальних рівнянь. Нестійкість оцінок параметрів для нелінійних моделей різко зростає. Як правило, такі задачі мають декілька розв’язків або не мають розв’язків взагалі.

Тема 2. Методи інтелектуального аналізу даних

Основні поняття та визначення інтелектуального аналізу даних. Мета використання технології. Сфера застосування. Класи систем інтелектуального аналізу даних. Методи використання навчальної інформації. Типи даних для роботи в Data Mining. Процес Data Mining. Методи та стадії Data Mining. Задачі класифікації та кластеризації. Задачі прогнозування. Задачі візуалізації. Методи класифікації та прогнозування. Дерева рішень. Метод опорних векторів. Метод «найближчий сусід». Байєсова класифікація. Методи кластерного аналізу. Ієрархічні і ітеративні методи. Алгоритм k-середніх. Нейронні мережі. Моделі нейронних мереж. Карти Кохонена. Методи пошуку асоціативних правил.

[7] с. 33-1273; [8] с. 6-364; [9] с. 15-72

***Лекція 2.1. (2 год)***

**Тема 2. Методи інтелектуального аналізу даних**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Основні поняття та визначення інтелектуального аналізу даних
2. Мета використання технології. Сфера застосування
3. Класи систем інтелектуального аналізу даних
4. Методи використання навчальної інформації

# ***1. ОСНОВНІ ПОНЯТТЯ ТА ВИЗНАЧЕННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ***

Корпоративна база даних будь-якого сучасного підприємства зазвичай містить набір таблиць, що зберігають записи про ті чи інші факти або об’єктах (наприклад, про товари, їх продажі, клієнтів, рахунки). Як правило, кожен запис в подібній таблиці описує якийсь конкретний об’єкт або факт. Наприклад, запис у таблиці продажів відображає той факт, що такий-то товар проданий такому-клієнтові тоді-таким-менеджером, і за великим рахунком нічого, крім цих відомостей, не містить. Однак сукупність великої кількості таких записів, накопичених за кілька років, може стати джерелом додаткової, набагато більш цінної інформації, яку неможливо одержати на основі однієї конкретної записи, а саме - відомостей про закономірності, тенденції або взаимозависимостях між якими-небудь даними. Прикладами подібної інформації є відомості про те, як залежать продажу певного товару від дня тижня, часу доби або пори року, які категорії покупців найчастіше купують той чи інший товар, яка частина покупців одного конкретного товару набуває інший конкретний товар, яка категорія клієнтів найчастіше вчасно не віддає наданий кредит.

Подібного роду інформація зазвичай використовується при прогнозуванні, стратегічному плануванні, аналізі ризиків, та цінність її для підприємства дуже висока. Мабуть, тому процес її пошуку і отримав назву Data Mining.

Data Mining переводиться як «видобуток» чи «розкопка даних». Нерідко поруч з Data Mining зустрічаються слова «виявлення знань у базах даних» (knowledge discovery in databases) і «інтелектуальний аналіз даних». Їх можна вважати синонімами Data Mining. Виникнення всіх зазначених термінів пов’язане з новим витком у розвитку засобів і методів обробки даних.

До початку 90-х років, здавалося, не було особливої потреби переосмислювати ситуацію в цій області. Все йшло своєю чергою в рамках напрямку, званого прикладної статистикою. Теоретики проводили конференції та семінари, писали значні статті та монографії, буяли аналітичними викладками.

Разом з тим, практики завжди знали, що спроби застосувати теоретичні екзерсиси для вирішення реальних завдань в більшості випадків виявляються марними. Але на заклопотаність практиків до пори до часу можна було не звертати особливої уваги – вони вирішували головним чином свої приватні проблеми обробки невеликих локальних баз даних.

І ось продзвенів дзвінок. У зв’язку з вдосконаленням технологій запису і зберігання даних на людей обрушилися колосальні потоки інформаційної руди в самих різних областях. Діяльність будь-якого підприємства (комерційного, виробничого, медичного, наукового і т.д.) тепер супроводжується реєстрацією та записом усіх подробиць його діяльності. Що робити з цією інформацією? Стало ясно, що без продуктивної переробки потоки сирих даних утворюють нікому не потрібну звалище.

Специфіка сучасних вимог до такої переробці наступні:

* дані мають необмежений обсяг;
* дані є різнорідними (кількісними, якісними, текстовими);
* результати мають бути конкретні та зрозумілі;
* інструменти для обробки сирих даних повинні бути прості у використанні.

Традиційна математична статистика, довгий час претендувала на роль основного інструмента аналізу даних, відверто спасувала перед обличчям виниклих проблем. Головна причина - концепція усереднювання по вибірці, яка призводить до операцій над фіктивними величинами (типу середньої температури пацієнтів по лікарні, середньої висоти будинку на вулиці, що складається з палаців і халуп і т.п.). Методи математичної статистики виявилися корисними головним чином для перевірки заздалегідь сформульованих гіпотез (verification-driven data mining) і для «грубого» розвідувального аналізу, що становить основу оперативної аналітичної обробки даних (online analytical processing, OLAP).

В основу сучасної технології Data Mining (discovery-driven data mining) покладена концепція шаблонів (патернів), що відображають фрагменти багатоаспектних взаємовідносин у даних. Ці шаблони являють собою закономірності, властиві підвибірки даних, які можуть бути компактно виражені в зрозумілій людині формі. Пошук шаблонів проводиться методами, не обмеженими рамками апріорних припущень про структуру вибірці і вигляді розподілів значень аналізованих показників. Приклади завдань на такий пошук при використанні Data Mining наведено в табл. 2.1.

Таблиця 2.1.

Приклади формулювань завдань при використанні методів OLAP і Data Mining

|  |  |
| --- | --- |
| OLAP | Data Mining |
| Які середні показники травматизму для курців і некурців? | Чи зустрічаються точні шаблони в описах людей, схильних до підвищеного травматизму? |
| Які середні розміри телефонних рахунків існуючих клієнтів у порівнянні з рахунками колишніх клієнтів (що відмовилися від послуг телефонної компанії)? | Чи є характерні портрети клієнтів, які, ймовірно, збираються відмовитися від послуг телефонної компанії? |
| Яка середня величина щоденних покупок за вкраденою і не вкраденої кредитній картці? | Чи існують стереотипні схеми покупок для випадків шахрайства з кредитними картками? |

Важливе положення Data Mining – нетривіальність розшукуваних шаблонів. Це означає, що знайдені шаблони повинні відображати неочевидні, несподівані (unexpected) регулярності в даних, складові так звані приховані знання (hidden knowledge). До товариства прийшло розуміння, що сирі дані (raw data) містять глибинний пласт знань, при грамотній розкопці якого можуть бути виявлені справжні самородки (рис. 2.1).



Рис. 1. Рівні знань, видобутих з даних

В цілому технологію Data Mining досить точно визначає Григорій Піатецкій-Шапіро – один із засновників цього напрямку:

Data Mining – це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень у різних сферах людської діяльності.

***2. МЕТА ВИКОРИСТАННЯ ТЕХНОЛОГІЇ. СФЕРА ЗАСТОСУВАННЯ***

**Суть та мету технології Data Mining** можна охарактеризувати так: це технологія, яка призначена для пошуку у великих обсягах даних неочевидних, об’єктивних і корисних на практиці закономірностей.

**Неочевидних** **–** це означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом.

**Об’єктивних** **–** це означає, що виявлені закономірності будуть повністю відповідати дійсності, на відміну від експертної думки, яке завжди є суб’єктивним.

**Практично корисних** **–** це означає, що висновки мають конкретне значення, котрому можна знайти практичне застосування.

**Знання** **–** сукупність відомостей, яка утворює цілісний опис, відповідне деякому рівню обізнаності про описуваному питанні, предметі, проблемі і т.д.

Використання знань означає дійсне застосування знайдених знань для досягнення конкретних переваг (наприклад, в конкурентній боротьбі за ринок).

Наведемо ще кілька визначень поняття Data Mining.

**Data Mining** **–** це процес виділення з даних неявної і неструктурованою інформації та представлення її у вигляді, придатному для використання.

**Data Mining** **–** це процес виділення, дослідження і моделювання великих обсягів даних для виявлення невідомих до цього структур (моделей ) з метою досягнення переваг у бізнесі (визначення SAS Institute).

**Data Mining** **–** це процес, мета якого **–** виявити нові значущі кореляції, зразки і тенденції в результаті просіювання великого обсягу збережених даних з використанням методик розпізнавання зразків плюс застосування статистичних і математичних методів (визначення Gartner Group).

**Сфера застосування** Data Mining нічим не обмежена **–** вона скрізь, де є будь-які дані. Але в першу чергу методи Data Mining сьогодні, м’яко кажучи, заінтригували комерційні підприємства, що розгортають проєкти на основі інформаційних сховищ даних (сховища даних). Досвід багатьох таких підприємств показує, що віддача від використання Data Mining може досягати 1000%. Наприклад, відомі повідомлення про економічний ефект, що в 10-70 разів перевищив початкові витрати від 350 до 750 тис. дол. Є відомості про проєкт в 20 млн. дол., який окупився всього за 4 місяці. Інший приклад **–** річна економія 700 тис. дол. за рахунок впровадження Data Mining в мережі універсамів у Великобританії.

Data Mining становлять велику цінність для керівників та аналітиків в їх повсякденній діяльності. Ділові люди усвідомили, що за допомогою методів Data Mining вони можуть отримати відчутні переваги в конкурентній боротьбі. Коротко охарактеризуємо деякі можливі бізнес–додатки інтелектуального аналізу даних.

**Роздрібна торгівля.** Підприємства роздрібної торгівлі сьогодні збирають докладну інформацію про кожну окрему покупку, використовуючи кредитні картки з маркою магазина і комп’ютеризовані системи контролю. Ось типові завдання, які можна вирішувати за допомогою Data Mining у сфері роздрібної торгівлі:

– *аналіз купівельної корзини* (аналіз подібності) призначений для виявлення товарів, які покупці прагнуть купувати разом. Володіння купівельної корзини необхідно для поліпшення реклами, вироблення стратегії створення запасів товарів і способів їх розкладки в торгових залах.

– *дослідження тимчасових шаблонів* допомагає торговим підприємствам приймати рішення про створення товарних запасів. Воно дає відповіді на питання типу «Якщо сьогодні покупець придбав відеокамеру, то через який час він найімовірніше купить нові батарейки і плівку?»

– *створення прогнозуючих моделей* дає можливість торговельним підприємствам дізнаватися характер потреб різних категорій клієнтів з певною поведінкою, наприклад, купують товари відомих дизайнерів або відвідують розпродажі. Ці знання потрібні для розробки точно спрямованих, економічних заходів щодо просування товарів.

**Банківська справа.** Досягнення технології Data Mining використовуються в банківській справі для вирішення наступних поширених завдань:

– *виявлення шахрайства з кредитними картками.* Шляхом аналізу минулих транзакцій, які згодом виявилися шахрайськими, банк виявляє деякі стереотипи такого шахрайства.

– *сегментація клієнтів.* Розбиваючи клієнтів на різні категорії, банки роблять свою маркетингову політику більш цілеспрямованою і результативною , пропонуючи різні види послуг різним групам клієнтів.

– *прогнозування змін клієнтури.* Data Mining допомагає банкам будувати прогнозні моделі цінності своїх клієнтів, і відповідним чином обслуговувати кожну категорію.

**Телекомунікації.** В області телекомунікацій методи Data Mining допомагають компаніям більш енергійно просувати свої програми маркетингу і ціноутворення, щоб утримувати існуючих клієнтів і залучати нових. Серед типових заходів відзначимо наступні:

*– аналіз записів про докладних характеристиках викликів.* Призначення такого аналізу **–** виявлення категорій клієнтів з схожими стереотипами користування їх послугами та розробка привабливих наборів цін і послуг;

– *виявлення лояльності клієнтів.* Data Mining можна використовувати для визначення характеристик клієнтів, які, один раз скориставшись послугами даної компанії, з великою часткою ймовірності залишаться їй вірними. У підсумку кошти, що виділяються на маркетинг, можна витрачати там, де віддача найбільше.

**Страхування.** Страхові компанії протягом ряду років накопичують великі обсяги даних. Тут широке поле діяльності для методів Data Mining:

– *виявлення шахрайства.* Страхові компанії можуть знизити рівень шахрайства, відшукуючи певні стереотипи в заявах про виплату страхового відшкодування, що характеризують взаємини між юристами, лікарями та заявниками.

– *аналіз ризику*. Шляхом виявлення поєднань факторів, пов’язаних з оплаченими заявами, страховики можуть зменшити свої втрати за зобов’язаннями. Відомий випадок, коли в США велика страхова компанія виявила, що суми, виплачені за заявами людей, одружених, вдвічі перевищує суми за заявами самотніх людей. Компанія відреагувала на це нове знання переглядом своєї загальної політики надання знижок сімейним клієнтам.

**Інші області в бізнесі.** Data Mining може застосовуватися в безлічі інших областей:

– *розвиток автомобільної промисловості.* При складанні автомобілів виробники повинні враховувати вимоги кожного окремого клієнта, тому їм потрібні можливість прогнозування популярності певних характеристик і знання того, які характеристики зазвичай замовляються разом;

– *політика гарантій.* Виробникам потрібно передбачати число клієнтів, які подадуть гарантійні заявки, і середню вартість заявок;

– *заохочення часто літаючих клієнтів.* Авіакомпанії можуть виявити групу клієнтів, яких даними заохочувальними заходами можна спонукати літати більше. Наприклад, одна авіакомпанія виявила категорію клієнтів, які здійснювали багато польотів на короткі відстані, що не накопичуючи досить миль для вступу в їхні клуби, тому вона таким чином змінила правила прийому до клубу, щоб заохочувати число польотів так само, як і милі.

**Медицина.** Відомо багато експертних систем для постановки медичних діагнозів. Вони побудовані головним чином на основі правил, що описують поєднання різних симптомів різних захворювань. За допомогою таких правил дізнаються не тільки, на що хворий пацієнт, але і як потрібно його лікувати. Правила допомагають вибирати засоби медикаментозного впливу, визначати показання **–** протипоказання, орієнтуватися в лікувальних процедурах, створювати умови найбільш ефективного лікування, пророкувати результати призначеного курсу лікування і т. п. Технології Data Mining дозволяють виявляти в медичних даних шаблони, що становлять основу зазначених правил.

**Молекулярна генетика і генна інженерія.** Мабуть, найбільш гостро і водночас чітко завдання виявлення закономірностей в експериментальних даних коштує в молекулярній генетиці та генної інженерії. Тут вона формулюється як визначення так званих маркерів, під якими розуміють генетичні коди, контролюючі ті чи інші фенотипічні ознаки живого організму. Такі коди можуть містити сотні, тисячі і більше пов’язаних елементів.

На розвиток генетичних досліджень виділяються великі кошти. Останнім часом в даній області виник особливий інтерес до застосування методів Data Mining. Відомо кілька великих фірм, що спеціалізуються на застосуванні цих методів для розшифровки генома людини і рослин.

**Прикладна хімія.** Методи Data Mining знаходять широке застосування в прикладної хімії (органічної та неорганічної). Тут нерідко виникає питання про з’ясування особливостей хімічної будови тих чи інших сполук, що визначають їх властивості. Особливо актуальна така задача при аналізі складних хімічних сполук, опис яких включає сотні і тисячі структурних елементів та їх зв’язків.

Можна навести ще багато прикладів різних областей знання, де методи Data Mining відіграють провідну роль. Особливість цих областей полягає в їх складній системної організації. Вони відносяться головним чином до над кібернетичному рівню організації систем, закономірності якого не можуть бути достатньо точно описані на мові статистичних чи інших аналітичних математичних моделей. Дані в зазначених областях неоднорідні, гетерогенні, нестаціонарні і часто відрізняються високою розмірністю.

***3. КЛАСИ СИСТЕМ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ***

***МЕТОДИ ВИКОРИСТАННЯ НАВЧАЛЬНОЇ ІНФОРМАЦІЇ***

**Типи закономірностей**

Виділяють п’ять стандартних типів закономірностей, які дозволяють виявляти методи Data Mining: ***асоціація, послідовність, класифікація, кластеризація і прогнозування****.*

**Асоціація** має місце в тому випадку, якщо кілька подій зв’язані один з одним. Наприклад, дослідження, проведене в супермаркеті, може показати, що 65 % купили кукурудзяні чіпси беруть також і «кока – колу», а за наявності знижки за такий комплект «колу» здобувають у 85% випадків. Маючи в своєму розпорядженні відомостями про подібну асоціацію, менеджерам легко оцінити, наскільки дієва надається знижка.

Якщо існує ланцюжок пов’язаних у часі подій, то говорять про **послідовність**. Так, наприклад, після покупки будинку в 45 % випадків протягом місяця купується і нова кухонна плита, а в межах двох тижнів 60 % новоселів обзаводяться холодильником.

За допомогою **класифікації** виявляються ознаки, що характеризують групу, до якої належить той чи інший об’єкт. Це робиться за допомогою аналізу вже класифікованих об’єктів і формулювання деякого набору правил.

**Кластеризація** відрізняється від класифікації тим, що самі групи заздалегідь не задані. За допомогою кластеризації засобів Data Mining самостійно виділяють різні однорідні групи даних.

Основою для всіляких систем **прогнозування** служить історична інформація, що зберігається в БД у вигляді часових рядів. Якщо вдається побудувати знайти шаблони, які адекватно відображають динаміку поведінки цільових показників, є ймовірність, що з їх допомогою можна передбачити і поведінку системи в майбутньому.

**Класи систем Data Mining**

Data Mining є мультидисциплінарною областю, яка виникла і розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних та ін. (рис. 2.2). Звідси велика кількість **методів і алгоритмів**, реалізованих у різних діючих системах Data Mining. Багато з таких систем інтегрують в собі відразу кілька підходів. Проте, як правило, в кожній системі є якась ключова компонента, на яку робиться головна ставка. Нижче наводиться класифікація зазначених ключових компонент на основі роботи. Виділеним класам дається коротка характеристика.

Data Mining

Нейронні мережі

Статистика

Ефективні обчислення

Сховища даних

Теорія без даних

Оперативна аналітична обробка

Інформаційний пошук

Експертні системи

Візуалізація даних

Розпізнавання образів

Рис. 2.2. Data Mining – мультидисциплінарна область

****

Рис. 2.3. Популярні продукти для Data Mining

***Предметно-орієнтовані аналітичні системи*** дуже різноманітні. Найбільш широкий підклас таких систем, що одержав поширення в галузі дослідження фінансових ринків, носить назву «технічний аналіз». Він являє собою сукупність декількох десятків методів прогнозу динаміки цін і вибору оптимальної структури інвестиційного портфеля, заснованих на різних емпіричних моделях динаміки ринку. Ці методи часто використовують нескладний статистичний апарат, але максимально враховують сформовану своїй області специфіку (професійна мова, системи різних індексів і пр.). На ринку є безліч програм цього класу. Як правило, вони досить дешеві (зазвичай $ 300-1000).

***Статистичні пакети****.*Останні версії майже всіх відомих статистичних пакетів включають поряд з традиційними статистичними методами також елементи Data Mining. Але основна увага в них приділяється все ж класичним методикам **–** кореляційному, регресійному, факторному аналізу і іншим. Недоліком систем цього класу вважають вимогу до спеціальної підготовки користувача. Також відзначають, що потужні сучасні статистичні пакети є занадто «ваговитими» для масового застосування у фінансах і бізнесі. До того ж часто ці системи досить дорогі – від $ 1000 до $ 15000.

Є ще більш серйозний принциповий недолік статистичних пакетів, що обмежує їх застосування в Data Mining. Більшість методів, що входять до складу пакетів спираються на статистичну парадигму, в якій головними фігурантами служать усереднені характеристики вибірки. А ці характеристики, як зазначалося вище, при дослідженні реальних складних життєвих феноменів часто є фіктивними величинами.

В якості прикладів найбільш потужних і поширених статистичних пакетів можна назвати SAS (компанія SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA та інші.

***Нейронні мережі.*** Це великий клас систем, архітектура яких має аналогію (як тепер відомо, досить слабку) з побудовою нервової тканини з нейронів. В одній з найбільш поширених архітектурі зі зворотним поширенням помилки, імітується робота нейронів у складі ієрархічної мережі, де кожен нейрон більш високого рівня з’єднаний своїми входами з виходами нейронів нижчого шару. На нейрони самого нижнього шару подаються значення вхідних параметрів, на основі яких потрібно приймати якісь рішення, прогнозувати розвиток ситуації і т. д. Ці значення розглядаються як сигнали, що передаються в наступний шар, ослабляючись або посилюючись в залежності від числових значень (ваг), приписуваних між нейронних зв’язків. У результаті на виході нейрона самого верхнього шару виробляється деяке значення, яке розглядається як відповідь **–** реакція всієї мережі на введені значення вхідних параметрів. Для того щоб мережу можна було застосовувати надалі, її колись треба «натренувати» на отриманих раніше даних, для яких відомі і значення вхідних параметрів, і правильні відповіді на них. Тренування полягає в підборі ваг між нейронних зв’язків, що забезпечують найбільшу близькість відповідей мережі до відомих правильних відповідей.

Основним недоліком нейронно-мережевої парадигми є необхідність мати дуже великий обсяг навчальної вибірки. Інший суттєвий недолік полягає в тому, що навіть натренована нейронна мережа являє собою чорний ящик. Знання, зафіксовані як ваги кількох сотень між нейронних зв’язків, абсолютно не піддаються аналізу та інтерпретації людиною (відомі спроби дати інтерпретацію структурі налаштованої нейронної мережі виглядають непереконливими **–** система « KINOsuite – PR»).

Приклади нейромережевих систем **–** BrainMaker (CSS ), NeuroShell (Ward Systems Group), OWL (HyperLogic ). Вартість їх досить значна: $ 1500-8000.

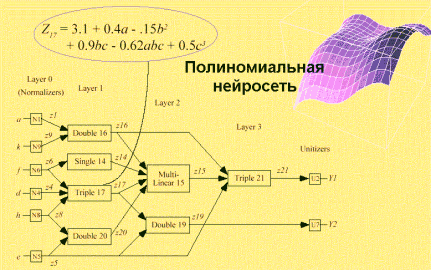


Рис. 2.4. Поліноміальна нейронна мережа

**Системи міркувань на основі аналогічних випадків.** Ідея систем case based reasoning **–** CBR **–** на перший погляд вкрай проста. Для того щоб зробити прогноз на майбутнє чи вибрати правильне рішення, ці системи знаходять у минулому близькі аналоги наявної ситуації і вибирають ту ж відповідь, який був для них правильним. Тому цей метод ще називають методом «найближчого сусіда» (nearest neighbour ). Останнім часом поширення отримав також термін memory based reasoning, який акцентує увагу, що рішення приймається на підставі всієї інформації, накопиченої в пам’яті.

Системи CBR показують непогані результати в найрізноманітніших задачах. Головним їх мінусом вважають те, що вони взагалі не створюють будь**–**яких моделей або правил, узагальнюючих попередній досвід у виборі рішення вони ґрунтуються на всьому масиві доступних історичних даних, тому неможливо сказати, на основі яких конкретно факторів CBR системи будують свої відповіді.

Інший мінус полягає в свавіллі, який допускають системи CBR при виборі міри «близькості». Від цієї міри найрішучішим чином залежить обсяг безлічі прецедентів, які потрібно зберігати в пам’яті для досягнення задовільною класифікації або прогнозу.

Приклади систем, що використовують CBR, **–** KATE tools (Acknosoft, Франція), Pattern Recognition Workbench (Unica, США).

**Дерева рішень (decision trees)** є одним з найбільш популярних підходів до вирішення завдань Data Mining. Вони створюють ієрархічну структуру класифікуючи правил типу «ЯКЩО ... ТО ...» (if – then), що має вигляд дерева. Для прийняття рішення, до якого класу віднести деякий об’єкт або ситуацію, потрібно відповісти на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Запитання мають вигляд «значення параметра A більше x?». Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня, якщо негативний **–** то до лівого вузла; потім знову слід питання, пов’язане з відповідним вузлом.

Популярність підходу пов’язана як би з наочністю і зрозумілістю. Але дерева рішень принципово не здатні знаходити «кращі» (найбільш повні і точні) правила в даних. Вони реалізують наївний принцип послідовного перегляду ознак і «чіпляють» фактично оскільки справжніх закономірностей, створюючи лише ілюзію логічного висновку.

Разом з тим, більшість сислтем використовують саме цей метод. Найвідомішими є See5/С5.0 (RuleQuest, Австралія), Clementine (Integral Solutions, Великобританія), SIPINA (University of Lyon, Франція), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада). Вартість цих систем варіюється від 1 до 10 тис. дол.

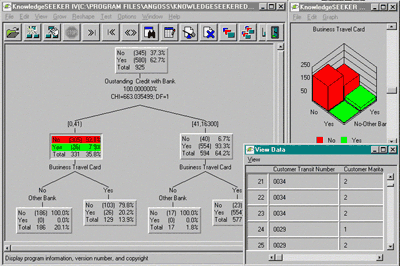


Рис. 2.5. Система KnowledgeSeeker обробляє банківську інформацію

**Еволюційне програмування.** Проілюструємо сучасний стан даного підходу на прикладі системи PolyAnalyst **–** вітчизняної розробки, що отримала сьогодні загальне визнання на ринку Data Mining. У даній системі гіпотези про вид залежності цільової змінної від інших змінних формулюються у вигляді програм на деякій внутрішній мові програмування. Процес побудови програм будується як еволюція у світі програм (цим підхід трохи схожий на генетичні алгоритми). Коли система знаходить програму, більш або менш задовільно виражає шукану залежність, вона починає вносити до неї невеликі модифікації і відбирає серед побудованих дочірніх програм ті, які підвищують точність. Таким чином система «вирощує» кілька генетичних ліній програм, які конкурують між собою в точності висловлювання шуканої залежності. Спеціальний модуль системи PolyAnalyst переводить знайдені залежності з внутрішньої мови системи на зрозумілу користувачеві мову (математичні формули, таблиці та ін.).

Інший напрям еволюційного програмування пов’язане з пошуком залежності цільових змінних від інших у формі функцій якогось певного виду. Наприклад, в одному з найбільш вдалих алгоритмів цього типу **–** методі групового урахування аргументів (МГУА) залежність шукають у формі поліномів. В даний час з продаються в Росії систем МГУА реалізований в системі NeuroShell компанії Ward Systems Group. Вартість систем до $500.

**Генетичні алгоритми.** Data Mining не основна область застосування генетичних алгоритмів. Їх потрібно розглядати скоріше як потужний засіб вирішення різноманітних комбінаторних завдань і завдань оптимізації. Проте генетичні алгоритми увійшли зараз в стандартний інструментарій методів Data Mining, тому вони і включені в даний огляд.

Перший крок при побудові генетичних алгоритмів **–** це кодування вихідних логічних закономірностей в базі даних, які іменують хромосомами, а весь набір таких закономірностей називають популяцією хромосом. Далі для реалізації концепції відбору вводиться спосіб зіставлення різних хромосом. Населення обробляється за допомогою процедур репродукції, мінливості (мутацій), генетичної композиції. Ці процедури імітують біологічні процеси. Найбільш важливі серед них: випадкові мутації даних в індивідуальних хромосомах, переходи (кросинговер) і рекомбінація генетичного матеріалу, що міститься в індивідуальних батьківських, та міграції генів. У ході роботи процедур на кожній стадії еволюції виходять популяції з усе більш досконалими індивідуумами.

Генетичні алгоритми зручні тим, що їх легко розпаралелювати. Наприклад, можна розбити покоління на кілька груп і працювати з кожною з них незалежно, обмінюючись час від часу кількома хромосомами. Існують також і інші методи розпаралелювання генетичних алгоритмів.

Генетичні алгоритми мають ряд недоліків. Критерій відбору хромосом і використовувані процедури є евристичними і далеко не гарантують знаходження «кращого» рішення. Як і в реальному житті, еволюцію може «заклинити» на який**-**небудь непродуктивною гілки. І, навпаки, можна навести приклади, як два неперспективних батька, які будуть виключені з еволюції генетичним алгоритмом, виявляються здатними призвести високоефективного нащадка. Це особливо стає помітно при вирішенні високо розмірних завдань зі складними внутрішніми зв’язками.

Прикладом може служити система GeneHunter фірми Ward Systems Group. Його вартість **–** близько $ 1000.

**Алгоритми обмеженого перебору** були запропоновані в середині 60-х років М.М. Бонгард для пошуку логічних закономірностей в даних. З тих пір вони продемонстрували свою ефективність при вирішенні безлічі завдань із всіляких областей.

Ці алгоритми обчислюють частоти комбінацій простих логічних подій у підгрупах даних. Приклади простих логічних подій: X = a ; X < a ; X a ; a < X < b та ін, де X – який або параметр, «a» і «b» – константи. Обмеженням служить довжина комбінації простих логічних подій (у М. Бонгард вона дорівнювала 3). На підставі аналізу обчислених частот робиться висновок про корисність тієї чи іншої комбінації для встановлення асоціації в даних, для класифікації, прогнозування.

Найбільш яскравим сучасним представником цього підходу є система WizWhy підприємства WizSoft. Хоча автор системи Абрахам Мейдан не розкриває специфіку алгоритму, покладеного в основу роботи WizWhy, за результатами ретельного тестування системи були зроблені висновки про наявність тут обмеженого перебору (вивчалися результати, залежно часу їх отримання від числа аналізованих параметрів та ін.)

Автор WizWhy стверджує, що його система виявляє ВСЕ логічні if **–** then правила в даних. Насправді це, звичайно, не так. По-перше, максимальна довжина комбінації в if **–** then правилі в системі WizWhy дорівнює 6, і, по-друге, з самого початку роботи алгоритму виробляється евристичний пошук простих логічних подій, на яких потім будується весь подальший аналіз. Зрозумівши ці особливості WizWhy, неважко було запропонувати найпростішу тестову задачу, яку система не змогла взагалі вирішити. Інший момент **–** система видає рішення за прийнятний час тільки для порівняно невеликої розмірності даних.

Проте, система WizWhy є на сьогоднішній день одним з лідерів на ринку продуктів Data Mining. Це не позбавлене підстав. Система постійно демонструє більш високі показники при вирішенні практичних завдань, ніж всі інші алгоритми. Вартість системи близько $4000, кількість продажів – 30000.

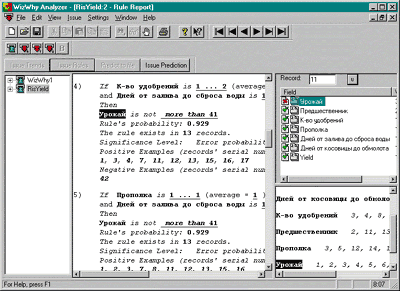


Рис. 2.6. Система WizWhy виявила правила, що пояснюють низьку врожайність деяких сільськогосподарських ділянок

**Системи для візуалізації багатовимірних даних.** В тій чи іншій мірі засоби для графічного відображення даних підтримуються всіма системами Data Mining. Разом з тим, досить значну частку ринку займають системи, що спеціалізуються виключно на цій функції. Прикладом тут може служити програма DataMiner 3D словацької фірми Dimension5 (5**-**й вимір).

У подібних системах основну увагу сконцентровано на доброзичливості користувацького інтерфейсу, що дозволяє асоціювати з аналізованими показниками різні параметри діаграми розсіювання об’єктів (записів) бази даних. До таких параметрів належать колір, форма, орієнтація щодо власної осі, розміри та інші властивості графічних елементів зображення. Крім того, системи візуалізації даних забезпечені зручними засобами для масштабування і обертання зображень. Вартість систем візуалізації може досягати декількох сотень доларів.



Рис. 2.7. Візуалізація даних системою DataMiner 3D

Загальний алгоритм аналізу Data Mining. Методика аналізу з використанням Data Mining базується на різних алгоритмах видобутку закономірностей у вхідних даних. Таких алгоритмів є багато, але вони не можуть гарантувати якісного кінцевого результату, бо існує багато чинників, що можуть вплинути на сам хід аналізу.

**Якісний аналіз даних з використанням DM**

Для якісного аналізу будь-яких даних слід дотримуватися загальної схеми використання DM:

1. Висування гіпотез;
2. Збір та систематизація даних;
3. Підбір адекватної моделі;
4. Тестування та інтерпретація отриманих даних;
5. Використання у реальних умовах.

Ця схема не залежить від предметної області та сфери діяльності. Вона є універсальною.

**1. Висування гіпотез.**

Гіпотезою тут будемо вважати припущення про вплив певних факторів на процес, що досліджується.

Автоматизувати процес висування гіпотез є вкрай складно, тому, цю задачу мають вирішувати експерти – фахівці в предметній області.

Слід довіритися їх досвіду та здоровому глузду, максимально використати ці знання про предмет досліджень і зібрати як найбільше гіпотез/припущень.

Зазвичай, добрі результати надають тактики «круглого столу» або «мозкової атаки». На початку слід зібрати та систематизувати всі ідеї, а оцінювати їх пізніше. В результаті повинен бути складений перелік з описів всіх факторів досліджуваного об’єкту.

Наприклад, Для задачі прогнозування попиту товару потрібно скласти перелік факторів, що впливатимуть на об’єкт і експертно**-**оцінити суттєвість кожного з них. Така оцінка не є вирішальною, але від неї починають відштовхуватися.

Сезон 100

День тижня 80

Об’єм продажів за попередні тижні 100

Об’єм продажів за аналогічний період минулого року 95

Рекламна компанія 60

Маркетингові заходи 40

Якість продукції 50

Бренд 25

Коливання ціни від середньо ринкової 60

Наявність подібного товару в конкурентів 15

Згодом, під час аналізу, може з’ясуватися, що фактор, який експерти оцінили як важливий, буде мати незначний вплив на процес і навпаки.

## **2. Збір та систематизація даних**

### 2.1. Збір даних.

Для аналізу потрібно як найбільше даних, бо це надає можливість оцінити вплив максимальної кількості показників. Згодом, простіше відхилити певну частину даних, аніж розпочинати новий збір.

### Методи збору:

#### 1. Отримання даних з внутрішніх джерел.

Це не складно, бо така інформація зазвичай зберігається в облікових системах у табличній формі, де існують різні механізми отримання звітів та експортування даних.

#### 2. Отримання відомостей з непрямих даних.

Наприклад, потрібно оцінити реальний фінансовий стан мешканців певного регіону. Існує кілька категорій товару (авто), що різняться за ціною – для незаможних, середнього класу, заможних. Якщо отримати звіт про продажі товару в цьому районі і проаналізувати пропорції, то робиться висновок: чим більшим є відсоток продажів дорогого товару, тим заможнішими є мешканці.

#### **3. Використання відкритих джерел.**

До широкого загалу надаються статистичні збірники, звіти корпорацій, результати маркетингових досліджень, соціологічні опитування.

#### **4. Влаштування власних маркетингових досліджень та подібних заходів по збору даних.**

Це зазвичай є дорогим заходом, але доволі ефективним.

#### **5. Наповнення даних згідно експертних оцінок співробітниками організації.**

Слід оцінити вартість збору даних, що потрібні для аналізу. Одні дані беруться з публічних інформаційних джерел, інші мають бути оплачені, дані про діяльність конкурентів можуть бути доволі дорогими.

Вартість збору інформації різними методами суттєво різниться за ціною та витраченим часом, тому, слід вважати на співвідношення теперішніх витрат з майбутніми результатами.

Від даних, які експерти вважають несуттєвими, певна річ, можна відмовитися, але від значущих даних не можна, бо аналіз буде базуватися у цьому випадку на другорядних факторах і відповідно, отримана модель буде надавати нестабільні та невірні результати.

**Дані, набір даних та їх атрибутів**

У широкому розумінні **дані – це** **факти, текст, графіки, картинки, звуки, аналогові або цифрові відео–сегменти**. Дані можуть бути отримані в результаті вимірювань, експериментів, арифметичних і логічних операцій. Дані повинні бути представлені у формі, придатній для зберігання, передачі і обробки. Іншими словами, дані – це необроблений матеріал, що надається постачальниками даних і використовується споживачами для формування інформації на основі даних.

У табл. 2.2 представлена двовимірна таблиця, що представляє собою набір даних.

По горизонталі таблиці розташовуються атрибути об’єкта або його ознаки. По вертикалі таблиці **–** об’єкти. Об’єкт описується як набір атрибутів. Об’єкт також відомий як запис, випадок, приклад, рядок таблиці і т.д.

**Атрибут –** властивість, що характеризує об’єкт. Наприклад : колір очей людини, температура води і т.д. Атрибут також називають змінною, полем таблиці, виміром, характеристикою.

Таблиця 2.2.

Двовимірна таблиця «об’єкт-атрибут»

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Атрибути** | | | | |
| **Об’єкти** | Код клієнта | Вік | Сімейний статус | Прибуток | Клас |
| 1 | 19 | неодр. | 1234 | 1 |
| 2 | 23 | одр. | 1222 | 1 |
| 3 | 34 | одр. | 2700 | 1 |
| 4 | 24 | неодр. | 2343 | 1 |
| 5 | 26 | одр. | 1765 | 2 |
| 6 | 32 | розл. | 2652 | 1 |
| 7 | 19 | неодр. | 1200 | 2 |
| 8 | 22 | неодр. | 1765 | 2 |
| 9 | 40 | одр. | 1998 | 1 |
| 10 | 43 | розл. | 4332 | 1 |

**Змінна** (variable) **–** властивість або характеристика, загальна для всіх досліджуваних об’єктів, прояв якої може змінюватися від об’єкта до об’єкта.

**Значення** (value) змінної є проявом ознаки.

При аналізі даних, як правило, немає можливості розглянути всю сукупність об’єктів, що нас цікавить. Вивчення дуже великих обсягів даних є дорогим процесом, що вимагає великих затрат часу, а також неминуче призводить до помилок, пов’язаних з людським фактором.

Цілком достатньо розглянути деяку частину всієї сукупності, тобто вибірку, і отримати цікаву для нас інформацію на її підставі.

Однак розмір вибірки повинен залежати від різноманітності об’єктів, представлених у генеральній сукупності. У вибірці повинні бути представлені різні комбінації і елементи генеральної сукупності.

**Генеральна сукупність** (population) **–** вся сукупність досліджуваних об’єктів, що цікавить дослідника.

**Вибірка** (sample) **–** частина генеральної сукупності, певним способом відібрана з метою дослідження та отримання висновків про властивості та характеристики генеральної сукупності.

**Параметри** **–** числові характеристики генеральної сукупності.

**Статистики** **–** числові характеристики вибірки. Часто дослідження ґрунтуються на гіпотезах. Гіпотези перевіряються за допомогою даних.

**Гіпотеза** **–** припущення щодо параметрів сукупності об’єктів, яке має бути перевірено на її частині.

**Гіпотеза** **–** частково обґрунтована закономірність знань, що служить або для зв’язку між різними емпіричними фактами, або для пояснення факту групи фактів.

Приклад гіпотези: між показниками тривалості життя та якістю харчування є зв’язок. У цьому випадку метою дослідження може бути пояснення змін конкретної змінної, в даному випадку **–** тривалості життя. Припустимо, існує гіпотеза, що залежна змінна (тривалість життя) змінюється залежно від деяких причин (якість харчування, спосіб життя, місце проживання і т.д.), які і є незалежними змінними.

Однак змінна першопочатково не є залежною або незалежною, вона стає такою після формулювання конкретної гіпотези. Залежна змінна в одній гіпотезі може бути незалежною в іншій.

**Вимірювання** **–** процес присвоєння чисел характеристикам досліджуваних об’єктів згідно певного правила.

У процесі підготовки даних вимірюється не сам об’єкт, а його характеристики.

**Шкала** **–** правило, відповідно до якого об’єктам присвоюються числа.

Багато інструментів Data Mining при імпорті даних з інших джерел пропонують вибрати тип шкали для кожної змінної та/або вибрати тип даних для вхідних і вихідних змінних (символьні, числові, дискретні і безперервні).

Користувачеві такого інструменту необхідно володіти цими поняттями.

**Змінні можуть бути числовими даними або символьними.**

Числові дані, у свою чергу, можуть бути дискретними і безперервними.

**Дискретні дані** являються значеннями ознаки, загальне число яких скінченне або нескінченне, але може бути підраховане за допомогою натуральних чисел від одного до нескінченності.

Приклад дискретних даних. Тривалість маршруту тролейбуса (кількість варіантів тривалості скінченне): 10, 15, 25 хв.

**Безперервні дані –** дані, значення яких можуть набувати якого**–**завгодно значення в деякому інтервалі. Вимірювання безперервних даних передбачає велику точність.

Приклад безперервних даних: температура, висота, вага, довжина і т.д.

**Шкали.** Існує п’ять типів шкал вимірювань: номінальна, порядкова, інтервальна, відносна і дихотомічна.

**Номінальна шкала** (nominal scale) **–** шкала, яка містить тільки категорії; дані в ній не можуть упорядковуватися, з ними не можуть бути зроблені ніякі арифметичні дії.

Номінальна шкала складається з назв, категорій, імен для класифікації і сортування об’єктів або спостережень за деякою ознакою.

Приклад такої шкали: професії, місто проживання, сімейний стан.

Для цієї шкали застосовні тільки такі операції: дорівнює (=), не дорівнює (≠).

**Порядкова шкала** (ordinal scale) **–** шкала, в якій числа присвоюють об’єктам для позначення відносної позиції об’єктів, але не величини відмінностей між ними.

Шкала вимірювань дає можливість ранжувати значення змінних. Вимірювання ж у порядковій шкалі містять інформацію лише про порядок проходження величин, але не дозволяють сказати «наскільки одна величина більше іншої, або» наскільки вона менше інший».

Приклад такої шкали: місце (1-ше, 2-ге, 3-є), яке команда отримала на змаганнях, номер студента в рейтингу успішності (1-й, 23-й, і т.д.), при цьому невідомо, наскільки один студент успішніше іншого, відомий лише його номер у рейтингу.

Для цієї шкали застосовні тільки такі операції: рівно (=), не дорівнює (≠), більше (>), менше (<).

**Інтервальна шкала** (interval scale) **–** шкала, різниці між значеннями якої можуть бути обчислені, проте їх відношення не мають сенсу.

Ця шкала дозволяє знаходити різницю між двома величинами, має властивості номінальної та порядкової шкал, а також дозволяє визначити кількісну зміну ознаки.

Приклад такої шкали: температура води в морі вранці – 19 градусів, ввечері – 24, тобто вечірня на 5 градусів вище, але не можна сказати, що вона в 1,26 разів вище.

Номінальна і порядкова шкали є дискретними, а інтервальна шкала – безперервною, вона дозволяє здійснювати точні вимірювання ознаки і виробляти арифметичні операції додавання, віднімання, множення, ділення.

Для цієї шкали застосовні тільки такі операції: одно (=), не дорівнює ( ), більше (>), менше (<), операції додавання (+) і віднімання ( –).

**Відносна шкала** (ratio scale) – шкала, в якій є певна точка відліку і можливі відносини між значеннями шкали.

Приклад такої шкали : вага новонародженої дитини (4 кг і 3 кг). Перша в 1,33 рази важче.

Ціна на картоплю в супермаркеті вище в 1,2 рази, ніж ціна на ринку.

Відносні та інтервальні шкали є числовими .

Для цієї шкали можуть бути застосовані тільки такі операції: рівно (=), не дорівнює (≠), більше (>), менше (<), операції додавання (+) і віднімання (–), множення (\*) і ділення (/ ).

**Дихотомічна шкала** ( dichotomous scale ) – шкала, яка містить тільки дві категорії.

Приклад такої шкали: стать (чоловіча і жіноча).

Приклад використання різних шкал для вимірювань властивостей різних об’єктів, в даному випадку температурних умов, наведено в таблиці даних, зображеної в табл. 2.3.

Таблиця 2.3.

Множина вимірювань властивостей різних об’єктів

|  |  |  |  |
| --- | --- | --- | --- |
| **Номер об’єкту** | **Професія (номінальна шкала)** | **Середній**  **бал (інтервальна шкала)** | **Освіта (порядкова шкала)** |
| **1** | Слюсар | 22 | середня |
| **2** | Вчений | 55 | вища |
| **3** | вчитель | 47 | вища |

Приклад використання різних шкал для вимірювань властивостей однієї системи, в даному випадку температурних умов, наведено в таблиці даних, зображеній в табл. 2.4.

Таблиця 2.4.

Множина вимірювань властивостей однієї системи

|  |  |  |  |
| --- | --- | --- | --- |
| **Дата змінення** | **Хмарність (номінальна шкала)** | **Температура о 7 годині (інтервальна шкала)** | **Сила вітру (порядкова шкала)** |
| **3 жовтня** | Хмарно | 22ºС | Сильний вітер |
| **4 жовтня** | напівхмарно | 17ºС | Слабий вітер |
| **5 жовтня** | Ясно | 23ºС | Дуже сильний вітер |

**Типи наборів даних.** Найбільш часто зустрічаються дані, що складаються з записів (record data).

Приклади таких наборів даних: табличні дані, матричні дані, документальні дані, транзакційні або операційні.

**Табличні дані** – дані, що складаються з записів, кожен з яких складається з фіксованого набору атрибутів.

**Транзакційні дані** представляють собою особливий тип даних, де кожен запис, що являється транзакцією, включає набір значень.

Приклад транзакційної бази даних, що містить перелік покупок клієнтів магазину, наведено на рис. 2.8.

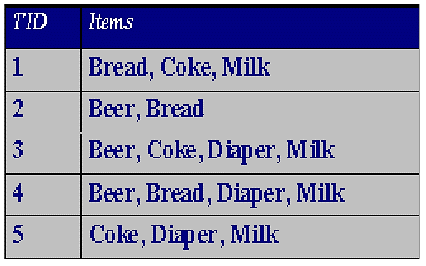


Рис. 2.8. Приклад транзакційних даних

**Графічні дані.** Приклади графічних даних: молекулярні структури; графи (рис. 2.9); карти.

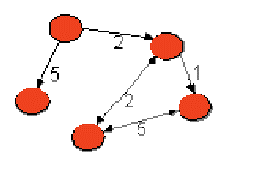


Рис. 2.9. Приклад графу

За допомогою карт, наприклад, можна відстежити зміни об’єктів в часі і

просторі, визначити характер їх розподілу на площині або в просторі.

Перевагою графічного представлення даних є велика простота їх сприйняття, ніж, наприклад, табличних даних.

Приклад карти, що є картою Кохонена (моделлю нейронних мереж), представлений на рис. 2.10.

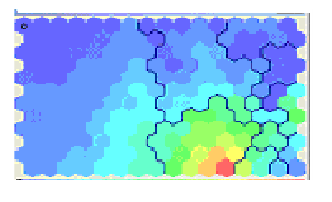


Рис. 2.10. Приклад даних типу «Карта Кохонена» хімічні дані

Хімічні дані представляють собою особливий тип даних. Приклад таких даних: Молекула бензолу C6H6

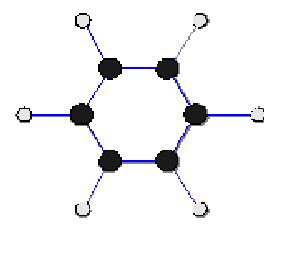


Рис. 2.11. Приклад хімічних даних

**Формати зберігання даних**

Одна з основних особливостей даних сучасного світу полягає в тому, що їх стає дуже багато.

*Можливі чотири аспекти роботи з даними:*

* визначення даних;
* обчислення;
* маніпулювання;
* обробка (збір, передача тощо).

При маніпулюванні даними використовується структура даних типу «файл». Файли можуть мати різні формати.

Більшість інструментів Data Mining дозволяють імпортувати дані з різних джерел, а також експортувати результуючі дані в різні формати.

Дані для експериментів зручно зберігати в якомусь одному форматі.

У деяких інструментах Data Mining ці процедури називаються імпорт / експорт даних, інші дозволяють напряму відкривати різні джерела даних і зберігати результати Data Mining в одному із запропонованих форматів.

Найбільш поширені формати, згідно з опитуванням «Формати зберігання даних», представлені на рис. 2.12.

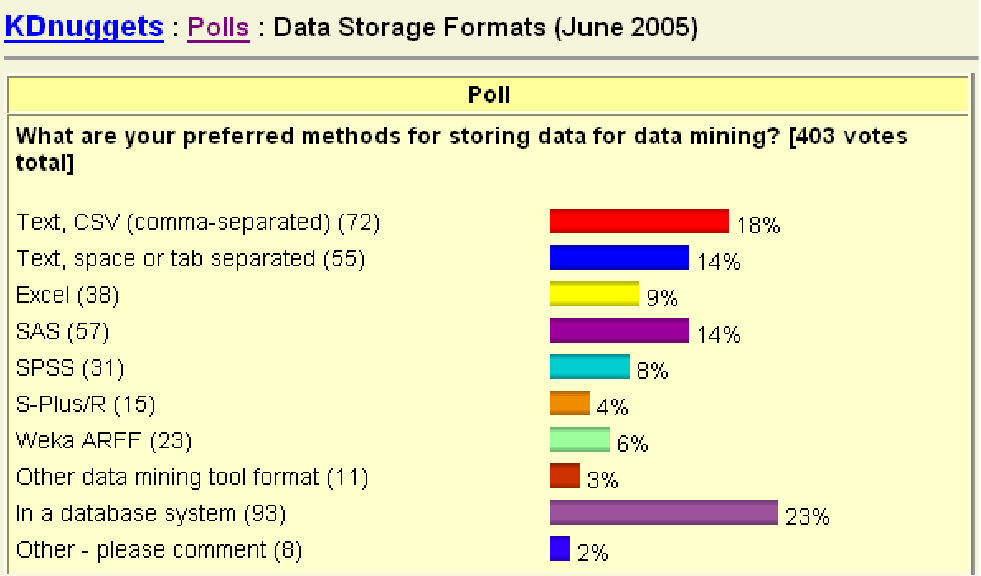


Рис. 2.12. Найбільш поширені формати зберігання даних

Найбільше число опитаних (23%) вважають за краще зберігати дані у форматі тієї бази даних, яку вони використовують. У форматі Text, CSV – 18%, по 14% опитаних зберігають дані у форматі Text, space or tab separated і SAS; в форматі Excel – 9%, SPSS – 8%, S–Plus/R – 4%, Weka ARFF – 6%, в ​​інших форматах інструментів Data Mining – 2%.

Як бачимо з результатів опитування, найбільш поширеним форматом зберігання даних для Data Mining виступають бази даних.

***Лекція 2.2. (2 год)***

**Тема 2. Методи інтелектуального аналізу даних**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Процес Data Mining
2. Методи та стадії Data Mining
3. Задачі класифікації та кластеризації
4. Задачі прогнозування
5. Задачі візуалізації

***1. ПРОЦЕС DATA MINING. МЕТОДИ ТА СТАДІЇ DATA MINING***

**Класифікація стадій Data Mining**

Основна особливість Data Mining – це поєднання широкого математичного інструментарію (від класичного статистичного аналізу до нових кібернетичних методів) і останніх досягнень у сфері інформаційних технологій. У технології Data Mining гармонійно об’єдналися строго формалізовані методи і методи неформального аналізу, тобто кількісний та якісний аналіз даних.

До методів і алгоритмів Data Mining належать наступні: штучні нейронні мережі, дерева рішень, символьні правила, методи найближчого сусіда і k – найближчого сусіда, метод опорних векторів, байєсовські мережі, лінійна регресія, кореляційно–регресійний аналіз; ієрархічні методи кластерного аналізу, неієрархічні методи кластерного аналізу, в тому числі алгоритми k – середніх і k – медіани; методи пошуку асоціативних правил, у тому числі алгоритм Apriori; метод обмеженого перебору, еволюційне програмування і генетичні алгоритми, різноманітні методи візуалізації даних і безліч інших методів.

Більшість аналітичних методів, що використовуються в технології Data Mining – це відомі математичні алгоритми і методи. Новим у їх застосуванні є можливість їх використання при вирішенні тих чи інших конкретних задач, обумовлена можливостями нових технічних і програмних засобів. Слід зазначити, що більшість методів Data Mining були розроблені в рамках теорії штучного інтелекту.

**Метод** (method) являє собою норму або правило, певний шлях, спосіб, прийом рішень задачі теоретичного, практичного, пізнавального, управлінського характеру.

Поняття алгоритму з’явилося задовго до створення електронних обчислювальних машин. Зараз алгоритми є основою для вирішення багатьох прикладних і теоретичних завдань у різних сферах людської діяльності, у більшості – це завдання, вирішення яких передбачено з використанням комп’ютера.

**Алгоритм** (algorithm) – точний припис щодо послідовності дій (кроків), що перетворюють вихідні дані в шуканий результат.

Data Mining може складатися з двох або трьох стадій:

**Стадія 1**. Виявлення закономірностей (**вільний пошук**).

**Стадія 2**. Використання виявлених закономірностей для передбачення невідомих значень (**прогностичне моделювання**).

На додаток до цих стадій іноді вводять **стадію** **валідації**, наступну за стадією вільного пошуку. **Мета валідації** – перевірка достовірності знайдених закономірностей. Однак, ми будемо вважати валідацію частиною першої стадії, оскільки в реалізації багатьох методів, зокрема, нейронних мереж і дерев рішень, передбачено поділ загальної множини даних на навчальну і перевірочну, і останнє дозволяє перевіряти достовірність отриманих результатів.

**Стадія 3**. Аналіз винятків – стадія призначена для виявлення і пояснення аномалій, знайдених в закономірностях.

Отже, процес Data Mining може бути представлений рядом таких послідовних стадій:

**ВІЛЬНИЙ ПОШУК** (у тому числі ВАЛІДАЦІЯ) →

→ **ПРОГНОСТИЧНЕ МОДЕЛЮВАННЯ** →

**→ АНАЛІЗ ВИНЯТКІВ**

*1. Вільний пошук (Discovery).*

На стадії вільного пошуку здійснюється дослідження набору даних з метою пошуку прихованих закономірностей. Попередні гіпотези щодо виду закономірностей тут не визначаються.

**Закономірність** (law) – істотний і постійно повторюваний взаємозв’язок, що визначає етапи і форми процесу становлення, розвитку різних явищ або процесів. Система Data Mining на цій стадії визначає шаблони, для отримання яких в системах **OLAP**, наприклад, аналітик повинен обдумувати і створювати безліч запитів. Тут же аналітик звільняється від такої роботи – шаблони шукає за нього система. Особливо корисне застосування даного підходу в надвеликих базах даних, де вловити закономірність шляхом створення запитів досить складно, для цього потрібно перепробувати безліч різноманітних варіантів.

**Вільний пошук** представлений такими **діями**:

– виявлення закономірностей умовної логіки (conditional logic);

– виявлення закономірностей асоціативної логіки (associations and affinities);

– виявлення трендів і коливань (trends and variations).

*Припустимо, є база даних кадрового агенства з даними про професії, стаж, вік і бажаний рівень винагороди. У разі самостійного задавання запитів аналітик може отримати приблизно такі результати: середній бажаний рівень винагороди фахівців у віці від 25 до 35 років дорівнює 1200 умовних одиниць. У разі вільного пошуку система сама шукає закономірності, необхідно лише задати цільову змінну. В результаті пошуку закономірностей система сформує набір логічних правил « якщо..., то...».*

*Можуть бути знайдені, наприклад, такі закономірності*

*«Якщо вік <20 років і бажаний рівень винагороди >700 умовних одиниць, то в 75% випадків здобувач шукає роботу програміста»*

*або*

*«Якщо вік > 35 років і бажаний рівень винагороди >1200 умовних одиниць, то в 90% випадків здобувач шукає роботу керівника». Цільовою змінною в описаних правилах виступає професія.*

*Задавши іншу цільову змінну, наприклад, вік, отримуємо такі правила: «Якщо здобувач шукає керівну роботу і його стаж >15 років, то вік здобувача >35 років у 65% випадків «*.

Описані дії, в рамках стадії вільного пошуку, виконуються за допомогою:

– індукції правил умовної логіки (задачі класифікації та кластеризації, опис в компактній формі близьких або схожих груп об’єктів);

– індукції правил асоціативної логіки (задачі асоціації та послідовності і одержувана за їх допомогою інформація);

– визначення трендів і коливань (вихідний етап задачі прогнозування).

На стадії вільного пошуку також повинна здійснюватися валідація закономірностей, тобто перевірка їх достовірності на частині даних, які не брали участь у формуванні закономірностей. Такий прийом розділення даних на навчальну і перевірочну множину часто використовується в методах нейронних мереж і дерев рішень.

*2. Прогностичне моделювання (Predictive Modeling).*

Друга стадія Data Mining – прогностичне моделювання – використовує результати роботи першої стадії. Тут виявлені закономірності використовуються безпосередньо для прогнозування.

Прогностичне моделювання включає такі дії:

– передбачення невідомих значень (outcome prediction);

– прогнозування розвитку процесів (forecasting).

У процесі прогностичного моделювання вирішуються завдання класифікації і прогнозування.

При вирішенні задачі класифікації результати роботи першої стадії (індукції правил) використовуються для віднесення нового об’єкта, з певною впевненістю, до одного з відомих, визначених класів на підставі відомих значень.

При вирішенні задачі прогнозування результати першої стадії (визначення тренда або коливань) використовуються для передбачення невідомих (пропущених або ж майбутніх) значень цільової змінної (змінних).

*Продовжуючи розглянутий приклад першої стадії, можемо зробити наступний висновок.*

*Знаючи, що здобувач шукає керівну роботу і його стаж > 15 років, на 65 % можна бути впевненим у тому, що вік здобувача > 35 років. Або ж, якщо вік здобувача > 35 років і бажаний рівень винагороди > 1200 умовних одиниць, на 90 % можна бути впевненим у тому, що здобувач шукає керівну роботу.*

Порівняння вільного пошуку і прогностичного моделювання з точки зору логіки.

Вільний пошук розкриває загальні закономірності. Він за своєю природою індуктивний. Закономірності, отримані на цій стадії, формуються від особистого до загального. У результаті ми отримуємо деяке загальне знання про деякий клас об’єктів на підставі дослідження окремих представників цього класу.

*Правило: « Якщо вік здобувача < 20 років і бажаний рівень винагороди > 700 умовних одиниць, то в 75 % випадків здобувач шукає роботу програміста» На підставі особистого, тобто інформації про деякі властивості класу « вік < 20 років» і « бажаний рівень винагороди > 700 умовних одиниць», ми робимо висновок про загальне, а саме: шукають роботу – програмісти.*

Прогностичне моделювання, навпаки, дедуктивне. Закономірності, отримані на цій стадії, формуються від загального до особистого і одиничного. Тут ми отримуємо нове знання про деякий об’єкт або ж групи об’єктів на підставі:

– знання класу, до якого належать досліджувані об’єкти;

– знання загального правила, діючого в межах даного класу об’єктів.

*Знаємо, що претендент шукає керівну роботу і його стаж > 15 років, на 65 % можна бути впевненим у тому, що вік здобувача > 35 років.*

*На підставі деяких загальних правил, а саме: мета здобувача – керівна робота і його стаж > 15 років, ми робимо висновок про одиничне – вік здобувача > 35 років.*

Слід зазначити, що отримані закономірності, а точніше, їх конструкції, можуть бути прозорими, тобто допускають тлумачення аналітика (розглянуті вище правила), і непрозорими, так званими «чорними ящиками». Типовий приклад останньої конструкції – нейронна мережа.

*3. Аналіз винятків (forensic analysis).*

На третій стадії Data Mining аналізуються виключення або аномалії, виявлені в знайдених закономірностях.

**Дія**, що виконується на цій стадії, – виявлення відхилень (deviation detection). Для виявлення відхилень необхідно визначити норму, яка розраховується на стадії вільного пошуку.

*Повернемося до одного з прикладів, розглянутих вище.*

*Знайдено правило « Якщо вік > 35 років і бажаний рівень винагороди > 1200 умовних одиниць, то в 90 % випадків здобувач шукає керівну роботу». Виникає питання – до чого віднести решту 10 % випадків? Тут можливі два варіанти. Перший з них – існує деякий логічне пояснення, яке також може бути оформлено у вигляді правила. Другий варіант для решти 10 % – це помилки вихідних даних. У цьому випадку стадія аналізу винятків може бути використана в якості очищення даних.*

Далі ми розглянемо кілька відомих класифікацій методів Data Mining за різними ознаками.

**Класифікація технологічних методів Data Mining**

Всі методи Data Mining поділяються на дві великі групи за принципом роботи з вихідними навчальними даними. У цій класифікації верхній рівень визначається на підставі того, зберігаються дані після Data Mining чи вони дистилюються для подальшого використання.

1. Безпосереднє використання даних, або збереження даних.

У цьому випадку вихідні дані зберігаються в явному деталізованому вигляді і безпосередньо використовуються на стадіях прогностичного моделювання та/або аналізу винятків. Проблема цієї групи методів – при їх використанні можуть виникнути складності аналізу надвеликих баз даних.

Методи цієї групи: **кластерний аналіз, метод найближчого сусіда, метод k – найближчого сусіда, міркування за аналогією.**

2. Виявлення і використання формалізованих закономірностей, або дистиляція шаблонів.

При технології дистиляції шаблонів один зразок (шаблон) інформації витягується з вихідних даних і перетворюється в якісь формальні конструкції, вид яких залежить від використовуваного методу Data Mining. Цей процес виконується на стадії вільного пошуку, у першій же групі методів дана стадія в принципі відсутня. На стадіях прогностичного моделювання та аналізу винятків використовуються результати стадії вільного пошуку, вони значно компактніше самих баз даних. Нагадаємо, що конструкції цих моделей можуть бути трактовані аналітиком або не трактовані («чорні ящики»).

Методи цієї групи: **логічні методи, методи візуалізації; методи крос – табуляції; методи, засновані на рівняннях.**

**Логічні методи, або методи логічної індукції, включають:**

* нечіткі запити і аналізи;
* символьні правила;
* дерева рішень;
* генетичні алгоритми.

Методи цієї групи є, мабуть, такими, що найкраще інтерпретуються – вони оформляють знайдені закономірності, в більшості випадків, у досить прозорому вигляді з точки зору користувача. Отримані правила можуть включати безперервні і дискретні змінні. Слід зауважити, що дерева рішень можуть бути легко перетворені в набори символьних правил шляхом генерації одного правила по шляху від кореня дерева до його термінальної вершини. Дерева рішень і правила фактично є різними способами вирішення однієї задачі і відрізняються лише за своїми можливостями. Крім того, реалізація правил здійснюється більш повільними алгоритмами, ніж індукція дерев рішень.

Методи крос–табуляції: агенти, баесовські (довірчі) мережі, крос – таблична візуалізація. Останній метод не зовсім відповідає одній з властивостей Data Mining – самостійного пошуку закономірностей аналітичною системою. Однак, надання інформації у вигляді крос – таблиць забезпечує реалізацію основного завдання Data Mining – пошук шаблонів, тому цей метод можна також вважати одним з методів Data Mining.

**Методи на основі рівнянь.** Методи цієї групи висловлюють виявлені закономірності у вигляді математичних виразів – рівнянь. Отже, вони можуть працювати лише з чисельними змінними, і змінні інших типів повинні бути закодовані відповідним чином. Це дещо обмежує застосування методів даної групи, проте вони широко використовуються при вирішенні різних завдань, особливо завдань прогнозування.

Основні методи цієї групи: **статистичні методи і нейронні мережі.**

**Статистичні методи** найбільш часто застосовуються для вирішення задач прогнозування. Існує безліч методів статистичного аналізу даних, серед них, наприклад, кореляційно – регресійний аналіз, кореляція рядів динаміки, виявлення тенденцій динамічних рядів, гармонійний аналіз.

Інша класифікація поділяє все різноманіття методів Data Mining на дві групи: **статистичні** та **кібернетичні** методи. Ця схема поділу заснована на різних підходах до навчання математичних моделей.

Слід зазначити, що існує два підходи віднесення статистичних методів до Data Mining. Перший з них протиставляє статистичні методи і Data Mining, його прихильники вважають класичні статистичні методи окремим напрямом аналізу даних. Відповідно до другого підходу, статистичні методи аналізу є частиною математичного інструментарію Data Mining. Більшість авторитетних джерел дотримується другого підходу.

**У цій класифікації розрізняють дві групи методів:**

– *статистичні методи*, засновані на використанні усередненого накопиченого досвіду, який відображений в ретроспективних даних;

– *кібернетичні методи*, що включають безліч різнорідних математичних підходів.

Недолік такої класифікації: і статистичні, і кібернетичні алгоритми тим чи іншим чином спираються на зіставлення статистичного досвіду з результатами моніторингу поточної ситуації.

Перевагою такої класифікації є її зручність для інтерпретації – вона використовується при описі математичних засобів сучасного підходу до вилучення знань з масивів вихідних спостережень (оперативних і ретроспективних), тобто в задачах Data Mining.

**Статистичні методи Data mining.** Ці методи являють собою чотири взаємопов’язаних розділи:

– попередній аналіз природи статистичних даних (перевірка гіпотез стаціонарності, нормальності, незалежності, однорідності, оцінка виду функції розподілу, її параметрів тощо);

– виявлення зв’язків і закономірностей (лінійний і нелінійний регресійний аналіз, кореляційний аналіз та ін);

– багатовимірний статистичний аналіз (лінійний і нелінійний дискримінантний аналіз, кластерний аналіз, компонентний аналіз, факторний аналіз та ін);

– динамічні моделі і прогноз на основі часових рядів.

**Арсенал статистичних методів Data Mining класифікований на чотири групи методів:**

1. Дескриптивний аналіз і опис вихідних даних.

2. Аналіз зв’язків (кореляційний та регресійний аналіз, факторний аналіз, дисперсійний аналіз).

3. Багатовимірний статистичний аналіз (компонентний аналіз, дискримінантний аналіз, багатовимірний регресійний аналіз, канонічні кореляції та ін.).

4. Аналіз часових рядів (динамічні моделі і прогнозування).

**Кібернетичні методи Data Mining.** Інший напрямок Data Mining – це безліч підходів, об’єднаних ідеєю комп’ютерної математики та використання теорії штучного інтелекту.

До цієї групи відносяться такі методи:

–штучні нейронні мережі (розпізнавання, кластеризація, прогнозування);

– еволюційне програмування (в т.ч. алгоритми методу групового обліку аргументів);

– генетичні алгоритми (оптимізація);

– асоціативна пам’ять (пошук аналогів, прототипів);

– нечітка логіка;

– дерева рішень;

– системи обробки експертних знань.

Методи Data Mining також можна класифікувати за задачами Data Mining. Відповідно до такої класифікації виділяємо дві групи. Перша з них – це підрозділ методів Data Mining на вирішальні завдання сегментації (тобто задачі класифікації та кластеризації) і завдання прогнозування.

У відповідності до другої класифікації по задачах методи Data Mining можуть бути спрямовані на отримання описових і прогнозуючих результатів.

Описові методи служать для знаходження шаблонів або зразків, що описують дані, які піддаються інтерпретації з точки зору аналітика.

