О.М. Мацуга, Ю.М. Архангельська, Н.М. Єрещенко

НАВЧАЛЬНИЙ ПОСІБНИК ДО ВИВЧЕННЯ КУРСУ «ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ РОЗПІЗНАВАННЯ ОБРАЗІВ»

2016

Міністерство освіти і науки України
Дніпропетровський національний університет
ім. Олеся Гончара

О.М. Мацуга, Ю.М. Архангельська, Н.М. Єрещенко

НАВЧАЛЬНИЙ ПОСІБНИК ДО ВИЧЕННЯ КУРСУ «ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ РОЗПІЗНАВАННЯ ОБРАЗІВ»

Дніпропетровськ
РВВ ДНУ
2016

УДК 004.93'1(075.8)
ББК 32.97я73
М 12

Рецензенти: д-р фіз.-мат. наук, проф. В.Б. Говоруха
д-р техн. наук, проф. О.Г. Байбуз

М 12 Мацуга, О.М. Навчальний посібник до вивчення курсу «Інформаційні технології розпізнавання образів» [Текст] / О.М. Мацуга, Ю.М. Архангельська, Н.М. Єрещенко. – Д.: РВВ ДНУ, 2016. – 60 с.

Розглянуто основи теорії розпізнавання образів. Подано методи класифікації і кластеризації даних, а також формування набору інформативних ознак. Роботу методів проілюстровано прикладами.

Для студентів ДНУ, які навчаються за напрямами підготовки «Програмна інженерія» та «Прикладна математика», а також студентів інших напрямів підготовки й аспірантів, які цікавляться теорією розпізнавання образів.

Темплан 2016, поз.

Навчальне видання

Ольга Миколаївна Мацуга
Юлія Михайлівна Архангельська
Наталія Миколаївна Єрещенко

**Навчальний посібник
до вивчення курсу
«Інформаційні технології
розпізнавання образів»**

Редактор Л.В. Дмитренко
Техредактор Л.П. Замятіна
Коректор Л.В. Дмитренко

Підписано до друку 29.01.16. Формат 60x84/16. Папір друкарський. Друк плоский.
Ум. друк. арк. 3,5. Ум. фарбовідб. 3,5. Обл. вид. арк. . Тираж 30 пр. Зам. №

РВВ ДНУ, просп. Гагаріна, 72, м. Дніпропетровськ, 49010.
ПП «Ліра ЛТД», вул. Погребняка, 25, м. Дніпропетровськ, 49010.
Свідоцтво про внесення до Державного реєстру
серія ДК №188 від 19.09.2000 р.
Фактична адреса: вул. Наукова, 5

ВСТУП

Задачу розпізнавання образів кожна людина розв’язує постійно. Читаючи текст, переходячи дорогу, ми розпізнаємо букви, цифри, колір сигналу світлофора, і цей процес здається настільки простим і природним, що ми навіть не замислюємося, що розв’язуємо якусь задачу. Взагалі, людина має дивовижну властивість досить легко розв’язувати задачі розпізнавання образів. Але з технічним прогресом і комп’ютеризацією суспільства виникла потреба в автоматизації процесу розпізнавання. У результаті стали розробляти методи розпізнавання образів. Вони – предмет розгляду в даному навчальному посібнику.

Під **розпізнаванням образів** розуміють процес віднесення деякого об’єкта до певного образу.

Розрізняють два види розпізнавання образів: із навчанням та без навчання.

Розпізнавання з навчанням (або класифікація) передбачає проведення навчання, у ході якого подають об’єкти, відносно яких відомо, до якого образу вони насправді належать. Таким чином, вивчаючи представників кожного образу, виявляючи їх особливості, навчають ці образи розрізняти. Після того як навчання завершене, починають процес розпізнавання, під час якого подають новий об’єкт (один або декілька) і розпізнають образ, до якого він належить.

Методи розпізнавання з навчанням застосовують для розпізнавання тексту (написаного від руки або друкованого), зображень (наприклад, облич, відбитків пальців, підписів, автомобільних номерів тощо), медичної та технічної діагностики, оцінки кредитоспроможності клієнтів банку та в багатьох інших галузях.

Розпізнавання з навчанням ще називають розпізнаванням із учителем, навчанням із учителем, дискримінантним аналізом, класифікацією. Саме останній термін ми використовуватимемо в цьому виданні.

Розпізнавання без навчання (або кластеризацію) можна порівняти з дитячою грою «знайди схожі об’єкти». Суть задачі у тому, щоб розділити об’єкти на групи за їх схожістю. На відміну від попереднього випадку, тут інформація про справжні групи-образи відсутня, їх потрібно визначити у ході розпізнавання. Такий вид навчання застосовують:

1) щоб вивчити структуру даних: з’ясувати, чи є множина об’єктів однорідна або в ній є декілька груп схожих між собою об’єктів, визначити кількість груп;

2) зменшити обсяг даних; якщо об’єктів багато, їх кількість можна скоротити, залишивши одного або декількох найбільш типових представників кожного образу;

3) побудувати нові класифікації для маловивчених явищ.

Замість «розпізнавання без навчання» часто використовують терміни: «розпізнавання без учителя», «навчання без учителя», «самонавчання», «автоматична класифікація», «кластеризація», «таксономія». Останній термін спочатку використовували лише для визначення класифікації видів тварин і рослин, але в наш час під таксономією розуміють розбиття на групи та впорядкування об’єктів різної природи.

Основними в обох видах розпізнавання є поняття образу та об’єкта.

Образ – це множна об’єктів, яким притаманні деякі спільні властивості, завдяки чому об’єкти одного образу можна розглядати як подібні й об’єднувати і в той же час відрізняти від об’єктів інших образів. У різних задачах термін «образ» заміняють на термін «клас» або «кластер».

Об’єкт (або предмет, явище, ситуація, сигнал) у кожній задачі має різну фізичну природу. Наприклад, у задачі медичної діагностики об’єктами є пацієнти, у задачі розпізнавання тексту або облич – відповідні зображення.

Застосування методів розпізнавання образів потребує, щоб кожен об’єкт, незалежно від його фізичної природи, був закодований, тобто поданий у вигляді числового вектора

$$X_i = \left(x_{i,1} \quad x_{i,2} \quad \dots \quad x_{i,p} \right),$$

де i – номер об’єкта; p – кількість ознак, що описують об’єкт; $x_{i,j}$ – значення j-ї ознаки для i-го об’єкта.

Для об’єкта може бути відомий його справжній образ, тобто вказано

$$X_i, y_i,$$

де y_i – назва образу, до якого він належить.

Такий об’єкт називають **прецедентом**, а їх множину – навчальною вибіркою.

Якщо загальну кількість об’єктів позначити через N, то дані спостереження над ними подають у вигляді матриці $N \times p$, у якій рядки відповідають об’єктам, а стовпці – ознакам. У випадку навчальної вибірки має місце матриця $N \times (p+1)$, у якій останній стовпець містить номери класів.

З геометричного погляду кожен об’єкт являє собою точку в p-вимірному просторі ознак. Сукупність об’єктів утворює у просторі «хмару» або декілька «хмар», форма та розташування яких можуть різнитися. Кожна «хмара» відповідає окремому образу. Якщо p = 2 або 3, можливе візуальне розпізнавання.

Ознаки x_j, $j = \overline{1, p}$, що характеризують об’єкти, можуть бути:

1) кількісні – значення яких можна виміряти, наприклад, температура пацієнта, рівень тиску, сума кредиту тощо;

2) якісні – нечислові за своєю природою (стать (чоловіча, жіноча); рід занять (студент, робітник, менеджер, тренер збірної), сімейний стан (холостий, одружений/заміжня)). Якісні ознаки завжди можна закодувати, приписавши їх можливим значенням певне число, наприклад, чоловіча/жіноча – 0/1.

Під час подальшого викладення, як правило, припускатимемо, що всі ознаки кількісні. Хоча більшість методів придатні для розпізнавання об’єктів і з якісними ознаками, потрібно лише обрати відповідну метрику відстані.

Розв’язання задачі розпізнавання образів у загальному випадку передбачає виконання таких дій:

1) описання об’єктів у формалізованому вигляді, тобто їх подання у вигляді, придатному для застосування в методах розпізнавання;

2) за необхідності скорочення кількості ознак та/або перетворення ознак;

3) власне розпізнавання;

4) оцінка якості розпізнавання.

1. КЛАСИФІКАЦІЯ ДАНИХ

Задача розпізнавання з навчанням, або класифікації, з формального погляду полягає у такому.

Мають місце K класів, які позначатимемо $S_1, S_2, \ldots S_K$. Задано множину прецедентів, тобто об'єктів, для кожного з яких відомо, до якого з цих класів він належить:

$$X = \left\{ X_i, y_i; i = \overline{1, N} \right\} = \left\{ \left(x_{i,1} \quad x_{i,2} \quad \ldots \quad x_{i,p} \right), y_i; i = \overline{1, N} \right\} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} & y_1 \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} & y_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{N,1} & x_{N,2} & \ldots & x_{N,p} & y_N \end{pmatrix},$$

де $X_i = \left(x_{i,1} \quad x_{i,2} \quad \ldots \quad x_{i,p} \right)$ – i-й об'єкт, описаний p ознаками; $x_{i,j}$ – значення j-ї ознаки для i-го об'єкта; y_i – назва класу, до якого належить i-й об'єкт; N – кількість об'єктів; p – кількість ознак.

Цю множину називають **навчальною вибіркою**.

На її основі потрібно побудувати правило, яке б дозволяло будь-який новий об'єкт $X_0 = \left(x_{0,1} \quad x_{0,2} \quad \ldots \quad x_{0,p} \right)$ відносити до одного з класів $S_1, S_2, \ldots S_K$.

Розв'язання задачі здійснюють двома етапами:

1) **навчання** – на цьому етапі будують вирішальне правило; він, як правило, більш складний і трудомісткий; геометрично на етапі навчання будують поверхню, що розділяє класи, хоча аналітичний вигляд цієї поверхні можна визначити не всіма методами;

2) **класифікація** – власне класифікація нового об'єкта на основі побудованого правила.

Серед методів розв'язання задачі класифікації можна виділити: метричні (ближнього сусіда та його модифікації, на основі порівняння з еталоном, потенціальних функцій), опорних векторів, дерева рішень, статистичні (байєсівський класифікатор, логістичну регресію), нейронні мережі, колективи методів (бустінг, беггінг).

В основу багатьох методів класифікації покладено **гіпотезу компактності**, яка стверджує, що класам відповідають компактні множини у просторі ознак. Це означає, що реалізації одного і того самого класу відображуються у просторі ознак у близькі точки, які утворюють «компактні» згустки. Дану гіпотезу не завжди можна підтвердити експериментально. Але найголовніше, що ті задачі, в межах яких вона виконується, усі без винятку мають простий розв'язок. І навпаки, задачі, для яких гіпотезу компактності не можна підтвердити, або зовсім не мають розв'язку, або їх розв'язують зі значними труднощами, із залученням додаткових прийомів. Тому гіпотезу компактності можна розглядати як ознаку можливості успішного розв'язання задачі.

1.1. Метричні методи класифікації

1.1.1. Метрики, або функції, відстаней

Методи, засновані на аналізі близькості (схожості) об’єктів, прийнято називати **метричними**.

У загальному випадку поняття близькості (схожості) об’єктів задають уведенням правила обчислення відстані $d(X_1, X_2)$ між об’єктами, або метрики відстані.

