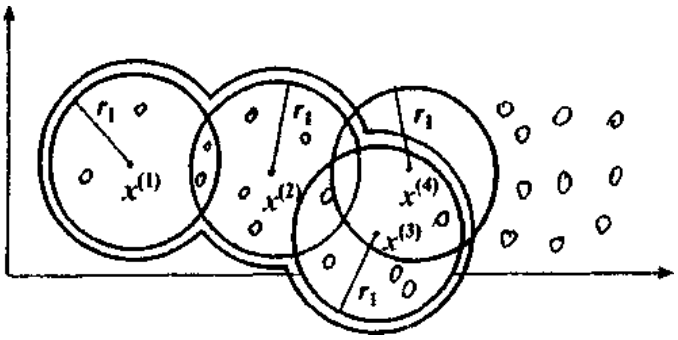
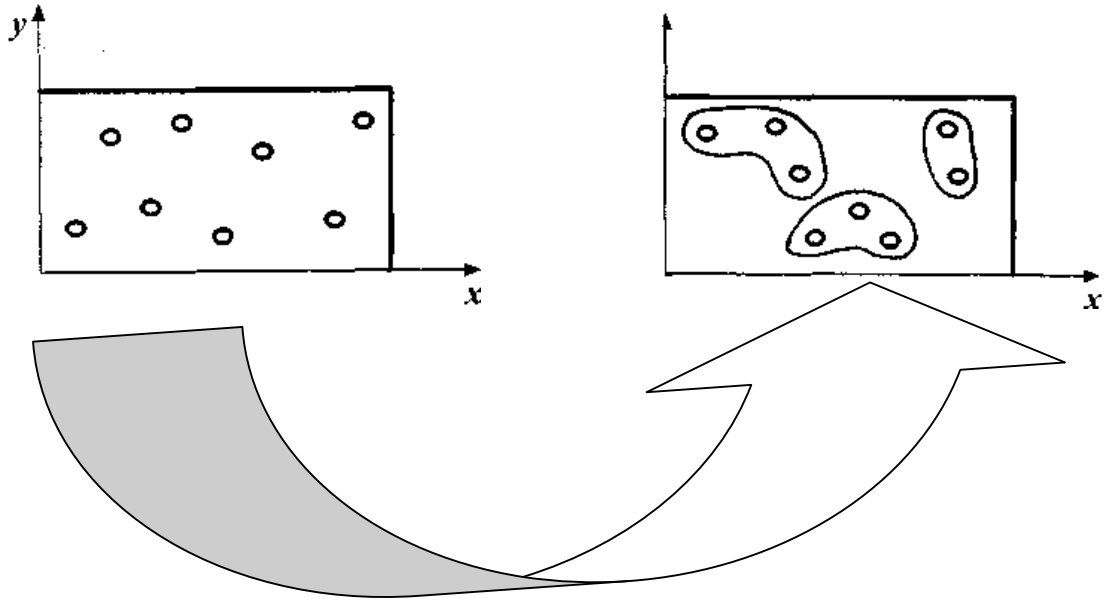
І.М. Пістунов, О.П. Антонюк,
І.Ю. Турчанінова



КЛАСТЕРНИЙ АНАЛІЗ В ЕКОНОМІЦІ



2008
Дніпропетровськ

Міністерство освіти і науки України
Національний гірничий університет

До 110 річчя НГУ

І.М. Пістунов, О.П. Антонюк, І.Ю. Турчанінова

КЛАСТЕРНИЙ АНАЛІЗ В ЕКОНОМІЦІ

(Навчальний посібник)

Дніпропетровськ
НГУ
2008

УДК 368.01 (075):51-7(075)
БКК 605.271
П34

Затверджено вченою радою університету як навчальний посібник по дисципліні "Інформаційний кластерний аналіз в економічний системах" для студентів очної та заочної форм навчання зі спеціальності 8.050102 "Економічна кібернетика" (Протокол № від).

Рецензенти:
Н. К. Васильєва , д-р екон. наук, проф., завідувач кафедри інформаційних систем (Дніпропетровський державний аграрний університет);
О.М. Марюта, д-р. техн. наук, проф., завідувач економічної інформатики і статистики (Дніпропетровський національний університет).

Пістунов І.М., Антонюк О.П., Турчанінова І.Ю.
П34 Кластерний аналіз в економіці: Навч. посібник – Дніпропетровськ: Національний гірничий університет, 2008.– 84 с.

Розглянуто теоретичні і практичні аспекти кластерного аналізу, такі як: розрахунок відстаней між економічними об'єктами, принципи поєднання їх у кластери, прийоми віднесення нових об'єктів до існуючих кластерів, порядок зміни параметрів кластерів, розглянуто критерії якості отриманих розбиттів. В додатках наведено словник спеціальних термінів. Подано перспективи розвитку цього напрямку в науково-технічній і економічній діяльності.

Кожен розділ супроводжують приклади вирішення задач із застосуванням електроннних таблиць Calc з пакету Open Office вільного програмного забезпечення, Microsoft Excel та статистичного пакету MathLab.

Після кожного розділу в посібнику подано завдання для самостійного вирішення, тому він може слугувати і як посібник для практичних чи лабораторних занять із застосуванням комп'ютерної техніки.

Призначено для студентів вищих учбових закладів і може бути корисним для економістів, плановиків, менеджерів та маркетологів.

Посібнику базується на літературних джерелах вітчизняних та зарубіжних авторів та на досвіді викладання дисципліни „Інформаційний кластерний аналіз в економічний системах" в Національному гірничому університеті.

БКК 605.271

ЗМІСТ

ВСТУП

Безсумнівно, класифікація – це основний процес в інтелектуальній діяльності людини. Зустрічаючись з новим явищем, ми намагаємося знайти йому аналог у відомій нам області. Розглядаючи групу яких-небудь об'єктів, ми мимоволі розділяємо їх на підгрупи близьких один одному елементів. Класифікація присутня при упорядкуванні відомих нам фактів, явищ, предметів. На підставі сказаного можна зробити висновок, що класифікація - це фундаментальне поняття науки і практики.

Кластерний аналіз – це сукупність методів, які дозволяють класифікувати багатомірні спостереження, кожне з яких описується набором вихідних перемінних $X_1, X_2, ..., X_m$. Метою кластерного аналізу є утворення груп схожих між собою об'єктів, що прийнято називати кластерами. Слово кластер англійського походження (cluster), переводиться як згусток, пучок, група. Споріднені поняття, використовувані в науковій літературі, – клас, таксон, згущення.

Кластерний аналіз з'явився наприкінці 30-х років нашого сторіччя, але активний розвиток цих методів і їхнє широке використання почався наприкінці 60-х – початку 70-х років.

Техніка кластеризації застосовується в найрізноманітніших областях. Хартиган (Hartigan, 1975) дав прекрасний огляд багатьох опублікованих досліджень, що містять результати, отримані методами кластерного аналізу. Наприклад, в області медицини кластеризація захворювань, лікування чи класифікація симптомів захворювань приводить до широко використовуваних таксономій. В області психіатрії правильна діагностика кластерів симптомів, таких як параноя, шизофренія і т.д., є вирішальної для успішної терапії. В археології за допомогою кластерного аналізу дослідники намагаються установити таксономії кам'яних знарядь, похоронних об'єктів і т.д. Відомі широкі застосування кластерного аналізу в маркетингових дослідженнях. Загалом, усякий раз, коли необхідно класифікувати "гори" інформації до придатного для подальшої обробки вигляду, кластерний аналіз виявляється дуже корисним і ефективним.

Кластерний аналіз – це загальна назва множини обчислювальних процедур, які використовують при створенні класифікації. У результаті роботи з процедурами утворюються класи чи групи подібних об’єктів. Більш точно, кластерний аналіз – це багатомірна статистична процедура, що виконує збір даних, що містять інформацію про вибірку об’єктів, і потім упорядковує об’єкти у порівняно однорідні групи.

Термін кластерний аналіз (уперше ввів *Tryon*, 1939) у дійсності містить у собі набір різних алгоритмів класифікації. Загальне питання, що задається дос-

лідниками в багатьох областях, полягає в тому, як організувати дані, що спостерігаються, у наочні структури, тобто розгорнути таксономії.

Надалі цей напрямок багатомірного аналізу дуже інтенсивно розвивався. З'явилися нові методи, нові модифікації уже відомих алгоритмів, істотно розширилася область застосування кластерного аналізу. Якщо спочатку методи багатомірної класифікації використовувалися в психології, археології, біології, то зараз вони стали активно застосовуватися в соціології, економіці, статистиці, в історичних дослідженнях. Особливо розширилося їхнє використання в зв'язку з появою і розвитком ЕОМ і, зокрема, персональних комп'ютерів. Це зв'язано насамперед із трудомісткістю обробки великих масивів інформації (обчислення і обертання матриць великих розмірностей).

Методи кластерного аналізу дозволяють вирішувати наступні задачі:

• проведення класифікації об'єктів з урахуванням ознак, що відбивають сутність, природу об'єктів. Рішення такої задачі, як правило, приводить до поглиблення знань про сукупності об'єктів, які піддаються класифікації;

• перевірка висунутих припущень про наявність деякої структури в досліджуваній сукупності об'єктів, тобто пошук існуючої структури;

• побудова нових класифікацій для явищ, які вивчені мало, коли необхідно установити наявність зв'язків усередині сукупності і спробувати привнести в неї структуру.

Критерій якості кластеризації тією чи іншою мірою відбиває наступні неформальні вимоги:

а) усередині груп об'єкти повинні бути тісно зв'язані між собою;

б) об'єкти різних груп повинні бути далекими один від іншого;

в) за інших рівних умов розподіл об'єктів по групах повинен бути рівномірним.

Вузловим моментом у кластерному аналізі вважається вибір метрики (або міри близькості об'єктів), від якого вирішальним чином залежить остаточний варіант розбивки об'єктів на групи при заданому алгоритмі розбивки. У кожній конкретній задачі цей вибір проводиться по-своєму, з урахуванням головної мети дослідження, фізичної і статистичної природи використовуваної інформації і т.п.

Алгоритми кластерного аналізу відрізняються великою розмаїтістю. Це можуть бути, наприклад, алгоритми, що реалізують повний перебір сполучень об'єктів або здійснюють випадкові розбивки множини об'єктів. У той же час більшість таких алгоритмів складається з двох етапів. На першому етапі задається початкове (можливо, штучне або навіть довільне) розбиття множини об'єктів

на класи і визначається деякий математичний критерій якості автоматичної класифікації. Потім, на другому етапі, об'єкти переносяться з класу в клас доти, поки значення критерію не перестане поліпшуватися.

Різноманіття алгоритмів кластерного аналізу обумовлено також безліччю різних критеріїв, що виражають ті або інші аспекти якості автоматичного групування. Найпростіший критерій якості безпосередньо базується на величині відстані між кластерами. Однак такий критерій не враховує "населеність" кластерів — відносну щільність розподілу об'єктів усередині виділюваних угруповань. Тому інші критерії ґрунтуються на обчисленні середніх відстаней між об'єктами усередині кластерів. Але найчастіше застосовуються критерії у виді відносин показників "населеності" кластерів до відстані між ними. Це, наприклад, може бути відношення суми відстаней поміж класами до суми внутрішньокласових (між об'єктами) відстаней або відношення загальної дисперсії даних до суми внутрішньокласових дисперсій і дисперсії центрів кластерів.

Кластерний аналіз – один з напрямків статистичного дослідження. Особливо важливе місце він займає в тих галузях науки, що зв'язані з вивченням масових явищ і процесів. Необхідність розвитку методів кластерного аналізу і їхнього використання продиктована насамперед тим, що вони допомагають побудувати науково обґрунтовані класифікації, виявити внутрішні зв'язки між одиницями сукупності, що спостерігаються. Крім того, методи кластерного аналізу можуть використовуватися з метою стиснення інформації, що є важливим чинником в умовах постійного збільшення й ускладнення потоків статистичних даних.

Саме тому велике значення цей тип статистичного аналізу має при аналізі економічної діяльності різних груп підприємств, економічних регіонів, різних країн.

Потреба в об’єктивному розділені різних економічних об’єктів на групи існує постійно, адже саме така класифікація дозволяє потім знайти методи ефективного керування цими об’єктами. А знайти такі методи значно легше, коли вони розробляються в межах однорідної групи.

1. РОЗРАХУНОК ВІДСТАНЕЙ МІЖ ОБ'ЄКТАМИ

Вивчення матеріалу цього розділу дозволить студенту навчитися розраховувати відстані між об'єктами за різними метриками.

1.1. Місце кластерного аналізу серед інших методів автоматичної класифікації

Як вже було сказано у вступі, кластерний аналіз дозволяє провести автоматичну класифікацію об'єктів. Розглянемо інші методи класифікації, щоб зрозуміти межу застосування кластерного аналізу.

РОЗПІЗНАВАННЯ ОБРАЗІВ – процес, при якому на підставі численних характеристик (ознак) якогось об'єкта визначається одна або кілька нових, найістотніших його характеристик, зокрема, його приналежність до певного класу об'єктів. Розв'язати задачу розпізнавання образів – значить за непрямими даними знайти правила, за якими кожному наборові значень ознак якогось об'єкта ставиться у відповідність одне рішення із заданої множини можливих рішень, що визначають істотні характеристики цього об'єкта. Задачами розпізнавання образів є, наприклад, задачі розпізнавання зорових сигналів (рукописних чи друкованих літер і цифр, фотографій реальних об'єктів тощо), звукових сигналів (напр., слів усного мовлення), задачі медичної і технічної діагностики тощо. Істотним тут є те, що одному й тому самому результатові розпізнавання або рішенню відповідає багато вхідних сигналів, відмінність між якими залежить від дії невідомих факторів. Автоматичне розпізнавання образів застосовують для введення інформації в автоматичні системи, наприклад у ЕОМ, та в тих випадках, коли людині важко прийняти рішення через надто велику кількість первісних даних, не пристосованих для людського розпізнавання, наприклад, при діагностиці несправностей механізмів за шумом. Основні поняття і термінологія. У кожній задачі розпізнавання первісними даними є результати спостережень або безпосередніх вимірювань, їх називають первинними ознаками, а сукупність усіх первинних ознак – вхідним сигналом. Наприклад, у випадку, коли розпізнаються звуки, за первинні ознаки можуть правити значення звукового тиску в дискретні моменти часу. Результатом одного акту розпізнавання є рішення, а результатом розв'язання задачі розпізнавання є алгоритм прийняття рішення, або вирішальна функція, яке визначає відображення множини сигналів на множину рішень, тобто для кожного сигналу вказує на певне рішення. Якщо множина рішень дискретна і різних рішень небагато, то розпізнавання можна розглядати як класифікацію.

НЕЙРОННІ СІТКИ – це сітки, що складаються зі зв'язаних між собою

простих елементів - формальних нейронів. Ядром використовуваних уявлень є ідея про те, що нейрони можна моделювати досить простими формулами, а вся складність процесу моделювання визначається зв'язками між нейронами. Кожен зв'язок представляється як зовсім простий елемент, що служить для передачі сигналу. Навчання нейтронної сітки звичайно будується так: існує задачник – набір прикладів із заданими відповідями. Ці приклади пред'являються системі. Нейрони одержують по вхідних зв'язках сигнали – «умови прикладу», перетворюють їх, кілька разів обмінюються перетвореними сигналами і, нарешті, видають відповідь – також набір сигналів. Відхилення від правильної відповіді штрафується. Навчання складається в мінімізації штрафу як (неявної) функції зв'язків. Неявне навчання приводить до того, що структура зв'язків стає «незрозумілою» – не існує іншого способу її прочитати, крім як запустити функціонування сітки. Стає складно побудувати зрозумілу людині логічну конструкцію, що відтворює дії сітки. Зате ця методика не вимагає вибору виду розділяючої функції, оскільки використовується невеликий набір типових функцій, якими сітка комбінує.

ДИСКРИМІНАНТНИЙ АНАЛІЗ – використовується для ухвалення рішення про те, які перемінні розрізняють (дискримінують) дві або більш виникаючі сукупності (групи). Наприклад, якийсь дослідник може захотіти досліджувати, які перемінні відносять випускника середньої школи до однієї з трьох категорій: (1) той, що вступає до коледжу, (2) той, що вступає до професійної школи або (3) той, що відмовляється від подальшої професійної підготовки. Для цієї мети дослідник може зібрати дані про різні змінні, зв'язані з учнями школи. Після випуску більшість учнів природно повинні потрапити в одну з названих категорій. Потім можна використовувати *дискримінантний аналіз* для визначення того, які перемінні дають найкраще пророкування вибору учнями подальшого шляху. Медик може реєструвати різні перемінні, стосовні до стану хворого, щоб з'ясувати, які перемінні краще пророкують, що пацієнт, імовірно, видужав цілком (група 1), частково (група 2) або зовсім не видужав (група 3). Економіст може записати різні характеристики подібних типів (груп) підприємств, щоб потім провести аналіз дискримінантної функції, що щонайкраще розділяє типи або групи.

ДИСПЕРСІЙНИЙ АНАЛІЗ – поставлена вище задача про дискримінантні функції може бути перефразована як задача одно входового дисперсійного аналізу. Можна запитати, зокрема, чи є дві або більш сукупності, що статистично значимо відрізняються одна від іншої за середнім значенням якої-небудь конкретної перемінної. Ясно, що коли середнє значення визначеної перемінної значиме по-різному для двох сукупностей, то ви можете сказати, що перемінна розділяє дані сукупності. У випадку однієї перемінної остаточний критерій значимості того, розділяє перемінна дві сукупності чи ні, дає *F-критерій*. *F*–статистика власне кажучи обчислюється, як відношення між групової дисперсії до об'єднаної внутрішньо групової дисперсії. Якщо між групова дисперсія виявляється істотно більше, тоді це повинно означати розходження

між середніми. При застосуванні дискримінантного аналізу звичайно маються трохи перемінних, і завдання полягає в тім, щоб установити, які з перемінних вносять свій внесок у дискримінацію між сукупностями. У цьому випадку ви маєте матрицю загальних дисперсій і коваріацій, а також матриці внутрішньо групових дисперсій і коваріацій. Ви можете порівняти ці дві матриці за допомогою багатомірного *F-критерію* для того, щоб визначити, чи маються значимі розходження між групами (з погляду всіх перемінних). Імовірно, найбільш загальним застосуванням дискримінантного аналізу є включення в дослідження багатьох перемінних з метою визначення тих з них, що щонайкраще розділяють сукупності між собою. Наприклад, дослідник, що цікавиться передбаченням вибору, який зроблять випускники середньої школи щодо свого подальшого навчання, зробить з метою одержання найбільш точних прогнозів реєстрацію можливо більшої кількості параметрів що навчаються, наприклад, мотивацію, академічну успішність і т.д.

Як видно з викладеного вище, кластерний аналіз корисний у випадках, коли практично невідомо наперед про можливу структуру класів об'єктів. А саме такою, частіше всього, є економіка, з її складністю зв'язків, розмаїтістю видів об'єктів, невизначеністю наслідків від керуючого впливу. Тому кластерний аналіз є найбільш прийнятним при дослідженні економічних систем.

1.2. Вимірювання відстаней між об'єктами

Кожен економічний об'єкт може бути представлений одним і тим же набором факторів або параметрів. Наприклад, торгове підприємство може бути охарактеризоване такими факторами як виторг, обсяг реалізації, валюта балансу, кількість працівників, кількість торгових точок тощо. Позначимо кожен з цих факторів чи параметрів економічного об'єкта як X_i. Тут i – номер фактора, який характеризує об'єкт. Тобто, кожен об'єкт може бути представлений вектором

$$X=(X_1,X_2,...,X_{Nf}) \qquad (1.1)$$

де N_f – кількість факторів. Цей вектор являє собою точку в гіперпросторі, який має розмірність N_f. Наприклад, якщо три об'єкти характеризуються двома факторами (1;2), (2;3) та (3;1) то їх можна представити як три точки на плоскому графіку (рис. 1.1).

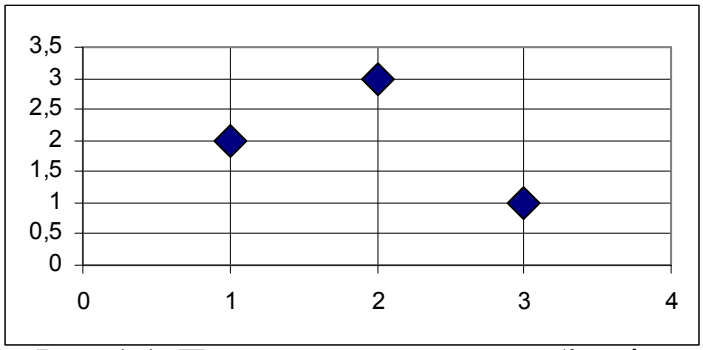


Рис. 1.1. Представлення трьох об'єктів, як точок на площині

І хоча фактори для всіх об'єктів, що розглядаються, є однаковими, їх числові значення будуть відрізнятися. Якщо охарактеризувати інший об'єкт аналогічним (1.1) вектором, який позначимо як Y, можна розрахувати міру близькості цих об'єктів.

Міра близькості або відстань між об'єктами розраховується за допомогою різних формул, які ще називаються метриками відстаней.

Відстань між двома об'єктами позначається як $d(x_i, y_i)$ – це ненегативна функція близькості задається при наступних умовах:

1) Вона завжди більше або дорівнює нулю.
2) Відстань від точки *X* до точки *Y* така сама, як і від *Y* до *X*.
3) Якщо числові значення факторів двох об'єктів однакові, відстань між ними дорівнює нулю.
4) Нехай існує третя точка *U*. Тоді сума відстаней між точками *XU* та *YU* завжди більша ніж відстань поміж точками *XY*.

Або у вигляді формули це записується так:

$$\left.\begin{array}{l} d(x_i, y_i) \geq 0 \\ d(x_i, y_i) = d(y_i, x_i) \\ d(x_i, y_i) = 0 \Leftrightarrow x_i = y_i \\ d(x_i, y_i) \leq d(x_i, u_i) + d(u_i, y_i) \end{array}\right\} \forall \{i\} \in N \qquad (1.2)$$

Найбільш розповсюдженою функцією відстані між двома об'єктами (X; Y) – є відстань у **метриці Евкліда** (d_E)

$$d_E(x_i; y_i) = \sqrt{\sum_{i=1}^{Nf} (x_i - y_i)^2} \qquad (1.3)$$

Метрика Евкліда дозволяє не враховувати знакові розходження, пропорційно збільшує відстань між об'єктами у випадку різних абсолютних значень показників. У результаті збільшується розмірність кластерного поля, об'єкти штучно віддаляються друг від друга, у результаті чого границі між кластерами стають більш чіткими і точними.

Другий по значимості функцією відстані прийнято вважати **міру несхожості Хемінга** (d_{XEM})

$$d_{XEM}(x_i; y_i) = \sum_{i=1}^{Nf} (x_i - y_i) \qquad (1.4)$$

Метрика Хемінга може використовуватися в тих випадках, коли знакові розходження характеристик об'єктів мають принципове значення. За рахунок нівелювання знакових розходжень показників об'єкти виявляються сконцентрованими до області ядра кластера, але при цьому утрачаються важливі знакові характеристики розходжень.

Несуттєво від метрики Евкліда відрізняється і функція відстані в метриці ***L*-норма** (d_L), яка інколи ще називається відстанню міських кварталів або манхеттенською відстанню

$$d_L(x_i; y_i) = \sum_{i=1}^{Nf} |x_i - y_i| \qquad (1.5)$$

Різниця у відстанях, обчислених за метриками у просторі Евкліда і *L*-норми, залежить від абсолютних числових значень і кількості розглянутих показників. У *L* - норми компактність вище, у середньому, на 3...10 відсотків, якщо $(x_i, y_i) \in [1;100]$, а при інтервалі $[0;1]$ компактність менше, у тих же пропорціях. Але з урахуванням визначеної невірогідності вихідних показників подібна різниця може бути визнана несуттєвою.

У деяких умовах класифікації, коли компактність кластерів занадто велика і розділити їх на підмножини досить складно, застосовують метрику «норма - верхня границя» або **метрика Чєбишева** – $(d_{\sup})$

$$d_{SUP}(x_i; y_i) = SUP|x_i - y_i| \tag{1.6}$$

Цей запис означає, що з усіх різниць значень факторів, взятих по модулю, потрібно вибрати одну – найбільшу. І саме вона буде характеристикою відстані між об'єктами. Використання цієї міри відстані може неправомірно змінити картину класифікації через зневагу усіма факторами крім одного. Тут необхідно теоретичне обґрунтування коректності обраної міри $(d_{\sup})$.

Узагалі, дослідник може використовувати безліч функцій відстані, запропонованих у різних роботах по кластерному аналізу, або власних. Відзначимо ще трохи функцій відстані, що згадуються часто.

Степенева відстань. Іноді бажають прогресивно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються. Це може бути досягнуто з використанням *степеневої відстані*:

$$d_S(x_i; y_i) = \left(\sum_{i=1}^{Nf} |x_i - y_i|^p \right)^{\frac{1}{r}} \tag{1.7}$$

де r і p - параметри, визначувані користувачем. Декілька прикладів обчислень можуть показати, як "працює" ця метрика. Параметр p відповідальний за поступове зважування різниць по окремих координатах, параметр r відповідальний за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри - r і p, рівні двом, то ця відстань співпадає з відстанню Евкліда.

Функція **відстані Джеффріса-Матусіти**

$$d_M(x_i; y_i) = \sqrt{\sum_{i=1}^{Nf} (\sqrt{x_i} - \sqrt{y_i})^2} \tag{1.8}$$

Функція Махаланобіса. Ця функція узагальнює можливі варіанти використання метрики Евкліда. Вона задана в матричній формі. Перетворення T (транспонування матриці) кореляційної матриці інваріантно-невиродженим лінійним перетворенням

$$D^2 = (X_i, X_j) = (X_i - X_j)^T W^{-1} (X_i - X_j), \tag{1.9}$$

де $\overline{X} = \sum_{i=1}^{n_1} x_i / n_1$, $\overline{Y} = \sum_{i=1}^{n_2} y_i / n_2$ – статистичні відстані між кластерами. W^{-1} – матриця, зворотна матриці розсіяння

$$W = \frac{n_1 n_2}{n_1 + n_2} (\overline{X} - \overline{Y})(\overline{X} - \overline{Y})^T \tag{1.10}$$

№ об'єкта	X_1	X_2

1	1	2
2	1	3
3	2	2
4	3	1

Приклад

Таблиця 1.1

Побудувати матрицю відстаней для 4-х об'єктів, представлених двома факторами у наведеній нижче таблиці, за метрикою Махаланобіса.

Рішення цієї задачі будемо виконувати в пакеті MATLAB. Задамо вхідні дані: *X* = [1 2; 1 3; 2 2; 3 1]

Для розрахунку відстані використовуємо функцію *Y=pdist*(*X*,'*metric*') – де *X* – вектор вхідних даних, '*metric*' – вид функції відстані (можливі значення '*Euclid*' – метрика Евкліда, '*Mahal*' – метрика Махаланобіса, '*CityBlock*'– манхетенсь

Таблиця 1.2

ка відстань). Для розрахунку відстані за метрикою Махаланобіса скористаємося *Y=pdist*(*X*,'*mahal*'), результати роботи функції наведені в наступній таблиці, де в заголовках рядку і стовпця стоять номери об'єктів.

	1	2	3	4
1	0	2.35	2.00	2.35
2	2.35	0	1.23	2.45
3	2.00	1.23	0	1.23
4	2.35	2.45	1.23	0

1.3. Нормування числових значень факторів

Якщо ми маємо тільки 2 об'єкта, то міра відстані поміж ними у нас буде одна. Вона може бути розрахована за будь-якою метрикою. Якщо ми будемо розглядати 3 об'єкти, то таких мір ми вже матимемо 3 – поміж першим і другим, поміж першим і третім, поміж другим і третім. Якщо будемо розглядати чотири об'єкти, то число мір відстані вже буде 6 – поміж першим та другим, третім, четвертим, поміж другим та третім, четвертим, поміж третім та четвертим. Загалом, кількість мір відстаней буде становити

$$M = \frac{N_0(N_0 - 1)}{2} \tag{1.11}$$

де N_0 – кількість об'єктів, відстань між якими ми розраховуємо.

Отже, найзручніше всі розраховані відстані вмістити у матрицю, розміром N_0xN_0. Вона матиме вигляд

$$D = \begin{vmatrix} d_{11} & d_{12} \ldots\ldots\ldots\ldots d_{1N_0} \\ d_{21} & d_{22} \ldots\ldots\ldots\ldots d_{2N_0} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ d_{N_0 1} & d_{N_0 2} \ldots\ldots d_{N_0 N_0} \end{vmatrix} \tag{1.12}$$

Діагональні елементи цієї матриці дорівнюють нулю, згідно третьої умови з (1.2). Вся матриця є діагонально симетричною, тобто, $d_{ji}=d_{ji}$. Це означає, що її можна представити у вигляді

$$D = \begin{vmatrix} 0 & & \\ d_{21} & 0 & \\ \ldots\ldots\ldots\ldots & 0 & \\ d_{N_0 1} & d_{N_0 2} & \ldots\ldots 0 \end{vmatrix} \tag{1.13}$$

Розглядаючи формули (1.3) – (1.10), можна помітити, що основу всіх метрик складають суми різностей чисельних значень відповідних факторів. Але ж абсолютні величини цих факторів можуть бути зовсім різними, тому, наприклад метрика Чєбишева (1.6) не буде давати об'єктивного результату.