До методів, спрямованих на отримання описових результатів, відносяться ітеративні методи кластерного аналізу, в тому числі: алгоритм k – середніх, k – медіани, ієрархічні методи кластерного аналізу, самоорганізуються карти Кохонена, методи крос – табличної візуалізації, різні методи візуалізації та інші.

Прогнозуючі методи використовують значення одних змінних для передбачення / прогнозування невідомих (пропущених) або майбутніх значень інших (цільових) змінних.

До методів, спрямованих на отримання прогнозуючих результатів, відносяться такі методи: нейронні мережі, дерева рішень, лінійна регресія, метод найближчого сусіда, метод опорних векторів та ін.

**Властивості методів Data Mining**

Різні методи Data Mining характеризуються певними властивостями, які можуть бути визначальними при виборі методу аналізу даних. Методи можна порівнювати між собою, оцінюючи характеристики їх властивостей.

Серед основних властивостей і характеристик методів Data Mining розглянемо такі: точність, масштабованість, інтерпретованість, здатність до перевірки, трудомісткість, гнучкість, швидкість і популярність.

**Масштабованість** – властивість обчислювальної системи, яка забезпечує передбачуваний ріст системних характеристик, наприклад, швидкості реакції, загальної продуктивності та ін., при додаванні до неї обчислювальних ресурсів.

Більшість інструментів Data Mining, пропонованих зараз на ринку програмного забезпечення, реалізують відразу кілька методів, наприклад, дерева рішень, індукцію правил і візуалізацію, або ж нейронні мережі, самоорганізовувані карти Кохонена та візуалізацію. В універсальних прикладних статистичних пакетах (наприклад, SPSS, SAS, STATGRAPHICS, Statistica, ін) реалізується широкий спектр найрізноманітніших методів (як статистичних, так і кібернетичних). Слід враховувати, що для можливості їх використання, а також для інтерпретації результатів роботи статистичних методів (кореляційного, регресійного, факторного, дисперсійного аналізу та ін) потрібні спеціальні знання в галузі статистики.

Універсальність того чи іншого інструмента часто накладає певні обмеження на його можливості. Перевагою використання таких універсальних пакетів є можливість відносно легко порівнювати результати побудованих моделей, отримані різними методами. Така можливість реалізована, наприклад, в пакеті Statistica, де порівняння засноване на так званій «конкурентній оцінці моделей». Ця оцінка полягає в застосуванні різних моделей до одного і того ж набору даних і в наступному порівнянні їх характеристик для вибору найкращої з них.

**Основні методи.** Кілька основних методів, які використовуються для інтелектуального аналізу даних, описують тип аналізу і операцію з відновлення даних. На жаль, різні компанії і рішення не завжди використовують одні й ті ж терміни, що може посилити плутанину і складність, що здається.

Розглянемо деякі ключові методи і приклади того, як використовувати ті чи інші інструменти для інтелектуального аналізу даних.

Асоціація (або відношення), ймовірно, найбільш відомий, знайомий і простий метод інтелектуального аналізу даних. Для виявлення моделей робиться просте зіставлення двох або більше елементів, часто одного і того ж типу. Наприклад, відстежуючи звички покупця, можна помітити, що разом з полуницею зазвичай купують вершки.

Створити інструменти інтелектуального аналізу даних на базі асоціацій або відносин неважко. Наприклад, в InfoSphere Warehouse є майстер, який видає конфігурації інформаційних потоків для створення асоціацій, досліджуючи джерело вхідної інформації, базис прийняття рішень і вихідну інформацію. [На рис 2.5](http://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques/#fig2). наведено відповідний приклад для зразка бази даних.

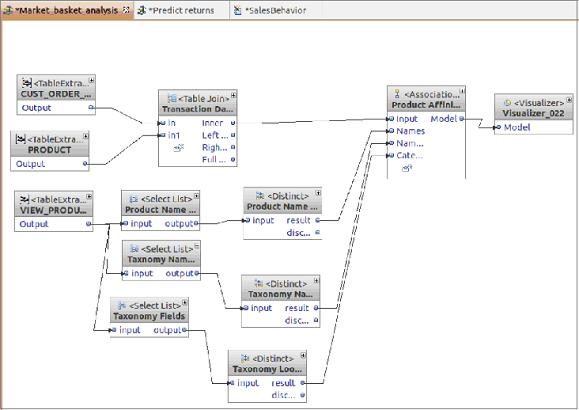


Рис. 2.5. Інформаційний потік, який використовується при підході асоціації

Класифікацію можна використовувати для отримання уявлення про тип покупців, товарів або об’єктів, описуючи кілька атрибутів для ідентифікації певного класу. Наприклад, автомобілі легко класифікувати по типу (седан, позашляховик, кабріолет), визначивши різні атрибути (кількість місць, форма кузова, ведучі колеса). Вивчаючи новий автомобіль, можна віднести його до певного класу, порівнюючи атрибути з відомим визначенням. Ті ж принципи можна застосувати і до покупців, наприклад, класифікуючи їх за віком та соціальною групою.

Крім того, класифікацію можна використовувати в якості вхідних даних для інших методів. Наприклад, для визначення класифікації можна застосовувати дерева прийняття рішень. Кластеризація дозволяє використовувати загальні атрибути різних класифікацій з метою виявлення кластерів.

Досліджуючи один або більше атрибутів або класів, можна згрупувати окремі елементи даних разом, отримуючи структурований висновок. На простому рівні при кластеризації використовується один або кілька атрибутів в якості основи для визначення кластера подібних результатів. Кластеризація корисна при визначенні різної інформації, тому що вона корелюється з іншими прикладами, так що можна побачити, де подібності і діапазони узгоджуються між собою.

Метод кластеризації працює в обидві сторони. Можна припустити, що в певній точці мається кластер, а потім використовувати свої критерії ідентифікації, щоб перевірити це. Графік, зображений на рис. [2](http://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques/#fig3).2., демонструє наочний приклад. Тут вік покупця порівнюється з вартістю покупки. Розумно очікувати, що люди у віці від двадцяти до тридцяти років (до вступу в шлюб і появи дітей), а також в 50–60 років (коли діти покинули будинок) мають більш високий наявний дохід.

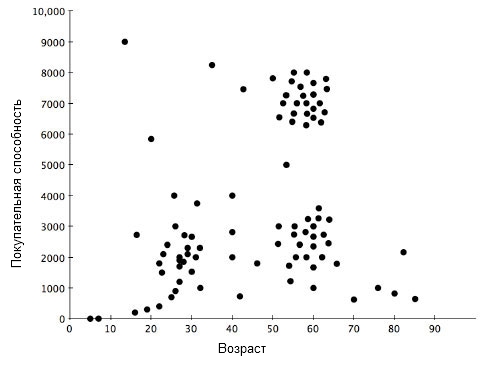


Рис. 2.6. Кластеризація

У цьому прикладі видно два кластери, один в районі $ 2000/20-30 років і інший в районі $7000-8000/50-65 років. У даному випадку ми висунули гіпотезу і перевірили її на простому графіку, який можна побудувати за допомогою будь-якого відповідного ПЗ для побудови графіків. Для більш складних комбінацій потрібен повний аналітичний пакет, особливо якщо потрібно автоматично засновувати рішення на інформації про найближчого сусіда.

Така побудова кластерів являє собою спрощений приклад так званого образу найближчого сусіда. Окремих покупців можна розрізняти за їх буквальною близькістю один до одного на графіку. Досить імовірно, що покупці з одного і того ж кластеру поділяють і інші загальні атрибути, і це припущення можна використовувати для пошуку, класифікації та інших видів аналізу членів набору даних.

Метод кластеризації можна застосувати і у зворотний бік: враховуючи певні вхідні атрибути, виявляти різні артефакти. Наприклад, недавнє дослідження чотиризначних PIN-кодів виявили кластери чисел у діапазонах 1-12 і 1-31 для першої та другої пар. Зобразивши ці пари на графіку, можна побачити кластери, пов’язані з датами (дні народження, ювілеї).

Прогнозування – це широка тема, яка простягається від передбачення відмов компонентів обладнання до виявлення шахрайства і навіть прогнозування прибутку компанії. У поєднанні з іншими методами інтелектуального аналізу даних прогнозування передбачає аналіз тенденцій, класифікацію, зіставлення з моделлю і відносини. Аналізуючи минулі події або примірники, можна передбачати майбутнє.

*Наприклад, використовуючи дані по авторизації кредитних карт, можна об’єднати аналіз дерева рішень минулих транзакцій людини з класифікацією і зіставленням з історичними моделями з метою виявлення шахрайських транзакцій. Якщо, наприклад, купівля авіаквитків збігається з транзакціями, то цілком імовірно, що ці транзакції справжні.*

Послідовні моделі, які часто використовуються для аналізу довгострокових даних, – корисний метод виявлення тенденцій, або регулярних повторень подібних подій.

*Наприклад, за даними про покупців можна визначити, що в різний час року вони купують певні набори продуктів. За цією інформацією додаток прогнозування купівельної корзини, ґрунтуючись на частоті та історії покупок, може автоматично припустити, що в кошик будуть додані ті чи інші продукти.*

Дерево рішень, пов’язане з більшістю інших методів (головним чином, класифікації та прогнозування), можна використовувати або в рамках критеріїв відбору, або для підтримки вибору певних даних в рамках загальної структури. Дерево рішень починають з простого питання, яке має дві відповіді (іноді більше). Кожна відповідь призводить до наступного питання, допомагаючи класифікувати та ідентифікувати дані або робити прогнози.

На [рис. 2.7](http://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques/#fig4) наведено приклад класифікації несправних станів.

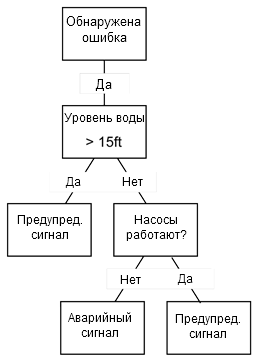


Рис. 2.7. Дерево рішень

Дерева рішень часто використовуються з системами класифікації інформації про властивості і з системами прогнозування, де різні прогнози можуть ґрунтуватися на минулому історичному досвіді, який допомагає побудувати структуру дерева рішень і отримати результат.

На практиці дуже рідко використовується тільки один з цих методів. Класифікація і кластеризація – подібні методи. Використовуючи кластеризацію для визначення найближчих сусідів, можна додатково уточнити класифікацію. Дерева рішень часто використовуються для побудови і виявлення класифікацій, які можна простежувати на історичних періодах для визначення послідовностей і моделей.

При всіх основних методах часто має сенс записувати і згодом вивчати отриману інформацію. Для деяких методів це абсолютно очевидно. Наприклад, при побудові послідовних моделей та навчанні з метою прогнозування аналізуються історичні дані з різних джерел і примірників інформації.

В інших випадках цей процес може бути більш яскраво вираженим. Дерева рішень рідко будуються один раз і ніколи не забуваються. При виявленні нової інформації, подій і точок даних може знадобитися побудова додаткових гілок або навіть зовсім нових дерев.

Деякі з цих процесів можна автоматизувати. Наприклад, побудова прогностичної моделі для виявлення шахрайства з кредитними картами зводиться до визначення ймовірностей, які можна використовувати для поточної транзакції, з подальшим оновленням цієї моделі при додаванні нових (підтверджених) транзакцій. Потім ця інформація реєструється, так що наступного разу рішення можна буде прийняти швидше.

Підготовка даних і очищення даних – часто забувають, але надзвичайно важливий крок у процесі «видобутку даних». У типових проєктах «видобутку даних» великі набори даних, зібрані за допомогою деяких автоматичних методів (наприклад, за допомогою Web), служать вхідними даними аналізу. Часто метод, за допомогою якого були зібрані дані, що не був жорстко регульованим, внаслідок чого дані можуть містити значення, що виходять за допустимі межі (наприклад, Дохід: –100), неможливі комбінації даних (наприклад, Стать: Чоловік, Вагітність: Так) і ін. Аналіз даних, що не були ретельно екрановані на предмет подібних ситуацій, може дати вводять в оману результати, особливо при [пророкує видобутку даних](http://www.statsoft.ru/home/textbook/glossary/GlossaryTwo/P/PredictiveDataMining.htm).

При [видобутку даних](http://www.statsoft.ru/home/textbook/glossary/GlossaryTwo/D/DataMining.htm) вхідні дані часто «зашумлені» – містять багато помилок і іноді інформацію в неструктурованою формі (наприклад, при [видобутку тексту](http://www.statsoft.ru/home/textbook/glossary/GlossaryTwo/T/TextMining.htm) ). Припустимо, що ви хочете проаналізувати велику базу даних, зібраних за допомогою Web в режимі он-лайн, ґрунтуючись на добровільних відповідях людей, які відвідують ваш Web-сайт (наприклад, потенційних клієнтів Web-продавця, який заповнив запропоновані анкети). У цьому прикладі дуже важливо спочатку перевірити і «очистити» дані на стадії підготовки даних, перед тим як застосовувати будь аналітичні процедури. Наприклад, деякі індивідууми можуть нарочито або випадково ввести завідомо помилкову інформацію (наприклад, вік = 300). У таких типах даних помилки не виявляються до стадії аналізу. Вони можуть сильно зміщувати результат і приводити до невиправданих висновків. Типово протягом стадії підготовки даних аналітик застосовує «фільтри» до даних для перевірки правильності їх діапазонів і виключення неможливих значень (наприклад, Вік = 5; Пенсіонер = Так).

***2. ЗАДАЧІ КЛАСИФІКАЦІЇ ТА КЛАСТЕРИЗАЦІЇ***

**Задачі класифікації**

**Класифікація** є найбільш простою і водночас найбільш часто розв’язуваною задачею Data Mining. Зважаючи на поширеність задач класифікації необхідно чітке розуміння суті цього поняття.

Наведемо кілька визначень.

**Класифікація** – системний розподіл досліджуваних предметів, явищ, процесів за родами, видами, типами, з якими-небудь істотними ознаками для зручності їх дослідження; угрупування вихідних понять і розташування їх у певному порядку, що відбиває ступінь цієї схожості.

**Класифікація** – впорядкована за деяким принципом множина об’єктів, які мають подібні класифікаційні ознаки (одна або декілька властивостей), обраних для визначення схожості або відмінності між цими об’єктами.

**Класифікація вимагає дотримання наступних правил:**

1. в кожному акті ділення необхідно застосовувати тільки одну основу;
2. ділення повинне бути пропорційним, тобто загальний обсяг видових понять повинен дорівнювати об’єму діленого родового поняття;
3. члени ділення повинні взаємно виключати один одного, їх об’єми не повинні перехрещуватися;
4. ділення повинне бути послідовним.

**Розрізняють:**

1. **допоміжну (штучну)** **класифікацію**, яка виробляється за зовнішньою ознакою і служить для надання множині предметів (процесів, явищ) потрібного порядку;
2. **природну** **класифікацію**, яка виробляється за істотними ознаками, що характеризують внутрішню спільність предметів і явищ. Вона є результатом і важливим засобом наукового дослідження, тому що передбачає і закріплює результати вивчення закономірностей об’єктів, що класифікуються.

Допоміжна класифікація створюється з метою найбільш швидкого відшукування якогось індивідуального предмету серед предметів, що класифікуються. Мета в цій класифікації визначає принцип її побудови. В основу допоміжної класифікації лягає яка-небудь зовнішня несуттєва ознака, яка, однак, виявляється корисною у процесі пошуку.

**Прикладами допоміжної класифікації** можуть бути розподіл студентів курсу в списку в алфавітному порядку або такий же розподіл бібліотечних карток в алфавітному каталозі і т.п. Знаючи порядок букв в алфавіті, ми можемо легко і швидко відшукати потрібне нам прізвище у списку або дані, що цікавлять нас в книзі, в каталозі.

Але знання того, яке місце в допоміжній класифікаційній системі займає той чи інший предмет, не дає можливості щось стверджувати про його властивості. Так, наприклад, те що студент Архипов записаний у списку першим, а студент Яковлєв – останнім, нічого не говорить про їх здібності і риси характеру . Тому допоміжна класифікація не є науковою.

На відміну від допоміжної **природна класифікація** являє собою розподіл предметів за класами на підставі їх найбільш суттєвих ознак. Найбільш істотними є такі ознаки предмета, які обумовлюють інші його ознаки. Наприклад, найбільш суттєвою ознакою людини є її здатність до праці. Ця ознака зумовлює наявність у людини таких ознак, як прямоходіння, здатність до спілкування (праця передбачає колектив), здатність до мислення та ін.

**Залежно від обраних ознак**, їх поєднання і процедури розподілу понять, **класифікація може бути**:

* **простою** – розподіл родового поняття тільки за ознакою і тільки один раз до розкриття всіх видів. Прикладом такої класифікації є дихотомія, при якій членами поділу бувають тільки два поняття, кожне з яких суперечить іншому (тобто дотримується принцип: «А і не А»);
* **складною** – застосовується для поділу одного поняття за різними основами і синтезу таких простих ділень в єдине ціле.

**Прикладом** такої класифікації є періодична система хімічних елементів.

Під **класифікацією** будемо розуміти віднесення об’єктів (спостережень, подій) до одного з заздалегідь відомих класів.

**Класифікація** – це закономірність, що дозволяє робити висновок щодо визначення характеристик конкретної групи. Таким чином, для проведення класифікації повинні бути присутні ознаки, що характеризують групу, до якої належить та чи інша подія або об’єкт (зазвичай при цьому на підставі аналізу вже класифікованих подій формулюються якісь правила).

**Класифікація** відноситься до стратегії навчання з вчителем (supervised learning), яку також іменують контрольованим або керованим навчанням.

**Маши́нне навчання** — узагальнена назва штучної генерації знань з досвіду. [Штучна система](http://uk.wikipedia.org/w/index.php?title=%D0%A8%D1%82%D1%83%D1%87%D0%BD%D0%B0_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0&action=edit&redlink=1) навчається на [прикладах](http://uk.wikipedia.org/wiki/%D0%9F%D1%80%D0%B8%D0%BA%D0%BB%D0%B0%D0%B4) і після закінчення фази [навчання](http://uk.wikipedia.org/wiki/%D0%9D%D0%B0%D0%B2%D1%87%D0%B0%D0%BD%D0%BD%D1%8F) може узагальнювати. Тобто [система](http://uk.wikipedia.org/wiki/%D0%A1%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0) не просто вивчає наведені приклади, а розпізнає певні закономірності в даних для навчання.

Серед багатьох програмних продуктів варто згадати [системи автоматичного діагностування](http://uk.wikipedia.org/w/index.php?title=%D0%A1%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B8_%D0%B0%D0%B2%D1%82%D0%BE%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%BD%D0%BE%D0%B3%D0%BE_%D0%B4%D1%96%D0%B0%D0%B3%D0%BD%D0%BE%D1%81%D1%82%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F&action=edit&redlink=1), розпізнавання шахрайства з [кредитними картками](http://uk.wikipedia.org/wiki/%D0%9A%D1%80%D0%B5%D0%B4%D0%B8%D1%82%D0%BD%D0%B0_%D0%BA%D0%B0%D1%80%D1%82%D0%BA%D0%B0), [аналіз](http://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B0%D0%BB%D1%96%D0%B7) [ринку цінних паперів](http://uk.wikipedia.org/wiki/%D0%A0%D0%B8%D0%BD%D0%BE%D0%BA_%D1%86%D1%96%D0%BD%D0%BD%D0%B8%D1%85_%D0%BF%D0%B0%D0%BF%D0%B5%D1%80%D1%96%D0%B2), [класифікація](http://uk.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D0%B8%D1%84%D1%96%D0%BA%D0%B0%D1%86%D1%96%D1%8F) ланцюжків [ДНК](http://uk.wikipedia.org/wiki/%D0%94%D0%9D%D0%9A), [розпізнавання мовлення](http://uk.wikipedia.org/wiki/%D0%A0%D0%BE%D0%B7%D0%BF%D1%96%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F_%D0%BC%D0%BE%D0%B2%D0%BB%D0%B5%D0%BD%D0%BD%D1%8F) та [тексту](http://uk.wikipedia.org/w/index.php?title=%D0%A0%D0%BE%D0%B7%D0%BF%D1%96%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D1%96%D0%B2&action=edit&redlink=1), [автономні системи](http://uk.wikipedia.org/w/index.php?title=%D0%90%D0%B2%D1%82%D0%BE%D0%BD%D0%BE%D0%BC%D0%BD%D1%96_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B8&action=edit&redlink=1).

Практичне використання відбувається, переважно, за допомогою [алгоритмів](http://uk.wikipedia.org/wiki/%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC). Різноманітні алгоритми машинного навчання можна грубо поділити за такою схемою:

* [**Навчання з вчителем**](http://uk.wikipedia.org/w/index.php?title=%D0%9D%D0%B0%D0%B2%D1%87%D0%B0%D0%BD%D0%BD%D1%8F_%D0%B7_%D0%B2%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D0%B5%D0%BC&action=edit&redlink=1) ([англ.](http://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0) Supervised learning): алгоритм вивчає функцію на основі наданих пар вхідних та вихідних даних. При цьому, в процесі навчання, «вчитель» вказує вірні вихідні дані для кожного значення вхідних даних. Одним з розділів навчання з вчителем є [машинна класифікація](http://uk.wikipedia.org/w/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%B0_%D0%BA%D0%BB%D0%B0%D1%81%D0%B8%D1%84%D1%96%D0%BA%D0%B0%D1%86%D1%96%D1%8F&action=edit&redlink=1). Такі алгоритми застосовуються для [розпізнавання текстів](http://uk.wikipedia.org/w/index.php?title=%D0%A0%D0%BE%D0%B7%D0%BF%D1%96%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D1%96%D0%B2&action=edit&redlink=1).
* [**Навчання без вчител**я](http://uk.wikipedia.org/w/index.php?title=%D0%9D%D0%B0%D0%B2%D1%87%D0%B0%D0%BD%D0%BD%D1%8F_%D0%B1%D0%B5%D0%B7_%D0%B2%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D1%8F&action=edit&redlink=1) ([англ.](http://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0) Unsupervised learning).
* [**Навчання з закріпленням**](http://uk.wikipedia.org/w/index.php?title=%D0%9D%D0%B0%D0%B2%D1%87%D0%B0%D0%BD%D0%BD%D1%8F_%D0%B7_%D0%B7%D0%B0%D0%BA%D1%80%D1%96%D0%BF%D0%BB%D0%B5%D0%BD%D0%BD%D1%8F%D0%BC&action=edit&redlink=1) ([англ.](http://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0) Reinforcement Learning): алгоритм навчається за допомогою тактики нагороди та покарання для максимізації вигоди для агентів (систем до яких належить компонента, що навчається)

Завданням класифікації часто називають передбачення категоріальної залежної змінної (тобто залежної змінної, що є категорією) на основі вибірки безперервних і/або категоріальних змінних.

*Наприклад,* можна передбачити, хто з клієнтів фірми є потенційним покупцем певного товару, а хто – ні, хто скористається послугою фірми, а хто – ні, і т.д. Цей тип завдань належить до завдань **бінарної класифікації**, в них залежна змінна може приймати тільки два значення (наприклад, так чи ні, 0 або 1).

Інший варіант класифікації виникає, якщо залежна **змінна** може приймати значення з деякої множини визначених класів. Наприклад, коли необхідно передбачити, яку марку автомобіля захоче купити клієнт. У цих випадках розглядається множина класів для залежної змінної.

Класифікація може бути **одновимірною** (за однією ознакою) і **багатовимірною** (за двома і більше ознаками).

**Багатовимірна класифікація** була розроблена біологами при вирішенні проблем дискримінації для класифікування організмів. Однією з перших робіт, присвячених цьому напрямку, вважають роботу Р. Фішера (1930 р.), в якій організми поділялися на підвиди залежно від результатів вимірювань їх фізичних параметрів. Біологія була і залишається найбільш затребуваним і зручним середовищем для розробки багатовимірних методів класифікації.

Розглянемо задачу класифікації на простому прикладі. Припустимо, є база даних про клієнтів туристичного агентства з інформацією про вік і доході за місяць. Є рекламний матеріал двох видів: більш дорогий і комфортний відпочинок і дешевший, молодіжний відпочинок. Відповідно, визначені два класи клієнтів: клас 1 і клас 2. База даних наведена в таблиці 2.8.

Таблиця 2.8.

База даних клієнтів туристичного агентства.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
| **Код клієнта** | **Вік** | **Дохід** | **Клас** |
| **1** | 18 | 25 | 1 |
| **2** | 22 | 100 | 1 |
| **3** | 30 | 70 | 1 |
| **4** | 32 | 120 | 1 |
| **5** | 24 | 15 | 2 |
| **6** | 25 | 22 | 1 |
| **7** | 32 | 50 | 2 |
| **8** | 19 | 45 | 2 |
| **9** | 22 | 75 | 1 |
| **10** | 40 | 90 | 2 |

**Завдання**. Визначити, до якого класу належить новий клієнт і який з двох видів рекламних матеріалів йому варто відсилати.

Для наочності представимо нашу базу даних у двомірному просторі (вік і дохід), у вигляді множини об’єктів, що належать класам 1 (помаранчева мітка) і 2 (сіра мітка). На [рис. 2.14](http://www.intuit.ru/studies/professional_skill_improvements/1210/courses/6/lecture/83?page=1#image.5.1) наведені об’єкти з двох класів.

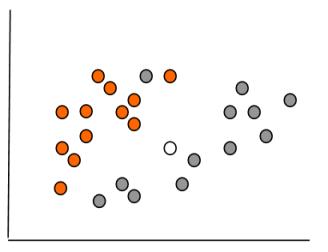


Рис. 2.14.Множина об’єктів бази даних у двомірному вимірі

Розв’язок нашої задачі буде полягати в тому, щоб визначити, до якого класу належить новий клієнт, на малюнку позначений білою міткою.

**Мета процесу** **класифікації** полягає в тому, щоб побудувати модель, яка використовує прогнозуючі атрибути в якості вхідних параметрів і отримує значення залежного атрибута. **Процес** **класифікації** полягає в розбитті множини об’єктів на класи за певним критерієм.

**Класифікатором** називається якась сутність, що визначає, якому з визначених класів належить об’єкт за вектором ознак.

Для проведення класифікації за допомогою математичних методів необхідно мати формальний опис об’єкта, яким можна оперувати, використовуючи математичний апарат класифікації. Таким описом у нашому випадку виступає база даних. Кожен об’єкт (запис бази даних) несе інформацію про деякі властивості об’єкта.

Набір вихідних даних (або вибірку даних) розбивають на **дв****і множини**: **навчальну** і **тестову**.

**Навчальна множина** **(****training set)** – множина, яка включає дані, що використовуються для навчання (конструювання) моделі.

Така множина містить вхідні та вихідні (цільові) значення прикладів. Вихідні значення призначені для навчання моделі.

**Тестова (test** **set) множина** також містить вхідні та вихідні значення прикладів. Тут вихідні значення використовуються для перевірки працездатності моделі.

Процес класифікації складається з **двох етапів**: **конструювання** **моделі** та її **використання**.

1. **Конструювання моделі:** опис множини визначених класів.

* Кожен приклад набору даних відноситься до одного визначеного класу.
* На цьому етапі використовується навчальна множина, на ньому відбувається конструювання моделі.
* Отримана модель **представлена класифікаційними правилами, деревом рішень або математичною формулою.**

1. **Використання моделі:** класифікація нових або невідомих значень.

* Оцінка правильності (точності) моделі.
* Відомі значення з тестового прикладу порівнюються з результатами використання отриманої моделі.
* Рівень точності – відсоток правильно класифікованих прикладів у тестовій множині.
* Тестова множина, тобто множина, на якій тестується побудована модель, не повинна залежати від навчальної множини.
* Якщо точність моделі допустима, можливе використання моделі для класифікації нових прикладів, клас яких невідомий.

Процес класифікації, а саме, конструювання моделі та її використання, представлений на [рис. 2.15.](http://www.intuit.ru/studies/professional_skill_improvements/1210/courses/6/lecture/83?page=2#image.5.2) – рис. 2.16.

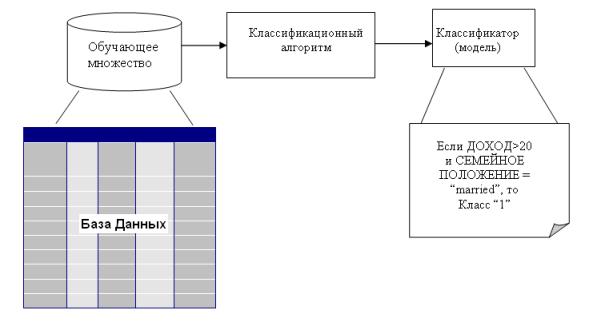


Рис. 2.15. Процес класифікації. Конструювання моделі

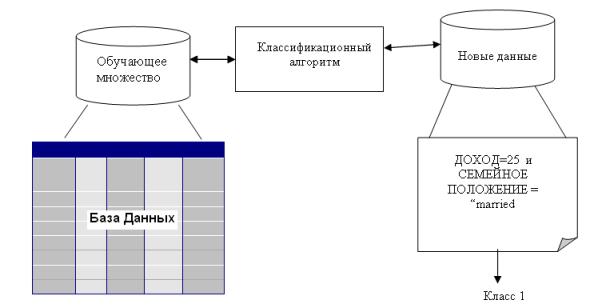


Рис. 2.16. Процес класифікації. Використання моделі

**Методи, що застосовуються для вирішення задач класифікації**

Для класифікації використовуються різні методи. Основні з них:

* Класифікація за допомогою **дерев рішень**;
* Байєсівська (наївна) класифікація;
* Класифікація за допомогою штучних нейронних мереж;
* Класифікація методом опорних векторів;
* Статистичні методи, зокрема, **лінійна регресія**;
* Класифікація за допомогою **методу найближчого сусіда**;
* Класифікація cbr – методом;
* Класифікація за допомогою генетичних алгоритмів.

Схематичне рішення задачі класифікації деякими методами (за допомогою лінійної регресії, дерев рішень і нейронних мереж) наведені на рис. 2.17.-2.19.

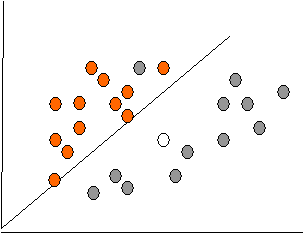


Рис. 2.17. Рішення задачі класифікації методом лінійної регресії

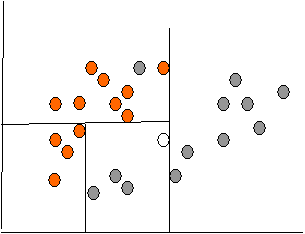


Рис. 2.18. Рішення задачі класифікації методом дерев рішень

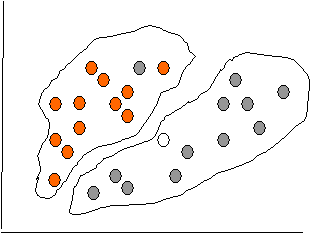


Рис. 2.19. Рішення задачі класифікації методом нейронних мереж

**Точність класифікації: оцінка рівня помилок.** Оцінка точності класифікації може проводитися за допомогою крос–перевірки. Крос–перевірка (Cross-validation) – це процедура оцінки точності класифікації на даних з тестової множини, яку також називають крос–перевірочною множиною. Точність класифікації тестової множини порівнюється з точністю класифікації навчаючої множини. Якщо класифікація тестової множини дає приблизно такі ж результати по точності, як і класифікація навчальної множини, вважається, що дана модель пройшла крос-перевірку.

Поділ на навчальну і тестову множину здійснюється шляхом ділення вибірки в певній пропорції, наприклад навчальна множина – дві третини даних і тестова – одна третина даних. Цей спосіб слід використовувати для вибірок з великою кількістю прикладів. Якщо ж вибірка має малі обсяги, рекомендується застосовувати спеціальні методи, при використанні яких навчальна і тестова вибірки можуть частково перетинатися.

**Оцінювання класифікаційних методів.**Оцінювання методів слід проводити, виходячи з таких характеристик: **швидкість,** **робастність, інтерпретованість,** **надійність.**

**Швидкість** характеризує час, який потрібен на створення моделі та її використання.

**Робастність**, тобто стійкість до будь–яких порушень вихідних передумов, означає можливість роботи з зашумленими даними і пропущеними значеннями в даних.

**Інтерпретованість** забезпечує можливість розуміння моделі аналітиком.

Властивості класифікаційних правил:

* розмір дерева рішень;
* компактність класифікаційних правил.

**Надійність** методів класифікації передбачає можливість роботи цих методів при наявності в наборі даних шумів і викидів.

**Завдання кластеризації**

Тільки що ми вивчили завдання класифікації, що відноситься до стратегії «навчання з учителем».

У цій частині лекції ми введемо поняття кластеризації, кластера, коротко розглянемо класи методів, за допомогою яких вирішується завдання кластеризації, деякі моменти процесу кластеризації, а також розберемо приклади застосування кластерного аналізу.

Завдання кластеризації схоже з завданням класифікації, є його логічним продовженням, але його відмінність в тому, що класи досліджуваного набору даних заздалегідь не зумовлені.

Синонімами терміну «кластеризація» є «автоматична класифікація», «навчання без вчителя» і «таксономія».

**Кластеризація** призначена для розбиття сукупності об’єктів на однорідні групи (кластери або класи). Якщо дані вибірки представити як точки в просторі ознак, то завдання кластеризації зводиться до визначення «згущувань точок».

Мета кластеризації – пошук існуючих структур. Кластеризація є описовою процедурою, вона не робить ніяких статистичних висновків, але дає можливість провести розвідувальний аналіз і вивчити «структуру даних».

Саме поняття «**кластер**» визначене неоднозначно. Перекладається поняття кластер (cluster) як «скупчення», «гроно».

**Кластер** можна охарактеризувати як групу об’єктів, що мають загальні властивості.

Характеристиками кластера можна назвати дві ознаки:

* внутрішня однорідність;
* зовнішня ізольованість.

Питання, що ставиться аналітиками при вирішенні багатьох завдань, полягає в тому, як організувати дані в наочні структури, тобто розгорнути таксономії.

Найбільше застосування кластеризація спочатку отримала в таких науках як біологія, антропологія, психологія. Для вирішення економічних завдань кластеризація тривалий час мало використовувалася через специфіку економічних даних і явищ.

У табл. 2.9. наведено порівняння деяких параметрів задач класифікації та кластеризації.

Таблиця 2.9.

Порівняння класифікації та кластерзаціі

|  |  |  |
| --- | --- | --- |
|  | | |
| **Характеристика** | **Класифікація** | **Кластеризація** |
| **Контрольованість навчання** | Контрольоване навчання | Неконтрольоване навчання |
| **Стратегія** | Навчання з вчителем | Навчання без вчителя |
| **Наявність позначки класу** | Навчальна множина супроводжується міткою, що вказує клас, до якого належить спостереження | Мітки класу навчальної множини невідомі |
| **Підстава для класифікації** | Нові дані класифікуються на підставі навчальної множини | Дано множину даних з метою встановлення існування класів або кластерів даних |

На рис. 2.20. схематично представлені завдання класифікації і кластеризації.

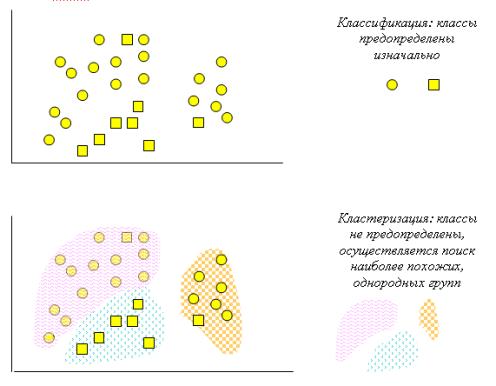


Рис. 2.20. Порівняння задач класифікації та кластеризації

Кластери можуть бути **такими, що не перетинаються**, або **ексклюзивними** (non-overlapping, exclusive), і такими, що **перетинаються** (overlapping). Схематичне зображення таких кластерів дано на рис. 2.21.

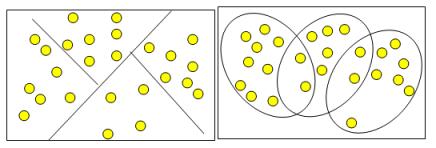


Рис. 2.21. Кластери, що не перетинаються і перетинаються

Слід зазначити, що в результаті застосування різних методів кластерного аналізу можуть бути отримані кластери різної форми. Наприклад, можливі кластери «ланцюжкового» типу, коли кластери представлені довгими «ланцюжками», кластери подовженої форми і т.д., а деякі методи можуть створювати кластери довільної форми.

Різні методи можуть прагнути створювати кластери певних розмірів (наприклад, малих або великих) або припускати в наборі даних наявність кластерів різного розміру.

Деякі методи кластерного аналізу особливо чутливі до шумів або викидів, інші – менш.

В результаті застосування різних методів кластеризації можуть бути отримані неоднакові результати, це нормально і є особливістю роботи того чи іншого алгоритму.

Дані особливості слід враховувати при виборі методу кластеризації.

Детальніше про всі властивості кластерного аналізу буде розказано в лекції, присвяченій його методам.

На сьогоднішній день розроблено більше сотні різних алгоритмів кластеризації. Деякі, найбільш часто використовувані, будуть детально описані в наступних лекціях.

Наведемо коротку характеристику підходів до кластеризації.

* *Алгоритми, засновані на поділі даних (Partitioning algorithms), в тому числі ітеративні:*
  + поділ об’єктів на k кластерів;
  + ітеративний перерозподіл об’єктів для поліпшення кластеризації.
* *Ієрархічні алгоритми (**Hierarchy Algorithms):*
* агломерація: кожен об’єкт спочатку є кластером, кластери, з’єднуючись один з одним, формують більший кластер і т.д.
* *Методи, засновані на концентрації об’єктів (**Density-based methods):*
  + засновані на можливості з’єднання об’єктів;
  + ігнорують шуми, знаходження кластерів довільної форми.
* *Грід-методи (Grid-based methods):*
  + квантування об’єктів в грід-структури.
* *Модельні методи (Model-based):*
  + використання моделі для знаходження кластерів, найбільш відповідних даним.

**Оцінка якості** **кластеризації** може бути проведена на основі таких процедур:

* ручна перевірка;
* встановлення контрольних точок та перевірка на отриманих кластерах;
* визначення стабільності кластеризації шляхом додавання в модель нових змінних;
* створення і порівняння кластерів з використанням різних методів. Різні методи кластеризації можуть створювати різні кластери, і це є нормальним явищем. Однак створення схожих кластерів різними методами вказує на правильність кластеризації.

**Процес** **кластеризації** залежить від обраного методу і майже завжди є ітеративним. Він може стати захоплюючим процесом і включати множину експериментів з вибору різноманітних параметрів, наприклад, міри відстані, типу стандартизації змінних, кількості кластерів і т.д. Однак експерименти не повинні бути самоціллю – адже кінцевою метою кластеризації є отримання змістовних відомостей про структуру досліджуваних даних. Отримані результати вимагають подальшої інтерпретації, дослідження і вивчення властивостей і характеристик об’єктів для можливості точного опису сформованих кластерів.

**Застосування кластерного аналізу**

Кластерний аналіз застосовується в різних областях. Він корисний, коли потрібно класифікувати велику кількість інформації. Огляд багатьох опублікованих досліджень, що проводяться за допомогою кластерного аналізу, дав Хартіган (Hartigan, 1975).

Так, в медицині використовується кластеризація захворювань, лікування захворювань або їх симптомів, а також таксономія пацієнтів, препаратів і т.д. В археології встановлюються таксономії кам’яних споруд і древніх об’єктів і т.д. У маркетингу це може бути задача сегментації конкурентів і споживачів. У менеджменті прикладом завдання кластеризації буде розбиття персоналу на різні групи, класифікація споживачів і постачальників, виявлення схожих виробничих ситуацій, при яких виникає шлюб. У медицині – класифікація симптомів. У соціології завдання кластеризації – розбиття респондентів на однорідні групи.

***3. ЗАДАЧІ ПРОГНОЗУВАННЯ***

**Задачі прогнозування** вирішуються в найрізноманітніших областях людської діяльності, таких як наука, економіка, виробництво й безліч інших сфер. Прогнозування є важливим елементом організації керування як окремими господарюючими суб’єктами, так і економіки в цілому.

Розвиток методів прогнозування безпосередньо пов’язаний з розвитком інформаційних технологій, зокрема, з ростом обсягів збережених даних і ускладненням методів і алгоритмів прогнозування, реалізованих в інструментах Data Mining.

Завдання прогнозування, мабуть, може вважатися однією з найбільш складних задач Data Mining, воно вимагає ретельного дослідження вихідного набору даних і методів, що підходять для аналізу.

**Прогнозування** ( від грецького Prognosis), у широкому розумінні цього слова, визначається як випереджаюче відображення майбутнього**.**

**Метою прогнозування** є передбачення майбутніх подій.

**Прогнозування** (forecasting) є однією з задач Data Mining і одночасно одним із ключових моментів при прийнятті рішень.

**Прогностика** (prognostics) – теорія й практика прогнозування.

**Прогнозування** спрямоване на визначення тенденцій динаміки конкретного об’єкта або події на основі ретроспективних даних, тобто аналізу його стану колись і тепер. Таким чином, розв’язок задачі прогнозування вимагає деякої навчальної вибірки даних.

**Прогнозування** – установлення функціональної залежності між залежними й незалежними змінними.

Прогнозування є розповсюдженим і затребуваним завданням у багатьох областях людської діяльності. У результаті прогнозування зменшується ризик прийняття невірних, необґрунтованих або суб’єктивних рішень.

**Приклади його задач**: прогноз руху грошових коштів, прогнозування врожайності агрокультури, прогнозування фінансової стабільності підприємства.

Крім економічної й фінансової сфери, задачі прогнозування постають в найрізноманітніших областях: медицині, фармакології; популярним зараз стає політичне прогнозування.

**Загалом розв’язок задачі прогнозування зводиться до розв’язку таких підзадач:**

* вибір моделі прогнозування;
* аналіз адекватності й точності побудованого прогнозу.

**Порівняння задач прогнозування і класифікації.**

У попередній темі нами було розглянуто задачу класифікації. Прогнозування подібне із задачею класифікації.

Багато методів Data Mining використовуються для розв’язку задач класифікації і прогнозування. Це, наприклад, лінійна регресія, нейронні мережі, дерева рішень (які іноді так і називають – дерева прогнозування й класифікації).

Завдання класифікації й прогнозування мають подібності й відмінності.

**Так у чому ж подібність завдань прогнозування й класифікації?** При розв’язку обох завдань використовується двоетапний процес побудови моделі на основі навчального набору і її використання для пророкування невідомих значень залежної змінної.

**Відмінність задач класифікації й прогнозування** полягає в тому, що в першій задачі передбачається клас залежної змінної, а в другій – числові значення залежної змінної, пропущені або невідомі (які відносяться до майбутнього).

Повертаючись до прикладу про туристичне агентство, розглянутого у попередній лекції, ми можемо сказати, що визначення класу клієнта є розв’язком задачі класифікації, а прогнозування доходу, який принесе цей клієнт наступного року, буде розв’язком задачі прогнозування.

**Прогнозування і часові ряди**

**Прогнозування і часові ряди.** Основою для прогнозування служить історична інформація, що зберігається в базі даних у вигляді **часових рядів**.

Існує поняття Data Mining **часових рядів** (Time-Series Data Mining).

На основі ретроспективної інформації у вигляді часових рядів можливий розв’язок різних задач Data Mining.

На рис. 2.22. представлені результати опитування відносно Data Mining часових рядів. Як бачимо, найбільший відсоток (23%) серед розв’язуваних задач займає прогнозування. Далі йдуть класифікація і кластеризація (по 14%), сегментація й виявлення аномалій (по 9%), виявлення правил (8%). На інші задачі доводиться менш, ніж по 6%.

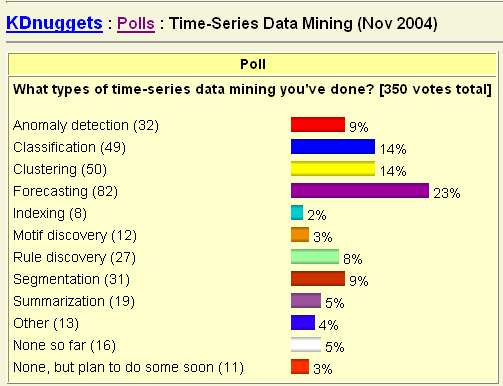


Рис. 2.22. Data Mining часових рядів

Однак щоб зосередитися на понятті прогнозування, ми будемо розглядати часові ряди лише в рамках розв’язку задачі прогнозування.

Приведемо **дві принципові відмінності часового ряду** **від простої послідовності спостережень:**

– Члени часового ряду, на відміну від елементів випадкової вибірки, не є статистично незалежними.

– Члени часового ряду не є однаково розподіленими.

**Часовий ряд** – послідовність спостережуваних значень якої–небудь ознаки, упорядкованих у невипадкові моменти часу.

Відмінністю аналізу часових рядів від аналізу випадкових вибірок є припущення про рівні проміжки часу між спостереженнями і їх хронологічний порядок. Прив’язка спостережень до часу відіграє тут ключову роль, тоді як при аналізі випадкової вибірки вона не має ніякого значення.

Типовий приклад часового ряду – дані біржових торгів.

**Інформація,** накопичена в різноманітних базах даних підприємства, **є** часовими рядами, якщо вона розташована в хронологічному порядку й зроблена в послідовні моменти часу.

**Аналіз часового ряду здійснюється з метою:**

* + визначення природи ряду;
  + прогнозування майбутніх значень ряду.

У процесі визначення структури й закономірностей часового ряду передбачається виявлення: шумів і викидів, тренду, сезонного компонента, циклічного компонента. Визначення природи часового ряду може бути використане як своєрідна «розвідка» даних. Знання аналітика про наявність сезонного компонента необхідне, наприклад, для визначення кількості записів вибірки, яка повинна брати участь у побудові прогнозу.

**Шуми й викиди** будуть докладно обговорюватися в наступних темах курсу. Вони ускладнюють аналіз часового ряду. Існують різні методи визначення й фільтрації викидів, що дають можливість виключити їх з метою більш якісного Data Mining.

**Тренд, сезонність і цикл**

Основними складовими часового ряду є тренд і сезонний компонент. Складові цих рядів можуть являти собою або тренд, або сезонний компонент.

Тренд є систематичним компонентом часового ряду, який може змінюватися в часі.

**Трендом** називають невипадкову функцію, яка формується під дією загальних або довгочасних тенденцій, що впливають на часовий ряд.

Прикладом тенденції може виступати, наприклад, фактор росту досліджуваного ринку.

Автоматичного способу виявлення трендів у часових рядах не існує. Але якщо часовий ряд включає монотонний тренд (тобто відзначене його стійке зростання або стійке спадання), аналізувати часовий ряд у більшості випадків неважко.

Існує велика різноманітність постановок задач прогнозування, які можна підрозділити на дві групи: прогнозування односерійних рядів і прогнозування мультисерійних, або взаємовпливаючих, рядів.

Група прогнозування односерійних рядів включає задачу побудови прогнозу однієї змінної за ретроспективним даними тільки цієї змінної, без врахування впливу інших змінних і факторів.

Група прогнозування мультисерійних, або взаємовплваючих, рядів включає задачу аналізу, де необхідно враховувати взаємовпливаючі фактори на одну або декілька змінних.

Крім розподілу на класи по односерійності й багатосерійності, ряди також бувають сезонними й несезонними.

Останній розподіл має на увазі наявність або відсутність у часового ряду такої складової як сезонність, тобто включення сезонного компонента.

Сезонна складова часового ряду є періодично повторюваним компонентом часового ряду.

Властивість сезонності означає, що через приблизно рівні проміжки часу форма кривої, яка описує поведінку залежної змінної, повторює свої характерні обриси.

Властивість сезонності важлива при визначенні кількості ретроспективних даних, які будуть використовуватися для прогнозування.

*Розглянемо простий приклад.* На рис. 2.23. наведений фрагмент ряду, який ілюструє поведінку змінної «обсяги продажу товару Х» за період, що становить один місяць. При вивченні кривої, наведеної на малюнку, аналітик не може зробити припущень щодо повторюваності форми кривої через рівні проміжки часу.

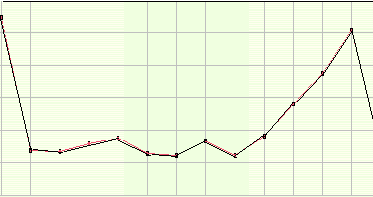


Рис. 2.23. Фрагмент часового ряду за сезонний період

Однак при розгляді більш тривалого ряду (за 12 місяців), зображеного на рис. 2.24., можна побачити явну наявність сезонного компонента. Отже, про сезонність продажів можна говорити тільки, коли розглядаються дані за кілька місяців.

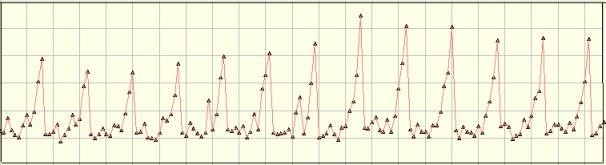


Рис. 2.24. Фрагмент часового ряду з 12–ти сезонних періодів

Таким чином, у процесі підготовки даних для прогнозування аналітикові слід визначити, чи володіє ряд, який він аналізує, властивістю сезонності.

Визначення наявності компоненти сезонності необхідно для того, щоб вхідна інформація мала властивість репрезентативності.

Ряд можна вважати несезонним, якщо при розгляді його зовнішнього вигляду не можна зробити припущень про повторюваність форми кривої через рівні проміжки часу.

Іноді по зовнішньому вигляду кривої ряду не можна визначити, є він сезонним чи ні.

**Існує поняття сезонного мультиряду.** У ньому кожний ряд описує поведінку факторів, які впливають на залежну (цільову) змінну.

Приклад такого ряду – ряди продажів декількох товарів, підданих сезонним коливанням.

При зборі даних і виборі факторів для розв’язку задачі по прогнозуванню в таких випадках слід ураховувати, що вплив обсягів продажів товарів один на одного тут набагато менше, ніж вплив фактору сезонності.

Важливо не плутати поняття сезонного компонента ряду й сезонів природи. Незважаючи на близькість їх звучання, ці поняття відрізняються. Так, наприклад, обсяги продажів морозива влітку набагато більше, ніж в інші сезони, однак це є тенденцією попиту на даний товар.

Дуже часто тренд і сезонність присутні в часовому ряді одночасно.

*Приклад.* Прибуток фірми росте протягом декількох років (тобто в часовому ряді присутній тренд); ряд також містить сезонний компонент.

**Відмінності циклічного компонента від сезонного:**

1. Тривалість циклу, як правило, більше, ніж один сезонний період;
2. Цикли, на відміну від сезонних періодів, не мають певної тривалості.

При виконанні яких–небудь перетворень зрозуміти природу часового ряду значно простіше, такими перетвореннями можуть бути, наприклад, видалення тренда й згладжування ряду.

**Перед початком прогнозування необхідно відповісти на наступні питання:**

1. Що потрібно прогнозувати?
2. У яких часових елементах (параметрах)?
3. З якою точністю прогнозу?

При відповіді на перше питання, ми визначаємо змінні, які будуть прогнозуватися. Це може бути, наприклад, рівень проведення конкретного виду продукції в наступному кварталі, прогноз суми продажу цієї продукції і т.д.

При виборі змінних слід ураховувати доступність ретроспективних даних, переваги осіб, що ухвалюють рішення, остаточну вартість Data Mining.

Часто при розв’язку задач прогнозування виникає необхідність пророкування не самої змінної, а зміни її значень.

**Друге питання при розв’язку задачі прогнозування – визначення наступних параметрів:**

* + періоду прогнозування;
  + горизонту прогнозування;
  + інтервалу прогнозування.

**Період прогнозування** – основна одиниця часу, на яку робиться прогноз.

*Наприклад*, ми прагнемо довідатися дохід компанії через місяць. Період прогнозування для цієї задачі – місяць.

**Горизонт прогнозування** – це число періодів у майбутньому, які покриває прогноз.

Якщо ми прагнемо дізнатися прогноз на 12 місяців уперед, з даними по кожному місяцю, то період прогнозування в цьому завданні – місяць, горизонт прогнозування – 12 місяців.

**Інтервал прогнозування** – частота, з якою робиться новий прогноз. Інтервал прогнолзування може збігатися з періодом прогнозування.

**Рекомендації з вибору параметрів прогнозування.** При виборі параметрів необхідно враховувати, що горизонт прогнозування повинен бути не менше, ніж час, який необхідний для реалізації розв’язку, прийнятого на основі цього прогнозу. Тільки в цьому випадку прогнозування буде мати сенс.

Зі збільшенням горизонту прогнозування точність прогнозу, як правило, знижується, а зі зменшенням горизонту – підвищується.

Ми можемо поліпшити якість прогнозування, зменшуючи час, необхідний на реалізацію розв’язку, для якого реалізується прогноз, і, отже, зменшивши при цьому горизонт і помилку прогнозування.

При виборі інтервалу прогнозування слід вибирати між двома ризиками: вчасно не визначити зміни в аналізованому процесі й високою вартістю прогнозу. При тривалому інтервалі прогнозування виникає ризик не ідентифікувати зміни, що відбувся в процесі, при короткому – зростають витрати на прогнозування.

При виборі інтервалу необхідно також ураховувати стабільність аналізованого процесу й вартість проведення прогнозу.

**Точність прогнозу**, необхідна для розв’язку конкретної задачі, дуже впливає на прогнозуючу систему. Помилка прогнозу залежить від використовуваної системи прогнозу.

Чим більше ресурсів має така система, тим більше шансів одержати більш точний прогноз. Однак прогнозування не може повністю усунути ризики при прийнятті розв’язків. Тому завжди враховується можлива помилка прогнозування.

**Види помилок та прогнозів**

Точність прогнозу характеризується помилкою прогнозу.

**Найпоширеніші види помилок:**

* + **Середня помилка (СП).** Вона обчислюється простим усередненням помилок на кожному кроціл2. Недолік цього виду помилки – позитивні й негативні помилки анулюють одна одну.
  + **Середня абсолютна помилка (САП)**. Вона розраховується як середнє абсолютних помилок. Якщо вона дорівнює нулю, то ми маємо досконалий прогноз. У порівнянні із середньою квадратичною помилкою, цей захід «не надає занадто великого значення» викидам.
  + **Сума квадратів помилок (SSE)**, средньоквадратична помилка. Вона обчислюється як сума (або середнє) квадратів помилок. Це найбільше часто використовувана оцінка точності прогнозу.
  + **Відносна помилка (ВП).** Попередні міри використовували дійсні значення помилок. Відносна помилка виражає якість припасування в термінах відносних помилок.

**Види прогнозів.** Прогноз може бути короткостроковим, середньостроковим і довгостроковим.

**Короткостроковий прогноз** являє собою прогноз на кілька кроків уперед, тобто здійснюється побудова прогнозу не більше ніж на 3% від обсягу спостережень або на 1–3 кроку вперед.

**Середньостроковий прогноз** – це прогноз на 3-5% від обсягу спостережень, але не більш 7-12 кроків уперед; також під цим типом прогнозу розуміють прогноз на один або половину сезонного циклу. Для побудови короткострокових і середньострокових прогнозів цілком підходять статистичні методи.

**Довгостроковий прогноз** – це прогноз більш ніж на 5% від обсягу спостережень.

При побудові даного типу прогнозів статистичні методи практично не використовуються, крім випадків дуже «гарних» рядів, для яких прогноз можна просто «намалювати».

Дотепер ми розглядали аспекти прогнозування, так чи інакше пов’язані із процесом ухвалення рішення. Існують і інші фактори, які необхідно враховувати при прогнозуванні.

**Задача 1.** Відомо, що аналізований процес відносно стабільний у часі, зміни відбуваються повільно, процес не залежить від зовнішніх факторів.

**Задача 2.** Аналізований процес нестабільний і дуже сильно залежить від зовнішніх факторів.

**Розв’язок першої задачі** повинен бути зосереджений на використанні великої кількості ретроспективних даних. **При розв’язку другої задачі** особливу увагу слід звернути на оцінки фахівця в предметній області, експерта, щоб мати можливість відбити в прогнозуючій моделі всі необхідні зовнішні фактори, а також приділити час для збору даних по цих факторах (збір зовнішніх даних часто набагато складніший збору внутрішніх даних інформаційної системи). Доступність даних, на основі яких буде здійснюватися прогнозування, – важливий фактор побудови прогнозної моделі. Для можливості виконання якісного прогнозу дані повинні бути представницькими, точними й достовірними.

**Методи прогнозування.** Серед розповсюджених методів Data Mining, використовуваних для прогнозування, відзначимо **нейронні мережі** й **лінійну регресію**.

Вибір методу прогнозування залежить від багатьох факторів, у тому числі від параметрів прогнозування. Вибір методу слід провадити з обліком усіх специфічних особливостей набору ретроспективних даних і цілей, з якими він будується.

Програмне забезпечення Data Mining, використовуване для прогнозування, повинне забезпечувати користувача точним і достовірним прогнозом. Однак одержання такого прогнозу залежить не тільки від програмного забезпечення й методів, закладених у його основу, але також і від інших факторів, серед яких повнота й вірогідність вихідних даних, своєчасність і оперативність їх поповнення, кваліфікація користувача.

**Завдання візуалізації. Візуалізація** – це інструментарій, який дозволяє побачити кінцевий результат обчислень, організувати керування обчислювальним процесом і навіть повернутися назад до вихідних даних, щоб визначити найбільш раціональний напрямок подальшого руху.

У результаті використання візуалізації створюється графічний образ даних. Застосування візуалізації допомагає в процесі аналізу даних побачити аномалії, структури, тренди. При розгляді завдання прогнозування ми використовували графічне представлення часового ряду й побачили, що в ньому є присутнім сезонний компонент. У попередній лекції ми розглядали завдання **класифікації й кластеризації**, і для ілюстрації розподілу об’єктів у двомірному просторі також використовували візуалізацію.

Можна говорити про те, що застосування візуалізації є більш економічним: лінія тренду або скупчення точок на діаграмі розсіювання дозволяє аналітикові набагато швидше визначити закономірності й прийти до потрібного розв’язку. Таким чином, тут мова йде про використання в Data Mining не символів, а образів.

**Головна перевага візуалізації** – практично повна відсутність необхідності в спеціальній підготовці користувача. За допомогою візуалізації ознайомитися з інформацією дуже легко, досить усього лише на неї подивитися.

Хоча найпростіші види візуалізації з’явилися досить давно, її використання зараз тільки набирає популярність. Візуалізація не спрямована винятково на вдосконалювання техніки аналізу – за словами Скотта Лейбса, у деяких випадках візуалізація може навіть замінити її.

**Візуалізація даних може бути представлена у вигляді: графіків, схем, гістограм, діаграм і т.д.**

**Коротко роль візуалізації можна описати такими її можливостями:**

* + підтримка інтерактивного й погодженого дослідження;
  + допомога у показі результатів;
  + використання очей (зору), щоб створювати зорові образи й осмислювати їх.

**Погана візуалізація.** Результати візуалізації іноді можуть вводити користувача в оману. Приведемо простий приклад поганої візуалізації. Допустимо, ми маємо базу «Прибуток компанії А» за період з 2000 по 2005 року, вона представлена в табличному виді в [табл. 2.10.](http://www.intuit.ru/department/database/datamining/6/datamining_6.html#table.6.1)

Таблиця 2.10.

Прибуток компанії А

Рік Прибуток

2000 1100

2001 1101

2002 1104

2003 1105

2004 1106

2005 1107

Побудуємо гістограмму в Excel за цим даними. Гістограмма являє собою візуальне зображення розподілу даних.

Ця інформація відображається за допомогою серії прямокутників або смуг однакової ширини, висота яких указує кількість даних у кожному класі.

Використовуючи всі значення побудови графіка, прийняті за замовчуванням, одержуємо гістограмму, наведену на рис. 2.25.

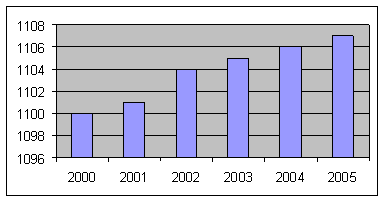


Рис. 2.25. Гістограмма, мінімальне значення осі y рівне 1096

Даний малюнок демонструє значне зростання прибутку компанії А за період з 2000 по 2005 року. Однак, якщо ми звернули увагу на вісь y, що показує величину прибутку, те побачимо, що ця вісь перетинає вісь x у значенні, рівному 1096. Фактично, вісь y зі значеннями від 1096 до 1108 вводить користувача в оману. Змінивши значення параметрів, відповідальних за формат осі y, одержуємо графік, наведений на рис. 2.26.

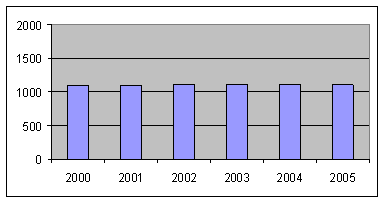


Рис. 2.26. Гістограма, мінімальне значення осі y рівне 0

Вісь у зі значеннями від 0 до 2000 дає користувачеві правильну інформацію про незначну зміну прибутків компанії.

Якщо мова йде про велику розмірність і складності вихідних даних, кошти візуалізації забезпечують їхнє різке зменшення, конденсуючи, можливо, мільйони записів даних у прості, легкі для розуміння й маніпулювання показники. Такі показники називають візуальним або графічним способом показу інформації. Візуалізацію можна вважати ключовим фактором у дослідженні даних, отриманих за допомогою інструментів Data Mining. У таких випадках говорять про візуальний Data Mining.

***4. ВІЗУАЛІЗАЦІЯ ІНСТРУМЕНТІВ DATA MINING***

Зі зростанням кількості даних, що накопичуються, навіть при використанні як завгодно потужних і різносторонніх алгоритмів Data Mining, стає усе складніше «переварювати» і інтерпретувати отримані результати. А, як відомо, одне з положень Data Mining – пошук практично корисних закономірностей. Закономірність може стати практично корисною, тільки якщо її можна осмислити й зрозуміти.

В 1987 році з ініціативи ACM SIGGRAPH IEEE Computer Society Technical Committee of Computer Graphics, у зв’язку з необхідністю використання нових методів, засобів і технологій даних, були сформульовані відповідні завдання напрямку візуалізації.

До способів візуального або графічного представлення даних відносять графіки, діаграми, таблиці, звіти, списки, структурні схеми, карти і т.д.

Візуалізація традиційно розглядалася як допоміжний засіб при аналізі даних, однак зараз усе більше досліджень говорить про її самостійну роль.

*Традиційні методи візуалізації можуть знаходити наступне застосування:*

* представляти користувачеві інформацію в наочному вигляді;
* компактно описувати закономірності, властиві вихідному набору даних;
* знижувати розмірність або стискати інформацію;
* відновлювати пробіли в наборі даних;
* знаходити шуми й викиди в наборі даних.

**Візуалізація інструментів Data Mining.** Кожний з алгоритмів Data Mining використовує певний підхід до візуалізації. У попередніх лекціях ми розглянули ряд методів Data Mining. У ході використання кожного з методів, а точніше, його програмної реалізації, ми одержували якісь візуалізатори, за допомогою яких нам вдавалося інтерпретувати результати, отримані в результаті роботи відповідних методів і алгоритмів.

Для дерев рішень це візуалізатор дерева рішень, список правил, таблиця спряженості.

Для нейронних мереж залежно від інструмента це може бути топологія мережі, графік зміни величини помилки, що демонструє процес навчання.

Для карт Кохонена: карти входів, виходів, інші специфічні карти.

Для лінійної регресії в якості візуалізатора виступає лінія регресії.

Для кластеризації: дендрограми, діаграми розсіювання.

Діаграми й графіки розсіювання часто використовуються для оцінки якості роботи того або іншого методу.

*Усі ці способи візуального представлення або відображення даних можуть виконувати одну з функцій:*

* є ілюстрацією побудови моделі (наприклад, представлення структури (графа) нейронної мережі);
* допомагають інтерпретувати отриманий результат;
* є засобом оцінки якості побудованої моделі;
* поєднують перераховані вище функції (дерево розв’язків, дендрограма).

Існує багато різних способів представлення моделей, але графічне їх представлення дає користувачеві максимальну «цінність».

Користувач, у більшості випадків, не є фахівцем у моделюванні, найчастіше він експерт у своїй предметній області. Тому модель Data Mining повинна бути представлена на найбільш природній для нього мові або, хоча б, містити мінімальну кількість різних математичних і технічних елементів.

Таким чином, доступність є однією з основних характеристик моделі Data Mining. Незважаючи на це, існує й такий розповсюджений і найбільш простий спосіб показу моделі, як «чорний ящик». У цьому випадку користувач не розуміє поведінки тієї моделі, якою користується. Однак, незважаючи на нерозуміння, він одержує результат – виявлені закономірності. Класичним прикладом такої моделі є модель нейронної мережі.

Інший спосіб представлення моделі – представлення її в інтуїтивному, зрозумілому виді. У цьому випадку користувач дійсно може розуміти те, що відбувається «усередині» моделі. Таким чином, можна забезпечити його особисту участь у процесі.

Такі моделі забезпечують користувачеві можливість обговорювати її логіку з колегами, клієнтами й іншими користувачами, або пояснювати її.

Розуміння моделі веде до розуміння її змісту. У результаті розуміння зростає довіра до моделі. Класичним прикладом є дерево рішень. Побудоване дерево рішень дійсно поліпшує розуміння моделі, тобто використовуваного інструмента Data Mining.

Крім розуміння, такі моделі забезпечують користувача можливістю взаємодіяти з моделлю, задавати їй питання й одержувати відповіді. Прикладом такої взаємодії є засіб «що, якщо». За допомогою діалогу «система–користувач» користувач може одержати розуміння моделі.

Тепер перейдемо до функцій, які допомагають інтерпретувати й оцінити результати побудови Data Mining моделей. Це всілякі графіки, діаграми, таблиці, списки і т.д.

Прикладами засобів візуалізації, за допомогою яких можна оцінити якість моделі, є діаграма розсіювання, таблиця спряженості, графік зміни величини помилки.

Діаграма розсіювання являє собою графік відхилення значень, прогнозованих за допомогою моделі, від реальних. Ці діаграми використовують для безперервних величин. Візуальна оцінка якості побудованої моделі можлива тільки по закінченню процесу побудови моделі.

Таблиця спряженості використовується для оцінки результатів класифікації. Такі таблиці застосовуються для різних методів класифікації. Оцінка якості побудованої моделі можлива тільки по закінченню процесу побудови моделі.

Графік зміни величини помилки. Графік демонструє зміну величини помилки в процесі роботи моделі. Наприклад, у процесі роботи нейронних мереж користувач може спостерігати за зміною помилки на навчальній й тестовій множинах і зупинити навчання для недопущення «перенавчання» мережі. Тут оцінка якості моделі і його зміни може оцінюватися безпосередньо в процесі побудови моделі.