Функцію $d(X_1, X_2)$ називають **метрикою відстані** або **функцією відстані**, якщо вона задовольняє умови:

1) невід’ємності:

$$d(X_1, X_2) \geq 0 \text{ для } \forall X_1, X_2;$$

2) симетрії:

$$d(X_1, X_2) = d(X_2, X_1) \text{ для } \forall X_1, X_2;$$

3) максимальної схожості об’єкта із самим собою:

$$d(X_1, X_2) = 0, \text{ коли } X_1 = X_2;$$

4) нерівності трикутника:

$$d(X_1, X_2) \leq d(X_1, X_3) + d(X_3, X_2) \text{ для } \forall X_1, X_2, X_3.$$

У численних виданнях описано близько 50 різних метрик відстаней, найбільш застосовні з них такі:

1) евклідова відстань, яка являє собою геометричну відстань у просторі ознак, її обчислюєть як

$$d(X_1, X_2) = \sqrt{\sum_{j=1}^{p} \left(x_{1,j} - x_{2,j}\right)^2};$$

2) зважена евклідова відстань, яку визначають за формулою

$$d(X_1, X_2) = \sqrt{\sum_{j=1}^{p} \omega_j \left(x_{1,j} - x_{2,j}\right)^2};$$

де $\omega_j > 0$ – ваговий коефіцієнт j-ї ознаки;

3) манхеттенська відстань (або відстань міських кварталів), вона дорівнює

$$d(X_1, X_2) = \sum_{j=1}^{p} \left| x_{1,j} - x_{2,j} \right|$$

і може бути застосована також у випадку якісних ознак (тоді її називають Хемінговою відстанню);

4) відстань Чебишева, яку визначають як

$$d(X_1, X_2) = \max_{1 \leq j \leq p} \left| x_{1,j} - x_{2,j} \right|.$$

Дану метрику застосовують для визначення двох об’єктів як «різних», навіть коли вони різняться значенням лише однієї ознаки;

5) узагальнена степенева відстань Мінковського

$$d\left(X_1, X_2\right) = \sqrt[m]{\sum_{j=1}^{p}\left|x_{1,j} - x_{2,j}\right|^m},$$

де m – показник степеня; як правило, застосовують значення $m = 1, 2, \infty$, які дають три попередні відстані: за $m = 2$ має місце евклідова відстань, коли $m = 1$ – манхеттенська, за $m = \infty$ – Чебишева;

6) відстань Махаланобіса, що визначають як

$$d\left(X_1, X_2\right) = \sqrt{\left(X_1 - X_2\right)\Sigma^{-1}\left(X_1 - X_2\right)^T},$$

де $\Sigma = \left\{v_{j,v};\, j, v = \overline{1, p}\right\}$ – матриця коваріацій між ознаками, елементи якої обчислюють за всією сукупністю об'єктів $\left\{X_i = \left(x_{i,1} \quad x_{i,2} \quad \ldots \quad x_{i,p}\right); i = \overline{1, N}\right\}$ за формулою

$$\Sigma = \sum_{i=1}^{N}\left(X_i - \overline{X}\right)^T\left(X_i - \overline{X}\right),$$

$$v_{j,v} = \sum_{i=1}^{N}\left(x_{i,j} - \overline{x}_j\right)\left(x_{i,v} - \overline{x}_v\right),\ j, v = \overline{1, p},$$

де $\overline{X} = \left(\overline{x}_1 \quad \overline{x}_2 \quad \ldots \quad \overline{x}_p\right)$ – вектор середніх

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i; \qquad \overline{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{i,j},\ j = \overline{1, p},$$

при цьому якщо кореляції між змінними відсутні, відстань Махаланобіса еквівалентна евклідовій відстані.

Поняттям, протилежним відстані, є **міра схожості** $s\left(X_1, X_2\right)$ між об'єктами X_1 та X_2. Якщо більші значення метрики відстані вказують на меншу схожість об'єктів, то для мір схожості, навпаки, більші значення свідчать про більшу схожість. Зазвичай дуже легко перейти від метрик відстаней до мір схожості й навпаки. Наприклад, міру схожості можна обчислювати так:

$$s\left(X_1, X_2\right) = \exp\left(-d\left(X_1, X_2\right)\right).$$

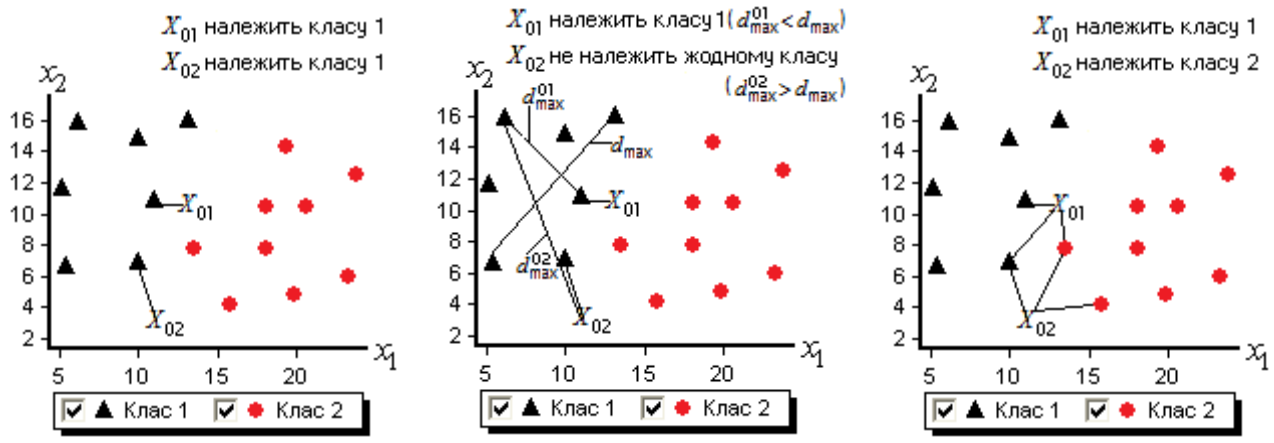
Мірою схожості часто є різноманітні коефіцієнти кореляції (Пірсона, Спірмена, Кендалла).

1.1.2. Метод найближчого сусіда та його модифікації

Метод найближчого сусіда (Nearest Neighbor – NN) – найпростіший метод класифікації. Процес навчання в ньому полягає в запам'ятовуванні всіх об'єктів навчальної вибірки. Класифікують новий об'єкт X_0 за таким правилом (рис. 1.1, *а*): класифікований об'єкт відносять до того класу, представник якого

найближче розташований до X_0, тобто

$$X_0 \in S_l: \qquad X_q \in S_l, \qquad d\left(X_0, X_q\right) = \min_{i=\overline{1,N}} d\left(X_0, X_i\right).$$



а *б* *в*

Рис. 1.1. Робота методів:
а – ближнього сусіда; *б* – модифікації ближнього сусіда; *в* – *k* ближніх сусідів (*k* = 3)

Перевагами методу є простота реалізації та наочна інтерпретація результатів.

Наведемо недоліки даного методу [3]:

1. Метод потребує зберігання усієї навчальної вибірки, що призводить до неефективних витрат пам'яті.

2. Нестійкість до шуму. Якщо серед навчальних об'єктів є викид, тобто об'єкт, розташований серед представників чужого класу, то всі об'єкти, для яких він виявиться найближчим, будуть класифіковані неправильно.

3. Пошук найближчого сусіда передбачає порівняння класифікованого об'єкта з усіма об'єктами навчальної вибірки, здійснюваного за $O(N)$ операцій, тобто потребує багато часу для задач із великими вибірками або високою частотою запитів. Для вирішення цієї проблеми слід застосовувати ефективні алгоритми пошуку найближчих сусідів, наприклад, застосовувати *k*-вимірні дерева (*k-d* tree) для зберігання навчальної вибірки.

4. Відсутні параметри для настройки.

Професор О.П. Приставка запропонував **модифікацію** даного **методу**, яка дозволяє «фільтрувати» викиди, що значно відрізняються від об'єктів навчальної вибірки. У всьому іншому переваги і недоліки модифікації аналогічні класичному методу найближчого сусіда. Процес навчання тут також зведено до запам'ятовування об'єктів навчальної вибірки. Класифікацію нового об'єкта X_0 проводять за таким правилом (рис. 1.1, *б*):

1. Знаходять найближчий до X_0 об'єкт навчальної вибірки. Нехай це об'єкт X_q із класу S_l:

$$X_q \in S_l, \qquad d\left(X_0, X_q\right) = \min_{i=\overline{1,N}} d\left(X_0, X_i\right).$$

2. Обчислюють ширину класу S_l як відстань між найбільш віддаленими об'єктами класу:

$$d_{\max} = \max_{X_i, X_a \in S_l} d\left(X_i, X_a\right).$$

3. Визначають відстань до найвіддаленішого від X_0 об'єкта класу S_l:

$$d^0_{\max} = \max_{X_i \in S_l} d\left(X_0, X_i\right).$$

4. Якщо $d^0_{\max} < d_{\max}$, то об'єкт X_0 відносять до класу S_l, інакше вважають, що X_0 не належить жодному класу.

Більш поширений і застосовуваний на практиці **метод *k* найближчих сусідів** (*k*NN), який стійкіший до шуму, ніж попередні. Його відмінність від методу найближчого сусіда полягає в тому, що на етапі класифікації знаходять не один, а *k* об'єктів, ближчих до X_0 в обраній метриці. Тоді об'єкт X_0 відносять до класу, чиїх представників виявилося більше серед *k* сусідів (рис. 1.1, *в*).

Існує альтернативний варіант методу, суть якого в такому. У кожному класі обирають *k* найближчих до X_0 об'єктів, і X_0 відносять до класу, для якого середня відстань до *k* найближчих сусідів мінімальна [3].

Важливо правильно вибрати параметр *k*. З одного боку, *k* має бути достатньо великий, щоб серед найближчих сусідів виявилися представники різних класів, з іншого – достатньо малий, щоб не стерти нюанси межі, що розділяє класи. На практиці оптимальне значення *k* можна визначити методом ковзного контролю. Для цього за фіксованого значення *k* класифікують кожен об'єкт навчальної вибірки (попередньо об'єкт, який класифікують, вилучають із навчальної вибірки, інакше найближчим сусідом для нього буде він сам). Оптимальним вважають *k*, за якого помилка класифікації найменша.

Під час застосування методу максимальна кількість сусідів може бути на декількох класах одночасно. У задачах із двома класами цього можна уникнути, якщо застосовувати непарне *k*. Більш загальна тактика, придатна і для випадку багатьох класів, передбачає введення ваг w_i, які визначають внесок *i*-го сусіда в класифікацію, наприклад, за правилом

$$w_i = d^{-2}\left(X_0, X_i^{\text{c}}\right),$$

де X_i^{c} – *i*-й сусід об'єкта X_0.

Тоді об'єкт X_0 відносять до класу, сумарний внесок представників якого максимальний:

$$X_0 \in S_l: \qquad \sum_{X_i^{\text{c}} \in S_l} w_i = \max_{h=1,K} \sum_{X_i^{\text{c}} \in S_h} w_i .$$

До переваг методу *k* найближчих сусідів слід віднести простоту реалізації, легку інтерпретацію результатів, а також стійкість до шуму порівняно з методом найближчого сусіда. Їх недоліки подібні: необхідність зберігати всю навчальну вибірку; значні витрати на пошук найближчих сусідів; мінімальні можливості з настроювання алгоритму, оскільки має місце лише один параметр *k*.

1.1.3. Методи на основі порівняння з еталоном

Еталоном називають узагальнене описання класу.

Основна суть методів на основі порівняння з еталоном полягає у побудові еталонів кожного класу на етапі навчання з подальшою класифікацією нових об'єктів за правилом найближчого або *k* найближчих еталонів.

Залежно від способу описання еталонів класів виділяють декілька варіантів методів.

1. **Еталонами обирають усереднені об'єкти класу**, які визначають за формулою

$$\overline{X}^{(l)} = \frac{1}{N_l} \sum_{X_i \in S_l} X_i \,,\; l = \overline{1,K} \,,$$

де N_l – кількість об'єктів у класі S_l.

Тоді новий об'єкт X_0 класифікують за правилом найближчого сусіда:

$$X_0 \in S_l : \qquad d\left(X_0, \overline{X}^{(l)}\right) = \min_{h=\overline{1,K}} d\left(X_0, \overline{X}^{(h)}\right).$$

Недоліком методу можна вважати те, що його застосування призводить до побудови у просторі ознак лінійних роздільних поверхонь.

2. **Еталонами можуть бути один або декілька навчальних об'єктів**, що забезпечують якість класифікації не гіршу, ніж за усією навчальною вибіркою. Задача навчання передбачає вибір еталонних об'єктів, наприклад, методом STOLP або FRiS-STOLP [8]. Таким чином, навчальна вибірка значно скорочується за рахунок видалення з неї неінформативних та шумових об'єктів. Класифікацію нових об'єктів здійснюють за методом найближчого або *k* найближчих сусідів.