Наприклад, якщо ми розглядаємо економічні об'єкти з двома факторами – валюта балансу та рентабельність – то числові їх значення можуть відрізнятися на 2-3 порядки. Скажімо, у першого об'єкта це відповідно 230 000 грн. та 15%, а у другого – 220 000 грн. та 10%. Різниці цих факторів будуть дорівнювати відповідно 10000 та 5. Метрика (1.6), за якою треба у якості міри відстані вибрати найбільшу різницю, дасть нам відповідь – 10000. Але це буде необ'єктивний результат. Адже для валюти балансу 10000 означає зміну показника тільки на 5%, а зміна рентабельності від 10% до 15% складає аж 50% різниці.

Безпосереднє використання змінних при аналізі може привести до того, що класифікацію будуть визначать змінні, що мають великий діапазон значень. Перед розрахунком міри відстані потрібно якимось чином привести всі фактори до одного масштабу. Тому застосовують наступні види нормалізації даних:

- *Z*–шкали (*Z*-Scores). Із значень змінних віднімається їх середнє і ці значення діляться на стандартне відхилення.

$$X_{ij}^{\text{н}} = \frac{X_{ij} - M_{Xi}}{\sigma_{Xi}} + 4 \tag{1.14}$$

- Розкид від -1 до 1. Лінійним перетворенням змінних добиваються розкиду значень від -1 до 1.

$$X_{ij}^{\text{н}} = \frac{X_{ij} - \overline{X}}{X_{\max} - X_{\min}} \tag{1.15}$$

- Розкид от 0 до 1. Лінійним перетворенням змінних добиваються розкиду значень від 0 до 1.

$$X_{ij}^{\text{н}} = \frac{X_{ij} - X_{\min}}{X_{\max} - X_{\min}} \tag{1.16}$$

- Максимум 1. Значення змінних діляться на їх максимум.

$$X_{ij}^{\text{н}} = \frac{X_{ij}}{X_{\max}} \tag{1.17}$$

- Середнє 1. Значення змінних діляться на їх середнє.

$$X_{ij}^{\text{і}} = \frac{X_{ij}}{\overline{X}} \tag{1.18}$$

- Стандартне відхилення 1. Значення змінних діляться на стандартне відхилення $$X_{ij}^{н} = \frac{X_{ij}}{\sigma_{X_i}} \quad (1.19)$$

Найпростіший з них, обрати серед факторів X_{ij} при фіксованому значенні i ($1 \le i \le N_f$ – кількість факторів, за якими розглядається об'єкти, $1 \le j \le N_O$ – кількість об'єктів) будь яке значення X_{ij}* – найбільше, найменше, середнє.

А потім для кожного фіксованого значення і виконати перетворення

$$X_{ij}^{н} = \frac{X_{ij}}{X_{ij}^{*}} \quad (1.20)$$

Таке перетворення є досить швидким, але не дає гарантії, що всі нормовані значення для всіх факторів попадуть в один діапазон значень. Тому пропонується наступна процедура:

1. Для кожного фактора розраховується його середнє значення для всіх об'єктів $$M_{Xi} = \frac{1}{N_0}\sum_{j=1}^{N_0} X_{ij} \quad (1.21)$$

2. Знаходиться його математичний стандарт (середнє квадратичне відхилення) $$\sigma_{Xi} = \sqrt{\frac{1}{N_0 - 1}\sum_{j=1}^{N_0} X_{ij}^2 - M_{Xi}^2} \quad (1.22)$$

3. Виконати нормування $$X_{ij}^{н} = \frac{X_{ij} - M_{Xi}}{\sigma_{Xi}} + 4 \quad (1.23)$$

Таке нормування з імовірністю 98% переведе всі значення у діапазон [+1;+7]. А отже, буде дотримана ще й умова підвищення компактності для L - метрики (1.5).

Приклад

Таблиця 1.1

№ об'єкта	X_1	X_2	X_3	X_4
1	87	0,39	560	2770
2	25	0,82	430	2590
3	67	0,29	270	2870
4	62	0,52	860	1920
5	53	0,54	790	2770

Побудувати матрицю відстаней для 5-ти об'єктів, представлених чотирма факторами у наведеній нижче таблиці, за метрикою Джеффріса-Матусіти. Рішення цієї задачі будемо виконувати у додатку Calc пакету Open Office. Спочатку проведемо нормування таблиці значень. Для цього розрахуємо середні, із застосуванням функції *AVERAGE*(B2:B6), де через двокрапку вказано діапазон адрес клітинок, які містять зміни значення першого фактора для всіх 5-ти об'єктів. Далі знаходимо стандарт, використовуючи функцію *STDEVA*(B2:B6), де так само подано діапазон клітинок для 1-го фактора. І нарешті, за допомогою формули *STANDARDIZE*(B2;B7;B8)+4 виконуємо нормування. Тут

Таблиця 1.3.

№ об'єкта	X_1	X_2	X_3	X_4
1	5,25	3,39	3,91	4,48
2	2,51	5,54	3,38	4,02
3	4,36	2,89	2,73	4,74
4	4,14	4,04	5,13	2,27
5	3,74	4,14	4,85	4,48

перше число – адреса клітинки, яка має бути нормована, 2-ге – адреса клітинки, де є середнє, 3-є – клітинка, де є стандарт.

Результат нормування представлено у наступній таблиці, а вікно функції *STANDARDIZE* – на рис. 1.2.

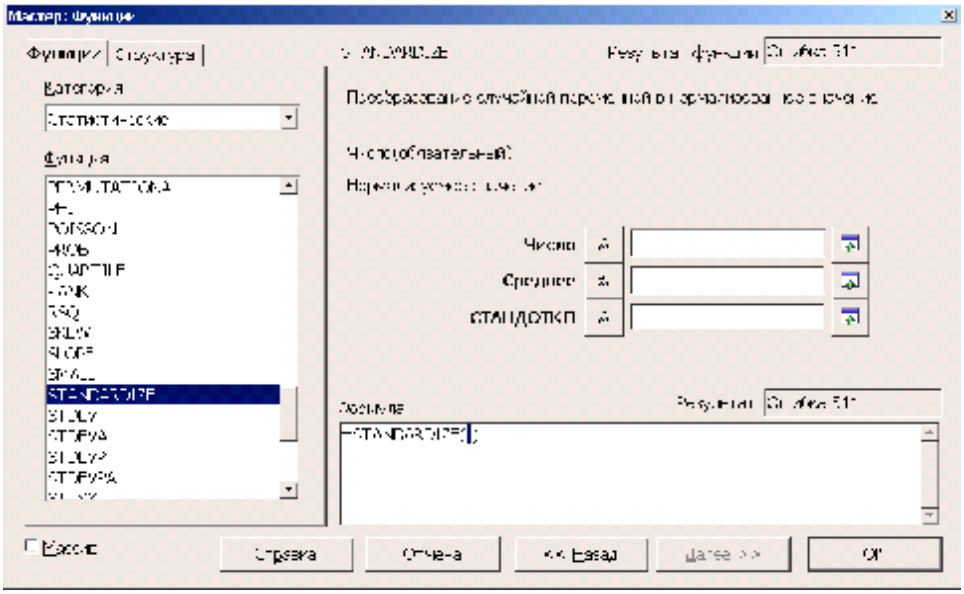


Рис. 1.2. Вікно функції *STANDARDIZE* електронних таблиць Calc

Знайдемо тепер матрицю відстаней за метрикою Джеффріса-Матусіти (1.8). Матриця відстаней буде мати розмір 5x5.

Спочатку знаходимо різницю коренів квадратних та зводимо їх у квадрат для кожної пари факторів за допомогою формули (*SQRT*(B13)-*SQRT*(C13))^2, тут B13 та C13 – адреси відповідних клітинок, які містять однакові фактори для різних об'єктів. Далі – знаходимо корінь квадратний з їх суми за формулою *SQRT*(*SUM*(B15:B19)). І задача розв'язана.

Таблиця 1.4

	1	2	3	4	5
1	0				
2	0,89	0			
3	0,41	0,86	0		
4	0,74	0,87	0,96	0	
5	0,46	0,61	0,66	0,62	0

Ці розрахунки показано у таблиці, де вказано номера об'єктів та відстані поміж ними за метрикою Джеффріса-Матусіти.

Як видно з розрахунків, найменшу відстань мають об'єкти 1-5 (0,46), а найбільшу – 3-4 (0,96).

Виконаємо тепер завдання цього прикладу в обчислювальному середовищі Microsoft Excel, тобто побудуємо матрицю відстаней для 5-ти об'єктів,

представлених чотирма факторами у наведеній нижче таблиці, за метрикою Джеффріса-Матусіти.

Спочатку проведемо нормування таблиці значень. Для цього розрахуємо середні, із застосуванням функції *СРЗНАЧ*(B2:B6), де через двокрапку вказано діапазон адрес клітинок, які містять зміни значення першого фактора для всіх 5-ти об'єктів. Далі знаходимо стандарт, використовуючи функцію *СТАНДОТКЛОН*(B2:B6), де так само подано діапазон клітинок для 1-го фактора. І нарешті, за допомогою формули *НОРМАЛИЗАЦИЯ*(B2;B7;B8)+4 виконуємо нормування. Тут перше число – адреса клітинки, яка має бути нормована, 2-ге – адреса клітинки, де є середнє, 3-є – клітинка, де є стандарт.

Таблиця 1.5

№ об'єкта	X_1	X_2	X_3	X_4
1	5,25	3,39	3,91	4,48
2	2,51	5,54	3,38	4,02
3	4,36	2,89	2,73	4,74
4	4,14	4,04	5,13	2,27
5	3,74	4,14	4,85	4,48

Результат нормування представлено у наступній таблиці, а вікно функції *НОРМАЛИЗАЦИЯ* – на рис. 1.3.

Знайдемо тепер матрицю відстаней за метрикою Джеффріса-Матусіти (1.8). Матриця відстаней буде симетричною і матиме розмір 5x5.

Спочатку знаходимо різницю коренів квадратних та зводимо їх у квадрат для кожної пари факторів за допомогою формули (*КОРЕНЬ*(B13)-*КОРЕНЬ*(C13))^2, тут B13 та C13 – адреси відповідних клітинок, які містять однакові фактори для різних об'єктів. Далі – знаходимо корінь квадратний з їх суми за формулою *КОРЕНЬ*(*СУММ*(B15:B19)). Ці розрахунки показано у наступній таблиці, де вказано номера об'єктів та відстані поміж ними за метрикою Джеффріса-Матусіти. Як видно з розрахунків, найменшу відстань мають об'єкти 1-5 (0,46), а найбільшу – 3-4 (0,96).

Таблиця 1.6

	1	2	3	4	5
1	0				
2	0,89	0			
3	0,41	0,86	0		
4	0,74	0,87	0,96	0	
5	0,46	0,61	0,66	0,62	0

1.4. Індивідуальне завдання №1. Розрахунок відстаней між об'єктами за різними метриками

Мета завдання: Вивчити методи розрахунків відстаней між об'єктами за допомогою різних метрик.

Кожному студенту надаються числові значення 9-ти параметрів для 10-ти об'єктів згідно з номера за списком групи. Застосувавши електронні таблиці Calc з пакету Open Office вільного програмного забезпечення, потрібно розрахувати матриці відстаней за метриками Евкліда, Чєбишева, Хемінга, Джеффрі-

са-Матусіти, степенної, „кварталів”, *L*-метрики. Результати представити в електронному вигляді, як елементи електронних таблиць Calc, придатні для подальших розрахунків. При розрахунку степенної метрики p дорівнює номеру варіанта, а r – останній цифрі у номері залікової книжки студента. Перед розрахунком проведіть нормування факторів X_i.

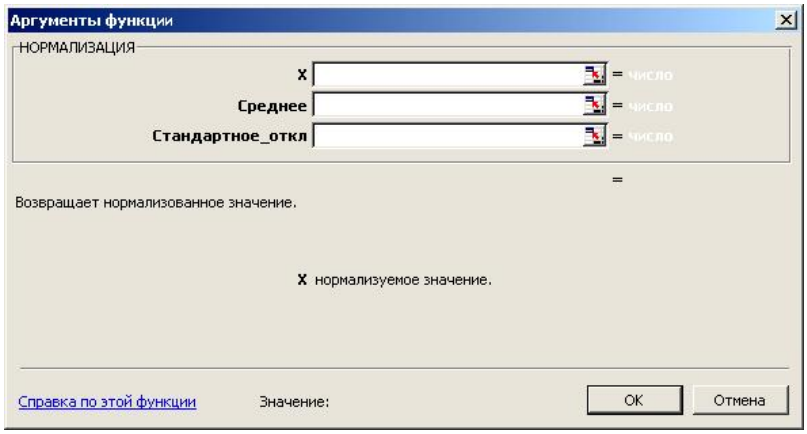


Рис. 1.3. Вікно функції *НОРМАЛИЗАЦИЯ* електронних таблиць Calc

Таблиця 1.6

Варіанти завдань

№ п/п	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	64	0,29	340	2710	304,92	8368,7	4,98953	4,43939	19602
	31	0,73	370	2770	279,51	8483,3	7,67620	3,10606	7194
	39	0,69	470	2780	415,03	4098,3	5,47801	1,90909	9504
	44	0,54	290	1280	145,68	4728,9	4,74529	1,86363	11814
	87	0,64	340	1460	272,73	5674,6	5,61758	1,84848	15378
	74	0,76	360	1220	448,91	2866	8,72295	1,27272	5148
	54	0,54	710	2480	149,07	8110,7	5,1291	4,10606	18084
	74	0,64	770	1790	220,22	7652,2	6,28053	2,80303	13992
	78	0,85	300	1690	150,76	2780,0	9,35101	4,75757	17622
	87	0,39	560	2770	501,42	8311,4	6,07117	3,96969	19074
2	25	0,82	430	2590	164,31	5416,7	4,50104	4,46969	20262
	67	0,29	270	2870	492,95	3525,1	10,8513	2,98484	7326
	62	0,52	860	1920	528,52	4069,7	9,24633	3,19697	13530
	53	0,54	790	2770	133,82	3926,4	4,67550	4,65151	13068
	42	0,42	610	1100	232,07	5474,0	4,29169	1,72727	6270
	64	0,29	700	2860	267,65	7107,6	9,69993	2,45454	14916
	30	0,57	550	2150	531,91	6305,2	8,72295	2,18181	8646
	41	0,88	800	1840	499,73	6276,5	6,28053	2,10606	12342
	68	0,37	740	1220	381,15	6047,2	5,82693	4,50000	20592
	49	0,58	490	2430	474,32	4384,9	4,57083	4,22727	8976
3	76	0,49	750	3100	247,32	5560,0	8,65317	3,62121	11616
	35	0,49	380	1850	152,46	5760,6	4,04745	1,62121	5214
	43	0,38	290	840	138,90	4958,1	10,5722	1,34848	14916
	52	0,57	860	2030	218,52	8769,9	10,9909	3,28787	10164
	72	0,26	620	1450	203,28	8827,2	9,76971	2,43939	6270

№ п/п	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
	73	0,28	460	1360	470,93	7824,1	6,62944	2,62121	8844
	88	0,37	750	2230	433,66	8712,6	10,8513	3,96969	9438
	42	0,59	770	1520	252,40	7795,5	10,6420	3,92424	5412
	69	0,72	740	990	320,16	3983,7	3,66364	4,45454	6204
	75	0,31	380	2830	135,52	4470,9	8,51360	4,33333	18150
4	33	0,26	390	2420	526,83	2636,7	3,24494	4,09090	19338
	41	0,29	760	1170	164,31	5846,6	8,09490	4,40909	20328
	69	0,75	640	1890	531,91	5216,1	10,1535	1,68181	11352
	66	0,54	860	2880	177,87	5846,6	3,21004	4,07575	13398
	38	0,52	360	2730	238,85	5531,3	3,66364	4,65151	19536
	57	0,5	680	1230	476,01	7709,5	8,44382	3,06060	8316
	58	0,59	410	2620	367,59	2264,1	2,82623	2,16666	7326
	33	0,69	310	1280	513,28	8110,7	5,23377	3,72727	18216
	60	0,89	800	1630	215,13	6878,4	8,79274	2,31818	14058
	32	0,76	500	2370	216,83	2350,1	6,14096	1,40909	17028
5	59	0,46	460	2300	442,13	4528,2	6,38520	3,83333	13464
	68	0,5	470	1560	282,89	3238,5	3,83810	3,09090	8316
	34	0,34	330	2070	430,27	7050,3	3,03559	2,39393	17754
	38	0,76	400	850	509,89	6706,4	8,19958	3,10606	19866
	56	0,76	890	1140	164,31	5674,6	5,58269	3,54545	5610
	84	0,43	870	1860	203,28	6591,8	4,08234	3,10606	13398
	86	0,65	540	820	166,01	2665,3	5,16399	4,09090	16500
	87	0,6	310	1740	481,09	7594,9	9,66503	3,33333	15840
	60	0,89	250	1230	370,98	7852,8	2,68667	4,43939	19602
	28	0,58	870	1670	210,05	4442,3	6,87369	4,03030	7062
6	86	0,41	510	2140	218,52	8941,9	10,2233	3,31818	17094
	81	0,33	400	1270	226,99	5359,4	5,26866	4,45454	18084
	58	0,74	630	2670	440,44	8769,9	4,67550	2,15151	6336
	40	0,57	490	810	159,23	5703,3	3,38450	2,92424	11352
	55	0,67	460	3060	181,25	4155,7	4,53593	1,98484	9306
	53	0,42	390	920	404,86	8110,7	5,09420	4,07575	9636
	44	0,78	400	2730	272,73	4155,7	4,81507	2,30303	11088
	59	0,79	310	980	143,99	4384,9	8,40893	1,59090	15246
	29	0,45	890	2230	164,31	4069,7	10,9211	1,60606	16962
	88	0,7	330	1890	274,42	2722,7	5,02442	4,18181	7062
7	25	0,42	580	1810	399,78	5101,4	8,12979	4,34848	20790
	69	0,7	800	3040	362,51	5646,0	2,79134	1,54545	10164
	74	0,46	570	2950	459,07	8082,1	3,52407	1,62121	17490
	72	0,52	360	1810	499,73	7766,8	3,83810	3,92424	7062
	86	0,68	790	1610	481,09	2206,8	10,0837	1,40909	20460
	63	0,51	790	2460	442,13	2436,1	10,7815	1,25757	9108
	55	0,62	770	2740	171,09	2579,4	6,21074	2,68181	11946
	84	0,83	820	990	433,66	6047,2	10,3628	2,60606	12540
	51	0,54	740	1610	447,21	4127,0	5,05931	4,77272	15708
	87	0,31	490	870	509,89	2780,0	7,25750	4,34848	15642
8	53	0,5	440	2510	333,71	5273,4	7,53663	4,77272	5214
	36	0,65	350	860	282,89	3840,4	9,63014	1,90909	19866
	32	0,88	550	2350	506,50	8282,7	8,33914	3,25757	9504
	71	0,46	430	1830	171,09	5187,4	7,11793	1,90909	7194

№ п/п	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
	67	0,89	750	3090	501,42	8082,1	6,35031	3,30303	5940
	77	0,43	330	1290	225,30	5388,0	5,23377	1,46969	5280
	40	0,64	290	2840	465,85	5072,8	5,02442	3,71212	13926
	54	0,86	850	2720	354,04	4872,2	4,46615	1,68181	18612
	49	0,68	820	3130	182,95	3152,6	10,6769	2,62121	17754
	49	0,71	250	2190	409,94	2436,1	7,85066	2,54545	19536
	48	0,54	800	1670	242,24	7222,3	10,5024	1,78787	18480
	29	0,54	250	1870	154,15	7136,3	8,82763	1,68181	16962
	82	0,62	540	790	169,4	4213,0	6,55966	4,66666	19932
	53	0,76	440	1750	169,4	7938,8	8,16468	3,06060	13200
	68	0,88	820	1000	262,57	4814,8	2,89602	3,75757	6534
9	36	0,69	530	2480	188,03	3295,9	5,37334	1,74242	14520
	82	0,86	550	2700	304,92	8827,2	2,96580	1,57575	12408
	82	0,86	820	2980	389,62	6161,9	7,99023	2,37878	10296
	27	0,73	410	1200	514,97	6563,1	7,67620	3,09090	8712
	80	0,34	690	1380	409,94	8368,7	9,59525	4,65151	14916
	80	0,7	520	2690	154,15	5273,4	10,3977	3,87878	11352
	75	0,39	600	1460	525,14	4213,0	4,53593	3,42424	15444
	83	0,38	490	3130	469,23	7021,7	9,24633	4,27272	6204
	38	0,64	870	890	499,73	6419,8	6,35031	4,39393	7524
	72	0,77	440	1900	398,09	4127,0	10,7117	3,33333	9240
10	68	0,31	690	1370	437,05	3525,1	10,8164	3,45454	9570
	40	0,68	460	1940	269,34	3754,4	3,52407	2,74242	8052
	71	0,67	730	1400	367,59	6849,7	8,40893	3,68181	17160
	83	0,75	330	2550	479,40	7365,6	5,16399	4,74242	20394
	89	0,51	440	3110	260,87	7537,5	5,1291	3,53030	8976
	89	0,51	610	2180	160,93	4786,2	7,78087	4,46969	7788
	36	0,81	270	1580	453,99	2522,0	4,71039	3,07575	7920
	83	0,79	730	2110	406,56	3353,2	8,09490	3,62121	20592
	82	0,54	280	2890	311,69	4814,8	7,32728	3,5	19404
	25	0,53	760	1190	176,17	7824,1	4,32658	3,98484	5940
11	67	0,69	730	2680	462,46	7365,6	4,08234	1,27272	18942
	85	0,51	400	1160	518,36	6849,7	10,0139	1,56060	13596
	62	0,54	330	2510	287,98	7623,5	7,92044	4,77272	10824
	35	0,36	390	1820	428,58	5588,7	4,95464	2,10606	16038
	53	0,28	660	2560	428,58	7652,2	6,00139	1,78787	18282
	44	0,45	440	1390	359,12	8426,0	7,57152	2,51515	7854
	44	0,83	540	1670	282,89	4012,4	10,3279	3,27272	13794
	58	0,51	330	1200	489,56	7967,4	10,3279	2,90909	11682
12	61	0,3	500	2220	506,50	7222,3	4,60572	2,98484	13596
	26	0,88	460	3060	237,16	6591,8	4,46615	1,77272	18810
	65	0,71	880	2220	262,57	7824,1	10,1186	1,78787	19404

№ п/п	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
	76	0,68	680	2160	308,30	6649,1	4,39637	4,42424	19866
	66	0,62	870	2330	343,88	7910,1	10,8513	2,51515	16038
	75	0,53	480	2170	208,36	3467,8	8,61828	2	10362
	66	0,51	410	2830	303,22	2464,7	9,59525	4,34848	6930
13	51	0,74	470	2290	411,64	3066,6	9,56036	1,63636	9570
	45	0,6	660	2820	374,37	4069,7	9,59525	4,71212	12738
	72	0,31	540	1290	425,19	4413,6	7,88555	1,80303	18216
	68	0,82	680	1020	477,70	6219,2	6,38520	1,83333	20328
	47	0,62	720	1690	340,49	6104,5	10,8513	3,63636	16170
	50	0,39	860	2410	430,27	8483,3	8,51360	1,46969	13926
	46	0,27	370	1260	213,44	6563,1	6,52477	4,28787	5082
	75	0,52	760	2020	177,87	6591,8	3,00069	1,25757	8778
	38	0,51	810	2110	250,71	3267,2	6,73412	4,74242	5544
	44	0,31	560	1540	294,75	3697,1	5,19888	1,83333	15840
14	27	0,67	750	1980	238,85	6161,9	8,19958	3,86363	11880
	47	0,5	800	2650	477,70	7766,8	9,69993	2,27272	16500
	36	0,73	810	1910	421,80	3381,8	4,98953	1,60606	16500
	54	0,67	700	2610	442,13	3496,5	9,66503	3,46969	10692
	51	0,31	630	1740	333,71	5416,7	9,42079	2,75757	12540
	43	0,34	850	1610	289,67	6534,4	3,87299	1,78787	8844
	76	0,68	530	1680	201,58	6419,8	6,38520	3,16666	9702
	53	0,85	250	1650	221,91	7394,2	4,53593	1,98484	13398
	26	0,3	370	1600	208,36	8512,0	2,86113	3,21212	12474
	39	0,33	610	2620	520,05	3037,9	8,75785	4,65151	10494
15	42	0,37	410	1260	362,51	8941,9	2,86113	2,28787	17226
	33	0,88	850	1170	216,83	2206,8	5,58269	4,60606	9042
	41	0,72	490	2410	413,33	7021,7	6,00139	2,30303	9372
	48	0,5	570	1960	496,34	3553,8	7,43196	4,22727	15312
	60	0,61	850	2380	513,28	4184,3	5,86182	2,16666	20262
	83	0,57	860	1430	238,85	2894,6	9,69993	1,36363	8250
	37	0,87	350	2390	243,93	7222,3	10,4326	4,18181	6138
	53	0,84	260	2150	282,89	4241,6	3,94277	3,81818	13992
	65	0,66	650	850	210,05	6362,5	8,51360	3,40909	7458
	71	0,26	590	1520	281,20	3123,9	9,42079	2,40909	8118

Контрольні запитання

1. Чому кластерний аналіз більш прийнятний для класифікації економічних об'єктів?
2. Що таке метрика відстані?
3. Яка з метрик є найбільш популярною?
4. Як виконати нормування чисельних значень факторів?

В розділі розглянуто порядок розрахунків метрик відстаней поміж економічними об'єктами з формуванням матриці відстаней.

2. АЛГОРИТМИ УТВОРЕННЯ КЛАСТЕРІВ

Вивчення матеріалу цього розділу дозволить студенту навчитися визначати різними методами групи кластерів, та їх характеристики: центр та радіус гіперсфери.

На відміну від комбінаційних угруповань кластерний аналіз приводить до розбивки на групи з обліком усіх ознак одночасно. Наприклад, якщо кожен об'єкт, що спостерігається, характеризується двома ознаками X_1 і X_2, то при виконанні комбінаційного угрупування вся сукупність об'єктів буде розбита на групи по X_1, а потім усередині кожної виділеної групи будуть утворені підгрупи по X_2. Такий підхід одержав назву *монотетичного*. У кластерному аналізі використовується інший принцип утворення груп, так називаний *політетичний* підхід. Усі ознаки одночасно беруть участь в угрупованні, тобто вони враховуються усі відразу при віднесенні спостереження в ту або іншу групу. При цьому, як правило, не зазначені чіткі границі кожної групи, а також невідомо заздалегідь, скільки ж груп доцільно виділити в досліджуваній сукупності.

Приміром, якщо в дослідженні беруть участь N характеристик (ознак, факторів), а метрична відстань між елементами множини визначається функцією $d(x;y)$, то цільова функція максимальної близькості має

$$Z = \sum_{k=1}^{K} d(x, y) \to \min, \qquad (2.1)$$

де K – кількість елементів досліджуваної множини (кластера); $\{x_i, y_i\}$ – елементи множини або їхньої координати, один із яких може бути прийнятий за центр згущення, але, за бажанням дослідника, ці елементи можуть бути і незалежними.

Розглянемо наочний випадок виміру відстаней між об'єктами, кожний з яких має координати $\{x_i, y_i\}$ (рис. 2.1-2.2).

На рис. 2.1 представлено кластери, що складаються з одного об'єкта, бо їх розподіл є однорідним. На рис. 2.2 однорідність груп нижче, ніж на рис. 2.1. Міжгрупові відстані досить великі для опису відмінностей кластерів одного від іншого.

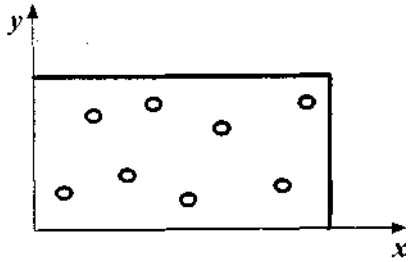


Рис. 2.1. Рівномірний розподіл об'єктів у кластерному полі

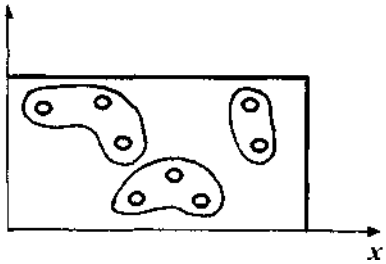


Рис. 2.2. Випадок нерівномірного розподілу

2.1. Кластеризація повним перебором об'єктів

Методично цей спосіб кластеризації найбільш простий, але досить трудомісткий.

1. Складемо вихідну матрицю спостережень над об'єктами.

2. Одержимо матрицю значень відстаней від довільно обраного об'єкта (його числової характеристики).