Прикладами засобів візуалізації, які допомагають інтерпретувати результат, є: лінія тренду в лінійній регресії, карти Кохонена, діаграма розсіювання в кластерному аналізі.

**Методи візуалізації**

Методи візуалізації, залежно від кількості використовуваних вимірів, прийнято класифікувати на дві групи:

* представлення даних в одному, двох і трьох вимірах;
* представлення даних у чотирьох і більше вимірах.

Представлення даних в одному, двох і трьох вимірах. До цієї групи методів ставляться добре відомі способи відображення інформації, які доступні для сприйняття людською уявою. Практично будь–який сучасний інструмент Data Mining включає способи візуального представлення із цієї групи.

*Відповідно до кількості вимірів представлення це можуть бути наступні способи:*

* одномірне (univariate) або 1-D;
* двовимірне (bivariate) або 2-D;
* тривимірне, проекційне (projection) або 3-D.

Слід відзначити, що найбільше природно людське око сприймає двомірні представлення інформації.

*При використанні двох- і тривимірного представлення інформації користувач має можливість побачити закономірності набору даних:*

* його кластерну структуру й розподіл об’єктів на класи (наприклад, на діаграмі розсіювання);
* топологічні особливості;
* наявність трендів;
* інформацію про взаємне розташування даних;
* існування інших залежностей, властивих досліджуваному набору даних.

*Якщо набір даних має більше трьох вимірів, то можливі такі варіанти:*

* використання багатомірних методів представлення інформації (вони розглянуті нижче);
* зниження розмірності до одно-, двох- або тривимірного представлення. Існують різні способи зниження розмірності, один з них – факторний аналіз – був розглянутий в одній з попередніх лекцій. Для зниження розмірністі й одночасного візуального представлення інформації на двовимірних картах використовуються карти, що самоорганізуються.

**Представлення даних в чотирьох і більше вимірах.** Представлення інформації в чотиривимірному й більш вимірах недоступні для людського сприйняття. Однак розроблені спеціальні методи для можливості відображення й сприйняття людиною такої інформації.

*Найбільш відомі способи багатомірного представлення інформації:*

* паралельні координати;
* «особи Чернова»;
* пелюсткові діаграми.

**Паралельні координати.** У паралельних координатах змінні кодуються по горизонталі, вертикальна лінія визначає значення змінної. Приклад набору даних, представленого в декартових координатах і паралельних координатах, даний на рис. 2.27. Цей метод представлення багатомірних даних був винайдений Альфредом Інселбергом (Alfred Inselberg ) в 1985 році.

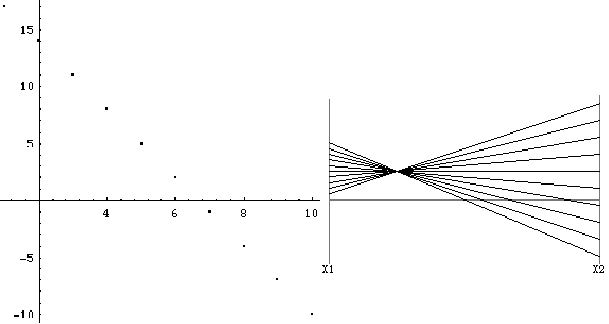


Рис. 2.27. Набір даних у декартових координатах і в паралельних координатах

**«Особи Чернова».** Основна ідея представлення інформації в «особах Чернова» полягає в кодуванні значень різних змінних у характеристиках або рисах людської особи. Приклад такого «особи» наведений на [рис](http://www.intuit.ru/department/database/datamining/16/datamining_16.html#image.16.2) 2.28.

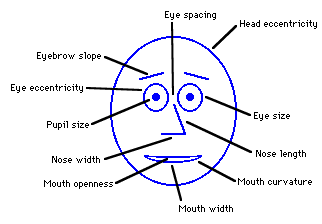


Рис. 2.28. «Особа Чернова»

Для кожного спостереження рисується окрема «особа». На кожній «особі» відносні значення змінних представлені як форми й розміри окремих рис особи (наприклад, довжина й ширина носа, розмір очей, розмір зіниці, кут між бровами).

Аналіз інформації за допомогою такого способу відображення заснований на здатності людини інтуїтивно знаходити подібності й відмінності в рисах особи.

На [рис. 2.29.](http://www.intuit.ru/department/database/datamining/16/datamining_16.html#image.16.3) представлений набір даних, кожний запис якого виражений у вигляді «Особи Чернова».

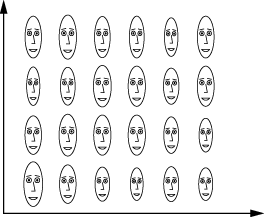


Рис. 2.29. Приклад багатомірного зображення даних за допомогою «осіб Чернова»

Перед використанням методів візуалізації необхідно:

* Проаналізувати, чи варто зображувати всі дані або ж лише якусь їхню частину.
* Вибрати розміри, пропорції й масштаб зображення.
* Вибрати метод, який може найбільше яскраво відобразити закономірності, властиві набору даних.

Багато сучасних засобів аналізу даних дозволяють будувати сотні типів різних графіків і діаграм. Тому вибір методу візуалізації, якщо він самостійно здійснюється користувачем, не такий простий і легкий, як може здатися на перший погляд. Наявність великої кількості засобів візуалізації, представлених в інструменті, який застосовує користувач, може навіть викликати розгубленість.

Ту саму інформацію можна представити за допомогою різних засобів. Для того щоб засіб візуалізації міг виконувати своє основне призначення – представляти інформацію в простому й доступному для людського сприйняття вигляді – необхідно дотримуватися законів відповідності обраного розв’язку змісту відображуваної інформації і її функціональному призначенню. Іншими словами, потрібно зробити так, щоб при погляді на візуальне представлення інформації можна було відразу виявити закономірності у вихідних даних і приймати на їхній основі рішення.

Серед двомірних і тривимірних засобів найбільше широко відомі лінійні графіки, лінійні, стовпчикові, кругові секторні й векторні діаграми.

Приведемо рекомендації з використання цих найбільш простих і популярних засобів візуалізації.

За допомогою лінійного графіка можна відобразити тенденцію, передати зміни якої–небудь ознаки в часі. Для порівняння декількох рядів чисел такі графіки наносяться на ті самі осі координат.

Гістограму застосовують для порівняння значень протягом деякого періоду або ж співвідношення величин.

Кругові діаграми використовують, якщо необхідно відобразити співвідношення частин і цілого, тобто для аналізу складу або структури явищ. Складові частини цілого зображуються секторами круга. Сектори рекомендують розміщати по їхній величині: угорі – найбільший, інші – по рухові годинної стрілки в порядку зменшення їх величини. Кругові діаграми також застосовують для відображення результатів факторного аналізу, якщо дії всіх факторів є односпрямованими. При цьому кожний фактор відображається у вигляді одного із секторів кола.

Вибір того або іншого засобу візуалізації залежить від поставленого завдання (наприклад, потрібно визначити структуру даних або ж динаміку процесу) і від характеру набору даних.

**Якість візуалізації.** Сучасні аналітичні засоби, у тому числі й Data Mining, немислимі без якісної візуалізації. У результаті використання засобів візуалізації повинні бути отримані наочні й виразні, ясні й прості зображення, за рахунок використання різноманітних засобів: кольору, контрасту, границь, пропорцій, масштабу і т.д.

У зв’язку з ростом вимог до засобів візуалізації, а також необхідності порівняння їх між собою, в останні роки був сформований ряд принципів якісної візуальної представлення інформації.

Принципи Тафта (Tufte’s Principles) графічне представлення даних високої якості говорить:

* надавайте користувачеві найбільшу кількість ідей, у найкоротший час, з найменшою кількістю чорнила на найменшому просторі;
* говоріть правду про дані.

**Принципи компонування візуальних засобів**

Основні принципи компонування візуальних засобів представлення інформації:

1. Принцип лаконічності.
2. Принцип узагальнення й уніфікації.
3. Принцип акценту на основних значеннєвих елементах.
4. Принцип автономності.
5. Принцип структурності.
6. Принцип стадійності.
7. Принцип використання звичних асоціацій і стереотипів.

Принцип лаконічності говорить про те, що засіб візуалізації повинен містити лише ті елементи, які необхідні для повідомлення користувачеві істотної інформації, точного розуміння її значення або прийняття (з імовірністю не нижче допустимої величини) відповідного оптимального розв’язку.

Крім позначених вище принципів, засіб візуалізації повинний мати високу надійність і швидкістю, яка влаштує користувача, що приймає на основі цієї інформації рішення.

**Представлення просторових характеристик.** Окремим напрямком візуалізації є наочне представлення просторових характеристик об’єктів. У більшості випадків такі засоби виділяють на карті окремі регіони й позначають їхніми різними кольорами залежно від значення аналізованого показника.

На [рис. 2.30.](http://www.intuit.ru/department/database/datamining/16/datamining_16.html#image.16.4) наведений приклад такої візуалізації в середовищі Mineset, що є, у цьому випадку, інструментом візуального Data Mining. Карта представлена у вигляді графічного інтерфейсу, що відображає дані у вигляді тривимірного ландшафту довільно визначених і позиціонованих форм (стовпчастих діаграм, кожна з індивідуальними висотою й кольором). Такий спосіб дозволяє наочно показувати кількісні й реляційні характеристики просторово–орієнтованих даних і швидко ідентифікувати в них тренди.

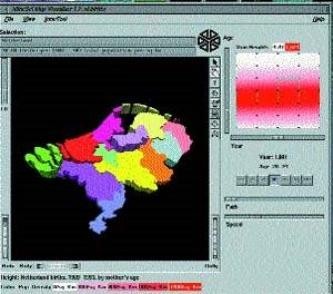


Рис. 2.30. Mineset. Ландшафтний візуалізатор

**Основні тенденції в області візуалізації**

Як ми вже відзначали, за допомогою засобів візуалізації підтримуються важливі завдання бізнесу, серед яких – процес прийняття рішень. У зв’язку із цим виникає необхідність переходу засобів візуалізації на більш якісний рівень, який характеризується появою абсолютно нових засобів візуалізації й поглядів на їх функції, а також розвитком ряду тенденцій у цій області.

Серед основних тенденцій в області візуалізації Філіп Рассом (Philip Russom) виділяє:

1. Розробка складних видів діаграм.
2. Підвищення рівня взаємодії з візуалізацією користувача.
3. Збільшення розмірів і складності структур даних, що представляються візуалізацією.

**Розробка складних видів діаграм.** Більшість візуалізацій даних побудовані на основі діаграм стандартного типу (секторні діаграми, графіки розсіювання і т.д.). Ці способи є одночасно найстаршими, найбільш елементарними й розповсюдженими. В останні роки перелік видів діаграм, підтримуваних інструментальними засобами візуалізації, суттєво розширився. Оскільки потреби користувачів досить різноманітні, інструменти візуалізації підтримують всілякі типи діаграм. Наприклад, відомо, що бізнес–користувачі віддають перевагу секторним діаграмам і гістограмам, тоді як вчених більше влаштовують візуалізації у вигляді графіків розсіювання й діаграм констеляції. Користувачі, що працюють із геопросторовими даними, сильніше зацікавлені в картах і інших тривимірних представленнях даних.

Електронні інструментальні панелі, у свою чергу, більш популярні серед керівників, що використовують бізнес–аналітичні технології для контролю над показниками роботи компанії. Такі користувачі потребують наочної візуалізації у вигляді «спідометрів», «термометрів» і «світлофорів».

Засоби створення діаграм і презентаційної графіки призначені головним чином для візуалізації даних. Однак можливості такої візуалізації звичайно вбудовані й у безліч різних інших програм і систем – в інструменти репортинга й OLAP, кошту для Text Mining і Data Mining, а також в Crm-Додатки й додатка для керування бізнесом. Для створення вбудованої візуалізації багато постачальників реалізують візуалізаційну функціональність у вигляді компонентів, що вбудовуються в різні інструменти, додатки, програми й web-сторінки ( у тому числі інструментальні панелі й персоналізовані сторінки порталів).

**Підвищення рівня взаємодії з візуалізацією користувача.** Ще зовсім недавно більша частина коштів візуалізації являла собою статичні діаграми, призначені винятково для перегляду. Зараз широко використовуються динамічні діаграми, уже самі по собі, що є користувацьким інтерфейсом, у якому користувач може прямо й інтерактивно маніпулювати візуалізацією, підбираючи нову виставу інформації.

*Наприклад,* базова взаємодія дозволяє користувачеві обертати діаграму або змінювати її тип у пошуках найбільш повної представлення даних. Крім того, користувач може міняти візуальні властивості – приміром, шрифти, кольори й рамки. У візуалізаціях складного типу (графіках розсіювання або діаграмах констеляції) користувач може вибирати інформаційні крапки за допомогою миші й переміщати їх, полегшуючи тим самим розуміння представлення даних.

Більш досконалі методи візуалізації даних часто містять у собі діаграму або будь–яку іншу візуалізацію як складений рівень. Користувач може глибшатися (drill down) у візуалізацію, досліджуючи подробиці

узагальнених нею даних, або глибшатися в OLAP, Data Mining або інші складні технології.

Складна взаємодія дозволяє користувачеві змінювати візуалізацію для знаходження альтернативних інтерпретацій даних. Взаємодія з візуалізацією має на увазі мінімальний по своїй складності користувацький інтерфейс, у якому користувач може управляти виставою даних, просто «кликая» на елементи візуалізації, перетаскуючи й поміщаючи представлення об’єктів даних або вибираючи пункти меню. Інструменти OLAP або Data Mining перетворюють безпосередня взаємодія з візуалізацією в один з етапів ітераційного аналізу даних. Кошту Text Mining або керування документами надають такій безпосередній взаємодії характер навігаційного механізму, що допомагає користувачеві досліджувати бібліотеки документів.

Візуальний запит є найбільш сучасною формою складної взаємодії користувача з даними. У ньому користувач може, наприклад, бачити крайні інформаційні крапки графіка розсіювання, вибирати їхньою мишкою й одержувати нові візуалізації, що представляють саме ці крапки. Додаток візуалізації даних генерує відповідна мова запиту, управляє прийняттям запиту базою даних і візуально представляє результуючу безліч. Користувач може сфокусуватися на аналізі, не відволікаючись на складання запиту.

**Збільшення розмірів і складності структур даних, що представляються візуалізацією.** Елементарна секторна діаграма або гістограма візуалізує прості послідовності числових інформаційних крапок. Однак нові вдосконалені типи діаграм здатні візуалізувати тисячі таких крапок і навіть складні структури даних – наприклад, нейронні мережі.

Скажемо, кошту OLAP (а також інструменти генерації запитів і випуску звітів) уже давно підтримують діаграми для своїх онлайнових звітів. Нові візуалізаційні програми обновляють контент за рахунок періодично повторюваного зчитування даних. Фактично користувачі візуалізаційних програм лінійні процеси, що відслідковують (коливання фондового ринку, показники роботи комп’ютерних систем, сейсмограми, сітки корисності й ін.), потребують завантаження даних у режимі реального часу або близькому до нього режимі.

Користувачі інструментів Data Mining звичайно аналізують дуже великі набори чисельних даних. Традиційні типи діаграм для бізнесу (секторні діаграми й гістограми) погано справляються з показом тисяч інформаційних точок. Тому інструменти Data Mining майже завжди підтримують якусь форму візуалізації даних, здатну відображати структури й закономірності досліджуваних наборів даних, відповідно до тих аналітичних підходів, які використовується в інструменті.

Крім того, що візуалізація підтримує обробку структурованих даних, вона також є ключовим засобом представлення схем так званих неструктурованих даних, наприклад текстових документів, тобто

Text Mining. Зокрема, засоби Text Mining можуть здійснювати парсинг більших пакетів документів і формувати предметні покажчики понять і тем, освітлених у цих документах. Коли предметні покажчики створені за допомогою нейронної мережевої технології, користувачеві непросто продемонструвати їх без деякої форми візуалізації даних.

***Лекція 2.3. (2 год)***

**Тема 2. Методи інтелектуального аналізу даних**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Методи класифікації та прогнозування
2. Методи кластерного аналізу
3. Алгоритм k-середніх
4. Нейронні мережі. Моделі нейронних мереж
5. Карти Кохонена. Методи пошуку асоціативних правил

***1. МЕТОДИ КЛАСИФІКАЦІЇ ТА ПРОГНОЗУВАННЯ***

Метод дерев рішень

*Метод дерев рішень* (decision trees) є одним з найбільш популярних методів розв’язку задач класифікації й прогнозування. Іноді цей метод Data Mining також називають деревами вирішальних правил, деревами класифікації і регресії.

Як видно з останньої назви, за допомогою даного методу вирішуються задачі класифікації й прогнозування.

Якщо залежна, тобто цільова змінна приймає дискретні значення, за допомогою методу дерева рішень вирішується задача класифікації.

Якщо ж залежна змінна приймає безперервні значення, то дерево рішень установлює залежність цієї змінної від незалежних змінних, тобто вирішує задачу чисельного прогнозування.

У найбільш простому вигляді дерево рішень – це спосіб показу правил в ієрархічній, послідовній структурі. Основа такої структури – відповіді «Так» або «Ні» на низку питань.

На [рис. 2.31.](http://www.intuit.ru/department/database/datamining/9/datamining_9.html#image.9.1) наведений приклад дерева рішень, задача якого – відповістити на запитання: «Чи грати в гольф?» Щоб розв’язати задачу, тобто прийняти рішення, чи грати в гольф, слід віднести поточну ситуацію до одного з відомих класів (у цьому випадку – «грати» або «не грати»). Для цього потрібно відповісти на низку питань, які є у вузлах цього дерева, починаючи з його кореня.

Перший вузол нашого дерева «Сонячно?» є вузлом перевірки, тобто умовою. При позитивній відповіді на запитання здійснюється перехід до лівої частини дерева, називаної лівою гілкою, при негативному – до правої частини дерева. Таким чином, внутрішній вузол дерева є вузлом перевірки певної умови. Далі йде наступне питання і т.д., поки не буде досягнутий кінцевий вузол дерева, що є вузлом розв’язку. Для нашого дерева існує два типи кінцевого вузла: «грати» і «не грати» у гольф.

У результаті проходження від кореня дерева (іноді називається кореневою вершиною) до його вершини вирішується задача класифікації, тобто вибирається один із класів – «грати» і «не грати» у гольф.

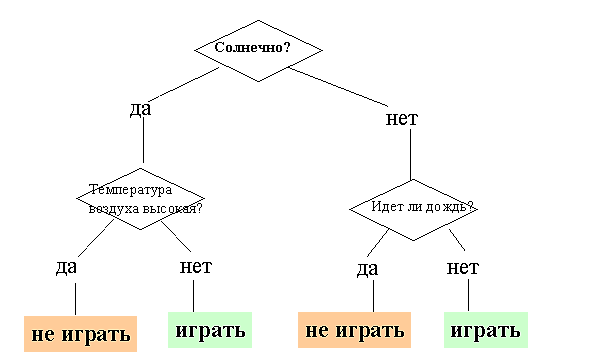


Рис. 2.31. Дерево рішень «Чи грати в гольф?»

Метою побудови дерева рішень в нашому випадку є визначення значення категоріальної залежної змінної.

Отже, для нашої задачі основними елементами дерева рішень є:

* Корінь дерева: «Сонячно?»
* Внутрішній вузол дерева або вузол перевірки: «Температура повітря висока?», « Чи йде дощ?»
* Листок, кінцевий вузол дерева, вузол розв’язку або вершина: «Грати», «Не грати» Гілки дерева (випадки відповіді): «Так», «Ні».

У розглянутому прикладі вирішується задача бінарної класифікації, тобто створюється дихотомічна класифікаційна модель. Приклад демонструє роботу так званих бінарних дерев.

У вузлах бінарних дерев розгалуження може відбуватися тільки у двох напрямках, тобто існує можливість тільки двох відповідей на поставлене питання («так» і «ні»).

Бінарні дерева є найпростішим, частковим випадком дерев рішень. В інших випадках, відповідей і, відповідно, гілок дерева, що виходять із його внутрішнього вузла, може бути більше двох.

Розглянемо більш складний приклад.

База даних, на основі якої повинне здійснюватися прогнозування, містить наступні ретроспективні дані про клієнтів банку, що є її атрибутами:

* вік,
* наявність нерухомості,
* освіта,
* середньомісячний дохід,
* чи повернув клієнт вчасно кредит.

Задача полягає в тому, щоб на підставі перерахованих вище даних (крім останнього атрибута) визначити, чи варто видавати кредит новому клієнтові.

Як ми вже розглядали в лекції, присвяченій задачі класифікації, така задача вирішується у два етапи:

* побудова класифікаційної моделі
* її використання.

На етапі побудови моделі, власне, і будується дерево класифікації або створюється набір якихось правил.

На етапі використання моделі побудоване дерево, або шлях від його кореня до однієї з вершин, що є набором правил для конкретного клієнта, використовується для відповіді на поставлене питання «чи видавати кредит?»

Правилом є логічна конструкція, представлена у вигляді «якщо : то :».

На [рис. 2.32](http://www.intuit.ru/department/database/datamining/9/datamining_9.html#image.9.2). наведений приклад дерева класифікації, за допомогою якого вирішується задача «Чи видавати кредит клієнтові?». Вона є типовою задачею класифікації, і за допомогою дерев рішень одержують досить хороші варіанти її розв’язку.

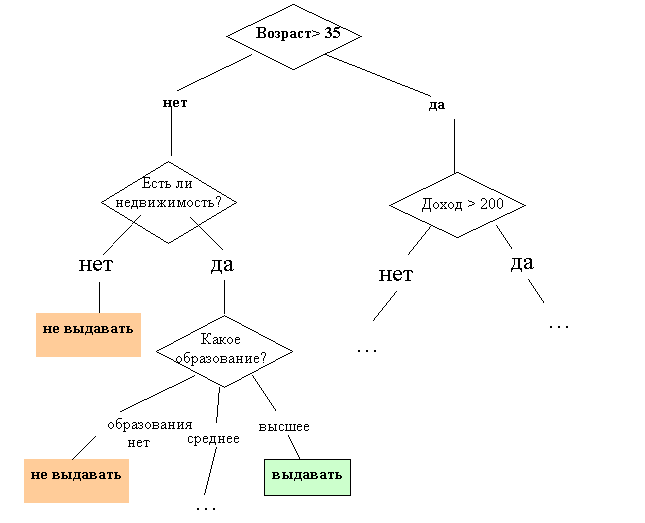


Рис. 2.32. Дерево рішень « чи Видавати кредит?».

Як ми бачимо, внутрішні вузли дерева (вік, наявність нерухомості, дохід і освіта) є атрибутами описаної вище бази даних.

Ці атрибути називають прогнозуючими, або атрибутами розщеплення (splitting attribute). Кінцеві вузли дерева, або листки, іменуються мітками класу, що є значеннями залежної категоріальної змінної «видавати» або «не видавати» кредит.

Кожна гілка дерева, що йде від внутрішнього вузла, відзначена предикатом розщеплення. Останній може відноситися лише до одного атрибуту розщеплення даного вузла.

Характерна риса предикатів розщеплення: кожний запис використовує унікальний шлях від кореня дерева тільки до одного вузла–розв’язку. Об’єднана інформація про атрибути розщеплення й предикати розщеплення у вузлі називається критерієм розщеплення (splitting criterion).

На [рис. 5.2](http://www.intuit.ru/department/database/datamining/9/datamining_9.html#image.9.2). зображене одне з можливих дерев рішень для розглянутої бази даних. Наприклад, критерій розщеплення «Яка освіта?», міг би мати два предикати розщеплення й виглядати інакше: освіта «вища» і «не вища». Тоді дерево рішень мало б інший вигляд.

Таким чином, для даної задачі (як і для будь–якої іншої) може бути побудовано множина дерев рішень різної якості, з різною прогнозуючою точністю.

Якість побудованого дерева розв’язку досить сильно залежить від правильного вибору критерію розщеплення. Над розробкою й удосконаленням критеріїв працюють багато дослідників.

Метод дерев рішень часто називають «наївним» підходом. Але завдяки цілому ряду переваг, даний метод є одним з найбільш популярних для розв’язку задач класифікації.

Переваги дерев рішень

Інтуїтивність дерев рішень. Класифікаційна модель, представлена у вигляді дерева рішень, є інтуїтивною й спрощує розуміння розв’язуваної задачі.

Результат роботи алгоритмів конструювання дерев рішень, *на відміну, наприклад, від нейронних мереж, що представляють собою «чорні ящики»,* легко інтерпретується користувачем. Ця властивість дерев рішень не тільки важлива при віднесенні до певного класу нового об’єкта, але й корисна при інтерпретації моделі класифікації в цілому. Дерево рішень дозволяє зрозуміти й пояснити, чому конкретний об’єкт відноситься до того або іншого класу.

Дерева рішень дають можливість витягати правила з бази даних природньою мовою. Приклад правила: Якщо Вік > 35 і Дохід > 200, то видати кредит.

Дерева рішень дозволяють створювати класифікаційні моделі в тих областях, де аналітикові досить складно формалізувати знання.

Алгоритм конструювання дерева рішень не вимагає від користувача вибору вхідних атрибутів (незалежних змінних). На вхід алгоритму можна подавати всі існуючі атрибути, алгоритм сам вибере найбільш значимі серед них, і тільки вони будуть використані для побудови дерева. У порівнянні, наприклад, з нейронними мережами, це значно полегшує користувачеві роботу, оскільки в нейронних мережах вибір кількості вхідних атрибутів суттєво впливає на час навчання.

Точність моделей, створених за допомогою дерев рішень, порівняно з іншими методами побудови класифікаційних моделей (статистичні методи, нейронні мережі).

Розроблений ряд масштабованих алгоритмів, які можуть бути використані для побудови дерев рішень на надвеликих базах даних; масштабованість тут означає, що з ростом числа прикладів або записів бази даних час, затрачуваний на навчання, тобто побудову дерев рішень, росте лінійно. Приклади таких алгоритмів: SLIQ, SPRINT.

Швидкий процес навчання. На побудову класифікаційних моделей за допомогою алгоритмів конструювання дерев рішень потрібно значно менше часу, ніж, наприклад, на навчання нейронних мереж.

Більшість алгоритмів конструювання дерев рішень мають можливість спеціальної обробки пропущених значень.

Багато класичних статистичних методів, за допомогою яких вирішуються задачі класифікації, можуть працювати тільки із числовими даними, у той час як дерева рішень працюють і із числовими, і з категоріальними типами даних.

Багато статистичних методів є параметричними, і користувач повинен заздалегідь володіти певною інформацією, наприклад, знати вид моделі, мати гіпотезу про вид залежності між змінними, припускати, який вид розподілу мають дані. Дерева рішень, на відміну від таких методів, будують непараметричні моделі. Таким чином, дерева рішень здатні вирішувати такі задачі Data Mining, у яких відсутня апріорна інформація про вид залежності між досліджуваними даними.

**Процес конструювання дерева рішень.** Нагадаємо, що розглянута нами задача класифікації відноситься до стратегії навчання із учителем, яке іноді називається індуктивним навчанням. У цих випадках усі об’єкти тренувального набору даних заздалегідь віднесені до одного з визначених класів.

Алгоритми конструювання дерев рішень складається з етапів «побудова» або «створення» дерева (tree building) і «скорочення» дерева (tree pruning). У ході створення дерева вирішуються питання вибору критерію розщеплення й зупинки навчання (якщо це передбачено алгоритмом). У ході етапу скорочення дерева вирішується питання відсікання деяких його гілок.

Розглянемо ці питання детальніше.

***Критерій розщеплення.*** Процес створення дерева відбувається зверху вниз, тобто є спадним. У ході процесу алгоритм повинен знайти такий критерій розщеплення, іноді також називається критерієм розбивки, щоб розбити множину на підмножини, які б асоціювалися з даним вузлом перевірки. Кожний вузол перевірки повинен бути позначений певним атрибутом. Існує правило вибору атрибута: він повинен розбивати вихідну множину даних таким чином, щоб об’єкти підмножин, одержуваних в результаті цієї розбивки, були представниками одного класу або ж були максимально наближені до такої розбивки. Остання фраза означає, що кількість об’єктів з інших класів, так званих «домішок», у кожному класі повинна прагнути до мінімуму.

Існують різні критерії розщеплення. Найбільш відомі – міра ентропії й індекс Gini.

У деяких методах для вибору атрибута розщеплення використовується так звана міра інформативності підпросторів атрибутів, яка ґрунтується на ентропійному підході й відома за назвою «захід інформаційного виграшу» (information gain measure) або захід ентропії.

Інший критерій розщеплення, запропонований Брейманом (Breiman) і ін., реалізований в алгоритмі CART і називається індексом Gini. За допомогою цього індексу атрибут вибирається на підставі відстаней між розподілами класів.

Якщо дана множина *T*, що включає приклади з *n* класів, індекс Gini, тобто *gini(T)*, визначається за формулою:



де T – поточний вузол, pj – імовірність класу j у вузлі T, n – кількість класів.

***Велике дерево не означає, що воно «підходяще».*** Чим більше окремих випадків описано в дереві рішень, тим менша кількість об’єктів потрапляє в кожний окремий випадок. Такі дерева називають «гіллястими» або «кущистими», вони складаються з невиправдано великої кількості вузлів і гілок, вихідна множина розбивається на велику кількість підмножин, що складаються із дуже малого числа об’єктів. У результаті «переповнення» таких дерев їх здатність до узагальнення зменшується, і побудовані моделі не можуть давати вірні відповіді.

У процесі побудови дерева, щоб його розміри не стали надмірно великими, використовують спеціальні процедури, які дозволяють створювати оптимальні дерева, так звані дерева «підходящих розмірів» (Breiman,1984).

Який розмір дерева може вважатися оптимальним? Дерево повинне бути досить складним, щоб ураховувати інформацію з досліджуваного набору даних, але одночасно воно повинне бути досить простим. Інакше кажучи, дерево повинне використовувати інформацію, що поліпшує якість моделі, і ігнорувати ту інформацію, яка її не поліпшує.

Отут існує дві можливі стратегії. Перша полягає в нарощуванні дерева до певного розміру відповідно до параметрів, заданих користувачем.

Визначення цих параметрів може ґрунтуватися на досвіді й інтуїції аналітика, а також на деяких «діагностичних повідомленнях» системи, що конструює дерево рішень.

Друга стратегія полягає у використанні набору процедур, що визначають «підходящий розмір» дерева, вони розроблені Бріманом, Куїлендом і ін. в 1984 році. Однак, як відзначають автори, не можна сказати, що ці процедури доступні починаючому користувачеві.

Процедури, які використовують для запобігання створення надмірно великих дерев, включають: скорочення дерева шляхом відсікання гілок; використання правил зупинки навчання.

Слід зазначити, що не всі алгоритми при конструюванні дерева працюють по одній схемі. Деякі алгоритми включають два окремі послідовні етапи: побудова дерева і його скорочення; інші чергують ці етапи в процесі своєї роботи для запобігання нарощування внутрішніх вузлів.

***Зупинка побудови дерева.*** Розглянемо правило зупинки. Воно повинне визначити, чи є розглянутий вузол внутрішнім вузлом, при цьому він буде розбиватися далі, або ж він є кінцевим вузлом, тобто вузлом розв’язком.

## Зупинка – **такий момент у процесі побудови дерева, коли слід припинити подальші розгалуження.**

Один з варіантів правил зупинки – «рання зупинка» (prepruning), вона визначає доцільність розбивки вузла. Перевага використання такого варіанта – зменшення часу на навчання моделі. Однак тут виникає ризик зниження точності класифікації. Тому рекомендується «замість зупинки використовувати відсікання» (Breiman, 1984).

Другий варіант зупинки навчання – обмеження глибини дерева. У цьому випадку побудова закінчується, якщо досягнута задана глибина.

Ще один варіант зупинки – задання мінімальної кількості прикладів, які будуть утримуватися в кінцевих вузлах дерева. При цьому варіанті розгалуження тривають до того моменту, поки всі кінцеві вузли дерева не будуть чистими або будуть містити не більш ніж задане число об’єктів.

Існує ще ряд правил, але слід зазначити, що жодне з них не має великої практичної цінності, а деякі можуть бути застосовні лише в окремих випадках.

***Скорочення дерева або відсікання гілок.*** Вирішенням проблеми занадто гіллястого дерева є його скорочення шляхом відсікання (pruning) деяких гілок.

Якість класифікаційної моделі, побудованої за допомогою дерева рішень, характеризується двома основними ознаками: точністю розпізнавання й помилкою.

Точність розпізнавання розраховується як відношення об’єктів, правильно класифікованих у процесі навчання, до загальної кількості об’єктів набору даних, які брали участь у навчанні.

Помилка розраховується як відношення об’єктів, неправильно класифікованих у процесі навчання, до загальної кількості об’єктів набору даних, які брали участь у навчанні.

Відсікання гілок або заміну деяких гілок під деревом слід проводити там, де ця процедура не приводить до зростання помилки. Процес проходить знизу вгору, тобто є висхідним. Це більш популярна процедура, ніж використання правил зупинки. Дерева, одержувані після відсікання деяких гілок, називають усіченими.

Якщо таке усічене дерево усе ще не є інтуїтивним і складне для розуміння, використовують витяг правил, які поєднують у набори для опису класів.

Кожний шлях від кореня дерева до його вершини або листка дає одне правило. Умовами правила є перевірки на внутрішніх вузлах дерева.

Алгоритми

На сьогоднішній день існує велика кількість алгоритмів, що реалізують дерева рішень: CART, C4.5, CHAID, CN2, Newid, Itrule і інші.

Алгоритм CART (Classification and Regression Tree), як видно з назви, вирішує задачу класифікації й регресії. Він розроблений в 1974–1984 роках чотирма професорами статистики – Leo Breiman (Berkeley), Jerry Friedman (Stanford), Charles Stone (Berkeley) і Richard Olshen (Stanford).

Атрибути набору даних можуть мати як дискретне, так і числове значення.

Алгоритм CART призначений для побудови бінарного дерева рішень. Бінарні дерева також називають двійковими. Приклад такого дерева розглядався на початку лекції.

*Інші особливості алгоритму CART:*

* функція оцінки якості розбивки;
* механізм відсікання дерева;
* алгоритм обробки пропущених значень;
* побудова дерев регресії.

Кожний вузол бінарного дерева при розбивці має тільки двох нащадків, що називаються дочірніми галузями. Подальший поділ гілок залежить від того, чи багато вихідних даних описує дана гілка. На кожному кроці побудови дерева правило, формоване у вузлі, ділить задану множину прикладів на дві частини. Права його частина (гілка right) – це та частина множини, у якій правило виконується; ліва (гілка left) – та, для якої правило не виконується.

Функція оцінки якості розбивки, яка використовується для вибору оптимального правила, – індекс Gini – був описаний вище. Відзначимо, що дана оціночна функція заснована на ідеї зменшення невизначеності у вузлі. Допустимо, є вузол, і він розбитий на два класи. Максимальна невизначеність у вузлі буде досягнута при розбивці його на дві підмножини по 50 прикладів, а максимальна визначеність – при розбивці на 100 і 0 прикладів.

Правила розбивки. Нагадаємо, що алгоритм CART працює із числовими й категоріальними атрибутами. У кожному вузлі розбивка може йти тільки по одному атрибуту. Якщо атрибут є числовим, то у внутрішньому вузлі формується правило виду xi <= c, Значення c у більшості випадків вибирається як середнє арифметичне двох сусідніх впорядкованих значень змінної xi навчального набору даних. Якщо ж атрибут відноситься до категоріального типу, то у внутрішньому вузлі формується правило xi V(xi), де V(xi) – деяка непорожня підмножина множин значень змінної xi у навчальному наборі даних.

Механізм відсікання. Цим механізмом, що має назву minimal cost–complexity tree pruning, алгоритм CART принципово відрізняється від інших алгоритмів конструювання дерев рішень. У розглянутому алгоритмі відсікання – це деякий компроміс між одержанням дерева «підходящого розміру» і одержанням найбільш точної оцінки класифікації. Метод полягає в одержанні послідовності зменшуваних дерев, але дерева розглядаються не всі, а тільки «кращі представники».

Перехресна перевірка (V-fold cross-validation) є найбільш складною й одночасно оригінальною частиною алгоритму CART. Вона являє собою шлях вибору остаточного дерева, за умови, що набір даних має невеликий обсяг або ж записи набору даних настільки специфічні, що розділити набір на навчальну й тестову вибірку не представляється можливим.

Отже, основні характеристики алгоритму CART: бінарне розщеплення, критерій розщеплення – індекс Gini, алгоритми minimal cost–complexity tree pruning і V-fold cross-validation, принцип «виростити дерево, а потім скоротити», висока швидкість побудови, обробка пропущених значень.

Алгоритм C4.5 будує дерево рішень з необмеженою кількістю гілок у вузла. Даний алгоритм може працювати тільки з дискретним залежним атрибутом і тому може вирішувати тільки задачу класифікації. C4.5 вважається одним з найвідоміших і широко використовуваних алгоритмів побудови дерев класифікації.

Для роботи алгоритму C4.5 необхідне дотримання наступних вимог:

* Кожний запис набору даних повинен бути асоційованим з одним з визначених класів, тобто один з атрибутів набору даних повинен бути міткою класу.
* Класи повинні бути дискретними. Кожний приклад повинен однозначно відноситися до одного із класів.
* Кількість класів повинна бути значно менше кількості записів у досліджуваному наборі даних.

Остання версія алгоритму – алгоритм C4.8 – реалізована в інструменті Weka як J4.8 (Java). Комерційна реалізація методу: C5.0, розроблювач Rulequest, Австралія.

Алгоритм C4.5 повільно працює на надвеликих й зашумлених наборах даних.

Ми розглянули два відомі алгоритми побудови дерев рішень CART і C4.5. Обидва алгоритми є робастними, тобто стійкими до шумів і викидів даних.

Алгоритми побудови дерев рішень відрізняються наступними характеристиками:

* Вид розщеплення – бінарне (binary), множинне (multi-way);
* Критерії розщеплення – ентропія, gini, інші;
* Можливість обробки пропущених значень;
* Процедура скорочення гілок або відсікання;
* Можливості витягування правил з дерев.

Жоден алгоритм побудови дерева не можна апріорі вважати найкращим або досконалим, підтвердження доцільності використання конкретного алгоритму повинне бути перевірене й підтверджене експериментом.

***Розробка нових масштабованих алгоритмів.*** Найбільш серйозна вимога, яка зараз пред’являється до алгоритмів конструювання дерев рішень – це масштабованість, тобто алгоритм повинен мати масштабований метод доступу до даних.

Розроблений ряд нових масштабованих алгоритмів, серед них – алгоритм Sprint, запропонований Джоном Боярином і його колегами. Sprint, що є масштабованим варіантом розглянутого в лекції алгоритму CART, висуває мінімальні вимоги до об’єму оперативної пам’яті.

**Метод опорних векторів**

*У попередніх темі ми розглянули такі методи класифікації й прогнозування як лінійна регресія й дерева рішень; у цій лекції ми продовжимо знайомство з методами цієї групи й розглянемо наступні з них: метод опорних векторів, метод найближчого сусіда (метод міркувань на основі прецедентів) і баєсовську класифікацію.*

Метод опорних векторів (Support Vector Machine – SVM) відноситься до групи граничних методів. Він визначає класи за допомогою границь областей.

За допомогою даного методу вирішуються задачі бінарної класифікації. В основі методу лежить поняття площин розв’язків.

Площина (plane) розв’язку розділяє об’єкти з різною класовою приналежністю.

На [рис. 2.33.](http://www.intuit.ru/department/database/datamining/10/datamining_10.html#image.10.1) наведений приклад, у якому беруть участь об’єкти двох типів. Поділяюча лінія задає границю, праворуч від якої – усі об’єкти типу brown (коричневий), а ліворуч – типу yellow (жовтий). Новий об’єкт, що потрапляє праворуч, класифікується як об’єкт класу brown або – як об’єкт класу yellow, якщо він розташувався ліворуч від поділяючої прямої. У цьому випадку кожний об’єкт характеризується двома вимірами.

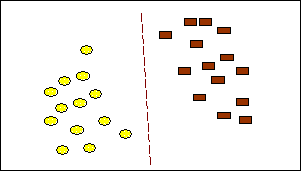


Рис. 2.33. Поділ класів прямою лінією

Ціль методу опорних векторів – знайти площину, що розділяє дві множини об’єктів; така площина показана на [рис.](http://www.intuit.ru/department/database/datamining/10/datamining_10.html#image.10.2) 2.34. На цьому малюнку множина зразків поділена на два класи: жовті об’єкти належать класу А, коричневі – класу В.

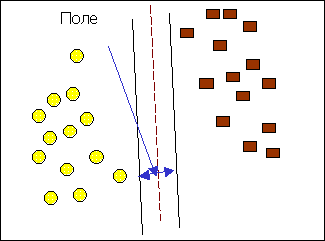


Рис. 2.34. До визначення опорних векторів

Метод відшукує зразки, що перебувають на границях між двома класами, тобто опорні вектори; вони зображені на [рис. 2.35](http://www.intuit.ru/department/database/datamining/10/datamining_10.html#image.10.3).

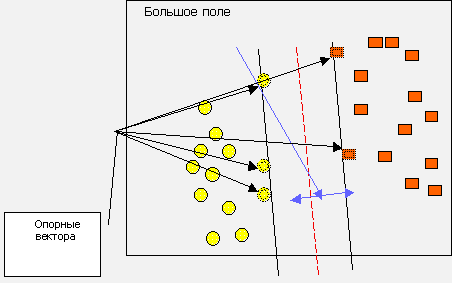


Рис. 2.35. Опорні вектори

Опорними векторами називаються об’єкти множини, що лежать на границях областей. Класифікація вважається гарною, якщо область між границями порожня.

На [рис. 2.35](http://www.intuit.ru/department/database/datamining/10/datamining_10.html#image.10.3). показано п’ять векторів, які є опорними для даної множини.

Лінійний SVM

Розв’язок задачі бінарної класифікації за допомогою методу опорних векторів полягає в пошуку деякої лінійної функції, яка правильно розділяє набір даних на два класи. Розглянемо задачу класифікації, де число класів рівне двом.

Задачу можна сформулювати як пошук функції *f(x)*, що приймає значення менше нуля для векторів одного класу й більше нуля – для векторів іншого класу. У якості вихідних даних для розв’язку поставленої задачі, тобто пошуку функції, що класифікує, *f(x)*, дано тренувальний набір векторів простору, для яких відома їхня приналежність до одного із класів. Сімейство функцій, що класифікують, можна описати через функцію *f(x)*. Гіперплощина визначена вектором *a* і значенням *b*, тобто *f(x)=ax+b*. Розв’язок даної задачі проілюстрований на [рис. 2.36](http://www.intuit.ru/department/database/datamining/10/datamining_10.html#image.10.4).

У результаті розв’язку задачі, тобто побудови Svm-Моделі, знайдена функція, що приймає значення менше нуля для векторів одного класу й більше нуля – для векторів іншого класу. Для кожного нового об’єкта негативне або позитивне значення визначає приналежність об’єкта до одного із класів.

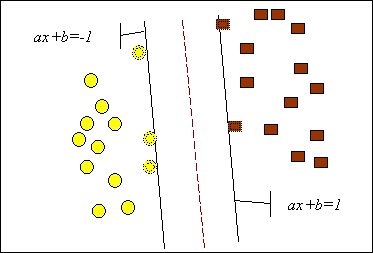


Рис. 2.36. Лінійний SVM

Найкращою функцією класифікації є функція, для якої очікуваний ризик мінімальний. Поняття очікуваного ризику в цьому випадку означає очікуваний рівень помилки класифікації.

Прямо оцінити очікуваний рівень помилки побудованої моделі неможливо, це можна зробити за допомогою поняття емпіричного ризику. Однак слід прийняти, що мінімізація останнього не завжди приводить до мінімізації очікуваного ризику. Цю обставину слід пам’ятати при роботі з відносно невеликими наборами тренувальних даних.

Емпіричний ризик – рівень помилки класифікації на тренувальному наборі.

Таким чином, у результаті розв’язку задачі методом опорних векторів для лінійно поділюваних даних ми одержуємо функцію класифікації, яка мінімізує верхню оцінку очікуваного ризику.

Однією із проблем, пов’язаних з розв’язком задач класифікації розглянутим методом, є та обставина, що не завжди можна легко знайти лінійну границю між двома класами.

У таких випадках один з варіантів – збільшення розмірності, тобто перенесення даних із площини в тривимірний простір, де можливо побудувати таку площину, яка ідеально розділить множину зразків на два класи. Опорними векторами в цьому випадку будуть служити об’єкти з обох класів, що є екстремальними.

Таким чином, за допомогою додавання так званого оператора ядра й додаткових розмірностей, знаходяться границі між класами у вигляді гіперплощин.

Однак слід пам’ятати: складність побудови Svm-Моделі полягає в тому, що чим вища розмірність простору, тим складніше з ним працювати. Один з варіантів роботи з даними високої розмірності – це попереднє застосування якого–небудь методу зниження розмірності даних для виявлення найбільш істотних компонентів, а потім використання методу опорних векторів.

Як і будь-який інший метод, метод SVM має свої сильні й слабкі сторони, які слід враховувати при виборі даного методу.

*Недолік методу* полягає в тому, що для класифікації використовується не вся множина зразків, а лише їхня невелика частина, яка перебуває на границях.

*Перевага методу* полягає в тому, що для класифікації методом опорних векторів, на відміну від більшості інших методів, достатньо невеликого набору даних. При правильній роботі моделі, побудованої на тестовій множині, цілком можливе застосування даного методу на реальних даних.

*Метод опорних векторів дозволяє:*

* одержати функцію класифікації з мінімальною верхньою оцінкою очікуваного ризику (рівня помилки класифікації);
* використовувати лінійний класифікатор для роботи з нелінійно поділюваними даними, поєднуючи простоту з ефективністю.

Метод «найближчого сусіда»

Метод «найближчого сусіда» або системи міркувань на основі аналогічних випадків.

Слід відразу зазначити, що метод «найближчого сусіда» («nearest neighbour») відноситься до класу методів, робота яких ґрунтується на зберіганні даних у пам’яті для порівняння з новими елементами. З появою нового запису для прогнозування мають місце відхилення між цим записом і подібними наборами даних, і ідентифікується найбільш подібний (або близький сусід).

Наприклад, при розгляді нового клієнта банку, його атрибути порівнюються з усіма існуючими клієнтами даного банку (дохід, вік і т.д.). Множина «найближчих сусідів» потенційного клієнта банку вибирається на підставі найближчого значення доходу, віку і т.д.

При такому підході використовується термін «*k-найближчий сусід*» («*k-nearest neighbour*»). Термін означає, що вибирається *k* «верхніх» (найближчих) сусідів для їхнього розгляду як множини «найближчих сусідів». Оскільки не завжди зручно зберігати всі дані, іноді зберігається тільки множина «типових» випадків. У такому випадку використовуваний метод називають міркуванням за аналогією (Case Based Reasoning, CBR), міркуванням на основі аналогічних випадків, міркуванням по прецедентах.

Прецедент – це опис ситуації в комбінації з докладною вказівкою дій, що застосовують у даній ситуації.

Підхід, заснований на прецедентах, умовно можна поділити на наступні етапи:

* збір докладної інформації про поставлене завдання;
* зіставлення цієї інформації з деталями прецедентів, що зберігаються в базі, для виявлення аналогічних випадків;
* вибір прецеденту, найбільш близького до поточної проблеми, з бази прецедентів;
* адаптація обраного розв’язку до поточної проблеми, якщо це необхідно;
* перевірка коректності кожного нового отриманого розв’язку;
* занесення детальної інформації про новий прецедент у базу прецедентів.

Таким чином, висновок, заснований на прецедентах, являє собою такий метод аналізу даних, який робить висновок щодо даної ситуації за результатами пошуку аналогій, що зберігаються в базі прецедентів.

Даний метод по своїй суті належить до категорії «навчання без вчителя», тобто являється технологією «що навчається самостійно», завдяки чому робочі характеристики кожної бази прецедентів з ходом часу і накопиченням прикладів покращуються. Розробка баз прецедентів по конкретній предметній області відбувається на природній для людини мові, отже, може бути виконана найбільш досвідченими співробітниками компанії – експертами або аналітиками, що працюють у даній предметній області.

Однак це не означає, що *CBR* – системи самостійно можуть ухвалювати рішення. Останнє завжди залишається за людиною, даний метод лише пропонує можливі варіанти розв’язку й указує на «найрозумніший» з її точки зору.

***Переваги методу:***

* Простота використання отриманих результатів.
* Розв’язки не унікальні для конкретної ситуації, можливе їх використання для інших випадків.
* Метою пошуку є не гарантовано вірний розв’язок, а кращий з можливих.

***Недоліки методу «найближчого сусіда»:***

* Даний метод не створює яких-небудь моделей або правил, що узагальнюють попередній досвід, – у виборі розв’язку вони ґрунтуються на всьому масиві доступних історичних даних, тому неможливо сказати, на якій підставі будуються відповіді.
* Існує складність вибору заходу «близькості» (метрики). Від цього заходу головним чином залежить обсяг множини записів, які потрібно зберігати в пам’яті для досягнення задовільної класифікації або прогнозу. Також існує висока залежність результатів класифікації від обраної метрики.
* При використанні методу виникає необхідність повного перебору навчальної вибірки при розпізнаванні, як наслідок цього – обчислювальна трудомісткість.
* Типові завдання даного методу – це завдання невеликої розмірності за кількістю класів і змінних.

За допомогою даного методу вирішуються задача класифікації й регресії.

Розглянемо докладно принципи роботи методу k–найближчих сусідів для розв’язку задач класифікації й регресії (прогнозування).

Розв’язок задачі класифікації нових об’єктів. Ця задача схематично зображена на [рис. 2.37](http://www.intuit.ru/department/database/datamining/10/datamining_10.html#image.10.5). Приклади (відомі екземпляри) відзначені знайомими «+» або «–», що визначають приналежність до відповідного класу («+» або «–»), а новий об’єкт, який потрібно класифікувати, позначений кружечком. Нові об’єкти також називають точками запиту.

Наша мета заключається в оцінці (класифікації) відгуку точок запиту з використанням спеціально обраного числа їх найближчих сусідів. Інакше кажучи, ми прагнемо довідатися, до якого класу слід віднести точку запиту: знак «+» або як знак «–».

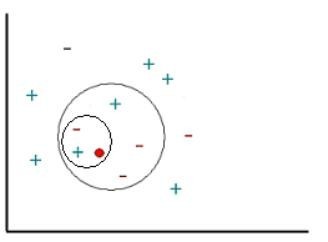


Рис. 2.37. Класифікація об’єктів множини при різному значенні параметра k

Для початку розглянемо результат роботи методу k–найближчих сусідів з використанням одного найближчого сусіда. У цьому випадку відгук точки запиту буде класифікований як знак плюс, тому що найближча сусідня точка має знак плюс.

Тепер збільшимо число використовуваних найближчих сусідів до двох. Цього разу метод k– найближчих сусідів не зможе класифікувати відгук точки запиту, оскільки друга найближча точка має знак мінус і обидва знаки рівноцінні (тобто перемога з однаковою кількістю голосів).

Далі збільшимо число використовуваних найближчих сусідів до 5. Таким чином, буде визначена ціла околиця точки запиту (на графіку її границя відзначена червоним (сірим) колом). Тому що в області утримується 2 точки зі знаком «+» і 3 точки зі знаком «–», алгоритм k-найближчих сусідів привласнить знак «–» відгуку точки запиту.

Розв’язок задачі прогнозування. Далі розглянемо принцип роботи методу *k*-найближчих сусідів для розв’язку задачі регресії. Регресійні задачі пов’язані із прогнозуванням значення залежної змінної за значеннями незалежних змінних набору даних.

Розглянемо графік, показаний на [рис. 2.38](http://www.intuit.ru/department/database/datamining/10/datamining_10.html#image.10.6). Зображений на ній набір точок (*зелені* прямокутники) отриманий по зв’язку між незалежною змінною *x* і залежною змінною *y* (*крива червоного кольору*). Заданий набір зелених об’єктів (тобто набір прикладів); ми використовуємо метод *k*–найближчих сусідів для прогнозування виходу точки запиту *X* по даному набору прикладів (зелені прямокутники).

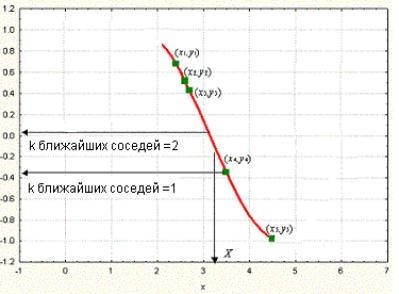


Рис. 2.38. Розв’язок задачі прогнозування при різних значеннях параметра k

Спочатку розглянемо як приклад метод k-найближчих сусідів з використанням одного найближчого сусіда, тобто при k, рівному одиниці. Ми шукаємо набір прикладів (зелені прямокутники) і виділяємо з їхнього числа найближчий до точки запиту *X*. Для нашого випадку найближчий приклад – точка (x4 ;y4). Вихід x4 (тобто y4), таким чином, приймається в якості результату прогнозування виходу *X* (тобто *Y*). Отже, для одного найближчого сусіда можемо записати: вихід Y рівний y4 (Y = y4).

Далі розглянемо ситуацію, коли *k* рівне двом, тобто розглянемо два найближчі сусіди. У цьому випадку ми виділяємо вже дві найближчі до *X* точки. На нашому графіку це точки y3 і y4 відповідно. Обчисливши середнє їхніх виходів, записуємо розв’язок для Y у вигляді Y = (y3 + y4)/2.

Розв’язок задачі прогнозування здійснюється шляхом перенесення описаних вище дій на використання довільного числа найближчих сусідів таким чином, що вихід Y точки запиту X обчислюється як середньоарифметичне значення виходів k–найближчих сусідів точки запиту.

Незалежні й залежні змінні набору даних можуть бути як безперервними, так і категоріальними. Для безперервних залежних змінних задача розглядається як задача прогнозування, для дискретних змінних – як задача класифікації.

Прогнозування в задачі прогнозування виходить усередненням виходів *k*-найближчих сусідів, а розв’язок задачі класифікації заснований на принципі «за більшістю голосів».

Критичним моментом у використанні методу *k*–найближчих сусідів є вибір параметра *k*. Він один з найбільш важливих факторів, що визначають якість прогнозної або класифікаційної моделі.

Якщо обране занадто мале значення параметра *k*, виникає ймовірність великого розкиду значень прогнозу. Якщо обране значення занадто велике, це може привести до сильного зміщення моделі. Таким чином, ми бачимо, що повинне бути обране оптимальне значення параметра k. Тобто це значення повинне бути настільки великим, щоб звести до мінімуму ймовірність неправильної класифікації, і одночасно, досить малим, щоб *k* сусідів були розташовані досить близько до точки запиту.

Таким чином, ми розглядаємо *k* параметр, як згладжуючий, для якого повинен бути знайдений компроміс між силою розмаху (розкиду) моделі і її зміщеністю.

Один з варіантів оцінки параметра *k* – проведення крос–перевірки (Bishop, 1995). Така процедура реалізована, наприклад, у пакеті STATISTICA (Statsoft).

Крос-перевірка – відомий метод одержання оцінок невідомих параметрів моделі. Основна ідея методу – поділ вибірки даних на *v* «складок». *V* «складки» тут є випадковим чином виділені ізольовані підвибірки.

За фіксованим значенням *k* будується модель *k*-найближчих сусідів для одержання прогнозів на *v*-му сегменті (інші сегменти при цьому використовуються як приклади) і оцінюється помилка класифікації. Для регресійних задач найбільш часто в якості оцінки помилки виступає сума квадратів, а для класифікаційних задач зручніше розглядати точність (відсоток коректно класифікованих спостережень).

Далі процес послідовно повторюється для всіх можливих варіантів вибору *v*. По вичерпанню *v* «складок» (циклів), обчислені помилки усереднюються й використовуються в якості міри стабільності моделі (тобто міри якості прогнозування в точках запиту). Вищеописані дії повторюються для різних *k*, і значення, що відповідає найменшій помилці (або найбільшій класифікаційній точності), приймається як оптимальне (оптимальне в сенсі методу крос-перевірки).

Слід враховувати, що крос–перевірка – ємнісна з точки зору обчислень процедура, і необхідно надати час для роботи алгоритму, особливо якщо обсяг вибірки досить великий.

Другий варіант вибору значення параметра *k* – самостійно задати його значення. Однак цей спосіб слід використовувати, якщо є обґрунтовані припущення щодо можливого значення параметра, наприклад, про попередні дослідження подібних наборів даних.

Метод *k*-найближчих сусідів показує досить непогані результати в найрізноманітніших задачах.

Прикладом реального використання описаного вище методу є програмне забезпечення центру технічної підтримки компанії Dell, розроблене компанією Inference. Ця система допомагає співробітникам центру відповідати на велике число запитів, відразу пропонуючи відповіді на розповсюджені питання й дозволяючи звертатися до бази під час розмови по телефону з користувачем. Співробітники центру технічної підтримки, завдяки реалізації цього методу, можуть відповідати одночасно на значне число дзвінків. Програмне забезпечення CBR зараз розгорнуте в мережі Intranet компанії Dell.

Інструментів Data Mining, що реалізують метод *k*–найближчих сусідів і *CBR*–Метод, не дуже багато. Серед найбільш відомих: CBR Express і Case Point (Inference Corp.), Apriori (Answer Systems), DP Umbrella (VYCOR Corp.), KATE tools (Acknosoft, Франція), Pattern Recognition Workbench (Unica, США), а також деякі статистичні пакети, наприклад, Statistica.

**Байєсова класифікація**

**Теорема Баєса** – одна з основних теорем [теорії ймовірностей](http://uk.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BE%D1%80%D1%96%D1%8F_%D0%B9%D0%BC%D0%BE%D0%B2%D1%96%D1%80%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9), яка визначає ймовірність настання події, коли відома тільки часткова інформація про подію. Названа на честь [Томаса Баєса](http://uk.wikipedia.org/wiki/%D0%A2%D0%BE%D0%BC%D0%B0%D1%81_%D0%91%D0%B0%D1%94%D1%81) (/beɪz/(1701–1761)). У теорії ймовірностей і статистиці, теорема Байеса (закон Байеса або ж правило Байєса) це теорема з двома різними інтерпретаціями. В байєсівсій інтерпретації, раціонально виразити, суб’єктивну ступінь віри у разі зміни показань для обліку . У «частотній інтерпретації», теорема стосується подання зворотної ймовірності двом подіям. В байєсівській інтерпретації, теорема Байеса має основоположне значення для байєсівської статистики та програми в полях у тому числі науки, техніки, медицини та права. Реалізація теореми Байеса для поновлення висновків про ймовірності називається байєсівським висновком. який запропонував використовувати першу теорему оновити вірування. Тим не менш, історичні праці, опубліковані посмертно WAS. Його ідеї отримали обмежений вплив, поки не були відкриті заново і далі Top самостійно розроблений Лапласа, хто вперше опублікований сучасної формулюванні в історію 1812 Аналітична теорія ймовірностей. До другої половини 20–го століття, байєсовський інтерпретації залучила широке незгоду [правити] від спільноти математики займав частотної переглядів цілому [правити] Відмова Bayesianism як ненаукові. Тим не менш, у даний час широко прийнято. Це може бути викликано були розвиток обчислювальної техніки, які дозволили успішно Застосування Bayesianism до багатьох складних проблем.

*Формула Баєса:*

Описание: P(A|B) = \frac{P(B|A)P(A)}{P(B)},

де Описание: P(A) – апріорна ймовірність гіпотези *A*; Описание: P(A|B) – ймовірність гіпотези *A* при настанні події *B* (апостеріорна ймовірність); Описание: P(B|A) – ймовірність настання події *B* при істинності гіпотези *A*; Описание: P(B) – ймовірність настання події *B*.

Формула виводиться із визначення [умовної ймовірності](http://uk.wikipedia.org/wiki/%D0%A3%D0%BC%D0%BE%D0%B2%D0%BD%D0%B0_%D0%B9%D0%BC%D0%BE%D0%B2%D1%96%D1%80%D0%BD%D1%96%D1%81%D1%82%D1%8C):

Описание:  P(A|B) = \frac{P(AB)}{P(B)} \quad \Rightarrow \quad P(AB) = P(A|B)P(B) = P(B|A)P(A).

**Байєсовський класифікатор** – широкий клас алгоритмів класифікації, заснований на принципі максимуму апостеріорної ймовірності. Для класифікованого об’єкта обчислюються функції правдоподібності кожного з класів, по них обчислюються апостеріорні ймовірності класів. Об’єкт відноситься до того класу, для якого апостеріорна ймовірність максимальна.

Апостеріорна ймовірність – умовна ймовірність випадкової події за умови того, що відомі апостеріорні дані, тобто отримані після досліду.

Альтернативні назви: байєсовске моделювання, байєсовська статистика, метод байєсовских мереж.

Споконвічно байєсовскьа класифікація використовувалася для формалізації знань експертів в експертних системах, зараз байєсовська класифікація також застосовується в якості одного з методів Data Mining.

Так звана наївна класифікація або наївно–байєсовський підхід (naive–bayes approach) є найбільш простим варіантом методу, що використовує байєсовські мережі. При цьому підході вирішуються задача класифікації, результатом роботи методу є так звані «прозорі» моделі.

«Наївна» класифікація – досить прозорий і зрозумілий метод класифікації. «Наївною» вона називається тому, що виходить із припущення про взаємну незалежність ознак.

Властивості наївної класифікації:

1. Використання всіх змінних і визначення всіх залежностей між ними.
2. Наявність двох припущень щодо змінних:
   * усі змінні є однаково важливими;
   * усі змінні є статистично незалежними, тобто значення однієї змінної нічого не говорить про значення іншої.

Більшість інших методів класифікації припускають, що перед початком класифікації ймовірність того, що об’єкт належить тому або іншому класу, однакова; але це не завжди правильно.

Допустимо, відомо, що певний відсоток даних належить конкретному класу. Виникає питання, чи можемо ми використовувати цю інформацію при побудові моделі класифікації? Існує множина реальних прикладів використання цих апріорних знань, що допомагають класифікувати об’єкти. Типовий приклад з медичної практики. Якщо лікар відправляє результати аналізів пацієнта на додаткове дослідження, він відносить пацієнта до якогось певного класу. Яким чином можна застосувати цю інформацію? Ми можемо використовувати її як додаткові дані при побудові класифікаційної моделі.

Відзначають такі переваги байєсовських мереж як методу Data Mining:

* у моделі визначаються залежності між усіма змінними, це дозволяє легко обробляти ситуації, у яких значення деяких змінних невідомі;
* байєсовські мережі досить просто інтерпретуються й дозволяють на етапі прогностичного моделювання легко проводити аналіз за сценарієм «що, якщо»;
* байєсовский метод дозволяє природно сполучати закономірності, виведені з даних, і, наприклад, експертні знання, отримані в явному вигляді;
* використання байєсовских мереж дозволяє уникнути проблеми переучування (overfitting), тобто надлишкового ускладнення моделі, що є слабкою стороною багатьох методів (наприклад, дерев рішень і нейронних мереж).

Наївно-байєсовський підхід має наступні недоліки:

* перемножувати умовні ймовірності коректно тільки тоді, коли всі вхідні змінні дійсно статистично незалежні; хоча часто даний метод показує досить гарні результати при недотриманні умови статистичної незалежності, але теоретично така ситуація повинна оброблятися більш складними методами, заснованими на навчанні байєсовських мереж;
* неможлива безпосередня обробка безперервних змінних – потрібно їхнє перетворення до інтервальнї шкали, щоб атрибути були дискретними; однак такі перетворення іноді можуть приводити до втрати значимих закономірностей.

***2. МЕТОДИ КЛАСТЕРНОГО АНАЛІЗУ***

Кластерний аналіз

З поняттям кластеризації ми познайомилися в п’ятій темі курсу. У цій лекції ми опишемо поняття «кластер» з математичної точки зору, а також розглянемо методи розв’язку задач кластеризації – методи кластерного аналізу.

Термін кластерний аналіз, уперше введений Тріоном (Tryon) в 1939 році, містить у собі більш 100 різних алгоритмів.

На відміну від завдань класифікації, кластерний аналіз не вимагає апріорних припущень про набір даних, не накладає обмеження на показ досліджуваних об’єктів, дозволяє аналізувати показники різних типів даних (інтервальні дані, частоти, бінарні дані). При цьому необхідно пам’ятати, що змінні повинні вимірюватися в порівнюваних шкалах.

Кластерний аналіз дозволяє скорочувати розмірність даних, робити їх наглядними.

Кластерний аналіз може застосовуватися до сукупностей тимчасових рядів, тут можуть виділятися періоди схожості деяких показників і визначатися групи тимчасових рядів зі схожою динамікою.

Кластерний аналіз паралельно розбудовувався в декількох напрямках, таких як біологія, психологія, ін., тому в більшості методів існує по дві й більш назв. Це суттєво утрудняє роботу при використанні кластерного аналізу.

*Завдання кластерного аналізу можна об’єднати в наступні групи:*

1. Розробка типології або класифікації.
2. Дослідження корисних концептуальних схем групування об’єктів.
3. Представлення гіпотез на основі дослідження даних.
4. Перевірка гіпотез або досліджень для визначення, чи дійсно типи (групи), виділені тем або іншим способом, присутні в наявних даних.

Як правило, при практичному використанні кластерного аналізу одночасно вирішується декілька із зазначених задач.

Розглянемо приклад процедури кластерного аналізу.

Допустимо, ми маємо набір даних А, що полягає з 14–ти прикладів, у яких є по дві ознаки X і Y.

Дані в табличній формі не носять інформативний характер. Представимо змінні X і Y у вигляді діаграми розсіювання, зображеної на [рис. 2.39](http://www.intuit.ru/department/database/datamining/13/datamining_13.html#image.13.1).

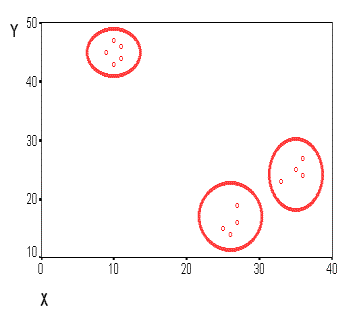


Рис. 2.39. Діаграма розсіювання змінних X і Y

На малюнку ми бачимо кілька груп «схожих» прикладів. Приклади (об’єкти), які за значеннями X і Y «схожі» один на одного, належать до одній групі (кластеру); об’єкти з різних кластерів не схожі один на одного.

Критерієм для визначення схожості й відмінності кластерів є відстань між точками на діаграмі розсіювання. Ця подібність можна «виміряти», воно дорівнює відстані між точками на графіку. Способів визначення заходу відстані між кластерами, називаною ще заходом близькості, існує трохи. Найпоширеніший спосіб – обчислення евклідової відстані між двома точками i і j на площині, коли відомі їхні координати X і Y:

 (2.1)

*Примітка:* щоб довідатися відстань між двома точками, треба взяти різницю їх координат по кожній осі, звести її у квадрат, скласти отримані значення для всіх осей і витягти квадратний корінь із суми.