Метод STOLP для випадку двох класів працює таким чином. Спочатку знаходять найбільш «напружені» примежові об'єкти. Для цього для кожного об'єкта обчислюють відстань до найближчого об'єкта свого класу (d_{in}) та найближчого об'єкта чужого класу (d_{out}). Відношення $W = d_{\text{in}} / d_{\text{out}}$ характеризує величину ризику для даного об'єкта бути розпізнаним як об'єкт чужого класу. Серед об'єктів кожного класу обирають по одному з максимальним значенням величини W, які заносять до списку еталонів. Далі виконують пробне розпізнавання всіх об'єктів навчальної вибірки за правилом найближчого сусіда за допомогою еталонів. Серед об'єктів, класифікованих неправильно, обирають один із максимальним значенням W, і ним поповнюють список еталонів. Після цього пробне розпізнавання всіх об'єктів навчальної вибірки повторюють. Процедуру продовжують, поки всі навчальні об'єкти не будуть розпізнаватися правильно.

Якщо кількість класів більша двох, описану процедуру застосовують для кожної пари класів. Для прискорення роботи метод рекомендують розпочинати застосовувати для пари двох найближчих класів. Після розв'язання задачі для них обирають наступну найближчу пару класів до тих пір, поки не буде розглянуто всі пари. Якщо під час розгляду поточної пари виявиться, що для одного з класів, який входить до неї, на попередніх кроках вже було сформовано список еталонів,

цей список включають із самого початку і за необхідності поповнюють. Як відстань між класами доцільно застосовувати відстань між двома найближчими об'єктами двох різних класів (відстань Хаусдорфа).

Приклад 1.1. Задано навчальну вибірку, що містить спостереження над 17-ма двовимірними об'єктами двох класів: $\{(1\ \ 5),1;\ (3\ \ 13),1;\ (9\ \ 3),1;\ (10\ \ 10),1;\ (13\ \ 5),1;\ (16\ \ 17),1;\ (21\ \ 16),1;\ (22\ \ 13),1;\ (17\ \ 11),2;\ (19\ \ 1),2;\ (21\ \ 7),2;\ (25\ \ 3),2;\ (26\ \ 9),2;\ (26\ \ 13),2;\ (30\ \ 7),2;\ (33\ \ 11),2;\ (34\ \ 2),2\}$. Знайдемо на її основі еталонні об'єкти за методом STOLP.

Спочатку визначимо найбільш «напружені» об'єкти кожного класу. Для першого класу це об'єкт із координатами $(16\ \ 17)$, для нього $W=0{,}838$, а для другого – $(17\ \ 11)$ із $W=1{,}05$ (рис. 1.2). На наступних кроках список еталонів буде поповнено об'єктами $(10\ \ 10)$, $W=0{,}825$; $(22\ \ 13)$, $W=0{,}79$; $(26\ \ 13)$, $W=1{,}0$, після чого процедуру буде завершено, оскільки знайдені еталони дозволяють безпомилково розпізнавати всі об'єкти навчальної вибірки.

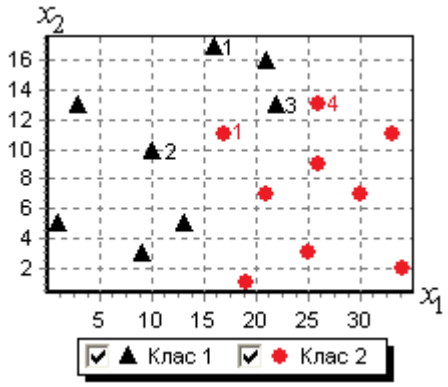


Рис. 1.2. Вибір еталонів за методом STOLP

3. **Еталонами також можуть бути геометричні фігури**, що покривають усі об'єкти свого класу. Прикладом навчального методу є метод еталонів, що подрібнюються (ДРЕТ) [8]. Один із варіантів цього методу передбачає застосування як покривних фігур набору гіперсфер.

Для кожного із K класів будують сферу мінімального радіуса, що накриває усі його навчальні реалізації. Центр сфери l-го класу ($l=\overline{1,K}$) визначають як усереднений об'єкт класу $\bar{X}^{(l)}$, а радіус обчислюють як відстань від центра до найбільш віддаленого об'єкта класу:

$$R_l=\max_{X_i\in S_l} d\left(X_i,\bar{X}^{(l)}\right).$$

Такі сфери вважають еталонними, а їх центри та радіуси запам'ятовують як еталони першого покоління. Область, яку накриває сфера, вважають такою, що належить відповідному класу (сфера 1 на рис. 1.3). Виняток становить випадок, коли сфери класів перетинаються. Тоді область перетину приписують одному з класів за нижченаведеним правилом.

Якщо сфери двох класів перетинаються, але в область перетину не потрапляє жоден об’єкт навчальної вибірки (сфери 2, 3), тоді область перетину належить сфері з меншим радіусом (сфері 2). Якщо в зону перетину потрапили об’єкти лише одного класу (сфери 5, 6), то ця зона належить відповідному класу (сфері 5). Якщо область перетину містить об’єкти різних класів (сфери 4, 5), то за даними об’єктами будують еталони другого покоління (сфери 7, 8). Якщо і вони перетинаються, то для об’єктів із їх зони перетину будують еталони третього покоління. Подрібнення еталонів продовжують до одержання заданої якості розпізнавання навчальної вибірки. Зазвичай у середньому достатньо не більше трьох поколінь еталонів [8].

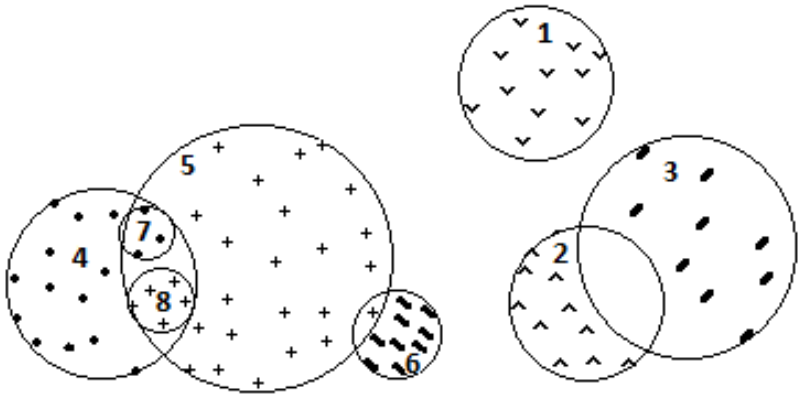


Рис. 1.3. Робота методу ДРЕТ

Правило класифікації нових об’єктів таке: нові об’єкти відносять до класу, у сферу якого вони потрапили. При цьому перевірку слід починати зі сфери з найменшим радіусом. Можлива ситуація, коли об’єкт не потрапляє в жодну гіперсферу. У такому випадку вирішальне правило слід доповнити, наприклад, передбачивши відмову від однозначного віднесення об’єкта до одного з класів або застосування критерію мінімуму відстані до еталонів даного або попереднього покоління.

Інший варіант методу ДРЕТ відрізняється від описаного застосуванням гіперпаралелепіпедів. Якщо їх сторони паралельні координатним осям, то визначення їх положення і перевірку умови потрапляння точок усередину паралелепіпедів або в область їх перетину здійснюють за допомогою дуже простих процедур, що дозволяє економити приблизно 30 % машинного часу порівняно з варіантом, який застосовує гіперсфери [8].

У цілому застосування методів на основі порівняння з еталоном дозволяє [3]:

1. Скоротити обсяг збережуваних даних.
2. Зменшити час класифікації, оскільки найближчих сусідів шукають лише серед еталонів, а не всіх об’єктів навчальної вибірки.
3. Підвищити якість та стійкість класифікації за рахунок того, що з навчальної вибірки вилучають неінформативні та шумові об’єкти і залишають найтиповіших представників класів.
4. Зрозуміти структуру класів за рахунок виділення найтиповіших представників або побудови фігур, що покривають класи.

1.1.4. Метод потенціальних функцій

У 60-х рр. XX ст. М.А. Айзерман, Е.М. Браверман, Л.І. Розоноер запропонували для розв'язання задач навчання розпізнаванню образів метод потенціальних функцій [2]. Метод реалізує ідею рекурентної процедури мінімізації середнього ризику та еквівалентний перцептрону Розенблатта.

Розглянемо функцію двох змінних $\Pi(X,Y)$, де X та Y – точки вихідного простору ознак. Якщо зафіксувати точку Y, поклавши її такою, що дорівнює X^*, то $\Pi(X,X^*)$ перетвориться на функцію однієї змінної X, яка залежить від параметра X^*. Прикладом такої функції у фізиці є потенціал, визначений для будь-якої точки простору, який залежить від того, де розташоване джерело потенціалу. З огляду на цю аналогію $\Pi(X,X^*)$ називають **потенціальною функцією**.

Нехай функція $\Pi(X,X^*)$ усюди додатна і монотонно спадає в разі віддалення X від X^*, досягаючи свого максимального значення за $X=X^*$ (рис. 1.4). Автори ввели ці умови для полегшення геометричної інтерпретації методу та його інтуїтивного розуміння, під час математичного обґрунтування методу їх не використовують.

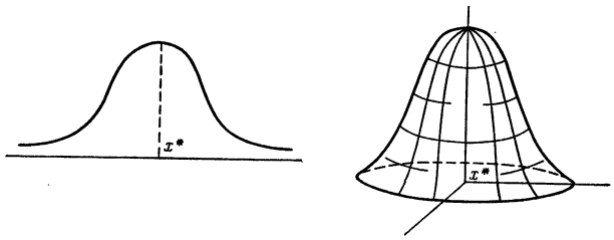


Рис. 1.4. Приклади потенціальних функцій [2, с. 31]

Потенціальну функцію можна задати у вигляді

$$\Pi(X,X^*)=\exp\left(-\alpha d^2(X,X^*)\right) \qquad \text{або} \qquad \Pi(X,X^*)=\frac{1}{1+\alpha d^2(X,X^*)},$$

де $\alpha>0$ – коефіцієнт, значення якого впливає на швидкість спадання потенціальної функції.

Розглянемо спочатку суть методу для випадку двох класів.

Нехай під час навчання подано об'єкт X_1 і вказано, якому класу він належить. Вважатимемо цей об'єкт джерелом потенціалу, поклавши $X^*=X_1$. Побудуємо функцію $\Pi(X,X_1)$ і запам'ятаємо, якому класу вона відповідає (рис. 1.5). Далі для навчання подано об'єкт X_2. Його також вважатимемо джерелом потенціалу, який описує функція $\Pi(X,X_2)$. Побудуємо цю функцію і запам'ятаємо, якому класу вона відповідає. У процесі подання чергового об'єкта X_i із навчальної вибірки будуватимемо потенціали $\Pi(X,X_i)$ і запам'ятовуватимемо, яким класам вони належать.

Після завершення процесу навчання підсумуємо потенціали, побудовані за об’єктами класу S_1 та S_2. Таким чином, побудуємо функції

$$\Pi_1(X)=\sum_{X_i\in S_1}\Pi(X,X_i),\qquad \Pi_2(X)=\sum_{X_i\in S_2}\Pi(X,X_i),$$

що являють собою сумарні потенціали класів S_1 і S_2 відповідно.

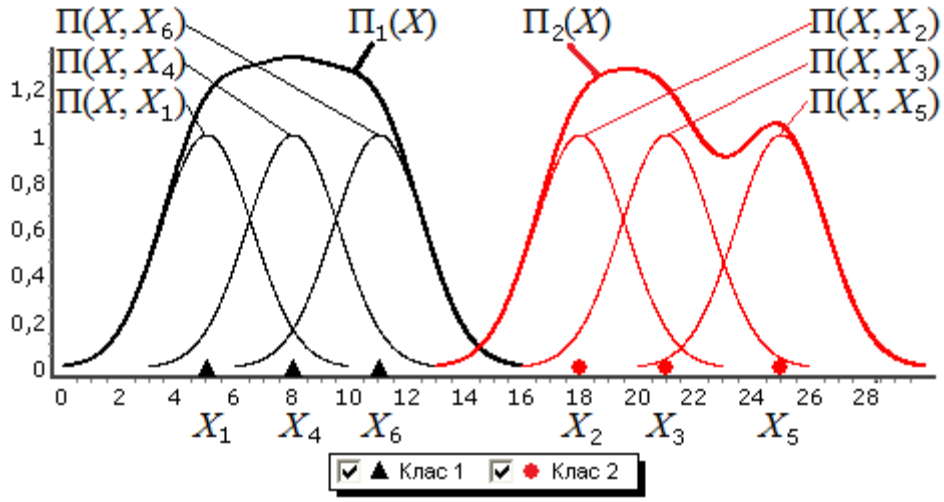


Рис. 1.5. Процес навчання за методом потенціальних функцій (одновимірний випадок)

Під час класифікації нового об’єкта X_0 його відносять до класу, чий сумарний потенціал у точці X_0 виявиться вищий:

$$\begin{aligned} &X_0\in S_1, \qquad \text{якщо } \Pi_1(X_0)>\Pi_2(X_0),\\ &X_0\in S_2, \qquad \text{якщо } \Pi_1(X_0)<\Pi_2(X_0). \end{aligned} \tag{1.1}$$

Рівність сумарних потенціалів $\Pi_1(X_0)=\Pi_2(X_0)$ має місце у випадку, коли об’єкт X_0 лежить на межі класів.