3. Введемо поняття приналежності *i*-го об’єкта до *k*-го кластера. Це буде матриця Q розмірності N_OxN_O, де N_O – кількість об’єктів, які розглядаються. В ній по стовпцям розташовані номера об’єктів, а по рядках – номера кластерів. Припускається, що кількість кластерів буде дорівнювати кількості об’єктів. Елементи цієї матриці представляють собою бінарні числа, тобто такі, які можуть приймати значення тільки 0 або 1.

4. Введемо цільову функцію, що відповідає обраному критерію внутрішньо групової однорідності об'єктів за прикладом (2.1)

$$Z = \sum_{i=1}^{N_0} \sum_{j=1}^{N_0} q_{ij} d_{ij} \to \min \tag{2.2}$$

де q_{ij} – елемент матриці Q, d_{ij} – метрика відстані поміж об’єктами ($1 \le i,j \le N_O$).

5. Додамо до цільової функції обмеження

$$\sum_{i=1}^{N_0} q_{ij} \le N_0 \tag{2.3}$$

$$\sum_{j=1}^{N_0} q_{ij} = 1 \tag{2.4}$$

Перше обмеження означає, що сума елементів q_{ij} по рядку не повинна перевищувати числа об’єктів, друге – що один і той же об’єкт не може бути включений до двох чи більше кластерів.

6. Останнє обмеження показує, що число об’єктів, включених до різних кластерів має дорівнювати їх загальній кількості

$$\sum_{i=1}^{N_0} \sum_{j=1}^{N_0} q_{ij} = N_0 \tag{2.5}$$

Ознакою того, скільки об’єктів включено до кластера буде значення суми (2.3). Змінними параметрами задачі оптимізації будуть елементи матриці Q.

Приклад

Для матриці відстаней, розрахованої у прикладі з розділу 1, побудувати кластери методом повного перебору.

Таблиця 2.1

	1	2	3	4	5
1	0	0,89	0,41	0,74	0,46
2	0,89	0	0,86	0,87	0,61
3	0,41	0,86	0	0,96	0,66
4	0,74	0,87	0,96	0	0,62
5	0,46	0,61	0,66	0,62	0

Перенесемо цю матрицю сюди та перетворимо її з трикутної на прямокутну, з урахуванням (1.12)

Перенесемо цю матрицю до електронних таблиць Calc. Там же утворимо чисту матрицю, теж розміром 5х5, в якій будуть знаходитися елементи матриці *Q*. При цьому ми припускаємо, що номери стовпців у ній відповідають номерам об'єктів, а номери рядків – номерам кластерів. У матриці *Q* знайдемо суму по рядкам, що буде відповідати лівій частині обмеження (2.3), та по стовпцях – лівій частині обмеження (2.4). Далі знайдемо загальну суму елементів матриці *Q* згідно з (2.5). Сума знаходиться за допомогою функції *SUM*(II:JJ), де II – адреса першої клітинки, яка містить масив чисел, а JJ – адреса останньої клітинки. Тепер сформуємо цільову функцію виду (2.2). Для цього скористаємося функцією *SUMPRODUCT*(II:JJ;NN:MM), де II:JJ – адреса масиву відстаней, а NN:MM – адреса масиву матриці *Q*. Спочатку всі суми дадуть нуль, оскільки всі елементи матриці *Q* дорівнюють нулю. Застосувавши функцію Solver електронних таблиць Calc, інтерфейс якої представлено на рис. 2.3-2.4, ми позначаємо елементи матриці *Q*, як змінні параметри, що мають бінарний характер, спрямовуємо функціонал до мінімуму, додаємо обмеження і отримуємо наступне оптимальне рішення, для якого функціонал дорівнює 4,36.

Таблиця 2.2

		Об'єкти					Сума по кластерам
		1	2	3	4	5	
Кластери	1	0	1	0	0	0	1
	2	1	0	0	0	0	1
	3	0	0	0	1	1	2
	4	0	0	1	0	0	1
	5	0	0	0	0	0	0
Сума по стовпцям		1	1	1	1	1	5

Аналіз результатів показує, що у перший кластер попав 2-й об'єкт, у другий кластер – 1-й, у третій кластер – 4-й та 5-й, у четвертий кластер – 3-й, у п'ятий – не включено жодного об'єкта.

На рис. 2.3-2.4 наведено порядок користування функцією Solver електронних таюлиць Calc.

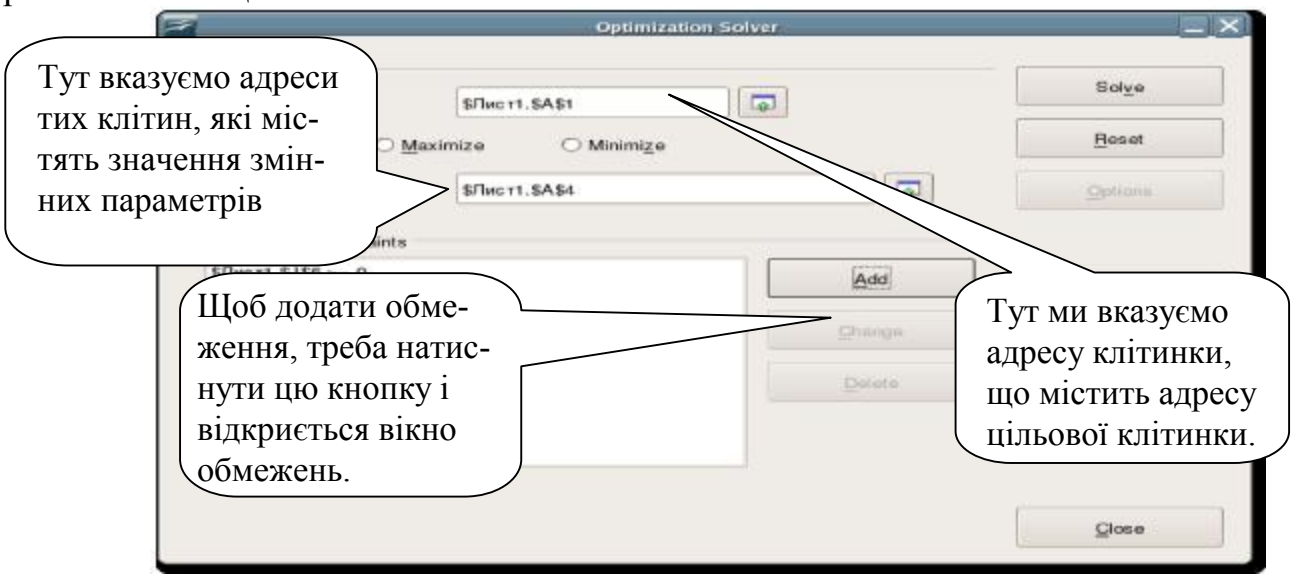


Рис. 2.3. Головне вікно функції Solver

Необхідно задати адресу цільової клітинки, вказати напрям оптимізації – minimize, ввести обмеження – кнопка «*Add*», вказати діапазон клітин, які містять значення змінних параметрів. Для розрахунку натиснути кнопку «*Solve*».

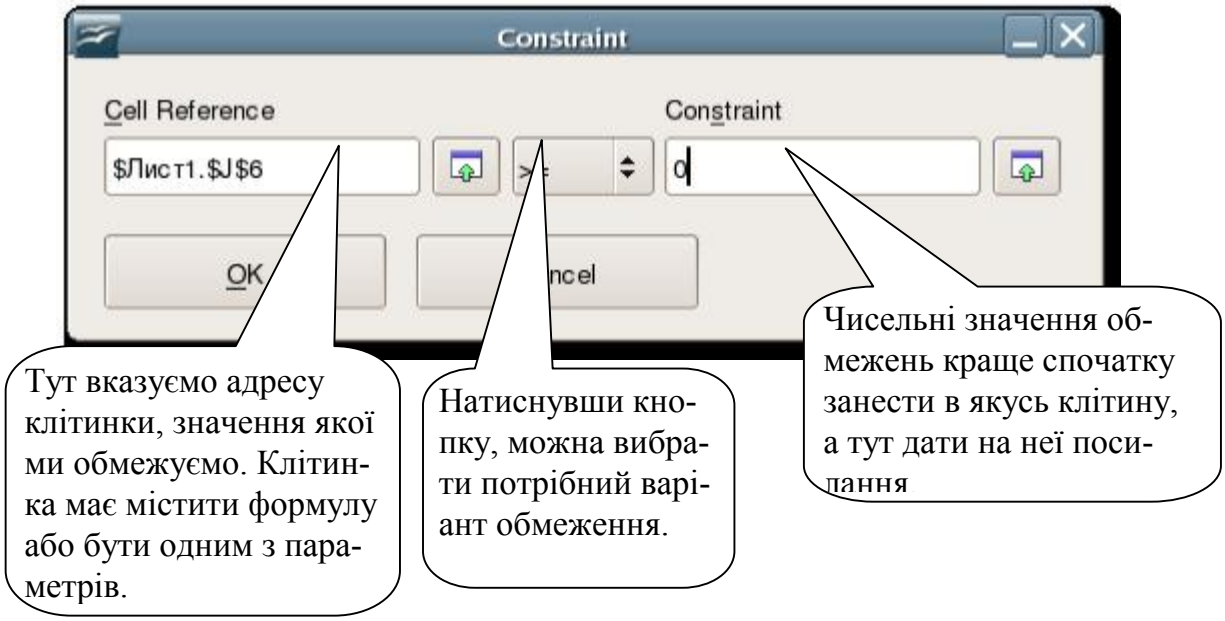


Рис. 2.4. Вікно введення обмежень функції Solver

Той же приклад вирішимо із застосуванням електронних таблиць Microsoft Excel, куди ми перенесемо цю матрицю. Там же утворимо чисту матрицю, теж розміром 5х5, в якій будуть знаходитися елементи матриці *Q*. При цьому ми припускаємо, що номери стовпців у ній відповідають номерам об'єктів, а номери рядків – номерам кластерів. У матриці *Q* знайдемо суму по рядкам, що буде відповідати лівій частині обмеження (2.3), та по стовпцях – лівій частині обмеження (2.4). Далі знайдемо загальну суму елементів матриці *Q* згідно з (2.5). Сума знаходиться за допомогою функції *СУММ*(II:JJ), де II – адреса першої клітинки, яка містить масив чисел, а JJ – адреса останньої клітинки. Тепер сформуємо цільову функцію виду (2.2). Для цього скористаємося функцією *СУММПРОИЗВ*(II:JJ;NN:MM), де II:JJ – адреса масиву відстаней, а NN:MM – адреса масиву матриці *Q*. Спочатку всі суми дадуть нуль, оскільки всі елементи матриці *Q* дорівнюють нулю. Застосувавши функцію *Solver* пункту меню «Сервис» електронних таблиць Excel, інтерфейс якої представлено на рис. 2.5 -2.6, ми позначаємо елементи матриці *Q*, як змінні параметри, що мають бінарний характер, спрямовуємо функціонал до мінімуму, додаємо обмеження і отримуємо наступне оптимальне рішення, для якого функціонал дорівнює 4,36.

Таблиця 2.3

		Об'єкти					Сума по кластерам
		1	2	3	4	5	
Кластери	1	0	1	0	0	0	1
	2	1	0	0	0	0	1
	3	0	0	0	1	1	2
	4	0	0	1	0	0	1
	5	0	0	0	0	0	0
Сума по стовпцям		1	1	1	1	1	5

Аналіз результатів показує, що у перший кластер попав 2-й об’єкт, у другий кластер – 1-й, у третій кластер – 4-й та 5-й, у четвертий кластер – 3-й, у п’ятий – не включено жодного об’єкта.

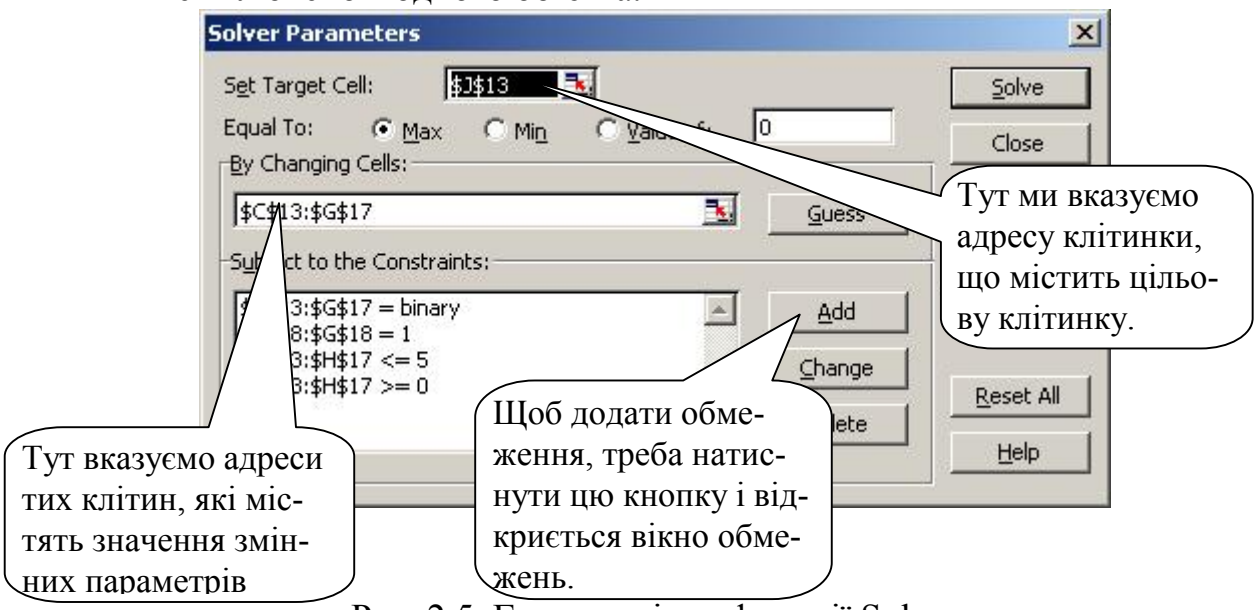


Рис. 2.5. Головне вікно функції Solver

В головному меню функції Solver необхідно задати: адресу цільової клітинки (Target Cell), вказати напрям оптимізації – Min, вказати діапазон клітин, які містять значення змінних параметрів, ввести обмеження – кнопка «Add» (значення змінних параметрів = binary – «0», якщо об’єкт X_i не належить до j-го кластеру; «1», якщо об’єкт X_i належить до j-го кластеру). Для розрахунку натиснути кнопку «Solve».

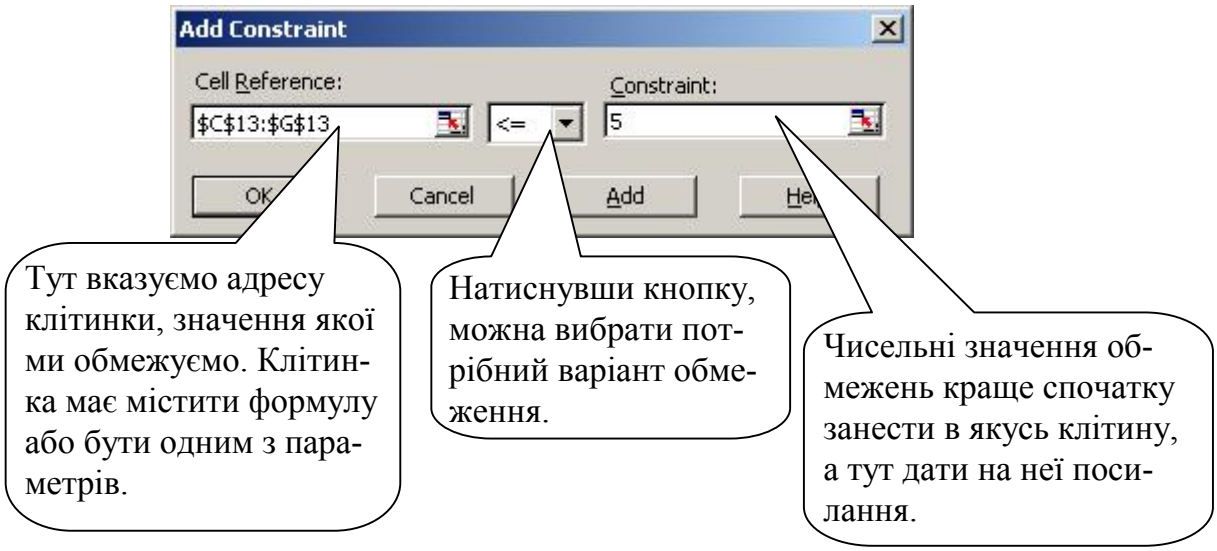


Рис. 2.6. Вікно уведення обмежень функції Solver

Результати одержані при кластеризації за допомогою електронних таблиць Microsoft Excel збігаються з результатами кластеризації в Open Office в додатку Calc.

2.2. Кластеризація методом перебору фіксованих відстаней від центрів сфер

Алгоритм такої кластеризації полягає у виконанні наступних кроків.

1. Випадковим-інтуїтивним способом вибирається точка (об'єкт класифікації) у деякому метричному просторі.

2. Обчислюються відстані від обраної точки до всіх інших об'єктів, потім ці відстані заносяться в матрицю в упорядкованому виді. Отримана матриця потрібна тільки для установлення фіксованого радіуса сфери, що є границею кластера.

3. Радіус сфери вибирається довільним чином. При виборі радіуса зручно дотримуватися принципу попадання у сферу визначеної кількості об'єктів, обчислених у попередній ітерації. Приміром, це може бути одна третина від загального числа об'єктів.

4. Маємо центр сфери $(x^1_{i,j,k})$ і r_1 - радіус метричного простору, постійний для всього циклу рішення задачі кластеризації.

5. Всі об'єкти, що потрапили в побудовану сферу, стають елементами кластера. Далі вибирається який-небудь новий елемент зі сфери, що стає центром $(x^2_{i,j,k})$ нової сфери того ж радіуса (r_1).

6. Для вибору координат нового центра $(x^{(n)}_{i,j,k})$ потрібно мати який-небудь формальний критерій. Одним з логічних критеріїв може бути мінімум відстані від обраного об'єкта до оболонки сфери (рис. 2.7).

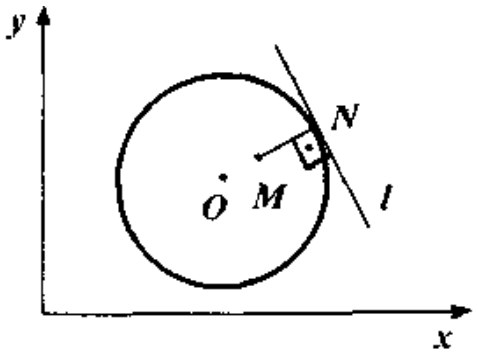


Рис. 2.7. Вибір нового центра при кластеризації у двовимірному просторі

7. Знову побудована сфера містить у собі як об'єкти з першої сфери $(x^{(1)}, r_1)$, так і нові. Старі об'єкти виключаються з розгляду, а потім процедура побудови сфер постійного радіуса повторюється доти поки в сферу не потрапить жоден новий об'єкт. Це відбудеться обов'язково, тому що або в кластер увійдуть усі розглянуті об'єкти, або відстань між якими-небудь об'єктами виявиться більше радіуса сфери.

8. Кластером у цьому алгоритмі будуть називатися всі об'єкти, що ввійшли хоча б в одну з побудованих сфер. Очевидно, незалежно від кількості влучень об'єкта в різні сфери, у кластері він враховується тільки один раз.

У тому випадку, коли кластери містять велику кількість об'єктів, досить проблематично підібрати радіус сфери таким чином, щоб у нього не ввійшло жодного об'єкта крім центра сфери. Щоб не затягувати процес формування кластера на тисячі ітерацій, раціонально спростити критерій закінчення кластеризації шляхом обмеження кількості об'єктів в останній сфері до якої-небудь межі (припустимо, 5 % від числа об'єктів у першій сфері).

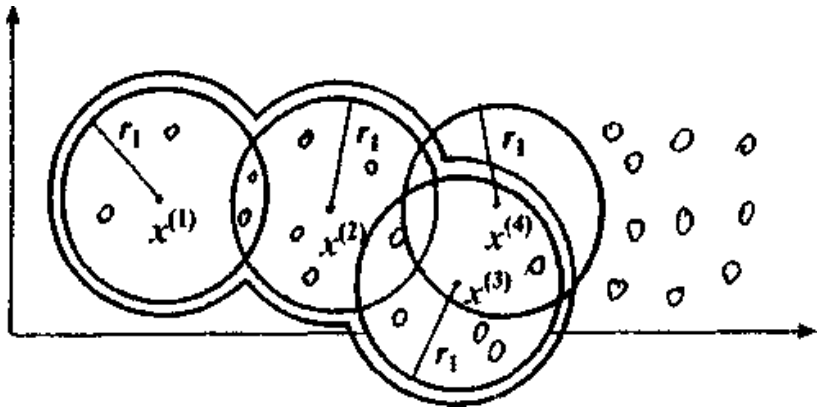


Рис. 2.7. Побудова кластерного метричного простору методом перебору фіксованих відстаней від центрів сфер

Алгоритм має істотні недоліки, оскільки модель передбачає закінчення процедури розбивки на групи об'єктів після першого знайдення порожнього кластера.

Приклад

Для прикладу з попереднього розділу провести кластеризацію методом перебору фіксованих відстаней від центра сфер. Ось наша матриця відстаней.

Таблиця 2.4

	1	2	3	4	5
1	0	0,89	0,41	0,74	0,46
2	0,89	0	0,86	0,87	0,61
3	0,41	0,86	0	0,96	0,66
4	0,74	0,87	0,96	0	0,62
5	0,46	0,61	0,66	0,62	0

Далі виконуємо пункти алгоритму:

1. Обираємо об’єкт 1, як центр кластеризації. Відстані до інших об’єктів знаходяться у першій колонці таблиці. Координати центру першого кластера беремо з нормованих числових значень його параметрів (факторів).

Таблиця 2.5

№ об’єкта	X_1	X_2	X_3	X_4
1	5,25	3,39	3,91	4,48

2. За радіус гіперсфери приймаємо значення r = 0,7. Тоді, до першого кластера увійдуть об’єкти 3 та 5.

3. Утворюємо центр нової гіперсфери за критерієм близькості до межі першої гіперсфери. Це визначиться за різницею відстаней об’єктів 3 та 5 з радіусом гіперсфери. Для третього об’єкта : 0,7 - 0,41 = 0,29. Для п’ятого об’єкта: 0,7 - 0,46 = 0,24. Отже п’ятий об’єкт утворить центр нової гіперсфери з координатами.

Таблиця 2.6

№ об’єкта	X_1	X_2	X_3	X_4
5	3,74	4,14	4,85	4,48

4. Оскільки старі об’єкти не можуть бути включені у нову гіперсферу, то для радіуса 0,7 згідно з матриці відстаней (відстань об’єкта 5 до інших представлена у п’ятій колонці матриці), можуть бути включені об’єкти 2 та 4.

5. Утворюємо центр нової гіперсфери за критерієм близькості до межі першої гіперсфери. Це визначиться за різницею відстаней об’єктів 2 та 3 з радіусом гіперсфери. Для другого об’єкта : 0,7 - 0,61 = 0,09. Для четвертого об’єкта: 0,7 - 0,62 = 0,08. Отже четвертий об’єкт утворить центр нової гіперсфери з координатами.

Таблиця 2.7

№ об’єкта	X_1	X_2	X_3	X_4
4	4,14	4,04	5,13	2,27

6. Оскільки старі об’єкти не можуть бути включені у нову гіперсферу, то для радіуса 0,7 згідно з матриці відстаней об’єктів більше не залишилося. Отже, розрахунок кластерів завершено. У перший кластер увійшли об’єкти 1 та 3, а у другий – 2, 4, 5. Центрами кластерів є відповідно координати об’єктів 1 та 5.

Таблиця 2.8

		Об'єкти					Сума по кластерам
		1	2	3	4	5	
Кластери	1	1	0	1	0	0	2
	2	0	1	0	1	1	3
Сума по стовпцям		1	1	1	1	1	5

Всі подальші методи теж базуються на подібних прийомах, які вимагають або ручної роботи або створення програм, які б реалізували ці обчислення.

2.3. Сферичний метод двоступінчастої кластеризації з виділенням ядра (згущення) об'єктів класифікації

Цей метод усуває деякі недоліки попереднього. У ньому теж використовується сферичний принцип побудови кластерів, але алгоритм послідовний і припускає мінімальне втручання дослідника в класифікацію на стадії обчислення й угрупування кластерів. Заздалегідь передбачається, що кластер буде мати сферичну форму, причому множина об'єктів у сфері (гіперсфері) розділяється на ядро (найбільше згущення) і менш щільну частину. Логічна модель кластеризації сферичним методом складається з наступних блоків.

1. Оцінюється інтервал, на якому розташовані можливі значення діамет-

ра сфери. Цей інтервал розділяється на ділянки і послідовно вибираються різні значення діаметра, найбільш удалий з яких фіксується при першій ітерації.

2. Зафіксувавши значення діаметра сфери, радіусом рівним чверті обраного раніше діаметра і центром у раніше обраній крапці описується допоміжна внутрішня сфера, площа якої складає не більш 25 % від площі великої сфери.

3. Всі об'єкти, що потрапили у внутрішню область малої сфери записуються в окрему матрицю і вважаються ядром кластера (згущенням).

4. Сам кластер є також сферою (гіперсферою), радіус якої дорівнює діаметрові сфери згущення (ядро кластера). Всі об'єкти, розташовані у внутрішній області великої сфери і є кластером.

5. У випадку ненасиченого кластера можливе об'єднання сфер за умови, якщо відстань між об'єктами не перевищує діаметра більшої сфери.

6. Отриману матрицю розбивки вихідної множини на кластери оцінюють зі змістовних позицій. У випадку незадовільних результатів мається можливість: а) зміни діаметра початкової сфери і перерахування всієї моделі кластеризації; б) зміни співвідношення діаметрів згущення й основної сфери; в) об'єднання декількох сфер у загальний кластер.

Сферичний метод кластеризації дозволяє строго окреслити границі між кластерами й однозначно привласнює кожному об'єктові приналежність до якої-небудь сфери. Але такі строгі границі залишають досить багато об'єктів (до 60 %) за межами класифікованих множин. Підвищення якості кластеризації вимагає значного зменшення діаметра сфер, що приводить до збільшення кластерів, аж до числа порівнянного з числом об'єктів. Але в цьому випадку кластеризація не спрощує, а ускладнює систему керування і втрачає практичний зміст.

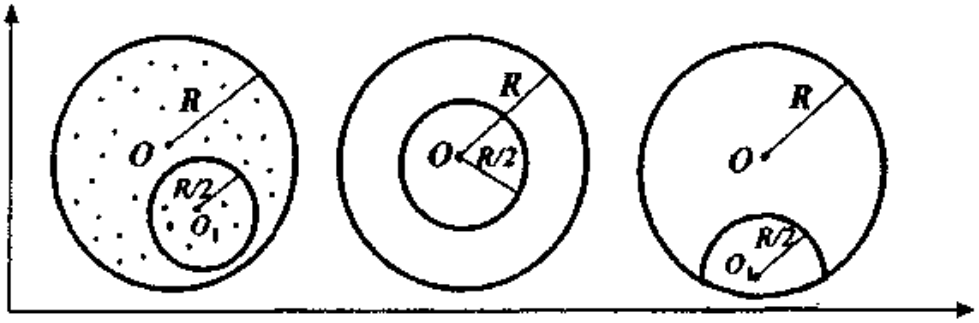


Рис. 2.8. Можливі варіанти розташування кластерів - сфер і згущень об'єктів у цих сферах

Тому метод сферичної кластеризації застосовується для такого розташування об'єктів, при якому існують щільні ядра об'єктів з малими відстанями між елементами і значні між групами, що дозволяє зневажити тими об'єктами, що неминуче виявляться поза сформованими кластерами.

Приклад

1.Аналізуючи матрицю відстаней між об’єктами, наведену нижче, знаходимо [0,7-0,9] інтервал на якому розташовані значення радіуса сфери. Тому діаметр майбутнього кластера покладемо рівним R=1,65 (середнє значення збі-

льшивши в 2 рази та деякий окіл).

№ об'єкту	**1**
2	0,89
3	0,41
4	0,74
5	0,46

2. Опишемо допоміжну внутрішню сферу $r=R/4$ (r=1,65/4=0,4125).

Обираємо об'єкт 1, як центр класифікації. Об'єкт 3 потрапив у внутрішню область малої сфери. Таким чином об'єкти 1 та 3 вважаються ядром кластера (згущенням).

Обчислимо координати центра гіперсфери $X_i = \frac{X_i^1 + X_i^3}{2}$, i=1,2,3,4:

	X_1	X_2	X_3	X_4
Координати центра 1-ї гіперсфери	4,8	3,14	3,32	4,61

3. Радіус гіперсфери *2r=0,825*. Обчислимо відстані від центра 1-ї гіперсфери до об'єктів 2,4,5.

№ об'єкту	Відстань від центру 1-ї гіперсфери
2	0,85
4	0,83
5	0,53

До 1-ї гіперсфери ввійде 5-й об'єкт (0,53<0,825).

Таким чином до першого кластеру належать об'єкти 1, 3 та 5. Вони виключаються з подальшого розгляду.