Коли осей більше, чим дві, відстань розраховується в такий спосіб: сума квадратів різниці координат складається зі стількох доданків, скільки осей (вимірів) є присутнім у нашому просторі. Наприклад, якщо нам потрібно знайти відстань між двома точками в просторі трьох вимірів (така ситуація представлена на [рис. 2.40.](http://www.intuit.ru/department/database/datamining/13/datamining_13.html#image.13.2)), формула (2.1) здобуває вид:

 (2.2)

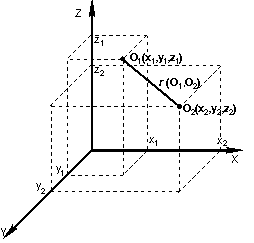


Рис. 2.40. Відстань між двома точками в просторі трьох вимірів

Кластер має наступні математичні характеристики: центр, радіус, середньоквадратичне відхилення, розмір кластера.

*Центр кластера* – це середнє геометричне місце точок у просторі змінних.

*Радіус кластера* – максимальна відстань точок від центру кластера.

Як було відзначено в одній з попередніх темі, кластери можуть бути, такими що перекриваються. Така ситуація виникає, коли виявляється перекриття кластерів. У цьому випадку неможливо за допомогою математичних процедур однозначно віднести об’єкт до одному із двох кластерів. Такі об’єкти називають спірними.

*Спірний об’єкт* – це об’єкт, який у міру подібності може бути віднесений до декільком кластерам.

*Розмір кластера* може бути визначений або по радіусу кластера, або по середньоквадратичному відхиленню об’єктів для цього кластера. Об’єкт ставиться до кластера, якщо відстань від об’єкта до центру кластера менше радіуса кластера. Якщо ця умова виконується для двох і більш кластерів, об’єкт є спірним.

Неоднозначність даного завдання може бути усунута експертом або аналітиком.

Робота кластерного аналізу опирається на два припущення. Перше припущення – розглянуті ознаки об’єкта в принципі допускають бажане розбиття сукупності об’єктів на кластери. Друге припущення – правильність вибору масштабу або одиниці вимірювання ознак.

Вибір масштабу в кластерному аналізі має велике значення. Розглянемо приклад. Уявимо собі, що дані ознаки х у наборі даних А на два порядки більші даних ознаки y: значення змінної х перебувають в діапазоні від 100 до 700, а значення змінної y – у діапазоні від 0 до 1.

Тоді, при розрахунках величини відстані між точками, що відображають положення об’єктів у просторі їх властивостей, змінна, що має більші значення, тобто змінна х, буде практично повністю домінувати над змінною з малими значеннями, тобто змінної y. У такий спосіб через неоднорідність одиниць виміру ознак стає неможливо коректно розрахувати відстані між точками.

Ця проблема вирішується за допомогою попередньої стандартизації змінних. Стандартизація (standardization) або нормування (normalization) приводить значення всіх перетворених змінних до єдиного діапазону значень шляхом вираження через відношення цих значень до якоїсь величини, що відображає певні властивості конкретної ознаки. Існують різні способи нормування вихідних даних.

Два найпоширеніші способи:

* розподіл вихідних даних на середньоквадратичне відхилення відповідних змінних;
* обчислення Z–внеску або стандартизованого внеску.

Поряд зі стандартизацією змінних, існує варіант додавання до кожної з них певного коефіцієнта важливості, або ваги, який би відображав значимість відповідної змінної. У якості ваг можуть виступати експертні оцінки, отримані в ході опитування експертів – фахівців предметної області. Отримані добутки нормованих змінних на відповідні ваги дозволяють одержувати відстані між точками в багатомірному просторі з урахуванням неоднакової ваги змінних.

У ході експериментів можливе порівняння результатів, отриманих з урахуванням експертних оцінок і без них, і вибір кращої з них.

**Ієрархічні і ітеративні методи**

*Методи кластерного аналізу можна розділити на дві групи:*

* ієрархічні;
* неієрархічні.

Кожна із груп включає безліч підходів і алгоритмів.Використовуючи різні методи кластерного аналізу, аналітик може одержати різні розв’язки для тих самих даних. Це вважається нормальним явищем.

Розглянемо ієрархічні й неієрархічні методи докладно.

***Ієрархічні методи кластерного аналізу.*** Суть ієрархічної кластеризації полягає в послідовному об’єднанні менших кластерів у більші або поділі більших кластерів на менші.

## Ієрархічні агломеративні методи (Agglomerative Nesting, AGNES). **Ця група методів характеризується послідовним об’єднанням вихідних елементів і відповідним зменшенням числа кластерів.**

На початку роботи алгоритму всі об’єкти є окремими кластерами. На першому кроці найбільш схожі об’єкти поєднуються в кластер. На наступних кроках об’єднання триває доти, поки всі об’єкти не будуть становити один кластер.

## Ієрархічні дивизимні (ділені) методи (Divisive Analysis, DIANA). **Ці методи є логічною протилежністю агломеративним методам. На початку роботи алгоритму всі об’єкти належать одному кластеру, який на наступних кроках ділиться на менші кластери, у результаті утворюється послідовність груп, що розщеплюються.**

Принцип роботи описаних вище груп методів у вигляді дендрограми показаний на [рис.](http://www.intuit.ru/department/database/datamining/13/datamining_13.html#image.13.3) 2.41.

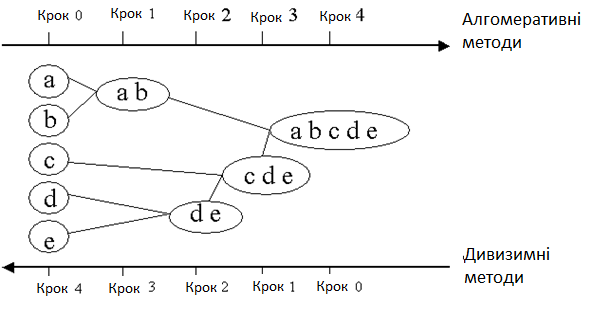


Рис. 2.41. Дендрограмма агломеративних і дивизимних методів

Програмна реалізація алгоритмів кластерного аналізу широко представлена в різних інструментах Data Mining, які дозволяють вирішувати завдання досить великої розмірності. Наприклад, агломеративні методи реалізовані в пакеті SPSS, дивизимні методи – у пакеті Statgraf.

Ієрархічні методи кластеризації різняться правилами побудови кластерів. У якості правил виступають критерії, які використовуються при розв’язку питання про «схожість» об’єктів при їхнім об’єднанні в групу (агломеративні методи) або поділу на групи (дивизимні методи).

Ієрархічні методи кластерного аналізу використовуються при невеликих обсягах наборів даних.

Перевагою ієрархічних методів кластеризації є їхню наочність.

Ієрархічні алгоритми пов’язані з побудовою дендрограмм ( від грецького dendron – «дерево»), які є результатом ієрархічного кластерного аналізу.

Дендрограмма описує близькість окремих точок і кластерів друг до друга, представляє в графічному виді послідовність об’єднання (поділу) кластерів.

Дендрограмма (dendrogram) – деревоподібна діаграма, що містить n рівнів, кожний з яких відповідає одному із кроків процесу послідовного укрупнення кластерів.

Дендрограмму також називають деревоподібною схемою, деревом об’єднання кластерів, деревом ієрархічної структури.

Дендрограмма являє собою вкладене угруповання об’єктів, яке змінюється на різних рівнях ієрархії.

Існує багато способів побудови дендограмм. У дендограмі об’єкти можуть розташовуватися вертикально або горизонтально. Приклад вертикальної дендрограми наведений на [рис. 2.4](http://www.intuit.ru/department/database/datamining/13/datamining_13.html#image.13.4)2.

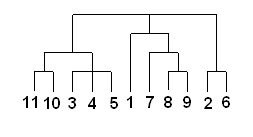


Рис. 2.42. Приклад дендрограмми

Числа 11, 10, 3 і т.д. відповідають номерам об’єктів або спостережень вихідної вибірки. Ми бачимо, що на першому кроці кожне спостереження представляє один кластер (вертикальна лінія), на другому кроці спостерігаємо об’єднання таких спостережень: 11 і 10; 3, 4 і 5; 8 і 9; 2 і 6. На другому кроці триває об’єднання в кластери: спостереження 11, 10, 3, 4, 5 і 7, 8, 9. Даний процес триває доти, поки всі спостереження не об’єднаються в один кластер.

Заходи подібності. Для обчислення відстані між об’єктами використовуються різні заходи подібності, їх називають також метриками або функціями відстаней. На початку теми ми розглянули евклідову відстань, це найбільш популярний захід подібності.

Квадрат евклідової відстані. Для надання більшої ваги більш віддаленим один від одного об’єктів можемо скористатися квадратом евклідової відстані шляхом піднесення у квадрат стандартної евклідової відстані.

Манхеттенска відстань (відстань міських кварталів), також називається «хемінговою» або « сіті-блок» відстанню.

Ця відстань розраховується як середня різниць по координатах. У більшості випадків ця міра відстані приводить до результатів, подібних розрахунків відстані евкліда. Однак, для цієї міри вплив окремих викидів менший, ніж при використанні евклидової відстані, оскільки тут координати не підносяться до квадру.

Відстань Чебишева. Цю відстань варто використовувати, коли необхідно визначити два об’єкти як «різні», якщо вони відрізняються за якимось одним виміром.

Відсоток незгоди. Ця відстань обчислюється, якщо дані є категоріальними.

Методи об’єднання або зв’язки. Коли кожний об’єкт являє собою окремий кластер, відстані між цими об’єктами визначаються обраною мірою. Виникає наступне питання – як визначити відстані між кластерами? Існують різні правила, називані методами об’єднання або зв’язками для двох кластерів.

Метод найближчого сусіда або одиночний зв’язок. Тут відстань між двома кластерами визначається відстанню між двома найбільш близькими об’єктами (найближчими сусідами) у різних кластерах. Цей метод дозволяє виділяти кластери як завгодно складної форми за умови, що різні частини таких кластерів з’єднані ланцюжками близьких один до одного елементів. У результаті роботи цього методу кластери представляються довгими «ланцюжками» або «волокнистими» кластерами, «зчепленими разом» тільки окремими елементами, які випадково виявилися ближче інших один до одного.

Метод найбільш віддалених сусідів або повний зв’язок. Тут відстані між кластерами визначаються найбільшою відстанню між будь–якими двома об’єктами в різних кластерах ( тобто « найбільш віддаленими сусідами»). Метод добре використовувати, коли об’єкти дійсно походять із різних «ділянок». Якщо ж кластери мають до певної міри подовжену форму або їх природній тип є «ланцюговим», то цей метод не слід використовувати.

Метод Варда (Ward’s method). У якості відстані між кластерами береться приріст суми квадратів відстаней об’єктів до центрів кластерів, одержуваний у результаті їх об’єднання (Ward, 1963). На відміну від інших методів кластерного аналізу для оцінки відстаней між кластерами, тут використовуються методи дисперсійного аналізу. На кожному кроці алгоритму поєднуються такі два кластери, які приводять до мінімального збільшення цільової функції, тобто внутрішньо групової суми квадратів.

Цей метод спрямований на об’єднання близько розташованих кластерів і «прагне» створювати кластери малого розміру.

Метод незваженого попарного середнього (метод незваженого попарного арифметичного середнього – unweighted pair–group method using arithmetic averages, UPGMA (Sneath, Sokal, 1973)).

У якості відстані між двома кластерами береться середня відстань між усіма парами об’єктів у них. Цей метод слід використовувати, якщо об’єкти дійсно походять із різних «ділянок», у випадках присутності кластерів «ланцюгового» типу, при припущенні нерівних розмірів кластерів.

Метод зваженого попарного середнього (метод зваженого попарного арифметичного середнього – weighted pair–group method using arithmetic averages, WPGM A (Sneath, Sokal, 1973)). Цей метод схожий на метод незваженого попарного середнього, різниця полягає лише в тому, що тут у якості вагового коефіцієнта використовується розмір кластера (число об’єктів, що втримуються в кластері).

Цей метод рекомендується використовувати саме при наявності припущення про кластери різних розмірів.

Незважений центроїдний метод (метод незваженого попарного центроїдного усереднення – unweighted pair–group method using the centroid average (Sneath and Sokal, 1973)).

У якості відстані між двома кластерами в цьому методі береться відстань між їхніми центрами ваги.

Зважений центроїдний метод (метод зваженого попарного центроїдного усереднення – weighted pair–group method using the centroid average, WPGMC (Sneath, Sokal 1973)). Цей метод схожий на попередній, різниця полягає в тому, що для обліку різниці між розмірами кластерів (числа об’єктів у них), використовуються ваги. Цей метод переважно використовують у випадках, якщо є припущення щодо істотних відмінностей у розмірах кластерів.

**Ієрархічний кластерний аналіз**

Розглянемо процедуру ієрархічного кластерного аналізу в пакеті SPSS (SPSS). Процедура ієрархічного кластерного аналізу в SPSS передбачає угруповання як об’єктів (рядків матриці даних), так і змінних (стовпців) [54]. Можна вважати, що в останньому випадку роль об’єктів відіграють змінні, а роль змінних – стовпці.

У цьому методі реалізується ієрархічний агломеративний алгоритм, зміст якого укладається в наступному. Перед початком кластеризації всі об’єкти вважаються окремими кластерами, у ході алгоритму вони поєднуються. Спочатку вибирається пара найближчих кластерів, які поєднуються в один кластер. У результаті кількість кластерів стає рівним N–1. Процедура повторюється, поки всі класи не об’єднаються. На будь–якому етапі об’єднання можна перервати, одержавши потрібне число кластерів. Таким чином, результат роботи алгоритму агрегування залежить від способів обчислення відстані між об’єктами й визначення близькості між кластерами.

Для визначення відстані між парою кластерів можуть бути сформульовані різні підходи. З обліком цього в SPSS передбачені наступні методи:

* Середня відстань між кластерами (Between–groups linkage), установлюється за замовчуванням.
* Середня відстань між усіма об’єктами пари кластерів з урахуванням відстаней усередині кластерів (Within–groups linkage).
* Відстань між найближчими сусідами – найближчими об’єктами кластерів (Nearest neighbor).
* Відстань між самими далекими сусідами (Furthest neighbor).
* Відстань між центрами кластерів (Centroid clustering) або центроидный метод. Недоліком цього методу є те, що центр об’єднаного кластера обчислюється як середнє центрів поєднуваних кластерів, без обліку їх обсягу.
* Метод медіан – той же центроидный метод, але центр об’єднаного кластера обчислюється як середнє всіх об’єктів (Median clustering).
* Метод Варда.

***Приклад ієрархічного кластерного аналізу.*** Порядок агломерації (протокол об’єднання кластерів) представлених раніше даних наведено в [табл. 2.11](http://www.intuit.ru/department/database/datamining/13/datamining_13.html#table.13.2). У протоколі зазначені такі позиції:

* Stage – стадії об’єднання (крок);
* Cluster Combined – поєднувані кластери (після об’єднання кластер ухвалює мінімальний номер з номерів поєднуваних кластерів);
* Coefficients – коефіцієнти.

Таблиця 2.11.

Порядок агломерації

**Cluster Combined Coefficients** 10,000

2 2 14 1,461E–02

3 3 9 1,461E–02

4 5 8 1,461E–02

5 6 7 1,461E–02

6 3 13 3,490E–02

7 2 11 3,651E–02

8 4 5 4,144E–02

9 2 6 5,118E–02

10 4 12 0,105

11 1 3 0,120

12 1 4 1,217

13 1 2 7,516

Так, у колонку Cluster Combined можна побачити порядок об’єднання в кластери: на першому кроці були об’єднані спостереження 9 і 10, вони утворюють кластер під номером 9, кластер 10 в оглядовій таблиці більше не з’являється. На наступному кроці відбувається об’єднання кластерів 2 і 14, далі 3 і 9, і т.д.

У колонку Coefficients наведена кількість кластерів, яка варто було б уважати оптимальним; під значенням цього показника мається на увазі відстань між двома кластерами, певне на підставі обраного заходу відстані. У нашому випадку це квадрат відстані, певний з використанням стандартизованих значень. Процедура стандартизації використовується для виключення ймовірності того, що класифікацію будуть визначати змін, що мають найбільший розкид значень. В SPSS застосовуються наступні види стандартизації:

* Z-Шкали (Z-Scores). Зі значень змінних віднімається їхнє середнє, і ці значення діляться на стандартне відхилення.
* Розкид від –1 до 1. Лінійним перетворенням змінних домагаються розкиду значень від –1 до 1.
* Розкид від 0 до 1. Лінійним перетворенням змінних домагаються розкиду значень від 0 до 1.
* Максимум 1. Значення змінних діляться на їхній максимум.
* Середнє 1. Значення змінних діляться на них середнє.
* Стандартне відхилення 1. Значення змінних діляться на стандартне відхилення.

Крім того, можливі перетворення самих відстаней, зокрема, можна відстані замінити їхніми абсолютними значеннями, це актуально для коефіцієнтів кореляції.

Можна також усі відстані перетворити так, щоб вони змінювалися від 0 до 1.

***Визначення кількості кластерів.*** Існує проблема визначення числа кластерів. Іноді можна апріорно визначити це число. Однак у більшості випадків число кластерів визначається в процесі агломерації/поділу безлічі об’єктів.

Процесу угруповання об’єктів в ієрархічному кластерному аналізі відповідає поступове зростання коефіцієнта, називаного критерієм Е. Стрибкоподібне збільшення значення критерію Е можна визначити як характеристику числа кластерів, які дійсно існують у досліджуваному наборі даних. Таким чином, цей спосіб зводиться до визначення стрибкоподібного збільшення деякого коефіцієнта, який характеризує перехід від сильно зв’язаного до слабко зв’язаного стану об’єктів.

У [табл. 2.11.](http://www.intuit.ru/department/database/datamining/13/datamining_13.html#table.13.2) ми бачимо, що значення поля Coefficients збільшується стрибкоподібно, отже, об’єднання в кластери слід зупинити, інакше буде відбуватися об’єднання кластерів, що перебувають на відносно великій відстані друг від друга.

У нашому прикладі це стрибок з 1,217 до 7,516. Оптимальним уважається кількість кластерів, рівне різниці кількості спостережень (14) і кількості кроків до стрибкоподібного збільшення коефіцієнта (12).

Отже, після створення двох кластерів об’єднань більше провадити не слід, хоча візуально ми очікували появи трьох кластерів.

Агрегування даних може бути презентовано графічно у вигляді дендрограмми. Вона визначає об’єднані кластери й значення коефіцієнтів на кожному кроці агломерації (відображені значення коефіцієнтів, наведені до шкали від 0 до 25).

Дендрограмма для нашого прикладу наведена на [рис. 2.43](http://www.intuit.ru/department/database/datamining/13/datamining_13.html#image.13.5). Розріз дерева агрегування вертикальною рисою дав нам два кластери, що полягають із 9 і 5 об’єктів.

На верхній лінії по горизонталі відзначені номери кроків алгоритму, усього алгоритму треба було 25 кроків для об’єднання всіх об’єктів в один кластер.

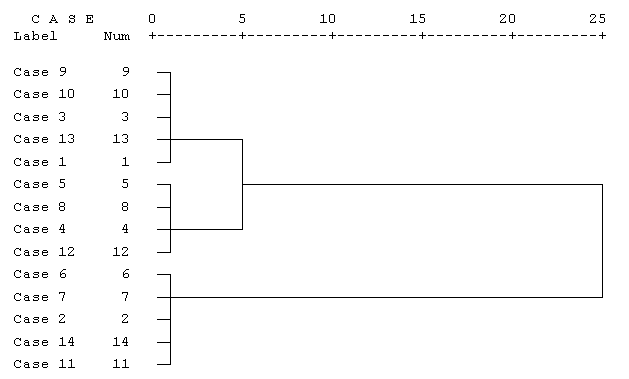


Рис. 2.43. Дендрограмма процесу злиття

**Алгоритми неієрархічної кластеризації**

При великій кількості спостережень ієрархічні методи кластерного аналізу непридатні.

У таких випадках використовують **неієрархічні методи**, засновані на поділі, які являють собою **ітеративні методи дроблення вихідної сукупності**. У процесі розподілу нові кластери формуються доти, поки не буде виконане **правило зупинки**.

Така неієрархічна кластеризація полягає в поділі набору даних на певну кількість окремих кластерів. **Існує два підходи**. Перший полягає у визначенні границь кластерів як найбільш щільних ділянок у багатомірному просторі вихідних даних, тобто визначення кластера там, де є велике «згущення точок». Другий підхід полягає в мінімізації міри відмінності об’єктів.

Найпоширеніший серед неієрархічних методів алгоритм k-середніх, також називають **швидким кластерним аналізом**. Повний опис алгоритму можна знайти в роботі Хартігана і Вонга (Hartigan and Wong, 1978). На відміну від ієрархічних методів, які не вимагають попередніх припущень щодо числа кластерів, для можливості використання цього методу необхідно мати гіпотезу про найбільш імовірну кількість кластерів.

Алгоритм k–середніх будує k кластерів, розташованих на максимально можливо великих відстанях один від одного. Основний тип задач, які вирішує алгоритм k-середніх, – наявність припущень (гіпотез) щодо числа кластерів, при цьому вони повинні бути різні настільки, наскільки це можливо. Вибір числа k може базуватися на результатах попередніх досліджень, теоретичних міркуваннях або інтуїції.

Загальна ідея алгоритму: задане фіксоване число k кластерів спостереження зіставляється кластерам так, що середні в кластері (для всіх змінних) максимально можливо відрізняються одна від одної.

***Опис алгоритму:***

1. Первісний розподіл об’єктів по кластерах.

Вибирається число k, і на першому кроці ці точки вважаються «центрами» кластерів. Кожному кластеру відповідає один центр.

Вибір початкових центроїдів може здійснюватися в такий спосіб:

* + вибір k-спостережень для максимізації початкової відстані;
  + випадковий вибір k-спостережень;
  + вибір перших k-спостережень.

У результаті кожний об’єкт призначений певному кластеру.

1. Ітеративний процес.

Обчислюються центри кластерів, якими потім і далі вважаються покоординатні середні кластерів. Об’єкти знову перерозподіляються.

Процес обчислення центрів і перерозподілу об’єктів триває доти, поки не виконана одна з умов:

* + кластерні центри стабілізувалися, тобто всі спостереження належать кластеру, якому належали до поточної ітерації;
  + число ітерацій дорівнює максимальному числу ітерацій.

На [рис. 2.44.](http://www.intuit.ru/department/database/datamining/14/datamining_14.html#image.14.1) наведений приклад роботи алгоритму k–середніх для k, рівного двом.

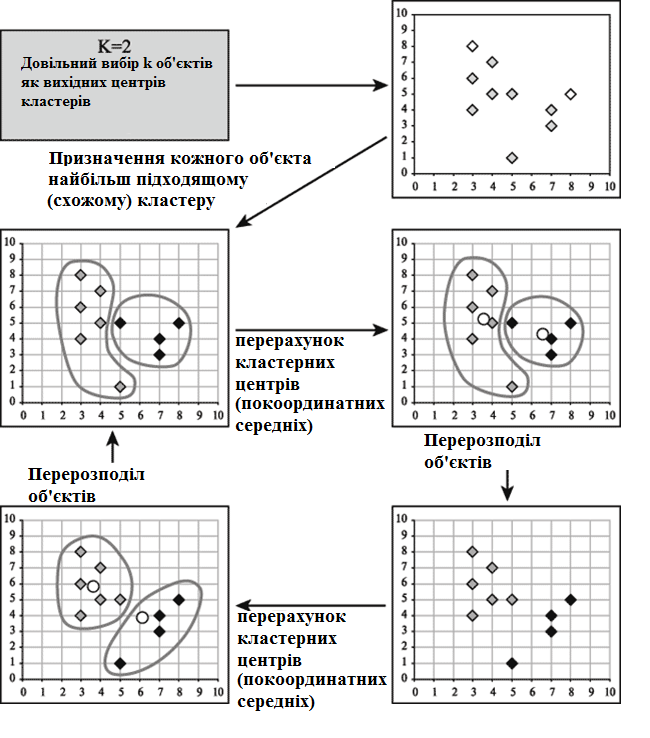


Рис. 2.44. Приклад роботи алгоритму k–середніх (k=2)

Вибір числа кластерів є складним питанням. Якщо немає припущень щодо цього числа, рекомендують створити 2 кластера, потім 3, 4, 5 і т.д., порівнюючи отримані результати.

***Перевірка якості кластеризації.*** Після одержання результатів кластерного аналізу методом k-середніх слід перевірити правильність кластеризації (тобто оцінити, наскільки кластери відрізняються один від одного). Для цього розраховуються середні значення для кожного кластера. При гарній кластеризації повинні бути отримані дуже відмінні середні для всіх вимірів або хоча б більшої їхньої частини.

*Переваги алгоритму k-середніх:*

* простота використання;
* швидкість використання;
* зрозумілість і прозорість алгоритму.

*Недоліки алгоритму k-середніх:*

* алгоритм занадто чутливий до викидів, які можуть спотворювати середнє. Можливим вирішенням цієї проблеми є використання модифікації алгоритму – алгоритм k-медіани;
* алгоритм може повільно працювати на великих базах даних. Можливим вирішенням даної проблеми є використання вибірки даних.

**Алгоритм PAM (partitioning around Medoids)** є модифікацією алгоритму k-середніх, алгоритмом k-медіани (k-medoids).

Алгоритм менш чутливий до шумів і викидів даних, ніж алгоритм k-means, оскільки медіана менше піддана впливам викидів.

PAM ефективний для невеликих баз даних, але його не слід використовувати для великих наборів даних.

Розглянемо приклад. Є база даних клієнтів фірми, яких слід розбити на однорідні групи. Кожний клієнт описується за допомогою 25 змінних.

Використання такого великого числа змінних приводить до виділення кластерів нечіткої структури. У результаті аналітикові досить складно інтерпретувати отримані кластери.

Більш зрозумілі й прозорі результати кластеризації можуть бути отримані, якщо замість множини вихідних змінних використовувати якісь узагальнені змінні або критерії, що містять у стислому вигляді інформацію про зв’язки між змінними. Тобто виникає задача зниження розмірності даних. Вона може вирішуватися за допомогою різних методів; один з найпоширеніших – факторний аналіз.

**Ітеративні методи кластеризації**

**Ітеративна кластеризація в SPSS.** Звичайно в статистичних пакетах реалізований широкий арсенал методів, що дозволяє спочатку провести скорочення розмірності набору даних (наприклад, за допомогою факторного аналізу), а потім уже властиво кластеризацію (наприклад, методом швидкого кластерного аналізу). Розглянемо цей варіант проведення кластеризації в пакеті SPSS.

Для скорочення розмірності вихідних даних скористаємося факторним аналізом. Для цього виберемо в меню: Analyze (Аналіз)/Data Reduction (Перетворення даних)/Factor (Факторний аналіз):

За допомогою кнопки Extraction:(Відбір) слід вибрати метод відбору. Ми залишимо обраний за замовчуванням аналіз головних компонентів, який згадувався вище. Також слід вибрати метод обертання – виберемо один з найбільш популярних – метод Варимакса. Для збереження значень факторів у вигляді змінних у закладці «Значення» необхідно поставити оцінку «Save as variables» (Зберегти як змінні).

У результаті цієї процедури користувач одержує звіт «Пояснена сумарна дисперсія», по якій видне число відібраних факторів – це ті компоненти, власні значення яких перевершують одиницю.

Отримані значення факторів, яким звичайно привласнюються назви fact1\_1, fact1\_2 і т.д., використовуємо для проведення кластерного аналізу методом k–середніх. Для проведення швидкого кластерного аналізу виберемо в меню:

Analyze (Аналіз)/Classify(Класифікувати)/K–Means Cluster: (Кластерный аналіз методом k–середніх).

У діалоговому вікні K Means Cluster Analysis (Кластерний аналіз методом k–середніх) необхідно помістити факторні змінні fact1\_1, fact1\_2 і т.д. у поле тестованих змінних. Тут же необхідно вказати кількість кластерів і кількість ітерацій.

У результаті цієї процедури одержуємо звіт з висновком значень центрів сформованих кластерів, кількості спостережень у кожному кластері, а також з додатковою інформацією, заданої користувачем.

Таким чином, алгоритм k–середніх ділить сукупність вихідних даних на задану кількість кластерів. Для можливості візуалізації отриманих результатів слід скористатися одним із графіків, наприклад, діаграмою розсіювання. Однак традиційна візуалізація можлива для обмеженої кількості вимірів, тому що, як відомо, людина може сприймати тільки тривимірний простір. Тому, якщо ми аналізуємо більш трьох змінних, слід використовувати спеціальні багатомірні методи представлення інформації.

**Ітеративні методи кластеризації** різняться вибором наступних параметрів:

* початкової точки;
* правилом формування нових кластерів;
* правилом зупинки.

Вибір методу кластеризації залежить від кількості даних і від того, чи є необхідність працювати одночасно з декількома типами даних.

У пакеті SPSS, наприклад, при необхідності роботи як з кількісними (наприклад, дохід), так і з категоріальними (наприклад, родиний стан) змінними, а також якщо обсяг даних досить великий, використовується метод Двоетапного кластерного аналізу, який являє собою масштабовану процедуру кластерного аналізу, що дозволяє працювати з даними різних типів.

Для цього на першому етапі роботи записи попередньо кластеризуются у велику кількість суб–кластерів. На другому етапі отримані суб–кластери групуються в необхідну кількість. Якщо ця кількість невідома, процедура сама автоматично визначає її. За допомогою цієї процедури банківський працівник може, наприклад, виділяти групи людей, одночасно використовуючи такі показники як вік, стать і рівень доходу. Отримані результати дозволяють визначити клієнтів, вхідних у групи ризику неповернення кредиту.

***Процес кластерного аналізу. Рекомендовані етапи.***

У загальному випадку всі етапи кластерного аналізу взаємозалежні, і розв’язки, прийняті на одному з них, визначають дії на наступних етапах.

Аналітикові слід розв’язати, чи використовувати всі спостереження або ж виключити деякі дані або вибірки з набору даних:

* Вибір метрики й методу стандартизації вихідних даних.
* Визначення кількості кластерів (для ітеративного кластерного аналізу).
* Визначення методу кластеризації (правила об’єднання або зв’язки).

На думку багатьох фахівців, вибір методу кластеризації є вирішальним при визначенні форми й специфіки кластерів.

**Аналіз результатів кластеризації**. Цей етап має на увазі вирішення таких питань: чи не є отримана розбивка на кластери випадковою; чи є розбивка надійною й стабільною на підвибірках даних; чи існує взаємозв’язок між результатами кластеризації й змінними, які не брали участь у процесі кластеризації; чи можна інтерпретувати отримані результати кластеризації.

**Перевірка результатів кластеризації**. Результати кластеризації також повинні бути перевірені формальними й неформальними методами. Формальні методи залежать від того методу, який використовувався для кластеризації. Неформальні включають наступні процедури перевірки якості кластеризації:

* аналіз результатів кластеризації, отриманих на певних вибірках набору даних;
* крос-перевірка;
* проведення кластеризації при зміні порядку спостережень у наборі даних;
* проведення кластеризації при видаленні деяких спостережень;
* проведення кластеризації на невеликих вибірках.

Один з варіантів перевірки якості кластеризації – використання декількох методів і порівняння отриманих результатів. Відсутність подібності не буде означати некоректність результатів, але присутність схожих груп вважається ознакою якісної кластеризації.

**Складності й проблеми, які можуть виникнути при застосуванні кластерного аналізу**

Як і будь-які інші методи, методи кластерного аналізу мають певні слабкі сторони, тобто деякі складності, проблеми й обмеження.

При проведенні кластерного аналізу слід враховувати, що результати кластеризації залежать від критеріїв розбивки сукупності вихідних даних. При зниженні розмірності даних можуть виникнути певні викривлення, за рахунок узагальнень можуть згубитися деякі індивідуальні характеристики об’єктів.

Існує ряд складностей, які слід продумати перед проведенням кластеризації.

* Складність вибору характеристик, на основі яких проводиться кластеризація. Необдуманий вибір приводить до неадекватної розбивки на кластери й, як наслідок, – до неправильного розв’язку завдання.
* Складність вибору методу кластеризації. Цей вибір вимагає непоганого знання методів і передумов їх використання. Щоб перевірити ефективність конкретного методу в певній предметній області, доцільно застосувати наступну процедуру: розглядають декілька апріорі різних між собою груп і перемішують їхніх представників між собою випадковим чином. Далі проводиться кластеризація для відновлення вихідної розбивки на кластери. Частка збігів об’єктів у виявлених і вихідних групах є показником ефективності роботи методу.
* Проблема вибору числа кластерів. Якщо немає ніяких відомостей щодо можливого числа кластерів, необхідно провести ряд експериментів і, у результаті перебору різного числа кластерів, вибрати оптимальне їхнє число.
* Проблема інтерпретації результатів кластеризації. Форма кластерів у більшості випадків визначається вибором методу об’єднання. Однак слід враховувати, що конкретні методи прагнуть створювати кластери певних форм, навіть якщо в досліджуваному наборі даних кластерів насправді немає.

**Порівняльний аналіз ієрархічних і неієрархічних методів кластеризації**

***Порівняльний аналіз ієрархічних і неієрархічних методів кластеризації.*** Перед проведенням кластеризації в аналітика може виникнути питання, якій групі методів кластерного аналізу віддати перевагу. Вибираючи між ієрархічними й неієрархічними методами, необхідно враховувати наступні їхні особливості.

**Неієрархічні методи** виявляють більш високу стабільність стосовно шумів і викидів, некоректного вибору метрики, включення незначущих змінних у набір, що брав участь у кластеризації. Ціною, яку доводиться платити за ці переваги методу, є слово «апріорі». Аналітик повинен заздалегідь визначити кількість кластерів, кількість ітерацій або правило зупинки, а також деякі інші параметри кластеризації. Це особливо складно починаючим фахівцям.

Якщо немає припущень щодо числа кластерів, рекомендують використовувати ієрархічні алгоритми. Однак якщо обсяг вибірки не дозволяє це зробити, можливий шлях – проведення ряду експериментів з різною кількістю кластерів, наприклад, почати розбивку сукупності даних із двох груп і, поступово збільшуючи їх кількість, порівнювати результати. За рахунок такого «варіювання» результатів досягається значно більша гнучкість кластеризації.

**Ієрархічні методи**, на відміну від неієрархічних, відмовляються від визначення числа кластерів, а будують повне дерево вкладених кластерів.

Складності ієрархічних методів кластеризації: обмеження обсягу набору даних; вибір міри близькості; негнучкість отриманих класифікацій.

Перевага цієї групи методів у порівнянні з неієрархічними методами – їх наочність і можливість одержати детальне представлення про структуру даних.

При використанні ієрархічних методів існує можливість досить легко ідентифікувати викиди в наборі даних і, у результаті, підвищити якість даних. Ця процедура лежить в основі двокрокового алгоритму кластеризації. Такий набір даних надалі може бути використаний для проведення неієрархічної кластеризації.

Існує ще один аспект, про який уже згадувалося в цій лекції. Це питання кластеризації всієї сукупності даних або ж її вибірки. Названий аспект вагомий для обох розглянутих груп методів, однак він більш критичний для ієрархічних методів. Ієрархічні методи не можуть працювати з більшими наборами даних, а використання деякої вибірки, тобто частини даних, могло б дозволити застосовувати ці методи.

Результати кластеризації можуть не мати достатнього статистичного обґрунтування. З іншого боку, при розв’язку задач кластеризації припустима нестатистична інтерпретація отриманих результатів, а також досить велика різноманітність варіантів поняття кластера. Така нестатистична інтерпретація дає можливість аналітикові одержати задовольняючі його результати кластеризації, що при використанні інших методів часто буває скрутним.

**Нові алгоритми й деякі модифікації алгоритмів кластерного аналізу.** Методи, які ми розглянули в цій і попередній лекціях, є «класикою» кластерного аналізу. До останнього часу основним критерієм, по якому оцінювався алгоритм кластеризації, була якість кластеризації: вважалося, щоб увесь набір даних вміщався в оперативній пам’яті.

Однак зараз, у зв’язку з появою надвеликих баз даних, з’явилися нові вимоги, яким повинен задовольняти алгоритм кластеризації. Основна з них, як уже згадувалося в попередній темі, – це масштабованість алгоритму.

Відзначимо також інші властивості, які повинен задовольняти алгоритм кластеризації: незалежність результатів від порядку вхідних даних; незалежність параметрів алгоритму від вхідних даних.

Останнім часом ведуться активні розробки нових алгоритмів кластеризації, здатних обробляти надвеликі бази даних. У них основна увага приділяється масштабованості. До таких алгоритмів відноситься узагальнене представлення кластерів (summarized cluster representation), а також вибірка й використання структур даних, підтримуваних СУБД.

Розроблені алгоритми, у яких методи ієрархічної кластеризації інтегровані з іншими методами. До таких алгоритмів ставляться: BIRCH, CURE, CHAMELEON, ROCK.

**Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)**запропонований Тьян Зангом і його колегами.

Завдяки узагальненим представленням кластерів, швидкість кластеризації збільшується, алгоритм при цьому має більше масштабування.

У цьому алгоритмі реалізований двоетапний процес кластеризації.

У ході першого етапу формується попередній набір кластерів. На другому етапі до виявлених кластерів застосовуються інші алгоритми кластеризації – придатні для роботи в оперативній пам’яті.

В наведена наступна аналогія, що описує цей алгоритм. Якщо кожний елемент даних уявити собі як бусинку, що лежить на поверхні стола, то кластери бусин можна «замінити» тенісними кульками й перейти до більш детального вивчення кластерів тенісних кульок. Число бусин може виявитися досить велике, однак діаметр тенісних кульок можна підібрати таким чином, щоб на другому етапі можна було, застосувавши традиційні алгоритми кластеризації, визначити дійсну складну форму кластерів.

***Алгоритм Wavecluster*** являє собою алгоритм кластеризації на основі хвильових перетворень. На початку роботи алгоритму дані узагальнюються шляхом накладення на простір даних багатомірних ґрат. На подальших кроках алгоритму аналізуються не окремі точки, а узагальнені характеристики точок, що потрапили в одне гніздо ґрат. У результаті такого узагальнення необхідна інформація поміщається в оперативній пам’яті. На наступних кроках для визначення кластерів алгоритм застосовує хвильове перетворення до узагальнених даних.

*Головні особливості Wavecluster:*

1. Складність реалізації;
2. Алгоритм може виявляти кластери довільних форм;
3. Алгоритм не чутливий до шумів;
4. Алгоритм застосуємо тільки до даних низкої розмірності.

***Алгоритм CLARA* (Clustering Large Applications)** був розроблений Kaufmann і Rousseeuw в 1990 році для кластеризації даних у великих базах даних. Даний алгоритм будується в статистичних аналітичних пакетах, наприклад, таких як S+.

Викладемо коротко суть алгоритму. Алгоритм CLARA витягає множину зразків з бази даних. Кластеризація застосовується до кожного зі зразків, на виході алгоритму пропонується краща кластеризація.

Для великих баз даних цей алгоритм ефективніший, ніж алгоритм PAM. Ефективність алгоритму залежить від обраного в якості зразка набору даних. Гарна кластеризація на обраному наборі може не дати гарну кластеризацію на всій множині даних.

***Алгоритми Clarans, CURE, Dbscan*** формулює задачу кластеризації як випадковий пошук у графові. У результаті роботи цього алгоритму сукупність вузлів графа являє собою розбивку множини даних на число кластерів, визначене користувачем. «Якість» отриманих кластерів визначається за допомогою критеріальної функції. Алгоритм Clarans сортує всі можливі розбивки множини даних у пошуках прийнятного розв’язку. Пошук розв’язку зупиняється в тому вузлі, де досягається мінімум серед визначеного числа локальних мінімумів.

Серед нових масштабованих алгоритмів також можна відзначити алгоритм CURE – алгоритм ієрархічної кластеризації, і алгоритм Dbscan, де поняття кластера формулюється з використанням концепції щільності (density).

Основним недоліком алгоритмів BIRCH, Clarans, CURE, Dbscan є та обставина, що вони вимагають задання деяких порогів щільності точок, а це не завжди прийнятне. Ці обмеження обумовлені тим, що описані алгоритми орієнтовані на надвеликі бази даних і не можуть користуватися великими обчислювальними ресурсами.

Над масштабованими методами зараз активно працюють багато дослідників, основне завдання яких – подолати недоліки алгоритмів, що існують на сьогоднішній день.

***3. АЛГОРИТМ k-СЕРЕДНІХ***

**Кластеризація методом k-середніх** (англ. *k-means clustering*) – популярний метод кластеризації, – впорядкування множини об’єктів в порівняно однорідні групи. Винайдений в 1950-х роках математиком Гуґо Штайнгаузом і майже одночасно Стюартом Ллойдом. Особливу популярність отримав після виходу роботи МакКвіна.

Мета методу – розділити *n* спостережень на *k* кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера.

Термін «k-середніх» був уперше вжитий Джеймсом МакКвіном (англ. *James MacQueen*) у 1967 році, хоча ідею методу вперше озвучив Гуґо Штайнгауз (англ. *Hugo Steinhaus*) у 1957 році. Стандартний алгоритм був вперше запропонований Стюартом Лойдом (англ. *Stuart Lloyd*) у 1957 р.

Маємо масив спостережень (об’єктів), кожен з яких має певні значення по ряду ознак. Відповідно до цих значень об’єкт розташовується у багатовимірному просторі.

1. Дослідник визначає кількість кластерів, що необхідно утворити
2. Випадковим чином обирається k спостережень, які на цьому кроці вважаються центрами кластерів
3. Кожне спостереження «приписується» до одного з n кластерів – того, відстань до якого найкоротша
4. Розраховується новий центр кожного кластера як елемент, ознаки якого розраховуються як середнє арифметичне ознак об’єктів, що входять у цей кластер
5. Відбувається така кількість ітерацій (повторюються кроки 3-4), поки кластерні центри стануть стійкими (тобто при кожній ітерації в кожному кластері опинятимуться одні й ті самі об’єкти), дисперсія всередині кластера буде мінімізована, а між кластерами – максимізована

Вибір кількості кластерів відбувається на основі дослідницької гіпотези. Якщо її немає, то рекомендують створити 2 кластери, далі 3,4,5, порівнюючи отримані результати.

**Демонстрація алгоритму**

|  |  |
| --- | --- |
| [https://upload.wikimedia.org/wikipedia/commons/thumb/5/5e/K_Means_Example_Step_1.svg/124px-K_Means_Example_Step_1.svg.png](https://uk.wikipedia.org/wiki/%D0%A4%D0%B0%D0%B9%D0%BB:K_Means_Example_Step_1.svg)  1. *k* початкових «середніх» (тут *k*=3) випадково згенеровані у межах домени даних (кольорові) | [https://upload.wikimedia.org/wikipedia/commons/thumb/a/a5/K_Means_Example_Step_2.svg/139px-K_Means_Example_Step_2.svg.png](https://uk.wikipedia.org/wiki/%D0%A4%D0%B0%D0%B9%D0%BB:K_Means_Example_Step_2.svg)  2. створено *k* кластерів, асоціюючи кожне спостереження з найближчим середнім. Розбиття відбувається згідно з діаграмою Вороного утвореною середніми |
| [https://upload.wikimedia.org/wikipedia/commons/thumb/3/3e/K_Means_Example_Step_3.svg/139px-K_Means_Example_Step_3.svg.png](https://uk.wikipedia.org/wiki/%D0%A4%D0%B0%D0%B9%D0%BB:K_Means_Example_Step_3.svg)  3. Центроїд кожного з *k* кластерів стає новим середнім | [https://upload.wikimedia.org/wikipedia/commons/thumb/d/d2/K_Means_Example_Step_4.svg/139px-K_Means_Example_Step_4.svg.png](https://uk.wikipedia.org/wiki/%D0%A4%D0%B0%D0%B9%D0%BB:K_Means_Example_Step_4.svg)  4. Кроки 2 і 3 повторюються до досягнення збіжності. |

Принцип алгоритму полягає в пошуку таких центрів кластерів та наборів елементів кожного кластера при наявності деякої функції Ф(°), що виражає якість поточного розбиття множини на k кластерів, коли сумарне квадратичне відхилення елементів кластерів від центрів цих кластерів буде найменшим.

В початковий момент роботи алгоритму довільним чином обираються центри кластерів, далі для кожного елемента множини ітеративно обраховується відстань від центрів з приєднанням кожного елемента до кластера з найближчим центром. Для кожного з отриманих кластерів обчислюються нові значення центрів, намагаючись при цьому мінімізувати функцію Ф(°), після чого повторюється процедура перерозподілу елементів між кластерами.

Алгоритм методу «Кластеризація за схемою к-середніх»:

* вибрати k інформаційних точок як центри кластерів поки не завершиться процес зміни центрів кластерів;
* зіставити кожну інформаційну точку з кластером, відстань до центра якого мінімальна;
* переконатися, що в кожному кластері міститься хоча б одна точка. Для цього кожний порожній кластер потрібно доповнити довільною точкою, що розташована «далеко» від центра кластера;
* центр кожного кластера замінити середнім від елементів кластера;
* кінець.

Головні переваги методу k-середніх — його простота та швидкість виконання. Метод k-середніх більш зручний для кластеризації великої кількості спостережень, ніж метод ієрархічного кластерного аналізу (у якому дендограми стають перевантаженими і втрачають наочність).

Одним із недоліків простого методу є порушення умови зв’язності елементів одного кластера, тому розвиваються різні модифікації методу, а також його нечіткі аналоги (англ. *fuzzy k-means methods*), у яких на першій стадії алгоритму допускається приналежність одного елемента множини до декількох кластерів (із різним ступенем приналежності).

Незважаючи на очевидні переваги методу, він має суттєві недоліки:

1. Результат класифікації сильно залежить від випадкових початкових позицій кластерних центрів
2. Алгоритм чутливий до викидів, які можуть викривлювати середнє
3. Кількість кластерів повинна бути заздалегідь визначена дослідником

Метод k-середніх є доволі простим і прозорим, тому успішно використовується у різноманітних сферах – маркетингових сегментаціях, геостатистиці, астрономії, сільському господарстві тощо.

**Існуючі реалізації алгоритму. Відкрите програмне забезпечення**

CrimeStat

Програмне забезпечення, створене для операційних систем Windows, що надає інструменти статистичного і просторового аналізу для вирішення завдання картування злочинності.

Julia

Високорівнева високопродуктивний вільний мову програмування з динамічною типізацією, створений для математичних обчислень, містить реалізацію k–means.

Mahout

Apache Mahout – Java бібліотека для роботи з алгоритмами машинного навчання з використанням MapReduce. Містить реалізацію k–means.

Octave

Написана на C ++ вільна система для математичних обчислень, яка використовує сумісний з MATLAB мову високого рівня, містить реалізацію k–means.

Spark

Розподілена реалізація k–means міститься в бібліотеці Mlib для роботи з алгоритмами машинного навчання, що взаємодіє з Python бібліотекою NumPy і бібліотека R.

Torch

MATLAB–подібна бібліотека для мови програмування Lua з відкритим вихідним кодом, надає велику кількість алгоритмів для глибинного навчання і наукових розрахунків. Ядро написано на Сі, прикладна частина виконується на LuaJIT, підтримується розпаралелювання обчислень засобами CUDA і OpenMP. Існують реалізації k–means.

Weka

Cвободное програмне забезпечення для аналізу даних, написане на Java. Містить k–means та x–means.

Accord.NET

C # реалізація алгоритмів k–means, k–means ++, k–modes.

OpenCV

Написана на С ++ бібліотека, спрямована в основному на вирішення завдань комп’ютерного зору. Містить реалізацію k–means.

MLPACK

Масштабна С ++ бібліотека для роботи з алгоритмами машинного навчання, містить реалізацію k–means.

SciPy

Бібліотека Python, містить безліч реалізацій k–means.

scikit–learn

Бібліотека Python, містить безліч реалізацій k–means.

R

Мова програмування для статистичної обробки даних і роботи з графікою, а також вільна програмне середовище обчислень з відкритим вихідним кодом в рамках проєкту GNU, містить три реалізації k–means.

ELKI

Java фреймворк, що містить реалізацію k–means, а також безліч інших алгоритмів кластеризації.

2.7.2 Пропрієтарне програмне забезпечення

Ayasdi

Stata

Mathematica

MATLAB

SAS

RapidMiner

SAP HANA

***4. НЕЙРОННІ МЕРЕЖІ. МОДЕЛІ НЕЙРОННИХ МЕРЕЖ***

**Класифікація нейронних мереж**

Одна з можливих класифікацій нейронних мереж – за спрямованістю зв’язків. Нейронні мережі бувають зі зворотними зв’язками й без зворотних зв’язків.

*Мережі без зворотних зв’язків:*

* Мережі зі зворотним поширенням помилки. Мережі цієї групи характеризуються фіксованою структурою, ітераційним навчанням, коректуванням ваг за помилками.
* Інші мережі (когнитрон, неокогнитрон, інші складні моделі).

Перевагами мереж без зворотних зв’язків є простота їх реалізації й гарантоване одержання відповіді після проходження даних по шарах.

Недоліком цього виду мереж вважається мінімізація розмірів мережі – нейрони багаторазово беруть участь в обробці даних.

Менший обсяг мережі полегшує процес навчання. Мережі зі зворотними зв’язками

* Мережі Хопфілда (задачі асоціативної пам’яті).
* Мережі Кохонена (задачі кластерного аналізу).

Перевагами мереж зі зворотними зв’язками є складність навчання, викликана більшим числом нейронів для алгоритмів того самого рівня складності.

Недоліки цього виду мереж – потрібні спеціальні умови, що гарантують збіжність обчислень.

Інша класифікація нейронних мереж: мережі прямого поширення й рекуррентні мережі.

*Мережі прямого поширення:*

* Персептрони.
* Мережа Back Propagation.
* Мережа зустрічного поширення.
* Карта Кохонена.

Рекуррентні мережі. Характерна риса таких мереж – наявність блоків динамічної затримки й зворотних зв’язків, що дозволяє їм обробляти динамічні моделі.

* Мережа Хопфілда.
* Мережа Елмана – мережа, що складається із двох шарів, у якій схований шар охоплений динамічним зворотним зв’язком, що дозволяє врахувати передісторію спостережуваних процесів і нагромадити інформацію для вироблення правильної стратегії керування. Ці мережі застосовуються в системах керування об’єктами, що рухаються.

Нейронні мережі можуть навчатися із учителем або без нього.

При навчанні із учителем для кожного навчального вхідного прикладу потрібне знання правильної відповіді або функції оцінки якості відповіді. Таке навчання називають керованим. Нейронній мережі пред’являються значення вхідних і вихідних сигналів, а вона за певним алгоритмом підбудовує ваги синоптичних зв’язків. У процесі навчання проводиться коректування ваг мережі за результатами порівняння фактичних вихідних значень із вхідними, відомими заздалегідь.

При навчанні без учителя розкривається внутрішня структура даних або кореляції між зразками в наборі даних. Виходи нейронної мережі формуються самостійно, а ваги змінюються за алгоритмом, що враховує тільки вхідні й похідні від них сигнали. Це навчання називають також некерованим. У результаті такого навчання об’єкти або приклади розподіляються по категоріях, самі категорії і їх кількість можуть бути заздалегідь не відомі.

Підготовка даних для навчання. При підготовці даних для навчання нейронної мережі необхідно звертати увагу на наступні істотні моменти.

Кількість спостережень у наборі даних. Слід враховувати той фактор, що чим більше розмірність даних, тим більше часу буде потрібно для навчання мережі.

Робота з викидами. Слід визначити наявність викидів і оцінити необхідність їх присутності у вибірці.

Навчальна вибірка повинна бути представницькою (репрезентативною).

Навчальна вибірка не повинна містити протиріч, тому що нейронна мережа однозначно зіставляє вихідні значення вхідним.

Нейронна мережа працює тільки із числовими вхідними даними, тому важливим етапом при підготовці даних є перетворення й кодування даних.

При використанні на вхід нейронної мережі слід подавати значення з того діапазону, на якому вона навчалася. Наприклад, якщо при навчанні нейронної мережі на один з її входів подавалися значення від 0 до 10, то при її застосуванні на вхід слід подавати значення із цього ж діапазону або прилеглих.

Існує поняття нормалізації даних. Метою нормалізації значень є перетворення даних до вигляду, який найбільше підходить для обробки, тобто дані, що надходять на вхід, повинні мати числовий тип, а їх значення повинні бути розподілені в певному діапазоні. Нормалізатор може приводити дискретні дані до набору унікальних індексів або перетворювати значення, що лежать в довільному діапазоні, у конкретний діапазон. Нормалізація виконується шляхом розподілу кожного компонента вхідного вектора на довжину вектора, що перетворює вхідний вектор в одиничний.

**Вибір структури нейронної мережі**

Вибір структури нейронної мережі обумовлюється специфікою й складністю розв’язуваної задачі. Для розв’язку деяких типів задач розроблені оптимальні конфігурації.

У більшості випадків вибір структури нейронної мережі визначається на основі об’єднання досвіду й інтуїції розроблювача.

Однак існують основні принципи, якими слід керуватися при розробці нової конфігурації:

1. можливості мережі зростають зі збільшенням числа гнізд мережі, щільності зв’язків між ними й числа виділених шарів;
2. введення зворотних зв’язків поряд зі збільшенням можливостей мережі піднімає питання про динамічну стабільність мережі;
3. складність алгоритмів функціонування мережі (у тому числі, наприклад, введення декількох типів синапсів – збуджуючих, гальмуючих та ін.) також сприяє посиленню потужності нейронної мережі.

Питання про необхідні й достатні властивості мережі для розв’язку того або іншого роду задач являє собою цілий напрямок нейронної комп’ютерної науки. Тому що проблема синтезу нейронної мережі сильно залежить від розв’язуваної задачі, дати загальні докладні рекомендації важко. Очевидно, що процес функціонування НМ (нейронної мережі), тобто сутність дій, які вона здатна виконувати, залежить від величин синаптичних зв’язків, тому, задавшись певною структурою НМ, що відповідає якому–небудь завданню, розроблювач мережі повинен знайти оптимальні значення всіх змінних вагових коефіцієнтів (деякі синаптичні зв’язки можуть бути постійними).

***5. КАРТИ КОХОНЕНА***

**Карти Кохонена****,** карти, що самоорганізуються (Self–Organizing Maps). Мережі, що називаються картами Кохонена, – це один з різновидів нейронних мереж, однак вони принципово відрізняються від розглянутих вище, оскільки використовують неконтрольоване навчання. Нагадаємо, що при такому навчанні навчальна множина складається лише зі значень вхідних змінних, у процесі навчання немає порівняння виходів нейронів з еталонними значеннями. Можна сказати, що така мережа вчиться розуміти структуру даних.

Ідея мережі Кохонена належить фінському вченому Тойво Кохонену (1982 рік). Основний принцип роботи мереж – введення в правило навчання нейрона інформації щодо його розташування.

В основі ідеї мережі Кохонена лежить аналогія із властивостями людського мозку. Кора головного мозку людину являє собою плоский аркуш зі згорнутими складками. Таким чином, можна сказати, що вона має певні топологічні властивості (ділянки, відповідальні за близькі частини тіла, примикають одна до однієї й усе зображення людського тіла відображається на цю двовимірну поверхню).

***Задачі, що розв’язуються за допомогою карт Кохонена.*** Карти, що самоорганізуються , можуть використовуватися для розв’язку таких задач, як моделювання, прогнозування, пошук закономірностей у великих масивах даних, виявлення наборів незалежних ознак і стискання інформації.

Найпоширеніше застосування мереж Кохонена – розв’язок задачі класифікації без учителя, тобто кластеризації.

Нагадаємо, що при такій постановці задачі нам даний набір об’єктів, кожному з яких зіставлений рядок таблиці (вектор значень ознак). Потрібно розбити вихідну множину на класи, тобто для кожного об’єкта знайти клас, до якого він належить.

У результаті одержання нової інформації про класи можлива корекція існуючих правил класифікації об’єктів.

Два з розповсюджених застосувань карт Кохонена: розвідницький аналіз даних і виявлення нових явищ.

Розвідницький аналіз даних. Мережа Кохонена здатна розпізнавати кластери в даних, а також установлювати близькість класів. Таким чином, користувач може поліпшити своє розуміння структури даних, щоб потім уточнити нейромережеву модель. Якщо в даних розпізнані класи, то їх можна позначити, після чого мережа зможе вирішувати задачу класифікації. Мережі Кохонена можна використовувати й у тих задачах класифікації, де класи вже задані, – тоді перевага буде в тому, що мережа зможе виявити подібність між різними класами.

Виявлення нових явищ. Мережа Кохонена розпізнає кластери в навчальних даних і відносить усі дані до тем або інших кластерів. Якщо після цього мережа зустрінеться з набором даних, несхожим ні на один з відомих зразків, то вона не зможе класифікувати такий набір і тим самим виявить його новизну.

Мережа Кохонена, на відміну від багатошарової нейронної мережі, дуже проста; вона являє собою два шари: вхідний і вихідний. Її також називають самоорганізованою картою. Елементи карти розташовуються в деякому просторі, як правило, двовимірному. Мережа Кохонена зображена на [рис. 2.45](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.1).

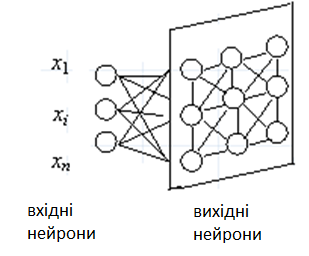


Рис. 2.45. Мережа Кохонена

Мережа Кохонена навчається методом послідовних наближень. У процесі навчання таких мереж на входи подаються дані, але мережа при цьому підбудовується не під еталонне значення виходу, а під закономірності у вхідних даних. Починається навчання з обраного випадковим чином вихідного розташування центрів.

У процесі послідовної подачі на вхід мережі навчальних прикладів визначається найбільш схожий нейрон (той, у якого скалярний добуток ваг і поданого на вхід вектора мінімальні). Цей нейрон оголошується переможцем і є центром при підстроюванні ваг у сусідніх нейронів. Таке правило навчання припускає «змагальне» навчання з урахуванням відстані нейронів від «нейрона-переможця».

Навчання при цьому полягає не в мінімізації помилки, а в підстроюванні ваг (внутрішніх параметрів нейронної мережі) для найбільшого збігу із вхідними даними.

Основний ітераційний алгоритм Кохонена послідовно проходить ряд епох, на кожній з яких обробляється один приклад з навчальної вибірки. Вхідні сигнали послідовно пред’являються мережі, при цьому бажані вихідні сигнали не визначаються. Після пред’явлення достатнього числа вхідних векторів синаптичні ваги мережі стають здатні визначити кластери. Ваги організують так, що топологічно близькі вузли чутливі до схожих вхідних сигналів.

У результаті роботи алгоритму центр кластера встановлюється в певній позиції, задовільним чином кластеризують приклади, для яких даний нейрон є «переможцем». У результаті навчання мережі необхідно визначити міру сусідства нейронів, тобто околицю нейрона-переможця.

Околиця являє собою кілька нейронів, які оточують нейрона-переможця.

Спочатку до околиці належить велика кількість нейронів, далі її розмір поступово зменшується. Мережа формує топологічну структуру, у якій схожі приклади утворюють групи прикладів, що близько перебувають на топологічній карті.

Отриману карту можна використовувати як засіб візуалізації при аналізі даних. У результаті навчання карта Кохонена класифікує вхідні приклади на кластери (групи схожих прикладів) і візуально відображає багатомірні вхідні дані на площині нейронів.

Унікальність методу карт, що самоорганизуються, полягає в перетворенні n–вимірного простору в двовимірний. Застосування двовимірних сіток пов’язане з тим, що існує проблема відображення просторових структур більшої розмірності.

Маючи таке представлення даних, можна візуально визначити наявність або відсутність взаємозв’язку у вхідних даних.

Нейрони карти Кохонена розташовують у вигляді двомірної матриці, розфарбовують цю матрицю залежно від аналізованих параметрів нейронів.

На рис. 2.46. наведений приклад карти Кохонена.

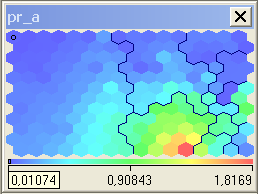


Рис. 2.46. Приклад карти Кохонена

Що ж означає її розфарбування? На [рис. 2.47.](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.3) наведене розфарбування карти, а точніше, її i-ї ознаки (показника pr a), у тривимірному представленні. Як ми бачимо, темно–сині ділянки на карті відповідають найменшим значенням показника, червоні – найвищим.

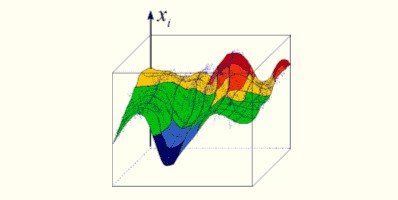


Рис. 2.47. Розфарбування *i*-ї ознаки в тривимірному просторі

Тепер, вертаючись до [рис. 2.46](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.2)., ми можемо сказати, які об’єкти мають найбільші значення розглянутого показника (група об’єктів, позначена червоним кольором), а які – найменші значення (група об’єктів, позначена синім кольором).

Таким чином, карти Кохонена (як і географічні карти) можна відображати:

* у двомірному вигляді, тоді карта розфарбовується відповідно до рівня виходу нейрона;
* у тривимірному вигляді.

У результаті роботи алгоритму одержуємо такі карти:

* карта входів нейронів;
* карта виходів нейронів;
* спеціальні карти.

Координати кожної карти визначають положення одного нейрона. Так, координати визначають нейрон, який перебуває на перетинанні 15-го стовпця з 30-м поруч у матриці нейронів. Розглянемо, що ж являють собою ці карти.

**Карта входів та виходів нейронів**

Карта входів нейронів. Ваги нейронів підбудовуються під значення вхідних змінних і відображають їхню внутрішню структуру. Для кожного входу малюється своя карта, розфарбована у відповідності зі значенням конкретної ваги нейрона.

При аналізі даних використовують кілька карт входів.

На одній з карт виділяють область певного кольору – це означає, що відповідні вхідні приклади мають приблизно однакове значення відповідного входу. Колірний розподіл нейронів із цієї області аналізується на інших картах для визначення схожих або відмітних характеристик. Приклад розглянутих карт входів буде наведений нижче.

Карта виходів нейронів. На карту виходів нейронів проєктується взаємне розташування досліджуваних вхідних даних. Нейрони з однаковими значеннями виходів утворюють кластери – замкнені області на карті, які включають нейрони з однаковими значеннями виходів.

Спеціальні карти. Це карта кластерів, матриця відстаней, матриця щільності потрапляння й інші карти, які характеризують кластери, отримані в результаті навчання мережі Кохонена.

Важливо розуміти, що між усіма розглянутими картами існує взаємозв’язок – усі вони є різними розфарбуваннями тих самих нейронів. Кожний приклад з навчальної вибірки має те саме розташування на всіх картах.

**Приклад розв’язку задачі.** Програмне забезпечення, що дозволяє працювати з картами Кохонена, зараз представлене великою кількістю інструментів. Це можуть бути як інструменти, що включають тільки реалізацію методу карт, що самоорганізуються, так і нейропакети з цілим набором структур нейронних мереж, серед котрих – і карти Кохонена; також даний метод реалізований у деяких універсальних інструментах.

До інструментарію, що включає реалізацію методу карт Кохонена, відносяться Somine, Statistica, Neuroshell, Neuroscalp, Deductor і багато інших. Для розв’язку задачі будемо використовувати аналітичний пакет Deductor.

Нехай є база даних комерційних банків з показниками діяльності за поточний період. Необхідно провести їх кластеризацію, тобто виділити однорідні групи банків на основі показників з бази даних, усього показників – 21.

Вихідна таблиця перебуває у файлі «banks.xls». Вона містить показники діяльності комерційних банків за звітний період.

Спочатку імпортуємо дані з xls-файлу в середовище аналітичного пакета.

На першому кроці майстра запускаємо майстер обробки й вибираємо зі списку метод обробки «Карта Кохонена». Далі слід настроїти призначення стовпців, тобто для кожного стовпця вибрати одне із призначень: вхідне, вихідне, не використовується й інформаційне. Вкажемо всім стовпцям, відповідним до показників діяльності банків, призначення «Вхідний». «Вихідний» не призначаємо.

Наступний крок пропонує розбити вихідну множину на навчальну, тестову й валідаційну. За замовчуванням,програма пропонує розбити множину на навчальну – 95% і тестову – 5%.

На кроці № 5, зображеному на [рис. 2.48.](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.4) пропонується настроїти параметри карти: кількість гнізд по Х і по Y, їх форму (шестикутну або чотирикутну).

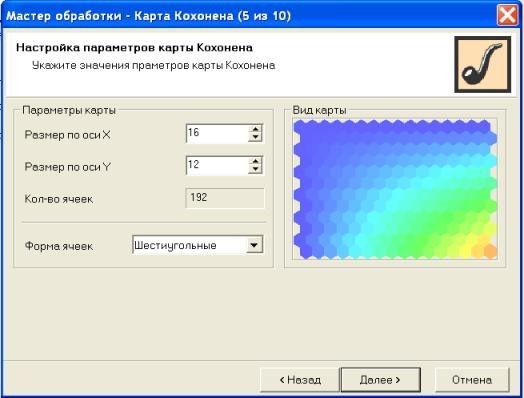


Рис. 2.48. Крок № 5 «Настроювання параметрів карти Кохонена»

На шостому кроці «Настроювання параметрів зупинки навчання», проілюстрованому на [рис. 2.49.](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.5), установлюємо параметри зупинки навчання й установлюємо епоху, по досягненню якої навчання буде припинено.

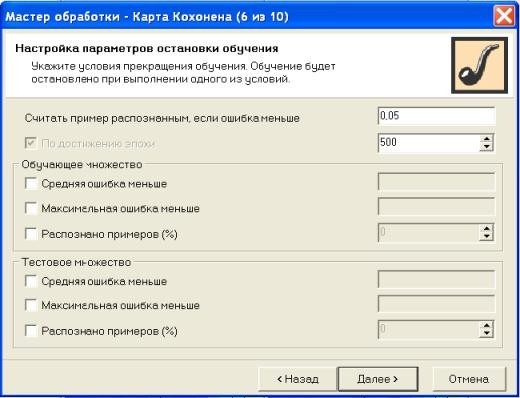


Рис. 2.50. Крок № 6 «Настроювання параметрів зупинки навчання»

На сьомому кроці, представленому на [рис. 2.51.](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.6), настроюються інші параметри навчання: спосіб початкової ініціалізації, тип функції сусідства. Можливі два варіанти кластеризації: автоматичне визначення числа кластерів з відповідним рівнем значимості й фіксована кількість кластерів (визначається користувачем). Оскільки нам невідома кількість кластерів, виберемо автоматичне визначення їх кількості.