Правило класифікації (1.1) можна переписати інакше, увівши до розгляду **відокремлювальну (дискримінантну) функцію**

$$g(X)=\Pi_1(X)-\Pi_2(X),$$

додатну для об’єктів S_1 і від’ємну для S_2, тобто таку, що знаком розділяє класи. Тоді правило класифікації нового об’єкта X_0 набуває вигляду

$$\begin{aligned} &X_0\in S_1, \qquad \text{якщо } g(X_0)>0,\\ &X_0\in S_2, \qquad \text{якщо } g(X_0)<0. \end{aligned}$$

Рівняння $g(X)=0$ задає поверхню, що розділяє класи.

Слід зауважити, що застосування такої процедури не завжди ефективне. Так, якщо кількість об’єктів класу S_1 суттєво перевищує кількість об’єктів класу S_2, це може призвести до побудови неправильної відокремлювальної поверхні. Але навіть якщо кількість представників кожного класу в навчальній вибірці майже однакова, найбільша щільність об’єктів може бути в тій області кожного класу, що межує із «малонаселеною» областю іншого класу (рис. 1.6). Тоді точки з «малонаселених» областей також класифікуватимуться неправильно.

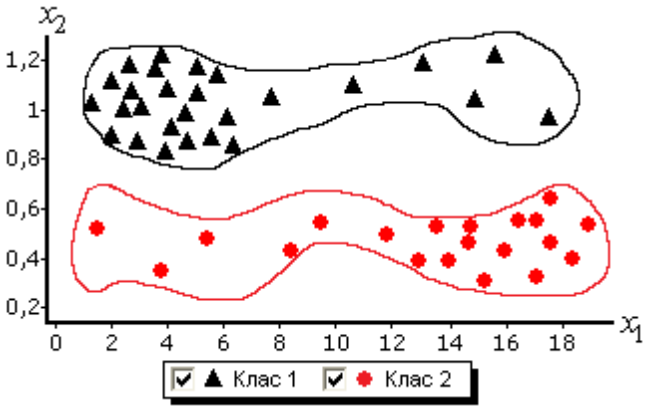


**Рис. 1.6. Приклад навчальної вибірки,
навчання за якою може виявитися неуспішним**

Для усунення цього недоліку процес навчання реалізують дещо інакше. Відокремлювальну функцію будують ітераційно. На кожному кроці подають черговий об’єкт навчальної вибірки і перевіряють, чи правильно він класифікований на основі поточної відокремлювальної функції. Якщо неправильно, то відокремлювальну функцію корегують: до поточного сумарного потенціалу класу, якому належить даний об’єкт, додають породжений ним потенціал. Коли об’єкт класифіковано правильно, він «забувається» і відокремлювальну функцію не змінюють. За результатами такої процедури будують функцію вигляду

$$g(X)=\sum_{\substack{X_i\in S_1,\\ X_i-\text{помилка}}}\Pi(X,X_i)-\sum_{\substack{X_i\in S_2,\\ X_i-\text{помилка}}}\Pi(X,X_i).$$

Формально описану **процедуру навчання за методом потенціальних функцій** зводять до ітераційної побудови відокремлювальної функції $g(X)$ згідно з виразом

$$g^{(i+1)}(X)=g^{(i)}(X)+\alpha_{i+1}\Pi(X,X_{i+1}),$$

де $g^{(i)}(X)$ – i-те наближення функції $g(X)$; X_{i+1} – об’єкт навчальної вибірки, поданий на $(i+1)$-му кроці;

$$\alpha_{i+1}=\begin{cases}1, & \text{якщо } g^{(i)}(X_{i+1})\le 0 \text{ і } X_{i+1}\in S_1,\\ -1, & \text{якщо } g^{(i)}(X_{i+1})\ge 0 \text{ і } X_{i+1}\in S_2,\\ 0 & \text{інакше.}\end{cases}$$

Початкове наближення $g^{(0)}\equiv 0$.

Після подання першого об’єкта X_1 будують функцію $g^{(1)}(X)$

$$g^{(1)}(X)=\begin{cases}\Pi(X,X_1), & \text{якщо } X_1\in S_1,\\ -\Pi(X,X_1), & \text{якщо } X_1\in S_2.\end{cases}$$

На $(i+1)$-му кроці, коли подають об’єкт X_{i+1}, має місце функція $g^{(i)}(X)$.

Відносно X_{i+1} можливі 4 ситуації:

1) $X_{i+1} \in S_1$, але $g^{(i)}(X_{i+1}) \le 0$, тобто виникла помилка; у такому разі беруть $\alpha_{i+1} = +1$, $g^{(i+1)}(X) = g^{(i)}(X) + \Pi(X, X_{i+1})$;

2) $X_{i+1} \in S_2$, але $g^{(i)}(X_{i+1}) \ge 0$, отже, виникла помилка; у даному випадку беруть $\alpha_{i+1} = -1$, $g^{(i+1)}(X) = g^{(i)}(X) - \Pi(X, X_{i+1})$;

3) $X_{i+1} \in S_1$ і $g^{(i)}(X_{i+1}) > 0$ – помилки немає;

4) $X_{i+1} \in S_2$ і $g^{(i)}(X_{i+1}) < 0$ – помилки немає.

У результаті такого навчання будують послідовність відокремлювальних функцій

$$g^{(0)}(X), g^{(1)}(X), g^{(2)}(X), \ldots, g^{(i)}(X), g^{(i+1)}(X), \ldots$$

Автори методу довели її збіжність до однієї з функцій, що розділяє класи, за скінченну кількість кроків (у разі виконання певних умов [2]).

Оскільки процес ітераційний, він має бути доповнений умовою зупинки. Процес навчання можна закінчити, як тільки після чергового виправлення помилки наступні M кроків не привели до нового виправлення. Величина M може:

1) бути числом, оціненим згідно з теоремою 1;

2) залежати від кількості m попередніх виправлень помилок, тобто дорівнювати $M = M_0 + m$; величину M_0 вибирають відповідно до теореми 2.

Теорема 1. Нехай p – імовірність помилки під час екзамену, який проводять після навчання. Тоді для $\forall \varepsilon > 0$ і $\forall \delta > 0$ імовірність події $p < \varepsilon$ буде більша за $1 - \delta$ ($\mathrm{P}\{p < \varepsilon\} > 1 - \delta$), якщо M задовольняє нерівність

$$M > \frac{\ln(\delta / k)}{\ln(1 - \varepsilon)},$$

де k – максимальна можлива кількість виправлень помилки, яка може бути оцінена за теоремою Новікова [2].

Теорема 2. Нехай p – імовірність помилки під час екзамену, який проводять після навчання. Тоді для $\forall \varepsilon > 0$ і $\forall \delta > 0$ імовірність події $p < \varepsilon$ буде більша за $1 - \delta$ ($\mathrm{P}\{p < \varepsilon\} > 1 - \delta$), якщо M_0 задовольняє нерівність

$$M_0 > \frac{\ln(\varepsilon\delta)}{\ln(1 - \varepsilon)}.$$

Доведення теорем можна знайти в роботі [2].

Під час **програмної реалізації описаного процесу навчання** необхідно в пам'яті зберігати координати об'єктів, на яких виправляли помилки (позначимо їх номери через $q_1, q_2, \ldots, q_L$), а також числа $\alpha_1, \alpha_2, \ldots, \alpha_L$, які вказують, до якого класу дійсно належать ці об'єкти ($\alpha_e = +1$, якщо $X_{q_e} \in S_1$, і $\alpha_e = -1$, коли $X_{q_e} \in S_2$). У процесі подання чергового об'єкта X_{i+1} із навчальної вибірки

обчислюють суму

$$g^{(i)}\left(X_{i+1}\right)=\sum_{e=1}^{L}\alpha_e\Pi\left(X_{i+1},X_{q_e}\right).$$

Якщо виявиться, що ($g^{(i)}\left(X_{i+1}\right)>0$ і $X_{i+1}\in S_1$) або ($g^{(i)}\left(X_{i+1}\right)<0$ і $X_{i+1}\in S_2$), то результатами обчислень на цьому кроці й самим об'єктом X_{i+1} нехтують, і розглядають наступний об'єкт навчальної вибірки. Якщо ($g^{(i)}\left(X_{i+1}\right)\le 0$ і $X_{i+1}\in S_1$) або ($g^{(i)}\left(X_{i+1}\right)\ge 0$ і $X_{i+1}\in S_2$), до пам'яті заносять номер чергового помилково класифікованого об'єкта $q_{L+1}=i+1$ і число α_{L+1}.

У ході реалізації даного процесу слід ураховувати, що навчальні вибірки, доступні досліднику, скінченні й навчальні об'єкти можуть закінчитися до моменту виконання умови зупинки. У такому випадку можна, наприклад, почати подавати об'єкти із тієї самої вибірки повторно.

Метод потенціальних функцій може бути легко **узагальнений у разі наявності багатьох класів**. На початку навчання значення сумарних потенціалів усіх класів вважають такими, що дорівнюють нулю:

$$\Pi_l^{(0)}\left(X\right)\equiv 0,\ l=\overline{1,K}.$$

На $(i+1)$-му кроці ітераційного процесу подають об'єкт навчальної вибірки X_{i+1} із класу S_h. Якщо

$$\Pi_h^{(i)}\left(X_{i+1}\right)=\max_{l=\overline{1,K}}\Pi_l^{(i)}\left(X_{i+1}\right),$$

це означає, що даний об'єкт правильно класифікують на основі поточних сумарних потенціалів, тому їх значення не змінюють:

$$\Pi_l^{(i+1)}\left(X\right)=\Pi_l^{(i)}\left(X\right)\ \text{для всіх}\ l=\overline{1,K}.$$

Якщо для об'єкта X_{i+1} із класу S_h максимального значення набуває потенціал v-го класу, тобто

$$\Pi_v^{(i)}\left(X_{i+1}\right)=\max_{l=\overline{1,K}}\Pi_l^{(i)}\left(X_{i+1}\right),\ v\ne h,$$

тоді потенціали класів S_h і S_v корегують у такий спосіб:

$$\Pi_h^{(i+1)}\left(X\right)=\Pi_h^{(i)}\left(X\right)+\Pi\left(X,X_{i+1}\right),$$
$$\Pi_v^{(i+1)}\left(X\right)=\Pi_v^{(i)}\left(X\right)-\Pi\left(X,X_{i+1}\right),$$
$$\Pi_l^{(i+1)}\left(X\right)=\Pi_l^{(i)}\left(X\right),\ l=\overline{1,K},\ l\ne h,v.$$

Під час екзамену новий об'єкт X_0 відносять до того класу, чий сумарний потенціал вищий:

$$X_0\in S_h:\qquad \Pi_h\left(X_0\right)=\max_{l=\overline{1,K}}\Pi_l\left(X_0\right).$$

Важливим моментом у ході застосування методу є вибір виду потенціальної функції $\Pi\left(X,X^*\right)$. Наприклад, якщо вона дуже швидко спадає зі збільшенням відстані, то можна одержати безпомилкову класифікацію навчальної вибірки й водночас незадовільну якість класифікації на нових об'єктах (ефект перенавчання

– див. підрозд. 1.4). У разі занадто «плоскої» потенціальної функції може збільшитися кількість помилок розпізнавання, у тому числі й на навчальних об’єктах. Певні рекомендації з цього приводу можна одержати, якщо розглядати метод зі статистичних позицій (відновлення щільності розподілу імовірності або відокремлювальної межі за вибіркою із застосуванням процедури типу стохастичної апроксимації). Питання вибору потенціальної функції автори методу розглядають у монографії [2].