4. Для побудови наступного кластеру, обираємо об'єкт номер 2. До внутрішньої області малої сфери не попадає жоден з тих об'єктів, які залишилися (відстань менше 0,4125). До 2-ї гіперсфери (відстань менше 0,4125) також об'єкт номер 4 не потрапляє (відстань між 2 та 4 об'єктами дорівнює 0,87>0,825). До другого кластеру належить 2-й об'єкт.

5. В якості центра 3-ї гіперсфери обирається об'єкт номер 4. Він єдиний з тих об'єктів, які залишилися. 3-й кластер складатиметься з 4-го об'єкта.

6. Отриману матрицю розбивки вихідної множини на кластери оцінимо зі змістовних позицій. В кластери 2 та 3 входить лише по одному об'єкту. У випадку ненасиченого кластера можливе об'єднання сфер за умови, якщо відстань між об'єктами не перевищує діаметра більшої сфери. Відстань між 2-м та 4-м об'єктами (2-м та 3-м кластерами) складає 0,87, що не перевищує 1,65 - діаметра більшої сфери. Об'єднаємо 2-й та 3-й кластери.

7. В результаті отримаємо до 1-го кластеру належать: 1, 3 та 5-й об'єкти, до 2-го кластеру належать 2-й та 4-й об'єкти.

Таблиця 2.9

		Об'єкти					Сума по кластерам
		1	2	3	4	5	
Кластери	1	1	0	1	0	1	3
	2	0	1	0	1	0	2
Сума по стовпцям		1	1	1	1	1	5

2.4. Кластеризація інтегральним методом геометризації інформаційного поля

Метод застосовується в тих випадках, коли всі характеристики об'єкта можна природним чином розділити на пари складових інформаційних даних: точні і наближені, числові і логічні, об'єктивні і суб'єктивні, або які-небудь інші змістовно поділені пари характеристик.

Типовим прикладом можливого поділу характеристик є обробка результатів експертного опитування, зокрема, будь-якого анкетування респондентів. Експертів (респондентів) опитують, як правило, за типовою анкетою, у якій питання розділяються на дві групи: точні не допускають подвійного тлумачення (вік, стать, громадянство, дата народження) і не потребуючих точних відповідей (переваги, відношення до чого-небудь). Перший тип питань називається альтернативними, а другий – питаннями типу «меню», оскільки відповіді на такі питання можуть бути неоднозначними, тобто експерт може вибрати одночасно кілька відповідей-підказок або не вибрати жодного.

Розглянемо змістовну модель кластеризації інтегральним методом геометризації інформаційного поля:

1) формування інформаційного поля, поділ характеристик на дві інтегральні складові: об'єктивні і суб'єктивні, якісні і кількісні і т.п.;

2) вибір міри подібності двох об'єктів, основних характеристик інтегральних складових;

3) рішення задачі про вибір системи координат інформаційного поля й умовного об'єкта, прийнятого за початок координат $A_0(0; 0)$. Цю операцію можна виконати обчисленням середньоарифметичних координат об'єктів $A_y \to A_0$, або призначенням точки A_0 у якому-небудь інтервалі значимих показників інформаційного поля;

4) обчислення абсолютних і нормованих відстаней між об'єктом A_0 (центр координат) і всіма іншими об'єктами. Розрахунок інтегральних показників близькості об'єктів;

5) розподіл об'єктів по інтервалах інформаційного поля відповідно до інтегральних характеристик віддаленості від об'єкта A_0 по двох вимірах;

6) уведення третьої характеристики інформаційного поля - кількості об'єктів з подібними показниками, геометрично розташованих в околицях точок розбивки осей координат. Приміром, якщо розглядається т. $A_1(x_1;y_1)$, її околиця $(x_1 + \Delta x);(y_1 + \Delta y)$, а кількість об'єктів, що увійшли в цю околицю, дорівнює n, то цю ситуацію можна інтерпретувати як тривимірний розподіл по базисних осях об'єктивних і суб'єктивних даних (розподіл на площині), а число влучень об'єктів у деяку елементарну площу додає цій конструкції опуклість. Число n визначає «висоту» розглянутої конструкції і створює рельєфне зображення двовимірного інформаційного поля (див. рис. 2.9);

7) відповідно до рельєфу інформаційного поля визначаються границі

кластерів: «вершини» найбільші в деякій околиці значення клітинок, які є центрами згущень, «западини» – мінімальні значення клітинок поля, можуть бути інтерпретовані як дискримінантні лінії, що окреслюють замкнутий контур кластера (рис. 2.9);

8) у такий спосіб відбувається класифікація об'єктів за рівнем їхньої однорідності, виділення центрів згущення кластерів (ядер) у залежності від евклідової міри далекості об'єктів точки A_0 за двома інтегральними характеристиками одночасно. На цьому закінчується цикл моделі, у якому відбувається виділення системи однорідних підмножин (кластерів) графічними методами;

9) на заключному етапі відбувається ідентифікація геометричних об'єктів кластерів з відповідними реальними економічними, соціальними, демографічними прототипами (можливо, експертами або респондентами опитувань);

10) по співпадаючих характеристиках реальних об'єктів кластера складається його узагальнений портрет. Ці характеристики, що збігаються, виділяються в особливу групу і є ознаками, що описують типові й особливі риси кластерів. Ці ознаки необхідні для точного і прогнозованого керування об'єктами кластерного поля;

11) опис у змістовних образах неформальної структури розглянутих об'єктів, їхніх числових характеристик і типології, можливих управлінських впливів, відповідної реакції на керування й інші види прогнозів стану системи.

Кластеризація методами геометризації висуває визначені вимоги до структури інформаційного масиву. Дані повинні мати два види змістовної інтерпретації, щоб їх можна було перенести на інтегральні оцінки об'єктів.

Геометричні методи кластеризації зручні ще і тим, що дозволяють «згорнути» багатомірний простір ознак у двовимірний, причому варіювання включення вихідних ознак в інтегральні дає можливість виключення малозначимих характеристик і ранжирування тих, що залишилися. У такий спосіб дослідник може визначити рівень впливу кожної ознаки на поводження системи.

На рис. 2.9 подано кластерне інформаційне поле однієї тисячі об'єктів (заповнених анкет), що мають дві інтегральні характеристики. Затемнені ділянки є геометричними кластерами, інтенсивність забарвлення вказує на потужність кластеру.

2.5. Метод визначення центра кластера за допомогою обчислення середньо арифметичних відстаней між об'єктами

Розглянутий метод припускає наявність визначених фактів про зміст кластерів до початку обчислювальних процедур. Природно, апріорні припущення можуть бути досить наближеними. Щоб уникнути помилкових припущень дослідник може розглянути кілька варіантів початкового угруповання об'єктів. Цей метод кластеризації не припускає яких-небудь обмежень на геометричну форму кластера.

Пропонований алгоритм кластеризації складається з наступних блоків:

1. Деякій точці, що належить множині досліджуваних об'єктів, привлас-

нюється геометрична ознака центра координатної системи, причому перший об'єкт у цій системі є початком відліку точка $A_1(x_1;y_1;z_1)$ – для тривимірного простору кластерного поля.

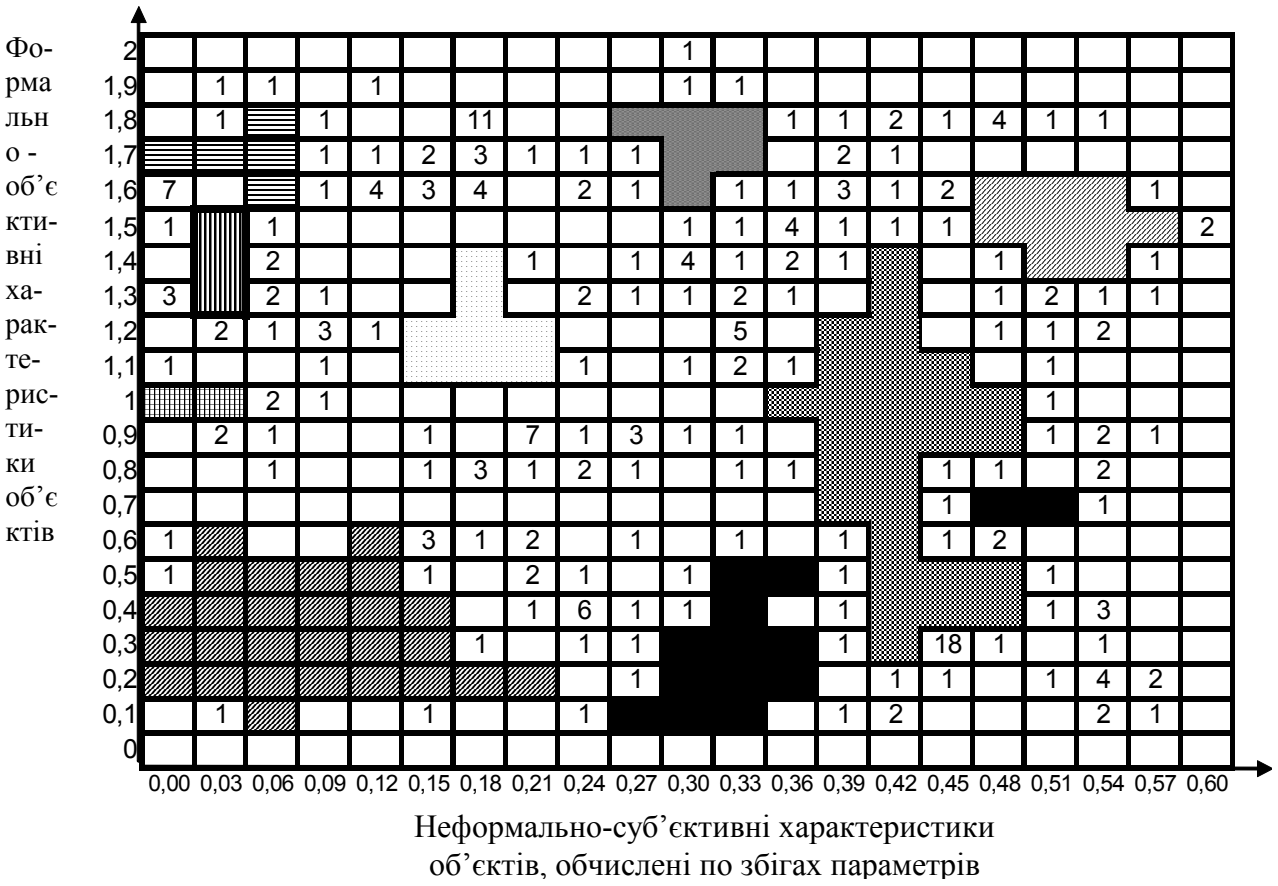


Рис. 2.9. Приклад застосування метода геометризації.

2. Вибирається визначене число об'єктів (тільки кількість) m, що будуть брати участь у розрахунках умовного центра кластера.

3. У випадку тривимірного простору розміщення об'єктів координати умовного центра кластера будуть рівні

$$A\text{ц} = \left(\frac{\mathrm{x}_1 + \sum_{i=1}^{m} x_i}{m+1}; \frac{\mathrm{y}_1 + \sum_{i=1}^{m} y_i}{m+1}; \frac{\mathrm{z}_1 + \sum_{i=1}^{m} z_i}{m+1} \right), \qquad (2.6)$$

де (x_i; y_i; z_i)–поточні координати об'єктів.

4. Далі необхідно задати який-небудь критерій, що обмежує об'єм кластера. Як приклад обмежень може бути використана гранична кількість об'єктів у кластері, або максимально припустима відстань від умовного центра до найбільш віддаленого об'єкта, або максимально можливий «вододіл» між найбільш близькими об'єктами. Дослідник, як правило, самостійно вирішує якому крите-

рієві віддати перевагу, або виробити власне обмеження на процес формування кластера.

5. У залежності від обраного алгоритму визначення критерію, необхідного для завершення формування кластера, обчислюється значення величини максимально припустимої відстані між об'єктами (d^*). Величина d^* може задаватися дослідником виходячи з аналізу змістовного образа кластера, або з розумінь насиченості кластера.

6. Методом перебору визначається об'єкт A_k найбільш близький до об'єкта $A_ц$. Далі перевіряється виконання нерівності $|A_ц - A_k| \le d^*$ і, у випадку безумовного виконання цієї нерівності, об'єкт A_k заноситься в матрицю даного кластера. З подальшого розгляду об'єкт $A_ц$ виключається.

7. Надалі операції перебору повторюються, причому об'єкт A_k стає центром розглянутої композиції. Ітерації попереднього блоку виконуються доти поки не залишиться жодного об'єкта поблизу кожної з розглянутих крапок, тобто буде перевищене граничне значення d^*.

Варіанти класифікації пропонованим методом визначаються перебором значень d^* і початкового об'єкта $A_ц$. У зв'язку з послідовним включенням об'єктів у кластери різниця між характеристиками початкових і кінцевих об'єктів може бути досить істотною. Це може послужити однією з причин неоднорідності кластерів.

Приклад

1. В якості центра координатної системи оберемо 1-й об'єкт. Аналізуючи матрицю відстаней між об'єктами з попереднього розділу, найбільш близькими до 1-го об'єкту є 3 та 5 – й об'єкти - вони будуть брати участь у розрахунках умовного центра кластера (m=2).

№ об'єкту	Відстань від $A_{ц1}$
1	0,1854
2	0,7399
3	0,3489
4	0,7205
5	0,3473

2. Обчислимо координати

$A_{ц1} = \{ (x_1^1 + x_1^3 + x_1^5)/3; \quad (x_2^1 + x_2^3 + x_2^5)/3; \quad (x_3^1 + x_3^3 + x_3^5)/3; \quad (x_4^1 + x_4^3 + x_4^5)/3; \}$

$A_{ц1}$={4,45; 3,47; 3,83; 4,57}.

Обчислимо відстань між $A_{ц1}$ та вихідною множиною об'єктів за метрикою Джффріса-Матусіти (1.8).

Аналізуючи матрицю відстаней, задамо d^*=0,5.

3. Найбільш близьким до $A_{ц1}$ є об'єкт номер 1. Нерівність $|A_ц - A_к| \le d^*$ виконується безумовно, 1-й об'єкт заноситься до 1-го кластеру. A_1 стає центром розглянутої композиції. Покладемо $A_{ц1} = A_1$.

№ об'єкту	Відстань від $A_{ц1}$
2	0,8912
3	0,4130
4	0,7410
5	0,4645

4. Обчислимо відстань між $A_{ц1}$ та об'єктами 2, 3, 4, 5, що залишились.

Найбільш близьким до $A_{ц1}$ є об'єкт номер 3.

Нерівність $|A_{ц1} - A_3| \le d^*$ виконується безумовно, 3-й об'єкт заноситься до 1-го кластеру. A_3 стає центром розглянутої композиції. Покладемо $A_{ц1} = A_3$.

5. Обчислимо відстань між $A_{ц1}$ та об'єктами 2, 4, 5, що залишились. Найбільш близьким до $A_{ц1}$ є об'єкт номер 5. Нерівність $|A_{ц1} - A_5| \le d^*$ не виконується. Формування 1-го кластеру завершено, до нього належать об'єкти 1 та 3.

№ об'єкту	Відстань від $A_{ц1}$
2	0,8634
4	0,9612
5	0,6648

6. В якості центра 2-го кластера оберемо 2-й об'єкт. Аналізуючи матрицю відстаней між об'єктами з попереднього розділу, найбільш близькими до 2-го об'єкту є 4 та 5 – й об'єкти - вони будуть брати участь у розрахунках умовного центра кластера (m=2).

7. Обчислимо координати

$$A_{ц2} = \{\ x_1^2 + (x_1^4 + x_1^5)/2;\quad x_2^2 + (x_2^4 + x_2^5)/2;\quad x_3^2 + (x_3^4 + x_3^5)/2;\quad x_4^2 + (x_4^4 + x_4^5)/2;\ \}$$

$A_{ц2}$={3,46; 4,57; 4,45; 3,59}.

Обчислимо відстань між $A_{ц2}$ та об'єктами 2, 4, 5.

№ об'єкту	Відстань від $A_{ц2}$
2	0,4564
4	0,4707
5	0,2718

Найбільш близьким до $A_{ц2}$ є об'єкт номер 5. Нерівність $|A_{ц2} - A_5| \le d^*$ виконується безумовно, тому 5-й об'єкт заноситься до 2-го кластеру. A_5 стає центром розглянутої композиції. Покладемо $A_{ц2} = A_5$.

8. Обчислимо відстань між $A_{ц2}$ та об'єктами 2, 4, що залишились.

№ об'єкту	Відстань від $A_{ц2}$
2	0,6073
4	0,6219

Нерівності $|A_{ц2} - A_2| \le d^*$ та $|A_{ц2} - A_4| \le d^*$ не виконуються. Формування 2-го кластеру завершено, до нього належить об'єкт номер 5.

9. В якості центра 3-го кластера оберемо 2-й об'єкт. Обчислимо координати $A_{ц3} = \{\ (x_1^2 + x_1^4)/2;\quad (x_2^2 + x_2^4)/2;\quad (x_3^2 + x_3^4)/2;\quad (x_4^2 + x_4^4)/2;\ \}$

10. Обчислимо відстань між $A_{ц3}$ та об'єктами 2, 4, що залишились.

№ об'єкту	Відстань від $A_{ц3}$
2	0,3108
4	0,6864

Найбільш близьким до $A_{ц3}$ є об'єкт номер 2. Нерівність $|A_{ц3} - A_2| \le d^*$ виконується безумовно, 2-й об'єкт заноситься до 3-го кластеру. A_2 стає центром розглянутої композиції. Покладемо $A_{ц3} = A_2$.

11. Обчислимо відстань між $A_{ц3}$ та об'єктом номер 4. Вона дорівнює 0,8667>d*. Об'єкт номер 4 до 3-го кластеру не належить. Він ввійде до 4-го кластеру.

Отримали таку класифікацію:

Таблиця 2.10

		Об'єкти					Сума по кластерам
		1	2	3	4	5	
Кластери	1	1	0	1	0	0	2
	2	0	0	0	0	1	1
	3	0	1	0	0	0	1

	4	0	0	0	1	0	1
Сума по стовпцям		1	1	1	1	1	5

2.6. Метод постійних кластерів і характеристик

Цей метод зручний у тих випадках, коли система, що піддається класифікації, добре вивчена й у дослідника існує визначена ясність щодо найбільш значимих характеристик кластерів. У цьому випадку дослідник може установити раціональні границі кількості кластерів і їхньої характеристики, не роблячи складних обчислень. Тоді розподіл об'єктів по кластерах відбувається в результаті простих арифметичних розрахунків, без циклічних повторень, в результаті однієї-двох ітерацій.

Алгоритм пропонованого методу заснований на змістовному аналізі інформаційного поля до початку процедури кластеризації. На цьому етапі дослідник повинний визначити раціональне число кластерів, виходячи з фізичних можливостей оцінки поля об'єктів, корисності для керування і можливості одержання нетривіальних результатів. Уведемо позначення: n – раціональна кількість кластерів; $A_ц(k, l, m)$ – центр «маси» l-го кластера, що має координати k, l, m.

Розглянемо алгоритм методу постійних кластерів і їхніх характеристик:

1. Вивчення змістовних характеристик інформаційного поля, аналіз даних і визначення кількості кластерів, їхніх основних характеристик – граничних умов кластерів.

2. Розподіл об'єктів по кластерах у залежності від включення координат об'єкта в граничні умови кластера.

3. Обчислення координат центрів «маси» кластерів $A_ц$ як середніх арифметичних координат об'єктів, що входять у розглянутий кластер.

4. Заміна апріорних характеристик границь кластерів на нові критерії кластеризації.

5. Приймаючи $A_{ц1}, A_{ц2}, \ldots, A_{цn}$ за центри «мас» кластерів (їхня кількість збереглася), починаємо формування кластерів заново. При цьому використовуємо принцип найбільшої близькості об'єкта до якого-небудь центра кластера.

6. Перерахувавши усі відстані від об'єктів до кожного з центрів і приєднавши їх до найближчого, одержимо нове поле кластерів, при якому всі об'єкти будуть належати якому-небудь кластерові з визначених заздалегідь.

7. Для знову сформованих кластерів виробляється перерахунок центрів «маси», обчислюються типові характеристики кластерів.

Розглянутий метод дозволяє включити всі об'єкти в кластери. Це серйозний недолік методу, тому що змістовний аналіз інформаційного поля класифікації свідчить про існування досить великої кількості об'єктів, що не можуть бути без перекручування властивостей включені до жодного з існуючих кластерів. Другим недоліком розглянутого методу є необхідність попереднього визначення кількості кластерів і їхніх типових характеристик. Клас подібних задач надзвичайно вузький і, швидше за все, може бути другим етапом класифікації. Тобто спочатку яким-небудь методом класифікація вже проведена, а пропонований метод використовується для підтвердження або спростування отриманих

результатів.

Приклад

Таблиця 2.11

№ об’єкта	Ознаки	
	X_1	X_2
1	0,7	9
2	1,2	15,5
3	2,5	24
4	0,5	5
5	3,58	7
6	2,45	18

Нехай маємо початкові дані наведені у наступній таблиці:

1. З змістовного аналізу даних та предметної області встановлено, що раціональне число кластерів дорівнює 4.
2. Аналізуючи чисельні значення ознак, встановимо такі граничні умови кластерів та приналежність об’єктів до кластерів:
3. Обчислимо координати центрів «мас» кластерів $A_ц$, як середніх арифметичних координат об'єктів, що входять у розглянутий кластер: $A_{ц1}$ ={0,6; 7}; $A_{ц2}$ ={3,58; 7}; $A_{ц3}$ ={1,2; 15,5}; $A_{ц4}$ ={2,48; 21}.

Таблиця 2.12

№ кластеру	Границі по X_1	Границі по X_2	№ об’єктів, що належать кластеру
1	0-2	0-10	1, 4
2	2-4	0-10	5
3	0-2	10-25	2
4	2-4	10-25	3, 6

4. Обчисливши центри кластерів, переформуємо кластери заново. При цьому використовуємо принцип найбільшої близькості об'єкта до якого-небудь центра кластера. Застосуємо міру відстані Джеффріса-Матусіти (1.8):

Таблиця 2.13

№ об’єкта	Відстань до 1 кластеру	Відстань до 2 кластеру	Відстань до 3 кластеру	Відстань до 4 кластеру
1	**0,36**	1,11	0,97	1,75
2	1,33	1,52	**0,00**	0,80
3	2,39	2,27	1,08	**0,32**
4	**0,42**	1,25	1,74	2,50
5	1,12	**0,00**	1,52	1,96
6	1,78	1,63	0,56	**0,34**

Аналізуючи одержану матрицю відстаней і приймаючи за порогове значення d=0,5 (якщо відстань від об'єкту до центру кластера менше ніж 0,5, то об'єкт включаємо у даний кластер), отримаємо:

Таблиця 2.14

		Об'єкти						Сума по кластерам
		1	2	3	4	5	6	
Кластери	**1**	1	0	0	1	0	0	2
	2	0	0	0	0	1	0	1
	3	0	1	0	0	0	0	1
	4	0	0	1	0	0	1	2

Сума по стовпцям	1	1	1	1	1	1	6

У випадку двовимірних ознак результат кластеризації можна представити графічно за допомогою можливостей пакету MathLab (рис. 2.10).

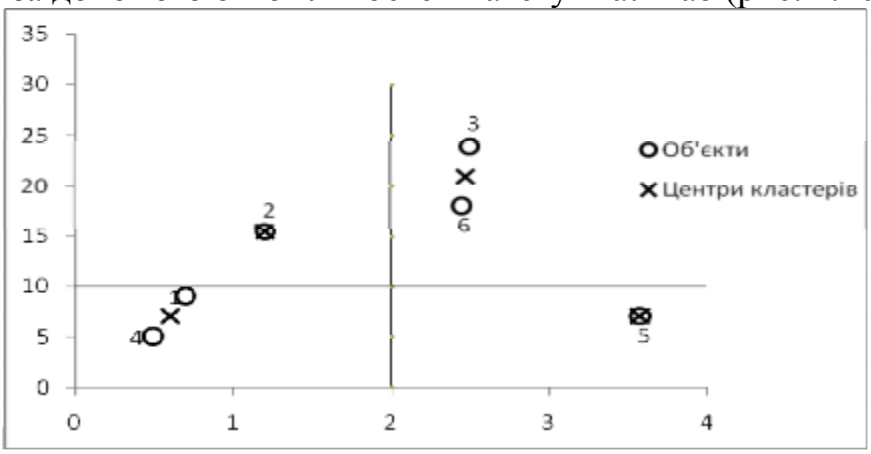


Рис 2.10 Результати кластеризації за методом постійних кластерів і характеристик

2.7. Кластеризація з урахуванням критерію якості і вибором кращого варіанта за цим критерієм

В алгоритмі передбачена процедура фільтрації при заданій цільовій функції добору кращого варіанта. Оптимальний набір кластерів має тверду, єдино можливу форму поля, причому контури окремих кластерів необов'язково повинні мати сферичну оболонку. Алгоритмом передбачається формулювання супермети, що композиційно включає в себе визначене число цільових функцій. У цьому випадку результатом класифікації буде єдиний, найкращий варіант розподілу об'єктів по кластерах.

Уведемо позначення: $M = \sum m$ – сумарна кількість об'єктів, розглянутих в умовах даної задачі класифікації; n - кількість кластерів, отриманих у результаті класифікації об'єктів $\alpha, \beta, \gamma, \psi; F$ – критерій оптимальності класифікації, причому F_{max} – суперціль.

Розглянемо послідовність процедур алгоритму:

1. Побудова інформаційного поля n об'єктів кластеризації, обчислення характеристик об'єктів і відстаней між ними.

2. Формування матриці мінімальних відстаней між об'єктами. Розмірність симетричної матриці (m-1).

3. Побудова контуру мінімальних відстаней між об'єктами і нанесення граничної дискримінантної лінії на кластерне поле. Перевірка замкнутості граничної лінії.

4. Поділ інформаційного масиву на можливі кластери, кількість яких знаходиться на відрізку [1; m - 1].

5. Порівняння якості кластеризації за заданими критеріями F. Формування інформаційних масивів кластерів і значень F.

6. Фільтрація варіантів кластеризації за локальними критеріями $F(z_1); F(z_2); F(z_3)....F(z_k)$, а також за комплексними критеріями $F(z_1, z_2)$ і т.д.

7. Визначення найкращого варіанта кластеризації за критерієм супермети F_{max}.

8. Опис характеристик об'єктів, що отримані в кластерах при найкращому варіанті F_{max}, або відповідних локальних критеріях оптимальної кластеризації.

У результаті обчислювальних процедур вихідні об'єкти розподіляються по кластерах досить простим способом, і відповідають сформульованим критеріям оптимальності розбивки. Але, незважаючи на простоту обчислювальних процедур, кількість варіантів кластеризації надзвичайно велика (C_{m-1}^{n}). Крім того, при розрахунку кожного варіанта приходиться розглядати множину значень відстаней між об'єктами усередині кластера для розрахунку критеріїв F. Тому дослідники звичайно вводять проміжні обмеження для скорочення числа розглянутих варіантів.

2.8. Ієрархічне угрупування

Ці процедури засновані на послідовному об'єднанні кластерів (агломеративні процедури) і на послідовній розбивці (дивизимні процедури). Найбільше поширення одержали агломеративні процедури. Розглянемо послідовність операцій у таких процедурах.

1. Всі об'єкти вважаються окремими кластерами.

2. Два найближчих кластери поєднуються в один кластер.

3. Кожне об'єднання зменшує число кластерів на один так, що зрештою всі об'єкти поєднуються в один кластер.

4. Найбільш підходящу розбивку вибирає найчастіше сам дослідник, за дендрограмою, що відображає результати групування об'єктів на всіх кроках алгоритму (рис. 2.7). Можуть одночасно також використовуватися і математичні критерії якості групування.

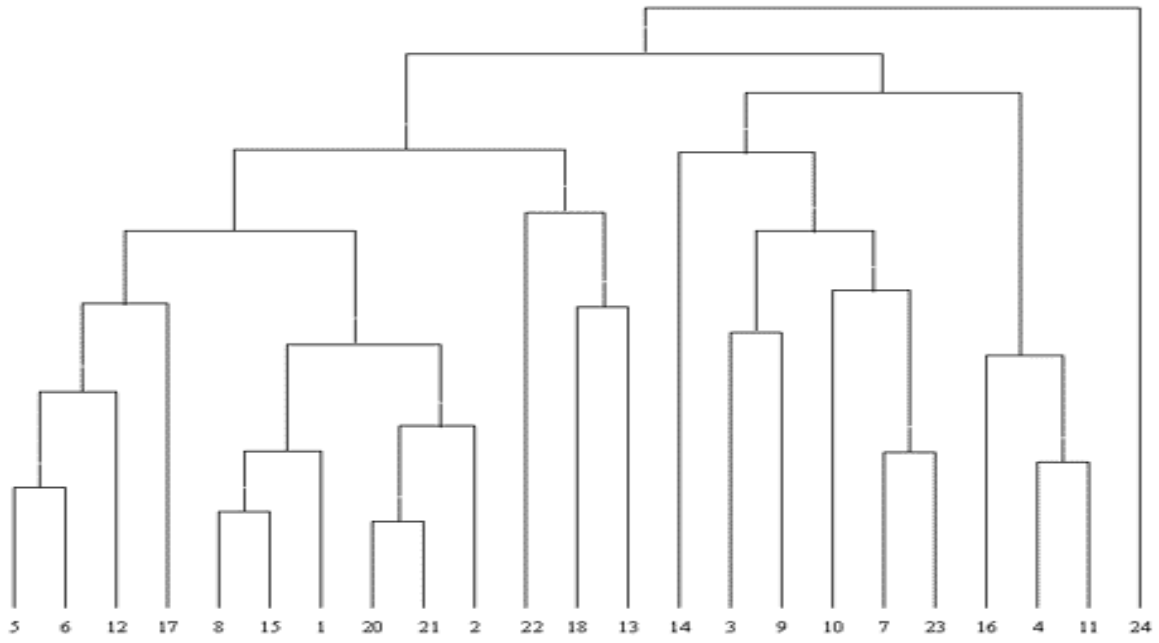


Рис. 2.11. Результати роботи ієрархічної агломеративної процедури групування 24 об'єктів, представлені у вигляді дендрограми

Приклад

Скориставшись матрицею відстаней з попереднього прикладу, побудувати дендрограму ієрархічного угрупування.