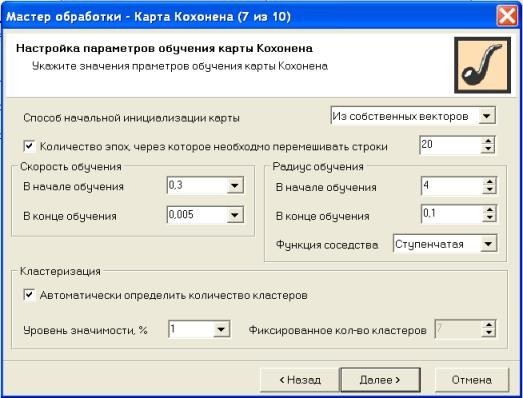


Рис. 2.51. Крок № 7 «Настроювання параметрів зупинки навчання»

На восьмому кроці запускаємо процес навчання мережі – необхідно натиснути на кнопку «Пуск» і дочекатися закінчення процесу навчання. Під час навчання можемо спостерігати зміну кількості розпізнаних прикладів і поточні значення помилок.

По закінченню навчання в списку візуалізаторів виберемо «Карту Кохонена» і візуалізатор Що–Якщо». На останньому кроці будуємо відображення карти Кохонена, цей крок проілюстрований на [рис. 2.52](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.7).

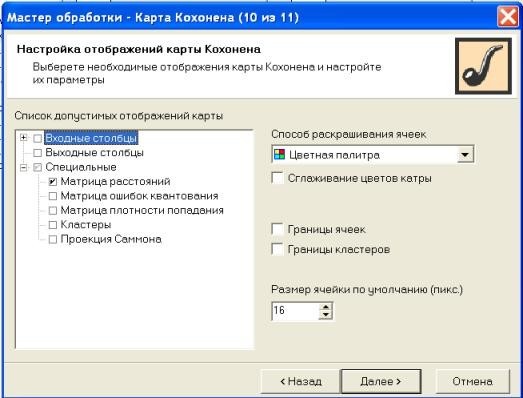


Рис. 2.52. «Крок № 10 Настроювання відображень карти Кохонена»

Укажемо відображення всіх вхідних, вихідних стовпців, кластерів, а також поставимо прапорець «Границі кластерів» для чіткого відображення границь.

***Карти входів.*** При аналізі карт входів рекомендують використовувати відразу кілька карт. Досліджуємо фрагмент карти, що складається з карт трьох входів, який наведений на [рис. 2.53](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.8).

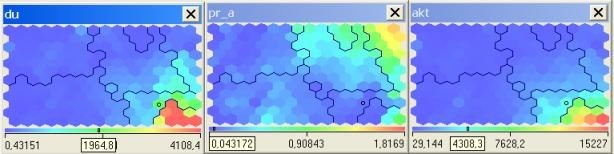


Рис. 2.53. Карти трьох входів

На одній з карт виділяємо область із найбільшими значеннями показника. Далі має сенс вивчити ці ж нейрони на інших картах.

На першій карті найбільші значення мають об’єкти, розташовані в правому нижньому куті. Розглядаючи одночасно три карти, ми можемо сказати, що ці ж об’єкти мають найбільші значення показника, зображеного на третій карті. Також по розфарбуванню першої й третьої карти можна зробити висновок, що існує взаємозв’язок між цими показниками.

Також ми можемо визначити, наприклад, таку характеристику: кластер, розташований у правому верхньому куті, характеризується низькими значеннями показників du (депозити юридичних осіб) і akt (активи банку) і високими значеннями показників pr\_a (прибутковість активів).

Ця інформація дозволяє наступним чином охарактеризувати кластер, що перебуває в правому верхньому куті: це банки з невеликими активами, невеликими притягнутими депозитними коштами від юридичних осіб, але з найбільш прибутковими активами, тобто це група невеликих, але найбільш прибуткових банків.

Це лише фрагмент висновку, який можна зробити, досліджуючи карту.

На наступному малюнку ([рис. 2.54](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.9).) наведена ілюстрація карт входів і виходів, остання – це карта кластерів. Тут ми бачимо кілька карт входів (показників діяльності банків) і сформовані кластери, кожний з яких виділений окремим кольором.

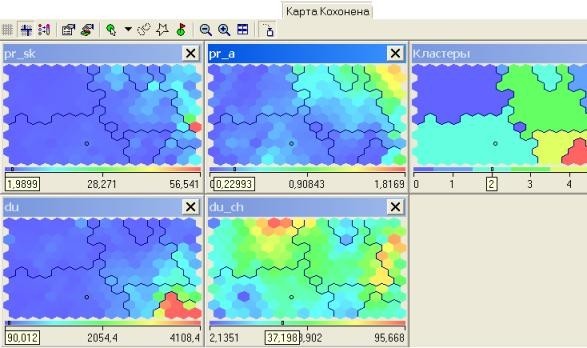


Рис. 2.54. Карти входів і виходів

Для знаходження конкретного об’єкта на карті необхідно натиснути правою кнопкою миші на досліджуваному об’єкті й вибрати пункт «Знайти гніздо на карті». Виконання цієї процедури показане на [рис. 2.55.](http://www.intuit.ru/department/database/datamining/12/datamining_12.html#image.12.10). У результаті ми можемо бачити як сам об’єкт, так і значення того виміру, який ми переглядаємо. Таким чином, ми можемо оцінити положення аналізованого об’єкта, а також зрівняти його з іншими об’єктами.

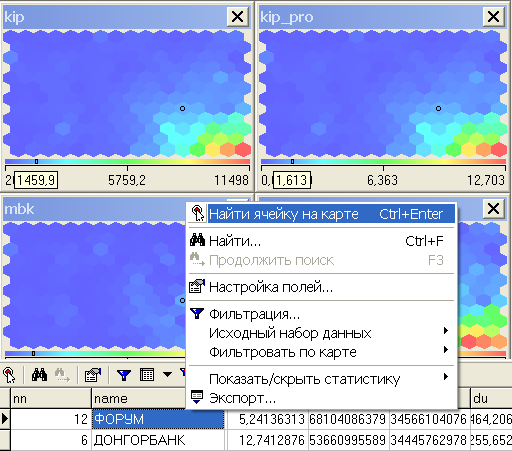


Рис. 2.55. Гніздо на карті

У результаті застосування карт, що самоорганізуються, багатовимірний простір вхідних факторів був представлений в двохвимірному вигляді, в якому його достатньо добре аналізувати.

Банки були класифіковані на 7 груп, для кожної з яких можливе визначення конкретних характеристик, виходячи з розфарбування відповідних показників.

***6. МЕТОДИ ПОШУКУ АСОЦІАТИВНИХ ПРАВИЛ***

**Що таке асоціативні правила?**

Асоціація – одна із задач Data Mining. Метою пошуку асоціативних правил (association rule) є знаходження закономірностей між зв’язаними подіями в базах даних.

Дуже часто покупці купують не один товар, а декілька. У більшості випадків між цими товарами існує взаємозв’язок. Так, наприклад, покупець, що купує ноутбук, швидше за все, захоче придбати також сумку. Ця інформація може бути використана для розміщення товару на прилавках.

Асоціативні правила, часто знаходять застосування:

* аналіз Web-блогів;
* у роздрібній торгівлі: визначення товарів, які варто просувати спільно; вибір місця розташування товару в магазині; аналіз споживчого кошика; прогнозування попиту;
* перехресні продажі: якщо є інформація про те, що клієнти придбали продукти A, Б і В, те які з них найімовірніше куплять продукт Г;
* маркетинг: пошук ринкових сегментів, тенденцій купівельної поведінки;
* сегментація клієнтів: виявлення загальних характеристик клієнтів компанії, виявлення груп покупців;
* оформлення каталогів, аналіз збутових кампаній фірми, визначення послідовностей покупок клієнтів (яка покупка піде за покупкою товару А);

Приведемо простий приклад асоціативного правила: покупець, що купує ноутбук, придбає до нього мишку з імовірністю 50%.

Введення в асоціативні правила. Уперше задача пошуку асоціативних правил (association rule mining) була запропонована для знаходження типових шаблонів покупок, здійснених у супермаркетах, тому іноді її ще називають аналізом ринкового кошика (market basket analysis).

Ринковий кошик – це набір товарів, придбаних покупцем у рамках однієї окремо взятої транзакції.

Транзакції є досить характерними операціями, ними, наприклад, можуть описуватися результати відвідувань різних магазинів.

Транзакція – це множина подій, які відбулися одночасно.

Реєструючи всі бізнес–операції протягом усього часу своєї діяльності, торговельні компанії накопичують величезні кількості транзакцій. Кожна така транзакція являє собою набір товарів, куплених покупцем за один візит.

Отримані в результаті аналізу шаблони включають перелік товарів і число транзакцій, які містять дані набори.

Транзакційна або операційна база даних (Transaction database) являє собою двовимірну таблицю, яка складається з номера транзакції (TID) і переліку покупок, придбаних під час цієї транзакції.

TID – унікальний ідентифікатор, що визначає кожну угоду або транзакцію.

*Приклад* транзакційної бази даних, що складається з купівельних транзакцій, наведл8ено в [табл. 2.12.](http://www.intuit.ru/department/database/datamining/15/datamining_15.html#table.15.1) У таблиці перший стовпчик (TID) визначає номер транзакції, у другому стовпчику таблиці наведені товари, придбані під час певної транзакції.

Таблиця 2.12.

Транзакційна база даних

|  |  |
| --- | --- |
| TID | Покупки |
| 100 | Карта пам’яті, DVD-диск, USB-подовжувач |
| 200 | DVD-диск, USB-подовжувач, WEB-камера |
| 300 | Комп’ютерна миша, DVD-диск, WEB-камера |
| 400 | USB-подовжувач, DVD-диск, Карта пам’яті, Комп’ютерна миша |
| 500 | DVD-диск, Карта пам’яті |
| 600 | VGA-кабель |

На основі наявної бази даних нам потрібно знайти закономірності між подіями, тобто покупками.

Шаблони, що часто зустрічаються, або зразки. Допустимо, є транзакційна база даних D. Присвоємо значенням товарів змінні ([табл. 2.13.](http://www.intuit.ru/department/database/datamining/15/datamining_15.html#table.15.2)).

Карта пам’яті = a ; DVD–диск = b ; USB–подовжувач = c ; Комп’ютерна миша = d ;

WEB– камера = e ; VGA– кабель = f

Таблиця 2.13.

Набори товарів, що часто зустрічаються

|  |  |  |  |
| --- | --- | --- | --- |
| TID | Покупки | TID | Покупки |
| 100 | Флешка, DVD-диск, USB-подовжувач | 100 | a, b, c |
| 200 | DVD-диск, USB-подовжувач, WEB-камера | 200 | b, c, e |
| 300 | Комп’ютерна миша, DVD-диск, WEB-камера | 300 | d, b, e |
| 400 | USB-подовжувач, DVD-диск, Флешка, Комп’ютерна миша | 400 | c, b, a, d |
| 500 | DVD-диск, Флешка, USB-подовжувач | 500 | b, a, c |
| 600 | VGA-кабель | 600 | f |

Розглянемо набір товарів (Itemset), що включає, наприклад, (флешка, DVD-диск, USB-подовжувач). Виразимо цей набір за допомогою змінних:

abc={a,b,c}

Підтримка. Цей набір товарів зустрічається в нашій базі даних три рази, тобто підтримка цього набору товарів рівна 3:

SUP(abc)=3.

При мінімальному рівні підтримки, рівному трьом, набір товарів abc є шаблоном, що часто зустрічається.

min\_sup=3, {Карта пам’яті, DVD– диск, USB– подовжувач} – частий шаблон, що зустрічається.

Підтримкою називають кількість або відсоток транзакцій, що містять певний набір даних.

Для даного набору товарів підтримка, виражена у відсотковому відношенні, рівна 50%.

SUP(abc)=(3/6)\*100%=50%

Підтримку іноді також називають забезпеченням набору.

Таким чином, набір становить інтерес, якщо його підтримка вище заданого користувачем мінімального значення (min support). Ці набори називають такими, що часто зустрічаються (frequent).

**Алгоритми пошуку асоціативних правил**

**Характеристики асоціативних правил.** Асоціативне правило має вигляд: «З події A випливає подія B». У результаті такого виду аналізу ми встановлюємо закономірність наступного виду: «Якщо в транзакції зустрівся набір товарів (або набір елементів) A, то можна зробити висновок, що в цій же транзакції повинен з’явитися набір елементів B)» Встановлення таких закономірностей дає нам можливість знаходити дуже прості й зрозумілі правила, називані асоціативними.

Основними характеристиками асоціативного правила є підтримка й вірогідність правила.

Розглянемо правило «з покупки флешки випливає покупка USB-подовжувача» для бази даних, яка була наведена вище в табл. 2.13.. Поняття підтримки набору ми вже розглянули. Існує поняття підтримки правила.

Правило має підтримку s, якщо s% транзакцій із усього набору містять одночасно набори елементів A і B або, інакше кажучи, містять обидва товари.

Флешка – це товар A, USB-подовжувач – це товар B. Підтримка правила «з покупки флешки випливає покупка USB-подовжувача» рівна 3, або 50%.

Вірогідність правила показує, яка ймовірність того, що з події A випливає подія B.

Правило «З A випливає B» справедливе з вірогідністю С, якщо c% транзакцій із усієї множини, що містить набір елементів A, також містять набір елементів B.

Якщо число транзакцій, що містять USB-подовжувач, рівне чотирьом, а число транзакцій, що містять також і флешку, рівне трьом, то вірогідність правила рівна (3/4)\*100%, тобто 75%.

Вірогідність правила «з покупки USB-подовжувача випливає покупка флешки рівна 75%, тобто 75% транзакцій, що містять товар А, також містять товар B.

**Границі підтримки й вірогідності асоціативного правила.** За допомогою використання алгоритмів пошуку асоціативних правил аналітик може одержати всі можливі правила виду « З A випливає B», з різними значеннями підтримки й вірогідності. Однак у більшості випадків, кількість правил необхідно обмежувати заздалегідь установленими мінімальними й максимальними значеннями підтримки й вірогідності.

Якщо значення підтримки правила занадто велике, то в результаті роботи алгоритму будуть знайдені правила очевидні й добре відомі. Занадто низьке значення підтримки приведе до знаходження дуже великої кількості правил, які, можливо, будуть у більшій частині необґрунтованими, але не відомими й не очевидними для аналітика. Таким чином, необхідно визначити такий інтервал, «золоту середину», який з однієї сторони забезпечить знаходження неочевидних правил, а з іншого – їх обґрунтованість.

Якщо рівень вірогідності занадто малий, то цінність правила викликає серйозні сумніви. Наприклад, правило з вірогідністю в 3% тільки умовно можна назвати правилом.

**Методи пошуку асоціативних правил**

**Алгоритм AIS**. Перший алгоритм пошуку асоціативних правил, що називався AIS, (запропонований Agrawal, Imielinski and Swami) був розроблений співробітниками дослідного центру IBM Almaden в 1993 році. Із цієї роботи почався інтерес до асоціативних правил; на середину 90-х років минулого століття припадає пік дослідницьких робіт у цій області, і з тих пор щороку з’являється кілька нових алгоритмів.

В алгоритмі AIS кандидати множини наборів генеруються й підраховуються «на льоту», під час сканування бази даних.

Алгоритм SETM. Створення цього алгоритму було мотивовано бажанням використовувати мову SQL для обчислення наборів товарів, що часто зустрічаються. Як і алгоритм AIS, SETM також формує кандидатів «на льоту», ґрунтуючись на перетвореннях бази даних. Щоб використовувати стандартну операцію об’єднання мови SQL для формування кандидата, SETM відокремлює формування кандидата від їхнього підрахунку.

Незручність алгоритмів AIS і SETM – надмірне генерування й підрахунок занадто багатьох кандидатів, які в результаті не є такими, що часто зустрічаються. Для поліпшення їх роботи був запропонований алгоритм Apriori.

Робота даного алгоритму складається з декількох етапів, кожний з етапів складається з наступних кроків:

* формування кандидатів;
* підрахунок кандидатів.

Формування кандидатів (candidate generation) – етап, на якому алгоритм, скануючи базу даних, створює множину *i*–елементних кандидатів (i – номер етапу). На цьому етапі підтримка кандидатів не розраховується.

Підрахунок кандидатів (candidate counting) – етап, на якому обчислюється підтримка кожного *i*–елементного кандидата. Тут же здійснюється відсікання кандидатів, підтримка яких менша мінімуму, установленого користувачем (min\_sup).

Решту *i*-елементних наборів називаємо такими, що часто зустрічаються.

Розглянемо роботу алгоритму Apriori на прикладі бази даних D. Ілюстрація роботи алгоритму наведена на [рис. 2.56](http://www.intuit.ru/department/database/datamining/15/datamining_15.html#image.15.1). Мінімальний рівень підтримки рівний 3.

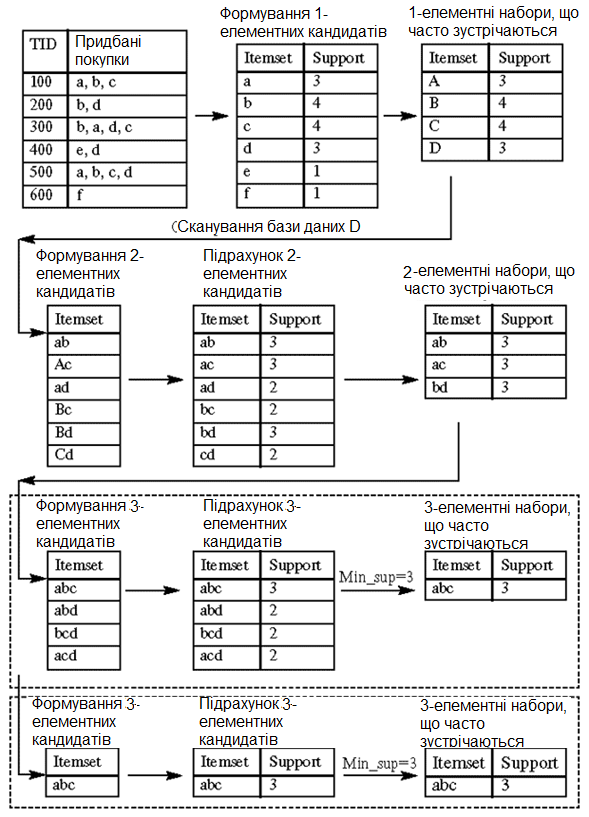


Рис. 2.56. Алгоритм Apriori

На першому етапі відбувається формування одноелементних кандидатів. Далі алгоритм підраховує підтримку одноелементних наборів. Набори з рівнем підтримки менше встановленого, тобто 3, відкидаються. У нашому прикладі це набори e і f, які мають підтримку, рівну 1. набори товарів, що залишилися, вважаються одноелементними наборами, що часто зустрічаються, товарів: це набори a, b, c, d.

Далі відбувається формування двоелементних кандидатів, підрахунок їх підтримки й відсікання наборів з рівнем підтримки, меншим 3. двоелементні набори товарів, що залишилися, вважаються двоелементними наборами, що часто зустрічаються, ab, ac, bd, беруть участь у подальшій роботі алгоритму.

Якщо дивитися на роботу алгоритму прямолінійно, на останньому етапі алгоритм формує трьохелементні набори товарів: abc, abd, bcd, acd, підраховує їхню підтримку й відтинає набори з рівнем підтримки, меншим 3. Набір товарів abc може бути названий таким, що часто зустрічається.

Однак алгоритм Apriori зменшує кількість кандидатів, відтинаючи – аprior – тих, які свідомо не можуть стати такими, що часто зустрічаються, на основі інформації про відсічених кандидатів на попередніх етапах роботи алгоритму.

Відсікання кандидатів відбувається на основі припущення про те, що в набору, що часто зустрічається, товарів усі підмножини повинні бути такими, що часто зустрічаються. Якщо в наборі перебуває підмножина, яку на попередньому етапі було визначено такою, що нечасто зустрічається, цей кандидат уже не включається у формування й підрахунок кандидатів.

Так набори товарів ad, bc, cd були відкинуті як такі, що нечасто зустрічаються, алгоритм не розглядав товарів abd, bcd, acd.

При розгляді цих наборів формування трьохелементних кандидатів відбувалося б за схемою, наведеною у верхньому пунктирному прямокутнику. Оскільки алгоритм апріорі відкинув набори, що свідомо нечасто зустрічаються, останній етап алгоритму відразу визначив набір abc як єдиний трьохелементний набір, що часто зустрічається (етап наведений у нижньому пунктирному прямокутнику).

Алгоритм Apriori розраховує також підтримку наборів, які не можуть бути відсічені апріорі. Це так звана негативна область (negative border), до неї належать набори-кандидати, які зустрічаються рідко, їх самих не можна віднести до таких, що часто зустрічаються, але всі підмножини даних наборів є такими, що часто зустрічаються.

**Різновиди алгоритму Apriori.** Залежно від розміру найдовшого набору, що часто зустрічається, алгоритм Apriori сканує базу даних певну кількість разів. Різновиди алгоритму Apriori, що є його оптимізацією, запропоновані для скорочення кількості сканувань бази даних, кількості наборів–кандидатів або того й іншого. Були запропоновані наступні різновиди алгоритму Apriori: Aprioritid і Apriorihybrid.

**Aprioritid.** Цікава особливість цього алгоритму – те, що база даних D не використовується для підрахунку підтримки кандидатів набору товарів після першого проходу.

Із цією метою використовується кодування кандидатів, виконане на попередніх проходах. У наступних проходах розмір закодованих наборів може бути набагато меншим, ніж база даних, і в такий спосіб заощаджуються значні ресурси.

**Apriorihybrid.** Аналіз часу роботи алгоритмів Apriori і Aprioritid показує, що в більш ранніх проходах Apriori досягає більшого успіху, ніж Aprioritid; однак Aprioritid працює краще Apriori у більш пізніх проходах. Крім того, вони використовують ту саму процедуру формування наборів-кандидатів. Заснований на цьому спостереженні, алгоритм Apriorihybrid запропонований, щоб об’єднати кращі властивості алгоритмів Apriori і Aprioritid. Apriorihybrid використовує алгоритм Apriori у початкових проходах і переходить до алгоритму Aprioritid, коли очікується, що закодований набір первісної множини наприкінці проходу буде відповідати можливостям пам’яті. Однак, перемикання від Apriori до Aprioritid вимагає залучення додаткових ресурсів.

Один з них – алгоритм DHP, також називаний алгоритмом хешування (J. Park, M. Chen and P. Yu, 1995 рік). В основі його роботи – імовірнісний підрахунок наборів–кандидатів, здійснюваний для скорочення числа підраховуваних кандидатів на кожному етапі виконання алгоритму Apriori. Скорочення забезпечується за рахунок того, що кожний з k-елементних наборів-кандидатів крім кроку скорочення проходить крок хешування. В алгоритмі на k-1 етапі під час вибору кандидата створюється так звана хеш-таблиця. Кожний запис хеш-таблиці є лічильником усіх підтримок k-елементних наборів, які відповідають цьому запису в хеш-таблиці. Алгоритм використовує цю інформацію на етапі k для скорочення множини k-елементних наборів-кандидатів. Після скорочення підмножини, як це відбувається в Apriori, алгоритм може вилучити набір-кандидат, якщо його значення в хеш-таблиці менше граничного значення, установленого для забезпечення.

До інших удосконалених алгоритмів відносяться: PARTITION, DIC, алгоритм «вибіркового аналізу».

PARTITION алгоритм (A. Savasere, E. Omiecinski and S. Navathe, 1995 рік). Цей алгоритм розбивки (поділу) полягає в скануванні транзакційної бази даних шляхом поділу її на розділи, які не перетинаються, кожний з яких може вміститися в оперативній пам’яті. На першому кроці в кожному з розділів за допомогою алгоритму Apriori визначаються «локальні» набори даних, що часто зустрічаються. На другому підраховується підтримка кожного такого набору щодо всієї бази даних. Таким чином, на другому етапі визначається множина усіх потенційних наборів даних, що зустрічаються.

Алгоритм DIC, Dynamic Itemset Counting (S. Brin R. Motwani, J. Ullman and S. Tsur, 1997 рік). Алгоритм розбиває базу даних на кілька блоків, кожний з яких відзначається так званими «початковими точками» (start point), і потім циклічно сканує базу даних.

**Змістовий модуль 2. Інформаційне підґрунтя розробки бізнес -моделі**

***Лекція 3.1. (2 год)***

**Тема 3. Аналітичне дослідження ринку**

**ІНФОРМАЦІЙНІ АСПЕКТИ ДОСЛІДЖЕННЯ РИНКУ**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Сутність, види та принципи ринкової інформації.

2. Носії та джерела ринкової інформації.

3. Методи збору ринкової інформації (первинної та вторинної).

***1. СУТНІСТЬ, ВИДИ ТА ПРИНЦИПИ РИНКОВОЇ ІНФОРМАЦІЇ***

***Інформація*** – форма спілкування, засіб отримання та передачі знань.

***Ринкова інформація*** являє собою сукупність цифр, фактів, відомостей, чуток, оцінок та інших даних, необхідних для аналізу й прогнозування ринкової діяльності.

У дослідженнях ринку формується інформаційне середовище, яке включає:

- інформаційні ресурси, тобто сукупність різних документів і окремих масивів документів та інших інформаційних систем (бібліотек, архівів, фондів, банків даних тощо), пов’язаних з ринковою діяльністю;

- інформаційну інфраструктуру, тобто сукупності інформаційних каналів і сховищ, а також інформаційних технологій – системи та методології формування, обробки, накопичення, пошуку, зберігання та використання інформації з метою дослідження ринку.

Важливим елементом інформаційної інфраструктури є інформаційна технологія. Вона поєднує різні методи збору, аналізу, зберігання та обробки інформації, зокрема статистичне і інше спостереження, банки і бази даних.

До інфраструктури досліджень ринку входять також інформаційні брокери, зайняті наданням стандартизованих інформаційних послуг. Характеристики, отримані в результаті маркетингового аналізу і прогнозу, також використовують інформаційну технологію.

**Структура ринкової інформації дозволяє виділяти й розрізняти кілька її видів, кожен із яких виконує певні функції**:

- ***факт***, що представляє собою найпростіший вид інформації: подія або умова, безпосередньо спостережувані;

- ***відомості***, що є різновидом фактів, представлених в систематизованій, узагальненій формі;

- ***чутки***, до яких відносяться непідтверджені, неперевірені факти;

- ***оцінки***, які включають інформацію, що базується на умовиводах, підрахунках і статистичних розрахунках (формою оцінки є також прогноз, тобто наукове передбачення);

- ***цифри***, що становлять форму відображення кількісної інформації.

Успішне функціонування будь-якої фірми у ринковому середовищі може бути забезпечене лише за наявності точної, повної і достовірної інформації. Хороша інформація дозволяє:

- отримувати конкурентні переваги;

- знижувати фінансовий ризик;

- визначати відношення споживачів;

- стежити за зовнішнім середовищем;

- координувати стратегію;

- оцінювати свою діяльність;

- підвищувати довір’я до реклами;

- отримувати підтримку в рішеннях.

Самостійною частиною інформаційної сукупності є статистична інформація, тобто сукупність цифрових відомостей, формування якої та інтерпретація підпорядковуються певним правилам і закономірностям. Значна частина кількісної інформації належить до категорії статистичних величин. З них виділяють:

- ***абсолютні*** дані, що дозволяють визначити розмірність, величину досліджуваного явища або процесу (кількість, товарна маса, обсяг продажів, число працівників і т. д.);

- ***відносні*** дані, що відображають співвідношення величин, результати їх порівняння (динаміка, структура, інтенсивність, координація і т. д.);

- ***середні величини***, які відображають узагальнений рівень, абстрагований від індивідуальних особливостей окремих одиниць (дохід на одну людину, середній розмір поставки і т. п.).

***Ринкову інформаційну систему*** можна визначити як сукупність процедур і методів, розроблених для створення, аналізу та поширення інформації для попереджувальних маркетингових рішень на регулярній постійній основі.

Інформація, яку використовують для досліджень ринку повинна бути:

– високоякісною;

– достовірною ( мінімальна помилковість);

– повною;

– точною (відповідність реальним значенням стану справ);

– актуальною (свіжа інформація, не застаріла);

– цінною і корисною (відповідність меті, ситуації на ринку).

Саме така інформація допомагає фірмам визначати відношення споживачів до товару та фірми, постійно стежити за зовнішнім середовищем, координувати стратегію і оцінювати діяльність, підвищувати рівень рекламної роботи, отримувати підтримку в прийнятих рішеннях, підтверджувати власну комерційну інтуїцію, підвищувати ефективність діяльності.

***Види ринкової інформації*** необхідно розглядати в розрізі таких ознак її класифікації:

1. За призначенням:

- вихідні дані – для конкретизації (визначення) проблеми та її вирішення;

- контрольні дані – для оцінки ефективності вирішення проблеми.

2. Залежно від рівня:

- макропланові дані – інформація про державну політику економічного регулювання фінансової системи, цін, позик, технологій тощо;

- мікропланові дані – інформація про величину попиту та пропозиції, споживачів, постачальників й ін.

3. Залежно від власності:

- власність підприємства (зібрана фірмою особисто);

- власність інших підприємств чи держави.

4. Залежно від міри відкритості:

- відкрита;

- приватна;

- секретна.

5. Залежно від ролі в діяльності підприємства:

- стратегічна;

- тактична;

- оперативна.

6. Залежно від змісту:

- ідеї, гіпотези, поняття;

- методи, підходи, методики;

- фактаж (статистика).

7. Залежно від джерел опитування:

- внутрішня (звіти підприємства, списки покупців, постачальників, посередників, конкурентів);

- зовнішня (періодика, дані опитувань, звіти посередників).

8. Залежно від потреб підприємства – інформація про:

- навколишнє середовище;

- наявні можливості впливу на ринок;

- існуючі обмеження впливу маркетингового інструментарію в різних умовах навколишнього середовища.

9. Залежно від аспектів маркетингової діяльності підприємства інформація щодо:

- попиту (що користується попитом, коли він є, хто його носії, яка їх поведінка, де є попит?);

- пропозиції (які продукти пропонуються, в яких обсягах, коли, де, хто пропонує?);

- якою мірою і в яких умовах здійснювалось урівноваження пропозиції та попиту;

- стану ринку (ринковий потенціал, реальний обсяг ринку, частки ринку);

- споживачів (характерні ознаки, сегменти, мотиви і т. д.);

- цін (рівень, динаміка, еластичність);

- конкуренції (кількість і сила конкурентів, їх цілі, стратегія, поведінка);

- макросистем (економічна, податкова, фінансова політика, система розподілу доходів, зовнішньоекономічні відносини, закони, норми та правила господарювання тощо).

10. Залежно від часу одержання:

- вторинна інформація (зібрана раніше для вирішення якихось інших проблем);

- первинна (дані, що збираються персонально спеціально для розв’язання поставленої проблеми).

Інформація, по суті, є одним з найбільш цінних ринкових продуктів. Цінність маркетингової інформації полягає в тому, що вона створює передумови одержання конкурентних переваг:

- допомагає зменшити міру ризику;

- визначає та попереджує про зміни в навколишньому середовищі;

- сприяє формуванню і координації стратегій;

- підтримує та обґрунтовує рішення;

- сприяє зростанню іміджу фірми;

- дає можливість аналізувати діяльність фірми з метою підвищення її ефективності

Види ринкової інформації залежно від джерел опитування відображені на рис. 3.1.

Види ринкової інформації залежно від джерел опитування

Внутрішня інформація

Первинна

Вторинна

Кабінетні дослідження

Польові дослідження

Зовнішня інформація

Офіційна

Довідкова

Огляд і листування

Чутки

Наукова та освітня

ЗМІ

Рекламна

Рис. 3.1. Види ринкової інформації залежно від джерел опитування

***Ринкова інформація базується на таких принципах***:

- реальність відомостей, що подаються в потрібний момент часу, коли це необхідно для конкретного маркетингового дослідження та отримання відповідних висновків;

- адекватність відомостей, що забезпечується дотриманням наукових принципів збору та обробки інформації, боротьбою з будь-якою тенденційністю;

- релевантність інформації (від англ. Relevant – доречний, що стосується справи), яка забезпечується отриманням інформації відповідно до поставлених завдань маркетингового дослідження;

- повнота інформації: планування дослідження, виявлення сутності досліджуваного явища або процесу, структурування дослідження виявлення і моделювання внутрішніх і зовнішніх зв’язків;

- відповідність інформації, що збирається і обробляється генеральній меті дослідження; - підпорядкування, формування та використання методології збору та обробки даних вимогам теорії інформатики.

***2. НОСІЇ І ДЖЕРЕЛА РИНКОВОЇ ІНФОРМАЦІЇ***

Інформація збирається серед її носіїв, тобто власників інформації, тобто осіб, які мають доступ до якихось відомостей та займаються певною ринковою діяльністю (як активною, так і пасивною). Деякі носії мають ексклюзивне право власності на інформацію, інші купують її або здобувають її будь-яким способом.

Класифікація інформації за типом носіїв відображена на рис. 3.2.

Особи, які збирають, накопичують і використовують інформацію, мають назву ***користувачів***.

Поняття носія інформації відрізняється від джерела інформації.

**Джерело інформації** – це концентрований і спеціалізований провідник різноманітних відомостей про ринкову діяльність, канал ринкової інформації, звідки її запозичують зацікавлені користувачі.

Носії інформації

Населення

Підприємство

Установи

Продавці

Покупці

Виробничі послуги

Торговельні послуги

Інше

Державні та муніципальні

Наукові та освітні

Суспільні

Рис. 3.2. Класифікація інформації за типами носіїв

У дослідженнях ринку сформувалися два канали добування інформації від її носія і користувача:

- носій інформації сам повідомляє користувачеві необхідні відомості;

- користувач отримує інформацію від інших джерел (безоплатно або в порядку купівлі або обміну).

Сформувалася самостійна класифікація джерел ринкової інформації. До них належать такі ***види інформації***:

- публікації: цифрові, фактографічні, теоретичні, графічні;

- інформація, що купується фірмою на комерційних засадах;

- обмін інформацією;

- дані спеціальних обстежень;

- інформація торгових кореспондентів;

- дані експертних оцінок;

- дані безпосереднього спостереження;

- дані включеного спостереження;

- дані опитування споживачів, підприємців, фахівців;

- дані панельних досліджень;

- дані експерименту (пробний маркетинг): польовий і лабораторний.

Особливе значення має класифікація інформації за її призначенням. Частина інформації становить об’єкт аналізу. Такий тип інформації називається аналітичним.

**Аналітична інформація** – дані, отримані в результаті характеристики ринкових процесів і явищ. До неї примикає рекомендаційна, або консалтингова, інформація, розроблена і запропонована спеціалізованими маркетинговими та консалтинговими фірмами.

**Консалтинг** – надання консультативних або маркетингових послуг іншим фірмам.

До аналітичної інформації можна віднести зібрані й розроблені фірмою або зовнішніми інформаційними службами бази даних, прогнози та сценарії.

**База даних** – інформація про існуючих або потенційних учасників ринкової діяльності зокрема покупців, продавців, конкурентів і т. п.

**Внутрішньофірмова звітність** – інформація, яку в установленому порядку і в установлені терміни подають підрозділи фірми.

**Статистична звітність** – інформація, що надається фірмою державі в обов’язковому порядку.

***3. МЕТОДИ ЗБИРАННЯ РИНКОВОЇ ІНФОРМАЦІЇ (ПЕРВИННОЇ ТА ВТОРИННОЇ)***

До методів збирання ринкової інформації (первинної та вторинної) відносять: метод кабінетних досліджень, опитування, спостереження, панельне обстеження, експеримент, методи аналізу інформації.

Для збирання ринкової інформації використовують два основні методи: ***метод кабінетних досліджень***, що передбачає збір вторинної інформації; ***метод польових досліджень***, за допомогою якого використовують первинну інформацію.

***Кабінетне дослідження*** – це сукупність методів збирання та оцінки ринкової інформації, що міститься в джерелах (статистичних даних або звітах), підготовлених для будь-яких інших цілей.

Кабінетні методи збирання інформації спираються на вторинні джерела, тому часто називаються методами роботи з документами. Документами можуть бути матеріали як вторинних (зовнішніх і внутрішніх) джерел, так і первинних документів: відповіді на відкриті питання анкет, матеріали фокус-груп і вільних інтерв’ю. Крім того, як документи розглядаються художні твори, наукові та монографічні видання, кіно-, відео-, аудіо-, фотоматеріали і т. п.

Кабінетні методи використовуються при підготовці польового дослідження (оскільки дозволяють попередньо ознайомитися з об’єктом дослідження, галуззю), при формуванні робочих гіпотез, при збиранні статистичної інформації для обґрунтування вибіркових процедур, при перевірці та інтерпретації інформації, отриманої за допомогою польових методів.

Кабінетні методи використовуються і як самостійні методи збору інформації при дослідженні ринку, вивченні традицій, динаміки громадської думки з якої-небудь проблеми, при дослідженні рекламних сюжетів і стимулів, спрямованих на активізацію споживачів.

***Методи позакабінетних (польових)*** досліджень припускають збір первинної інформації, одержуваної безпосередньо від споживача й покупця або іншого досліджуваного суб’єкта. Первинна інформація збирається згідно цілям дослідження. Збір первинної інформації пов’язаний зі значними витратами. Методи польових досліджень поділяються на опитування, спостереження, експеримент, тестування та імітацію.

***Опитування*** – це отримання первинної інформації від респондента шляхом виявлення думки або дій респондента.

Опитування передбачає систематичний збір інформації в опитуваних особисто, а також телефоном або поштою. Найбільш точний метод опитування, що зменшує ступінь невизначеності, – особисте інтерв’ювання. Однак це дорогий спосіб вивчення ринку, він займає багато часу й потребує високої кваліфікації інтерв’юера.

Опитування телефоном відносно недороге й обмежене в часі. Однак не в усіх покупців є телефон. Опитування поштою – найдешевший із зазначених методів. Але мінуси його – низький відсоток повернення анкет (якщо повернулося близько 12-14 % заповнених анкет, то це вважається успіхом) і затримка відповідей, а також участь не тих осіб, на які розраховував дослідник.

***Спостереження*** – науково організований збір й використання інформації, призначеної для задоволення інформаційно-аналітичних потреб дослідження. Методологія спостереження в ринковій сфері використовує і поєднує прийоми і методи маркетингового, статистичного та соціологічного спостереження.

Виділяють дві основні ***форми спостереження***, що використовуються в дослідженнях ринку:

- *безпосереднє (пряме)* спостереження: стеження за об’єктом спостереження (часто візуальне), яке може бути відкритим, коли об’єкт спостереження знає, що за ним спостерігають, і прихованим, коли досліджуваний об’єкт не знає, що він стає предметом спостереження;

- *непряме* спостереження: вивчення результатів будь-якої діяльності тощо. В непрямому спостереженні головне місце відводиться документальному способу спостереження.

В організації спостереження виділяються два види: за часом реєстрації фактів; за охопленням досліджуваних одиниць.

У свою чергу, в спостереженні, здійснюваному за часом реєстрації фактів, виділяють: безперервне (поточне), періодичне і одноразове спостереження. В поточному спостереженні фіксуються всі можливі зміни у міру їх виникнення. Спостереження може бути постійним, що проводиться впродовж тривалого часу, періодичним, що регулярно повторюється, й одноразовим, яке може бути разовим, або приуроченим до яких-небудь подій.

За ознакою охоплення одиниць досліджуваної сукупності спостереження ділиться на суцільне, що забезпечує інформацію про всі одиниці досліджуваної сукупності, й несуцільне (часткове), коли обстежується лише частина сукупності. Найбільш інформативною є вибіркове спостереження, або вибірка, коли спеціальні статистичні методи дають можливість з високим ступенем імовірності робити висновки за деякою частиою сукупності про всю сукупність.

Досить поширений метод основного масиву. Зазвичай при ньому обстежуються найбільш представницькі підприємства та організації, що мають найбільшу питому вагу в досліджуваній сукупності. Іноді використовується метод моментних спостережень, коли одиниці сукупності реєструються в заздалегідь визначені моменти часу (вибірка в часі). В дослідженнях ринку іноді відбираються і ретельно досліджуються найбільш типові підприємства або регіони. За ними можна робити висновки про закономірності стану й розвитку всього явища або процесу. Таке несуцільне спостереження називається монографічним.

***Спеціально організовані спостереження*** – інформація, зібрана і призначена для будь-якої певної мети в дослідженнях ринку.

***Експеримент*** (англ. experiment) – метод вивчення, що полягає у викликанні визначеного явища (чи його зміну) в штучно створених (лабораторних) умовах з метою дослідження і з’ясування процесу його розвитку.

Експеримент – спроба, дослід, що потребує підтвердження чи спростування, форма пізнання, один з основних методів наукового дослідження, в якому вивчення явищ відбувається в доцільно вибраних або штучно створених умовах, що забезпечують появу тих процесів, спостереження яких необхідне для встановлення закономірних зв’язків між явищами. Важливими характеристиками експерименту є його надійність та валідність.

Експеримент відрізняється від спостереження активною взаємодією з досліджуваним об’єктом. Зазвичай, експеримент проводиться в рамках наукового дослідження і слугує для перевірки гіпотези, встановлення причинно-наслідкових зв’язків між явищами.

***Експеримент*** – це не лише спосіб отримання інформації, він є активним елементом маркетингового менеджменту, цілеспрямованого впливу на ринок.

Розрізняють такі ***види експериментів***:

*Пасивний експеримент*. При пасивному експерименті інформація про досліджуваний об’єкт накопичується шляхом пасивного спостереження, тобто інформацію отримують в умовах звичайного функціонування об’єкта. Активний експеримент проводиться з застосуванням штучного впливу на об’єкт за спеціальною програмою. При пасивному експерименті існують лише фактори у вигляді вхідних контрольованих, але некерованих змінних, й експериментатор знаходиться в становищі пасивного спостерігача. Завдання планування в цьому разі зводиться до оптимальної організації збору інформації та вирішення таких питань, як вибір кількості та частоти вимірювань, вибір методу обробки результатів вимірювань.

*Активний експеримент*. Активний експеримент дозволяє швидше й ефективніше вирішувати завдання дослідження, але складніший, вимагає великих матеріальних витрат і може перешкодити нормальному ходу технологічного процесу. Іноді відсутня можливість проведення активного експерименту. Проте, враховуючи переваги активного експерименту, тоді, коли це можливо, перевагу віддають йому. При активному експерименті фактори повинні бути керованими і незалежними.

*Польовий експеримент* – оцінка та моделювання результатів деяких видів ринкової діяльності (наприклад, характеру продажу невеликих партій товарів, вивчення реакції споживачів на зміну факторів);

*Лабораторний* – побудова за даними польових спостережень імітаційної моделі.

Експерименти з однією змінною передбачають вивчення впливу зміни одного фактора на збутову, рекламну та іншу діяльність фірми.

Експерименти з декількома змінними передбачають вивчення залежності змін результатів діяльності фірми від дії декількох чинників одночасно. Такі експерименти більш складні і дозволяють вимірювати не лише вплив окремих факторів на ті чи інші процеси та явища на ринку, але взаємовплив і взаємодію цих факторів, і їх загальний кумулятивний підсумок впливу на ринкові механізми та попит.

*Тестування (виміри)* буває прямим (за реальних умов) і проміжним (у лабораторних умовах), попереднім і заключним. Попередні виміри здійснюються на стадії розроблення нового товару, нової стратегії, нового методу продажів тощо, в той час як заключні виміри здійснюються вже в реальних ринкових умовах, під час продажу розробленого товару, проведення в життя розробленої ринкової стратегії і т. п.

*Метод імітації* – це різновид експерименту в лабораторних умовах. Застосування цього методу стало можливим у зв’язку з розвитком електронно-обчислювальної техніки, оскільки він передбачає відтворення ситуації з використанням низки факторів маркетингової діяльності не в реальних умовах ринку, а з використанням програмного забезпечення. Спочатку будується модель контрольованих і неконтрольованих факторів, з якими стикається фірма. Потім їх різні поєднання закладаються в програму комп’ютера, щоб визначити вплив на загальну стратегію досліджень ринку. Для імітації не потрібно співпраці з боку споживачів, і вона може враховувати безліч взаємопов’язаних факторів.

*Розвідка в дослідженнях ринку* – діяльність по збиранню поточної інформації про зміну зовнішнього середовища ринку, необхідної для розроблення і коректування планів з досліджень ринку. Розвідка в дослідженнях ринку – інформація, що має пряме і безпосереднє відношення до ринку компанії, де зібрані й проаналізовані дані для прийняття рішення за важливою для компанії проблематики.

Дані, отримані в ході ринкової розвідки, дозволяють визначити показники розвитку даного ринку, ринкові можливості компанії, в кінцевому рахунку - стратегію її поведінки на ринку. Розвідка передбачає отримання інформації про навколишнє середовище організації з різних джерел легальним способом. Доцільно розрізняти імпліцитну (видобуту) і експліцитну (комунікативну) інформацію про навколишній світ. Цим різновидам інформації відповідають адекватні методи, інструменти збору та аналізу інформації.

Менеджери, що відповідають за ринкову розвідку, ставлять перед собою два головні завдання:

- оцінити ситуацію;

- побачити нові можливості для розвитку своєї організації (бізнесу).

Для вирішення цих цілей і завдань використовуються основні два ***види розвідки***:

- пасивну;

- активну.

***Пасивна розвідка***. Забезпечує організацію необхідними знаннями для здійснення об’єктивної оцінки. Наприклад, накопичення інформації про діяльність конкурентів.

***Наступальна, активна розвідка***. Забезпечує організації виявлення сприятливих шансів і можливостей. Ринкова розвідка не існує як самостійний вид діяльності або як певна функція організації. Ринкова розвідка є складовою частиною складного процесу прийняття рішення керівництвом у відповідь на виклики зовнішнього середовища. Її необхідно розглядати як один із етапів процесу прийняття рішень керівництвом організації.

Ринкова розвідка використовує переважно так звані, PEST- і SWOT-аналізи.

***Лекція 3.2. (2 год)***

**Тема 3. Аналітичне дослідження ринку**

**АНАЛІЗ МАКРО І МІКРОСЕРЕДОВИЩА БІЗНЕСУ**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Можливості і загрози бізнесу
2. Фактори макро- і мікро середовища
3. Аналіз внутрішнього середовища суб’єктів бізнесу

***1. МОЖЛИВОСТІ І ЗАГРОЗИ БІЗНЕСУ***

Середовище бізнесу постійно змінюється: можливості і загрози виникають, змінюються і зникають. Фактори середовища можуть становити як можливості, так і загрози для діяльності суб’єктів бізнесу на ринку.

***Можливості*** – це наслідок позитивного впливу на діяльність ринкових суб’єктів факторів зовнішнього середовища, які проявляються (чи можуть проявитися) в розриві фактичних результатів діяльності з потенційними можливостями.

На існування проблем чи можливостей впливають як існуючі фактори ринкового середовища так і поява нових.

**Можливості**

**Ресурси**

**Цілі**

**Ринкові можливості**

Рис. 3.3. Можливості суб’єкту бізнесу

***Загрози*** – це наслідок негативного впливу на діяльність ринкових суб’єктів тих чи інших факторів зовнішнього середовища, що проявляється (чи може проявитися) в розриві між запланованими і фактичними результатами діяльності. Наприклад: низькі рівні обсягів продажу у порівняні з запланованими, зниження прибутку чи долі ринку тощо.

Для ринкових суб’єктів підприємництва важливо вчасно розпізнавати можливості і загрози, використовувати можливості і уникати загроз.

Для бізнес-середовища характерними є певні риси, які відображають особливості його функціонування та впливу. До них відносять:

* взаємозв’язок факторів середовища (рівень сили, з якою зміна одного фактору впливає на інші фактори);
* складність середовища (кількість факторів, на які організація зобов’язана реагувати, а також рівень варіативності кожного фактору);
* мінливість середовища (швидкість, з якою відбуваються зміни в середовищі);
* невизначеність середовища (відносна кількість інформації про оточення і певність в її точності).

Основним завданням аналізу даних в бізнесі є підбір і оцінка тих факторів макро-середовища, що можуть впливати і впливають на його діяльність.

Для аналізу можливостей потрібно проаналізувати фактори макро- та мікро-середовища (зовнішні фактори).

***2. ФАКТОРИ МАКРО- ТА МІКРО-СЕРЕДОВИЩА***

До факторів макро-середовища відносяться: політико-правові, економічні, демографічні, науково-технічні, природні та культурні фактори.

До факторів мікро-середовища належать наступні фактори: споживачі, конкуренти, постачальники та контактні аудиторії.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Макро-середовище*** | | | | | | | | | | | | | | | | | |
|  | **Політико-правові фактори**   * фактори політичної стабільності чи нестабільності * політична система * законодавчі основи бізнесу * методи державного регулювання економіки, інше | | | | | |  | **Економічні фактори**   * доходи населення * рівень цін * доступність кредитів * фаза розвитку країни та окремих галузей * рівень інфляції валовий національний * продукт, інше | | |  | **Демографічні фактори**   * рівень народжуваності та старіння * життєвий цикл сім’ї * рівень міграції * віковий, статевий склад населення рівень освіти, соціальна стратифікація, інше | | | | |  |
|  | | | | | |  | | |  | | | | |
|  | | ***Мікро-середовище*** | | | | | | | | | | | | |  |
|  | | | | | Посередники | | |  | | | | |
|  | | |
|  | Споживачі | |  | | **Суб’єкт** | | |  | | Конкуренти | |  |
|  | |  | | |  | |
|  | Постачальники | | | |  | Контактні аудиторії | | | |  |
|  | | | |  | | | |
|  | | | | | | | | | | | | |
|  | **Науково-технічні фактори**   * технічні і технологічні інновації * рівень оновлення продукції * інформаційні і комунікаційні можливості * захист прав інтелектуальної власності * життєвий цикл товарів і технологій | | | | | | **Природні фактори**   * доступність сировини * кліматичні умови * екологія * джерела енергопостачання * розміщення продуктивних сил * геодезичні характеристики * інше | | | |  | **Культурні фактори**   * традиції * культура, субкультура * спосіб життя * звичаї * мовні особливості * інше | | | | |
|  | | | | | | | | | |  | | | | |

Рис. 3.4. Зовнішні фактори середовища бізнесу

До факторів глобального бізнес-середовища відносяться: рівень глобальної конкуренції, розвиток глобальних ринків, рівень транснаціоналізації світової економіки, інтенсивність міграційних процесів, розвиток міжнародних телекомунікацій, регулювання економіки з боку наднаціональних економічних організацій, поглиблення глобальних проблем сучасності тощо.

***Споживачі*** – суб’єкти ринкових відносин, які є реальними або потенційними покупцями товарів в конкретному акті ринкового обміну з метою задоволення своїх потреб.

Критерії класифікації ***cпоживачів***:

1. Лояльність:

• Реальні споживачі

– ті, які купують товари компанії протягом тривалого періоду часу;

– нові споживачі (споживачі, які вперше почали купувати товари);

– споживачі, які купують аналогічні товари конкурентів.

• Потенційні споживачі

* споживачі, які усвідомили свої мотивації щодо вибору виробника;
* споживачі, які перебувають в процесі усвідомлення свого економічного інтересу;
* індиферентні споживачі.

2. Типи ринків:

* Споживачі споживчого ринку(В2C)
* Споживачі промислового ринку (В2В)

3. Інтенсивність споживання:

– Великі оптові закупки

– Середні роздрібні закупки

– Одиничні закупки

4. Географічні межі ринків:

– Національні споживачі

– Споживачі міжнародних ринків

– Споживачі глобальних ринків

5. Джерела фінансування покупки:

– Державні підприємства і установи

– Громадські організації

– Представники приватного підприємництва

– Домогосподарства

***Конкуренти*** – це ринкові суб’єкти які виробляють або продають аналогічні товари, товари-замінники і економічні інтереси яких співпадають відносно споживачів і прямо протилежні один одному. Головна їх мета – завоювання споживачів. Компанії-конкурентів можна класифікувати за товарно-родовими ознаками – це конкуренція з боку товарів-замінників та товарно-видова конкуренція між товарами одного виду.

***Постачальники*** – це ринкові суб’єкти, які забезпечують компанію сукупністю необхідних для її діяльності товарів та послуг.

***Контактні аудиторії*** – це групи осіб (фізичних чи юридичних), які безпосередньо не приймають участі в процесі виробництва чи купівлі-продажу товарів, але можуть впливати на діяльність ринкових суб’єктів. До них відносять: фінансові інститути, засоби масової інформації, державні установи, цивільні групи дій, широка громадськість, внутрішні контактні аудиторії.

***3. ФАКТОРИ ВНУТРІШНЬОГО СЕРЕДОВИЩА СУБ’ЄКТІВ БІЗНЕСУ***

Необхідно розрізняти мікро-фактори зовнішнього середовища і мікро-фактори внутрішнього середовища суб’єктів господарювання. Існують різні критерії класифікації внутрішніх ресурсів. На рис. 3.5. наведено класифікацію внутрішніх ресурсів суб’єктів господарювання за критерієм можливості відображення у бухгалтерській звітності.

На стадії оцінки сильних і слабких сторін в ринковій діяльності суб’єктів господарювання здійснюється аналіз відповідності факторів внутрішнього мікро-середовища вимогам зовнішнього макро- і мікро-середовища.

Аналіз факторів макро-середовища можна формалізувати шляхом виконання наступного алгоритму: для того, щоб проаналізувати вплив кожної групи факторів необхідно перш за все підібрати ті фактори по кожній групі, які можуть вплинути на діяльність суб’єкта бізнесу; класифікувати відібрані фактори за критеріями- загроза (-), чи можливість (+) та впливу їх на попит чи пропозицію.

Після підбору імовірних факторів, необхідно провести їх детальну аргументацію за наступною схемою:

- підбір фактів, цифрових даних, які підтверджують вплив кожного фактору на діяльність суб’єкту господарювання в якості загрози чи можливості (треба мати на увазі, що один і той же фактор може впливати і як загроза і як можливість, але сила впливу може бути різною);

- аргументація, щодо впливу фактору на попит чи пропозицію;

- розробка гіпотези можливої реакції суб’єкта господарювання на дію фактору.

- експертна оцінка значущості дії фактору за допомогою оціночної шкали (від 1-10).

***Внутрішні фактори***

|  |  |
| --- | --- |
| **Ресурси, що не відображаються у**  **бухгалтерській звітності**   * організаційно-управлінські (організаційна форма підприємства, структура управління, стратегії, система планування, контроль, мотивації); * інформаційні ресурси (джерела інформації, бази даних, методи обробки інформації); * кадрові ресурси (керівники, фахівці, компетенції, корпоративна культура, етика взаємовідносин); * комерційні ресурси (система партнерських відносин, збутові мережі, ресурси часу); * адміністративні ресурси (держзамовлення, зв’язки з державними і місцевими органами влади) | **Ресурси, що відображаються у фінансовій звітності**   * виробничо-технічні ресурси (земля, природні ресурси, будівлі. споруди, засоби виробництва, інфраструктура, технології); * матеріальні ресурси (сировина, матеріали, комплектуючі, готові товари); * нематеріальні ресурси (ліцензії, патенти, бренд, ноу-хау, програмні продукти); * фінансові ресурси (власний капітал, запозичений капітал, грошові кошти, відкладені платежі, гудвіл) |

Рис. 3.5. Внутрішні мікро-фактори суб’єктів бізнесу

Аналітичні таблиці і аргументація складається по всіх шести групах факторів: політично-правові, економічні, демографічні, науково-технічні, природні, соціо-культурні. Необхідно зазначити, що деякі групи факторів можуть взагалі не впливати на діяльність того чи іншого суб’єкта господарювання, що досліджується, в такому випадку дана група не аналізується, але аргументується причина.

Після того, як підібрані, аргументовані та оцінені всі групи факторів макро-середовища необхідно підсумувати цій аналіз шляхом складання двох підсумкових зведених таблиць, - ринкових загроз та можливостей. Для складання цих таблиць необхідно визначити вагомість кожної з шести груп факторів за допомогою шкали з постійною сумою (1, 10, 100).

Вагові коефіцієнти розподіляються між групами факторів за наступними критеріями:

- кількість факторів в кожній групі;

- вагомість факторів в кожній групі;

- співвідношення між факторами проблем і можливостей в групі;

- співвідношення між факторами попиту і пропозиції в групі;

- складність варіанту вирішення проблем чи реалізації можливості. Експертні оцінки за групами факторів необхідно скорегувати на отримані вагові коефіцієнти.

Зведені підсумкові таблиці факторів макро-середовища складаються за наступною формою (табл. 3.1. – 3.2.).

Таблиця 3.1.

Зведена підсумкова таблиця факторів загроз

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Фактори | Експертна бальна оцінка значущості фактору з урахуванням вагового коефіцієнту | Гіпотетичні варіанти вирішення проблем | Фактор попиту/ пропозиції |
| 1. |  |  |  |  |
| 2. |  |  |  |  |

Таблиця 3.2.

Зведена підсумкова таблиця факторів можливостей

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Фактори | Експертна бальна оцінка значущості фактору з урахуванням вагового коефіцієнту | Гіпотетичні варіанти реалізації можливостей | Фактор попиту/ пропозиції |
| 1. |  |  |  |  |
| 2. |  |  |  |  |

Аналіз факторів мікро-середовища передбачає дослідження таких факторів як: споживачі, конкуренти, постачальники, контактні аудиторії.

Аналіз споживачів передбачає розробку гіпотези можливої стратегії охоплення ринку (масовий маркетинг, множинна сегментація, концентрований маркетинг) виходячи з попереднього аналізу факторів макро-середовища.

При цьому слід аргументувати привабливість обраної стратегії, описати критерії сегментації (у випадку обрання цієї стратегії), описати сегменти, можливі мотиви, дати попередню оцінку ємкості цільового ринку.

При аналізі конкурентів необхідно розробити гіпотези конкурентної стратегії, та факторів, що можуть бути основою формування ключових компетенцій. Дуже важливим при цьому відібрати декількох безпосередніх конкурентів для здійснення аналізу сильних і слабих сторін компанії.

Аналізуючи постачальників, необхідно виявити загрози та можливості, що виникають у фірми під час роботи з постачальниками. Якщо необхідно, то слід провести рейтингування постачальників. За такою ж схемою аналізують контактні аудиторії.

Таблиця 3.3.

Аналіз сильних і слабких сторін суб’єкта бізнесу порівнянні з безпосередніми конкурентами

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Показники | Cуб’єкт бізнесу | Конкурент 1 | Конкурент 2 |
|  | Виробничі потужності | Оціночні шкали | Оціночні шкали | Оціночні шкали |
| 1. |  |  |  |  |
| 2. |  |  |  |  |
|  | Товар |  |  |  |
| 1. |  |  |  |  |
| 2. |  |  |  |  |
|  | Кадри |  |  |  |
| 1. |  |  |  |  |
| 2. |  |  |  |  |
|  | Маркетинг |  |  |  |
| 1. |  |  |  |  |
| 2. |  |  |  |  |
|  | Ціна |  |  |  |
| 1. |  |  |  |  |
| 2. |  |  |  |  |
|  | Канали розподілу |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Після завершення аналізу факторів макро- і мікро зовнішнього середовища здійснюється аналітичне заключення, щодо ступені привабливості певного ринку для даного суб’єкта бізнесу і приймається рішення щодо доцільності подальшого дослідження даного ринку і роботи на ньому.

***Лекція 3.3. (2 год)***

**Тема 3. Аналітичне дослідження ринку**

**СИСТЕМА БІЗНЕС-ІНФОРМАЦІЇ. ПОСЛІДОВНІСТЬ ЇЇ АНАЛІЗУ**

Для отримання достовірної і необхідної для прийняття конкретних управлінських рішень інформації необхідно застосовувати різні типи досліджень.

***Типи досліджень*** вирізняються за певними критеріями.

1. За критерієм – мета дослідження розрізняють академічні (ринкові) і комерційні дослідження.

***Академічні або ринкові дослідження*** – це збір інформації з метою вивчення тенденцій, закономірностей розвитку різноманітних ринків, моделей споживчої поведінки, культурних особливостей тощо, не для вирішення конкретної управлінської проблеми.

***Комерційні дослідження*** – це дослідження, які здійснюються суб’єктом бізнесу чи дослідницькою організацією для прийняття конкретних управлінських рішень.

1. За степенем проникнення в проблему розрізняють: теоретичні; дискрептивні або описово- аналітичні дослідження; причино-наслідкові дослідження.

***Теоретичні дослідження*** використовуються на ранніх етапах прийняття управлінських рішень для того, щоб з мінімальними витратами коштів і часу зібрати попередню інформацію про ситуацію, яка виникла.

***Дискрептивні*** або ***описово-теоретичні дослідження*** передбачають виявлення факторів макро- та мікрорівня, що впливають на діяльність компанії в умовах плинного ринкового середовища.

***Причинно-наслідкові дослідження*** використовується у випадку, якщо необхідно визначити вплив одного фактору на інший при можливості його кількісного визначення.

1. За типом ринку розрізняють ***дослідження ринку споживчих товарів (В2С)*** *і* ***промислового ринку (В2В*)**.
2. За критерієм періодичності дослідження поділяються на: спеціальні, що здійснюються одноразово під задану мету дослідження; систематичні, що передбачають постійне спостереження за певними об’єктами дослідження (моніторні дослідження, «таємничі дослідження”).
3. За критерієм – кількість об’єктів дослідження розрізняють: якісні і кількісні дослідження.

***Якісні дослідження*** – це неструктурований або частково структурований тип отримання первинної інформації, від обмеженої кількості респондентів , що є носіями знань про об’єкти , що досліджуються і має на меті зрозуміти проблему та зібрати інформацію щодо глибинних якісних характеристик об’єкту.

***Кількісні дослідження*** – це структурований тип отримання первинної інформації від значної кількості респондентів в ході проведення вибіркових спостережень; він передбачає представлення інформації в кількісній формі з використанням методів статистичної обробки інформації.

1. За географічними межами розрізняють дослідження національних та міжнародних ринків ( більше одного ринку).

***Цінність інформації*** – це рівень її значущості для прийняття управлінських рішень в умовах конкретної бізнес-ситуації.

***Основні методологічні принципи****,* ***що забезпечують ефективність досліджень*:**

* ***об***’***єктивність*** – необхідність збору інформації, яка адекватно відтворює реальну ринкову ситуацію за допомогою чіткої систематизації процесу досліджень та запобігання можливих помилок в процесі збору і обробки даних;
* ***систематичність*** – здійснення досліджень у відповідності до розроблених планів у строгій відповідності до необхідних етапів;
* ***комплексність*** – здійснення комплексу взаємопов’язаних дій, одні з яких забезпечують ефективність наступних;
* ***безперервність*** –дослідження є процесом, що постійно повторюється у зв’язку з постійною динамічною зміною факторів середовища;
* ***оперативність*** – передбачає здатність швидко та своєчасно виправити процес досліджень або направити їх у бажаному руслі.

Етапи досліджень представлені в табл. 3.4.

**Перший етап: визначення цілей дослідження.**

Даний етап дослідження є найбільш складним і тим, що вимагає високої кваліфікації експертів-дослідників. Суть даного етапу полягає в розробці найбільш прийнятного варіанту вирішення управлінської проблеми, що виникла у конкретного суб’єкта господарювання.

***Управлінська проблема*** – це конкретна господарська ситуація, яка виникла у суб’єкта господарювання в ході його ринкової діяльності і вимагає вирішення (зниження обсягів продажу, зниження ефективності реклами, проблема виведення на ринок нового товару тощо).

Для виявлення найбільш ефективних варіантів вирішення управлінської проблеми необхідно виявити проблеми і можливості, розробити гіпотези варіантів вирішення проблем чи реалізації можливостей. На основі зведених підсумкових таблиць проблем та можливостей (табл. 3.1. і 3.2.), а саме гіпотез варіантів вирішення проблем чи реалізації можливостей, після їх узагальнення і групування, складається таблиця варіантів вирішення управлінської проблеми (табл. 3.5.).

Таблиця 3.4.

Характеристика етапів дослідження бізнес-інформації

|  |  |
| --- | --- |
| Етапи | Зміст |
| **1.** | **Визначення цілей дослідження**   * аналіз проблеми і можливостей; * визначення альтернативних варіантів вирішення проблеми; * встановлення меж дослідження. |
| **2.** | **Розробка дослідницьких завдань:**   * складання пошукових питань; * розробка гіпотез; * визначення джерел інформації; * оцінка цінності інформації. |
| **3.** | **Організація і планування дослідження:**   * вибір методів дослідження; * складання плану якісного та кількісного дослідження * розробка анкет; * складання графіку та бюджету дослідження. |
| **4.** | **Збір інформації (польові роботи)** |
| **5.** | **Обробка і аналіз інформації** |
| **6.** | **Підготовка звіту та презентація результатів дослідження** |
| **7.** | **Впровадження результатів дослідження** |

Таблиця 3.5.

Альтернативні варіанти вирішення управлінської проблеми

|  |  |  |  |
| --- | --- | --- | --- |
| *№* | *Альтернативи* | *Слабкі сторони* | *Сильні сторони* |
| 1. |  |  |  |
| 2. |  |  |  |
| 3. |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

З даної таблиці відбирається один варіант, який є найбільш прийнятним для суб’єкта бізнесу, за наступними критеріями:

- витрати на реалізацію варіанту;

- тривалість реалізації варіанту;

- ефективність варіанту.

В ході одного дослідження перевіряється лише один варіант вирішення управлінської проблеми. На основі відібраного варіанту формується мета дослідження і встановлюються його межі.

***Межі дослідження*** встановлюються за наступними критеріями:

- територіальні межі (міжнародні ринки, національний ринок, ринок регіону, міста тощо);

- часові межі (короткостроковий, середньостроковий, довгостроковий період).

- широта продуктової лінії (в одному дослідженні одна продуктова лінія у випадку масового маркетингу, наприклад, пральні порошки, жувальні гумки тощо чи продуктова лінія кожного сегменту);

-кількість ринкових сегментів.

**Другий етап: розробка дослідницьких завдань**

Основне завдання даного етапу визначення обсягу інформації необхідної для прийняття управлінських рішень.

***Пошукові питання*** – це перелік запитань, відповіді на які мають бути отримані в ході досліджень для реалізації його мети. Для розробки дослідницьких завдань складається таблиця пошукових питань, гіпотез на пошукові питання і джерел інформації.

Таблиця 3.6.