Переваги методу потенціальних функцій:

1. Ефективний, якщо навчальні об’єкти надходять потоком і зберігати їх у пам’яті немає можливості або необхідності.

2. Добре адаптований для застосування в разі незбалансованих класів.

3. Еквівалентний одній із перших нейронних мереж – перцептрону Розенблатта; одержані для нього теоретичні обґрунтування можуть бути використані під час дослідження нейронних мереж.

Недоліки методу:

1. Повільна збіжність; потребує досить значних за обсягом вибірок.

2. Результат навчання залежить від черговості надходження об’єктів.

3. Якщо розглядати метод як різновид нейронної мережі, то сучасні нейронні мережі мають значно більше можливостей із налаштування та здатні забезпечити вищу якість розпізнавання.

1.2. Метод на основі дерева рішень

Дерево рішень (decision tree) – це правило класифікації у вигляді дерева (рис. 1.7). Внутрішні вузли дерева називають вузлами перевірки, оскільки в них перевіряють умову щодо значення деякої ознаки. Листи дерева містять рішення у вигляді назв класів.

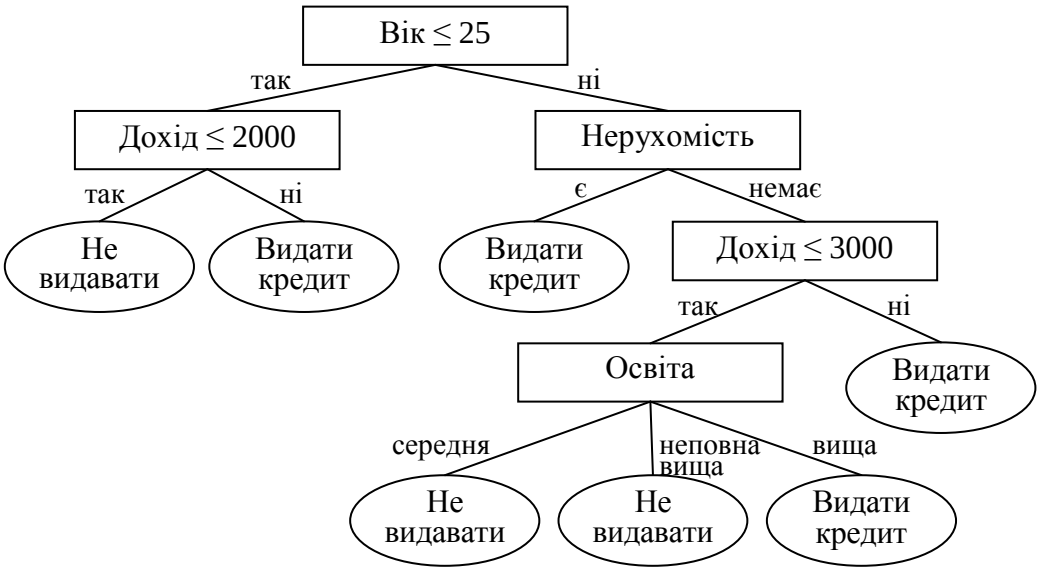


Рис. 1.7. Приклад дерева рішень

На етапі навчання відбувається побудова дерева рішень за навчальною вибіркою. Щоб класифікувати новий об’єкт на основі побудованого дерева, проходять шлях від кореня до листа. На кожному внутрішньому вузлі перевіряють значення певної ознаки об’єкта. Залежно від отриманої відповіді знаходять відповідне розгалуження, за яким рухаються до вузла, що знаходиться на рівень нижче. Процедуру продовжують до моменту виявлення листа, який і містить назву класу об’єкта.

Для побудови дерева рішень розроблено різні алгоритми, серед яких можна виділити ID3, C4.5, CART тощо. В їх основі покладено таку загальну ідею [14].

Процес побудови дерева відбувається зверху вниз. Спочатку створюють корінь дерева. Для цього серед усіх ознак шляхом перебору обирають одну, найкращу за деяким критерієм. На основі обраної ознаки формують умову для кореня, яка розбиває множину об’єктів, що з ним асоціюється (для кореня це вся навчальна вибірка), на декілька підмножин, і відповідно породжує нові вузли дерева (рис. 1.8). До кожного з породжених вузлів рекурсивно застосовують аналогічну процедуру. Якщо з вузлом асоціюється множина об’єктів, що належать одному класу, то вузол перетворюють на лист, і його побудову завершують. Рішенням листа стає клас, об’єкти якого з ним асоціюються. Вузол також перетворюють на лист, коли з ним асоціюється пуста множина об’єктів. У такому разі рішенням вузла стає клас, представників якого більше у вузла-предка.

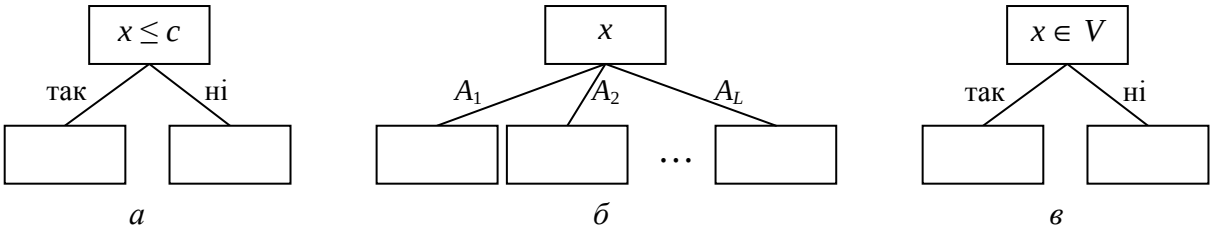


Рис. 1.8. Приклади умов перевірки у вузлі за ознакою *x*:
а – кількісною; *б–в* – якісною

Відмінності в алгоритмах побудови дерева рішень полягають, головним чином, у критерії вибору ознаки, вигляді умови перевірки у вузлі, можливості врахування пропущених значень ознак, наявності механізму відсікання гілок.

Дуже часто алгоритми будують досить складні дерева, які мають велику кількість вузлів та гілок. Їх називають гіллястими і вони небажані. По-перше, у таких деревах вершини, розташовані далеко від кореня, асоціюються з незначною кількістю об’єктів навчальної вибірки, а отже, правила в цих вершинах ненадійні. По-друге, такі дерева досить складні для інтерпретації.

Щоб одержати дерево з меншою кількістю вузлів, більш просте та наочне, застосовують одну з технік:

1. Рання зупинка (prepruning). Уводять критерій, який визначає доцільність побудови вузла. У разі виконання критерію вузол перетворюють на лист і далі не будують, рішенням листа стає клас, представників якого більше у цьому вузлі. Критерієм зупинки може бути досягнення заданої глибини дерева, кількості об’єктів у вузлі, величини ентропії у вузлі тощо. Застосування ранньої зупинки може скоротити час побудови дерева, але водночас знизити якість класифікації.

2. Відсікання гілок (pruning). Дана техніка ефективніша за попередню. Відсікання здійснюють після того, як дерево повністю побудовано. У деяких алгоритмах, наприклад CART, передбачений власний механізм відсікання гілок. У разі його відсутності можна застосувати таку процедуру. Починаючи з вузлів найнижчого рівня, кожен вузол перетворюють на лист, якщо це не призводить до суттєвого збільшення помилки класифікації. Цю процедуру виконують, поки існують вузли, які можна перетворити на листя.

Застосування дерев рішень дозволяє:

1. Будувати правила класифікації на природній мові, за допомогою термінів, зрозумілих спеціалістам предметної галузі.

2. Одночасно враховувати кількісні та якісні ознаки в правилі.

3. Відбирати інформативні ознаки в ході побудови дерева, що дозволяє не застосовувати спеціальні методи (розділ 3).

4. Будувати правила за даними з пропусками.

Недоліки дерев рішень:

1. Усі алгоритми побудови дерев рішень жадібні, а отже, побудовані дерева часто неоптимальні щодо якості класифікації та розміру.

2. Алгоритми побудови дерев рішень можуть ускладнювати структуру дерева і будувати досить гіллясті дерева. Через це правила в деяких вершинах і самі дерева стають менш надійні. Також це може призвести до ефекту перенавчання (підрозд. 1.4). Крім того, нівелюється така перевага, як наочність і простота інтерпретації.

1.2.1. Алгоритм C4.5

Алгоритм C4.5 розроблений Дж. Р. Квінланом. Він – удосконалена версія алгоритму ID3 того ж автора. У ньому додано можливість будувати дерево рішень за вибірками, що містять і якісні, і кількісні ознаки, а також працювати з даними з пропусками.

Алгоритм передбачає побудову дерева за описаною вище схемою. Окремого розгляду потребує вигляд умови перевірки у вузлі та критерій вибору ознаки.

Якщо ознака якісна й набуває L значень $A_1, A_2, \ldots, A_L$, перевірка у вузлі зводиться до встановлення того, якого саме значення набуває ознака. Така перевірка породжує для даного вузла стільки нащадків, скільки може бути значень в ознаки (рис. 1.8, *б*).

За кількісною ознакою у вузлі завжди формують умову вигляду «$x \le c$», де c – деякий поріг. Така умова розбиває множину об’єктів, що асоціюється з вузлом, на дві підмножини: одну утворюють об’єкти, для яких умова виконується, іншу – ті, для яких відповідно не виконується (рис. 1.8, *а*).

Формування умови за кількісною ознакою потребує вибору порогу c, який здійснюють таким чином. Нехай ознака набуває L значень (розглядають лише значення в об’єктів, що асоціюються з вузлом), їх сортують за зростанням і

позначають $A_1, A_2, \ldots, A_L$. Порогом може бути будь-яке значення між двома сусідніми A_t та A_{t+1}, $t = \overline{1, L-1}$, наприклад середнє арифметичне між ними:

$$c_t = 0{,}5\left(A_t + A_{t+1}\right),\ t = \overline{1, L-1}.$$

Серед усіх можливих порогів c_t, $t = \overline{1, L-1}$ шляхом перебору обирають найкращий за критерієм, який застосовують для вибору ознаки.

В алгоритмі C4.5 для вибору ознаки застосовують ентропійний критерій. Нехай із поточним вузлом дерева асоціюється множина об'єктів *U*. На основі ознаки *x* цю множину розбивають на *L* підмножин $U_1, U_2, \ldots, U_L$. Тоді критерій вибору ознаки ґрунтуватиметься на функції

$$\mathrm{Gain}(x) = \mathrm{Info}(U) - \mathrm{Info}_x(U), \tag{1.2}$$

де $\mathrm{Info}(U)$ – ентропія множини *U*

$$\mathrm{Info}(U) = -\sum_{l=1}^{K} \frac{\mathrm{freq}(S_l, U)}{|U|} \log_2\left(\frac{\mathrm{freq}(S_l, U)}{|U|}\right),$$

$|U|$ – потужність множини *U*, тобто кількість об'єктів, що входять до неї; $\mathrm{freq}(S_l, U)$ – кількість об'єктів класу S_l, які входять до множини *U*;

$$\mathrm{Info}_x(U) = \sum_{t=1}^{L} \frac{|U_t|}{|U|} \mathrm{Info}(U_t).$$

Перевагу надають ознаці, для якої $\mathrm{Gain}(x)$ максимальна.

У критерію, заснованого на функції (1.2), один недолік. Серед якісних ознак він надає перевагу тій, яка має більшу кількість значень, що ускладнює структуру дерева. Для усунення недоліку можна застосувати модифікований критерій, заснований на функції

$$\mathrm{GainModify}(x) = \frac{\mathrm{Gain}(x)}{\mathrm{Split}(x)}, \tag{1.3}$$

де $\mathrm{Split}(x)$ являє собою оцінку потенційної інформації, одержуваної в разі розбиття множини *U* на *L* підмножин

$$\mathrm{Split}(x) = -\sum_{t=1}^{L} \frac{|U_t|}{|U|} \log_2\left(\frac{|U_t|}{|U|}\right).$$

Приклад 1.2. Задано навчальну вибірку, за якою потрібно побудувати дерево рішень за алгоритмом C4.5: $\{(3\ \ 159), 1;\ (2\ \ 173), 1;\ (1\ \ 141), 2;\ (2\ \ 151), 1;\ (1\ \ 152), 2;\ (3\ \ 177), 2\}$. Ознаку обиратимемо за функцією (1.2).