1. Найменша відстань (0,41) у об’єктів 1 та 3. Отже, першими поєднуємо їх.
2. Далі йде об’єкт 5 (0,46), який є найближчим до 1.
3. До нього ближче всього об’єкти 2 (0,61) та 4(0,62).

Класифікацію закінчено. Дендрограма має вигляд, представлений на рис. 2.11.

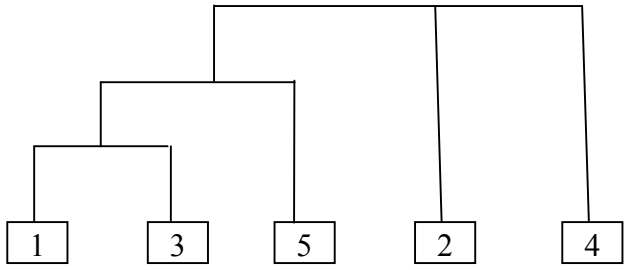


Рис. 2.12. Дендрограма з прикладу

На першому етапі об’єднуються об’єкти 1 та 3, на другому етапі приєднується 5 –й об’єкт, далі 2–й та 4–й. Ієрархічне угрупування дозволяє простежити етапи формування класів.

2.9. Алгоритм нечіткої кластеризації, методом *c*-середніх

Нехай досліджувана сукупність даних являє собою скінчену множину елементів $A=(a_1, a_2, \ldots, a_n)$, яка називається множиною кластеризації. Нехай також $P=(p_1, p_2, \ldots, p_q)$ – скінчена множина ознак або атрибутів, кожна з яких кількісно представляє деяку властивість або характеристику елементів пробле-

мної області. При цьому натуральне *n* визначає загальну кількість об`єктів даних, а натуральне *q* – загальну кількість ознак об`єктів.

Припустимо, що для кожного з об`єктів кластеризації деяким чином знайдені всі ознаки множини *P* у деякій кількісній шкалі. Таким чином, кожному елементу $a_i \in A$ поставлено у відповідність вектор $x_i = (x_1^i, x_2^i, \ldots, x_q^i)$, де $x_i$ – кількісне значення ознаки $p_j \in P$ для об`єкта $a_i \in A$. Для визначеності припустимо, що усі x_j^i приймають дійсні значення.

Нехай задана матриця вигляду

$$X = \begin{pmatrix} x_1^1 \ldots x_j^1 \ldots x_q^1 \\ \ldots\ldots\ldots\ldots\ldots \\ x_1^i \ldots x_j^i \ldots x_q^i \\ \ldots\ldots\ldots\ldots\ldots \\ x_1^n \ldots x_j^n \ldots x_q^n \end{pmatrix} \tag{2.7}$$

де x_i^j – значення *j*–го параметра для *i* – го об`єкта.

Нехай *c* – кількість кластерів, на яку потрібно розбити множину об`єктів. Нечіткі кластери опишемо наступною матрицею нечіткого розбиття:

$U = [\mu_{Ak}(a_i)], \quad \mu_{Ak}(a_i) \in [0,1], \quad k = 1...c, \; i = 1..n]$, в якій *k*– рядок містить ступені приналежності об`єкта $(x_1^k, x_2^k, \ldots, x_q^k)$ до кластерів $A_1, A_2, \ldots, A_c$.

Далі для кожного нечіткого кластера розглянемо центри v_k нечітких кластерів A_k, які розраховуються для кожного з нечітких кластерів для кожної з ознак за формулою

$$v_j^k = \frac{\sum_{i=1}^{n} (\mu_{Ak}(a_i))^m \cdot x_j^i}{\sum_{i=1}^{n} (\mu_{Ak}(a_i))^m}, \tag{2.8}$$

де *m* – параметр, що називається експоненціальною вагою і дорівнює деякому дійсному числу (*m*>1). Кожний з центрів кластерів являє собою вектор $v_i = (v_1^k, v_2^k, \ldots, v_q^k)$.

В якості цільової функції будемо розглядати суму квадратів зважених відхилень координат об`єктів кластеризації від центрів нечітких кластерів:

$$f(A_k, v_j^k) = \sum_{i=1}^{n} \sum_{k=1}^{c} (\mu_{Ak}(a_i))^m \sum_{j=1}^{q} (x_j^i - v_j^k)^2, \tag{2.9}$$

Де *m* – експоненціальна вага нечіткої кластеризації ($m \in R$, m>1), значення якої задається залежно від кількості елементів (потужності) множини *X*. Чим більшу кількість елементів містить множина *X*, тим менше значення обирається для *m*.

Розглянемо сам алгоритм нечіткої кластеризації. Він складається з наступних кроків:

1. Задати кількість нечітких кластерів *c*, максимальну кількість ітерацій алгоритму *s*, параметр збіжності ε, а також експоненціальну вагу *m*, (зазвичай *m*=2). В якості поточного нечіткого розбиття на першій ітерації для матриці

даних задати деяке поточне нечітке розбиття на *c* не порожніх нечітких кластерів, які описуються сукупністю функцій приналежності $\mu_{Ak}(a_i)$.

2. Для поточного нечіткого розбиття обчислити центри нечітких кластерів v_j^k та значення цільової функції $f(A_k, v_j^k)$.

3. Сформувати нове нечітке розбиття множини об`єктів кластеризації на с не порожніх нечітких кластерів, які характеризуються сукупністю функцій приналежності $\mu'_{Ak}(a_i)$, що обчислюються за формулою:

$$\mu'_{Ak}(a_i) = \left(\sum_{l=1}^{c} \left(\frac{\left(\sum_{j=1}^{q} (x_j^i - v_j^k)^2 \right)^{1/2}}{\left(\sum_{j=1}^{q} (x_j^i - v_j^l)^2 \right)^{1/2}} \right)^{\frac{2}{m-1}} \right)^{-1} \qquad (2.10)$$

4. При цьому якщо для деякого $k \in \{2..c\}$ і деякого $a_i \in A$ значення $\sum_{j=1}^{q} (x_j^i - v_j^k)^2 = 0$, то для відповідного нечіткого кластера покладемо $\mu'_{Ak}(a_i) = 1$, а для інших покладемо $\mu'_{Ak}(a_i) = 0$, $\forall l \in \{2...c\}$, $l \neq k$. Якщо ж таких $k \in \{2...c\}$, для деякого $a_i \in A$ буде декілька, тобто для них $\sum_{j=1}^{q} (x_j^i - v_j^k)^2 = 0$, то евристично для меншого з *k* покладемо $\mu'_{Ak}(a_i) = 1$, а для інших $l \in \{2...c\}$, $l \neq k$ покладемо $\mu'_{Ak}(a_i) = 0$.

5. Для нового нечіткого розбиття розрахувати центри нечітких кластерів і значення нечіткої функції.

6. Якщо кількість виконаних ітерацій більша за *s* або модуль різниці $\left| f(A_k, v_j^k) - f'(A_k, v_j^k) \right| \leq \varepsilon$, то в якості результату нечіткої кластеризації прийняти поточне нечітке розбиття і закінчити виконання алгоритму. У протилежному випадку вважати поточним знайдене нечітке розбиття і перейти на крок 3 алгоритму, збільшивши на 1 кількість виконаних ітерацій.

Слід відзначити, що дані отримані в результаті виконання алгоритму визначають, взагалі кажучи, локально – оптимальне нечітке розбиття.

Для різних апріорно заданих початкових значень матриці ступіней приналежності кінцевий розподіл характеризується тим, що порядкові номера кластерів можуть змінюватись, але значення ступіней приналежності в них зберігаються. Таким чином, нечітка кластеризація дозволяє отримати чіткі результати.

Приклад

Для реалізації алгоритму нечіткої кластеризації методом с-середніх використовується в середовищі MathLab 6.1 команда *fcm()*.

Функція *fcm()* визначає для кожної крапки вихідної множини ступень приналежності до кожного з кінцевих кластерів. Функція *fcm()* ітеративно просуває центри кластерів у «правильному» напрямку згідно вихідних даних. Ці

ітерації базуються на мінімізації цільової функції, що представляє собою відстань від вихідної множини точок до центрів кластерів зважену ступенями приналежності вихідних даних.

Синтаксис функції

fcm: [center, U, obj_fcn] = *fcm*(data, cluster_n)

де вхідні параметри:

data– множина даних для кластеризації, дані містяться в рядках.

cluster_n – кількість кластерів (більше 1).

результуючі дані:

center– матриця кінцевих центрів кластерів, де кожен рядок містить координати центрів

U– кінцева нечітка матриця розподілу (матриця ступенів приналежності)

obj_fcn– значення цільової функції на кожній ітерації

Функція *fcm(data, cluster_n, options)* може використовуватись з іншими параметрами, що впливають на кластеризацію:

options(1): показник ступеня приналежності для матриці розподілу U (по замовчуванню: 2.0)

options(2): мінімальна кількість ітерацій (по замовчуванню: 100)

options(3): мінімальна величина досягнення точності

Процес кластеризації зупиняється, коли досягнуто максимальну кількість ітерацій або значення цільової функції між двома послідовними ітераціями менше ніж задана величина точності.

Розглянемо роботу функції *fcm()* на прикладі дворозмірної випадкової величини. Дані нормовані в діапазоні [0,1].

load fcmdata.dat –завантажили дані з файлу *fcmdata.dat*

plot(fcmdata(:,1),fcmdata(:,2),'o'); – побудували графік вихідних даних

[center,U,objFcn] = fcm(fcmdata,2); – побудувати 2 кластери

Iteration count = 1, obj.fcn = 8.941176 – на першій ітерації значення цільової функції дорівнює 8.941176

Iteration count = 2, obj.fcn = 7.277177– на другій ітерації значення цільової функції дорівнює 7.277177.

Графічна інтерпретацію розрахунків покетом MathLab представлена на рис. 2.13 – 2.16.

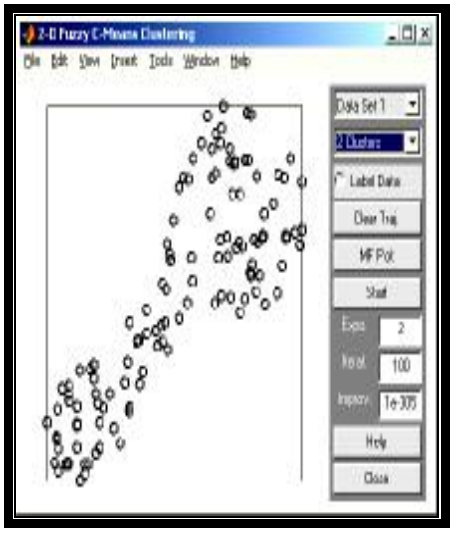


Рис. 2.13. Вихідні дані

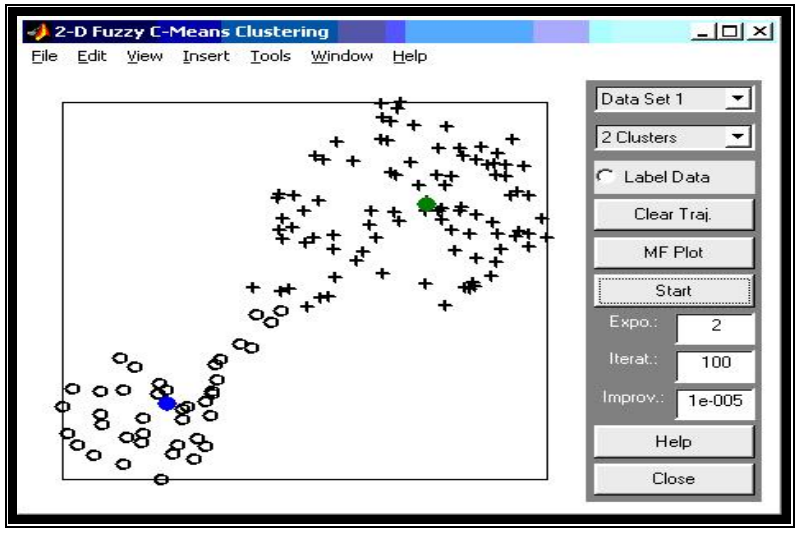


Рис. 2.14. Результат кластеризації, розбиття вихідної множини на 2 кластери ('+' –елементи 1–го кластера, 'o'– елементи 2– го кластера, '•'– центри кластерів)

2.10. Вибір найкращого розбиття

В результаті застосування різних алгоритмів та параметрів кластеризації з пунктів 2.1-2.9 одержуємо, в загальному випадку, розбиття, що відрізняються. Необхідно визначити, яке розбиття на класи найбільш адекватно описує вихідну множину даних. Природно спробувати визначити якість різних способів розбивки заданої сукупності об'єктів на класи, тобто визначити той кількісний

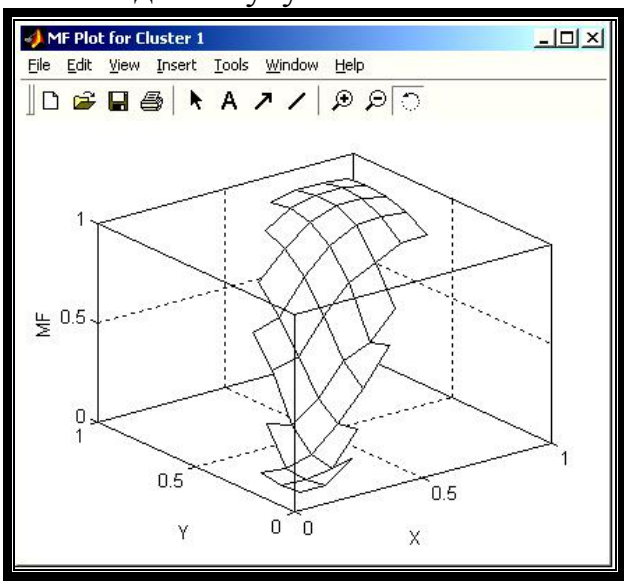


Рис. 2.15. Функція приналежності до 1-го кластера

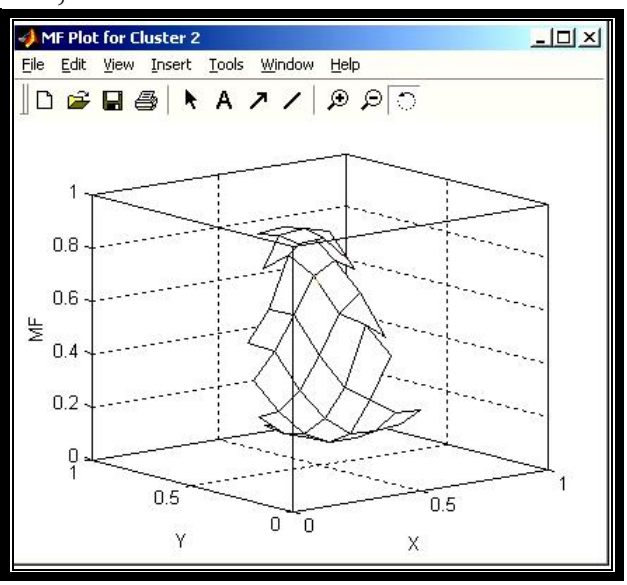


Рис. 2.16. Функція приналежності до 2-го кластера.

критерій, на підставі якого можна було б зволіти одну розбивку іншій. Для цієї мети використовуються *функціонали якості розбивки Q(S)*, визначені на множині всіх можливих розбивок, і найкращим вважається та розбивка, на якому досягаються екстремуми обраних функціоналів якості. Під критеріями якості

класифікації будемо розуміти деякі функціонали, що залежать від обсягів класів і відстаней між об'єктами, що ввійшли у виділені класи.

Нехай сукупність об'єктів $X_1, X_2, \ldots, X_n$ за допомогою метрики ρ розбита на k класів $S_1, S_2, \ldots, S_k$ – уведемо наступні величини, що характеризують ступінь розсіювання об'єктів з X:

• загальне розсіювання (розкид) –

$$Q(S) = \sum_{i=1}^{N} \rho^2(Xi, \overline{X}), \qquad (2.11)$$

де $\rho^2(Xi, \overline{X})$ - обрана метрика в просторі об'єктів, тобто правило обчислення відстані між будь-якою парою об'єктів досліджуваної множини $\{X_1, X_2, \ldots, X_N\}$, $\overline{X}$ - це загальний центр ваги сукупності;

• сума ("зважена") внутрішньокласових дисперсій (внутрішньокласовий розкид)

$$Q_1(S) = \sum_{l=1}^{k} \sum_{X_i \in S_l} \rho^2(X_i, \overline{X}(l)) \qquad (2.12)$$

де $\overline{X}_i(l)$ - центр ваги l–го класу.

• узагальнена внутрішньокласова дисперсія

$$Q3(S) = \det(\sum_{i=1}^{K} n_i, W_i) \qquad (2.13)$$

де під *det(A)* розуміють «визначник матриці *A*», а елементи $W_{qm}(l)$ вибіркової коваріаційної матриці W_l класу S_l, що підраховуються за формулою:

$$w_{qm}(l) = \frac{1}{n_l} \sum_{X_i \in S_l} (x_i^{(q)} - \overline{x}^{(q)}(l))(x_i^{(m)} - \overline{x}^{(m)}(l)), \quad \text{де } q,m=1,\ldots,p.$$

$x_i^{(q)}$ – q-та компонента багатомірного спостереження X_i, $\overline{x}^{(q)}(l)$ – середнє значення q-ї компоненти, підраховане за спостереженнями l-го класу. Чим менше Q_3, тим краще розбиття.

У випадку, якщо використовується евклідова чи "зважена " евклідова відстань загальне розсіювання являє собою суму між класового і внутрішньокласового розкиду.

Розглянемо наступний вираз: $$Q_2 = 1 - \frac{Q(S)}{Q1(S)} \qquad (2.14)$$

Чим більше величина Q_2, тим більша частка загального розкиду точок пояснюється між класовим розкидом, і, можна вважати з визначеною підставою, тим краще якість поділу.

Таким чином, краща розбивка буде вибиратися як розбивка, для якого виконується $$Q_2^* = max\ Q_2 \qquad (2.15)$$

Приклад

Розрахуємо якість розбиття отриманого в пункті 2.1 за функціоналом якості (2.14).

Таблиця 2.15

№ об'єкта	X_1	X_2	X_3	X_4

1	87	0,39	560	2770
2	25	0,82	430	2590
3	67	0,29	270	2870
4	62	0,52	860	1920
5	53	0,54	790	2770
$\overline{X}$	58,8	0,512	582	2584

Розрахуємо *Q(S)* загальне розсіювання (розкид) за (2.11) відстань за метрикою Джеффріса-Матусіти (1.8) від об'єктів досліджуваної множини $\{X_1, X_2,..., X_N\}$ до $\overline{X}$ загального центру ваги сукупності. Координати $\overline{X} = \{58,8;0,512;582;2584\}$ знайдемо як середнє відповідних координат всіх елементів сукупності. Для розрахунку *Q(S)* за метрикою Джеффріса-Матусіти (1.8) в середовищі OpenOffice Calc використовуються функції *(POWER(B$13;0,5)-POWER(B15;0,5))^2* – різницю квадратних коренів між відповідними координатами центра сукупності та кожного елементу сукупності піднести до другого степеню. В середовищі Microsoft Excel ця формула має вигляд *(КОРЕНЬ(C2)-КОРЕНЬ(C7))^2*. По ко

Таблиця 2.16

					Сума	Відстань від $\overline{X}$
X_1	2,75	0,008	0,212	3,23	6,205	2,49
X_2	7,12	0,036	11,48	0,004	18,64	4,32
X_3	0,27	0,031	59,18	7,50	66,99	8,18
X_4	0,04	0	27,05	49,21	76,31	8,74
X_5	0,15	0,004	15,86	3,23	19,24	4,39

жному рядку знайдемо суму:

За формулою (2.11) отримаємо *Q*=187,38.

В пункті 2.1 було отримано чотири кластери. До 1–го класу належить один 2–й елемент, до 2–го – 1–й елемент, до 4–го один 3–й елемент – внутрішньокласовий розкид (відстань між елементами класу та центром) дорівнює нулю. До 3–го класу належать два елементи 4–й та 5–й, тому розрахуємо координати центру кластеру $x_i^{\text{ц}} = \frac{x_i^4 + x_i^5}{2}$, $i = \overline{1,4}$, отримаємо (57,50; 0,53; 825,00; 2345,00). Далі за вище наведеною формулою розрахуємо відстань між елементами класу та центром кластеру та знайдемо за (2.12) *Q1*=39,83. З (2.14) $Q2 = 1 - \frac{Q(S)}{Q1(S)} = 1 - \frac{187,38}{39,83} = -3,7$.

Виконавши вище викладені розрахунки для розбиття отриманого в пункті 2.2, де перший кластер увійшли об’єкти 1 та 3, а у другий – 2, 4, 5. Отримаємо *Q*=187,38; *Q1*=99,5;

$$Q2 = 1 - \frac{Q(S)}{Q1(S)} = 1 - \frac{187,38}{99,5} = -0,88.$$

Для результатів розбиття множини на класи сферичним методом двоступінчастої кластеризації з виділенням ядра (згущення) об'єктів класифікації пункт 2.3 отримаємо *Q*=187,38;

$$Q1=139,35;\ Q2 = 1 - \frac{Q(S)}{Q1(S)} = 1 - \frac{187,38}{139,35} = -0,34.$$

Використавши в якості критерію кластеризації (2.14) можна зробити висновок, що найкраще розбиття отримано в пункті 2.1 за алгоритмом кластеризація повним перебором об'єктів.

2.11. Індивідуальне завдання №2. Кластеризація об’єктів

Мета завдання: Вивчити метод повного перебору при кластеризації групи об’єктів та метод перебору фіксованих відстаней від центра сфер.

Використовуючи результати розрахунків відстаней за різними метриками з індивідуального завдання №1, розрахувати включення об’єктів до кластерів. Застосувавши функцію Solve електронних таблиць Calc з пакету Open Office вільного програмного забезпечення, потрібно вирішити оптимальні задачі включення до кластерів для матриць відстаней, розрахованих за метриками Евкліда, Чєбишева, Хемінга, Джеффріса-Матусіти, степенної, „кварталів”, *L*-метрики. Результати представити в електронному вигляді, як елементи електронних таблиць Calc, придатні для подальших розрахунків.

Контрольні запитання

1. У чому кластерний аналіз відрізняється від інших методів угрупування?
2. Що таке центр кластера? Радіус кластера?
3. Які недоліки алгоритму повного перебору?
4. Назвіть найбільш прийнятний, на вашу думку, метод кластеризації. Обґрунтуйте ваш вибір.
5. Які з алгоритмів вимагають програмування на мовах високого рівня?

В розділі розглянуто порядок застосування метрик відстаней для проведення угрупування об’єктів у кластери сімома різними методами з визначенням числових характеристик кордонів кластерів.

3. ВІДНЕСЕННЯ НОВИХ ОБ'ЄКТІВ ДО ІСНУЮЧИХ КЛАСТЕРІВ

Вивчення матеріалу цього розділу дозволить студенту визначати статистичні характеристики кластерів та відносити нові об'єкти до раніше утворених кластерів.

3.1. Визначення оптимального числа кластерів

Одним з найважливіших питань при рішенні проблеми кластеризації є вибір необхідного числа кластерів. В деяких випадках число кластерів K може бути вибране апріорно при аналізі предметної області, проте в загальному випадку це число визначається в процесі розбиття множини на кластери.

Закони простого випадкового добору можуть бути застосовані для визначення числа кластерів, яке має бути прийняте для досягнення ймовірності α того, що знайдене найкраще розбиття. Таким чином, оптимальне число розбиття є функцією заданої частки β «найкращих» або в деякому сенсі допустимих розбиттів на множині всіх можливих. Загальне розсіювання множини кластерів буде тим більше, чим вища частка β «допустимих» розбиттів. У табл. 3.1 можна найти необхідне число розбиттів $S(\alpha,\beta)$ в залежності від значень α і β.

Таблиця 3.1

Визначення оптимального числа кластерів для заданого рівня достовірності

β \ α	0,20	0,10	0,05	0,01	0,001	0,0001
0,2	8	11	14	21	31	42
0,1	16	22	29	44	66	88
0,05	32	45	59	90	135	180
0,01	160	230	299	459	689	918
0,001	1626	2326	3026	4652	6977	9303
0,0001	17475	25000	32526	55000	75000	100000

3.2. Визначення статистичних характеристик кластерів

Коли кластери вже сформовані, виникає потреба у визначенні їх основних характеристик, якими є:

– середнє значення кожного фактора $$M_{kj} = \frac{1}{N_k}\sum_{i=1}^{Nk} X_{kji} \qquad (3.1)$$

– середнє квадратичне відхилення (стандарт) для кожного фактора

$$\sigma_{kj} = \frac{1}{N_k - 1} \sum_{i=1}^{Nk} X_{kji}^2 - M_{kj}^2 \qquad (3.2)$$

де N_k – кількість об'єктів, які увійшли у k-й кластер, j – номер фактора, для якого знаходиться ця характеристика ($1 \le j \le N_f$ – загальна кількість факторів, якими характеризуються об'єкти). Отже, якщо ми маємо 2 кластери, а кількість факторів становить 4, то потрібно розрахувати 8 статистичних характеристик.

Сукупність середніх дає нам координати центра кластера у гіперпросторі, а сукупність стандартів – міру розсіювання центра кластера.

Саме ці характеристики будуть потім слугувати як міра для прийняття чи відхилення нового об'єкта до певного кластера.

Важливо також перевірити статистичну достовірність відмінності середніх для кожного фактора для різних кластерів. Найзручніше скористатися *t*-тестом (або критерієм Стьюдента), який полягає у обчисленні ймовірності того, що значення середніх двох різних вибірок статистично є однаковими.

Щільність розподілу Стьюдента має вигляд

$$f(t) = S(t,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\cdot\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < t < \infty \qquad (3.3)$$

Другою важливою перевіркою є визначення статистичної достовірності того, що дисперсії кожного фактора для різних кластерів відносяться до однієї генеральної сукупності. Це досягається розрахунком *F*-тесту за критерієм Фішера.

Щільність розподілу Фішера має вигляд:

$$f_\xi(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & x \ge 0, \\ 0, & x < 0, \end{cases} \qquad \lambda > 0 \qquad (3.4)$$

Важливим є те, щоб ця ймовірність для кожного фактора була якомога меншою (<0,5). Це означатиме, що розділення класів було проведено правильно, бо їх середні статистично є відмінними, а дисперсії не належать до однієї генеральної сукупності. Якщо для якогось фактора будь-якої пари кластерів ці ймовірності виявляться значними (>0,6-0,7), то необхідно або виключити цей фактор з характеристик об'єктів або повторити кластеризацію за іншим алгоритмом, щоб досягти низької ймовірності для всіх пар кластерів.

Для автоматизації розрахунків доцільно використовувати наступні функції електронних таблиць Calc з пакету Open Office:

– для розрахунку середніх за (3.1) – *AVERAGEA(масив)*;

– для розрахунку стандартів за (3.2) – *STDEVA(масив)*;

– для розрахунків за критерієм Стьюдента (*t*-тест) –

TTECT(масив 1; масив 2; Режим; Тип)

масив 1 – перший масив даних.

масив 2 – другий масив даних.

Режим = 1, то функція використовує односторонній розподіл, якщо Режим = 2, то двосторонній.

Тип – Тип = 1 означає двосторонній. Тип =2 означає дві вибірки, рівні ймовірності. Тип = 3 означає дві вибірки, нерівні ймовірності;

– для розрахунків за критерієм Фішера (*F*-тест) – *FTECT(масив 1; масив 2)*, де „масив" – означає адреси клітинок, які містять чисельні значення фактора певного кластера. Якщо вказано 2 масиви, це означає, що для одного і того самого фактора потрібно вказати дані з одного і другого кластера.

Також можна скористатися функціями електронних таблиць Excel з пакету Microsoft Office:

– для розрахунку середніх за (3.1) – *СРЗНАЧ (масив)*;

– для розрахунку стандартів за (3.2) – *СТАНДОТКЛОН (масив)*;

– для розрахунків за критерієм Стьюдента (*t*-тест) –

TTECT(масив 1; масив 2; Режим; Тип)

масив 1 – перший масив даних, масив 2 – другий масив даних.