Перелік пошукових питань, гіпотез та джерел інформації

|  |  |  |  |
| --- | --- | --- | --- |
| № | Пошукові питання | Гіпотеза | Джерела інформації |
| 1.  2. |  |  |  |

Інформація для відповіді на пошукові питання міститься в джерелах вторинної інформації та у відповідях експертів. На деякі пошукові питання можна розробити гіпотези, правильність яких буде перевірятись у ході збору первинної інформації.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Визначення цілей дослідження** | | | | |
|  | | | | |
| **Визначення потреби в інформації** | | | | |
|  | | | | |
| **Пошук та аналіз внутрішніх джерел інформації** | | | | |
|  | | | | |
| **Пошук та аналіз джерел вторинної інформації** | | | | |
|  | | | | |
| **Збір первинної інформації** | | | | |
|  | | | | |
| **Опитування** | **Спостереження** | **Експеримент** | **Моделювання** | **Аналіз аналогічних ситуацій** |

Рис. 3.6. Схема збору інформації

Інформація поділяється на: первинну та вторинну.

***Первинна інформація* –** це інформація, яка збирається вперше (суб’єктом бізнесу чи сторонньою дослідницькою організацією) для досягнення поставлених цілей суб’єктів господарювання, і для прийняття управлінських рішень.

***Вторинна інформація*** – це така інформація, яка існує в опублікованому вигляді чи у вигляді електронних баз даних і збиралась для цілей інших, ніж ті, які має на меті суб’єкт господарювання в конкретних ринкових умовах. Переваги: невисока вартість, економія часу. Недоліки: може бути застарілою, не повною, суперечливою, може не підходити до цілі дослідження, методика збору цієї інформації може бути невідома.

Вторинна інформація поділяється на внутрішню та зовнішню інформацію.

***Внутрішня інформація*** – це дані, які збираються й аналізуються суб’єктом бізнесу. Вона може бути у вигляді бухгалтерських звітів, звітів про обсяг закупівель та збуту та інше.

***Зовнішня інформація*** – це опублікована інформація, яка збирається за межами конкретного суб’єкта бізнесу (періодичні видання, довідники, книги, комп’ютерні бази даних та інше).

Первинна інформація може класифікуватися за критерієм – кількість об’єктів дослідження. За даним критерієм первинна інформація поділяється на якісну і кількісну.

***Якісна первинна інформація*** – це бази даних, які збираються на основі дослідження обмеженої кількості респондентів, які є носіями глибоких, професійних знань щодо об’єкту дослідження.

Для продовження процесу досліджень експерт має оцінити цінність додаткової інформації необхідної для прийняття управлінських рішень. Якщо у дослідника існує достатня кількість вторинної інформації для відповідей на пошукові питання цінність додаткової інформації знижується. Цінність інформації залежить від рівня стабільності факторів ринкового середовища, типу ринку, наявності достовірної вторинної інформації, масштабів проєкту, обсягів витрат на дослідження, ступені ймовірного ринкового успіху товару тощо.

***Цінність інформації*** – це рівень її значущості для прийняття управлінських рішень умовах конкретного бізнес-середовища. Для прийняття рішення про доцільність подальших досліджень використовуються, зокрема, байесовський метод.

**Третій етап: організація і планування маркетингових досліджень**

Цей етап передбачає вибір методів збору первинної інформації; розробку анкети; планування вибіркового спостереження чи експерименту; розробку графіку та кошторису дослідження.

До методів збору якісної первинної інформації відносять: індивідуальні та групові експертні опитування; фокус-групи; глибинні інтерв’ю; метод Дельфі.

**Експертні опитування** – це структуроване, або частково структуроване опитування експертів, відібраних на основі певних критеріїв.

**Глибинне інтерв’ю** – це неструктурований метод збору якісної інформації, що передбачає проведення особистого інтерв’ю одним з .висококваліфікованим інтерв’юером для з’ясування глибинних, часто підсвідомих мотивацій щодо об’єкту дослідження.

**Метод Дельфі** – це структурований метод збору якісної інформації, шляхом опитування групи експертів, що здійснюється у декілька етапів і побудований за принципом зворотного зв’язку. Експертні оцінки передбачають взаємний аналіз поглядів експертів при відсутності безпосередніх контактів між ними і збереженні анонімності.

**Проекційні методи** – це неструктуровані, непрямі форми опитування, які спонукають респондентів висловлювати свої приховані мотивації по відношенню до об’єкту дослідження шляхом оцінки поведінки третіх осіб (асоціативні методи, метод завершення ситуації, методи конструювання ситуації, тощо).

***Основними методами збору кількісної первинної інформації*** є: опитування, спостереження, експеримент, моделювання, аналіз аналогічних ситуацій.

**Опитування** – це метод збору первинної інформації в процесі якого вивчаються погляди респондентів шляхом відповідей на запитання за допомогою різних форм комунікацій (особисті інтерв’ю, телефоні інтерв’ю, поштове інтерв’ю, електронне інтерв’ю).

Таблиця 3.7.

Порівняльні характеристики методів комунікацій при опитуванні

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Фактори* | *Особисте інтерв’ю* | *Телефонне інтерв’ю* | *Поштове інтерв’ю* | *Електронне інтерв’ю* |
| 1.Витрати часу | Значні | Помірні | Помірні | Значні |
| 2.Витрати коштів | Високі | Помірні | Низькі | Низькі |
| 3.Гнучкість | Висока | Помірна | Низька | Низька |
| 4.Складність адміністративного управління | Висока | Помірна | Низька | Низька |
| 5.Обсяг даних | Значний | Низький | Помірний | Помірний |
| 6.Відсоток відкликів | Високий | Помірний | Низький | Низький |
| 7.Контроль вибірки | Значний | Помірний | Помірний | Низький |

Методи комунікацій при опитувані обираються в залежності до наявних фінансових ресурсів, часу, та вимог до ступеню достовірності інформації.

Таблиця 3.8.

Переваги і недоліки методів комунікації при опитуванні

|  |  |
| --- | --- |
| **Переваги** | **Недоліки** |
| **Особисте інтерв’ю** | |
| * можливість зацікавити опитуваних, можливість зменшити число відмов * можливість задавати складні питання, що вимагають додаткових пояснень * можливість прояснити те, що незрозуміло * можливість варіювати питаннями | * вимагають значних витрат часу на проведення * потребують значних фінансових витрат * складні в адміністративному управління і здійснені контролю |
| **Телефонне інтерв’ю** | |
| * простіше в адміністративному управлінні (можливість здійснювати з єдиного центру) * менші витрати часу * є можливість одержання повторної відповіді на питання | * унеможливлює перевіряти складні питання * неможливість спостерігати реакцію опитуваного * складності у формуванні репрезентативної вибірки * можливість не якісного телефонного зв’язку |
| **Поштове інтерв’ю** | |
| * можливість отримання більш продуманих відповідей * економія фінансових ресурсів * можливість задавати конфіденційні питання | * малий відсоток повернення заповнених анкет * неможливість визначити число учасників, що заповнювали анкету * неможливість пояснення складних питань |

Таблиця 3.9.

Типи помилок при опитуванні

|  |  |  |
| --- | --- | --- |
| **Типи помилок при опитуванні** | | |
| **Помилки вибіркового спостереження** | **Помилки, пов’язані з проведенням інтерв’ю** | **Помилки пов’язані з особою опитуваного** |
| * Помилки з причин опитування респондентів, які не є цільовими * Помилки з причин відсутності на місці або відмови від участі в інтерв’ю | * Нездатність інтерв’юера установити   контакт із респондентом   * Невміння точно ставити запитання * Помилки при фіксуванні відповідей | * Неточність відповідей * Неспроможність сформулювати відповідь * Незрозуміла відповідь |

**Спостереження** – це метод збору первинної інформації шляхом спостереження за поведінкою респондентів дослідником або технічними засобами.

Таблиця 3.10.

Переваги та обмеження методу спостереження

|  |  |
| --- | --- |
| *Переваги* | *Обмеження* |
| Об’єктивність в оцінці ситуації, що спостерігається | Неможливість дістати інформацію про глибинні мотивації |
| Відсутність фактору відмов | Значні витрати часу і коштів |
| Відсутність помилок пов’язаних з нерозумінням запитань |  |

**Комп’ютерне моделювання** – це метод збору первинної інформації за допомогою комп’ютерних програм. В моделі вводяться різноманітні параметри, які відображають зміни системи, що стосуються комплексу маркетингу чи зовнішніх факторів середовища. Під час моделювання визначається дія незалежних факторів на залежні. При цьому моделюється і порівнюються різні варіанти можливого впливу факторів.

Таблиця 3.11.

Переваги та обмеження моделювання

|  |  |
| --- | --- |
| *Переваги* | *Обмеження* |
| Вимагає незначних витрат коштів і часу | Складність і інколи неможливість відтворення реальної ситуації |
| Дає можливість отримати конкретні цифрові показники | Плинність факторів маркетингового середовища |

**Аналіз аналогічних ситуацій** – це метод збору первинної інформації, за допомогою матеріалу у вигляді господарських ситуацій (кейсів), які допомагають виявити критерії прийняття рішень і перенести їх на ситуацію, що досліджується.

Таблиця 3.12.

Переваги і обмеження аналізу аналогічних ситуацій

|  |  |
| --- | --- |
| *Переваги* | *Обмеження* |
| Невисока вартість | Обмежений банк ситуацій |
| Економія часу | Не повна відповідність ситуації, що була тій, що досліджується |

**Анкети в бізнес-дослідженнях**

***Анкета*** – це структурований, логічний письмовий перелік запитань (відкритих і закритих), які дають змогу отримати достовірну інформацію для відповідей на пошукові питання для досягнення конкретної мети дослідження.

**Процедура складання анкети**

* відбір зі списку пошукових питань тих питань, джерелами інформації для яких зазначено збір первинної інформації
* визначення структури анкети
* обґрунтування логіки і обсягу анкети;
* вибір типу запитань та їх редагування
* вибір методів шкалювання поглядів респондентів
* складання таблиці відповідності пошукових питань питанням анкети
* апробація анкети, корегування анкети і її оформлення

**Типова анкета складається з наступних трьох змістовних частин:**

*1. Вступ.* У вступі зазначається суб’єкт, який здійснює дослідження, мета дослідження, комунікаційний місток між дослідником та респондентом для заохочення останнього до співпраці та порядок заповнення анкети.

*2. Основна частина* містить комунікаційні питання, питання, які дають змогу відповісти на пошукові питання. Комунікаційні запитання не відповідають на питання, що стосується мети дослідження, але саме вони встановлюють контакт з респондентами та задають йому тему для спілкування. Після цього в чіткій послідовності розгортаються питання, що дозволяють реалізувати мету дослідження. Логіка розгортається по схемі: від питань, що розкривають ідеальне відношення респондента до певного товару (від загального) до питань, які потребують пригадування, підрахунків, намірів та відношення до конкретних торгових марок (конкретного).

*3. Відомості про респондента.* У випадку, коли застосовується стратегія масового маркетингу відомості про респондента практично відсутні, перевіряються лише критерії сегментації.

**Типи питань в анкеті**

***Відкриті питання*** – це вербальні конструкції (запитання, розповідь, текст) чи малюнок без заданої структури відповіді, при цьому респондент в довільній формі висловлює свою думку.

***Закриті питання*** – це вербальні, графічні, цифрові конструкції (запитання, твердження, шкали) з чітко.

**Види відкритих запитань в анкеті**

питання без заданої структури відповіді

підбір словесних асоціацій

завершення речення

завершення розповіді

завершення рисунку

тематичний тест

Рис. 3.7. Види відкритих запитань в анкеті

**Види закритих запитань в анкеті**

Альтернативне запитання (відповідь «так» чи «ні»)

Запитання з варіантами відповідей

Запитання зі шкалою Лайкерта (передбачає можливість висловити ступінь згоди чи незгоди з висунутим твердженням)

Семантичний диференціал (розмірна шкала, межами якої є біполярні твердження)

Шкала важливості (респондент визначає ступінь важливості з запропонованих йому тверджень)

Оціночна шкала (бальна оцінка запропонованого твердження)

Рис. 3.8. Види закритих запитань в анкеті

Для оцінки поглядів споживачів використовуються номінальні, порядкові, інтервальні та відносні шкали. Після чого анкети редагуються і оформлюються у відповідності до правил.

**Вибіркове спостереження** – це кількісний метод збору первинної інформації, в ході якого спостереженню підлягають не всі одиниці генеральної сукупності, а лише їх частина, відібрана на основі певних процедур і критеріїв. Формування вибіркової сукупності включає:

* визначення генеральної сукупності;
* встановлення меж вибірки;
* визначення процедури відбору одиниць вибіркового спостереження;
* визначення методів доступу до респондентів;
* визначення розміру вибірки;
* визначення критеріїв відбору елементів спостереження;
* розробка плану управління відмовами.

**Процедури вибору одиниць вибіркового спостереження**: за *принципом випадковості* (імовірнісна процедура) або *принципу детермінованості* (неімовірнісна процедура).

**Види імовірнісних вибірок**: проста випадкова вибірка, систематична вибірка, типова вибірка, групова вибірка.

**Проста випадкова вибірка** передбачає те, що кожен член генеральної сукупності має рівну можливість бути відібраним для спостереження з таблиці випадкових чисел.

**Систематична вибірка** передбачає вибір елементів генеральної сукупності на певній визначеній основі, – наприклад, кожний десятий.

**Типова або стратифікована вибірка** використовується за умов, що у вибірці є підгрупи, які якісно відрізняються одна від одної (наприклад, ринкові сегменти). Відбір здійснюється в кожній групі пропорційно її розміру і міри розбіжності поглядів.

**Групова вибірка** використовується в тому випадку, коли генеральна сукупність ділиться на групи, але не за стратами (наприклад, студентські групи).

**Види неімовірнісних вибіркових спостережень:**

* *експертна вибірка,* за якою дослідники самі встановлюють критерії відбору експертів;
* *відбір за принципом снігової кулі* (один фахівець рекомендує інших), використовується для невеликих вибірок, де респонденти частіше спеціалісти вузького профілю;
* *принцип зручності* використовується тоді, коли необхідно швидко і дешево отримати інформацію (студенти в групі, жінки в магазині, відвідувачі кафе тощо);
* *квотовані вибірки* використовуються у випадку, коли необхідно обстежити невеликі підгрупи генеральної сукупності, які, як правило, відбираються за демографічними критеріями у відсотках до їх реальної структури в рамках генеральної сукупності.

**Експеримент** є методом збору первинної інформації в ході здійснення причинно-наслідкових досліджень у випадках, коли необхідно дослідити і виміряти вплив одного фактору на інший.

Експерименти можуть бути лабораторні і такі, що здійснюються в реальних ринкових умовах.

Для збору первинної інформації слід визначитись з попереднім графіком збору первинної інформації. ***Необхідні заходи*** щодо збору інформації:

1. Розробка попередньої анкети
2. Попередній відбір вибірки
3. Попередній відбір інтерв’юерів
4. Підготовка навчального матеріалу для інтерв’юерів.
5. Підготовка інтерв’юерів.
6. Проведення перевірочних Інтерв’ю.
7. Аналіз попередніх результатів інтерв’ю.
8. Остаточна підготовка всіх необхідних форм, тиражування анкет.
9. Остаточний відбір інтерв’юерів.
10. Підготовка інтерв’юерів.
11. Проведення інтерв’ю.
12. Проведення перевірки анкетування.
13. Підготовка звіту про польові роботи

В ході здійснення польових робіт треба мати на увазі, що не завжди графік робіт виконується так, як заплановано. Це пов’язано з тим, що досить важко передбачити кількість відмов або труднощів у контакті з респондентами, чи інші ситуації. Тому в ході проведення польових робіт необхідно контролювати ці показники.

**Четвертий етап: збір інформації (польові роботи)**

Основна мета цього етапу здійснити збір інформації в ринкових умовах. Цей етап передбачає підготовку інтерв’юєрів інструкцій для інтерв’юерів, місця та графіку збору даних.

**П’ятий етап: обробка та аналіз даних.**

В рамках даного етапу здійснюється кодування, редагування, статистична обробка даних за допомогою спеціалізованих комп’ютерних програм, зокрема SPSS. Здійснюється аналіз імовірних помилок та проблеми відмов.

**Шостий етап: підготовка звіту та презентація результатів**

**Звіт** – це чітко структурований письмовий звіт в якому висвітлюються результати дослідження у вигляді інформації, що відповідає поставленій меті дослідження. Структура звіту узгоджується з замовником.

**Типова структура звіту** включає наступні розділи:

1. Титульний лист, який включає назву проєкту, термін виконання, дату замовлення, зазначення виконавця та замовника.

2. Лист-супроводження, який включає характеристику типу дослідження, виконавців, підсумки роботи без висновків, пропозиції щодо подальших дій замовника.

3. Договір на проведення дослідження.

4. Зміст з перерахуванням розділів, таблиць, графіків, додатків.

5. Резюме для вищого керівництва, яке включає управлінську проблему, основні результати, висновки і рекомендації.

6. Вступ в якому зазначаються причини виникнення управлінської проблеми та мета дослідження.

7. Основні розділи: методологія дослідження, результати дослідження, обмеження, висновки і рекомендації, додатки.

Презентація готується залежно від характеру аудиторії відповідно до загальноприйнятих вимог.

**Сьомий етап: впровадження результатів.**

Даний етап може здійснювати, як замовник так і стороння дослідницька організація в залежності від домовленості між ними. Якщо впровадження рекомендацій здійснює стороння організація, це вимагає додаткової домовленості.

***Лекція 3.4. (2 год)***

**Тема 3. Аналітичне дослідження ринку**

**ІНСТРУМЕНТИ ДЛЯ БІЗНЕС-АНАЛІЗУ**

1. ***Визначення критеріїв приймання та оцінки*** – визначення вимог, які повинні бути задоволені для того, щоб рішення вважалося прийнятним для ключових зацікавлених сторін.

Визначити, які вимоги можуть бути найбільш ефективно використані в якості критеріїв приймання або оцінки.

* критерій приймання описує мінімальний набір вимог, які повинні бути задоволені для того, щоб конкретне рішення коштувало реалізовувати.
* критерій оцінки є набором вимог, які будуть використовуватися для вибору між декількома рішеннями.

Критерій приймання та оцінки може бути визначений, якщо може бути продемонстровано, що рішення або компонент рішення відповідає вимозі. Критерій приймання зазвичай використовується, коли тільки одне можливе рішення в даний час оцінюється і, як правило, виражається у вигляді позитивного або негативного рішення. Критерії оцінки використовуються для порівняння декількох рішень або компонентів рішення і допускають використання діапазону можливих балів.

**Елементи метода**

1. Тестування
2. Скоринг (рейтинг)
3. ***Беклог (backlog) –*** це список всіх робіт, записаний у вигляді коротких описів і сформований за пріоритетами.

Кожен елемент беклога повинен відповідати таким **принципам**:

* **Detailed Appropriately** – мати належну деталізацію;
* **Estimated** – мати оцінку;
* **Emergent –** повинен бути таким, що несподівано виникає. Емерджентність – наявність у будь-якої системи особливих властивостей, які не притаманні її елементам, а також сумі елементів, не пов’язаних особливими системоутворюючими зв’язками, не зведення властивостей системи до суми властивостей її компонентів. Тобто беклог не статичний, він змінюється з плином часу, в міру того, як ми більше дізнаємося про рішення, історії в беклог додаються, видаляються елементи або змінюються їх пріоритети;
* **Prioritized** – елементи беклога мають бути пріоретизованими.

**Управління беклогом** – це «наука» ковзної роботи через етапи послідовного життєвого циклу продукту / рішення в ефективній формі.

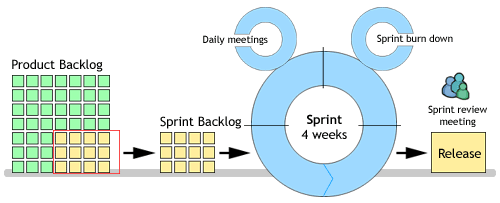
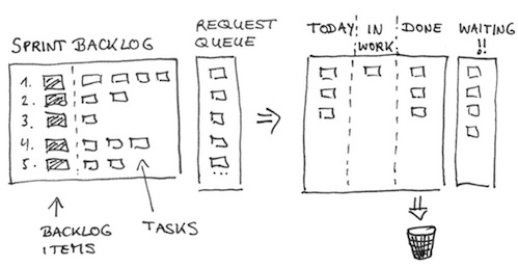
**Ключовими аспектами успішного управління беклогом є:**

* Постановка цілей;
* Призначення і узгодження пріоритетів;
* Встановлення організаційних зобов’язань і відносин;
* Реалізація процесу досягнення цілей;
* Вимірювання продуктивності;
* Аналіз продуктивності;
* Аудит процесу.

**Управління беклогом включає в себе наступні процеси:**

* Управління робочими запитами;
* Розробка замовлень на виконання робіт, підготовка до роботи і ремонтні процедури;
* Планування робіт;
* Виконання роботи і моніторинг незавершених етапів робіт.

**Product Backlog:**

[](http://analytics.infozone.pro/wp-content/uploads/2014/12/product_backlog.gif)  
**Sprint Backlog:**  
[](http://analytics.infozone.pro/wp-content/uploads/2014/12/sprint_backlog.jpg)

1. ***Збалансована система показників (ЗСП, Balanced Scorecard, BSC)*** – концепція перенесення і декомпозиції стратегічних цілей для планування операційної діяльності та контролю досягнення цілей. По суті ЗСП – це механізм взаємозв’язку стратегічних задумів і рішень з щоденними завданнями, спосіб напрямки діяльності всієї компанії на досягнення стратегічних цілей.

**Основна структурна ідея BSC** полягає в тому, щоб збалансувати систему показників у вигляді чотирьох груп.

**Перша група** включає традиційні фінансові показники. Як би ми не доводили важливість ринкової орієнтації підприємства і досконалості внутрішніх процесів, власника завжди в першу чергу будуть цікавити показники фінансової віддачі на вкладені кошти. Тому збалансована система повинна починатися (в класифікації) і закінчуватися (в кінцевій оцінці) фінансовими показниками.

**Друга група** описує зовнішнє оточення підприємства, його ставлення з клієнтами. Основними фокусами уваги виступають:

* здатність підприємства до задоволення клієнта;
* здатність підприємства до утримання клієнта;
* здатність придбання нового клієнта;
* прибутковість клієнта;
* обсяг ринку;
* ринкова частка в цільовому сегменті.

**Третя група** характеризує внутрішні процеси підприємства:

* інноваційний процес;
* розробка продукту;
* підготовка виробництва;
* постачання основних ресурсів;
* виготовлення;
* збут;
* післяпродажне обслуговування.

**Четверта група** дозволяє описати здатність підприємства до навчання і росту, яка фокусується в наступні чинники:

* люди з їх здібностями, навичками і мотивацією;
* інформаційні системи, що дозволяють постачати критичну інформацію в режимі реального часу;
* організаційні процедури, що забезпечують взаємодію між учасниками процесу і визначають систему прийняття рішення.

**6 обов’язкових елементів ЗСП:**

* 1. **Перспективи (perspectives)** – компоненти, за допомогою яких проводиться декомпозиція стратегії з метою її реалізації. Зазвичай використовуються 4 базові перспективи, однак їх список можна доповнити відповідно до специфіки стратегії компанії. Базовими перспективами є:
* **Фінанси** (отримання стабільно зростаючого прибутку – «як бачать нас акціонери компанії»);
* **Клієнти** (поліпшення знання кожного клієнта – «як бачать нас клієнти»);
* **Процеси** (внутрішні процеси компанії – «чим ми виділяємося серед конкурентів»);
* **Персонал** (навчання і розвиток) і **інновації** («як ми створюємо і збільшуємо цінність для наших клієнтів»).
  1. **Стратегічні цілі (objectives)** визначають, в яких напрямках буде реалізовуватися стратегія.
  + Фінанси
    - * Зростання прибутку
      * Скорочення видатків
      * …
  + Клієнти
    - * Збільшення задоволеності клієнтів
      * Залучення клієнтів
      * Збереження клієнтської бази
      * …
  + Внутрішні бізнес-процеси
    - * Закупівля якісних ТМЦ та інструментів
      * Оптимізація запасів
      * Зростання якості робіт
      * Дотримання своєчасності бізнес-процесів
      * Впровадження обліку
      * Створення системи контролю
      * …
  + Навчання та розвиток
    - * Підвищення кваліфікації персоналу
      * Запровадження автоматизації управління
      * …
  1. **Показники (measures) –** це метрики досягнень, які повинні відображати прогрес у русі до стратегічної мети. Показники мають на увазі певні дії, необхідні для досягнення мети, і вказують на те, як стратегія буде реалізована на операційному рівні.
  2. **Цільові значення (targets)** – кількісні вираження рівня, якому повинен відповідати той чи інший показник.
  3. **Причинно-наслідкові зв’язки (cause and effect linkages)** повинні пов’язувати в єдиний ланцюжок стратегічні цілі компанії таким чином, що досягнення однієї з них обумовлює прогрес в досягненні інший (зв’язок по типу «якщо-то»).
  4. **Стратегічні ініціативи (strategic initiatives)** – проєкти або програми, які сприяють досягненню стратегічних цілей.

1. **Бенчмаркинг** – це процес вивчення господарської діяльності конкурентів, продуктів, послуг і процесів діяльності конкурентів, з метою використання їх позитивного досвіду в роботі своєї компанії (реалізації змін для досягнення і збереження конкурентоспроможності).

**Залежно від об’єктів порівняння бенчмаркінг може поділятися на кілька видів:**

* **Внутрішній бенчмаркінг** – при цьому виді бенчмаркінгу здійснюється порівняння процесів (продуктів, послуг) всередині організації. Як об’єкти вибираються близькі або схожі процеси (продукти, послуги). При внутрішньому бенчмаркингу досить легко зібрати дані, проте можливості для порівняння обмежені, і результати можуть виявитися упередженими.
* **Конкурентний бенчмаркінг** – порівняння проводиться з прямими конкурентами (за певними продуктами або послугами), які працюють на місцевому, регіональному або міжнародному ринку. Для цього виду бенчмаркінгу необхідно вибирати конкурентів, що знаходяться на іншому «рівні» ринку. Наприклад, організація, що працює на місцевому ринку, може вибрати для порівняння організацію, яка працює на міжнародному ринку. У цьому випадку дані, отримані при порівнянні, будуть більш обґрунтованими і важливими, але їх досить важко отримати.
* **Функціональний бенчмаркінг** – порівнюються процеси власної організації зі схожими процесами іншої організації, але яка працює в іншій сфері діяльності. При такому вигляді бенчмаркінгу можна отримати об’єктивні і важливі дані з меншими зусиллями, застосовуючи етичні і легальні методи отримання інформації.
* **Узагальнений бенчмаркінг –** для цього виду бенчмаркінгу відбираються організації, які мають кращі в своєму сегменті процеси і підходи. Такі організації відкрито публікують інформацію про діяльність (прикладами можуть служити публікації з виробничої системи Toyota, або з системи 6-сигм компанії Motorola). З цих процесів і підходів вибираються для вивчення і порівняння найбільш підходящі. Після чого вони адаптуються для умов своєї власної організації.

**Основні етапи бенчмаркингу включають:**

**1. Визначення, аналіз і деталізація об’єкта бенчмаркінгу.** Як об’єкт можуть бути обрані процеси, послуги або продукти організації. На цьому етапі важливо зрозуміти, скільки ресурсів і зусиль організація готова витратити на процес бенчмаркінгу - чи буде це разовий захід або бенчмаркінг стане постійною практикою організації.

**2. Виявлення і визначення характеристик, за якими буде проводитися бенчмаркінг**. Це можуть бути важливі споживчі властивості продукту або послуги, або параметри якості процесу.

**3. Формування команди бенчмаркінгу.** В команду краще включати фахівців з різних підрозділів організації, щоб була можливість більш широко і об’єктивно оцінити можливості як своїх процесів (продуктів, послуг), так і процесів (продуктів, послуг) партнерів з бенчмаркінгу.

**4. Вибір партнерів по бенчмаркінгу.** Партнерами можуть виступати організації-лідери, які досягли успіху в реалізації характеристик, які цікавлять (визначені на етапі 2). Партнером може бути одна організація або кілька. Якщо виконується внутрішній бенчмаркінг, то такими партнерами будуть суміжні підрозділи, процеси або продукти надаються самою організацією.

**5. Збір та аналіз інформації, необхідної для порівняння.** Щоб провести порівняння може знадобитися представити отриману інформацію в тому ж вигляді, як вона представляється всередині організації. Наприклад, якщо порівнюються технічні характеристики продукту, то у різних виробників набір цих характеристик може відрізнятися. Характеристики необхідно буде привести до єдиної «бази».

**6. Проведення оцінки можливості організації в досягненні необхідних характеристик в порівнянні з партнером (або партнерами) з бенчмаркінгу.** Оцінка може проводитися різними методами, які дозволяють оцінити існуючий «розрив» між роботою власної організації і роботою партнера з бенчмаркінгу (наприклад, за допомогою GAP - аналізу).

**7. Визначення можливих змін існуючої практики роботи.** Створюється «бачення» майбутнього стану організації. Це бачення повинно бути засноване на результатах адаптації процесів партнера з бенчмаркінгу до умов своєї організації.

**8. Розробка стратегічних цілей і планів по їх реалізації для досягнення бажаного рівня характеристик.** Залежно від масштабності змін плани можуть зачіпати зміна процесів, системи управління, організаційної системи, культуру виконання робіт та ін. аспекти.

**9. Реалізація запланованих змін і постійний контроль за ходом перетворень в організації.** Якщо необхідно, то виконуються коригування планів.

**10. Після досягнення встановлених цілей та реалізації планів приймається рішення про повторення циклу і реалізації всіх етапів бенчмаркінгу для нових умов**.

1. **Аналіз ринку (Market Analysis)** – зовнішній аналіз попиту, потреб, конкуренції, дистрибуції, факторів навколишнього середовища, які впливають на попит і поведінку споживачів.

**Опис етапів аналізу ринку:**

**Етап 1.** Визначити цілі і основні завдання аналізу ринку**.**

**Етап 2.** Скласти послідовний план маркетингового аналізу ринку**.**

**Етап 3.** Визначити можливі терміни і максимальний бюджет на аналіз ринку.

**Етап 4.** Визначити методи аналізу ринку і джерела отримання інформації щодо ринку.

**Етап 5.** Провести необхідні маркетингові дослідження ринку товарів або послуг.

**Етап 6.** Підготувати наочний аналіз всієї зібраної інформацією з висновками.

**Етап 7.** Скласти зведений звіт з аналізу ринку.

**Етап 8.** За необхідності підготувати презентацію з проведеного аналізу ринку.

Перед тим, як приступити до аналізу ринку, необхідно визначити предмет аналізу (що будемо аналізувати?) і описати цілі аналізу.

|  |  |
| --- | --- |
| **Предмет аналізу** | **Описання цілей** |
| Структура ринка | Проведення аналізу ємності і кон’юнктури ринку, оцінка ринкових тенденції |
| Товар компанії | Проведення аналізу розвитку ринку і ринкової частки товару компанії в сегменті |
| Цільовий сегмент | Проведення аналізу привабливості сегментів ринку з метою вибору цільового ринку |
| Споживач | Проведення аналізу попиту на ринку і аналізу ключових потреб ринку, докладне вивчення поведінки, вимог цільової аудиторії до продукту |
| Ціни | Проведення аналізу цінового позиціонування конкурентів, діючої структури цін в галузі |
| Вільні ніші | Аналіз сегментів ринку з метою пошуку вільних ринкових ніш і нових джерел продажів |
| Конкуренти | Проведення конкурентного аналізу ринку з метою аналізу конкурентних переваг товару і визначення слабких боку компанії |

**Опис доступних джерел інформації про ринок:**

**Особисті інтерв’ю.** Поговоріть особисто з цільовою аудиторією ринку, проведіть 5-10 інтерв’ю. Включіть в інтерв’ю користувачів різних торгових марок, споживачів і неспоживачів ринку. Опитайте тих, хто приймає рішення і впливає на покупку і тих, хто користується купленим товаром. На таке опитування ви витратите менше тижня і отримаєте багато корисної інформації

**Форуми і соцмережі.** Використовуйте можливості інтернет: можливість запитати споживачів на форумах, в соціальних мережах, по електронній пошті, зв’язатися по Skype - все це знижує витрати на дослідження

**Ресурси інтернет.** Вивчіть наявну інформацію в інтернет щодо необхідної тематики, в тому числі інформацію про суміжних ринках.

**Співробітники компаній.** Опитайте співробітників компанії про питання, які вас цікавлять, дізнайтеся їх думку; окремо поговоріть з представниками відділу збуту. Якщо ви проводите дослідження ринку як незалежна сторона - проведіть інтерв’ю з керівниками компаній.

**Особисте спостереження.** Самі поспостерігайте за поведінкою покупців в місцях продажів: як він робить вибір, як вибирає.

**Особистий досвід.** Спробуйте самі стати покупцем свого продукту і опишіть свої враження.

1. **Бізнес-модель «Канвас»** – являє собою спосіб візуалізації бізнес-моделі, яка за замовчуванням складається з 9 структурних блоків:

* Ключові партнери;
* Ключові види діяльності;
* Ціннісні пропозиції;
* Взаємини з клієнтами;
* Споживчі сегменти;
* Ключові ресурси;
* Канали збуту;
* Структура витрат;
* Потоки надходження доходів.

**Ілюстрація структури бізнес-моделі «Канвас»**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 8. Ключові партнери  Хто наші ключові партнери?  Хто наші ключові постачальники?  Які ключові ресурси ми від них отримали?  Які заходи вони для нас проводять? | 7. Ключові дії  Які ключові дії нам необхідні для роботи?  Для каналів збуту?  Для побудови відносин з клієнтами?  Отримання і обліку потоків доходів? | 2. Ключові цінності  Які проблеми клієнта ми розв’язуємо?  Що цінного в нашій пропозиції?  Які послуги ми можемо запропонувати кожному сегменту споживачів? | | 4. Взаємовідносини з клієнтами  Які у нас відносини з кожним сегментом?  Як вони інтегровані?  Наскільки вони дорогі для нас? | 1. Сегменти споживачів  Для кого ми працюємо?  Хто найбільш важливий клієнт для нас? |
| 6. Ключові ресурси  Які ключові ресурси нам потрібні для створення ключових цінностей?  Наші канали поширення?  Відносини з клієнтами?  Потоки доходів? | 3. Канали  Через які канали наші клієнти хочуть отримувати наші цінності?  Через які канили вони отримують їх зараз?  Які з них найбільш ефективні? |
| 9. Структура витрат  Які найбільш важливі витрати, пов’язані з бізнес-моделлю?  Які ключові ресурси є найбільш дорогими?  Які основні дії є найбільш дорогими? | | | 5. Потоки доходів  За що наші клієнти готові платити?  За що вони платять зараз? Як вони платять?  Яка питома вага кожного з потоків в загальній сумі доходів? | | |

1. ***Metrics and Key Performance Indicators (KPIs)***

**Ціль**

Метою метрик і ключових показників ефективності є оцінка ефективності рішень, компонентів рішення та інших питань, які становлять інтерес для зацікавлених сторін.

**Опис**

**Метрика** – це кількісний рівень показника, який організація використовує для вимірювання прогресу. Показник вказує на певне числове вимір, яке характеризує ступінь прогресу на шляху до досягнення мети, завдання, результату, виконання виду діяльності або додаткового вкладу. Ключовим показником ефективності є той, який вимірює прогрес на шляху до досягнення стратегічної мети або завдання. Звітність – це процес інформування зацікавлених сторін за допомогою метрик індикаторів в зазначеному форматі через певні проміжки часу.

Метрики і звітність є ключовими компонентами моніторингу та оцінки. Моніторинг являє собою безперервний процес збору даних для визначення того, наскільки добре рішення було реалізовано в порівнянні з очікуваними результатами. Оцінка – це систематична і об’єктивна оцінка рішення для визначення його статусу реалізації та ефективності досягнення цілей з плином часу, а також для того, щоб визначити шляхи поліпшення рішення для кращої відповідності цілям. Найбільш пріоритетними напрямками системи моніторингу та оцінки є поставлені цілі і ефекти рішення, а також вхідні ресурси / дані, види діяльності та вихідні результати.

**Елементи**

**1. Indicators (показники)**

Показник (індикатор) визначає конкретний чисельний вимір для мети, впливу, результат, виду діяльності, вхідних ресурсів (матеріали, дані і т.п.). Кожен з цих факторів вигоди (користі) має показник для зміни його належним чином, але деяким факторам може знадобитися кілька показників. Хороший показник задовольняє п’яти характеристикам:

* Clear (Зрозумілість): показник повинен бути точним і однозначним;
* Relevant (Відповідність або релевантність): повинен відповідати вимірюваному фактору;
* Economical (Економічність): витрати на показник повинні бути помірними;
* Adequate (Адекватність): показник повинен забезпечувати достатню основу для оцінки продуктивності;
* Quantifiable (Вимірність): показник може бути незалежно підтверджений.

На додаток до цих характеристик, інтереси зацікавлених сторін можуть бути також важливі. Деякі показники можуть допомагати зацікавленим сторонам виконувати або покращувати процеси більше, ніж інші. Згодом, недоліки в деяких показниках можуть бути ідентифіковані і поліпшені.

Не всі фактори можуть бути виміряні безпосередньо. Непрямі показники можуть використовуватися коли дані для прямих показників не доступні або недоцільно збирати дані на регулярній основі. Наприклад, при відсутності процесів опитування задоволення клієнтів, організація може використовувати в якості непрямого показника частку відновлених (пролонгованих) договорів в загальному обсязі договорів.

При створенні індикатора необхідно враховувати його характеристики: джерело даних, метод збору, хто збирає дані, а також вартість, частоту і складність збору. Вторинні джерела даних можуть бути найбільш економічними, але для задоволення інших характеристик хорошого індикатора може знадобитися проведення первинних досліджень, таких як опитування, інтерв’ю, прямі спостереження. Метод збору даних – це ключовий фактор моніторингу, оцінки та звітності за вартістю системи.

**2. Metrics (Метрики)**

Метрики – це кількісний рівень показника, який оцінюється на вказаний момент часу. Цільова метрика – це завдання, яке повинно бути виконане протягом певного періоду. При налаштуванні метрики (зазвичай однією) для показника важливо мати чітке розуміння базової відправної точки, ресурси, які можуть бути спрямовані на поліпшення чинників, які охоплюються показниками, а також мати чітке розуміння політичних проблем.

Метрика може бути конкретним пунктом, граничним значенням або діапазоном. Діапазон може бути корисним, якщо показник є новим. Рамки часу для досягнення цільової метрики можуть представлятися в багаторічному, річному або квартальному вигляді, або частіше, залежно від необхідності.

**3. Structure (Структура)**

Створення системи моніторингу та оцінки вимагає процедури збору даних, процедури аналізу даних, процедури створення звітності і збору вихідних даних. Процедура збору даних охоплює блоки аналізу, процедури відбору зразків, вибір інструментів збору даних для використання, частоту збору даних і відповідальність за збір даних. Метод аналізу визначає порядок проведення аналізу і споживачів даних, які можуть мати сильну зацікавленість в тому, як проводиться аналіз. Процедура створення звітності охоплює шаблони звітів, одержувачів звітів, частоту створення звітів і засоби комунікації (комунікаційні канали). Вихідна інформація полягає в тому, які дані надаються безпосередньо перед вимірами або на початку періоду вимірювання. Вихідні дані використовуються для того, щоб дізнатися про останні успіхи та оцінки прогресу в відправної точки. Вихідна інформація повинна бути зібрана для кожного показника, який буде аналізуватися або використовуватися в звітності.

Існує три ключові чинники в оцінці якості показників і їх метрик – це надійність, достовірність (валідність) і своєчасність. Надійність – це ступінь, в якій підхід до збору даних є стабільним і послідовним у часі і просторі. Достовірність – це ступінь, в якій дані зрозуміло (чітко) і безпосередньо вимірюють продуктивність, яку організація має намір виміряти. Своєчасність – це відповідність частоти і часу очікування даних, в яких потребує менеджмент.

**4. Reporting (звітність)**

Як правило звіти порівнюють базові, поточні та цільові показники один з одним, розраховані відмінності представлені як в абсолютному вираженні, так і у відносному. У більшості випадків, тенденції є більш авторитетними і важливими, ніж абсолютні показники. Візуальні представлення тренду, як правило, більш ефективно, ніж таблиці, особливо при використанні якісних описів для пояснення даних.

1. ***SWOT Analysis***

**SWOT-аналіз** - це базова аналітична модель пошуку стратегії, яка ґрунтується на загальній оцінці сильних і слабких сторін компанії, а також можливостей і загроз зовнішнього середовища. SWOT-аналіз досить абстрактний, оскільки його висновки носять чисто описовий характер: вони не припускають ніяких рекомендацій і не розставляють пріоритети. Але цього і не потрібно. Його завдання - просто уявити нам структурований опис ситуації, щодо якої потрібно прийняти рішення.

**АНАЛІЗ ЗОВНІШНЬОГО СЕРЕДОВИЩА (МОЖЛИВОСТЕЙ І ЗАГРОЗ).** Бізнес-одиниця повинна постійно відстежувати основні фактори макросередовища (демографічні, економічні, природні, технологічні, політичні, правові, соціальні, культурні), а також значущі елементи мікросередовища (покупці, конкуренти, постачальники, дистриб’ютори, дилери), які впливають на отримання прибутку. Виявити нові тенденції макро- і мікросередовища і відбуваються в них зміни дозволяє створення маркетингової інформаційної системи. Для кожної нової тенденції керівництво компанії має виявити відповідні можливості і загрози.

**Маркетингова можливість -** область купівельних потреб і інтересів, задоволення яких з високою часткою ймовірності принесе компанії прибуток.

**Загрози зовнішнього середовища –** це негативний вплив деяких тенденцій або несприятливий розвиток подій, які за відсутності захисних маркетингових заходів призводять до скорочення обсягів продажів або прибутку компанії.

**АНАЛІЗ ВНУТРІШНЬОГО СЕРЕДОВИЩА (СИЛЬНИХ І слабких сторін).** Оцінка внутрішніх сильних і слабких сторін виконується за наступними напрямками діяльності компанії - організація, виробництво, постачання і логістика, інновації, фінанси, маркетинг, персонал, безпека, матеріально-технічне забезпечення.

На основі можливостей і загроз, сильної і слабкої сторін складаються 4 типи стратегії:

1. Стратегія зіставлення сильних сторін і можливостей;

2. Стратегія зіставлення сильних сторін і загроз;

3. Стратегія зіставлення слабких сторін і можливостей;

4. Стратегія зіставлення слабких сторін і загроз.

1. ***Портрет споживача***

Цільова аудиторія – це люди, які зацікавлені у вашому продукті. Вони вже купували його, хочуть купити повторно, планують покупку, купують суміжні товари і поглядають на ваш, словом, все це реальні та потенційні клієнти вашої компанії.

Портрет клієнта (ЦА) – це збірний образ покупця, що складається з демографічних, соціальних, поведінкових факторів та інших характеристик. Чим детальніше і точніше складено цей портрет, тим точніше буде бити в ціль ваш бізнес.

Однією з найпопулярніших є класика жанру – методика 5W (What, Who, Why, When, Where). Інтерпретуючи відповіді на п’ять основних питань, ви отримаєте досить об’ємну картину:

* У чому зацікавлений клієнт?
* Хто він такий?
* Чому він робить покупку (оцінка мотивації)?
* Коли і в яких умовах клієнт робить покупку?
* Де можна знайти покупців і, відповідно, залучити їх?

Взявши за основу цю просту схему, ми заглибимося в детальне вивчення інформації та її подальше структурування. Питання: “Хто такий наш клієнт?” – має на увазі, що ми повинні знати про нього буквально все – від статі, віку, місця проживання та сімейного стану до захоплень і фобій.

Скористайтеся аналітичними сервісами, щоб зібрати дані клієнтів. Використовуйте Google Analytics, вбудовану аналітику Facebook Audience Insights та інші спеціальні можливості.

Проводьте опитування і уважно аналізуйте відповіді. Намагайтеся максимально точно формулювати відкриті питання в анкетах, щоб отримувати розгорнуті відповіді та коректну інформацію (яку потім ви вдало використаєте в рекламних повідомленнях). Дізнавайтеся чим незадоволені ваші клієнти, що вони хочуть змінити, чого їм взагалі не вистачає для щастя.

Аналізуйте існуючих клієнтів і канали комунікацій для вибору найбільш ефективних. Постійне дослідження аудиторії гарантує глибоке її розуміння.

Займіться “шпигунством” серед конкурентів. Оцінюйте їх рекламні кампанії, переглядайте сторінки в соцмережах, вивчайте контент. На основі отриманих даних будуйте гіпотези і перевіряйте їх на практиці. Сміливо задійте найефективніші методики конкурентів для залучення нових сегментів ЦА.

Розділіть широку цільову аудиторію на більш вузькі сегменти, визначте основний сегмент (який приносить максимум прибутку) і створіть детальний портрет клієнта – встановіть його особистість, риси характеру, відтворити історію життя, щоб повністю перейнятися бажаннями, інтересами і страхами свого клієнта.

1. ***PEST-аналіз*** (іноді позначають як STEP) – це маркетинговий інструмент, призначений для виявлення політичних (P – political), економічних (E – economic), соціальних (S – social) і технологічних (T – technological) аспектів зовнішнього середовища, які впливають на бізнес компанії.

Методика PEST аналізу часто використовується для оцінки ключових ринкових тенденцій галузі, а результати PEST аналізу можна використовувати для визначення списку загроз і можливостей при складанні SWOT аналізу компанії.

Зазвичай виконується у форматі таблиці, що складає 4 квадранти, кожен з яких відповідає за окремий напрямок: політичний, економічний, соціальний, технологічний.

Проводячи PEST аналіз, рекомендовано описувати не лише поточний стан кожного фактора, а прогнозувати його зміни на найближчі 3-5 років. Саме оцінка впливу фактора в довгостроковій перспективі на прибуток компанії, дозволяє застосовувати отримані дані для формування стратегії.

Якщо компанія реалізує свої товари на різних ринках і функціонує в різних галузях - рекомендується проводити PEST аналіз для кожної галузі та для кожного ринку.

PEST – P:

Стійкість політичної влади та існуючого уряду.

Бюрократизація і рівень корупції.

Податкова політика.

Незалежність ЗМІ.

Прагнення до протекціонізму галузі, наявність державних компаній у галузі.

Ступінь захисту інтелектуальної власності, закон про авторське право.

Законодавство з охорони навколишнього середовища.

Майбутнє і поточне законодавство, що регулює правила роботи у галузі.

Можливість розвитку військових дій у країні.

PEST – E:

Темпи зростання економіки.

Рівень інфляції та процентні ставки.

Курси основних валют.

Рівень безробіття, розмір та умови оплати праці.

Рівень розвитку підприємництва та бізнес-середовища.

Кредитно-грошова і податково-бюджетна політика країни.

Ступінь глобалізації та відкритості економіки.

Рівень розвитку банківської сфери.

PEST – S:

Рівень охорони здоров’я та освіти.

Ставлення до імпортних товарів і послуг.

Ставлення до роботи, дозвілля, виходу на пенсію.

Вимоги до якості продукції та рівня сервісу.

Спосіб життя і звички споживання.

Розвиток релігії та інших вірувань.

Ставлення до натуральних і екологічно чистих продуктів.

Темпи зростання населення.

Імміграційні настрої.

Розмір і структура сім’ї.

PEST – T:

Рівень інновацій і технологічного розвитку галузі.

Витрати на дослідження і розробку.

Законодавство у сфері технологічного оснащення галузі.

Розвиток і поширення інтернету, розвиток мобільних пристроїв.

Доступ до новітніх технологій.

Ступінь використання, впровадження та передачі технологій.

***Лекція 4.1. (2 год)***

**Тема 4. Оцінка й обґрунтування вибору організаційно-правової форми бізнесу ОЦІНКА Й ОБҐРУНТУВАННЯ ВИБОРУ ОРГАНІЗАЦІЙНО-ПРАВОВОЇ ФОРМИ БІЗНЕСУ**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

* 1. Які організаційно-правові форми існують
  2. Переваги і недоліки деяких організаційно-правових форм
  3. Як вибрати оптимальну організаційно-правову форму (ОПФ) для ведення бізнесу

Джерело: <https://gc.ua/uk/fop-abo-tov-shho-vibrati/>

***1. ЯКІ ОРГАНІЗАЦІЙНО-ПРАВОВІ ФОРМИ ІСНУЮТЬ***

Згідно зі ст. 42 Господарського кодексу України , підприємництво - це самостійна, ініціативна, систематична, на власний ризик господарська діяльність, що здійснюється суб’єктами господарювання (підприємцями) з метою досягнення економічних і соціальних результатів та одержання прибутку. Підприємці мають право без обмежень самостійно здійснювати будь-яку підприємницьку діяльність, яка не заборонена законом.

Отже, в разі здійснення господарської діяльності систематично, а не одноразово або час від часу, людина стикається з необхідністю легалізувати свою діяльність. Несистематичне отримання доходів внаслідок разового виконання робіт або надання послуг, навіть на підставі письмових договорів, не зобов’язує особу реєструвати офіційно ведення господарської діяльності.

Законодавство надає можливість вибирати організаційно-правові форми ведення господарської діяльності. Можна вести підприємницьку діяльність в наступних формах:

* Фізична особа-підприємець (ФОП);
* Господарські товариства: повне товариство, командитне товариство, товариство з обмеженою (ТОВ) або додатковою відповідальністю, акціонерне товариство;
* Кооперативи;
* Приватні підприємства (ПП);
* Громадські об’єднання (відповідно до Закону України «Про громадські об’єднання» від 22.03.2012 р. їм надається право здійснювати підприємницьку діяльність безпосередньо).

Кожна з цих форм має свої переваги і недоліки на етапі заснування, на етапі становлення і розвитку бізнесу. Для деяких видів бізнесу законодавство визначає обов’язкові вимоги до організаційно-правової форми (наприклад, банки можуть бути тільки публічними акціонерними товариствами або кооперативними банками), але в більшості випадків вибір засновників нічим не обмежений. Для того, щоб використовувати законодавчі пільги щодо обмеження відповідальності та спрощеного оподаткування можна структурувати бізнес і замість однієї компанії зробити кілька. Одна і та ж особа в більшості випадків може бути одночасно засновником і директором необмеженої кількості юридичних осіб, а також бути фізичною особою-підприємцем (за винятком рідкісних обмежень).Головне в такому випадку не укладати угод між різними фірмами за підписом однієї і тієї ж особи, оскільки такі угоди можуть бути оскаржені.

***2. ПЕРЕВАГИ І НЕДОЛІКИ ДЕЯКИХ ОРГАНІЗАЦІЙНО-ПРАВОВИХ ФОРМ***

На етапі створення і становлення бізнесу найпростішими для державної реєстрації та управління і найпопулярнішими сьогодні є ФОП, ТОВ та ПП. Аби не заглиблюватися в їх основні характеристики, які нескладно знайти в законодавстві, перейдемо відразу до аналізу їх позитивних і негативних рис з точки зору «молодого» бізнесу.

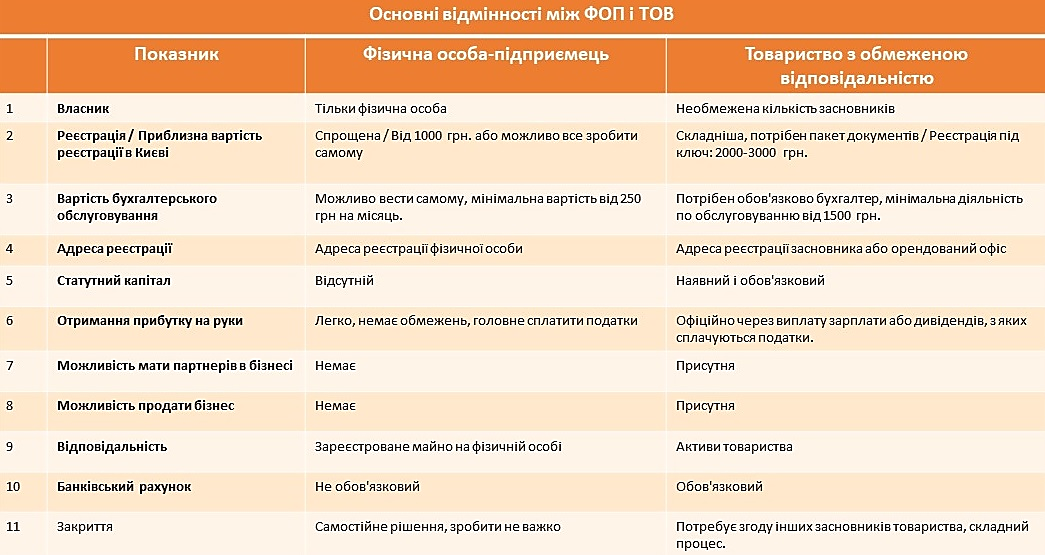
Перевагами такої організаційно - правової форми як ФОП є:

* простота реєстрації,
* знижені податки,
* спрощена система звітності та діловодства, яка не вимагає найму професійного бухгалтера на постійну роботу.
* Недоліками є:
* складність юридичної процедури припинення підприємницької діяльності в результаті її фактичного припинення,
* підвищена відповідальність
* складність залучення в бізнес компаньйонів та інвесторів. Не кожен компаньйон повірить на слово, що з грошей, вкладених в бізнес, він матиме свою частину доходу, тому документально підтвердити це досить складно, а укладання договорів позики та спільної діяльності має суттєві недоліки для самого ФОП.

Що стосується ТОВ, то на сьогодні це одна з найбільш врегульованих форм бізнесу в Україні, наділена широким спектром прав і чітко визначених обов’язків. Законодавство детально регламентує внутрішні відносини в ТОВ, а також відносини з третіми особами. Всі ризики за результати своєї діяльності несе ТОВ, а не його учасники. З червня 2011 мінімального розміру статутного капіталу ТОВ законодавством не встановлено (для порівняння: для акціонерного товариства мінімальний статутний капітал становить 1250 мінімальних заробітних плат; попередньо мінімальний статутний капітал ТОВ встановлювався в розмірі 1 мінімальної зарплати).Також перевагами ТОВ є повне розмежування між відповідальністю ТОВ і відповідальністю його учасників, підвищена відповідальність директора і можливість безперешкодного відсторонення останнього від обов’язків за рішенням загальних зборів засновників; простота залучення зовнішніх інвесторів. ТОВ як аналог іноземної Limited Liability Company буде повністю зрозуміло іноземним партнерам (на відміну від ПП, яке є, на мою думку, пережитком радянського права).

Переваги ПП - це мінімальні законодавчі вимоги до його реєстрації і діяльності, розмежування між відповідальністю ПП і відповідальністю його учасників, відсутність вимог до мінімального статутного капіталу, в певних випадках - можливість залучення сторонніх інвесторів і використання деяких законодавчих переваг.

Але недоліки ПП перевищують його позитивні сторони. Серед них: поверхневе, недостатньо докладне правове регулювання (яке не змінюється так динамічно, як у відношенні ТОВ), неврегульованість в законі проблемних питань діяльності ПП, відсутність достатніх роз’яснювальних джерел судової практики про надзвичайну подію та інші недоліки, які не дають можливості порадити таку організаційно-правову форму в порівнянні з її альтернативою, а саме ТОВ.



***3. ЯК ВИБРАТИ ОПТИМАЛЬНУ ОРГАНІЗАЦІЙНО-ПРАВОВУ ФОРМУ (ОПФ)***

***ДЛЯ ВЕДЕННЯ БІЗНЕСУ***

Відомості про фізичних осіб-підприємців та юридичних осіб включаються до Єдиного державного реєстру, який є відкритим для ознайомлення. У ньому можна в будь-який час отримати основну інформацію про будь-яку юридичну особу або ФОП-контрагента, ці відомості є загальнодоступними.

Основними нормативно-правовими актами, що встановлюють правовий статус юридичних осіб і фізичних осіб-підприємців, які регулюють основи господарської діяльності, з якими обов’язково потрібно ознайомитися кожному підприємцю, а не тільки його юристу, є: Цивільний кодекс України від 16.01.2003, Господарський кодекс України від 16.01.2003, Закон України «Про господарські товариства» від 19.09.1991 р, Закон України «Про товариства з обмеженою та додатковою відповідальністю», Податковий кодекс України від 02.12.2010 р.

Важливо, що 28.08.2011 р вступив в силу Закон «Про внесення змін до деяких законодавчих актів України щодо запровадження принципу державної реєстрації юридичних осіб на підставі модельного статуту», який надає можливість створювати юридичну особу за рішенням її засновника (ів) без затвердження статуту на підставі модельного статуту, затвердженого Кабінетом Міністрів України. Постановою КМ України від 16 листопада 2011 р N 1182 затверджено модельний статут ТОВ .З першого погляду, діяльність при підставі модельного статуту дозволяє спростити і здешевити державну реєстрацію новостворених ТОВ. Однак для розвиненого, ефективного бізнесу необхідно мати власний унікальний статут, який враховує індивідуальні особливості та потреби юридичної особи та її засновників, адже в разі виникнення будь-якого внутрішнього конфлікту або рейдерської атаки саме норми досконало виписаного статуту стануть стовпами, які дозволять утримати Ваш бізнес.

Крім того, зі вступом в силу нового Закону «Про товариства з обмеженою та додатковою відповідальністю» редакція Модельного статуту від 2011 року не відповідає чинному законодавству, тому до затвердження Модельного статуту ТОВ в новій редакції, зареєструвати товариство з обмеженою відповідальністю із застосуванням Модельного статуту або перейти на його використання, не представляється можливим (примітка редакції).

Підсумовуючи вищевикладене, пропонуємо розібратися в перевагах ТОВ та ФОП, а також зрозуміти які ж все таки недоліки вони мають перед один одним.

### **Переваги ФОП**

* До ведення бухгалтерського обліку підприємця не висуваються серйозні законодавчі вимоги, вести свою бухгалтерію підприємець може і самостійно.
* Можливість ведення бізнесу без обов’язкового відкриття рахунку в банку.
* Фізична особа-підприємець, може на свій розсуд розпоряджатися своїми доходами та прибутком.
* ФОП має повне право коли завгодно ліквідувати свою діяльність.
* Ціна на реєстрацію набагато нижча, в порівнянні з ТОВ.

### **Недоліки ФОП**

* Підприємець не може вибирати для свого бізнесу назву.
* Є обмеження по видам діяльності.
* У ситуації, коли діяльність призупинилася, ФОП, в разі, якщо він знаходиться на ЄП, щомісяця повинен продовжувати оплачувати ЄСВ
* Бізнес, який створений «навколо» ФОП не може бути переоформлений на іншого інвестора.
* ***Відповідальність власним майном при банкрутстві.***

### **Переваги ТОВ**

* ТОВ може вибрати будь-яке ім’я, якого нема в державному реєстрі.
* ***Легкий продаж частки чи всього бізнесу***.
* ***Можливість «почати з чистого аркуша»***.

### **Недоліки ТОВ**

* ***Бухгалтерський облік ТОВ вимагає кваліфікованого бухгалтера*** чи кількох.
* За ТОВ податкові органи значно пильніше слідкують та контролюють, ніж за ФОП.
* Якщо бізнес створено в форматі ТОВ, то інвестор не може в будь-який момент вилучити кошти з бізнесу за власним бажанням.
* **Складна ліквідація.**

### **Оподаткування: ФОП або ТОВ?**

Перед початком роботи бізнесу потрібно визначитися з планованими (влаштовують інвестора) рівнями податкових платежів.

Наприклад:

* ФОП платник єдиного податку 2 групи сплачує податок незалежно від обсягів діяльності за ставкою до 20% мінімальної зарплати на місяць. Додатково сплачується ЄСВ «за себе».
* ФОП платник єдиного податку 3 групи, неплатник ПДВ сплачує податок в залежності від обсягів діяльності за ставкою 5% від суми доходів;
* ФОП на загальній системі сплачує 18% податку з доходів, 1.5% військового збору і 22% ЄСВ від суми чистого доходу (прибутку)
* ТОВ-платник єдиного податку, неплатник ПДВ сплачує податок в залежності від обсягів діяльності за ставкою 5% від суми доходів;
* ТОВ на загальній системі сплачує 18% податку на прибуток з суми прибутку, визначеного за даними бухобліку.

Звичайно, вибрати правильну систему оподаткування бізнесу дуже важливо, адже, в подальшому, не завжди її можна поміняти швидко  (перейти на єдиний податок можна тільки з початку кварталу), а в ряді випадків змінити не вийде – ТОВ не зможе використовувати 2 групу єдиного податку як ФОП.

### **Чи потрібен статус платника ПДВ?**

Перед створенням бізнесу необхідно чітко розуміти – потрібен для його успішної роботи статус платника податку на додану вартість (ПДВ). Платником ПДВ може бути як юридична особа, так і підприємець. Однак в ряді випадків система оподаткування ФОП не дозволяє бути одночасно і платником ПДВ (наприклад, 2 група єдиного податку або 3 група зі ставкою 5%).

В основному статус платника ПДВ може бути необхідний у разі якщо:

* Планується в великих обсягах імпортувати товар або обладнання зі сплатою ПДВ на митниці;
* Контрагенти – покупці або замовники готові працювати тільки з платником ПДВ;
* Плановані обсяги операцій за перший рік перевищують 5 000 000 грн.

Статус платника ПДВ – додатковий обсяг бухгалтерської роботи, відповідальності ФОП або ТОВ, а також – істотне податкове навантаження, яке бажано врахувати і прорахувати при прийнятті рішення про реєстрацію платником ПДВ.

### **Заробітна плата у ФОП або ТОВ**

ФОП самостійно веде підприємницьку діяльність на свій страх і ризик. При цьому його винагородою є прибуток від такої діяльності. Заробітної плати він не отримує, хоча і може встановити собі якусь «ставку», яку буде стабільно забирати з бізнесу.

У ТОВ в будь-якому випадку буде керівник, якому повинна регулярно, два рази на місяць виплачуватися заробітна плата в розмірі не менше мінімальної. Звичайно при «припиненні» діяльності ТОВ співробітники можуть бути звільнені, а функції з представництва інтересів ТОВ можуть бути покладені на засновника, проте це вимагає оформлення документально.

Найманим працівникам ТОВ та ФОП виплачують заробітну плату і сплачують податки з неї за рівними правилами.

**Змістовий модуль 3. Економічне обґрунтування бізнес-моделі**

Тема 5. Фінансово-економічна оцінка бізнес-ідеї

Оцінка комерційної ефективності підприємницької ідеї в цілому. Витрати виробництва і реалізації продукції (послуг). Визначення потреби в оборотних коштах (оборотному капіталі). Загальні капіталовкладення (інвестиції). Звіт про прибутки і збитки. Грошові потоки та показники ефективності. Фінансовий профіль бізнес-плану. Оцінка ефективності участі в підприємницькій діяльності. Оцінка фінансової можливості реалізації бізнес-ідеї і розрахунок показників її ефективності.

[1] c. 87-89; [3] c. 177-190; [4] c. 168-190

***Лекція 5.1. (2 год)***

**Тема 5. Фінансово-економічна оцінка бізнес-ідеї**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Оцінка комерційної ефективності підприємницької ідеї в цілому
2. Витрати виробництва і реалізації продукції (послуг)
3. Визначення потреби в оборотних коштах (оборотному капіталі)

***1. ОЦІНКА КОМЕРЦІЙНОЇ ЕФЕКТИВНОСТІ ІНВЕСТИЦІЙНОГО ПРОЄКТУ В ЦІЛОМУ***

Комерційна ефективність проєкту в цілому оцінюється з метою визначення його потенційної привабливості для можливих учасників та пошуків джерел фінансування.

Показники комерційної ефективності проєкту в цьому випадку враховують фінансові наслідки його здійснення для єдиного учасника (організатора), що реалізує інвестиційний проєкт в припущенні, що він робить всі необхідні для реалізації проєкту витрати і користується усіма його результатами.

Показники комерційної ефективності проєкту визначаються на основі грошових потоків, розрахунок яких проводиться з урахуванням даних, що визначаються по кроках розрахункового періоду:

* витрати виробництва і реалізації продукції;
* потреба в оборотних коштах і приріст оборотних коштів;
* загальні капіталовкладення;
* звіт про доходи та витрати і показники рентабельності.

***2. ВИТРАТИ ВИРОБНИЦТВА І РЕАЛІЗАЦІЇ ПРОДУКЦІЇ (ПОСЛУГ)***

При розрахунку витрат виробництва і реалізації продукції (послуг) угруповання витрат не має принципового значення. Однак при виконанні цих розрахунків необхідно дотримуватися таких правил:

* виділяти амортизаційні відрахування в окрему статтю;
* угруповання витрат виробляти таким чином, щоб їх можна було розділити на умовно-постійні та умовно-змінні.

Нижче наведена приблизна форма таблиці для розрахунку витрат виробництва і реалізації продукції (послуг) (табл. 5.1.).

Таблиця 5.1.

Витрати виробництва і реалізації продукції (тис. грн.)

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість (в роках або частках року) |  |  |  |
| Статті:  1. Обсяг виробництва (шт.) |  |  |  |
| 2. Прямі матеріальні витрати (підсумок табл. 12) |  |  |  |
| 3. Витрати на оплату праці виробничих робітників і фахівців, безпосередньо зайнятих виготовленням продукції (підсумок табл. 14) |  |  |  |
| 4. Відрахування на соціальні потреби (% від стор.3) |  |  |  |
| 5. Амортизаційні відрахування (підсумок табл. 15) |  |  |  |
| 6. Податки, які відносять на собівартість |  |  |  |
| 7. Загальновиробничі і загальногосподарські витрати без амортизаційних відрахувань і податків, що відносяться на собівартість |  |  |  |
| 8. Витрати виробництва (сума рядків 2 / 7) |  |  |  |
| 9. Витрати виробництва без амортизації (стр.8 – стр.5) |  |  |  |
| 10. Витрати зі збуту продукції |  |  |  |
| Витрати виробництва і збуту продукції (стр.8 + стр.10) |  |  |  |
| 12. Витрати виробництва і збуту продукції без амортизації (стр.9 + стр.10) |  |  |  |
| 13. Виробнича собівартість одиниці продукції (стр.8 / стр.1) |  |  |  |
| 14. Повна собівартість одиниці продукції (стр. 11 / стр.1) |  |  |  |
| 15. Умовно-постійні витрати |  |  |  |
| 16. Умовно-змінні витрати на одиницю продукції |  |  |  |

Витрати по збуту продукції складаються з наступних статей витрат:

* витрати на тару і упаковку виробів (на складах відділу збуту);
* витрати на транспортування продукції;
* витрати на заробітну плату працівників, зайнятих збутом продукції (вантажно-розвантажувальні роботи та ін.), і відрахування на соціальні потреби;
* комісійні відрахування збутовим організаціям;
* гарантійний ремонт і обслуговування;
* витрати на рекламу;
* витрати на участь у виставках;
* підготовка продавців і торговців даною продукцією;
* інші витрати по збуту.

***3. ВИЗНАЧЕННЯ ПОТРЕБИ В ОБОРОТНИХ КОШТАХ*** ***(ОБОРОТНОМУ КАПІТАЛІ)***

***Загальні положення***

Необхідність досить докладного обліку оборотного капіталу визначається декількома факторами.