Спочатку побудуємо корінь дерева. Із ним асоціюються усі 6 об'єктів навчальної вибірки. Значення функції критерію (1.2) для кожної з ознак такі: $\mathrm{Gain}(x_1) = 0{,}667$, $\mathrm{Gain}(x_2) = 0{,}191$ (у разі порогів 146 і 171), 0 (у випадку порогів 151,5 і 166) та 0,082 (у разі порогу 155,5). Отже, побудуємо корінь за ознакою x_1, умова перевірка за нею породжує три вузли-нащадки. Два з них перетворюються на листи, а третій потрібно побудувати.

Із вузлом, який слід побудувати, асоціюються два об’єкти навчальної вибірки (1-й і 6-й). На їх основі обчислимо значення функції критерію для кожної ознаки: $\mathrm{Gain}(x_1)=0$, $\mathrm{Gain}(x_2)=1$ (за єдиного можливого порогу 168). Отже, правило у вузлі визначають за ознакою x_2 і воно породжує два листи.

Побудоване дерево зображено нижче (рис. 1.9).

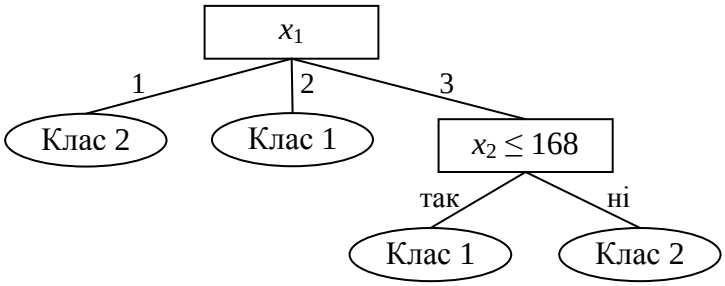


Рис. 1.9. Приклад побудованого дерева рішень

Якщо значення якоїсь ознаки в деяких об’єктів відсутні, алгоритм побудови дерева рішень залишається незмінний. Лише під час застосування критерію (1.2) або (1.3) за такою ознакою потрібно ігнорувати об’єкти, у яких її значення відсутні.

У разі класифікації нового об’єкта з пропущеними значеннями деяких ознак виконують такі дії. Якщо на якомусь вузлі дерева з’ясовують, що значення ознаки відсутнє, тоді досліджують усі можливі шляхи вниз по дереву і визначають найімовірніший клас. Його і обирають як відповідь дерева.

1.2.2. Алгоритм CART

Алгоритм CART (Classification And Regression Tree) розробляли із 1974 по1984 р. Л. Брейман, Дж. Фрідман, Ч. Стоун та Р. Олсен. Основні відмінності від алгоритму C4.5 полягають у тому, що будують бінарне дерево, застосовують інший критерій вибору ознаки та оригінальний метод відсікання гілок.

Для побудови бінарного дерева в алгоритмі передбачене відповідне формування умови перевірки у вузлі.

Для кількісної ознаки умову формують так само, як і в алгоритмі C4.5, у вигляді «$x \le c$». Така умова завжди забезпечує бінарне розбиття вузла.

У разі якісної ознаки, яка набуває L значень $A_1, A_2, \ldots, A_L$, формують умову вигляду «$x \in V$», де V – непуста підмножина значень ознаки (рис. 1.8, *в*). V обирають серед підмножин множини $\{A_1, A_2, \ldots, A_L\}$ шляхом перебирання на основі критерію, застосовного для вибору ознаки.

В основі критерію вибору ознаки в алгоритмі CART покладено індекс Gini. Для множини *U*, що містить об’єкти K класів, індекс Gini визначають як

$$\mathrm{Gini}(U) = 1 - \sum_{l=1}^{K} p_l^2 ,$$

де p_l – відносна частота об’єктів l-го класу в множині *U*.

Нехай із поточним вузлом дерева асоціюється множина об'єктів U. На основі ознаки x цю множину розбивають на 2 підмножини U_{Left} і U_{Right}. Тоді функція критерію матиме вигляд

$$\text{GiniSplit}(x) = \frac{|U_{\text{Left}}|}{N}\text{Gini}(U_{\text{Left}}) + \frac{|U_{\text{Right}}|}{N}\text{Gini}(U_{\text{Right}}). \quad (1.4)$$

Найкращою вважають ознаку, за якої (1.4) набуває мінімального значення.

Якщо позначити через nl_l та nr_l кількість об'єктів l-го класу, що входять до підмножин U_{Left} і U_{Right} відповідно, функцію критерію можна спростити і подати у вигляді

$$\text{GiniSplit}(x) = \frac{1}{|U_{\text{Left}}|}\sum_{l=1}^{K} nl_l^2 + \frac{1}{|U_{\text{Right}}|}\sum_{l=1}^{K} nr_l^2 . \quad (1.5)$$

Тоді перевагу надають ознаці, для якої функція (1.5) максимальна.

Важлива відмінність алгоритму CART полягає у методі відсікання гілок, який називають minimal cost-complexity tree pruning і розглядає відсікання як компроміс між одержанням дерева оптимального розміру і з мінімальною помилкою класифікації.

Перш ніж розглянути метод, уведемо такі позначення:

T – деяке дерево;

$|T|$ – кількість листя дерева;

$R(T)$ – помилка класифікації дерева, яку визначають як відносну кількість помилково класифікованих навчальних об'єктів на усьому листі дерева;

$C(T,\alpha) = R(T) + \alpha|T|$ – вартість дерева, де $\alpha \in [0;+\infty]$ – параметр;

$T_{\max}$ – максимальне за розміром дерево, яке потрібно обрізати.

Для будь-якого фіксованого значення параметра α існує найменше мінімізоване піддерево T_α дерева $T_{\max}$, таке що $C(T_\alpha,\alpha) = \min\limits_{T \le T_{\max}} C(T,\alpha)$, а у разі існування декількох піддерев однакової вартості T_α буде найменше з них.

Кожному $\alpha \in [0;+\infty]$ відповідає своє найменше мінімізоване піддерево дерева $T_{\max}$. Хоча α набуває нескінченну кількість значень, у $T_{\max}$ існує скінченна кількість піддерев, із яких можна побудувати таку послідовність:

$$\underset{[0;\alpha_1)}{T_1} > \underset{[\alpha_1;\alpha_2)}{T_2} > \underset{[\alpha_2;\alpha_3)}{T_3} > \ldots > \underset{[\alpha_m;+\infty)}{\{\text{корінь } T_{\max}\}}. \quad (1.6)$$

Перше дерево цієї послідовності $T_1 = T_{\alpha=0}$ – найменше піддерево $T_{\max}$, помилка класифікації якого така сама, що і у $T_{\max}$. Кожне наступне піддерево – відсікання попереднього. Останнє дерево кореневе, воно містить лише корінь $T_{\max}$.

Метод відсікання гілок передбачає побудову послідовності (1.6) із подальшим вибором із неї дерева, яке забезпечує найкращу якість класифікації на об'єктах контрольної вибірки (підрозд. 1.4).

Послідовності будують у такий спосіб.

На першому кроці знаходять дерево T_1. При цьому, якщо $T_{\max}$ було побудоване без застосування ранньої зупинки, то $T_1 = T_{\max}$. Коли ранню зупинку застосовували, щоб одержати T_1, вузли в дереві $T_{\max}$, починаючи з самого нижнього рівня, намагаються перетворити на листя, не змінивши при цьому помилку класифікації на навчальній вибірці.

На наступному кроці будують піддерево T_2. Для цього в кожному вузлі t дерева T_1 обчислюють величину

$$\alpha(t) = \frac{R(t) - R(T_{1,t})}{|T_{1,t}| - 1},$$

де $T_{1,t}$ – піддерево T_1 із коренем у вузлі t; $R(t)$ – помилка класифікації вузла t, яка дорівнює відносній частоті об'єктів навчальної вибірки, які неправильно класифікує вузол; $R(T_{1,t})$ – помилка класифікації дерева, що дорівнює сумарній помилці класифікації на усьому листі дерева.

У вузлах, у яких $\alpha(t)$ набуває найменшого значення, проводять відсікання. Таким чином формують наступне дерево послідовності T_2, до якого застосовують таку саму процедуру.

Процедуру продовжують до моменту одержання кореневого дерева.

1.3. Метод опорних векторів

Метод опорних векторів (Support Vector Machines, SVM) – популярний метод побудови лінійного класифікатора [15].

Якщо два класи розділяє гіперплощина

$$wX^T + w_0 = 0, \tag{1.7}$$

де $w = (w_1 \;\; w_2 \;\; \ldots \;\; w_p)$ та w_0 – параметри гіперплощини, то правило класифікації нового об'єкта X_0 має вигляд

$$X_0 \in S_1, \qquad \text{якщо } wX_0^T + w_0 > 0,$$
$$X_0 \in S_2, \qquad \text{якщо } wX_0^T + w_0 < 0$$

і його називають **лінійним класифікатором**.

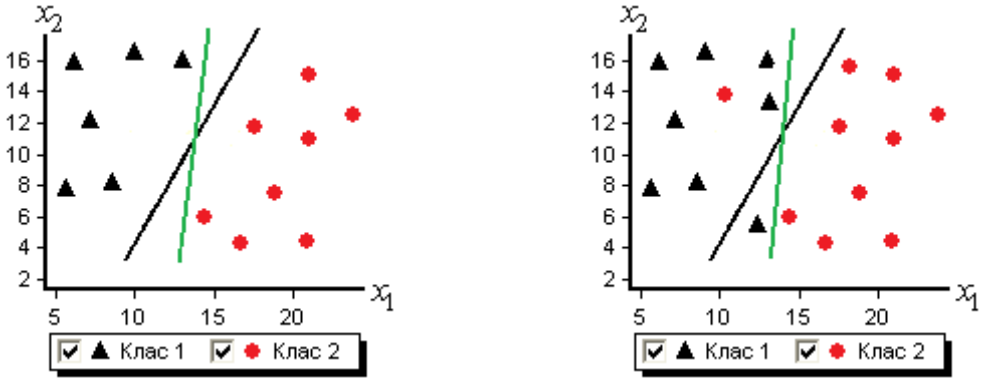
Рівність $wX_0^T + w_0 = 0$ має місце, коли об'єкт X_0 лежить на гіперплощині.

Побудувати за навчальною вибіркою гіперплощину (1.7) і на її основі лінійний класифікатор можна за допомогою методу опорних векторів.

Метод розроблено для випадку двох класів. Для задачі з багатьма класами можливо будувати відокремлювальні гіперплощини для кожної пари.

Для спрощення викладок припускають, що класи позначені як «+1» та «–1», тобто задано навчальну вибірку $\{X_i, y_i; i = \overline{1, N}\}$, де $y_i = +1$ або –1.

Розглянемо спочатку більш простий та наочний **випадок лінійно відокремлювальних класів**, тобто випадок, коли існує гіперплощина, що усі об'єкти навчальної вибірки з одного класу лежать з одного боку від неї, а об'єкти з іншого класу – з іншого (рис. 1.10).



а *б*

Рис. 1.10. Випадки:
а – лінійно відокремлювальних класів; *б* – лінійно невідокремлювальних класів

Нехай має місце відокремлювальна гіперплощина $wX^T + w_0 = 0$. Знайдемо найближчі до неї об'єкти кожного класу і проведемо через них гіперплощини, паралельні відокремлювальній. Вони утворять коридор між класами (рис. 1.11, *а*). Ширина коридору дорівнює

$$\frac{2}{\|w\|},$$

де $\|w\| = \sqrt{w \cdot w^T}$ – евклідова норма вектора w.