Режим = 1, то функція використовує односторонній розподіл, якщо Режим = 2, то двосторонній.

Тип – Тип = 1 означає двосторонній. Тип =2 означає дві вибірки, рівні ймовірності. Тип = 3 означає дві вибірки, нерівні ймовірності;

– для розрахунків за критерієм Фішера (*F*-тест) – *ФТЕСТ(масив 1; масив 2)*, де „масив" – означає адреси клітинок, які містять чисельні значення фактора певного кластера. Якщо вказано 2 масиви, це означає, що для одного і того самого фактора потрібно вказати дані з одного і другого кластера.

Якщо кластерів утворено більше двох, потрібно зробити статистичні розрахунки для всіх пар кластерів, щоб пересвідчитися, що кластери утворені вірно.

Приклад

Для кластерів з п. 2.2 провести статистичні розрахунки.

Результати розрахунків вміщено в табл. 3.1- 3.2. В першій таблиці були взяті оригінальні значення факторів, а у другій – їх нормовані величини.

Оскільки в п.2.2 було утворено тільки 2 кластери, то всі розрахунки проводилися для однієї пари кластерів, отже, для кожного фактора ми маємо одне число для *t*- тесту і одне число для *F*- тесту.

Аналіз отриманих результатів показує, що ймовірність тотожності для середніх (*t*–тест) значно міняється. І якщо для 2 та 4 факторів вона мала (0,22; 0,57), то факторів 1 та 3 близька до одиниці. Зате ймовірність віднесення дисперсії для однієї вибірки (*F*–тест) для факторів 1 та 3 дуже мала 0,08 - 0,16. Тобто, дисперсії не відносяться до однієї генеральної сукупності. Отже, потрі-

бен новий розрахунок кластерів для цих об'єктів. Звертає увагу те, що значення тестів не відрізняються для абсолютних і нормованих значень факторів. Але координати центрів кластерів та мір їх розсіювань звичайно мають різні значення для абсолютних і нормованих значень факторів.

Таблиця 3.2

Статистичні розрахунки для ненормованих значень факторів

№ об'єкта	X_1	X_2	X_3	X_4	Номер кластера
1	87	0,39	560	2770	**1**
3	67	0,29	270	2870	
2	25	0,82	430	2590	**2**
4	62	0,52	860	1920	
5	53	0,54	790	2770	
FTEST	0,08	0,06	0,13	0,16	
TTEST	0,92	0,57	1,0	0,22	
M_{kj}	77	0,34	415	2820	**1**
σ_{kj}	14,14	0,07	205,06	70,71	
M_{kj}	46,67	0,63	693,33	2426,67	**2**
σ_{kj}	19,3	0,17	230,72	447,92	

Таблиця 3.3

Статистичні розрахунки для нормованих значень факторів

№ об'єкта	X_1	X_2	X_3	X_4	Номер кластера
1	5,25	3,39	3,91	4,48	1
3	4,36	2,89	2,73	4,74	
2	2,51	5,54	3,38	4,02	2
4	4,14	4,04	5,13	2,27	
5	3,74	4,14	4,85	4,48	
FTEST	**0,08**	**0,06**	**0,13**	**0,16**	
TTEST	**0,93**	**0,57**	**1,0**	**0,22**	
M_{kj}	4,81	3,14	3,32	4,61	1
σ_{kj}	0,63	0,35	0,83	0,18	
M_{kj}	3,46	4,57	4,45	3,59	2
σ_{kj}	0,85	0,84	0,94	1,17	

3.3. Критерії віднесення нового об'єкта до існуючого кластера

Коли кластеризація об'єктів проведена і статистично визначено, що вона є достовірною, кластери можна вважати стабільними. Але в економіці часто з'являються нові об'єкти. Важливим є їх класифікація, тобто, віднесення до певного, раніше утвореного кластеру.

Така процедура виконується за наступних умов:

1. Новий об'єкт характеризується тими ж факторами, що і об'єкти в кластерах, утворених раніше.
2. Фактори нового об'єкту нормуються за значеннями середнього та стандарту, за якими нормувалися об'єкти раніше утворених кластерів.

Для кожного кластера знайдемо середньозважену довжину вектора його центра

$$M_k = \sqrt{\frac{1}{N_f}\sum_{j=1}^{Nf} M_{kj}^2}\ . \qquad (3.5)$$

та середньозважену міру розсіяння (стандарт) довжини цього вектора

$$\sigma_k = \sqrt{\frac{1}{N_f}\sum_{j=1}^{Nf} \sigma_{kj}^2}\ . \qquad (3.6)$$

За формулою (3.5) знайдемо також довжину вектора нового об'єкта $M_н$.

За значеннями відстаней розрахуємо тепер різниці

$$\Delta_k = M_k - M_н\ . \qquad (3.7)$$

Далі процедура виконується за наступним алгоритмом:

1. Якщо всі різниці мають один знак, то новий об'єкт відноситься до того кластера, для якого різниця є найменшою і процедура закінчується.
2. Якщо різниці мають різні знаки, то серед них обираються дві різного знаку, для яких значення по модулю є найменшими

$$|\Delta_k| \to \min\ , \qquad (3.8)$$

які назвемо Δ_+ та Δ_-.

3. До негативної різниці додамо, а від позитивної віднімемо значення середньозваженої міри розсіяння (стандарту) з утворенням двох нових різниць

$$\Delta\Delta_+ = \Delta_+ - \sigma_{k+}\ . \qquad (3.9)$$

$$\Delta\Delta_- = \Delta_- + \sigma_{k-}\ . \qquad (3.10)$$

4. Якщо $\Delta\Delta_+$ буде позитивним, а $\Delta\Delta_-$ – негативним, або вони поміняють знаки на протилежні, об'єкт відноситься до того кластера, для якого цей параметр по модулю є мінімальним

$$|\Delta\Delta_{\substack{+\\-}}| \to \min \qquad (3.11)$$

5. Якщо обидва $\Delta\Delta_+$ та $\Delta\Delta_-$ будуть позитивними, об'єкт відноситься до кластера, для якого різниця Δ_k була негативною.

6. Якщо обидва $\Delta\Delta_+$ та $\Delta\Delta_-$ будуть негативними, об'єкт відноситься до кластера, для якого різниця Δ_k була позитивною.

Приклад

За даними та розрахунками прикладу з п.3.2 визначити, до якого з двох кластерів можна віднести об'єкт, значення факторів якого наступні

X_1	X_2	X_3	X_4
71	0,26	590	1520

Рішення задачі починаємо з нормування факторів нового об'єкта параметрами, які ми знайшли у прикладі п. 1.3.

Таблиця 3.4

	X_1	X_2	X_3	X_4
M_{kj}	58,8	0,51	582	2584
σ_{kj}	22,63	0,2	245,7	384,68

Тепер, як і в п.1.3, скористаємося функцією *STANDARDIZE()* електронних таблиць Calc (функцією *НОРМАЛИЗАЦИЯ()* електронних таблиць Excel), щоб нормувати значення факторів нового об'єкта

X_1	X_2	X_3	X_4
4,17	3,03	4,01	3,3

Для кожного кластера знайдемо середньозважену довжину вектора його центра за (3.5) та середньозважену міру розсіяння (стандарт) довжини цього вектора за (3.6). Знайдемо також довжину вектора нового об'єкта.

Таблиця 3.5

Номер кластера	Чисельні значення	Позначення параметрів
1	4,04	M_{kj}
	0,56	σ_{kj}
2	4,05	M_{kj}
	0,96	σ_{kj}
Новий об'єкт	3,66	M_{kj}

Визначимо різниці Δ_k за (3.7). $\Delta_1 = 4{,}04 - 3{,}66 = 0{,}38$, $\Delta_2 = 4{,}05 - 3{,}66 = 0{,}39$. Отже, всі різниці позитивні. Тоді, згідно п.1 алгоритму віднесення об'єкта до кластера, об'єкт треба віднести до першого кластера, бо там різниця найменша.

3.4. Дискрімінантні функції для класифікації багатовимірних об'єктів

Хай заданий простір ознак X розмірністю $m>1$, точками якого є конкретні вимірювання $X=(X_1,X_2,...,X_m)$, де X_i – фактор чи параметр економічного об'єкта. Початкова таблиця спостережень розбита на p непересічних підмножин рядків, де кожному рядку X поставлений у відповідність деякий клас якості y_k, $k=1,2, ..., p$, причому будь-якому з p класів належить не менше одного об'єкту. Змістовний сенс системи класифікації $\{y_1, y_2,..., y_p\}$, що задається, стосовно економічних досліджень може мати цілком довільне тлумачення (наприклад, будь-які градації організаційно-правових форм власності, платоспроможності, класів якості продукції, потужностей виробництва).

Необхідно визначити набір формальних вирішальних правил, що дозволяють для довільного вимірювання X вказати клас y_k, до якого воно належить.

Завдання дискримінантного аналізу можна розділити на три типи. Завдання першого типу часто зустрічаються в медичній практиці. Допустимо, що ми маємо в своєму розпорядженні інформацію про деяке число індивідуумів, хвороба кожного з яких відноситься до одного з двох або більш діагнозів. На основі цієї інформації потрібно знайти функцію, що дозволяє поставити у відповідність новим індивідуумам характерні для них діагнози. Побудова такої функції і складає завдання дискримінації.

Другий тип завдання відноситься до ситуації, коли ознаки приналежності об'єкту до тієї або іншої групи втрачені, і їх потрібно відновити. Прикладом може служити визначення статі давно померлої людини по її останках, знайдених при археологічних розкопках.

Завдання третього типу пов'язані з прогнозом майбутніх подій на підставі наявних даних. Такі завдання виникають при прогнозі віддалених результатів лікування, наприклад, прогноз реабілітації прооперованих хворих.

Методи класифікації пов'язані з отриманням однієї або декількох функцій, що забезпечують можливість віднесення даного об'єкту до однієї з груп. Ці функції називаються класифікуючими і залежать від значень змінних таким чином, що з'являється можливість віднести кожен об'єкт до однієї з груп.

3.4.1. Дискримінація

Основною метою дискримінації є знаходження такої лінійної комбінації змінних (надалі ці змінні називатимемо змінними дискримінантів), яка б оптимально розділила дані групи. Лінійна функція

$$d_{km} = \beta_0 + \beta_1 x_{1km} + ... + \beta_p x_{pkm}, \quad m = 1,...,n, \quad k = 1,...,g \qquad (3.12)$$

називається *канонічною розділяючою функцією* з невідомими коефіцієнтами β_i. Тут d_{km} – значення розділяючої функції для m–го об'єкту в групі k; x_{ikm} значення дискримінантної змінної X_i для m–го об'єкту в групі k. З геометричної точки зору розділяючі функції визначають гіперповерхні в p–вимірному просторі. У окремому випадку при $p=2$ вони є прямими, а при $p=3$ площинами.

Коефіцієнти β_i першої канонічної розділяючої функції вибираються так, щоб центроїди різних груп якомога більше відрізнялися один від одного. Кое-

фіцієнти другої групи вибираються так само, але при цьому накладається додаткова умова, щоб значення другої функції були некорельовані із значеннями першої. Аналогічно визначаються і інші функції. Звідси витікає, що будь-яка канонічна розділяюча функція дискримінанта d має нульову внутрішньо групову кореляцію з $d_1, d_2,\ldots, d_{g-1}$. Якщо число груп рівне g, то число канонічних функцій дискримінантів буде на одиницю менше числа груп. Проте з багатьох причин практичного характеру корисно мати одну, дві або ж три розділяючі функції. Тоді графічне зображення об'єктів буде представлено в одно –, двох – і тривимірних просторах. Таке уявлення особливе корисно у разі, коли число змінних дискримінантів p велике в порівнянні з числом груп g.

3.4.2. Коефіцієнти канонічної функції дискримінанта

Для отримання коефіцієнтів β_i канонічної розділяючої функції потрібен статистичний критерій розрізнення груп. Очевидно, що класифікація змінних здійснюватиметься тим краще, чим менше розсіяння крапок відносно центроїда усередині групи і чим більша відстань між центроїдами груп. Зрозуміло, що велика внутрішньо групова варіація небажана, оскільки в цьому випадку будь-яка задана відстань між двома середніми тим менш значуща в статистичному сенсі, чим більша варіація розподілів, відповідних цим середнім. Один з методів пошуку якнайкращої дискримінації даних полягає в знаходженні такої канонічної розділяючої функції d, яка б максимізувала відношення між групової варіації до внутрішньо групової

$$\lambda = \frac{B(d)}{W(d)} \qquad (3.13)$$

де B – між групова, W – внутрішньо групова матриці розсіяння спостережуваних змінних від середніх.

Розглянемо максимізацію відношення (2.17) для довільного числа класів. Введемо наступні позначення: g–число класів; p – число змінних дискримінантів; n_k – число спостережень в k–й групі; n–загальне число спостережень по всіх групах; x_{ikm} – величина перемінної i для m–го спостереження в k–й групі; $\overline{x_{ik}}$ – середня величина змінної i в k–й групі; $\overline{F_i}$ – середнє значення змінної – i по всіх групах; $T(u,v)$ – загальна сума перехресних добутків для змінних – u та v; $W(u,v)$ – внутрішньо групова сума перехресних добутків для змінних – u та v.

У моделі дискримінація повинні дотримуватися наступні умови:

1. число груп: $g \geq 2$;
2. число об'єктів в кожній групі: $n_i \geq 2$;
3. число змінних дискримінантів: $0<p<(n-2)$;
4. змінні дискримінантів вимірюються в інтервальній шкалі;
5. змінні дискримінантів лінійно незалежні;
6. коваріаційні матриці груп приблизно рівні;
7. змінні дискримінантів в кожній групі підкоряються багатовимірному нормальному закону розподілу.

Розглянемо завдання максимізації відношення (3.13) коли є g груп. Оцінимо спочатку інформацію, що характеризує ступінь відмінності між об'єктами

по всьому простору точок, що визначаються змінними груп. Для цього обчислимо матрицю розсіяння *T*, яка рівна сумі квадратів відхилень і попарних добутків спостережень від загальних середніх $\overline{x_i},\ i = 1, ..., p$ по кожній змінній. Елементи матриці *T* визначаються виразом

$$t_{ij} = \sum_{k=1}^{g}\sum_{m=1}^{n}(x_{ikm} - \bar{x}_i)(x_{jkm} - \bar{x}_j) \qquad (3.14)$$

де $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{g} n_i \bar{x}_{ik},\ \ i = 1,..., p,\ \ \bar{x}_{ik} = \frac{1}{n_i}\sum_{m=1}^{n_k} \bar{x}_{ikm},\ \ i = 1,..., p;\ \ k = 1,..., g$

Запишемо цей вираз в матричній формі. Позначимо *p–вимірну* випадкову векторну змінну *k-ї* групи таким чином

$$X_k = \{x_{ikm}\},\quad i = 1,..., p,\quad k = 1,..., g,\quad m = 1,..., n_k .$$

Тоді об'єднана *p-вимірна* випадкова векторна змінна всіх груп матиме вигляд $X = [X_1 X_2 ...\ X_g]$.

Загальне середнє цієї *p-вимірної* випадкової векторної змінної буде рівне вектору середніх окремих ознак $\bar{x} = [\bar{x}_1 \bar{x}_2 ...\ \bar{x}_p]$.

Матриця розсіяння від середнього при цьому запишеться у вигляді

$$T = \sum_{k=1}^{g}(X_k - \bar{x})(X_k - \bar{x})' \qquad (3.15)$$

Матриця *T* містить повну інформацію про розподіл крапок по простору змінних. Діагональні елементи є сумою квадратів відхилень від загального середнього і показують як поводяться спостереження по окремо узятій змінній. Позадіагональні елементи рівні сумі добутків відхилень по одній змінній на відхилення по іншій.

Якщо розділити матрицю *T* на (*n*-1), то отримаємо коваріаційну матрицю. Для перевірки умови лінійної незалежності змінних корисно розглянути замість *T* нормовану кореляційну матрицю.

Для вимірювання ступеня розкиду об'єктів усередині груп розглянемо матрицю *W*, яка відрізняється від *T* тільки тим, що її елементи визначаються векторами середніх для окремих груп, а не вектором середніх для загальних даних. Елементи внутрішньо групового розсіяння визначаться виразом

$$w_{ij} = \sum_{k=1}^{g}\sum_{m=1}^{n_k}(x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk}) \qquad (3.16)$$

Якщо розділити кожен елемент матриці *W* (*n*–*g*), то отримаємо оцінку коваріаційної матриці внутрішньо групових даних.

Коли центроїди різних груп співпадають, то елементи матриць *T* і *W* будуть рівні. Якщо ж центроїди груп різні, то різниця *B=T–W* визначатиме між групову суму квадратів відхилень і попарних добутків. Якщо розташування груп в просторі розрізняється (тобто їх центроїди не співпадають), то ступінь розкиду спостережень усередині груп буде менше між групо-

вого розкиду. Відзначимо, що елементи матриці B можна обчислити і за даними середніх. $$b_{ij} = \sum_{k=1}^{g} n_k (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j),\ \ i,j = 1,...,p \qquad (3.17)$$

Матриці W і B містять всю основну інформацію про залежність усередині груп і між групами. Для кращого розділення спостережень на групи потрібно підібрати коефіцієнти розділяючої функції з умови *максимізації відношення між групової матриці розсіяння до внутрішньо групової матриці розсіяння за умови ортогональності площин дискримінантів*. Тоді знаходження коефіцієнтів функцій дискримінантів зводиться до рішення задачі про власні значення і вектори. При цьому функції дискримінантів можна отримувати: по *нестандартизованим* і *стандартизованим коефіцієнтам.*

3.4.3. Нестандартизовані коефіцієнти

Хай $\lambda_1 \geq ... \geq \lambda_p$ та $v_1,...,v_p$ відповідно власні значення і вектори. Тоді умова (2.17) в термінах власних чисел і векторів запишеться у вигляді

$$\lambda = \frac{\sum_k b_{jk} v_j v_k}{\sum_k w_{jk} v_j v_k} \qquad (3.18)$$

Звідки витікає $\sum_k (b_{jk} - \lambda w_{jk}) v_k = 0$, або в матричному записі

$$(B - \lambda_i W) v_i = 0, \quad v_i' W v_j = \delta_{ij} \qquad (3.19)$$

де δ_{ij} – символ Кронекера. Таким чином, розв'язання рівняння $|B - \lambda W| = 0$ дозволяє нам визначити компоненти власних векторів, відповідних функціям дискримінантів. Якщо B і W не вироджені матриці, то власні корені рівняння $|B - \lambda W| = 0$ такі ж, як і у $|W^{-1}B - \lambda I| = 0$.

Кожне рішення, яке має своє власне значення λ_i і власний вектор v_i, відповідає одній розділяючій функції. Компоненти власного вектора v_i можна використовувати як коефіцієнти функції дискримінанта. Проте при такому підході початок координат не співпадатиме з головним центроїдом. Для того, щоб початок координат співпав з головним центроїдом потрібно нормувати компоненти власного вектора $$\beta_i = v_i \sqrt{n - g},\ \ \beta_0 = -\sum_{i=1}^{p} \beta_i \bar{x}_i \qquad (3.20)$$

Нормовані коефіцієнти (2.24) отримані по нестандартизованим початковим даним, тому вони називаються ***нестандартизованими.*** Нормовані коефіцієнти приводять до таких значень дискримінантів, одиницею вимірювання яких є стандартне квадратичне відхилення. При такому підході кожна вісь в перетвореному просторі стискається або розтягується таким чином, що відповідне значення дискримінанта для даного об'єкту є числом стандартних відхилень крапки від головного центроїда.

Стандартизовані коефіцієнти можна отримати двома способами:

- за формулою (2.24), якщо початкові дані були приведені до стандартизованої форми;
- перетворенням нестандартизованих коефіцієнтів до стандартизованої форми

$$c_i = \beta_i \sqrt{\frac{w_{ii}}{n-g}} \qquad (3.21)$$

де w_{ii} – сума внутрішньо групових квадратів i змінної, визначуваної за формулою (2.20).

Стандартизовані коефіцієнти корисно застосовувати для зменшення розмірності початкового признакового простору змінних. Якщо абсолютна величина коефіцієнта для даної змінної для всіх функцій дискримінантів мала, то цю змінну можна виключити, тим самим скоротивши число змінних.

3.4.4. Число функцій дискримінантів

Загальне число функцій дискримінантів не перевищує числа дискримінантних змінних і, принаймні, на 1 менше числа груп. Ступінь розділення вибіркових груп залежить від величини власних чисел: чим більше власне число, тим сильніше розділення. Найбільшою розділовою здатністю володіє перша розділяюча функція, відповідна найбільшому власному числу λ_1, друга забезпечує максимальне розрізнення після першої і так далі. Розділяючу здатність *i–ої* функції оцінюють по відносній величині у відсотках власного числа λ_i від суми всіх λ.

Класифікуючі функції

До цих пір ми розглядали отримання канонічних розділяючих функцій при відомій приналежності об'єктів до того або іншого класу. Основна увага приділялася визначенню числа і значущості цих функцій, і використанню їх для пояснення відмінностей між класами. Проте найбільший інтерес представляє завдання прогнозу класу, якому належить деякий випадково вибраний об'єкт. Цю задачу можна вирішити, використовуючи інформацію, що міститься в змінних дискримінантів. Існують різні способи класифікації.

У процедурах класифікації можуть використовуватися як самі змінні дискримінантів, так і *канонічні функції дискримінантів*. У першому випадку застосовується *метод максимізації відмінностей між класами для отримання функції класифікації*, відмінність же класів на значущість не перевіряється і, отже, аналіз дискримінанта не проводиться. У другому випадку для класифікації використовуються безпосередньо *функції дискримінантів* і проводиться глибший аналіз.

Значення *класифікуючої функції* обчислюється за формулою:

$$d_{ik} = b_{k0} + b_{k1}x_{i1} + ... + \beta_{kp}x_{ki} + \ln q_k, \;\; k = 1,...,g \qquad (3.22)$$

Об'єкт $X_i=(x_{i1}...x_{ip})$ відноситься до класу, у якого значення d виявляється найбільшим. Коефіцієнти класифікуючих функцій зручніше обчислювати по скалярних виразах

$$b_{ki} = (n-g)\sum_{j=1}^{p}(w^{-1})_{ij}\bar{x}_{jk},\ \ k = 1,...,g \qquad (3.23)$$

де b_{ki} –коефіцієнт для змінної i у виразі, відповідному класу k, $(w^{-1})_{ij}$ – зворотний елемент внутрішньо групової матриці сум попарних добутків W. Постійний член знаходиться по формулі

$$b_{k0} = -0.5 \cdot \sum_{j=1}^{p} b_{kj}\bar{x}_{jk},\ \ k = 1,...,g \qquad (3.24)$$

Функції, що визначаються співвідношенням (2.26), називаються *простими класифікуючими функціями* тому, що вони припускають лише рівність групових ковариаційних матриць і не вимагають інших додаткових властивостей.

3.4.5. Класифікація об'єктів за допомогою функції відстані

Вибір функцій відстані між об'єктами для класифікації є найбільш очевидним способом введення міри схожості для векторів об'єктів, які інтерпретуються як крапки в евклідовому просторі. Як міру схожості можна використовувати евклидову відстань між об'єктами. Чим менше відстань між об'єктами, тим більше схожість. Проте в тих випадках, коли змінні корельовані, зміряні в різних одиницях і мають різні стандартні відхилення, важко чітко визначити поняття "відстані". В цьому випадку корисно застосувати не евклидову відстань, а *вибіркову відстань Махаланобіса*

$$D^2(x/G_k) = (n-g)\cdot\sum_{v=1}^{p}\sum_{j=1}^{p}(w^{-1})_{vj}(x_{iv}-\bar{x}_{vk})(x_{ij}-\bar{x}_{jk}),\ \ k = 1,...,g \qquad (3.25)$$

де x представляє об'єкт з p змінними, $\overline{X}_k$ – вектор середніх для змінних k–*ої* групи об'єктів.

При використанні функції відстані, об'єкт відносять до тієї групи, для якої відстань D^2 найменша.

Відзначимо, той факт, що апріорна вірогідність робить найбільший вплив при перекритті груп і, отже, багато об'єктів з великою вірогідністю можуть належати до багатьом групам. Якщо групи сильно розрізняються, то облік апріорної вірогідності практично не впливає на результат класифікації, оскільки між класами знаходитиметься дуже мало об'єктів.

3.4.6. Класифікаційна матриця

У дискримінантному аналізі процедура класифікації використовується для визначення приналежності до тієї або іншої групи випадково вибраних об'єктів, які не були включені при обчисленні дискримінанту і класифікуючих функцій. Для перевірки точності класифікації застосуємо класифікуючі функції до тих об'єктів, по яких вони були отримані. По частці правильно класифікованих об'єктів можна оцінити точність процедури класифікації. Результати такої класифікації представляють у вигляді *класифікаційної матриці*.

Розглянемо приклад класифікаційної матриці, приведеної в таблиці 3.4.

Таблиця 3.4

Класифікаційна матриця

Групи	**Передбачені групи(число/відсоток)**								
	1		2		3		4		Всього
1	**9**	**90.0**	0	0.0	0	0.0	1	0.0	10
2	0	0.0	**4**	**80.0**	1	20.0	0	0.0	5
3	8	14.8	4	7.4	**37**	**68.5**	5	9.3	54
4	1	7.7	0	0.0	1	7.7	**11**	**84.6**	13

У першій групі точно передбачені з 10 об'єктів 9, що складає 90%, один об'єкт віднесений до 4–й групи. У другій групі правильно передбачено 80% об'єктів, один об'єкт (20%) віднесений до третьої групи. У третій групі відсоток правильного прогнозу найнижчий і складає 68,5%, причому з 54 об'єктів до четвертої групи віднесено лише 37 об'єктів. У четвертій правильно передбачені 84,6%, по одному об'єкту віднесено до першої і третьої груп.

Відсоток правильної класифікації об'єктів є додатковою мірою відмінностей між групами і її можна вважати найбільш відповідною мірою дискримінації. Слід зазначити, що величина процентного змісту придатна для думки про правильний прогноз тільки тоді, коли розподіл об'єктів по групах проводився випадково. Наприклад, для двох груп при випадковій класифікації можна правильно передбачити 50%, а для чотирьох груп ця величина складає 25%. Тому якщо для двох груп маємо 60% правильного прогнозу, то потрібно вважати цю величину дуже малою, тоді як для чотирьох груп ця величина говорить про хорошу роздільчу здатність.

Приклад.

Підприємства у передбанкрутному стані, загальним числом 23, були розділені на три групи:

- *Група 1*. Санація підприємтсва виявилася успішною; проведене через деякий проміжок часу обстеження показало, що підприємство знаходиться у задовільному стані.
- *Група 2*. Санація безуспішна, тобто стан підприємство неодмінно збанкутіє.
- *Група 3*. Результат санації успішний, але надалі можливі повторне набуття передбанкрутного стану.

За наслідками обстеження 23 підприємтв було розраховано наступні коефіццєнти оцінки перебанкрутного стану, значення яких подано у відносних величинах:

- Y_6 – Інтегральний коефіцієнт рівня загрози банкрутства Альтмана;
- Y_9 Коефіцієнт Бівера;
- Y_{10} – Модель Спрінгейта

Конкретні значення цих коефіцієнтів приведені в таблиці 3.5

Таблиця 3.5

Дані про 23 підприємства у передбанкрутному стані

N	Гр.	Y_6	Y_9	Y_{10}
1	1	14.4	25.1	0.20
2	1	20.1	40.1	0.11
3	1	24.1	32.1	0.17
4	1	11.1	16.9	0.12
5	1	16.3	32.1	0.36
6	1	40.5	64.4	0.21
7	1	52.7	50.0	0.53
8	1	20.8	22.3	0.13
9	1	14.0	3.1	0.18
10	1	27.0	41.7	0.19
11	1	44.3	63.8	0.22
12	1	47.5	50.1	0.29
13	1	54.0	57.0	0.19
14	1	16.1	20.6	0.22
15	1	57.5	74.5	0.49
16	1	37.8	63.0	0.32
17	2	55.8	48.0	2.74
18	2	75.0	60.0	1.37
19	2	72.0	65.0	0.70
20	2	70.6	45.0	1.40
21	3	24.1	45.0	0.22
22	3	33.2	55.0	0.01
23	3	30.4	44.6	0.09

За матрицею початкових даних знаходяться средні стандартні відхилення змінних (таблиці 3.6 і 3.7) дискримінантів, загальна ***T*** і внутрішньогрупові ***W*** матриці сум квадратів і перехресних добутків).