Це в першу чергу :

* обсяг оборотного капіталу, що залежить від типу інвестиційного проєкту;
* рівень прогнозованої інфляції;
* ступінь невизначеності термінів надходження необхідних матеріалів та оплати готової продукції.

У зв’язку зі специфікою процесу виробництва і обігу існують певні особливості в складі і структурі оборотних коштів підприємств різних сфер діяльності.

У підприємств видобувних галузей практично відсутні сировину і основні матеріали, покупні напівфабрикати, значна частка допоміжних матеріалів, витрат майбутніх періодів (витрати на гірничо-підготовчі роботи). Наприклад, більшість інвестиційних проєктів у газовій промисловості (крім деяких, пов’язаних з газопереробних заводах, підземними сховищами газу і ін.) Не пов’язані з помітним обсягом оборотного капіталу і тому не вимагають докладної оцінки потреби в ньому. При розробці таких проєктів досить використовувати спрощені оцінки.

На підприємствах обробних галузей значними є виробничі запаси. В силу тривалості виробничого циклу висока питома вага незавершеного виробництва.

До складу оборотних виробничих фондів сільськогосподарських підприємств входять: молодняк тварин, тварини на вирощуванні і відгодівлі, корми, насіння, запасні частини, паливно-мастильні матеріали, добрива, допоміжні матеріали, посіви озимих культур, зяб. Найбільшу питому вагу мають виробничі запаси.

Основним елементом оборотних активів у будівельних організацій є виробничі запаси і незавершене будівельне виробництво. До складу виробничих запасів входять запаси будівельних конструкцій, деталей, блоків, будівельних матеріалів.

Продукція підприємств транспорту не має речового виразу, і тому відсутній такий елемент оборотних активів, як незавершене виробництво. Оборотні виробничі фонди включають в себе допоміжні матеріали, запасні частини, паливо, мастильні матеріали.

Оборотні виробничі фонди підприємств торгівлі представлені запасами товарів, допоміжними матеріалами. Фонди обігу включають кошти в розрахунках, грошові кошти в касі та на рахунках в банку.

Особливістю діяльності організацій сфери науки є тривалий характер виконання науково-дослідних і дослідно-конструкторських робіт. Висока частка заробітної плати в складі витрат, при відносно невеликому питомій вазі матеріальних витрат. У зв’язку з цим у складі оборотних виробничих фондів значну питому вагу має незавершене виробництво.

Розрахунки потреби в оборотному капіталі при розробці інвестиційних проєктів та оцінки їх ефективності дещо відрізняються від аналогічних бухгалтерських розрахунків, що обумовлено відмінностями в обліку витрат і результатів, а також необхідністю більш точного обліку фактора часу .

Розрахунки потреби в оборотних коштах представляють певну складність і вимагають великого обсягу вихідної інформації на початкових стадіях розробки проєкту. Тому для попередніх розрахунків або в разі, коли оборотний капітал малий і істотного впливу не робить, потреба в ньому може визначатися укрупнено, наприклад, як певний відсоток від середньомісячних (на даному етапі) чистих операційних витрат або собівартості. Значення такого відсотка може визначатися за даними аналогічних діючих підприємств. При розрахунках оборотного капіталу окремі його складові можуть не враховуватися, якщо фахівець, який здійснює розрахунок, визнає, що враховувати їх годі було, і призведе обґрунтування своєї думки.

На рис. 5.1. наведені методи розрахунку оборотного капіталу для діючих і новостворюваних підприємств.

Методи розрахунку оборотного капіталу

Для діючих підприємств

Як % до собівартості чи операційним витратам

Для новостворюваних підприємств

Коефіцієнтний

Аналітичний

Методи прямого рахунку

Визначення нормованої частини ОЗ

Різниця між оборотними активами і пасивами

Рис. 5.1. Методи розрахунку оборотного капіталу

**Аналітичний метод визначення потреби в оборотних засобах**

Аналітичний метод передбачає визначення потреби в оборотних коштах на основі їх середньо фактичних залишків з урахуванням зміни обсягу виробництва.

Умова застосування аналітичного і коефіцієнтних методів - наявність сформованої виробничої програми, статистичних даних за минулі періоди про зміну залишків нормованих оборотних активів.

Потреба в оборотних коштах визначається за формулою

де – оборотні кошти планованого виробництва; – среднефактіческіх залишки оборотних коштів у базовому періоді; – обсяг виробництва в планованому періоді; – обсяг виробництва в базовому періоді.

**Коефіцієнтний метод визначення потреби в оборотних засобах**

Коефіцієнтний метод передбачає поділ оборотних активів на три групи:

* оборотні активи, потреба в яких залежить від обсягів виробництва (сировина, основні матеріали, паливо, готова продукція, комплектуючі вироби і т.д.) (У.пер.Об);
* оборотні активи, потреба в яких не знаходиться в прямій залежності від обсягів виробництва (запасні частини, енергія і паливо для утримання основних засобів і т.д.) (У.пос.Об);
* оборотні активи, потреба в яких визначається за особливим розрахунком (витрати майбутніх періодів) ().

За першою групою потреба в оборотних коштах у планованому періоді (У.пер.ОбП) визначається виходячи з розмірів у базисному періоді (У.пер.ОбБ) і темпів зростання обсягів виробництва в планованому періоді:

;

По другій групі потреба в оборотних коштах планується на рівні їх середньо фактичних залишків за кілька періодів (У.пос.ОбП).

По третій групі потреба в оборотних коштах визначається з урахуванням особливостей кожного елемента (витрати майбутніх періодів ().

Таким чином, потреба в оборотних коштах у планованому періоді можна визначити за формулою

**Визначення потреби в оборотних коштах як відсоток до собівартості або операційним витратам**

Як вже було сказано вище, для попередніх розрахунків або в разі, коли оборотний капітал малий і істотного впливу на грошові потоки не робить, потреба в ньому може визначатися укрупнено, як певний відсоток від середньомісячних (на даному етапі) чистих операційних витрат або собівартості. Значення такого відсотка може визначатися експертно виходячи з даних аналогічних діючих підприємств. При цьому, однак, слід враховувати, що на момент введення в експлуатацію такий розрахунок може привести до великих помилок, особливо за такими статтями балансу, як дебіторська заборгованість, запаси готової продукції, незавершене виробництво, розрахунки з бюджетом і позабюджетними фондами, розрахунки з персоналом.

**Визначення потреби в оборотних коштах методом розрахунку їх нормованої частини**

Цей метод прямого рахунку для новостворюваних підприємств передбачає наявність вихідної інформації, наведеної в табл. 5.2.

Таблиця 5.2.

Вихідна інформація для розрахунку потреби в оборотних коштах

(нормованої частини)

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість (в роках або частках року) |  |  |  |
| Статті  1. Обсяг виробництва, V (шт.) |  |  |  |
| 2. Середньодобовий випуск продукції, VСД  (стр. 1 / тривалість кроку в календарних днях) |  |  |  |
| 3. Вартість сировини, матеріалів, комплектуючих виробів на одиницю продукції, См (грн.) |  |  |  |
| 4. Норма запасу сировини, матеріалів, комплектуючих виробів, НЗМ (в календарних днях) |  |  |  |
| 5. Виробнича собівартість одиниці продукції, Sвироб. (грн.) |  |  |  |
| 6. Тривалість виробничого циклу, Тц (в календарних днях) |  |  |  |
| 7. Коефіцієнт наростання витрат, Кнв (визначається за статистичними даними підприємств, які виробляють подібну продукцію. При відсутності статистичних даних розраховується за формулою  (3) |  |  |  |
| 8. Норма запасу готової продукції, Нз.г.п (в календарних днях) |  |  |  |

Потреба в оборотних коштах по кроках розрахунку (О) визначається як сума активів нормованої частини оборотних коштів:

(5.4.)

де – запаси сировини, матеріалів, комплектуючих виробів,

(5.5.)

(Позначення див. в табл. 2);

– витрати в незавершеному виробництві,

(5.6.)

– запаси готової продукції,

(5.7.)

Результати розрахунків зводяться в табл. 5.3.

Таблиця 5.3.

Потреба в оборотних коштах і приріст оборотних коштів (тис. грн.)

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість (в роках або частках року) |  |  |  |
| статті:  1. Запаси сировини, матеріалів і комплектуючих виробів () |  |  |  |
| 2. Витрати в незавершеному виробництві () |  |  |  |
| 3. Запаси готової продукції () |  |  |  |
| 4. Потреба в оборотних коштах (сума рядків 1 : 3) |  |  |  |
| 5. Приріст оборотних коштів |  |  |  |

**Визначення потреби в оборотних коштах як різниці їх активів і пасивів**

**Загальні положення**

***Оборотний капітал являє собою різницю між оборотними активами і оборотними пасивами.***Збільшення оборотного капіталу є частиною інвестиційних витрат (відтоків) ІП, зменшення оборотного капіталу - частиною інвестиційних притік.

Оскільки виручка від реалізації продукції зазвичай надходить не з самого початку реалізації ІП і не безперервно з плином часу, виробництво потребує деяких запасах оборотних коштів для задоволення своїх короткострокових потреб. З необхідністю цих запасів і пов’язані ***оборотні активи***.

Якщо обсяг і витрати виробництва не змінюються, оборотні активи по кроках розрахунку залишаються постійними і їх приріст дорівнює нулю; в іншому випадку вони або збільшуються (позитивний приріст), або зменшуються (негативний приріст).

***Оборотні пасиви***пов’язані з тим, що платежі, пов’язані з виробництвом продукції в певний момент часу, здійснюються не в той же, а в більш пізній момент часу, тобто з деякою затримкою (наприклад, заробітна плата працівникам виплачується не щодня, а два або один раз на місяць, тобто затримується в середньому на одну або два тижні; витрачена електроенергія оплачується також не щодня, а раз на місяць і т.д. ). У межах зазначеної затримки відповідні кошти можуть бути використані з метою покриття короткострокової заборгованості, і зокрема для (повного або часткового) покриття потреби в оборотних активах.

Таким чином, оборотні пасиви (за винятком передоплати) – це не джерело коштів, а можливість їх економії.

**Початкові дані**

Вихідні дані для розрахунку потреби в оборотному капіталі представлені в табл. 5.4.

Слід мати на увазі, що зазначені в таблиці норми можуть бути взаємопов’язаними. Наприклад, частина оборотних активів на створення запасів комплектуючих може бути врахована за рядком 5 «Аванси постачальникам за послуги», рядок 3 «Готова продукція» і рядок 4 «Дебіторська заборгованість» також можуть виявитися пов’язаними між собою (хоча, як, буде показано нижче, неідентичних).

Оборотний капітал двояко впливає на ефективність і можливість реалізувати ІП.

По-перше, позитивний приріст оборотного капіталу на кожному кроці є елементом грошового відтоку (негативний – притоки) від інвестиційної діяльності і тому впливає як на ефективність, так і на можливість реалізувати проєкту (або на потребу у фінансуванні).

Таблиця 5.4.

Вихідна інформація для розрахунку потреби в оборотному капіталі

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Номер рядка | Показники | | Номер кроку | | |
| Структура оборотного капіталу | Норми\* (в днях) | 0 | 1 | … |
| 1 | **Сукупні активи:**  Сировина, матеріали, комплектуючі та ін. \*\* | Страховий запас  Періодичність поставок |  |  |  |
| 2 | Незавершене виробництво | цикл виробництва |  |  |  |
| 3 | Готова продукція | періодичність відвантаження |  |  |  |
| 4 | Дебіторська заборгованість | затримка платежів |  |  |  |
| 5 | Аванси постачальникам за послуги | Термін передоплати  Частка ціни, оплачувана попередньо |  |  |  |
| 6 | Резерв коштів | покриття потреби |  |  |  |
|  | **Пасиви (розрахунки з кредиторами)** |  |  |  |  |
| 7 | Розрахунки за товари, роботи і послуги | затримка платежів |  |  |  |
| 8 | Авансові платежі (передоплата) | Термін передоплати  Частка платежів, оплачуваних попередньо |  |  |  |
| 9 | Розрахунки з оплати праці | Періодичність виплат |  |  |  |
| 10 | Розрахунки з бюджетом та позабюджетними фондами | Періодичність виплат |  |  |  |
| 11 | Погашення та обслуговування позик, орендні (лізингові) платежі | Періодичність виплат |  |  |  |

Залежно від величини запасів і норм зберігання.

По-друге, приріст деяких елементів оборотного капіталу, що відносяться до даного кроку розрахунку (запаси сировини, матеріалів і комплектуючих, запаси готової продукції, дебіторська заборгованість, передоплата, кредиторська заборгованість), відбувається не одночасно з іншими надходженнями і витратами, що впливає на ефективність проєкту за рахунок зміни коефіцієнта дисконтування і зміни цін (інфляція, сезонні ціни і т.д.). У випадках, коли цей вплив виявляється помітним, його необхідно враховувати.

Стандартний бухгалтерський розрахунок враховує тільки перший вид впливу (на потребу у фінансуванні). Для розрахунку ефективності необхідно враховувати обидва види впливу.

Спеціаліст, який розраховує інвестиційний проєкт, повинен обґрунтувати вибір вихідних даних для розрахунку потреб в оборотному капіталі і визначити, які з його складових необхідно враховувати в даному проєкті, а які не грають ролі і тому можуть бути опущені.

**Формули для розрахунку потреби в оборотному капіталі**

Розрахунок потреб в оборотному капіталі для кожного кроку розрахункового періоду проводиться за наведеними нижче формулами (для спрощення номер кроку в них опущений).

**Оборотні активи за статтями:**

1) «Сировина, матеріали, комплектуючі та ін.»

(5.8.)

де ВМ – витрати на матеріали даного виду на даному етапі; Р1 – тривалість кроку в днях; d – величина страхового запасу в днях (стор. 1 табл. 4); q – періодичність поставок в днях (стр.1 табл. 5.4.);

2) «Незавершене виробництво»

(5.9.)

де Зпр – сума прямих витрат (прямі матеріальні витрати + витрати на оплату праці основного і допоміжного виробничого персоналу з нарахуваннями) на кроці; Тц – тривалість виробничого циклу в днях (стор.2 табл. 5.4.);

3) «Готова продукція»

(5.10.)

де В0 – виручка без ПДВ на кроці; г0 – періодичність відвантаження в днях (стор.3 табл. 5.4.);

4) «Дебіторська заборгованість» (інша уживане назва: «Рахунки до отримання»)

(5.11.)

де Вир[[1]](#footnote-1) – виручка на кроці, що включає ПДВ в складі виручки або (і) експортні тарифи та інші податки, що нараховуються на виручку[[2]](#footnote-2); – величина затримки платежів в днях (стор. 4 табл.5.4.);

5) «Аванси постачальникам за послуги»

(5.12.)

де Пст – вартість послуг сторонніх організацій (постачальників) на кроці; Ру – частка передоплати постачальникам (стор. 5 табл. 5.5.); Су – термін передоплати послуг в днях (стор. 5 табл. 5.5.);

6) «Резерв коштів»

(5.13.)

де Впс – витрати на виробництво і збут за винятком прямих матеріальних витрат на кроці; s – покриття потреби в коштах в днях (стор. 6 табл. 5.4.).

**Оборотні пасиви за статтями:**

1) «Розрахунки за товари, роботи і послуги» (інша уживане назва: «Кредиторська заборгованість»)

(5.14.)

де ВИТ – прямі матеріальні витрати + величина відкладених виплат стороннім організаціям; ВП – відстрочка платежів у днях[[3]](#footnote-3) (стор. 7 табл. 5.4.);

2) «Авансові платежі (передоплата)»

(5.15.)

де Вир – виручка, що залишається у підприємства на кроці (після сплати ПДВ, акцизів і імпортних зборів); рпр – частка передоплати: відсоток від виручки в частках (стор. 8 табл. 4); Тп – термін передоплати за реалізовану продукцію в днях (стор. 8 табл. 4);

3) «Розрахунки з оплати праці»

(5.16.)

де ЗП - загальна зарплата за крок розрахунку; k – періодичність виплат: число виплат зарплати в місяць;

4) «Розрахунки з бюджетом та позабюджетними фондами» представляють собою суму пасивів за статтями:

* нарахування на зарплату;
* виплати по ПДВ, що вноситься до бюджету;
* виплати по податку на прибуток;
* виплати за іншими податками (зборами).

По кожному з податків (зборів, нарахувань) відповідна величина оборотних пасивів (ППі) визначається за формулою

(5.17.)

де ВП – величина податку (збору), що відноситься до даного кроку; ПВ – періодичність виплат цього податку (збору) в днях.

Загальна величина оборотних пасивів за статтею «Розрахунки з бюджетом та позабюджетними фондами» визначається підсумовуванням розрахованих значень за всіма видами податків (зборів, нарахувань);

5) «Розрахунки за кредитами, позиками, оренді та лізингу» є сумою оборотних пасивів за кожним кредитом (позикою, з орендною угодою, договором лізингу). Цей рядок визначається тільки при розрахунку з урахуванням схеми фінансування.

Оборотні пасиви Пк за кожним кредитом (позикою, з орендною угодою, договором лізингу) визначаються за формулою

(5.18.)

де ПРк – величина процентної виплати за кредитом (позикою) орендного або лізингового платежу на кроці; ПВк – періодичність цієї виплати в днях.

Загальна величина оборотних пасивів по новому рядку «Розрахунки за кредитами, позиками, оренді та лізингу» визначається підсумовуванням розрахованих значень за всіма кредитами (позиками, орендним або лізинговим договорами).

**Особливості розрахунків потреби в оборотному капіталі при оцінці різних видів ефективності**

**Комерційна ефективність проєкту**

При визначенні комерційної ефективності проєкту розрахунок потреби в оборотному капіталі проводиться за формулами (5.8.) – (5.17.) без урахування залучених джерел фінансування (позики, оренда, лізинг).

Як уже зазначалося, окремі складові оборотного капіталу можуть не враховуватися, якщо фахівець, який здійснює розрахунок, визнає, що враховувати їх годі було, і призведе обгрунтування своєї думки.

Результати розрахунків представляються у вигляді табл. 5.5.

Таблиця 5.5.

Розрахунок приросту оборотного капіталу при оцінці комерційної ефективності

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Номер рядки | Структура оборотних коштів для розрахунку комерційної ефективності | Номер кроку розрахунку | | |
| 0 | 1 | 2 |
| 1 | **Активи**  Сировина, матеріали, комплектуючі та ін. |  |  |  |
| 2 | Незавершене виробництво |  |  |  |
| 3 | Готова продукція |  |  |  |
| 4 | Дебіторська заборгованість |  |  |  |
| 5 | Аванси постачальникам за послуги |  |  |  |
| 6 | Резерв коштів |  |  |  |
| 7 | Разом: активи (сума стр.1 6) |  |  |  |
| 8 | **Пасиви**  Розрахунки за товари, роботи і послуги |  |  |  |
| 9 | Авансові платежі (передоплата) |  |  |  |
| 10 | Розрахунки з оплати праці |  |  |  |
| 11 | Розрахунки з бюджетом та позабюджетними фондами |  |  |  |
| 12 | Разом: пасиви (сума стор.8 11) |  |  |  |
| 13 | Оборотний капітал (стор.7 – стор.12) |  |  |  |
| 14 | Приріст оборотного капіталу |  |  |  |

При розрахунку ефективності акціонерного капіталу потреба в оборотному капіталі визначається так само, як і при розрахунку комерційної ефективності проєкту, але з наступними відмінностями:

* при розрахунку оборотних активів по рядку «Дебіторська заборгованість» в величині ВИП, що входить в формулу (11), враховуються всі складові, в тому числі і значення платежів за оренду, лізинг і за позиками;
* до оборотних пасивів додається новий рядок – «Розрахунки за кредитами, позиками, оренді та лізингу», яка розраховується за формулою (18). Нижче наводиться таблиця для розрахунку приросту оборотного капіталу (табл. 5.6).

Таблиця 5.6.

Розрахунок приросту оборотного капіталу при оцінці ефективності акціонерного капіталу

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Номер рядки | Структура оборотних коштів для розрахунку ефективності акціонерного капіталу | Номер кроку розрахунку | | | |
| 0 | 1 | 2 | … |
| 1 | **Активи**  Сировина, матеріали, комплектуючі та ін. |  |  |  |  |
| 2 | Незавершене виробництво |  |  |  |  |
| 3 | Готова продукція |  |  |  |  |
| 4 | Дебіторська заборгованість |  |  |  |  |
| 5 | Аванси постачальникам за послуги |  |  |  |  |
| 6 | Резерв коштів |  |  |  |  |
| 7 | Разом: активи (сума стр.1 6) |  |  |  |  |
| 8 | **Пасиви**  Розрахунки за товари, роботи і послуги |  |  |  |  |
| 9 | Авансові платежі (передоплата) |  |  |  |  |
| 10 | Розрахунки з оплати праці |  |  |  |  |
| 11 | Розрахунки з бюджетом та позабюджетними фондами |  |  |  |  |
| 12 | Розрахунки за кредитами, позиками, оренді та лізингу |  |  |  |  |
| 13 | **Разом: пасиви (сума стор.8 + 12)** |  |  |  |  |
| 14 | **Оборотний капітал (стор.7 – стор.13)** |  |  |  |  |
| 15 | **Приріст оборотного капіталу** |  |  |  |  |

***Лекція 5.2. (2 год)***

**Тема 5. Фінансово-економічна оцінка бізнес-ідеї**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Загальні капіталовкладення (інвестиції)
2. Звіт про прибутки і збитки
3. Грошові потоки та показники ефективності
4. Фінансовий профіль бізнес-плану

***1. ЗАГАЛЬНІ КАПІТАЛОВКЛАДЕННЯ (ІНВЕСТИЦІЇ)***

Загальні капіталовкладення складаються з наступних статей витрат:

* основні засоби;
* нематеріальні активи;
* інші (некапіталізовані) роботи і витрати;
* пуско-налагоджувальні роботи;
* заміщення основних засобів і нематеріальних активів;
* приріст оборотних коштів.

Детальніше склад капіталовкладень наведено в табл. П 1.2 додатка 1. Використовуючи табл. 6, 7, 8, 15 бізнес-плану, 5.3., 5.5. даного розділу, а також графік реалізації проєкту, складається табл. 5.7. загальних капіталовкладень (інвестицій).

Таблиця 5.7.

Загальні капіталовкладення (інвестиції)

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість (в роках або частках року) |  |  |  |
| Статті:  1. Основні засоби (стор.10 табл.6 БП) |  |  |  |
| 2. Нематеріальні активи (стр.6 табл.7 БП) |  |  |  |
| **3. Разом: витрати на необоротні активи (стр.1 + стор. 2)** |  |  |  |
| 4. Інші (некапіталізовані) роботи і витрати (стор.4 табл.8 БП) |  |  |  |
| 5. Пусконалагоджувальні роботи (стор.5 табл. 8 БП) |  |  |  |
| 6. Заміщення основних засобів і нематеріальних активів (стр.3 табл.15 БП) |  |  |  |
| 7. Приріст обігових коштів (підсумок табл.3 або 5) |  |  |  |
| **8. Загальні капіталовкладення (інвестиції) (сума стор.3 : 7)** |  |  |  |

***2.*** ***ЗВІТ ПРО ПРИБУТКИ І ЗБИТКИ***

Раніше проведені розрахунки зводимо в табл. 5.8.

При необхідності табл. 8 доповнюється показниками рентабельності, розрахованими щодо балансового прибутку, чистого прибутку або доходу:

* рентабельності продажів;
* рентабельності продукції;
* рентабельності активів;
* рентабельності основного капіталу і т.д.

Таблиця 5.8.

Звіт про прибутки і збитки (фінансові результати)

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість кроку (роки, частки року) |  |  |  |
| Показники:  1. Обсяг продажів (шт.) |  |  |  |
| 2. Ціна одиниці |  |  |  |
| **3. Виручка (стр.1 х стор.2) + інші і позареалізаційні доходи** |  |  |  |
| 4. Витрати виробництва і збуту без амортизації (стор.11 табл.1) |  |  |  |
| 5. Амортизаційні відрахування (стор.5 табл.1) |  |  |  |
| **6. Витрати виробництва і збуту продукції (стор.4 + стор.5)** |  |  |  |
| 7. Прибуток балансова (стр.3 – стр.6) |  |  |  |
| 8. Податки і збори з балансового прибутку |  |  |  |
| 9. База податку на прибуток (стор.7 – стор.8) |  |  |  |
| 10. Податок на прибуток |  |  |  |
| **11. Прибуток чиста (стор.9 – стор.10)** |  |  |  |
| **12.Прибуток чистий наростаючим підсумком** |  |  |  |
| **13. Дохід (стор.11 + стор.5)** |  |  |  |
| **14. Дохід наростаючим підсумком** |  |  |  |

***3. ГРОШОВІ ПОТОКИ ТА ПОКАЗНИКИ ЕФЕКТИВНОСТІ***

**Грошовий потік від інвестиційної діяльності**

У грошовий потік від інвестиційної діяльності в якості відтоку включаються, перш за все, розподілені по кроках розрахункового періоду витрати по створенню і введенню в експлуатацію нових основних фондів і ліквідації, заміщенню вибувають існуючих основних фондів. Сюди ж відносяться некапіталізовані витрати (наприклад, сплата податку на земельну ділянку, що використовується в ході будівництва, витрати по будівництву об’єктів зовнішньої інфраструктури та ін.). Крім того, в грошовий потік від інвестиційної діяльності включаються зміни оборотного капіталу (збільшення розглядається як відтік грошових коштів, зменшення – як приплив). Як відтоку включаються також власні кошти, вкладені на депозит, а також витрати на покупку цінних паперів інших господарюючих суб’єктів, призначені для фінансування даного інвестиційного проєкту. Як припливу в грошовий потік від інвестиційної діяльності включаються доходи від реалізації вибувають активів. У цьому випадку, однак, необхідно передбачити сплату відповідних податків.

Відомості про інвестиційні витрати повинні включати інформацію, класифіковану за видами витрат.

Оцінка витрат на придбання окремих видів основних фондів може проводитися також на основі результатів оцінки відповідного майна. Розподіл інвестиційних витрат по періоду будівництва має бути ув’язано з графіком будівництва.

Орієнтовна форма подання інформації про капітальні вкладення представлена в табл. 7 (позиції в рядках таблиці при необхідності можуть бути деталізовані). При підготовці інформації враховуються такі обставини:

1. Відсотки за кредит, взятий на фінансування будівництва об’єктів, що сплачуються до введення об’єктів в експлуатацію, в вартість об’єкта не включаються, а враховуються окремо і тільки при оцінці ефективності проєкту в цілому.
2. Обсяги витрат заносяться в таблицю в поточних цінах у валюті, в якій вони здійснюються.
3. У таблицю заносяться як початкові капіталовкладення, так і наступні, в тому числі на рекультивацію земель після початку експлуатації і на заміну вибуває обладнання, яка визначається на підставі його термінів служби, які можуть не кореспондуватися з нормами амортизації.
4. На останніх кроках розрахунку в складі капітальних вкладень повинні враховуватися витрати, пов’язані з ліквідацією підприємства, включаючи витрати на демонтаж обладнання, захист і відновлення довкілля і т.д. (Здійснення таких витрат може займати кілька кроків).
5. Величину доходів від продажу основних фондів у разі припинення проєкту рекомендується визначати за даними прогнозної оцінки. Вона може не збігатися з залишковою вартістю такого майна.

Терміни (кроки розрахунку) введення в дію, вартість, знос і структура основних фондів показуються окремо по пускових комплексах (черг).

Дані про вартість основних засобів рекомендується наводити в тій валюті, в якій здійснюються витрати, і в підсумковій валюті в поточних цінах, з урахуванням прогнозної оцінки.

Орієнтовна форма запису вихідної інформації для розрахунку потреби в оборотних коштах і приросту оборотного капіталу наведена в табл. 5.3., 5.5. За діючим підприємствам–учасникам проєкту додатковою інформацією є їх баланси за останні звітні квартали і за останній рік.

Орієнтовна таблиця розрахунку грошових потоків від інвестиційної діяльності наведена нижче (табл. 5.9.).

Таблиця 5.9.

Грошовий потік від інвестиційної діяльності

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість кроку (роки, частки року) |  |  |  |
| Показники  **1. Грошові притоки всього,**  в тому числі:   1. Доходи (за вирахуванням податків) від іншої реалізації (основних фондів) 2. Доходи (за вирахуванням податків) від реалізації основних засобів і нематеріальних активів у разі припинення ІП (ліквідаційна вартість) 3. Повернення (в кінці ІП) оборотних активів (ліквідаційна вартість) |  |  |  |
| **2. Грошові відтоки всього,**  в тому числі:   1. Загальні капіталовкладення (інвестиції) (табл. 7) 2. ліквідаційні витрати 3. Вкладення коштів в додаткові фонди (депозит, витрати на покупку цінних паперів) |  |  |  |
| **3. Сальдо грошового потоку від інвестиційної діяльності (стр.1 – стр.2)** |  |  |  |

**Грошовий потік від операційної діяльності**

Основним результатом операційної діяльності є отримання прибутку на вкладені кошти. Відповідно в грошових потоках при цьому враховуються всі види доходів і витрат, пов’язаних з виробництвом і реалізацією продукції, і податки, що сплачуються з зазначених доходів. Зокрема, тут враховуються притоки коштів за рахунок надання власного майна в оренду, доходів за депозитними вкладами і з цінних паперів.

**Обсяги виробництва і реалізації продукції та інші доходи**

Обсяги виробництва рекомендується вказувати в натуральному і вартісному виразах.

Ціни на вироблену продукцію, передбачені в проєкті, повинні враховувати вплив реалізації проєкту на загальний обсяг пропозиції даної продукції (і, отже, на ціни цієї продукції) на відповідному ринку.

Джерелом інформації є передпроєктні і проєктні матеріали, дослідження російського і зарубіжних ринків, що підтверджуються, наприклад, міжурядовими угодами, угодами про наміри, укладених договорів та ін.

Вихідна інформація для визначення виручки від продажу продукції задається за підсумками розрахунку для кожного виду продукції, окремо для реалізації на внутрішньому і зовнішньому ринках. Орієнтовна форма подання такої інформації представлена в табл. П1.1 додатка 1.

Крім виручки від реалізації в притоках і відтоках реальних грошей необхідно враховувати доходи і витрати від позареалізаційних операцій, безпосередньо не пов’язаних з виробництвом продукції. До них, зокрема, відносяться:

* доходи від здачі майна в оренду, або лізингу (якщо ця операція не є основною діяльністю);
* надходження коштів за депозитними вкладами і за придбаними цінними паперами інших господарюючих суб’єктів;
* повернення позик, представлених іншим учасникам.

**Витрати на виробництво і збут продукції**

Орієнтовна форма подання вихідної інформації для розрахунку поточних витрат на виробництво і збут продукції показана в табл. П1.3 – П1.8 додатка 1 і в табл. 5.1.

Джерелами інформації є передпроєктні і проєктні матеріали.

Для кожного виду споживаних при реалізації проєкту ресурсів ціни повинні бути обґрунтовані. У разі необхідності слід враховувати вплив проєкту на загальний обсяг попиту на цей вид ресурсів (і, отже, на його ціну) на відповідному ринку.

Всі показники рекомендується вказувати без ПДВ та інших податків і зборів, що включаються в ціну.

У разі, якщо підприємство здійснює кілька видів діяльності, щодо яких установлено різні ставки податків, доходи і витрати по кожному з таких видів діяльності визначаються окремо.

**Розрахункова таблиця грошового потоку від операційної діяльності**

Орієнтовна таблиця розрахунку грошових потоків від операційної діяльності наводиться нижче (табл. 5.10.).

Таблиця 5.10.

Грошовий потік від операційної діяльності

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість кроку (роки, частки року) |  |  |  |
| **показники**  **1. Грошові притоки всього,**  в тому числі:   1. Виручка від реалізації продукції 2. Інші і позареалізаційні доходи |  |  |  |
| **2. Грошові відтоки всього,**  в тому числі:   1. Витрати виробництва і збуту продукції без амортизаційних відрахувань (стор.11 табл. 1) 2. Податки і збори (стор.8 + стор.11 табл.8) 3. позареалізаційні витрати |  |  |  |
| **3. Сальдо грошового потоку від операційної діяльності (стр.1 – стр.2)** |  |  |  |

**Розрахункові таблиці для оцінки комерційної ефективності інвестиційного проєкту в цілому**

Нижче наводиться зведена таблиця грошових потоків для оцінки комерційної ефективності проєкту (табл. 5.11.).

Таблиця 5.11.

Грошові потоки для оцінки комерційної ефективності проєкту в цілому

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість кроку (роки, частки року) |  |  |  |
| **Показники**  **Операційна діяльність 1. Грошові притоки (стр.1.1 + стор. 1.2)**   1. Виручка від реалізації продукції 2. Інші і позареалізаційні доходи |  |  |  |
| **2. Грошові відтоки (сума рядків 2.1 + 2.3)**  2.1. Витрати виробництва і збуту продукції без амортизаційних відрахувань  2.2. Податки і збори   1. 2.3. Позареалізаційні витрати |  |  |  |
| **3. Сальдо грошового потоку від операційної діяльності (стр.1 – стр.2)** |  |  |  |
| **Інвестиційна діяльність**  **4. Грошові притоки** (ліквідаційна вартість) |  |  |  |
| **5. Грошові відтоки** (стр.5.1 + стр.5.2)   1. Загальні капіталовкладення (інвестиції) 2. Вкладення коштів в додаткові фонди |  |  |  |
| **6. Сальдо грошового потоку від інвестиційної діяльності (стор.4 – стор.5)** |  |  |  |
| **7. Сальдо двох потоків (чисті грошові надходження проєкту) (стор.3 + стр.6)** |  |  |  |
| 8. Те ж наростаючим підсумком () |  |  |  |
| 9. Коефіцієнт дисконтування |  |  |  |
| 10. Чиста поточна вартість (стор.7 х стор.9) |  |  |  |
| 11. Чиста поточна вартість наростаючим підсумком () |  |  |  |

Для ряду інвестиційних проєктів суворо розмежувати потоки з різних видів діяльності може виявитися скрутним. У цих випадках, а також на попередніх стадіях інвестиційного проєктування можна об’єднати деякі (або всі) потоки .

Розрахункові показники об’єднаних грошових потоків для оцінки комерційної ефективності проєкту в цілому наведені в табл. 5.12.

Таблиця 5.12.

Модель дисконтування грошових потоків для оцінки комерційної ефективності інвестиційного проєкту в цілому

|  |  |  |  |
| --- | --- | --- | --- |
| Номер кроку | 0 | 1 | … |
| Тривалість кроку (роки, частки року) |  |  |  |
|  |  |  |  |
| **Показники**  **1. Грошові притоки (стр.1.1 + стр.1.2)**   1. Виручка від реалізації продукції + інші позареалізаційні доходи 2. Ліквідаційна вартість активів |  |  |  |
| **2. Грошові відтоки (сума рядків 2.1 + 2.5)**   1. Загальні капіталовкладення (інвестиції) 2. Витрати виробництва і збуту продукції без амортизації 3. Податки і збори 4. позареалізаційні витрати 5. Вкладення коштів в додаткові фонди |  |  |  |
| **3. Чисті грошові надходження (стр.1 – стр.2)** |  |  |  |
| **4. Чисті грошові надходження наростаючим підсумком** |  |  |  |
| 5. Коефіцієнт дисконтування при нормі дисконту Е |  |  |  |
| 6. Чиста поточна вартість (стор.3 х стор.5) |  |  |  |
| **7. Інтегральний економічний ефект (чиста поточна вартість наростаючим підсумком)** |  |  |  |

***Показники комерційної ефективності інвестиційного проєкту в цілому***

За розрахунковими даними таблиць грошових потоків (табл. 5.11. або 5.12.) визначаємо такі показники ефективності:

* максимальний грошовий відтік (або величина зовнішнього фінансування, необхідна для реалізації проєкту);
* термін окупності капіталовкладень без дисконтування;
* термін окупності капіталовкладень з дисконтуванням;
* кумулятивні чисті грошові надходження;
* інтегральний економічний ефект;
* внутрішня норма прибутковості;
* Індекс прибутковості дисконтованих інвестицій.

За вказаними показниками можна судити про привабливість (доцільності) реалізації проєкту.

**Оцінка стійкості проєкту**

Оцінка стійкості та ефективності проєкту в умовах невизначеності і ризику може бути проведена наступними методами:

1. метод укрупненої оцінки стійкості;
2. метод розрахунку рівнів беззбитковості;
3. метод варіації параметрів (сценаріїв);
4. метод оцінки очікуваного ефекту проєкту з урахуванням кількісних характеристик невизначеності.

***4. ФІНАНСОВИЙ ПРОФІЛЬ ПРОЄКТУ***

Графічне відображення динаміки показників кумулятивних грошових надходжень, показників чистої поточної вартості наростаючим підсумком при нормі дисконту Е, чистої поточної вартості наростаючим підсумком при нормі дисконту Евн (рівній внутрішній нормі прибутковості) являє собою так званий фінансовий профіль проєкту (Cumulative Cash Flow Diagram). Як правило, він має форму, представлену на рис. 2.

Такі узагальнюючі показники, як чисті грошові надходження, максимальний грошовий відтік, термін окупності капіталовкладень, інтегральний економічний ефект, внутрішня норма прибутковості отримують за допомогою фінансового профілю проєкту наочну графічну інтерпретацію.

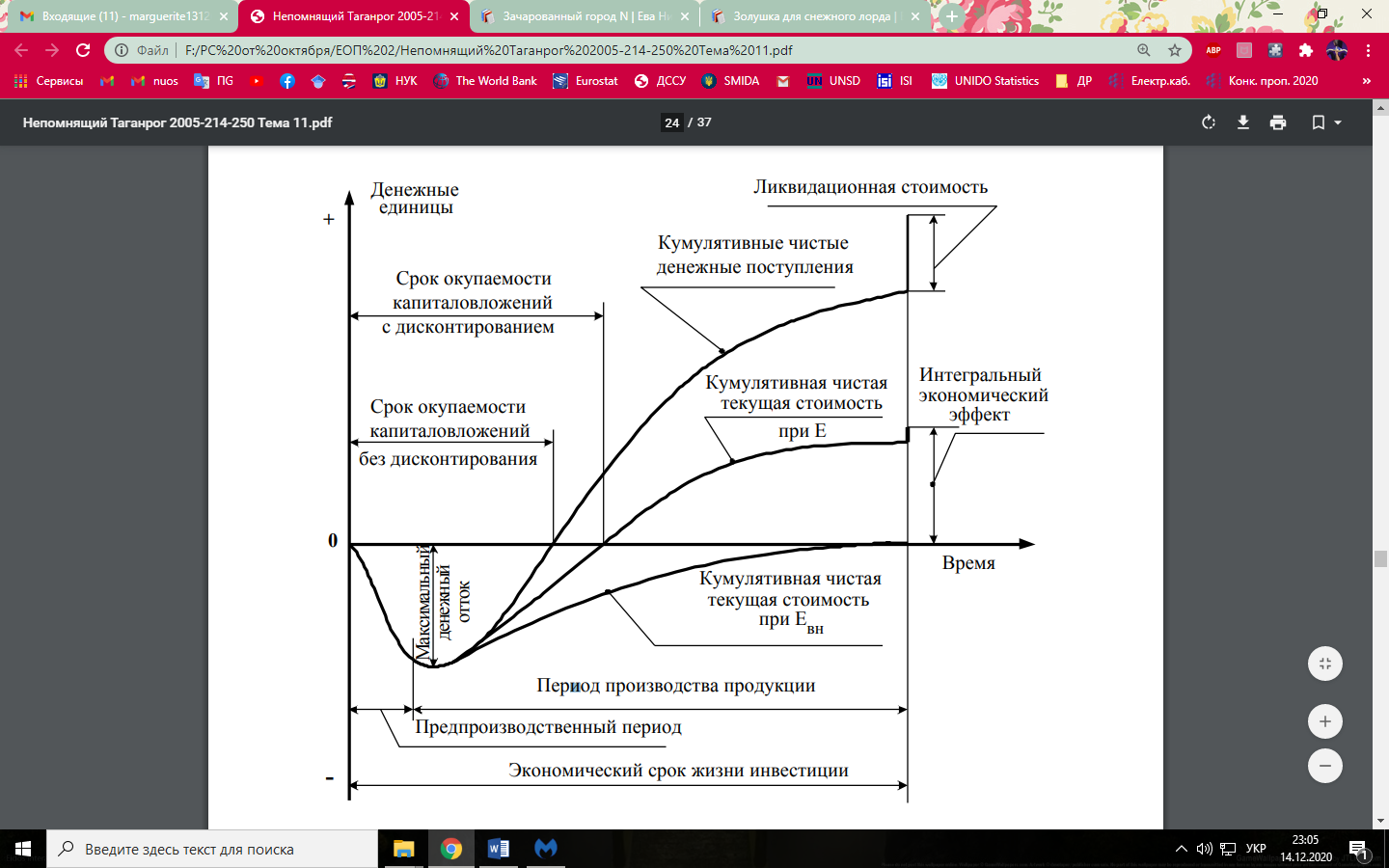


Рис. 5.2. Фінансовий профіль проєкту

Тема 6. Основні показники ефективності підприємницької діяльності та методи їх оцінки

Класифікація показників ефективності підприємницької діяльності. Прості методи оцінки доцільності інвестицій. Чисті грошові надходження. Максимальний грошовий відтік. Показники ефективності підприємницької діяльності. Чиста поточна вартість. Індекс прибутковості інвестицій. Внутрішня норма прибутковості. Термін окупності інвестицій з урахуванням дисконтування. Максимальний грошовий відтік (потреба у фінансуванні з урахуванням дисконту, ДПФ).

[1] c. 120-292; [4] c. 85-96, 101-123

***Лекція 6.1. (2 год)***

**Тема 6. Основні показники ефективності підприємницької діяльності та методи їх оцінки**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Класифікація показників ефективності підприємницької діяльності
2. Прості методи оцінки доцільності інвестицій
3. Чисті грошові надходження
4. Максимальний грошовий відтік
5. Показники ефективності підприємницької діяльності
6. Чиста поточна вартість
7. Індекс прибутковості інвестицій
8. Внутрішня норма прибутковості
9. Термін окупності інвестицій з урахуванням дисконтування
10. Максимальний грошовий відтік (потреба у фінансуванні з урахуванням дисконту, ДПФ)

***1. КЛАСИФІКАЦІЯ ПОКАЗНИКІВ ЕФЕКТИВНОСТІ ПІДПРИЄМНИЦЬКОЇ ДІЯЛЬНОСТІ***

Міжнародна практика обґрунтування інвестиційних проєктів використовує кілька показників, що дозволяють підготувати рішення про доцільність (недоцільність) вкладення коштів.

Ці показники можна об’єднати в дві групи:

1. Показники, що не припускають використання концепції дисконтування:

– простий термін окупності інвестицій;

– показники простий рентабельності інвестицій; – чисті грошові надходження; –індекс прибутковості інвестицій; максимальний грошовий відтік.

1. Показники, що визначаються на підставі використання концепції дисконтування:

– чиста поточна вартість;

– Індекс прибутковості дисконтованих інвестицій; – внутрішня норма прибутковості;

– термін окупності інвестицій з урахуванням дисконтування; – максимальний грошовий відтік з урахуванням дисконтування.

Класифікація основних показників ефективності інвестиційних проєктів приведена в табл. 6.1.

Таблиця 6.1

Основні показники ефективності інвестиційних проєктів

|  |  |  |  |
| --- | --- | --- | --- |
| Абсолютні показники | | Відносні показники | Тимчасові показники |
| Метод  наведеної  вартості | Метод ануїтету | Метод  рентабельності | Метод ліквідності |
| Способи, засновані на застосуванні концепції дисконтування | | | |
| Інтегральний економічний ефект (чиста поточна вартість, NPV) | Дисконтований річний економічний ефект (AN PV) | Внутрішня норма прибутковості (JRR) індекс прибутковості інвестицій | Термін окупності інвестицій з урахуванням дисконтування |
| Спрощені (рутинні) способи | | | |
|  | Приблизний ануїтет | Показники простий рентабельності індекс прибутковості інвестицій | Приблизний (простий) термін окупності інвестицій |

***2. ПРОСТІ МЕТОДИ ОЦІНКИ ДОЦІЛЬНОСТІ ІНВЕСТИЦІЙ***

Прості (рутинні) методи оцінки інвестицій широко використовувалися ще до того, як концепція дисконтування грошових потоків отримала загальне визнання в якості способу отримання самої точної оцінки прийнятності інвестицій. Однак вони залишаються в арсеналі розроблювачів і аналітиків інвестиційних проєктів, оскільки дають можливість отримати деяку додаткову інформацію, що, в свою чергу, дозволяє знижувати ризики невдалого вкладення грошових коштів.

***Простий термін окупності інвестицій***

**Простим терміном окупності інвестицій** (Payback period) називається тривалість періоду від початкового моменту до моменту окупності. Початковим моментом зазвичай є початок першого кроку або початок операційної діяльності. Моментом окупності називається той найбільш ранній момент часу в розрахунковому періоді, після якого кумулятивні поточні чисті грошові надходження NV (k) стають і надалі залишаються невід’ємними.

Метод розрахунку терміну окупності РР інвестицій полягає у визначенні того терміну, який знадобиться для відшкодування суми початкових інвестицій. Якщо сформулювати суть цього методу більш точно, то він припускає обчислення того періоду, за який кумулятивна сума (сума наростаючим підсумком) грошових надходжень порівнюється з сумою початкових інвестицій.

Формула розрахунку терміну окупності має вигляд

(6.1)

де РР –термін окупності інвестицій (років);  – початкові інвестиції;  – середньорічні грошові надходження від реалізації інвестиційного проєкту.

Простий термін окупності є широко використовуваним показником для оцінки того, відшкодуються чи початкові інвестиції протягом терміну їх економічного життєвого циклу інвестиційного проєкту.

Хоча простий термін окупності інвестицій розрахувати досить просто, що, ймовірно, і є причиною популярності цього показника, при його застосуванні виникає ряд проблем. Показник передбачає відшкодування початкових інвестицій на так званій «взносовой» основі, тобто в даному випадку мається на увазі питання: «Скільки часу пройде, поки ми не отримаємо свої гроші назад?».

Просто відшкодувати капітал буде недостатньо, тому що з економічної точки зору інвестор додатково сподівається отримати прибуток на вкладений в справу капітал.

**Приклад 6.1.**

Припустимо, що для придбання нового обладнання необхідні кошти в сумі 100 тис. грн., Які забезпечать щорічне отримання грошових надходжень після сплати податків в сумі 25 тис. грн. протягом шести років без істотних щорічних коливань. Хоча обладнання після шести років експлуатації не буде повністю зношене, тим не менш, навряд чи можливо припускати, що на цей момент часу його вартість буде перевищувати вартість брухту. Витрати на ліквідацію будуть відшкодовані за рахунок виручки від продажу брухту. Лінійна амортизація за ці шість років (16667 грн. за рік) буде відповідно включена в суму грошових надходжень 25 тис. грн.

Щоб оцінити цей проєкт в самому першому наближенні, досить прикинути, чи покриє поточна (сучасна) вартість майбутніх грошових надходжень ті витрати, які фірмі доведеться здійснити зараз. Реально це означає, що нам слід визначити чисту поточну вартість, яку фірма отримає від реалізації такого проєкту.

При проведенні розрахунків будемо виходити із ставки прибутковості (норми дисконту) на рівні 10% річних.

Результати розрахунків зведені в табл. 6.2.

Таблиця 6.2.

Розрахунок поточної вартості грошових потоків

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Період часу (роки) | Інвестиції (грн.) | Грошові надходження | Поточна вартість 1 грн. грошових потоків різних років (коефіцієнти дисконтування при ставці 10%) | Поточна вартість грошових потоків різних років, грн. |
| 0 (сьогодні) | 100000 |  | 1,000 | –100000 |
| 1 |  | 25000 | 0,909 | +22725 |
| 2 |  | 25000 | 0,826 | +20650 |
| 3 |  | 25000 | 0,751 | +18775 |
| 4 |  | 25000 | 0,683 | +17075 |
| 5 |  | 25000 | 0,621 | +15525 |
| 6 |  | 25000 | 0,564 | +14100 |
| Чиста поточна вартість проєкту | | | | +8850 |

Як показали розрахунки, для даного проєкту чиста поточна вартість виявилася позитивною. Іншими словами, здійснення інвестиційного проєкту призвело до зростання капіталу фірми на 8850 грн. в сучасному обчисленні. Отже, інвестиції виявилися корисними і привели до зростання цінності фірми.

У нашому простому прикладі видно, що окупність відбудеться імовірно за 4 роки при оціненого строку економічного життєвого циклу інвестицій 6 років



Хоча простий термін окупності інвестицій розрахувати досить просто, що, ймовірно, і є причиною популярності цього показника, при його застосуванні виникає ряд проблем. Показник передбачає відшкодування початкових інвестицій на так званій «взносовой» основі, тобто в даному випадку мається на увазі питання: «Скільки часу пройде, поки ми не отримаємо свої гроші назад?».

Просто відшкодувати капітал буде недостатньо, тому що з економічної точки зору інвестор додатково сподівається отримати прибуток на вкладений в справу капітал.

Проведемо аналогію з ощадним рахунком, на який покладено 100 тис. грн., з яких 25 тис. грн. вилучаються в кінці кожного року. Після 4 років основна сума буде виплачена з рахунку. Однак вкладник буде розчарований, якщо дізнається, що в даному випадку його рахунок повністю вичерпаний. Очікувана норма прибутку на вкладені інвестиції становить 6 або 8 відсотків на рік по зменшуваного залишку коштів на рахунку. Інвестор буде, наприклад, додатково вимагати виплати накопиченого відсотка.

**Приклад 6.1 (**продовження). У нашому прикладі з інвестиціями в нове обладнання термін окупності розрахований виходячи з початкового припущення про те, що не намічається отримання будь-якого прибутку на вкладені кошти. Чотири роки якраз досить для того, щоб передбачити отримання прибутку. Справді, якщо термін економічного життєвого циклу і термін окупності точно збігаються, то це означає, що інвестор зазнає потенційний збиток, оскільки ті ж самі грошові кошти, інвестовані будь-яким іншим чином, ймовірно, приносили б певний прибуток щороку – принаймні на рівні не нижче банківського відсотка по депозитах.

Це показано в табл. 6.3. Тут знову передбачається, що інвестиції капіталу в сумі 100 тис. грн. забезпечують щорічні надходження грошових коштів в сумі 25 тис. грн. Якщо підприємство зазвичай заробляє від своїх інвестицій 10% річних після сплати податків, то частина щорічних грошових надходжень повинна розглядатися в якості цієї нормальної норми прибутку, в той час як залишок (по графі 5 табл. 6.3.) буде показувати скорочення залишку заборгованості.

Таблиця 6.3.

Відшкодування інвестицій в сумі 100 тис. Грн. при 10% річних

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Рік | Вхідний залишок | Нормативна прибуток | Грошові надходження від поточної господарської діяльності | Заключний залишок, що підлягає відшкодуванню |
| 1 | 2 | 3 | 4 | 5 |
| 1 | 100000 | 10000 | 25000 | 85000 |
| 2 | 85000 | 8500 | 25000 | 68500 |
| 3 | 68500 | 6850 | 25000 | 50350 |
| 4 | 50350 | 5035 | 25000 | 30385 (окупність) |
| 5 | 30385 | 3039 | 25000 | 8424 |
| 6 | 8424 | 842 | 25000 | 15734 (прибуток) |

Друга колонка показує входить баланс (залишок) інвестицій для кожного року. Нормативна прибуток при нормі 10% розрахована на основі цих залишків у третій колонці. Грошові надходження від поточної господарської діяльності показані в четвертій колонці. Сума вхідного залишку інвестицій і нормативного прибутку, зменшена на суму надходжень грошових коштів від поточної господарської діяльності, дозволяє отримати залишок на кінець кожного року. В результаті графік відшкодування інвестицій перевищить 5 років, що більше, ніж передбачався нами термін окупності капітальних вкладень.

З даних табл. 6.3. видно, що окупність терміном на 4 роки означає можливість отримання збитку в сумі приблизно 30 тис. грн., Якщо проєкт завершиться в цей термін. При терміні економічного життя в 5 років можливі втрати, при нормативній прибутку виходячи з 10% річних скоротяться до 8400 грн., В той час як при 6 роках буде отримано прибуток у сумі 15700 грн.

Цей простий приклад показує найбільш істотні недоліки показника простий окупності:

– він не пов’язаний з економічним терміном життя інвестицій і тому не може бути реальним критерієм прибутковості. З нашого прикладу випливає висновок, що проєкт окупиться за 4 роки плюс два додаткових роки для отримання прибутку. Більш того, показник окупності дасть той же результат «чотири роки плюс щось понад» по іншим проєктам з такою ж річною сумою надходжень грошових коштів, але з 8- або 10-річним терміном економічного життя інвестицій;

– інший недолік показника простий окупності полягає в тому, що він внутрішньо передбачає однаковий рівень щорічних грошових надходжень від поточної господарської діяльності. Проєкти зі зростаючими або падінням надходженнями грошових коштів не можуть належним чином бути оцінені за допомогою цього показника. Інвестиції в новий продукт, наприклад, можуть приносити грошові надходження, які будуть повільно зростати на ранніх стадіях, але які далі, на наступних стадіях економічного життєвого циклу продукту зростають більш стрімко. Заміна машин, навпаки, зазвичай буде породжувати постійно приріст операційних витрат, у міру того, як існуюча машина буде зношуватися. Більш того, будь-які додаткові наступні інвестиції протягом періоду або відшкодування капіталу в кінці економічного життєвого циклу будуть викликати невідповідності в цьому показнику.

Табл. 6.3. ілюструє нееластичність показника простий окупності до різних варіацій грошових надходжень.

Якщо припускати однаковий ступінь ризику для кожного з трьох проєктів, показаних в табл. 3, то проєкт 2 є кращим у порівнянні з проєктом один з-за більш тривалого економічного терміну життя інвестицій. Це дозволяє отримати прибуток на 50 тис. грн. більше, ніж за проєктом 1.

Проєкт 3, з іншого боку, може здатися найкращим, якщо судити тільки по показнику окупності. Хоча видно, що проєкт 3 допускає можливість втрат, тому що грошові надходження від господарської діяльності протягом 3-річного періоду здійснення проєкту є достатніми лише для того, щоб відшкодувати початкові витрати капіталу, не забезпечуючи отримання прибутку. Відмінності в поведінці грошових надходжень в проєктах один і 2 також затушовуються при використанні критерію окупності вкладень. Хоча обидва проєкти передбачають окупність початкових інвестицій за 4 роки, проте, загальна (накопичена) сума грошових надходжень від поточної господарської діяльності за проєктом 2 вище, ніж за проєктом 1. Середньорічні грошові надходження становлять 35000 грн. проти 25000 грн. протягом перших 4 років. Отже, проєкт 2 забезпечує більш високу суму грошових надходжень від господарської діяльності і тому більш привабливий для інвестора.

Таблиця 6.4.

Показники простий окупності при різних умовах здійснення проєктів

(показники в грн.)

|  |  |  |  |
| --- | --- | --- | --- |
| Показники | Проєкт 1 | Проєкт 2 | Проєкт 3 |
| Початкові інвестиції | 100000 | 100000 | 100000 |
| Середньорічні грошові надходження | 25000 | 25000 | 33333 |
| Економічний тривалість життя інвестицій | 6 років | 8 років | 3 роки |
| Простий термін окупності інвестицій | 4 роки | 4 роки | 3 роки |
| Щорічні грошові надходження: | | | |
| 1 | 25000 | 25000 | 16667 |
| 2 | 25000 | 30000 | 33333 |
| 3 | 25000 | 50000 | 50000 |
| 4 | 25000 | 40000 | 0 |
| 5 | 25000 | 30000 | 0 |
| 6 | 25000 | 15000 | 0 |
| 7 | 0 | 10000 | 0 |
| 8 | 0 | 5000 | 0 |
| Всього: | 150000 | 200000 | 100000 |
| Підсумок за перші чотири роки | 100000 | 140000 | \* |
| В середньому за перші чотири роки | 25000 | 35000 | \* |

\* – не визначаються.

*Модифікацією показника простий окупності є показник, який використовує в знаменнику величину середнього чистого прибутку (тобто після сплати податків) замість загальної суми надходжень грошових коштів після сплати податків*

(6.2)

де  – середньорічна чистий прибуток.

Перевага цього показника полягає в тому, що приріст прибутку на вкладені інвестиції не враховує щорічних амортизаційних відрахувань.

**Приклад 6.1** (Продовження). У нашому прикладі середньорічні грошові надходження після сплати податків склали 25000 грн., Включаючи середньорічну чистий прибуток в сумі 8333 грн. і щорічну амортизацію в сумі 16667 грн. (Амортизація 100000 грн. За лінійним способом при терміні корисного використання 6 років). Використання цього значення середньорічної чистого прибутку в даній формулі (6.2) показує, що проста окупність збільшиться до 12 років, відображаючи очікувану економічну вигоду:



Цей показник показує, за скільки років буде отримано чистий прибуток, яка за сумою буде дорівнює величині спочатку авансованого капіталу.

Показник простий окупності інвестицій завоював широке визнання завдяки своїй простоті і легкості розрахунку навіть тими фахівцями, які не володіють фінансової підготовкою.

Користуючись показником простий окупності, треба завжди пам’ятати, що він добре працює тільки при справедливості таких припущень:

1. все зіставляються з його допомогою інвестиційні проєкти мають однаковий економічний термін життя;
2. всі проєкти передбачають разове вкладення початкових інвестицій;
3. після завершення вкладення коштів інвестор починає отримувати приблизно однакові щорічні грошові надходження протягом усього періоду економічного життя інвестиційних проєктів.

Використання показника простої окупності як одного з критеріїв оцінки інвестицій має серйозну підставу поряд з простотою розрахунку та ясністю для розуміння. Цей показник досить точно сигналізує про ступінь ризикованості проєкту.

Резон тут простий: керуючі вважають, що чим більший термін потрібен хоча б для повернення інвестованих сум, тим більше шансів на несприятливий розвиток ситуації, здатне перевернути все попередні аналітичні розрахунки. Крім того, чим коротше термін окупності, тим більше грошові надходження в перші роки реалізації інвестиційного проєкту, а значить, і кращі умови для підтримування ліквідності фірми.

*Поряд із зазначеними достоїнствами метод розрахунку простого терміну окупності має дуже серйозними недоліками, так як ігнорує три важливих обставини*:

1. *відмінність цінності грошей у часі;*
2. *існування грошових надходжень і після закінчення терміну окупності;*
3. *різні за величиною грошові надходження від господарської діяльності по роках реалізації інвестиційного проєкту.*

Саме тому розрахунок терміну окупності не рекомендується використовувати як основний метод оцінки прийнятності інвестицій. До нього доцільно звертатися тільки заради отримання додаткової інформації, що розширює уявлення про різні аспекти оцінюваного інвестиційного проєкту.

**Показники простої рентабельності інвестицій**

Показник розрахункової норми прибутку (Accounting Rate of Return) є зворотним за змістом терміну окупності капітальних вкладень.

***Розрахункова норма прибутку відображає ефективність інвестицій у вигляді процентного відношення грошових надходжень до суми початкових інвестицій:***

 (6.3)

де  – розрахункова норма прибутку інвестицій;  – середньорічні грошові надходження від господарської діяльності;  – вартість початкових інвестицій.

Цьому показнику притаманні всі недоліки, властиві показнику терміну окупності. Він бере до уваги тільки два критичних аспекту, інвестиції та грошові надходження від поточної

господарської діяльності та ігнорує тривалість економічного терміну життя інвестицій.

**Приклад 6.1** (Продовження).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Розрахункова |  | Середньорічні грошові надходження |  |  |  |  |  |
| рентабельність | = | від господарської діяльності | = | 25000 | 100% | = | 25% |
| інвестицій |  | початкові інвестиції |  | 100000 |

Цей показник говорить про те, що для отримання щорічних грошових надходжень норма прибутковості на інвестиції в сумі 100000 грн. повинна становити 25%. Це єдине, що можна сказати про проєкт, тому що тут відсутня інформація про тривалість економічного терміну життя інвестицій і не передбачається нарахування відсотків на що залишається невідшкодування суму інвестицій. Зауважимо, що, використовуючи розрахунок норми прибутку, буде отримано однакове рішення в різних ситуаціях, тобто, коли період економічного терміну життя інвестицій складе 1, 10 або 20 років. Насправді прибутковість, відображена у формулі (6.3), буде мати економічний сенс тільки якщо інвестиції, щоб забезпечити отримання 25000 грн., Будуть безперервними. Тільки тоді ми зможемо сказати, що прибутковість дійсно становить 25%*.*

**Різновидом показника розрахункової рентабельності інвестицій є показник, де в якості чисельника у формулі (6.3) варто середньорічна чистий прибуток (після сплати податків, але до процентних платежів):**

 (6.4)

де  – середньорічна чистий прибуток[[4]](#footnote-4).

Розрахункова рентабельність інвестицій може бути визначена також на основі середньорічної валового прибутку () До сплати процентних і податкових платежів:

 (6.5)

Що стосується вартості первинних інвестицій, по відношенню до якої визначається рентабельність, то вона може в формулах (6.4) і (6.5) мати два значення:

– первісна вартість;

– середня вартість між вартістю на початок  і кінець  розрахункового періоду.

Звідси формула для розрахунку рентабельності інвестицій матиме вигляд

 (6.6)

 (6.6а)

ARR, розрахована за формулами (6.4) – (6.6а), буде мати різні значення, тому при підготовці або аналізі інвестиційного проєкту необхідно обумовлювати, за якою методикою цей показник розраховується.

Застосування показників розрахункової рентабельності ґрунтується на зіставленні його розрахункового рівня зі стандартними для фірми рівнями рентабельності.

Використання ARR донині в багатьох фірмах і країнах світу пояснюється рядом переваг цього показника.

**По перше**, Він простий і очевидний при розрахунку, а також не вимагає використання таких витончених прийомів, як дисконтування грошових потоків.

**По-друге**, Показник ARR зручний для вбудовування його в систему стимулювання керівного персоналу фірм. Саме тому ті фірми, які пов’язують системи заохочення керівників своїх філій і підрозділів з результативністю їх інвестицій, звертаються до ARR. Це дозволяє задати керівникам середньої ланки, систему орієнтирів інвестиційної діяльності, яку вони легко піднімають.

Слабкості ж показника розрахункової рентабельності інвестицій є зворотним боком його достоїнств.

**По перше**, Так само, як показник періоду окупності, ARR не враховує рівноцінності грошових коштів у часі, оскільки кошти, що надходять, скажімо на 10 рік після вкладення коштів, оцінюються за тим самим рівнем рентабельності, що і надходження в першому році.

**По-друге**, Цей метод ігнорує відмінності в тривалості експлуатації активів, створених завдяки інвестуванню.

**По-третє**, Розрахунки на основі ARR носять більш «вітринний» характер, ніж розрахунки на основі показників, які використовують дані про грошові потоки. Останні показують реальну зміну цінності фірми в результаті інвестицій, тоді як ARR орієнтована переважно на отримання оцінки проєктів, адекватної очікуванням і вимогам акціонерів та інших осіб і фірм «з боку».

***3. ЧИСТІ ГРОШОВІ НАДХОДЖЕННЯ***

**Чистими грошовими надходженнями** (Net Value, NV) (інші назви – ЧДП, чистий дохід, чистий грошовий потік) називається накопичений ефект (сальдо грошового потоку) за розрахунковий період:

(6.7)

де  –приплив грошових коштів на m-му кроці;  – відтік грошових коштів на m-му кроці.

Підсумовування поширюється на всі кроки розрахункового періоду. Для оцінки ефективності інвестиційного проєкту за перші  кроків розрахункового періоду рекомендується використовувати показник поточних чистих грошових надходжень (накопиченого сальдо):

(6.8)

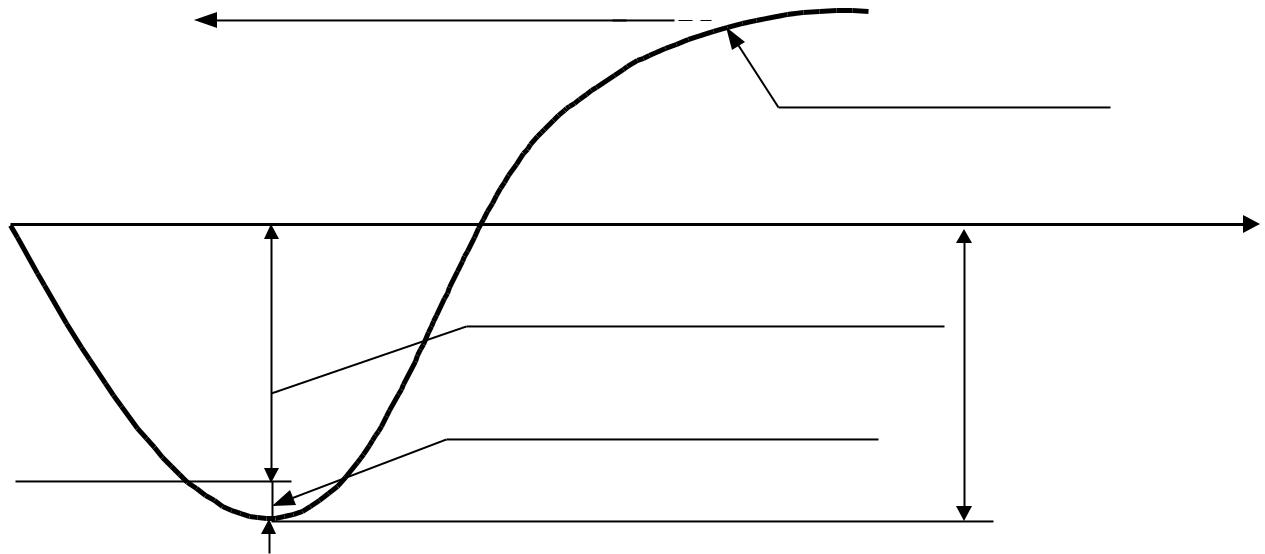
***4. МАКСИМАЛЬНИЙ ГРОШОВИЙ ВІДТІК***

Максимальний грошовий відтік (Cash Outflow), званий в вітчизняних джерелах потребою фінансування (ПФ), – це максимальне значення абсолютної величини негативного накопиченого сальдо від інвестиційної та операційної діяльності.

Величина ПФ показує мінімальний обсяг зовнішнього фінансування проєкту, необхідний для забезпечення його фінансової можливості бути реалізованим. Тому ПФ називають ще капіталом ризику.

Термін зовнішнє фінансування на відміну від внутрішнього передбачає будь-які джерела фінансування (власні та залучені), зовнішні по відношенню до проєкту, тоді як внутрішнє фінансування здійснюється в процесі реалізації проєкту за рахунок отримання чистого прибутку і амортизаційних відрахувань.

На рис. 6.1. показана графічна інтерпретація максимального грошового відтоку.