При цьому всі об'єкти навчальної вибірки лежать або за межами коридору, або на його межах, тобто для них правдиві нерівності

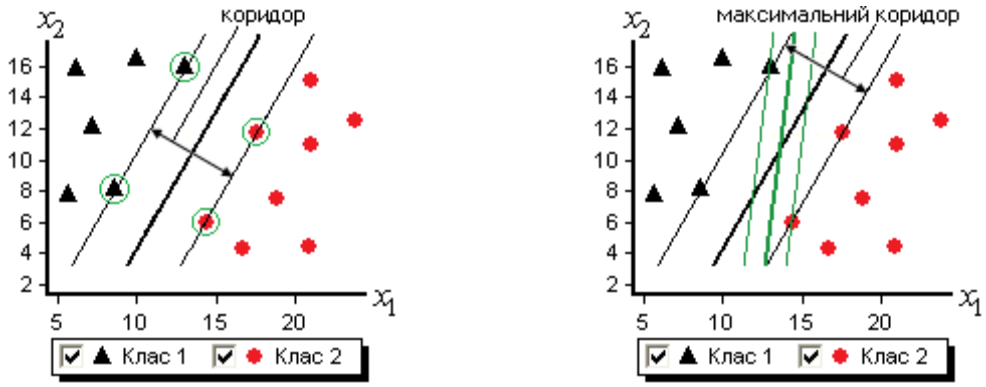
$$\begin{cases} wX_i^T + w_0 \geq 1, & \text{якщо } y_i = +1, \\ wX_i^T + w_0 \leq -1, & \text{якщо } y_i = -1, \end{cases}$$

або інакше

$$y_i\left(wX_i^T + w_0\right) \geq 1,\ i = \overline{1, N}.$$

Рівності в наведених виразах мають місце для об'єктів, що лежать на межах коридору. Саме такі об'єкти називають **опорними векторами** (на рис. 1.11 вони обведені кружечками).

Різним відокремлювальним гіперплощинам відповідають різні за шириною коридори (рис. 1.11, *б*). Гіперплощину, якій відповідає коридор максимальної ширини (інакше кажучи, мінімальне значення величини $\|w\|$, або, що аналогічно, $\|w\|^2$), вважають оптимальною. Її і намагаються побудувати за методом опорних векторів.



а *б*

Рис. 1.11. Ілюстрація:
а – коридору між класами; *б* – максимального коридору між класами

Формально задача побудови оптимальної гіперплощини полягає в розв'язанні такої оптимізаційної задачі: знайти w і w_0, які мінімізують функцію

$$\Phi(w)=\frac{1}{2}\|w\|^2 \tag{1.8}$$

у разі виконання умов

$$y_i\left(wX_i^T+w_0\right)\geq 1,\ i=\overline{1,N}. \tag{1.9}$$

Цю задачу називають прямою задачею квадратичної оптимізації.

Цільова функція (1.8) квадратична, опукла, а обмеження (1.9) лінійні за шуканими параметрами. Тому для розв'язання задачі можна застосувати метод множників Лагранжа, згідно з якого вводять N (за кількістю об'єктів навчальної вибірки) множників Лагранжа $\lambda_i\geq 0$, $i=\overline{1,N}$, які утворюють вектор $\lambda=(\lambda_1\ \ \lambda_2\ \ \ldots\ \ \lambda_N)$. Застосування методу передбачає пошук сідлової точки функції Лагранжа

$$L(w,w_0,\lambda)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\lambda_i\left(y_i\left(wX_i^T+w_0\right)-1\right) \tag{1.10}$$

у випадку виконання умов Каруша–Куна–Таккера

$$\lambda_i=0 \qquad \text{або} \qquad y_i\left(wX_i^T+w_0\right)=1,\ i=\overline{1,N}. \tag{1.11}$$

Пошук сідлової точки полягає в знаходженні таких w і w_0, що надають мінімуму функції (1.10), та вектора λ, який максимізує функцію (1.10).

Із необхідної умови мінімуму функції Лагранжа за величинами w та w_0, тобто рівності нулю відповідних частинних похідних, випливають два співвідношення

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{N}\lambda_i y_i X_i=0 \quad\Rightarrow\quad w=\sum_{i=1}^{N}\lambda_i y_i X_i, \tag{1.12}$$

$$\frac{\partial L}{\partial w_0}=-\sum_{i=1}^{N}\lambda_i y_i=0 \quad\Rightarrow\quad \sum_{i=1}^{N}\lambda_i y_i=0. \tag{1.13}$$

Підставивши вирази (1.12) і (1.13) у функцію (1.10), можна одержати двоїсту задачу квадратичної оптимізації: необхідно знайти вектор множників Лагранжа λ, який надає максимуму функції

$$L(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{q=1}^{N} \lambda_i \lambda_q y_i y_q X_i X_q^T \qquad (1.14)$$

за умов

$$\lambda_i \geq 0,\ i = \overline{1, N}; \qquad \sum_{i=1}^{N} \lambda_i y_i = 0. \qquad (1.15)$$

Остання задача має єдиний розв'язок, оскільки функція (1.14) та область, визначена нерівностями (1.15), опуклі. Її розв'язують за допомогою методів квадратичної оптимізації.

Отже, побудову оптимальної гіперплощини у випадку лінійно відокремлювальних класів здійснюють так:

1. Знаходять оптимальні значення множників Лагранжа λ_i^*, $i = \overline{1, N}$, які надають максимуму функції (1.14) за виконання умов (1.15). Слід зауважити, що частина визначених множників Лагранжа дорівнюватиме нулю. Відмінними від нуля, з огляду на правдивість умов Каруша–Куна–Таккера (1.11), будуть лише множники Лагранжа, які відповідають опорним векторам.

2. Обчислюють оптимальний вектор w^* за виразом (1.12):

$$w^* = \sum_{i=1}^{N} \lambda_i^* y_i X_i .$$

Фактично вектор w^* являє собою лінійну комбінацію лише об'єктів навчальної вибірки, для яких $\lambda_i^* \neq 0$, тобто опорних векторів.

3. Розраховують, відповідно до умови (1.11), оптимальне значення w_0^* на основі будь-якого опорного вектора за виразом

$$w_0^* = y_{\text{опорний}} - w^* X_{\text{опорний}}^T .$$

Для підвищення стійкості рекомендовано знаходити w_0^* за усіма опорними векторами і як розв'язок обирати їх середнє арифметичне.

У реальних задачах лінійно відокремлювальні класи зустрічаються рідко. Тому перейдемо до реалістичнішого **випадку лінійно невідокремлювальних класів.**

У даному разі неможливо побудувати відокремлювальну гіперплощину, повністю виключивши помилки класифікації. Тому допустимо їх наявність, тобто можливість ситуацій, коли об'єкт лежить:

1) не з того боку гіперплощини, де інші представники його класу;

2) із боку свого класу, але потрапляє в коридор (рис. 1.12).

В обох випадках об'єкти вважатимемо порушниками.

Поставимо у відповідність кожному об'єкту навчальної вибірки X_i, $i = \overline{1, N}$ змінну $\xi_i \geq 0$, $i = \overline{1, N}$, яка характеризує ступінь порушення ним умов

«правильного розташування», інакше кажучи, величину помилки на цьому об'єкті.

Змінна ξ_i може набувати таких значень (рис. 1.12):

1) $\xi_i = 0$ для об'єктів, що лежать із правильного боку від гіперплощини, у тому числі й для опорних векторів;

2) $0 < \xi_i < 1$ для об'єктів, які також лежать із правильного боку, але потрапляють у коридор;

3) $\xi_i = 1$ для об'єктів, що лежать на гіперплощині;

4) $\xi_i > 1$ для об'єктів, які лежать не з того боку від гіперплощини, що відповідає їх класу.

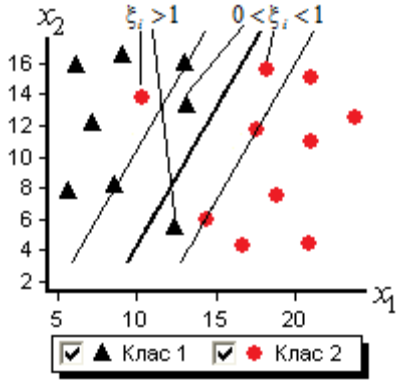


Рис. 1.12. Об'єкти-порушники

Помилка класифікації на всіх об'єктах навчальної вибірки становить $\sum_{i=1}^{N} \xi_i$.

Ураховуючи наявність об'єктів-порушників, потрібно пом'якшити умови, які мають задовольняти об'єкти навчальної вибірки:

$$y_i \left(w X_i^T + w_0 \right) \geq 1 - \xi_i, \; i = \overline{1, N}.$$

Зрозуміло, що чим більший коридор між класами, тим більша кількість об'єктів-порушників. Тому необхідно знайти певний компроміс між максимальним коридором та мінімальною помилкою класифікації.

Формально задача побудови оптимальної гіперплощини для лінійно невідокремлювальних класів полягає у розв'язанні такої задачі квадратичної оптимізації: знайти w і w_0, які мінімізують функцію

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \qquad (1.16)$$

за виконання умов

$$\xi_i \geq 0, \; i = \overline{1, N};$$

$$y_i \left(w X_i^T + w_0 \right) \geq 1 - \xi_i, \; i = \overline{1, N},$$

де $C > 0$ – параметр, що дозволяє регулювати відношення між максимізацією ширини зазору та мінімізацією сумарної помилки класифікації.

Параметр C можна обрати експериментально на основі, наприклад, ковзного контролю (підрозд. 1.4).

Для розв’язання цієї задачі також застосовують метод множників Лагранжа. У даному випадку його застосування дозволяє одержати таку двоїсту задачу: знайти вектор множників Лагранжа λ, який надає максимуму функції

$$L(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{q=1}^{N} \lambda_i \lambda_q y_i y_q X_i X_q^T, \qquad (1.17)$$

за умов

$$0 \le \lambda_i \le C,\ i = \overline{1, N}; \qquad \sum_{i=1}^{N} \lambda_i y_i = 0. \qquad (1.18)$$

Відмінність від випадку лінійно відокремлювальних класів полягає лише у вигляді обмежень на множники Лагранжа. У випадку лінійно відокремлювальних класів необхідна їх невід’ємність, а в даному разі вони додатково мають бути обмежені зверху константою *C*.

У цілому побудову оптимальної гіперплощини для лінійно невідокремлювальних класів здійснюють за аналогією з попереднім випадком:

1. Вираховують оптимальні значення множників Лагранжа λ_i^*, $i = \overline{1, N}$, які надають максимуму функції (1.17) за виконання умов (1.18). Але відмінними від нуля будуть множники Лагранжа, які відповідають опорним векторам та об’єктам-порушникам.

2. Обчислюють оптимальний вектор w^* згідно з виразом

$$w^* = \sum_{i=1}^{N} \lambda_i^* y_i X_i.$$

Вектор w^* являє собою лінійну комбінацію об’єктів навчальної вибірки, для яких $\lambda_i^* \ne 0$, тобто опорних та порушників.

3. Розраховують оптимальне значення w_0^* на основі будь-якого опорного вектора за виразом

$$w_0^* = y_{\text{опорний}} - w^* X_{\text{опорний}}^T$$

або за усіма опорними векторами з подальшим усередненням.

Суттєвою перевагою методу опорних векторів є те, що на його основі можна **побудувати нелінійну відокремлювальну поверхню**. Ідея її побудови полягає в переході до простору H більш високої розмірності, у якому класи можна розділити гіперплощиною. Для цього вводять нелінійне відображення $\varphi: X \to \mathrm{H}$, яке ставить у відповідність об’єктам X_i вихідного простору ознак об’єкти $\varphi(X_i)$ розширеного простору ознак. Тоді в розширеному просторі застосовують метод опорних векторів.

Щоб побудувати у розширеному просторі оптимальну гіперплощину методом опорних векторів, розв’язують задачу максимізації функції

$$L(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{q=1}^{N} \lambda_i \lambda_q y_i y_q \varphi(X_i) \varphi^T(X_q) \qquad (1.19)$$

за умов (1.18).

Після її розв'язання знаходять оптимальне значення вектора

$$w^* = \sum_{i=1}^{N} \lambda_i^* y_i \varphi(X_i),$$

а отже, відокремлювальну гіперплощину

$$w\varphi^T(X) + w_0 = 0$$

у вигляді

$$\sum_{i=1}^{N} \lambda_i^* y_i \varphi(X_i)\varphi^T(X) + w_0 = 0. \qquad (1.20)$$

Можна помітити, що в розширеному просторі максимізована функція (1.19) та відокремлювальна гіперплощина (1.20) залежать не від самих об'єктів нового простору, а від їх скалярного добутку $\varphi(X_i)\varphi^T(X_q)$. Позначимо

$$\mathrm{K}(X_i, X_q) = \varphi(X_i)\varphi^T(X_q).$$

Функцію $\mathrm{K}(X_i, X_q)$ називають **ядром скалярного добутку**.

Таким чином, у розширеному просторі розв'язують задачу максимізації функції

$$L(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{q=1}^{N} \lambda_i \lambda_q y_i y_q \mathrm{K}(X_i, X_q)$$

за умов (1.18) і будують гіперплощину

$$\sum_{i=1}^{N} \lambda_i^* y_i \mathrm{K}(X_i, X_q) + w_0 = 0. \qquad (1.21)$$

У вихідному просторі ознак вираз (1.21) відповідає нелінійній поверхні.