Таблиця 3.6

Середні дискримінантних зміних

Групи	Y_6	Y_9	Y_{10}	К-ть
1 ($\bar{x}_{i1}$)	31,1375	41,0500	0,2456	16
2 ($\bar{x}_{i2}$)	68,3500	54,5000	1,5525	4
3 ($\bar{x}_{i3}$)	29,2333	48,2000	0,1067	3
Всі групи ($\bar{x}_i$)	37,3609	44,3217	0,4548	23

Таблиця 3.7

Стандартні відхилення

Групи	Y_6	Y_9	Y_{10}	К-ть
1	16,2739	20,4760	0,1237	16
2	8,5656	9,5394	0,8551	4
3	4,б608	5,8924	0,1060	3
Всі групи				23

Таблиця 3.8

Матриця загальної суми перехресних добутків ***T***

Змінна	Y_6	Y_9	Y_{10}
Y_6	8895,3148	6025,1896	163,2293
Y_9	6025,1896	7262,2391	53,5466
Y_{10}	163,2293	53,5466	8,3290

Таблиця 3.9

Матриця внутрішньогрупової суми перехресних добутків W

Змінна	Y_6	Y_9	Y_{10}
Y_6	4236,1542	4532,3100	–2,1545
Y_9	4532,3100	6631,4600	1,9565
Y_{10}	–2,1545	1,9565	2,4455

Якщо розділити кожен елемент *T* (*n*-1), а кожен елемент *W* на (*n*-*g*), то отримаємо ковариационные матриці. Для оцінки міри зв'язку між дискримінантими змінними матриці *T* і *W* перетворені в кореляційні матриці, які приведені в таблицях 3.10 і 3.11. Елементи цих матриць знайдені за формулами

$$r_{ij}^{(t)} = \frac{T_{ij}}{(n-1)S_i S_j} \text{ і } r_{ij}^{(w)} = \frac{W_{ij}}{(n-g)S_i S_j}$$

Із загальної кореляційної матриці (табл. 3.10) видно, що змінні некорельовані на рівні 0,01. Звідси витікає, що жодна змінна не може бути передбачена за значенням, відповідному іншій змінній.

Таблиця 3.10

Загальна кореляційна матриця

Змінна	Y_6	Y_9	Y_{10}
Y_6	1,0000	-0,1759	0,0664
Y_9	-0,1759	1,0000	0,3480
Y_{10}	0,0664	0,3480	1,0000

Для вимірювання міри розкиду спостережень усередині класів використовується внутрішньогрупова кореляційна матриця, яка приведена в таблиці 2.8. Ця матриця не співпадає із загальною кореляційною матрицею. З таблиці видно, що багато коефіцієнтів відрізняються від значень, приведених в таблиці 2.7.

Таблиця 3.11

Внутрішньогрупова кореляційна матриця

Змінна	Y_6	Y_9	Y_{10}
Y_6	1,0000	0,8551	–0,0212
Y_9	0,8551	1,0000	0,0154
Y_{10}	–0,0212	0,0154	1,00

З таблиць 3.8 і 3.9 видно, що велика частина елементів матриці W менше відповідних елементів матриці T. Різниця цих матриць $B=T–W$ визначає міжгрупову суму квадратів відхилень і попарних добутків. Ця матриця приведена в табл. 3.12.

Таблиця 3.12

Матриця міжгрупової суми перехресних добутків ***B***

Змінна	Y_6	Y_9	Y_{10}
Y_6	4659,1606	1492,8796	165,3838
Y_9	1492,8796	630,7791	51,5901
Y_{10}	165,3838	51,5901	5,8834

Для знаходження коефіцієнтів канонічної функції дискримінанта вирішуємо задачу (3.13) в термінах власних чисел і векторів, яка в матричному записі має вигляд (3.19). Систему рівнянь (3.19) вирішуємо за допомогою розкладання матриці Холецького $W^{-1} = LL'$, $(L'BL - \lambda_i I)v_i = 0$, ,$v_i'v_i = \delta_{ij}$.

Найбільше власне значення для системи рівне λ_1=5,3514 і λ_3=0,0452, яким відповідають власні вектори v_1=[0,7360 0,0990 0,6697]` і $v_3$=[–0,4368 0,8252 0,3581]`. Поклавши $b=Lv$, отримуємо коефіцієнти канонічної функції дискримінанта b_1=[0,0219 –0,0137 0,4585]`, $b_3$=[–0,0130 0,0190 0,2024]`.

При використанні коефіцієнтів b початок координат не співпадатиме з головним центроїдом. Для того, щоб початок координат співпав з головним центроїдом потрібно нормувати компоненти вектора b, використовуючи формули (3.20). Для оцінки відносного внеску кожної змінної в значення функції дискримінанта обчислимо стандартизовані коефіцієнти дискримінантів по формулі (3.21). Результати обчислень приведені в таблиці 3.13 і таблиці 3.14. З таблиці 3.14 видно, що дві найбільш значущо корельовані змінні Y_6 і Y_9 мають приблизно однакові стандартизовані коефіцієнти. Значення нестандартизованої канонічної функції для кожного підприємства зведені в табл. 3.19. Координати центроїдів першої, другої і третьої груп відповідно становлять:

$$\begin{bmatrix} -0{,}8363 & 4{,}6553 & -1{,}7466 \\ -0{,}1063 & 0{,}0604 & 0{,}4862 \end{bmatrix}$$

Таблиця 3.13

Нестандартізовані дискримінантні коефіцієнти

Змінна	**Коефіцієнти**	
Y_6	0,0978	-0,0580
Y_9	-0,0614	0,0850
Y_{10}	2,0504	0,9050
Константа	-1,8628	-0,20112

Таблиця 3.14

Стандартизовані дискримінантні коефіцієнти

Змінна	**Коефіцієнти**	
Y_6	1,4228	-0,8445
Y_9	-1,1184	1,5479
Y_{10}	0,7170	0,33165
Власні значення	5,3514	0,0452

Для визначення взаємної залежності окремої змінної і функції дискримінанта розглянемо внутрішньогрупові структурні коефіцієнти. Результати обчислень представлені в таблиці 3.15.

Таблиця 3.15

Внутрішньогрупові структурні коефіцієнти

Змінна	Коефіцієнт	
Y_6	1,4580	–0,8653
Y_9	–1,1460	1,5861
Y_{10}	0,7347	0,3243

Змінні Y_6 і Y_9 мають невеликі структурні коефіцієнти, але у них відносно великі стандартизовані коефіцієнти. Це пояснюється значущою кореляцією змінної Y_6 з іншими змінними і може опинитися, що внесок змінних Y_6 і Y_9 в дискримінантні значення невеликий. Для оцінки реальної корисності канонічної функції дискримінанта обчислюємо:

1. Коефіцієнт канонічної кореляції для і-*ї* функції: $r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$.Чим більша величина r_i , тим краща роздільна здатність дискримінантної функції.

1. Λ– статистику Уїлкса: $\Lambda = \prod_{i=k+1}^{g} \frac{1}{1+\lambda_i}$.

2. Статистику хи – квадрат (χ^2): $\chi^2 = -[n-(p+g)/2-1]\ln\Lambda_k$, $k = 0,1,...,g-1$ з $(p–k)(g–k–1)$ ступенями свободи.

3. Рівень значущості.

Обчислення проводимо в наступному порядку:

1. Знаходимо значення критерію χ^2 при *k=0.* Значущість критерію підтверджує існування відмінностей між групами. Крім того, це доводить, що перша функція дискримінанта значуща і має сенс її обчислювати.

2. Визначаємо першу функцію дискримінанта і перевіряємо значущість критерію при *k=1.* Якщо критерій значущий, то обчислюємо другу функцію дискримінанта і продовжуємо процес до тих пір, поки не буде вичерпана вся значуща формація. Результати обчислень приведені в табл. 3.16.

Таблиця 3.16

Основні статистики

Дискримінантна функція	Власне значення	Канонічна кореляція R	Λ– статистика Уїлкса	Статистика *χ2*	Ступені свободи	Рівень значущості
1	5,3514	0,9179	0,1506	35,9655	6	$4{,}076*10^{-6}$
2	0,0452	0,2080	0,9567	0,8405	2	0,6569

Дані таблиці указують на хорошу дискримінацію груп: велика величина канонічної кореляції відповідає тісному зв'язку функції дискримінанта з групами; мала величина *Λ*– статистика Уїлкса означає, що чотири використані змінні ефективно беруть участь в розрізненні груп і, нарешті, статистика χ–квадрат значуща з рівнем $1{,}6*10^{-8}$.

Процедура класифікації. Процедури класифікації можуть використовувати канонічні функції дискримінантів або самі змінні дискримінантів. Для

класифікації за допомогою змінних дискримінантів коефіцієнти класифікуючої функції обчислюємо за формулою (3.24). Результати обчислень приведені в таблиці 3.17. Значення класифікуючої функції для кожного хворого обчислені за формулою (3.22), результати класифікації у вигляді класифікаційної матриці представлені в таблиці 3.18. Оскільки відсоток правильної класифікації складає 100%, то таблицю класифікуючих функцій для окремих пацієнтів можна не представляти. Результати класифікації за допомогою відстані Махаланобіса (формула (3.25)) приведені в таблиці 3.19.

Таблиця 3.17

Коефіцієнти класифікуючих функцій

Змінна	**Група 1**	**Група 2**	**Група 3**
Y_6	0,0603	0,5875	–0,0631
Y_9	0,0820	–2,4110	0,1883
Y_{10}	1,9962	13,4071	0,6661
Константа	–2,8760	–23,9141	–3,6512

Таблиця 2.18

Класифікаційна матриця

Групи	**Передбачені групи(число/відсоток)**						
	1		2		3		Всього
1	**10**	**62.50**	0	0.0	6	37.50	16
2	0	0.0	**4**	**100.0**	0	0.0	4
3	0	0.0	0	0.0	**3**	**100.0**	3

Таблиця 2.19

Зведення результатів классификації

№ хворого	Нестандартізовані канонічні функції d_i			Квадрат відстані Махалано-біса $D^2(x/G_k)$		
	Група	Значення		Група 1	Група 2	Група 3
1	1	–1,6258	–0,5453	1,3941	39,9613	1,7126
2	1	–2,1879	0,3389	2,1281	46,4330	0,4254
3	1	–1,1576	–0,5402	0,3037	33,8515	1,4480
4	1	–1,6083	–1,1376	2,1155	40,6888	3,1499
5	1	–1,5398	0,0998	1,6444	39,0807	1,3698
6	1	–1,4635	1,3352	2,4410	38,6575	0,8729
7	1	–1,3373	–0,3477	5,3223	12,0657	10,6765
8	1	–1,2347	–0,9555	1,2544	32,8613	3,5611
9	1	–2,4564	–0,3223	5,7100	30,9378	10,5528
10	1	0,1421	–1,4293	0,4101	36,6478	0,2827
11	1	1,0663	–1,0241	1,6739	33,2676	1,1976
12	1	–0,2524	0,3058	0,1102	19,8784	5,5216
13	1	–0,1306	0,3126	3,2852	20,941	6,5678
14	1	–1,0198	–1,1302	1,2853	34,6955	3,0330
15	1	1,4639	0,1921	4,0840	22,5124	5,3097
16	1	1,4759	–1,4148	2,6895	38,3378	1,0454

№ хворого	Нестандартізовані канонічні функції d_i			Квадрат відстані Махаланобіса $D^2(x/G_k)$		
	Група	Значення		Група 1	Група 2	Група 3
17	2	1,3432	6,4170	60,6784	12,4824	73,1019
18	2	–0,0236	4,7068	29,9684	0,4904	40,9341
19	2	–0,0311	2,6839	14,5114	6,785	21,8918
20	2	–1,0408	5,2731	36,9560	1,7390	50,1042
21	3	0,6296	–1,8645	1,7390	42,4824	0,2744
22	3	0,7651	–2,0234	2,1344	44,5377	0,2310
23	3	0,0998	–1,4813	0,4413	37,2501	0,2704

З результуючої таблиці 2.16 видно розподілення хворих до 2–*ї* та 3–*ї* груп відбувається з 100% ймовірністю як за допомогою канонічних розділяючих функцій (значення d_i має бути максимальним) та за допомогою методу квадрату відстані Махаланобіса (значення d_i має бути мінімальним).

3.5. Індивідуальне завдання №3. Розрахунок статистичних характеристик кластерів та віднесення до них нових об'єктів

Мета завдання: Вивчити методи віднесення нових об'єктів до існуючих кластерів.

Використовуючи результати утворення кластерів за різними методами з індивідуального завдання № 2, розрахувати включення нових об'єктів до створених раніше кластерів. Застосувавши електронні таблиці Calc з пакету Open Office вільного програмного забезпечення, провести такі розрахунки для об'єктів, числові значення яких подані у наступному варіанті. Для студента з останнім номером варіанта треба взяти перший варіант.

Результати представити в електронному вигляді, як елементи електронних таблиць Calc, придатні для подальших розрахунків. Зробити висновки.

Контрольні запитання

1. Які принципи визначення оптимальної кількості кластерів?
2. Як знайти середньозважену довжину вектора центру кластера?
3. Що таке міра розсіювання координат центру кластера?
4. Для чого слугує *t*-тест?
5. Коли може бути припинено процедуру віднесення нового об'єкта до існуючих кластерів?

В розділі розглянуто порядок застосування метрик відстаней для проведення угрупування об'єктів у кластери сімома різними методами з визначенням числових характеристик кордонів кластерів.

4. ПРИКЛАДИ ЗАСТОСУВАННЯ КЛАСТЕРНОГО АНАЛІЗУ В ЕКОНОМІЦІ

Після ознайомлення з матеріалами розділу студенти отримають приклади застосування кластерного аналізу в економічних задачах.

В усіх наведених прикладах при характеристиці кластерів, подано середні значення факторів. При створенні кластерів використовувалася метрика Евкліда, а кількість кластерів призначалася дослідниками наперед.

4.1. Оцінка ступеня відмінності регіонів в діяльності туристсько-готельного господарства

Застосування кластерного аналізу дозволило оцінити ступінь відмінності російських регіонів в діяльності туристсько-готельного господарства Росії. Для цього були вибрані дані за 2004 рік.

Всі дані перед початком кластерного аналізу були нормалізовані для усунення впливу одиниць вимірювання показників. При розрахунках були використані дані по 79 регіонам Росії. В результаті реалізації процедури кластерного аналізу початкова сукупність з 79 регіонів розбита на 3 основні кластери, і виділені 4 аномальні регіони: м. Москва, Московська область, м. Санкт-Петербург і Краснодарський край. Результати кластерного аналізу показують значну диференціацію регіонів Росії по рівню розвитку туризму (табл. 4.1).

У Росії є 4 високо розвинуті туристські регіони: Москва, Московська область, Санкт-Петербург, Краснодарський край. Але навіть серед цих аномальних регіонів значно виділяється м. Москва.

До першого кластера відносяться 18 з 79 досліджуваних регіонів. Ці регіони характеризуються найвищим рівнем виробництва послуг або вигідним географічним розташуванням, в місцях мають значне туристське значення.

Регіони 1-го кластера разом з аномальними регіонами складають основу туризму в Росії. У цих регіонах надано майже 87% загального об'єму Росії туристсько-екскурсійних послуг, 77,2% санаторно-оздоровчих послуг і 87,1% всіх послуг готелів Росії.

Другий кластер є найчисленнішим по кількості регіонів Росії, що увійшли до нього (їх 39). У регіонах цього кластера рівень розвитку туризму значно менший, ніж в регіонах першого кластера. Середні значення показників регіонів цього кластера дозволяють зробити висновок про те, що ці регіони займа-

ють проміжне положення між якнайменше розвиненими (3 кластер) і найрозвиненішими (1 кластер) регіонами у сфері туризму.

Таблиця 4.1.

Середні значення показників кожного кластера

Показники	1 кластер	2 кластер	3 кластер	Аномальні регіони	Всі регіони (за виключенням аномальних)	Всі регіони
Об'єм окремих видів платних послуг в 2004 р., млн. крб., зокрема:						
Туристсько-екскурсійні	177,6	42,6	9,1	2243,6	67,0	177,2
Санаторно-оздоровчі	458,1	103,5	52,7	2168,7	176,4	277,3
Послуги готелів	217,8	57,0	27,5	3624,5	88,5	267,6
Чисельність розміщених осіб, 2004 р.:						
у підприємствах готельного типу (тис. чол.);	305,3	122,6	40,5	1521,0	146,7	216,3
у спеціалізованих засобах розміщення (тис. чол.)	176,7	55,8	15,1	754,3	75,0	109,4
Число регіонів	18	39	14	4	75	79

Третій кластер містить 14 регіонів, що характеризуються найнижчими показниками розвитку туризму в Росії. У цих регіонах вироблено тільки 5,3% послуг, що свідчить про їх економічну відсталість, і, як наслідок, про наявність низького рівня життя в цих регіонах. Середнє значення вартості туристських послуг третього кластера приблизно в 10 разів нижче в порівнянні з показниками 1-го кластера, що свідчить про практичну відсутність туристської діяльності в цих регіонах.

4.2. Чинники, що впливають на прибуток і рентабельність крупних сільськогосподарських підприємств

Використовувалися дані Омської області Росії. Нормування даних не проводилося, використовувалася метрика Евкліда. У табл. 4.2 - 4.3 показано об’єднання у кластери групи сільськогосподарських об’єктів, для яких обрано різні фактори, що характеризують їх діяльність. Поєднання неекономічних та економічних показників дозволило об’єктивно розділити їх на 3 кластери. Цікавим є те, що врахування прибутку до оподаткування змінило склад кластерів. Деякі об’єкти перемістилися з першого у третій кластер.

Таблиця 4.2

Результати кластерного аналізу при обліку валового прибутку

Показники	Од. вим.	Кластери, середні показники		
		1	2	3
Кількість об'єктів		27	7	1
Рілля	Га	6340,89	7226,86	17895,00
Луги	Га	919,11	5219,14	1350,00
Пасовища	Га	1206,48	4331,86	1420,00
Ліси	Га	1804,26	7362,86	0,00
Трактори	Штук	39,11	54,43	130,00
Зернозбиральні комбайни	Штук	11,37	14,14	52,00
Паливо	Тонн	357,19	387,71	1900,00
Врожайність зернових	ц/га	12,61	13,17	23,50
Дійні корови	Голів	486,48	525,29	1200,00
Надої на корову в рік	Літрів	1977,81	1626,29	2890,00
Яловичина в живій масі	Тонн	86,23	109,93	309,00
Кількість постійно зайнятих у господарстві	Чол.	182,26	222,43	690,00
Валовий прибуток	Тис. крб.	797,41	-86,14	27300,00

Таблиця 4.3

Результати кластерного аналізу при обліку до оподаткування

Показники	Од. вим.	Кластери, середні показники		
		1	2	3
Кількість об'єктів		11	13	3
Рілля	Га	4400,45	8160,15	5095,67
Луги	Га	233,09	1564,23	6819,00
Пасовища	Га	987,64	1849,96	4911,00
Ліси	Га	1116,91	2767,77	9792,00
Трактори	Штук	33,18	47,77	42,67
Зернозбиральні комбайни	Штук	9,36	13,08	11,67
Паливо	Тонн	215,82	465.15	203,67
Врожайність зернових	ц/га	13,05	13,48	12,67
Дійні корови	Голів	359,55	615,54	255,00
Надої на корову в рік	Літрів	1899,27	2116,38	1780,00
Яловичина в живій масі	Тонн	61,40	124,25	44,60
Кількість постійно зайнятих у господарстві	Чол.	133,91	223,15	141,33
Прибуток до оподаткування	Тис. крб.	-113,82	1893,08	-338,00
Валовий прибуток	Тис. крб.		621,77	-1347,00

4.3. Відносини власності, рентабельність і борги крупних сільськогосподарських підприємств

Розглядалися дані по 100 господарствам. Щоб запобігти впливу змінних з великими значеннями інтервалів між кластерними центрами по відношенню до змінних з малими значеннями, було застосовано нормалізацію даних.

Ієрархічний кластерний аналіз проводився для двох кластерів. У табл. 4.4 представлені результати утворення кластерів.

Таблиця 4.4

Результати кластерного аналізу при двох кластерах

Показники	Кластери	
	1	2
Віддаленість	-0,06	+0,08
С/г угіддя	-0,67	+0,70
Посіви	-0,40	+0,33
Величина земельних ділянок	-0,45	+0,50
Частка земельних ділянок у незайнятих	+0,39	+0,40
Частка капіталу у незайнятих у господарстві	+0,13	-0,09
Рентабельність	+0,24	-0,28
Борги	+0,22	-0,20
Дебіторська заборгованість	+0,11	-0,11
Кількість об'єктів	50	46

Кластер 1 охоплює підприємства, що мають в рентабельність вищу від середньої. Крім того, вони характеризуються нижче середнього віддаленістю від крупних міст, забезпеченістю землею і величиною земельних часток. У цих підприємств також вище середнього розмір землі і капіталу, що знаходиться в руках незайнятого на підприємстві населення, боргові зобов'язання і дебіторська заборгованість. Для підприємств 2-го кластера – навпаки. Оскільки нами були одержані детальні результати залежно від рентабельності і величини підприємства, доцільним стало утворити більшу кількість кластерів. При заданій величині в п'ять кластерів були одержані наступні результати (табл. 4.5).

Найбільше значення рентабельності має одне з дуже крупних підприємств (кластер 5). Крім того, у 31 крупного підприємства рентабельність витрат перевищує середню величину (кластер 1). Для цих підприємств перш за все характерно, що частка ріллі і капіталу, що знаходиться в руках незайнятого на підприємстві населення, а також сума боргів менше середнього значення. Значення рентабельності вище середнього мають також 17 дрібних (найдрібніших) підприємств (кластер 4), у яких частка ріллі і капіталу, що знаходиться в руках незайнятого на підприємстві населення, також знаходиться на рівні нижче середнього, проте при великих боргових зобов'язаннях. 45 крупних підприємств з невеликими площами досягли низької, в порівнянні з середнім значенням, рен-

табельності, пов'язаної з вищою часткою ріллі і капіталу, що знаходиться у власності незайнятого на підприємстві населення (кластер 2). Крім того, було виявлено 2 великі підприємства з рентабельністю набагато нижче середнього рівня. Очевидно це свідчить про вплив того, чи знаходиться земля і капітал в основному в руках зайнятих на підприємстві.

Таблиця 4.5

Результати кластерного аналізу при п'яти кластерах

Показники	Кластери				
	1	2	3	4	5
Віддаленість	0,03	-0,27	-0,02	0,67	1,09
С/г угіддя	0,86	-0,42	0,66	-0,69	1,44
Посіви	0,75	-0,47	0,71	-0,50	0,58
Величина земельних ділянок	0,22	-0,19	0,32	-0,34	7,55
Частка земельних ділянок у незайнятих у господарстві	-0,57	0,63	-0,54	-0,45	-1,10
Частка капіталу у незайнятих	-0,43	0,58	-0,66	-0,57	0,66
Рентабельність	0,19	-0,18	-0,25	0,05	0,64
Борги	-0,32	-0,29	0,77	1,34	0,25
Дебіторська заборгованість	-0,23	-0,10	6,19	0,01	-0,12
Кількість об’єктів	31	45	2	17	1

Для порівняння було проведено кластерний аналіз на основі ненормованих даних. Ієрархічний кластерний аналіз показав, як і для нормованих даних, необхідність утворення двох кластерів. У табл. 4.6 представлені змінні для обох кластерів, а також кількість впорядкованих в кластери підприємств.

Таблиця 4.6

Результати кластерного аналізу при двох кластерах

Показники	Міра	Кластери	
		1	2
Віддаленість	Км	249,65	229,28
С/г угіддя	Га	22017,75	8632,22
Посіви	Га	5639,20	2691,45
Величина земельних ділянок	Га	35,77	20,25
Частка земельних ділянок у незайнятих	%	45,00	53,00
Частка капіталу у незайнятих	%	44,55	48,51
Рентабельність	%	369,34	522,23
Борги	Крб./га	81,58	75,49
Дебіторська заборгованість	Крб./га	22,7	46,3
Кількість об’єктів		**20**	**76**

Кластер 1 містить 20 підприємств, які мають в порівнянні з підприємствами другого кластера великі площі сільськогосподарських угідь і посівів, вищу рентабельність, невелику частку ріллі і капіталу у власності не зайнятих на підприємстві, а також невеликі борги. Таким чином, з цієї таблиці можна було б зробити висновок, що для того, щоб досягти високої рентабельності, підприємство повинне бути якомога крупніше. З метою отримання детальніших результатів, потім був проведений кластерний аналіз з трьома, чотирма або п'ятьма кластерами. При аналізі з трьома кластерами колишній кластер 2 був розбитий на два кластери, і в один із знов утворених кластерів було додане підприємство з кластера 1 (табл. 4.7)

Таблиця 4.7

Результати кластерного аналізу при трьох кластерах

Показники	Міра	Кластери		
		1	2	3
Віддаленість	Км	256,47	225,09	231,55
С/г угіддя	Га	22.374,47	6.135,48	12.161,45
Посіви	Га	5.475,32	2.138,11	3.612,91
Величина земельних ділянок	Га	36,35	18,43	22,81
Частка земельних ділянок у незайнятих у господарстві	%	44,00	57,00	49,00
Частка капіталу у незайнятих у господарстві	%	44,16	47,34	50,10
Рентабельність	%	5,48	8,33	-6,68
Борги	Крб/га	388,78	537,59	485,93
Дебіторська заборгованість	Крб/га	84,76	48,13	110,32
Кількість об’єктів		19	33	44

При трьох кластерах стає ясно, що маленькі підприємства мають щонайвищу рентабельність (кластер 2), за ними слідують найбільші (кластер 1), потім середні. Найбільша величина боргів спостерігається у підприємств з найвищою рентабельністю, що може указувати на великий об'єм інвестицій в цій групі. Крім того, в руках не зайнятих на підприємстві зосереджена частка землі і капіталу, перевищуюча середнє значення. У табл. 4.8 представлені результати кластерного аналізу, при якому було утворено 5 кластерів. Тут ще більш помітно, що найвищої рентабельності досягають як малі, так і великі підприємства. Подібно табл. 4.7 стає очевидним, що щонайвищої рентабельності досягають як малі, так і великі підприємства, тоді як середні підприємства знаходяться на рівні нижче середнього. На малих підприємствах велика частина капіталу знаходиться у володінні зайнятих на підприємстві, переважно земельними частками володіють не зайняті на підприємстві. На крупних підприємствах, де рента-

бельність вище середнього рівня – навпаки. Цей результат суперечить даним одержаним при застосуванні нормування і наочно показує, що групи були утворені іншим чином.

Таблиця 4.8

Результати кластерного аналізу при п'яти кластерах

Показники	Міра	Кластери				
		1	2	3	4	5
Віддаленість	Км	218,47	233,86	267,60	252,50	235,13
С/г угіддя	Га	8485,4	4118,9	25915,8	21109,7	13052,4
Посіви	Га	3363,8	1259,2	3658,80	6124,07	3300,63
Величина земельних ділянок	Га	20,08	17,01	33,31	37,44	23,50
Частка земельних у незайнятих у господарстві	%	54,00	58,00	34,00	48,00	49,00
Частка капіталу у незайнятих у господарстві	%	47,31	47,57	26,00	50,64	50,98
Рентабельність	%	4,12	16,55	-14,40	12,58	-13,89
Борги	Крб/га	450,50	669,09	312,84	415,90	467,61
Дебіторська заборгованість	Крб/га	101,54	60,80	21,46	107,37	51,33
Кількість об’єктів		32	21	4	14	24

У таблиці 4.9 представлений розподіл підприємств по кластерах згідно організаційно-правовим формам.

Таблиця 4.9

Розподіл підприємств по кластерах згідно організаційно-правовим формам

ОПФ	Кластери				
	1	2	3	4	5
СПК	16	10	4	9	8
Колгоспи	6	1	0	3	12
ТОВ	3	5	0	0	0
АТ	6	5	1	0	3
Інші	1	0	0	2	1

СПК знаходяться перш за все в кластерах 1,2,4 (рентабельні підприємства) і 5 (нерентабельні підприємства), колгоспи – в кластерах 5 і 1. ТОВ і АТ переважають в кластерах 1 і 2.

4.5. Дослідження даних хімічного моніторингу за допомогою технології кластерного аналізу

Район Кривбасу є зоною екологічного лиха, що обумовлено забрудненням грунтових, поверхнех вод та особливо підземних вод під впливом гірновидобувної промисловості, металургійної та хімічної промисловості. Маємо дані гідрохімічного моніторингу з сведловин Криворізького гірнозбагачувального комбінату, вони містять заміри вмісту хімічних речовин у 14 свердловинах.

Перед застосуванням сферичного методу двоступінчастої кластеризації з виділенням ядра (згущення) об'єктів класифікації вихідні дані пронормували та перевірили на некорельованість.

Таблиця 4.10

Дані замірів хімічного моніторингу району Криворізького гірнозбагачувального комбінату у березні 1990 року.