Збитки в період освоєння виробництва або ринків збуту продукції

Початкові інвестиції

ЧДП (Накопичене сальдо)

Період виробництва

Передвиробничий період

Час

Грошові одиниці

Рис. 6.1. Графічна інтерпретація максимального грошового відтоку (потреба у фінансуванні)

Показники ефективності підприємницької діяльності

***6. ЧИСТА ПОТОЧНА ВАРТІСТЬ***

Найважливішим показником ефективності інвестиційного проєкту є чиста поточна вартість (інші назви – ЧТС, інтегральний економічний ефект, чиста поточна приведена вартість, чистий дисконтований дохід, Net Present Value, NPV) – накопичений дисконтований ефект за розрахунковий період. ЧТС розраховується за такою формулою:

 (6.10)

де  –приплив грошових коштів на m-му кроці;  – відтік грошових коштів на m-му кроці;  – коефіцієнт дисконтування на m-му кроці.

На практиці часто користуються модифікованою формулою

 (6.11)

де  – величина відтоку грошових коштів на m-му кроці без капіталовкладень (інвестицій)  на тому ж кроці.

Для оцінки ефективності інвестиційного проєкту за перші До кроків розрахункового періоду рекомендується використовувати показник поточної ЧТС (накопичене дисконтоване сальдо):

(6.12)

Чиста поточна вартість використовується для зіставлення інвестиційних витрат і майбутніх надходжень грошових коштів, наведених в еквівалентні умови.

Для визначення чистої поточної вартості перш за все необхідно підібрати норму дисконтування і виходячи з її значення знайти відповідні коефіцієнти дисконтування за аналізований розрахунковий період.

Після визначення поточної вартості припливів і відтоків грошових коштів чиста поточна вартість визначається як різниця між зазначеними двома величинами. Отриманий результат може бути як позитивним, так і негативним.

***Таким чином, чиста поточна вартість показує, чи досягнуто інвестиції за економічний термін їх життя бажаного рівня віддачі:***

***– позитивне значення чистої поточної вартості показує, що за розрахунковий період дисконтовані грошові надходження перевищать дисконтовану суму капітальних вкладень і тим самим забезпечать збільшення цінності фірми;***

***– навпаки, від’ємне значення чистої поточної вартості показує, що проєкт не забезпечить отримання нормативної (стандартної) норми прибутку і, отже, призведе до потенційних збитків.***

**Приклад 6.1** (Продовження)/ Інвестиції в сумі 100000 грн. при щорічних протягом 6 років грошові надходження (ануїтетs) в сумі 25000 грн. дозволяють отримати чисту поточну вартість в сумі майже 16000 грн. виходячи з припущення про те, що фірма передбачає застосування норми дисконту (тобто стандартної норми прибутку) на рівні 8% після сплати податку. Всі початкові інвестиції будуть відшкодовані протягом – 5-річного періоду. Чиста поточна вартість проєкту 15575 грн. збільшила капітал фірми на цю суму в сучасному обчисленні, що може захистити інвестора від можливого ризику, в разі, якщо грошові надходження оцінені неточно, а проєкт не завершить свою економічну життя раніше наміченого терміну (табл. 6.5).

Таблиця 6.5

Чиста поточна вартість при нормі дисконту Е = 8%, грн.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Період часу | Інвестиції | Грошові надходження | Коефіцієнт дисконтування при ставці 8% | Чистий поточна вартість різних років | Кумулятивна чиста поточна вартість |
| 0 | 100000 | - | 1,000 | –100000 | –100000 |
| 1 | - | 25000 | 0,926 | +23150 | –76850 |
| 2 | - | 25000 | 0,857 | +21425 | –55425 |
| 3 | - | 25000 | 0,794 | +19850 | –35575 |
| 4 | - | 25000 | 0,735 | +18375 | –17200 |
| 5 | - | 25000 | 0,681 | +17025 | –175 |
| 6 | - | 25000 | 0,630 | +15750 | **+15575** |
|  | 100000 | 150000 |  | **+15575** |  |

**Приклад 6.1** (Продовження). Зробимо розрахунок чистої поточної вартості при збільшенні норми дисконту, що дорівнює 12% (табл. 6.6).

Чиста поточна вартість залишається позитивною, однак її величина скоротилася до 2800 грн. При збільшенні норми дисконту при інших рівних умовах чиста поточна вартість знижується. При нормі дисконту Е = 14% чиста поточна вартість зменшиться ще більше і стане негативною величиною (–2775 грн.).

Забігаючи трохи наперед, відзначимо, що термін окупності інвестицій з дисконтуванням (тобто проміжок часу, необхідний для того, щоб кумулятивна чиста поточна вартість стала позитивною величиною) збільшується (див. Останні колонки табл. 6.5. і 6.6.).

При нормі дисконту 8% термін окупності складе близько 5 років, в той час як при Е = 12% – майже 6 років.

Таблиця 6.6

Чиста поточна вартість при нормі дисконту Е = 12%, грн.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Період часу | Інвестиції | Грошові надходження | Коефіцієнт дисконтування при ставці 8% | Чистий поточна вартість різних років | Кумулятивна чиста поточна вартість |
| 0 | 100000 | - | 1,000 | –100000 | –100000 |
| 1 | - | 25000 | 0,893 | +22325 | –77675 |
| 2 | - | 25000 | 0,797 | +19995 | –57750 |
| 3 | - | 25000 | 0,712 | +17800 | –39950 |
| 4 | - | 25000 | 0,636 | +15900 | –24050 |
| 5 | - | 25000 | 0,567 | +14175 | –9875 |
| 6 | - | 25000 | 0,507 | +12675 | **+2800** |
|  | 100000 | 150000 |  | **+2800** |  |

Найбільш ефективним є застосування показника чистої поточної вартості в якості критеріального механізму, що показує мінімальну нормативну рентабельність (норму дисконту) інвестицій за економічний термін їх життя. Якщо ЧТС є позитивною величиною, то це означає можливість отримання додаткового доходу понад нормативного прибутку, при негативній величині чистої поточної вартості прогнозовані грошові надходження не забезпечують отримання мінімальної нормативного прибутку і відшкодування інвестицій. При чистої поточної вартості, близькою до 0, нормативний прибуток ледве забезпечується (але тільки в разі, якщо оцінки грошових надходжень і прогнозованого економічного строку життя інвестицій виявляться точними).

Незважаючи на всі ці переваги оцінки інвестицій, метод чистої поточної вартості не дає відповіді на всі питання, пов’язані з економічною ефективністю капіталовкладень. Цей метод дає відповідь лише на питання, чи сприяє аналізований варіант інвестування зростання цінності фірми або багатства інвестора взагалі, але ніяк не говорить про відносну міру такого зростання.

A цей захід завжди має велике значення для будь-якого інвестора. Для заповнення такого пробілу використовується інший показник– метод розрахунку рентабельності інвестицій.

***7. ІНДЕКС ПРИБУТКОВОСТІ ІНВЕСТИЦІЙ***

Індекс прибутковості інвестицій (ВД) – відношення суми елементів грошового потоку від операційної діяльності до абсолютній величині суми елементів грошового потоку від інвестиційної діяльності. Він дорівнює збільшеному на одиницю відношенню ЧДП до накопиченого обсягу інвестицій.

Формулу для розрахунку ІД можна визначити, використовуючи формулу (6.7), попередньо перетворивши її в такий вигляд:

(6.8)

де  – величина відтоку грошових коштів на -тому кроці без капіталовкладень (К) (інвестицій) на тому ж кроці.

Тоді формулу для визначення індексу доходності можна представити у вигляді

 (6.9)

Якщо прийняти ряд припущень, то можна показати графічну інтерпретацію індексу прибутковості (рис. 6.1).

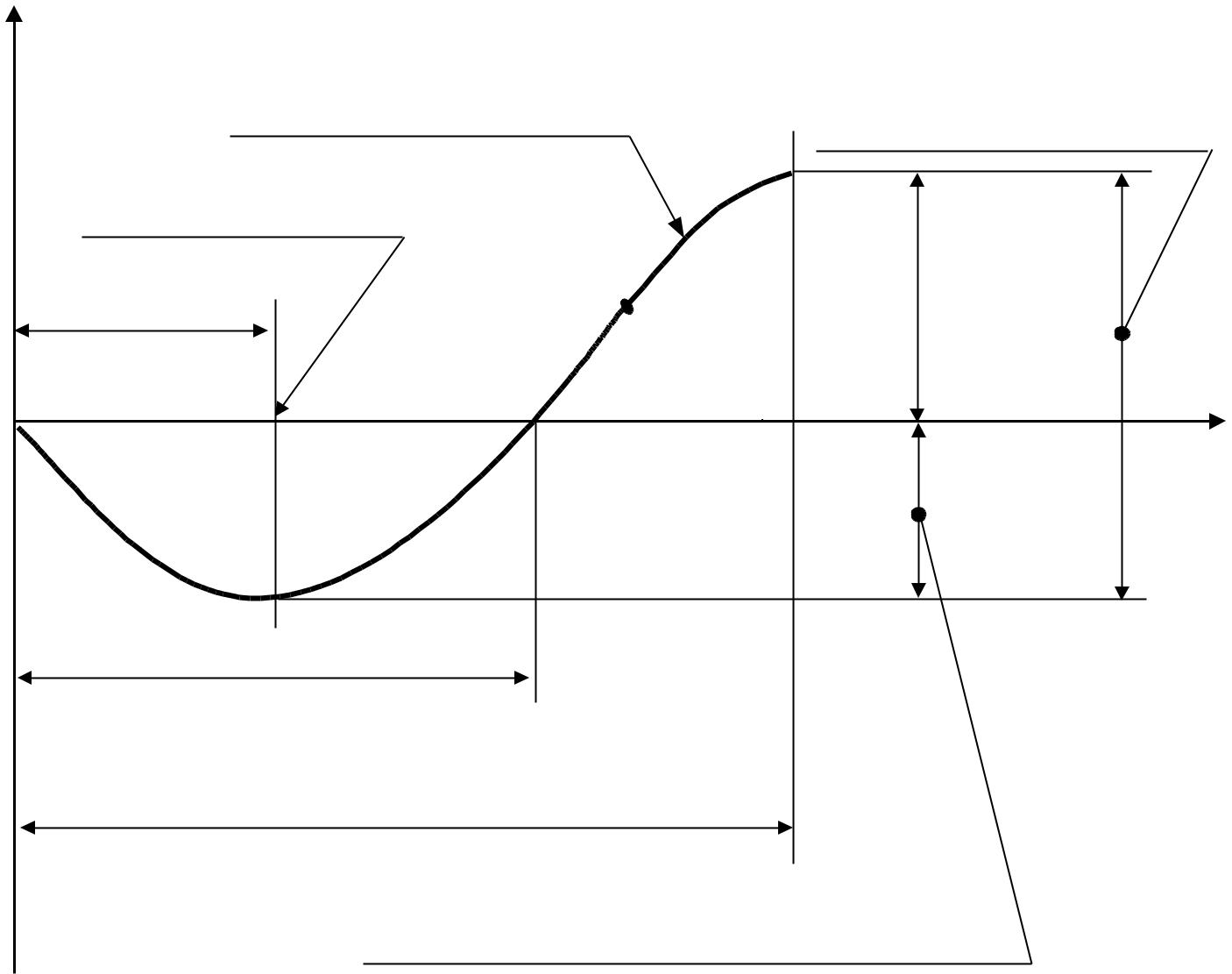
Ці припущення полягають у наступному:

– інвестиції в проєкт для створення активів виробляються тільки в початковий період ();

– підприємство не має збитків в період освоєння виробництва або освоєння ринку збуту виробленої продукції;

– ми нехтуємо ліквідаційною вартістю активів. Використовуючи формулу (6.9) і рис. 6.2, можна уявити формулу для визначення індексу прибутковості у вигляді

 або  (6.9а)



Чисті грошові надходження наростаючим підсумком

Сума елементів грошового потоку від операційної діяльності

Термін окупності інвестицій без дисконтування

Економічний тривалість життя інвестицій

Сума елементів грошового потоку від інвестиційної діяльності

Початок операційної діяльності

Період інвестування









Рис. 6.2. Графічна інтерпретація ВД інвестицій

Таким чином, індекс прибутковості (ІД) є не що інше, як показник рентабельності інвестицій, визначений щодо сумарних показників ЧДП та інвестицій за економічний термін їх життя.

При розрахунку індексу прибутковості можуть враховуватися або всі капіталовкладення за розрахунковий період, включаючи вкладення в заміщення вибувають основних фондів, або тільки початкові капіталовкладення, здійснювані до введення підприємства в експлуатацію (відповідні показники будуть, звичайно, мати різні значення).

Індекс прибутковості інвестицій перевищує 1, якщо і тільки якщо для цього потоку чисті грошові надходження мають позитивні значення.

***Індекс прибутковості дисконтованих інвестицій*** *(Інші назви – ІДД, рентабельність інвестицій, Profitability Index, PI)* – відношення суми дисконтованих елементів грошового потоку від операційної діяльності до абсолютній величині дисконтованою суми елементів грошового потоку від інвестиційної діяльності. ІДД дорівнює збільшеному на одиницю відношенню ЧТС (NPV) до накопиченого дисконтованого обсягу інвестицій.

Формула для визначення ІДД має наступний вигляд:

******(6.13)

або

 (6.14)

При розрахунку ІДД можуть враховуватися або всі капіталовкладення за розрахунковий період, включаючи вкладення в заміщення вибувають основних фондів, або тільки початкові капітальні вкладення, здійснювані до введення підприємства в експлуатацію. У цьому випадку відповідні показники будуть мати різні значення.

Графічна інтерпретація індексу дисконтованої доходності (з урахуванням припущень, зазначених в підрозділ. 6.2.3) наведено на рис. 6.3.

Індекси прибутковості дисконтованих інвестицій перевищують 10, якщо і тільки якщо для цього потоку чиста поточна вартість позитивна.

В процесі визначення чистої поточної вартості декількох інвестиційних проєктів стикаються з проблемою вибору з альтернативних інвестицій, що відрізняються за своїми розмірами. При цьому чиста поточна вартість може виявитися рівною для цих

інвестиційних проєктів. Наприклад, чи будуть однаково привабливими інвестиційні проєкти, які передбачають отримання чистої поточної вартості в сумі 100000 грн. або 250000 грн., навіть якщо ми припустимо однакові для кожного проєкту тривалості економічного терміну життя інвестицій і ступінь ризику.

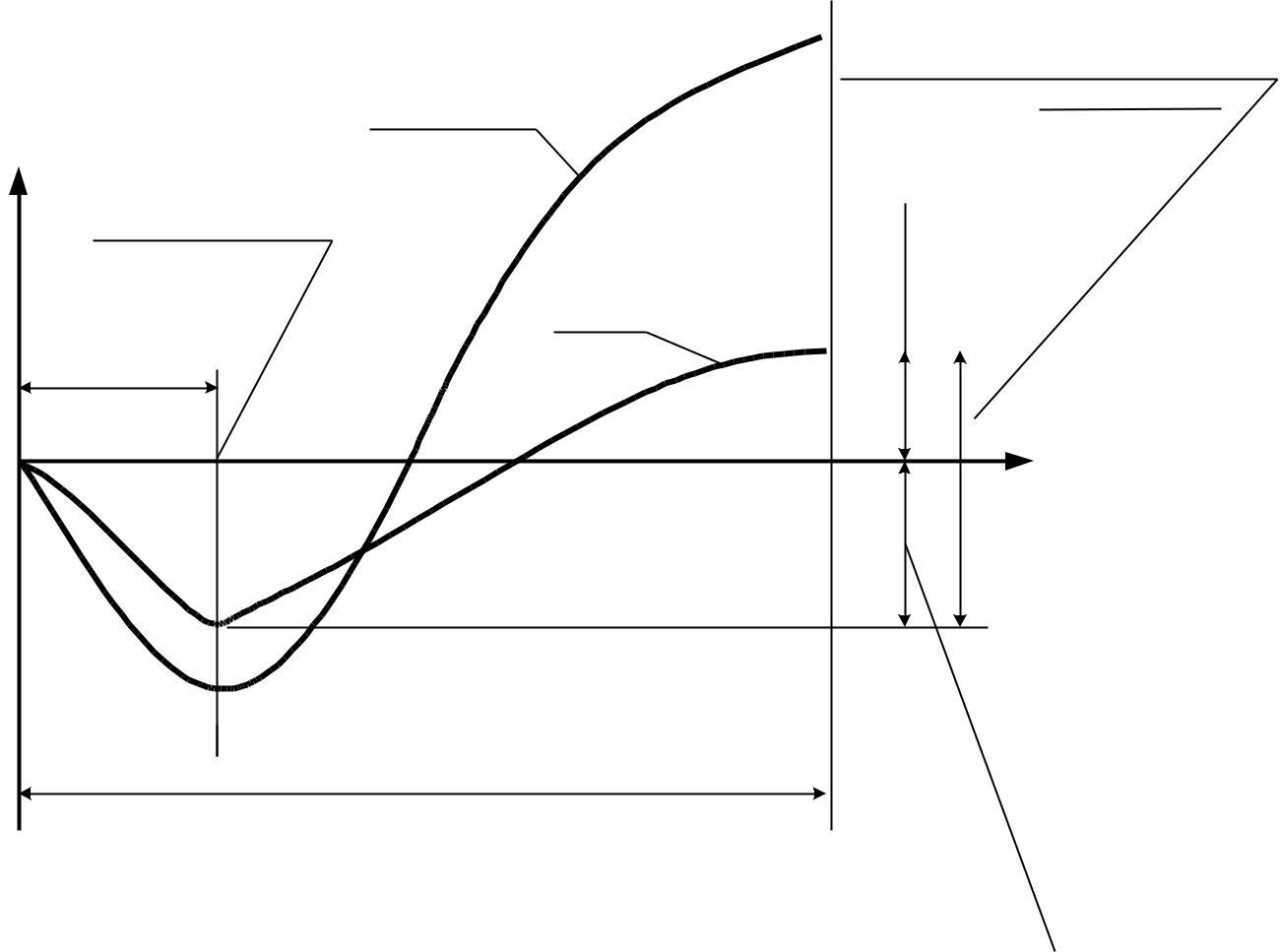
Розрахуємо ІДД для цих двох інвестиційних проєктів:

1)  2) 

Чим вище індекс прибутковості дисконтованих інвестицій, тим краще проєкт. У нашому випадку перший проєкт виявився кращим.

Якщо ІДД дорівнює 1,0, то проєкт ледь забезпечує отримання мінімальної нормативного прибутку. При ІДД менше 1,0 проєкт не забезпечує одержання мінімальної нормативного прибутку.

ІДД, що дорівнює 1,0, висловлює нульову чисту поточну вартість.



Початок операційної діяльності

Чистий грошовий потік наростаючим підсумком

Період інвестування

Економічний тривалість життя інвестицій

Сума дисконтованих елементів грошового потоку від операційної діяльності

Грошові одиниці

Час

Сума дисконтованих елементів грошового потоку від інвестиційної діяльності

Чиста поточна вартість наростаючим підсумком





******



Рис. 6.3. Графічна інтерпретація індексу прибутковості дисконтованих інвестицій

***8. ВНУТРІШНЯ НОРМА ПРИБУТКОВОСТІ***

Внутрішня норма прибутковості (інші назви – ВНД, внутрішня норма дисконту, внутрішня норма прибутку, внутрішній коефіцієнт ефективності, Internal Rate of Return, IRR).

В найпоширенішому випадку інвестиційних проєктів, що починаються з (інвестиційних) витрат і мають позитивне значення чистих грошових надходжень, внутрішньою нормою прибутковості називається позитивне число ЕВ, Якщо:

– *при нормі дисконту* Е = ЕВ *чиста поточна вартість проєкту звертається в* 0,

– *це число єдине.*

*В більш загальному випадку внутрішньою нормою прибутковості називається таке позитивне число* ЕВ*, Що при нормі дисконту* Е = ЕВ *чиста поточна вартість проєкту звертається в* 0, при всіх великих значеннях Е – *негативна, при всіх менших значеннях* Е – *позитивна. Якщо не виконано хоча б одне з цих умов, вважається, що ВНД не існує.*

Якщо повернутися до описаних вище рівнянням (6.10) і (6.11), то ВНД – це значення норми дисконту (Е) в цих рівняннях, при якій чиста поточна вартість буде дорівнює нулю, т. е .:

 (6.15)

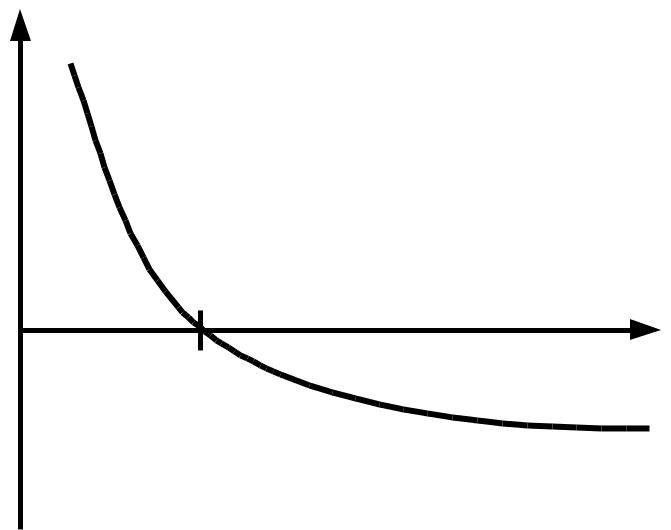
 (6.15а)

Для того щоб легше розібратися в категорії ВНД, домовимося, що поки ми будемо вести мову про такі інвестиційні проєкти, при реалізації яких:

– треба спочатку здійснити витрати грошових коштів (допустити відтік коштів) і лише потім можна розраховувати на грошові надходження (притоки коштів);

– грошові надходження носять кумулятивний характер, причому їх знак змінюється лише один раз (тобто спочатку вони можуть бути негативними, але, ставши потім позитивними, залишатимуться такими протягом усього розрахункового періоду).

Для таких інвестицій справедливо твердження про те, що чим вище норма дисконту (Е), тим менше величина інтегрального ефекту (NPV), що як раз і ілюструє рис. 6.4.



ЧТС (NPV)

ВНД (JRR)

E

–

+

Рис. 6.4. Залежність величини ЧТС (NPV) від рівня норми дисконту (Е)

Як видно з рис. 6.4, ВНД – це та величина норми дисконту (Е), при якій крива зміни ЧТС перетинає горизонтальну вісь, тобто інтегральний економічний ефект (NPV) виявляється рівним нулю. Знайти величину ВНД можна за допомогою таблиць коефіцієнтів приведення (дод. 4).

**Приклад 6.2**. Припустимо, що транспортна фірма планує здійснити закупівлю судна «річка –море «вартістю 40 млн грн. Його експлуатація, як показують розрахунки, може забезпечити протягом 20 років щорічні (якщо вважати на кінець року) грошові надходження на рівні 3,2 млн грн. Якщо виходити з того, що ми маємо справу з ануїтетом, то можна скористатися формулою (5.6) для поточної вартості:

,

тобто 

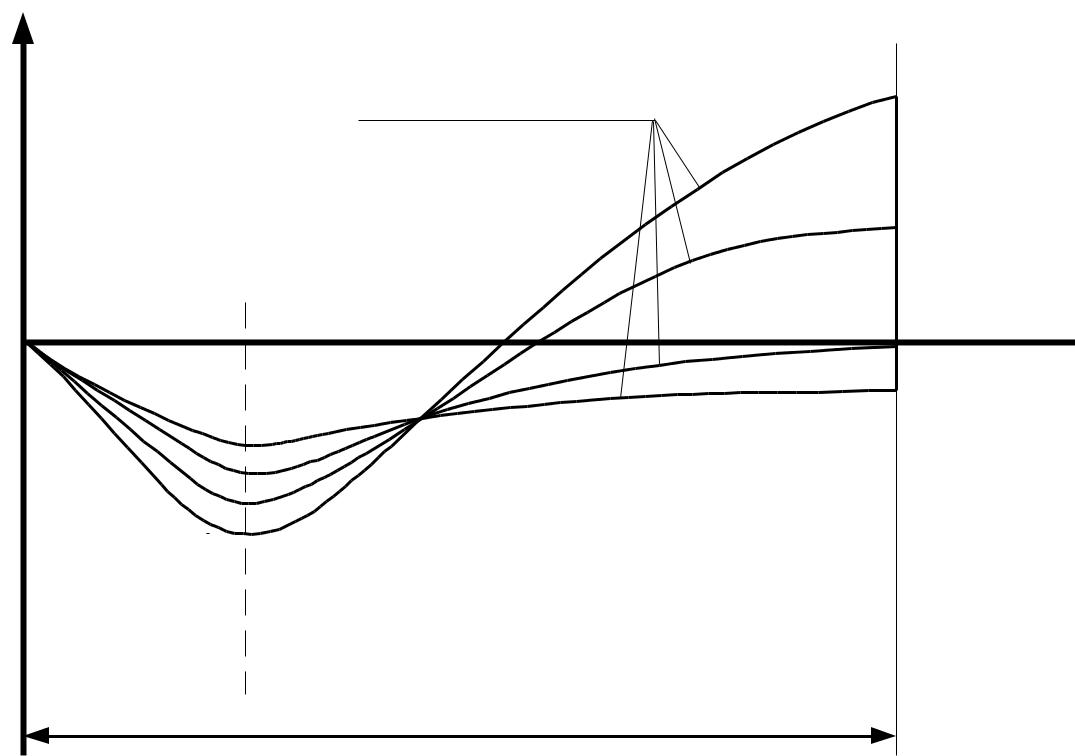
і тоді .

Звернувшись до довідкової таблиці коефіцієнтів розрахунку поточної вартості для ануїтету (дод. 4), ми виявимо, що по рядку з номером 20 (тобто для 20-річного періоду) найбільш близьким значенням буде величина 12,462, відповідна нормі дисконту на рівні 5% .

Саме в такому значенні норми дисконту поточна вартість майбутніх грошових надходжень від експлуатації судна буде дорівнює сучасним інвестиціям, а ЧТС (NPV) – нулю. Отже, в даному прикладі ВНД (JRR) дорівнює 0,05*.*

Рішення задачі визначення ВНД стає особливо важкою в тих випадках, коли майбутні грошові надходження можуть бути не однаковими за величиною. Суть завдання залишається колишньою– знайти значення ВНД, при якому ЧТС (NPV) буде дорівнює нулю. В цьому випадку процес розрахунку зводиться до методу проб і помилок, щоб шляхом декількох послідовних наближень, ітерацій знайти шукане значення ВНД. При цьому спочатку ЧТС (NPV) визначається за допомогою експертно-обраної величини норми дисконту. Якщо при цьому ЧТС виявляється позитивною, то розрахунок повторюється з використанням більшої величини норми дисконту (або, навпаки, при негативному значенні ЧТС –- з використанням менші за розміром норми дисконту), поки не вдасться підібрати таку норму дисконту, при якій ЧТС буде дорівнює нулю.

На рис. 6.5 показана графічна інтерпретація визначення ВНД методом ітерацій, з якого видно, що ВНД для розглянутого проєкту дорівнює ≈ 0,18.



Чиста поточна вартість

наростаючим підсумком

Економічний тривалість життя інвестицій

Грошові одиниці

при E = 0,1

при E = 0,15

при E = 0,18

при E = 0,2

–

0

+

***Рис. 6.5. Графічна інтерпретація визначення ВНД методом ітерацій***

Отже, ВНД визначається як така норма дисконту (E), при якій чиста поточна вартість дорівнює нулю, тобто інвестиційний проєкт не забезпечує зростання цінності фірми, але і не веде до її зниження. Саме тому у вітчизняній літературі ВНД іноді називають перевірочним дисконтом, так як вона дозволяє знайти граничне значення норми дисконту (Eв), Що розділяє граничні інвестиції на прийнятні і невигідні. Для цього ВНД порівнюють з прийнятою для проєкту нормою дисконту (E).

Принцип порівняння цих показників такої:

* якщо ВНД (JRR) < E – проєкт прийнятний (так як ЧТС в цьому випадку має позитивне значення);
* якщо ВНД > E – проєкт не прийнятний (так якЧТС негативна);
* якщо ВНД = E – можна приймати будь-яке рішення.

Таким чином, ВНД стає як би ситом, що відсіває невигідні проєкти.

Крім того, цей показник може служити основою для ранжирування проєктів за ступенем вигідності, при інших рівних умовах, тобто при тотожності основних вихідних параметрів порівнюваних проєктів:

* рівній сумі інвестицій;
* однакової тривалості розрахункового періоду;
* однаковому рівні ризику.

Внутрішня норма прибутковості може бути використана також:

* для економічної оцінки проєктних рішень, якщо відомі прийнятні значення ВНД (залежні від області застосування) у проєктів даного типу;
* для оцінки ступеня стійкості інвестиційних проєктів по різниці ВНД – Е;
* для встановлення учасниками проєкту норми дисконту Е за даними про внутрішній нормі прибутковості альтернативних напрямків вкладення ними власних коштів.

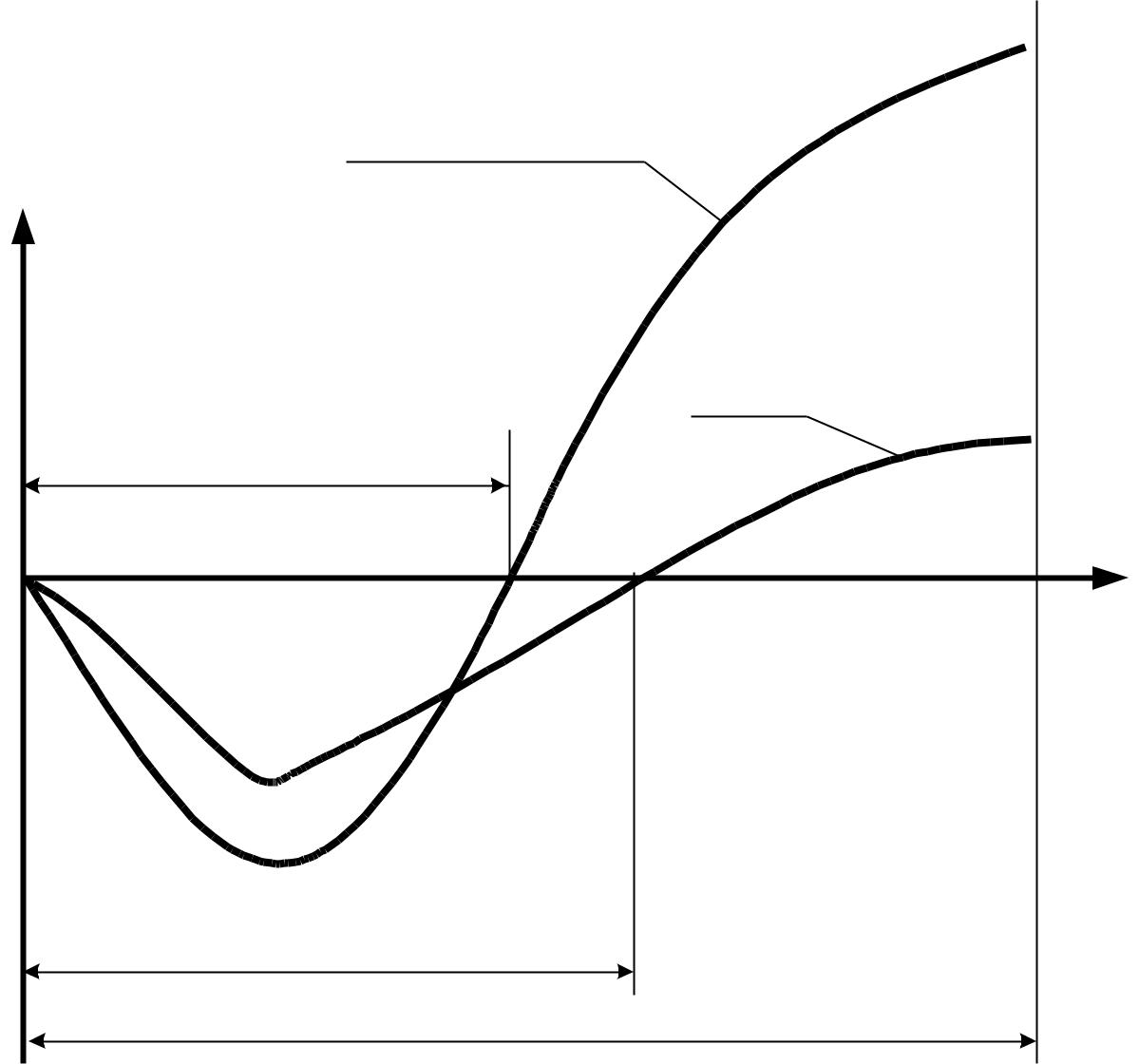
Для оцінки ефективності інвестиційних проєктів за перші До кроків розрахункового періоду використовується показник поточної внутрішньої норми прибутковості (поточна ВНД), яка визначається як таке число ВНД (K), що при нормі дисконту Е = ВНД (K) величина ЧТС (K) звертається до 0, при всіх великих значеннях E – негативна, при всіх менших значеннях E – позитивна. Для окремих проєктів і значень K поточна ВНД може не існувати.

***9. ТЕРМІН ОКУПНОСТІ ІНВЕСТИЦІЙ З УРАХУВАННЯМ ДИСКОНТУВАННЯ***

Терміном окупності інвестицій з урахуванням дисконтування називається тривалість періоду від початкового моменту до моменту окупності з урахуванням дисконтування. моментом окупності з урахуванням дисконтування називається той найбільш ранній момент часу в розрахунковому періоді, після якого чиста поточна вартість ЧТС (K) стає і надалі залишається неотрицательной (рис. 6.6).

***10. МАКСИМАЛЬНИЙ ГРОШОВИЙ ВІДТІК З УРАХУВАННЯМ ДИСКОНТУВАННЯ (ПОТРЕБА У ФІНАНСУВАННІ З УРАХУВАННЯМ ДИСКОНТУ, ДПФ)***

Максимальний грошовий відтік з урахуванням дисконтування (потреба у фінансуванні з урахуванням дисконту, ДПФ) – максимальне значення абсолютної величини негативного накопиченого дисконтованого сальдо від інвестиційної та операційної діяльності. Величина ДПФ показує мінімальний дисконтований обсяг зовнішнього (по відношенню до проєкту) фінансування проєкту, необхідний для забезпечення його фінансової реалізованості (рис. 6.7).



Чистий грошовий потік наростаючим підсумком

Грошові одиниці

Час

Термін окупності інвестицій (без дисконтування)

Чиста поточна вартість наростаючим підсумком

Термін окупності інвестицій з урахуванням дисконтування

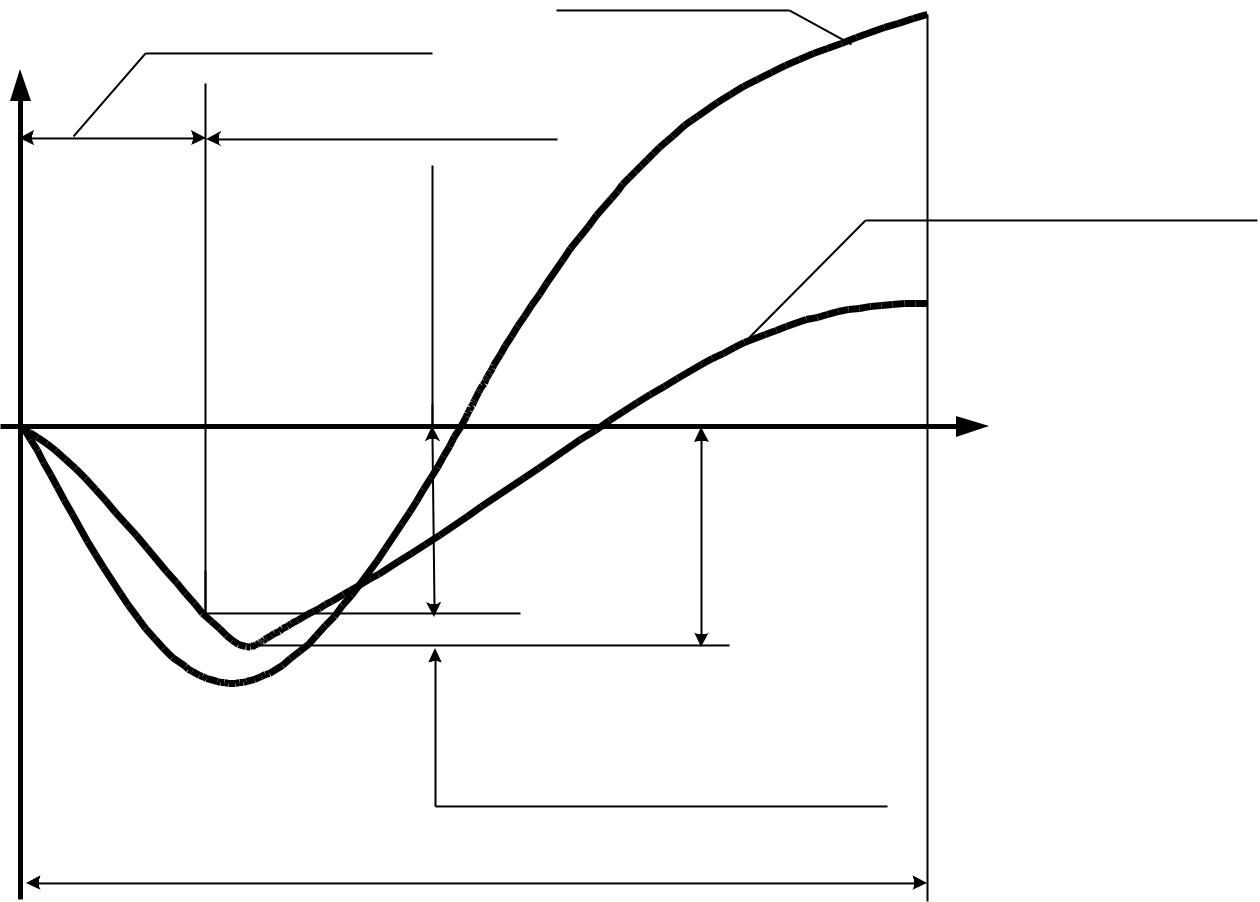
Економічний тривалість життя інвестицій

+

0

–

Рис. 6.6. Графічна інтерпретація терміну окупності інвестицій з урахуванням дисконтування



Передвиробничий період

Грошові одиниці

Початкові дисконтовані інвестиції

Максимальний грошовий відтік з врахуванням дисконтування (ДПФ)

Економічний тривалість життя інвестицій

Час

Дисконтовані збитки в період освоєння виробництва або ринку збуту продукції

Період виробництва

Чиста поточна вартість (накопичене сальдо)

Чисті грошові надходження (накопичене сальдо)

Рис.6.7. Графічна інтерпретація максимального грошового відтоку з урахуванням дисконтування (ДПФ)

Тема 7. Оцінка невизначеності і ризику бізнес-ідеї

Збільшена оцінка стійкості бізнес-ідеї. Збільшена оцінка стійкості бізнес-ідеї з точки зору її учасників. Розрахунок меж беззбитковості. Метод варіації параметрів. Граничні значення параметрів. Оцінка очікуваного ефекту бізнес-ідеї з урахуванням кількісних характеристик невизначеності. Ймовірнісна невизначеність. Інтервальна невизначеність.

[3] c. 159-168; [4] c. 232-258

***Лекція 7.1. (2 год)***

**Тема 7. Оцінка невизначеності і ризику бізнес-ідеї**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Збільшена оцінка стійкості бізнес-ідеї
2. Збільшена оцінка стійкості бізнес-ідеї з точки зору її учасників
3. Розрахунок меж беззбитковості

***1. ЗБІЛЬШЕНА ОЦІНКА СТІЙКОСТІ БІЗНЕС-ІДЕЇ***

В розрахунках ефективності проєктів рекомендується враховувати невизначеність, тобто неповноту і неточність інформації про умови реалізації проєкту, і ***ризик***, тобто можливість виникнення таких умов, які призведуть до негативних наслідків для всіх або окремих учасників проєкту. Показники ефективності проєкту, обчислені з урахуванням факторів ризику і невизначеності, іменуються ***очікуваними***.

При цьому сценарій реалізації проєкту, для якого були виконані розрахунки ефективності (тобто поєднання умов, до якого відносяться ці розрахунки), розглядається як основний (базисний), всі інші можливі сценарії – як сценарії, які викликають ті чи інші позитивні або негативні відхилення від відповідних базисних (проєктних) значень показників ефективності. Наявність або відсутність ризику, пов’язане із здійсненням того чи іншого сценарію, визначається кожним учасником за величиною і знаку відповідних відхилень. Ризик, пов’язаний з виникненням тих чи інших умов реалізації проєкту, залежить від того, з точки зору чиїх інтересів він оцінюється.

Окремі фактори невизначеності підлягають обліку в розрахунках ефективності, якщо при різних значеннях цих факторів витрати і результати за проєктом істотно розрізняються.

***Проєкт вважається стійким, якщо при всіх сценаріях він виявляється ефективним і фінансово-реалізованим, а можливі несприятливі наслідки усуваються заходами, передбаченими організаційно-економічним механізмом проєкту.***

З метою оцінки стійкості та ефективності проєкту в умовах невизначеності рекомендується використовувати такі методи (кожен наступний метод є більш точним, хоча і більш трудомістким, і тому застосування кожного з них робить непотрібним застосування попередніх):

1. укрупненої оцінки стійкості;
2. розрахунку рівнів беззбитковості;
3. варіації параметрів;
4. оцінки очікуваного ефекту проєкту з урахуванням кількісних характеристик невизначеності.

Всі методи, крім першого, передбачають розробку сценаріїв реалізації проєкту в найбільш ймовірних або найбільш небезпечних для будь-яких учасників умовах і оцінку фінансових наслідків здійснення таких сценаріїв. Це дає можливість при необхідності передбачити в проєкті заходи щодо запобігання або перерозподілу виникають втрат.

При виявленні нестійкості проєкту рекомендується внести необхідні корективи в організаційно-економічний механізм його реалізації, в тому числі:

* змінити розміри і / або умови надання позик;
* передбачити створення необхідних запасів, резервів грошових коштів, відрахувань до додаткового фонду;
* скорегувати умови взаєморозрахунків між учасниками проєкту;
* передбачити страхування учасників проєкту на ті чи інші страхові випадки.

У тих випадках, коли і при цих коректив проєкт залишається нестійким, його реалізація визнається недоцільною, якщо відсутня додаткова інформація, достатня для застосування четвертого з перерахованих вище методів. В іншому випадку вирішення питання реалізації проєкту проводиться на підставі цього методу без урахування результатів усіх попередніх.

При використанні цього методу з метою забезпечення стійкості проєкту рекомендується:

* використовувати помірно песимістичні прогнози техніко-економічних параметрів проєкту, цін, ставок податків, обмінних курсів валют і інших параметрів економічного оточення проєкту, обсягу виробництва та цін на продукцію, термінів виконання та вартості окремих видів робіт і т.д. (при цьому позитивні відхилення вказаних параметрів будуть більш ймовірними, ніж негативні);
* передбачати резерви коштів на непередбачені інвестиційні та операційні витрати, обумовлені можливими помилками проєктної організації, переглядом проєктних рішень в ході будівництва, непередбаченими затримками платежів за поставлену продукцію і т.п .;
* збільшити норму дисконту в розрахунках комерційної ефективності на величину ***поправки на ризик***.

При дотриманні цих умов проєкт рекомендується розглядати як стійкий в цілому, якщо він має досить високі значення інтегральних показників, зокрема позитивне значення очікуваної чистої поточної вартості.

***2. ЗБІЛЬШЕНА ОЦІНКА СТІЙКОСТІ БІЗНЕС-ІДЕЇ З ТОЧКИ ЗОРУ ЙОГО УЧАСНИКІВ***

Стійкість ІП з точки зору підприємства – учасника проєкту при можливих змінах умов його реалізації – може бути збільшено перевірена за результатами розрахунків комерційної ефективності для основного (базисного) сценарію реалізації проєкту шляхом аналізу динаміки потоків реальних коштів. Вхідні в розрахунок потоки реальних коштів при цьому обчислюються за всіма видами діяльності учасника з урахуванням умов надання та погашення позик.

Якщо на тому чи іншому етапі розрахункового періоду можлива аварія, ліквідація наслідків якої, включаючи відшкодування збитків, вимагає додаткових витрат, до складу грошових відтоків включаються відповідні ***очікувані втрати***. Вони визначаються як добуток витрат по ліквідації наслідків аварії на ймовірність виникнення аварії на даному етапі.

Для укрупненої оцінки стійкості проєкту можуть використовуватися показники внутрішньої норми комерційної прибутковості та індексу прибутковості дисконтованих інвестицій. При цьому ІП вважається стійким, якщо значення ВНД досить велике (не менше 25-30%), значення норми дисконту не перевищує рівня для малих і середніх ризиків, і при цьому не передбачається позик за реальними ставками, що перевищують ВНД, а індекс прибутковості дисконтованих інвестицій перевищує 1,2.

При дотриманні вимог підрозділу 2 до параметрів основного сценарію реалізації проєкту, проєкт рекомендується оцінити як стійкий тільки при наявності певного фінансового резерву. З огляду на, що вільні фінансові кошти підприємства включають не тільки накопичене сальдо грошового потоку від усіх видів діяльності, а й резерв коштів у складі активів підприємства, умова стійкості проєкту може бути сформульовано таким чином.

***На кожному кроці розрахункового періоду сума накопиченого сальдо грошового потоку від усіх видів діяльності (накопиченого ефекту) і фінансових резервів повинна бути невід’ємною***.

Рекомендується, щоб вона становила не менше 5% суми чистих операційних витрат і здійснюваних на цьому кроці інвестицій.

Для виконання даної рекомендації може знадобитися змінити передбачені проєктом норми резерву фінансових коштів, передбачити відрахування в резервний капітал або скорегувати схему фінансування проєкту. Якщо подібні заходи не забезпечать виконання зазначеної вимоги, необхідно більш детальне дослідження впливу невизначеності на достовірність, і ефективність ІП (див. нижче).

***3. РОЗРАХУНОК МЕЖ БЕЗЗБИТКОВОСТІ***

Ступінь стійкості проєкту по відношенню до можливих змін умов реалізації може бути охарактеризована показниками ***меж беззбитковості*** і ***граничних значень*** таких параметрів проєкту, як обсяг виробництва, ціни продукції, що виробляється та ін. Подібні показники використовуються тільки для оцінки впливу можливої зміни параметрів проєкту на його фінансову реалізованість та ефективність , але самі вони не відносяться до показників ефективності інвестиційного проєкту, і їх обчислення не замінює розрахунків інтегральних показників ефективності.

Кордон беззбитковості параметра проєкту для деякого кроку розрахункового періоду визначається як такий коефіцієнт до значення цього параметра на даному етапі, при застосуванні якого чистий прибуток, отриманий в проєкті на цьому кроці, стає нульовою. Одним з найбільш поширених показників цього типу є рівень беззбитковості. Він зазвичай визначається для проєкту в цілому, чому і відповідає приводиться нижче формула (7.1).

Рівнем беззбитковості на кроці *m* називається відношення обсягу продажів (виробництва), що відповідає «точці беззбитковості» (), до проєктного () на цьому кроці. Під «точкою беззбитковості» розуміється обсяг продажів, при якому чистий прибуток стає рівною нулю. При визначенні цього показника приймається, що на кроці *m*:

* обсяг виробництва дорівнює обсягу продажів;
* обсяг виручки змінюється пропорційно обсягу продажів;
* доходи від позареалізаційної діяльності і витрати по цій діяльності не залежать від обсягів продажів;
* повні поточні витрати виробництва можуть бути розділені на умовно-постійні (Які не змінюються при зміні обсягу виробництва) і умовно-змінні, що змінюються прямо пропорційно обсягам виробництва;
* розрахунок рівня беззбитковості здійснюється за формулою

(7.1.)

Точка беззбитковості визначається за формулою

(7.2.)

де – умовно-постійні витрати на кроці т, включаючи амортизацію, податки і інші відрахування, що відносяться на собівартість і фінансові результати, які не залежать від обсягу виробництва; – доходи від позареалізаційної діяльності за вирахуванням витрат по цій діяльності на цьому кроці; Р – ціна одиниці продукції; – умовно-змінні витрати на одиницю продукції (послуг), включаючи податки і інші відрахування, що відносяться на собівартість і фінансові результати, пропорційні виручці за винятком податку на прибуток на *m*-мe кроці.

На практиці використовується також формула для визначення рівня беззбитковості такого вигляду:

(7.3.)

де – обсяг виручки на т-м кроці; – повні поточні витрати виробництва продукції (виробничі витрати плюс амортизація, податки і інші відрахування, що відносяться як на собівартість, так і на фінансові результати, крім податку на прибуток) на *m*-му кроці; – умовно-змінна частина повних поточних витрат виробництва (що включає поряд зі змінною частиною виробничих витрат і, можливо, амортизації податки та інші відрахування, пропорційні виручці) на *m*-му кроці; – доходи від позареалізаційної діяльності за вирахуванням витрат по цій діяльності на *m* -му кроці.

Якщо проєкт передбачає виробництво декількох видів продукції, формула (7.3.) не змінюється, а всі вхідні в нею величини беруться з усього проєкту (без поділу за видами продукції).

При використання формул (7.2.), (7.3.) всі ціни і витрати слід враховувати без ПДВ.

На рис. 7.1. наведено графічний спосіб визначення точки беззбитковості.

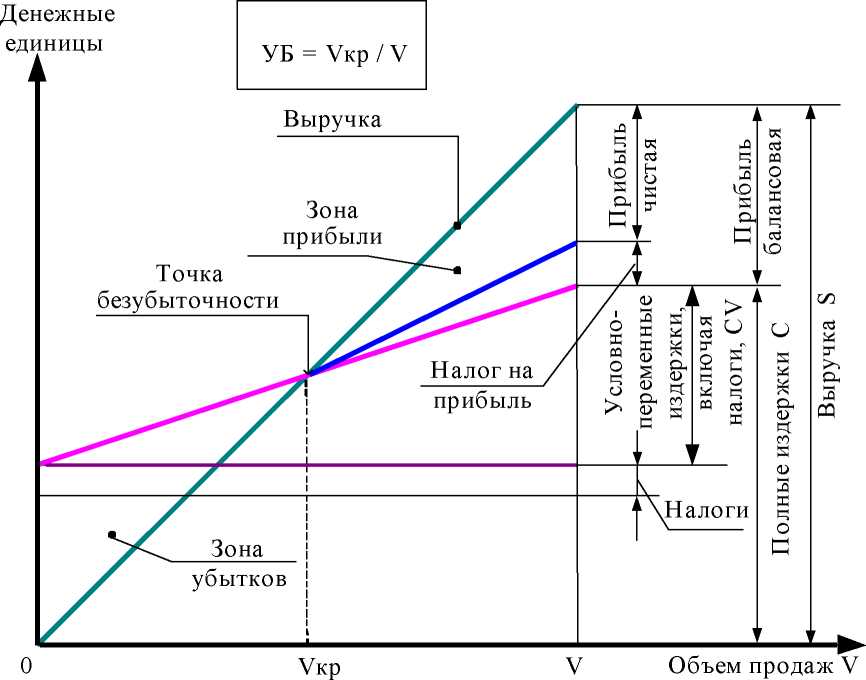


Рис. 7.1. Графік точки беззбитковості

Зазвичай проєкт вважається стійким, якщо в розрахунках за проєктом в цілому рівень беззбитковості не перевищує 0,6 ÷ 0,7 після освоєння проєктних потужностей. Близькість рівня беззбитковості до 1 (100%), як правило, свідчить про недостатню стійкості проєкту до коливань попиту на продукцію на даному етапі. Навіть задовільні значення рівня беззбитковості на кожному кроці не гарантують ефективність проєкту (позитивність ЧТВ). У той же час, високі значення рівня беззбитковості на окремих кроках не можуть розглядатися як ознака нереалізованості проєкту (наприклад, на етапі освоєння вводять експлуатацію або в період капітального ремонту дорогого високопродуктивного обладнання вони можуть перевищувати 100%).

Якщо припущення про пропорційність або / та на кроці *m* обсягом продажів (виробництва) на тому ж кроці не виконуються, замість використання формул (7.2.), (7.3.) слід визначати рівень беззбитковості варіантних розрахунками (підбором) чистого прибутку при різних обсягах виробництва.

Поряд з розрахунками рівнів беззбитковості, для оцінки стійкості проєкту можна оцінювати кордону беззбитковості для інших параметрів проєкту – граничних рівнів цін на продукцію і основні види сировини, граничної частки продажів без передоплати і ін. Для подібних розрахунків необхідно враховувати вплив змін відповідного параметра на різні складові грошових надходжень і витрат. Близькість проєктних значень параметрів до кордону беззбитковості може свідчити про недостатню стійкості проєкту на відповідному кроці.

Межі беззбитковості можна визначати і для кожного учасника проєкту (критерій досягнення кордону - звернення в нуль чистого прибутку цього учасника). Для цього необхідно визначити, як змінюються доходи і витрати цього учасника при зміні значень параметра, для якого визначаються значення кордону.

***Лекція 7.2. (2 год)***

**Тема 7. Оцінка невизначеності і ризику бізнес-ідеї**

СТРУКТУРНИЙ ПЛАН ЛЕКЦІЇ

1. Метод варіації параметрів. Граничні значення параметрів
2. Оцінка очікуваного ефекту бізнес-ідеї з урахуванням кількісних характеристик невизначеності
3. Ймовірнісна невизначеність
4. Інтервальна невизначеність

***1. МЕТОД ВАРІАЦІЇ ПАРАМЕТРІВ. ГРАНИЧНІ ЗНАЧЕННЯ ПАРАМЕТРІВ***

Вихідні показники проєкту можуть суттєво змінитися при несприятливій зміні (відхиленні від проєктних) деяких параметрів.

Рекомендується перевіряти достовірність, і оцінювати ефективність проєкту в залежності від зміни наступних параметрів:

* інвестиційних витрат (або їх окремих складових);
* обсягу виробництва;
* витрат виробництва і збуту (або їх окремих складових);
* відсотка за кредит;
* прогнозів загального індексу інфляції, індексів цін та індексу внутрішньої інфляції (або іншої характеристики зміни купівельної спроможності) іноземної валюти;
* затримок платежів;
* тривалості розрахункового періоду (моменту припинення реалізації проєкту);
* інших параметрів.

При відсутності інформації про можливих, з точки зору учасника проєкту, не більше зміни значень зазначених параметрів рекомендується провести варіантні розрахунки можливості бути реалізованим і ефективності проєкту послідовно для наступних сценаріїв:

1. збільшення інвестицій. При цьому вартість робіт, виконуваних російськими підрядниками, і вартість обладнання російської поставки збільшуються на 20%, вартість робіт і устаткування інофірм - на 10%. Відповідно змінюються вартість основних фондів і розміри амортизації в собівартості;
2. збільшення на 20% від проєктного рівня непрямих виробничих витрат і на 30% питомих (на одиницю продукції) прямих матеріальних витрат на виробництво і збут продукції. Відповідно змінюється вартість запасів сировини, матеріалів, незавершеного виробництва і готової продукції в складі оборотних коштів;
3. зменшення обсягу виручки до 80% її проєктного значення;
4. збільшення на 100% часу затримок платежів за продукцію, що поставляється без передоплати;
5. збільшення відсотка за кредит на 40% його проєктного значення за кредитами в рублях і на 20% по кредитах у СКВ.

Ці сценарії рекомендується розглядати на тлі несприятливого розвитку інфляції, що задається експертно.

Якщо проєкт передбачає страхування на випадок зміни відповідних параметрів проєкту, або значення цих параметрів фіксовані в підготовлених до висновку контрактах, відповідні цих випадків сценарії не розглядаються.

***Примітка.*** Зазвичай при не дуже великі зміни параметрів проєкту відповідні зміни елементів грошових потоків і узагальнюючих показників ефективності проєкту виражаються залежностями, близькими до лінійних. В цьому випадку проєкт, який реалізується і ефективний при декількох сценаріях, буде реалізованим і ефективним при будь-яких «середніх» сценаріях. Наприклад, з реалізованим проєкту при сценаріях 2 і 3 слід його достовірність, і ефективність при одночасному збільшенні виробничих витрат на 10% і зменшення обсягу виручки на 10%.

Проєкт вважається стійким по відношенню до можливих змін параметрів, якщо при всіх розглянутих сценаріях:

* ЧТВ позитивна;
* забезпечується необхідний резерв фінансової можливості бути реалізованим проєкту.

Якщо при будь-якому з розглянутих сценаріїв хоча б одне із зазначених умов не виконується, рекомендується провести більш детальний аналіз меж можливих коливань відповідного параметра і при можливості уточнити верхні межі цих коливань. Якщо і після такого уточнення умови стійкості проєкту не дотримуються, рекомендується:

* при відсутності додаткової інформації відхилити проєкт;
* при наявності інформації, зазначеної в підрозділ. 7.6, оцінювати ефективність ІП більш точними викладеними там методами.

Оцінка стійкості може проводитися також шляхом визначення граничних значень параметрів проєкту, т е. Таких їх значень, при яких інтегральний комерційний ефект учасника стає рівним нулю. Одним з таких показників є ВНД, що відображає ***граничне значення норми дисконту***. Для оцінки граничних значень параметрів, мінливих по кроках розрахунку (ціни продукції і основного технологічного обладнання, обсяги виробництва, обсяг кредитних ресурсів, ставки найбільш істотних податків і ін.), рекомендується обчислювати ***граничні інтегральні рівні*** цих параметрів, тобто такі коефіцієнти (постійні для всіх кроків розрахунку) до значень цих параметрів, при застосуванні яких ЧТВ проєкту (або учасника) стає нульовою.

***2. ОЦІНКА ОЧІКУВАНОГО ЕФЕКТУ БІЗНЕС-ІДЕЇ З УРАХУВАННЯМ КІЛЬКІСНИХ ХАРАКТЕРИСТИК НЕВИЗНАЧЕНОСТІ***

При наявності більш детальної інформації про різні сценарії реалізації проєкту, ймовірності їх здійснення і про значення основних техніко-економічних показників проєкту для кожного із сценаріїв при оцінці ефективності проєкту може бути використаний більш точний метод. Він дозволяє безпосередньо розрахувати узагальнюючий показник ефективності проєкту - очікуваний інтегральний ефект (очікувану ЧТВ). Оцінка очікуваної ефективності проєкту з урахуванням невизначеності проводиться за наявності більш детальної інформації про різні сценарії реалізації проєкту, ймовірності їх здійснення і про значення основних техніко-економічних показників проєкту для кожного із сценаріїв. Така оцінка може проводитися як з урахуванням, так і без урахування схеми фінансування проєкту.

Розрахунки проводяться в наступному порядку:

* описується все безліч всіляких сценаріїв реалізації проєкту (або у формі перерахування, або у вигляді системи обмежень на значення основних технічних, економічних і тому подібних параметрів проєкту);
* по кожному сценарієм досліджується, як буде діяти у відповідних умовах організаційно-економічний механізм реалізації проєкту, як при цьому зміняться грошові потоки учасників;
* для кожного сценарію по кожному кроці розрахункового періоду визначаються (розраховуються або задаються аналітичними виразами) притоки і відтоки реальних грошей і узагальнюючі показники ефективності. За сценаріями, які передбачають «нештатні» ситуації (аварії, стихійні лиха, різкі зміни ринкової кон’юнктури і т.п.), враховуються виникають при цьому додаткові витрати. При визначенні ЧТВ по кожному сценарієм норма дисконту приймається безризиковою;
* перевіряється фінансова реалізація проєкту. Порушення умов реалізованості розглядається як необхідна умова припинення проєкту (при цьому враховуються втрати і доходи учасників, пов’язані з ліквідацією підприємства через його фінансової неспроможності);
* вихідна інформація про фактори невизначеності представляється в формі ймовірностей окремих сценаріїв або інтервалів зміни цих ймовірностей. Тим самим визначається певний клас допустимих (узгоджених з наявною інформацією) імовірнісних розподілів показників ефективності проєкту;
* оцінюється ризик нереализуемости проєкту – сумарна ймовірність сценаріїв, при яких порушуються умови фінансової можливості бути реалізованим проєкту;
* оцінюється ризик неефективності проєкту – сумарна ймовірність сценаріїв, при яких інтегральний ефект (ЧТВ) стає негативним;
* оцінюється середній збиток від реалізації проєкту в разі його неефективності;
* на основі показників окремих сценаріїв визначаються узагальнюючі показники ефективності проєкту з урахуванням факторів невизначеності - показники очікуваної ***ефективності.***

Основними такими показниками, використовуваними для порівняння різних проєктів (варіантів проєкту) і вибору кращого з них, є ***показники очікуваного інтегрального ефекту*** (ЧТВ) (для окремої ділянки). Ці ж показники використовуються для обґрунтування раціональних розмірів і форм резервування і страхування.

Методи визначення показників очікуваного ефекту залежать від наявної інформації про невизначених умовах реалізації проєкту.

***3. ЙМОВІРНІСНА НЕВИЗНАЧЕНІСТЬ***

При ймовірнісної невизначеності по кожному сценарієм вважається відомої (заданої) ймовірність його реалізації. Розподіл усіх опис умов реалізації проєкту виправдано і доцільно, коли ефективність проєкту обумовлена насамперед невизначеністю природно-кліматичних умов (погода, характеристики грунту або запаси корисних копалин, можливість землетрусів чи повеней і т.д.) або процесами експлуатації та зносу основних засобів (зниження міцності конструкцій будівель і споруд, відмови обладнання і т.п.). З певною часткою умовності коливання дефлірованних цін на вироблену продукцію і споживані ресурси можуть описуватися також в імовірнісних термінах.

У разі, коли є кінцеве кількість сценаріїв і ймовірності їх задано, очікуваний інтегральний ефект проєкту розраховується за формулою математичного очікування:

(7.4.)

де – очікуваний інтегральний ефект проєкту; – інтегральний ефект (ЧТВ) при k-м сценарії; – ймовірність реалізації цього сценарію.

При цьому ризик неефективності проєкту () і середній збиток від реалізації проєкту в разі його неефективності () визначаються за формулами

(7.5.)

де підсумовування ведеться тільки за тими сценаріями (k), для яких інтегральні ефекти (ЧТВ) негативні.

Інтегральні ефекти сценаріїв і очікуваний ефект залежать від значення норми дисконту (Е). Премія (g) за ризик неодержання доходів, передбачених основним сценарієм проєкту, визначається з умови рівності між очікуваним ефектом проєкту (Е), розрахованим при безризиковою нормі дисконту Е, і ефектом основного сценарію (Е +g), розрахованим при нормі дисконту Е + g, що включає поправку на ризик:

(Е) = (Е + g).

В цьому випадку середні втрати від неотримання передбачених основним сценарієм доходів при несприятливих сценаріях покриваються середнім виграшем від отримання більш високих доходів при сприятливих сценаріях.

**Приклад 7.1.** Процес функціонування об’єкта розглядається як дискретний і починається з кроку (року) 1. Термін служби об’єкта необмежений. На кожному т-м кроці об’єкт забезпечує отримання невипадкового (річного) ефекту . У той же час проєкт припиняється на певному етапі, якщо на цьому кроці відбувається «катастрофа» (стихійне лихо, серйозна аварія обладнання або поява на ринку більш дешевого продукту-замінника). Імовірність того, що катастрофа станеться на певному етапі за умови, що її не було на попередніх кроках, не залежить від номера кроку і дорівнює .

Очікуваний інтегральний ефект тут визначається наступним чином. Зауважимо насамперед, що ймовірність того, що на кроці 1 катастрофи не станеться, дорівнює . Імовірність того, що її не відбудеться ні на першому, ні на другому кроці, за правилом твори ймовірностей дорівнює і т.д. Тому або до кінця кроку т катастрофи не станеться і ефект проєкту на цьому етапі буде дорівнює , або така подія відбудеться і тоді цей ефект буде дорівнює нулю. Це означає, що математичне очікування (середнє значення) ефекту на даному етапі дорівнюватиме . Підсумовуючи ці величини з урахуванням різночасності, знайдемо математичне сподівання ЧТВ проєкту:

З отриманої формули видно, що різночасові ефекти Фт, що забезпечуються «в нормальних умовах» (тобто за відсутності катастроф), наводяться до базового моменту часу за допомогою коефіцієнтів , що не збігаються з «звичайними» коефіцієнтами дисконтування . Для того щоб «звичайне» дисконтування без урахування факторів ризику і розрахунок з урахуванням цих факторів дали один і той же результат, необхідно, щоб в якості норми дисконту було прийнято інше значення , таке, що . Звідси отримуємо, що . При малих значеннях ця формула набирає вигляду , підтверджуючи, що в даній ситуації облік ризику зводиться до розрахунку ЧТВ «в нормальних умовах», але з нормою дисконту, що перевищує безризикову на величину «премії за ризик»,

***4. ІНТЕРВАЛЬНА НЕВИЗНАЧЕНІСТЬ***

У разі, коли будь-яка інформація про можливості сценаріїв відсутня (відомо тільки, що вони є позитивними і в сумі складають 1), розрахунок очікуваного інтегрального ефекту здійснюється за формулою

(7.6.)

де і – найбільший і найменший інтегральний ефект (ЧТВ) за розглянутими сценаріями; – спеціальний норматив для обліку невизначеності ефекту, що відображає систему переваг відповідного господарюючого суб’єкта в умовах невизначеності.

У загальному випадку, при наявності додаткових обмежень на ймовірності окремих сценаріїв (), розрахунок очікуваного інтегрального ефекту рекомендується проводити за формулою

(7.7.)

де – інтегральний ефект (ЧТВ) при k-м сценарії, а максимум і мінімум розраховуються за всіма допустимим (узгодженим з наявною інформацією) сполученням ймовірностей окремих сценаріїв.

1. Як вже зазначалося вище, віднесення оборотного активу до запасів готової продукції або до дебіторської заборгованості – це певною мірою питання до експерта. Різниця полягає втому, що

   * до тих пір, поки продукція зберігається на складі, на неї принципово може бути змінена ціна (через інфляцію та (або) сезонних змін);
   * на відвантажену продукцію незалежно від її оплати може бути поширена частина податків (відповідно до конкретними законами).

   [↑](#footnote-ref-1)
2. У зв’язку зі зміною податкових норм перелік податків може відрізнятися від наведеного. Наведений перелік відповідає положенню, що всі перераховані податки платяться при відвантаженні товару споживачеві. Якщо якісь з цих податків сплачуються після отримання оплати від споживача, вони повинні бути виключені з величини Вир. [↑](#footnote-ref-2)
3. Якщо в проєкті приймаються різні величини відстрочки платежів для різних їх видів, формула (4) записується у вигляді суми за окремими видами платежів. [↑](#footnote-ref-3)
4. Інші назви: бухгалтерська рентабельність інвестицій ROI (return on investment); середня норма прибутку на інвестиції ARR (average rate of return). [↑](#footnote-ref-4)