№ свердловини	Вміст хімічних речовин														
	CO_2	CO_3	NH_4	Fe	Fep	Go	Gu	S	SO	HCO_3	Cl	SO_4	Ca	Mg	Na
3	17,8	13,2	0,8	0,5	0,3	11,5	11,5	1926	1988	880	373	127	49	110	370
4	13,3	0	0,6	0,5	0,3	7,84	7,84	1044	1112	562	186	24,7	29,4	77,5	162
5	0	25,2	0,6	0,3	0,5	13,7	10,7	2368	2432	655	171	868,	58,8	131	457
7	21	0	0,8	0,3	0,5	9,07	9,07	1846	1922	729	186	399,	49	80,5	378
8	0	14,4	0,6	0,2	0,3	15,4	15,4	2031	2122	993	249	250	88,2	134	300
11	8,9	0	0,6	0,3	0,5	22,0	2,44	3307	3350	148	264	1913	161	169	647
14	0	23,4	0,6	0	0	7,35	7,35	1257	1344	683	155	74	39,2	65,5	215
15	0	212	0	1	1	1,23	1,23	3719	3965	784	218	1400	19,6	3	1078
21	0	300	0,8	0,2	0,8	12,5	12,5	3592	3876	1678	327	382	24,6	137	740
22	0	22,2	0,6	0,5	0,5	25,9	7,16	4271	4344	436	268	2332	107	250	855
23	0	32,4	0,8	0,3	0,5	6,38	6,24	1770	1834	380	28	559	44,2	50,7	437
25	8,7	15	0,8	0,5	0,3	8,09	5,95	1602	1670	363	218	543	29,4	80,5	352
27	10,2	0	0,8	0,5	0,5	23,0	1,68	3593	3634	102	358	2077	98	220	733
31	17,8	0	0,8	0,3	0,5	14,7	14,7	2261	2302	1317	327	32,9	49	149	384

Таблиця 4.11

Розподіл свердловин по кластерах згідно вмісту хімічних представляються у вигляді текстових файлів з структурою:

Координата X	Координата У	Координата Z	№ свердловини	Концентрація хім. речовини	Номер класа
162	98	131.80	3	1926	2
161	78	130.10	4	1044	1
178	98	132.71	5	2368	2
187	108	130.46	7	1846	2
163	122	124.54	8	2031	2
148	123	121.42	11	3307	1
183	62	137.39	14	1257	3
135	78	132.12	15	3719	3
223	117	114.21	21	3592	1
86	114	122.34	22	4271	2
95	113	126.43	23	1770	1

Координата X	Координата У	Координата Z	№ свердловини	Концентрація хім. речовини	Номер класа
88	137	123.87	25	1602	1
76	153	121.02	27	3593	2
177	108	131.32	31	2261	1

Графічно результаити результати кластеризації представлені на рис. 4.1.

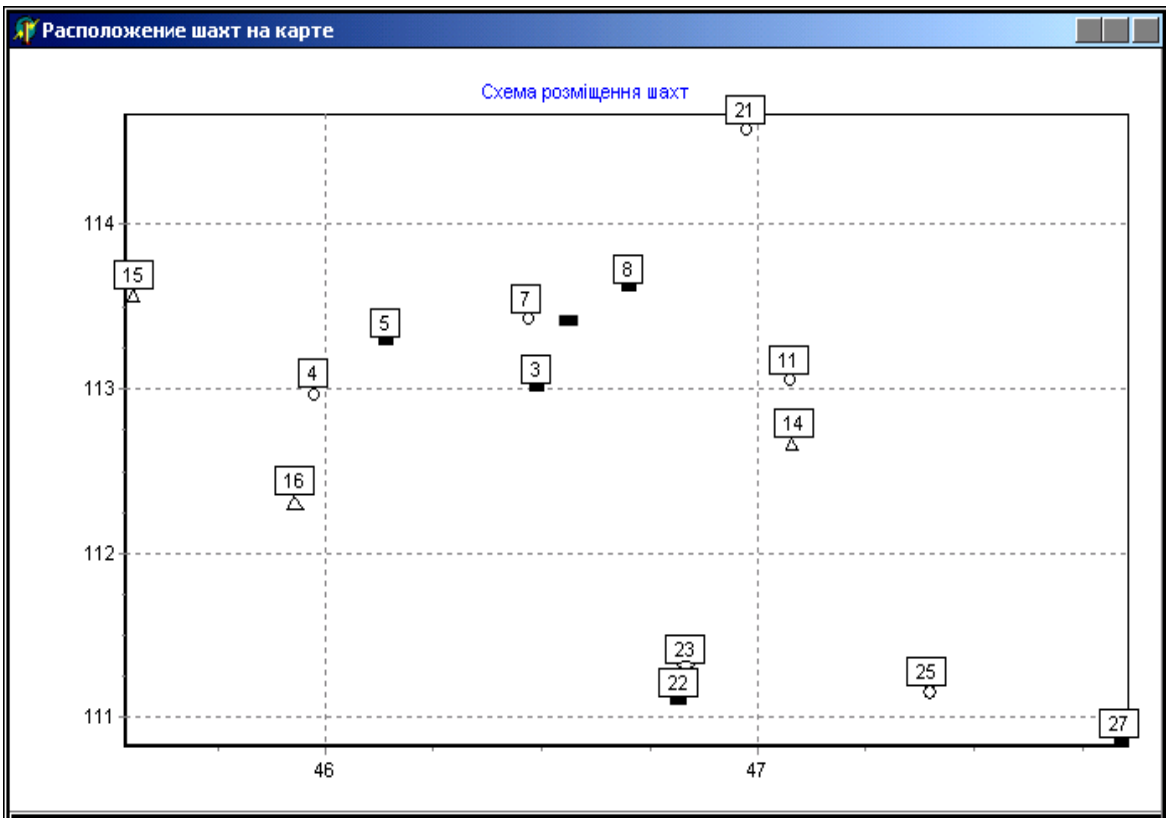


Рис. 4.1. Графічне представлення результатів кластеризації свердловин.

Порівнюючи розбиття на класи отримані в різні пори, на результати класифікації впливають різні гідрохімічні процеси, що приймають участь в накопиченні елементів в підземних водах, осадок, дифузія, накопичення, сорбція. Висновок, по основним хімічним елементам (по хлору, сульфатам, карбонатам та мінералізації), в залежності від сезону року міняються розбиття, тобто на концентрацію цих речовин впливають сезонні коливання і розташування свердловин. Суттєвим є діаметр класу, тобто окіл в межах якого скважини потрапляють в один клас. Якщо його обрати великим, то до одного класу потрапить більша кількість скважин, в яких значення вмісту хімічних елементів будуть мати більший діапазон. Якщо малий – то кількість класів буде більшою, і значення їх ознак більш унікальні.

У розділі показані приклади застосування кластерного аналізу діяльності туристичних організацій, сільськогосподарських підприємств і показана можливість глибоких висновків про їх фінансових стан.

ПІДСУМКИ

В посібнику весь матеріал подано у чотирьох розділах.

У першому з них ви дізналися про місце кластерного аналізу серед інших методів класифікації. Виявляється, оскільки класифікація в цьому методі виконується по всіх ознаках водночас, то він є найбільш прийнятним для класифікації економічних об’єктів. У цьому ж розділі подано дев’ять різних метрик відстаней у гіперпросторі. Показано, що для об’єктивної класифікації всі параметри потрібно спочатку нормувати. Наведено два правила нормування.

У другому розділі розглядаються дев’ять алгоритмів кластеризації об’єктів. Для першого з них, методу повного перебору, автором розроблено методику застосування лінійного цілочисельного 0-1 програмування. Інші, такі як кластеризація методом перебору фіксованих відстаней від центрів сфер, сферичний метод двоступінчастої кластеризації з виділенням ядра (згущення) об'єктів класифікації, кластеризація інтегральним методом геометризації інформаційного поля, метод визначення центра кластера за допомогою обчислення середньо арифметичних відстаней між об'єктами, метод постійних кластерів і характеристик, кластеризація з урахуванням критерію якості і вибором кращого варіанта за цим критерієм, ієрархічне угрупування та алгоритм нечіткої кластеризації, методом *c*-середніх вимагають покрокового застосування або розробки спеціалізованих програм.

В третьому розділі подано розрахунок статистично достовірного числа кластерів, розрахунок статистичних характеристик утворених кластерів, які дозволяють додатково пересвідчитися про те, що утворені кластери не пересікаються. Автором наведена розроблена ним оригінальна методика віднесення нового об’єкта до раніше утворених кластерів.

У четвертому розділі подано приклади застосування кластерного аналізу в економіці. Зокрема для оцінки ступеня відмінності регіонів в діяльності туристсько-готельного господарства, визначення чинників, що впливають на прибуток і рентабельність крупних сільськогосподарських підприємств, розрахунку відносин власності, рентабельності з боргами крупних сільськогосподарських підприємств.

Додатки містять основні терміни та визначення по предмету.

Кожен розділ містить приклади розрахунку за запропонованими методиками із застосування статистичних функцій електронних таблиць Calc та Excel, що значно прискорює всі розрахунки.

Індивідуальні завдання, якими закінчується кожен розділ, дають можливість глибше засвоїти матеріал посібника, набути вміння робити розрахунки на комп’ютері.

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Айвазян С.А. Классификация многомерных наблюдений. – М.: Статистика. – 1974. – 240 с.
2. Апраушева Н.Н. Предварительное обнаружение идеальных кластеров. – М.: ВЦ АН СССР. – 1986. – 20 с.
3. Беллман Р Дрейфус С. Прикладные задачи динамического программирования. – М.: Наука. – 1965. – 406 с.
4. Белов А.Г. Вероятностно-статистические методы при экспериментальном разделении множественных процессов. – М.: МГУ. – 1985. – 141 с.
5. Гиттис Л.Х. Кластерный анализ. Основные идеи и методы. – М.: МГУ. – 2000. – 62 с.
6. Данилов Б.С. Кластерный анализ в EXCEL / тез. докл. XIX Всероссийская молодежная конференция "Строение литосферы и геодинамика", г. Москва, 25-30 октября 2002 г. – С. 18.
7. Дюран Б., Оделл П. Кластерный анализ. –М.: Статистика. – 1977. – 128 с.
8. Жамбю М. Иерархический кластерный анализ и соответствие. – М.: Финансы и статистика. – 1988. – 342 с.
9. Зенкин А.И., Ленский А.А. Ускорение решения задач классификации в факторном пространстве. – М.: ВЦ АН СССР. – 1990. – 24 с.
10. Иберла К. Факторный анализ. – М.: Статистика. – 1980. – 398 с.
11. Ивахненко А.Г. Объективная кластеризация на основе теории самоорганизации моделей. /Автоматика. – 1987. – № 5. – С. 6-15.
12. Кильджиев Г.С. Аболенцев Ю.И. Многомерные группировки. – М.: Статистика. – 1978. – 160 с.
13. Клигер С.А. Шкалирование при сборе и анализе социологической информации. – М.: Наука, 1978. – 112 с.
14. Пиотровский А. Денисов А. Кластерный анализ как инструмент подготовки эффективных маркетинговых решений /Практический маркетинг. – №5. – 2001 – С.211-232.
15. Прикладная статистика. Классификация и снижение размерности /С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика. – 1989. – 607 с.
16. Стиранка А.И. Решение кластерной задачи большой размерности в нечеткой постановке./Кибернетика. – 1991.– № 1. – С. 116 -121.
17. Терехина А.Ю. Анализ данных методами многомерного шкалирования. – М.: Наука. – 1986. – 166 с.
18. Томашевич В.Н.Многомерный статистический анализ в экономике. – М.: Юнити, 1999. - 599 с.

Додаток.
Основні терміни та визначення

Абстрактна (неформальна) група об'єктів – множина об'єктів, розділених штучними методами (наприклад, кластерним аналізом). Підмножини (кластери) містять у собі об'єкти, об'єднані близькістю характеристик, але по формальних ознаках приналежним різним множинам. Не виключено, що в деяких випадках формальна й абстрактна структура можуть збігатися.

Адекватність кластерного аналізу – рівень відповідності результатів рішення задачі кластерного аналізу й еталонної моделі класифікації тих же об'єктів, отриманих іншими, найчастіше змістовними методами.

Альтернативні характеристики об'єкта – окремі характеристики, що можуть приймати тільки одне фіксоване значення. Наприклад, у соціологічних опитуваннях часто використовуються такі питання до респондентів, на які можна давати тільки однозначні взаємовиключні відповіді. Для обробки альтернативних характеристик розроблені спеціальні процедури.

Ботриологія (термін введений И. Дж. Гудом) – одна з маловідомих назв кластерного аналізу.

Варіаційні методи кластеризації – комплекс формалізованих методів класифікацій об'єктів, заснованих на розгляді цільових функцій, відстаней між об’єктами та групами, інших статистичних показників.

Виділення факторів (extraction of factors) – первісний етап факторного аналізу; ко-варіаційна матриця відтворюється за допомогою невеликого числа схованих факторів або компонент.

Вирішальне правило кластеризації – визначена формальна процедура, як правило реалізована на комп'ютері, що дозволяє вибирати оптимальний варіант або зону раціональних рішень задачі кластеризації.

Відсівання критерій (scree-test) – евристичний критерій визначення числа факторів, заснований на графічному зображенні усіх власних значень кореляційної матриці; застосовується при впливі другорядних факторів.

Відстані функція (міра) - у теорії розпізнавання образів функція відстані вимірює подібність і розходження між досліджуваними об'єктами в деякій реальній або умовній метриці. Ця функція може носити не тільки кількісний, але і характер переваг, експертних оцінок, функцій «меню», результатів соціологічних опитувань.

Відстань між двома соціально-економічними об'єктами - метрична величина (або функція), що визначає подібність між даними об'єктами. Може вимірятися як у метриці Евкліда, так і в довільно визначеній.

Власне число (proper number) – характеристика матриці, використовується одночасно як критерій визначення числа виділюваних факторів і як мара дисперсії, що відповідає даному факторові.

Власний вектор (eigenvector) – вектор, зв'язаний з відповідним власним числом; виходить у процесі виділення первісних факторів; ці вектори, представлені в нормованій формі, є факторними навантаженнями.

Головні компоненти (principal components) – лінійна комбінація змінних, що володіє властивістю ортогональності; перший головний компонент відтворює найбільшу частку дисперсії вихідних даних, другий – наступну по величині частку і т.д.

Дискримінантна функція (discriminate function) – статистика, що служить для побудови правила класифікації об'єктів по групах.

Дискримінантна множина (discriminate score) – це база для віднесення об'єктів і індивідуумів до якої-небудь визначеної групи.

Дискримінантні ваги або коефіцієнти дискримінантної функції (discriminate weights) – параметри рівняння дискримінантної функції, що дозволяють оцінити здатність конкретних незалежних змінних визначати розходження в групах об'єктів або індивідуумів. Незалежні змінні, що істотно впливають на розходження в групах, мають великі ваги, а ті змінні, котрі мають незначний вплив, - маленькі ваги. У результаті аналізу необхідно вибрати ті змінні, котрі в більшій мірі визначають імовірність влучення якого-небудь об'єкта в конкретну групу. Коефіцієнти дискримінантної функції можуть бути представлені в стандартизованій і не стандартизованій формах.

Дискримінантні змінні (discriminate variables) – характеристики, застосовувані для того, щоб відрізняти один клас від іншого; вони повинні вимірятися або за інтервальною шкалою або за шкалою відносин. Таким чином, стає можливим обчислення математичних сподівань, дисперсій і правомірне використання математичних рівнянь.

Дискримінантний аналіз (discriminate analysis) – статистичний метод, що дозволяє вивчати розходження між двома і більш об'єктів по декількох ознаках одночасно.

Дискримінантні лінії – визначають границі кластерів або з'єднують які-небудь спеціально позначені межі інформаційного поля.

Дискримінантні функції – формальні правила віднесення об'єктів до якої-небудь підмножини і описують контури цих підмножин.

Дисперсія (variance) – ступінь коливання ознаки, його варіація, породжувана всією сукупністю діючих на нього факторів, обчислюється як середній квадрат відхилень значень ознаки від загальної середньої.

Довільний кластер – підмножина, сформована на основі довільного трактування характеристик об'єктів, що не враховують об'єктивних показників.

Евристичний метод кластеризації – інтуїтивний метод поділу множини об'єктів на кластери, заснований на досвіді дослідника, апріорних представленнях про кількість кластерів і їхньому змісті.

Загальний фактор (common factor) – не вимірювана (гіпотетична) схована величина, що враховує кореляцію принаймні між двома значеннями змінних.

Зважена ознака – характеристика об'єкта вивчення (класифікації), що складається з двох величин: абсолютного значення цієї величини і деякого до-

вільного коефіцієнта (ваги змінної), що відбиває об'єктивну або суб'єктивну оцінку значимості цієї характеристики в досліджуваному процесі. Застосування зважених ознак при кластеризації є дискусійним і вимагає обґрунтування.

Згоди показник – рівень відповідності змістовної структури множини й отриманого на його основі кластерного поля.

Ієрархічна кластеризація – ранжирування кластерів, отриманих у результаті класифікації деякої множині, по визначеним у постановці задачі критеріям.

Кількісна ознака – характеристика об'єкта вивчення, що має числові значення, для яких справедлива аксіоматика Евкліда.

Класифікація – упорядкування об'єктів по їхній схожості.

Кластер (таксон, образ, клас) – множина умовно однорідних (схожих) елементів (об'єктів). Ступінь однорідності (подібності) може бути різним і визначається цілями класифікації. Ці поняття використовуються при різних видах класифікації і розпізнавання образів.

Кластерний аналіз (кластер-аналіз) – задача поділу неоднорідної множини на деяку кількість умовно однорідних підмножин (кластерів), обґрунтування цього поділу й інтерпретація отриманих результатів.

Коефіцієнт асоціації (coefficient of association) – оцінка ступеня тісноти зв'язку між якісними ознаками, кожна з яких представлена у виді альтернативної ознаки. Однак у тих випадках, коли один з чотирьох показників відсутній, величина коефіцієнта дорівнює 1, що дає перебільшену оцінку зв'язку між ознаками.

Коефіцієнт взаємної спряженості Пірсона (Pearson’s mutual coefficient of a contingency) – оцінка ступеня тісноти зв'язку між якісними, але не альтернативними ознаками.

Коефіцієнт контингенції (contingent coefficient) – оцінка ступеня тісноти зв'язку між якісними ознаками, кожна з яких, також як і для коефіцієнта асоціації, повинна бути представлена у виді альтернативної ознаки. Однак коефіцієнт контингенції за абсолютною величиною менше коефіцієнта асоціації.

Коефіцієнт кореляції (correlation coefficient) – числова характеристика спільного розподілу двох випадкових величин, що виражає їхній взаємозв'язок. Якщо коефіцієнт більше 0, то при збільшенні значень однієї з величин, друга має тенденцію до збільшення; якщо менше нуля – до зниження.

Коефіцієнт кореляції рангів Спірмена (Spearman’s rank correlation coefficient) – непараметрична оцінка, що дозволяє виміряти тісноту зв'язку як між кількісними, так і між якісними ознаками. Вона заснована на розгляді різниці рангів значень факторної і результативної ознак.

Коефіцієнт множинної детермінації (multiple determinant coefficient) – загальний показник тісноти зв'язку усіх вхідних у рівняння регресії факторів з результативним показником. Він являє собою відношення частини варіації результативної ознаки, що пояснюється за рахунок варіації включених у рівняння факторів, до загальної варіації результативної ознаки.

Коефіцієнт розходження – числова характеристика двох об'єктів або

двох кластерів, що мають фіксований набір показників. Обчислюється по алгоритмах подібним з обчисленнями функцій відстані між об'єктами.

Конгруентні (подібні) об'єкти – два або більш різні об'єкти, що мають цілком співпадаючі характеристики.

Кореляція (correlation) - лінійна залежність між випадковими змінними, що не має строго функціонального характеру, при якому зміна однієї з випадкових величин приводить до зміни математичного сподівання іншої.

Косокутне обертання (oblique rotation) – перетворення у факторному аналізі, за допомогою якого виходить проста структура; фактори обертаються без накладення умови ортогональності, і результуючі фактори корелюють один з одним.

Критерій Стьюдента (t-критерій) (Student's t-test) – статистичний критерій, що використовує при оцінці несуперечності статистичної гіпотези результатам спостережень функцію, що має розподіл Стьюдента.

Лінійна комбінація (linear combination) – сума, у яку змінні входять з постійними вагами.

Лінійна система (linear system) – лінійна залежність між перемінними, у факторному аналізі – модель, у якій вимірювані величини лінійно зв'язані зі схованими факторами.

Меню функція – найбільш важлива нечислова функція, що характеризує переваги людини (експерта, респондента, покупця і т. п). Використовується при опитуваннях. Відповіді на питання типу «меню» можуть мати кілька варіантів, що обробляються в деякому нормованому числовому просторі.

Метод максимальної правдоподібності (method of maximum likelihood) – метод статистичного оцінювання, у якому визначається значення перемінних генеральної сукупності з використанням вибіркового розподілу; у факторному аналізі – метод одержання первісного факторного рішення.

Міра близькості двох об'єктів – числова характеристика, що визначає збіг показників (у відносних числах), що володіє властивостями геометричної відстані.

Монотетичні кластери – умовно однорідні об'єкти, згруповані по ознаці збігу однієї характеристики. При цьому передбачається, що обрана характеристика є основною і вирішальною.

Навчання розпізнаванню образів – задача вибору комп'ютером найкращого варіанта класифікації об'єктів і можливого автоматичного удосконалювання формального і змістовного апарата алгоритму розпізнавання.

Найменших квадратів метод (least-squares solution) – рішення, для якого мінімізується сума квадратів відхилень між величинами, що спостерігаються і передбачуваними значеннями; у факторному аналізі – метод одержання первісного факторного рішення.

Непараметричні методи класифікації – такі методи угруповання об'єктів, що не припускають споконвічного знання властивостей розділених груп, їхніх характеристик і розподілу елементів по групах. Таким чином, кластерний аналіз відноситься до непараметричних методів класифікації.

Нестандартизовані коефіцієнти (raw coefficients) – коефіцієнти, що надають інформацію про абсолютний внесок перемінної в значення дискримінантної функції.

Об'єкт – у теорії розпізнавання образів елементарна одиниця, що поєднується з іншими об'єктами, утворюючи класи, таксони, кластери та інші складні множині. Фізична, економічна, соціальна або інша природа об'єктів для класифікації об'єктів значення не має. Узагалі говорячи, «об'єктом можна назвати усе, що завгодно, включаючи процеси і дії - усе, чому можна приписати вектор дескрипторів» (Р.Р. Сокал).

Об'єктивний метод кластеризації – метод, заснований на дотриманні точних формалізованих правил.

Ординація - один з методів класифікації, що використовує безперервну шкалу вимірів і не враховує дискретних характеристик об'єктів. Ординація в деякому змісті, є альтернативою кластерному аналізові.

Ортогональне обертання (orthogonal rotation) – перетворення, за допомогою якого виходить проста структура при виконанні обмеження ортогональності (не корельованості) факторів; фактори, виділювані за допомогою цього обертання, по визначенню, некорельовані.

Політетичні кластери – класифікація по принципах групи емпіричних ознак, що діють тільки у комплексі, причому кожна з цих ознак, використана на самоті або іншій комбінації не дають подібної класифікації. Таким чином, однорідність політетичних кластерів залежить від числа ознак. Найбільше поширення політетичні кластери одержали в тих задачах, де шкала відстаней між елементами має нечисловий характер, а самі кластери відповідають аксіоматиці переваг вхідних реальних об'єктів.

Часткові коефіцієнти кореляції (individual correlation coefficients) – показники тісного зв'язку між результативною ознакою й одною з факторних ознак, тобто не обумовленої дією інших факторів, включених у модель.

Природна кластеризація – класифікація, заснована на максимально великій кількості ознак, що забезпечують найбільшу подібність кластерів. При такому підході приналежність об'єкта кластерові визначається природним образом у результаті їхнього об'єктивного порівняння.

Проста структура (simple structure) – спеціальний термін, що відноситься до факторної структури, що має визначені властивості: перемінні повинні мати навантаження на мінімальне число загальних факторів, кожен загальний фактор повинен навантажувати деякі перемінні і не навантажувати інші.

Простір ознак – *n* - вимірний простір, що складається з незалежних векторів - значимих ознак об'єкта. Розмірність простору може бути перетворена в меншу за рахунок скорочення лінійно залежних ознак. При цьому можлива втрата інформації.

Ранжирування – розташування елементів множині у визначеному порядку, відповідно до їх рангів (по значимості, убуванню, подільності, привабливості і т.п.)

Рівняння регресії (regression equation) – спосіб апроксимації тісної регресійної залежності. Воно описує зміну умовного середнього значення результуючого показника в залежності від зміни факторної ознаки.

Розпізнавання образів теорія - науково-практичний напрямок, зв'язаний з необхідністю класифікації деякої множини об'єктів, установлення приналежності об'єкта якому-небудь класові, що має відомі загальні характеристики. Розпізнавання виконується за допомогою зіставлення характеристик (ознак) об'єктів, класів. Порівняння цих ознак і дає підставу робити висновки про ступінь відповідності розглянутих образів. Теорія розпізнавання образів може застосовуватися в гірничо-геологічних дослідженнях, економіці, соціології, діагностиці й інших науках.

Специфічність (specific component) – частка дисперсії змінної, що спостерігається відповідному специфічному факторові; застосовується для позначення частини характерності, одержуваної при виключенні дисперсії помилки.

Спільність (communality) – частка дисперсії перемінних, обумовлена загальними факторами, у моделі з ортогональними факторами вона дорівнює сумі квадратів факторних навантажень.

Стандартизовані коефіцієнти (standardized coefficients) – характеристики відносного внеску дискримінантних змінних у значення дискримінантної функції.

Статистика (W) Уілкса (Wilks' Lambda) – міра розходжень між класами по декількох дискримінантним змінним. Величини *W*, близькі до нуля, свідчать про високий ступінь розходження між класами. Максимальне значення *W*, рівне 1, характеризує відсутність розходжень між класами.

Статистика F – включення (*F* – enter) – є частковою *F*- статистикою з числом ступенів волі (g – 1) і (n – p – g + 1); оцінює поліпшення розходження від використання розглянутої змінної в порівнянні з розходженням, досягнутим за допомогою інших уже відібраних змінних. Якщо величина *F* – включення мала, то навряд чи необхідно відбирати таку змінну, тому що вона не дає досить великого внеску в розходження.

Статистика F- видалення (*F* – remove) – також є частковою *F*- статистикою з числом ступенів волі (g – 1) і (n – p – g + 1). Однак вона оцінює значимість погіршення розходження після видалення перемінної зі списку уже відібраних перемінних. Статистика *F*- видалення використовується для ранжирування дискримінантних можливостей відібраних змінних.

Структурні коефіцієнти (structure coefficients) – коефіцієнти кореляції між окремою змінною і дискримінантною функцією. Їх називають "повними структурними коефіцієнтами". Коли абсолютна величина структурного коефіцієнта велика, вся інформація про дискримінантної функції укладена в цій змінній. Якщо коефіцієнт близький до нуля, то зв'язок між ними малий.

Суб'єктивний метод кластеризації – метод класифікації, заснований на особистих представленнях і перевагах дослідника.

Таксон – те ж, що і кластер.

Таксономія (грець.) – теорія класифікації і систематизації складних об'єктів, що ма-

ють числову і нечислову метрику, але доступну для ієрархічного поділу.

Толерантність (tolerance) – тест по перевірці відібраних для аналізу перемінних. Якщо перемінна, що перевіряється, є лінійною комбінацією (або приблизно дорівнює лінійній комбінації) однієї або декількох відібраних перемінних, то її толерантність дорівнює нулеві (або близька до нуля). Таку перемінну небажано використовувати в розрахунках, тому що вона не дає ніякої нової інформації, але створює обчислювальні проблеми. Толерантність ще не відібраної перемінної дорівнює одиниці мінус квадрат множинної кореляції між цією перемінною і усіма уже відібраними перемінними.

Упорядкована кластеризація – розташування кластерів у порядку їхньої значимості за критеріями поставленої задачі.

Фактори (factors) – гіпотетичні, безпосередньо не вимірювані, сховані перемінні, підрозділяються на загальні і характерні.

Факторне навантаження (factor loading) – загальний термін, що означає коефіцієнти матриці факторного відображення або структури.

Факторного відображення матриця (factor pattern matrix) – матриця коефіцієнтів, у якій стовпці відповідають загальним факторам, а рядки – перемінним, що спостерігаються.

Факторної структури матриця (factor structure matrix) – матриця коефіцієнтів кореляції між змінними і факторами, у випадку ортогональних факторів збігається з матрицею факторного відображення.

Характерність (unique component) – частка дисперсії змінної, не зв'язана з загальними факторами і властива саме даній змінній.

Цільова функція (для кластерного аналізу) – деякий критерій оптимальності розбивки вихідної множини на підмножини відповідно до представлень дослідника щодо оптимальності.

Шкала – вибір одиниці виміру відстаней між двома об'єктами (кластерами, таксонами). Обрана шкала може вимірятися в абсолютних величинах (м, кг, руб.) або відносних (відсотки).

Ядро кластера – область найбільшого згущення об'єктів класифікації усередині кластера. Як правило, кластер має одне ядро, що займає не більш 10-25% простору (гіперпростору) кластера. Питання співвідношення ядра і самого кластера теоретично не прояснені і вирішуються емпіричними або інтуїтивними методами.

Якісна ознака - характеристика об'єкта вивчення, що не має числових значень. Процедури, у яких використовуються якісні ознаки, використовують функції збігу або інше поле ознак.

ПРЕДМЕТНИЙ ПОКАЖЧИК

Навчальне видання

Пістунов Ігор Миколайович
Антонюк Оксана Петрівна
Турчанінова Інна Юріївна

КЛАСТЕРНИЙ АНАЛІЗ В ЕКОНОМІЦІ

(Навчальний посібник)

Редакційно-видавничий комплекс

У редакції авторів

Комп’ютерна верстка І.М. Пістунова

Електронне видання