Функція $\mathrm{K}(X_i, X_q)$ може бути ядром скалярного добутку, якщо вона задовольняє умови теореми Мерсера [15].

Існує декілька стандартних ядер, застосовних у задачі класифікації:

1) поліномне однорідне $\mathrm{K}(X_i, X_q) = \left(X_i X_q^T\right)^d$;

2) поліномне неоднорідне $\mathrm{K}(X_i, X_q) = \left(1 + X_i X_q^T\right)^d$;

3) радіальна базисна функція Гаусса $\mathrm{K}(X_i, X_q) = \exp\left(-\frac{1}{2\sigma^2}\left\|X_i - X_q\right\|^2\right)$, $\sigma > 0$;

4) сигмоїд $\mathrm{K}(X_i, X_q) = \mathrm{th}\left(kX_i X_q^T + \delta\right)$ (але не за всіх значень k та δ).

До переваг методу опорних векторів можна віднести те, що правило класифікації будують, виходячи не з евристичних міркувань, а шляхом розв'язання математичної задачі, яка до того ж має єдиний розв'язок. Крім того, відповідний вибір ядра робить метод опорних векторів еквівалентним іншим відомим методам класифікації, зокрема різним видам нейронних мереж. Є у методу і недолік – чутливість до шуму в навчальних вибірках. Якщо у вибірці є аномальні викиди, вони будуть об'єктами-порушниками, а саме на них, як було зазначено вище, будують оптимальну гіперплощину.

1.4. Оцінка якості класифікації

Нехай проведено навчання за обраним методом, і на його основі побудовано правило класифікації. Але постає питання: якщо застосувати це правило для класифікації нових об’єктів, як часто воно буде «помилятися», тобто наскільки воно якісне?

Під **якістю класифікації** будемо розуміти відсоток помилок, які допускає правило на об’єктах контрольної вибірки. Чим менший відсоток помилок, тим якісніша класифікація за побудованим правилом.

На різних контрольних вибірках можна одержати різний відсоток помилок, тобто різну якість класифікації. Це свідчить про те, що задану якість можна гарантувати лише з деякою імовірністю. Імовірність одержання заданої якості називають **надійністю класифікації**.

Під час оцінки якості важливо, щоб об’єкти контрольної вибірки були незалежні від навчальних, інакше неможливо виявити перенавчання. **Перенавчання** (overtraining) – це небажане явище, яке виникає внаслідок застосування надзвичайно складних моделей для навчання, тоді правило занадто добре налаштовується на розпізнавання об’єктів навчальної вибірки, але при цьому втрачає здатність до узагальнення і якісного розпізнавання нових об’єктів. Про наявність перенавчання свідчить те, що відсоток помилок класифікації на об’єктах контрольної вибірки значно вищий, ніж на навчальних.

На практиці контрольна вибірка, як правило, відсутня, і оцінку якості потрібно проводити, маючи лише одну навчальну вибірку. У такому разі застосовують ковзний контроль (або крос-перевірку, крос-валідацію, cross-validation, CV). Існує декілька варіантів реалізації ковзного контролю:

1. Навчальну вибірку випадковим чином розбивають на 2 частини, наприклад, у співвідношенні 70:30 або 60:40. За першою (більшою) частиною проводять навчання, об’єкти з другої частини застосовують як контрольні, і на них обчислюють відсоток помилок. Таку процедуру повторюють певну кількість раз. Таким чином формують масив відсотків помилок, за яким розраховують середню арифметичну помилку, її і обирають як показник якості класифікації. Для аналізу надійності одержаної оцінки доцільно також обчислювати середньоквадратичне відхилення та розмах за масивом відсотків.

2. Навчальну вибірку розбивають на K блоків. Здійснюють навчання із застосування $(K-1)$-го блока, а об’єкти з того блока, що не був задіяний під час навчання, класифікують, для них підраховують відсоток помилок. Таку процедуру виконують K раз, кожного разу різний блок застосовують як контрольний. Таким способом формують масив відсотків помилок, за яким, як у першому варіанті, розраховують середню помилку.

Якщо навчальна вибірка невелика за розміром, то можна взяти кількість блоків K такою, що дорівнюватиме кількості об’єктів навчальної вибірки N. Тоді кожного разу навчання проходитиме на всій навчальній вибірці, крім одного об’єкта, який будуть класифікувати. Відсоток помилково класифікованих контрольних об’єктів слугуватиме оцінкою якості.

2. КЛАСТЕРНИЙ АНАЛІЗ ДАНИХ

Кластеризація – це розбиття множини об’єктів на однорідні підмножини за схожістю опису ознак об’єктів. При цьому необхідно, щоб об’єкти однієї підмножини були більш схожі між собою, ніж із об’єктами інших підмножин. Одержані в результаті розбиття підмножини називають **кластерами** (англ. cluster – гроно, в’язка). Термін «кластеризація» запропонований К. Тріоном у 1939 р.

Нехай задано множину об’єктів, яку слід кластеризувати:

$$X=\left\{X_i;i=\overline{1,N}\right\}=\left\{\left(x_{i,1}\quad x_{i,2}\quad\dots\quad x_{i,p}\right);i=\overline{1,N}\right\}=\begin{pmatrix}x_{1,1}&x_{1,2}&\dots&x_{1,p}\\x_{2,1}&x_{2,2}&\dots&x_{2,p}\\\dots&\dots&\dots&\dots\\x_{N,1}&x_{N,2}&\dots&x_{N,p}\end{pmatrix}.$$

Формально кластеризація являє собою одержання розбиття $S=\left\{S_l;l=\overline{1,K}\right\}$, тобто одержання сукупності непустих кластерів $S_l=\left\{X_i^{(l)};i=\overline{1,N_l}\right\}$, $l=\overline{1,K}$, таких, що $S_l\bigcap S_h=\varnothing$, коли $l\neq h$, $l,h=\overline{1,K}$, $X=\bigcup\limits_{l=1}^{K}S_l$. Кількість кластерів K заздалегідь може бути відома або її також необхідно визначити.

Слід зауважити, що таке визначення відповідає поняттю «чіткої» кластеризації, яку ми розглядаємо в даному навчальному посібнику. Виділяють також підходи, за яких кластери перетинаються ($S_l\bigcap S_h\neq\varnothing$). Це, наприклад, кластеризація на основі статистичної теорії (EM алгоритм та його модифікації) [11; 13] і теорії нечітких множин (методи *c*-середніх, Густафсона–Кесселя та ін.) [4].

У цілому універсального методу кластеризації не існує. Під час застосування кожного важливо враховувати, яку структуру даних він найкраще здатен розпізнавати. Існуючі методи можна розділити на такі групи: ієрархічні («чіткі» та «нечіткі»); засновані на оптимізації деякої цільової функції (*K*-середніх, FOREL, *c*-середніх, Густафсона–Кесселя тощо); графові; на основі оцінювання функції щільності (статистичні); нейронні мережі.

Перед кластеризацією часто необхідно **перетворювати ознаки**, приводячи їх значення до єдиного масштабу, щоб осмислено порівнювати об’єкти за ними. Найбільш застосовуваним перетворенням є стандартизація

$$x_{i,j}^{*}=\frac{x_{i,j}-\overline{x}_j}{\hat{\sigma}_j},\ j=\overline{1,p},\ i=\overline{1,N},$$

де $\overline{x}_j=\frac{1}{N}\sum\limits_{i=1}^{N}x_{i,j}$; $\hat{\sigma}_j=\sqrt{\frac{1}{N-1}\sum\limits_{i=1}^{N}\left(x_{i,j}-\overline{x}_j\right)^2}$,

також можна застосовувати й такі типи перетворень:

$$x_{i,j}^{*}=\frac{x_{i,j}}{x_{j,\max}-x_{j,\min}},\qquad x_{i,j}^{*}=\frac{x_{i,j}}{x_{j,\max}},\ j=\overline{1,p},\ i=\overline{1,N},$$

де $x_{j,\max}=\max\left\{x_{j,1},\dots,x_{j,N}\right\}$; $x_{j,\min}=\min\left\{x_{j,1},\dots,x_{j,N}\right\}$.

2.1. Ієрархічні методи кластеризації

Ієрархічні методи [1; 6; 7; 9] – найпоширеніші методи кластеризації. Розрізняють два види методів: агломеративні (об'єднуючі) та дивізимні (роз'єднуючі). В агломеративних ієрархічних методах спочатку кожен об'єкт розглядають як окремий кластер, далі знаходять два найближче розташовані кластери і об'єднують їх в один. Процес об'єднання продовжують, поки всі об'єкти не утворять один кластер. Логічна протилежність – дивізимні методи. У них на початковому етапі всі об'єкти вважають належними одному кластеру, який ділять на складові частини, поки кожен об'єкт на утворить окремий кластер.

Застосування ієрархічних методів потребує введення поняття **відстані між кластерами**, тобто між цілими групами об'єктів. Нехай є два кластери: $S_1 = \left\{ X_i^{(1)}, i = \overline{1, N_1} \right\}$ та $S_2 = \left\{ X_i^{(2)}, i = \overline{1, N_2} \right\}$. Відстань $D(S_1, S_2)$ між ними можна визначати:

1) як відстань найближчого сусіда – відстань між найближчими об'єктами кластерів (рис. 2.1, *а*)

$$D(S_1, S_2) = \min_{\substack{i_1 = \overline{1, N_1}; \\ i_2 = \overline{1, N_2}}} d\left(X_{i_1}^{(1)}, X_{i_2}^{(2)} \right);$$

2) відстань найвіддаленішого сусіда, яка дорівнює відстані між найвіддаленішими об'єктами кластерів (рис. 2.1, *б*)

$$D(S_1, S_2) = \max_{\substack{i_1 = \overline{1, N_1}; \\ i_2 = \overline{1, N_2}}} d\left(X_{i_1}^{(1)}, X_{i_2}^{(2)} \right);$$

3) середню зважену відстань, що дорівнює середньому значенню попарних відстаней між об'єктами різних кластерів

$$D(S_1, S_2) = \frac{1}{N_1 N_2} \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} d\left(X_{i_1}^{(1)}, X_{i_2}^{(2)} \right);$$

4) середню незважену відстань, яка відрізняється від попередньої тим, що в ній не враховують розміри кластерів

$$D(S_1, S_2) = \frac{1}{4} \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} d\left(X_{i_1}^{(1)}, X_{i_2}^{(2)} \right);$$

5) медіанна відстань (рис. 2.1, *в*)

$$D(S_1, S_2) = \frac{1}{2} d\left(Me^{(1)}, Me^{(2)} \right),$$

де $Me^{(l)} = \left(me_1^{(l)} \quad me_2^{(l)} \quad \ldots \quad me_p^{(l)} \right)$ – об'єкт із медіанними значеннями ознак у l-му кластері, $l = 1, 2$; значення $me_j^{(l)}$, $j = \overline{1, p}$ визначають за відсортованою

вибіркою $\left\{x_{1,j}^{(l)}, x_{2,j}^{(l)}, \ldots, x_{N_l,j}^{(l)}\right\}$ за формулою

$$me_j^{(l)} = \begin{cases} x_{(N_l+1)/2,j}^{(l)}, & \text{коли } N_l - \text{ непарне,} \\ \dfrac{1}{2}\left(x_{N_l/2,j}^{(l)} + x_{N_l/2+1,j}^{(l)}\right), & \text{коли } N_l - \text{ парне;} \end{cases}$$

6) відстань між центрами, що дорівнює відстані між об’єктами із середніми значеннями усіх ознак (рис. 2.1, *г*)

$$D(S_1, S_2) = d\left(\bar{X}^{(1)}, \bar{X}^{(2)}\right),$$

де $\bar{X}^{(l)} = \left(\bar{x}_1^{(l)} \quad \bar{x}_2^{(l)} \quad \ldots \quad \bar{x}_p^{(l)}\right)$, $\bar{x}_j^{(l)} = \dfrac{1}{N_l}\sum_{i=1}^{N_l} x_{i,j}^{(l)}$, $l = 1,2$;

7) відстань Уорда

$$D(S_1, S_2) = \frac{N_1 N_2}{N_1 + N_2} d^2\left(\bar{X}^{(1)}, \bar{X}^{(2)}\right).$$

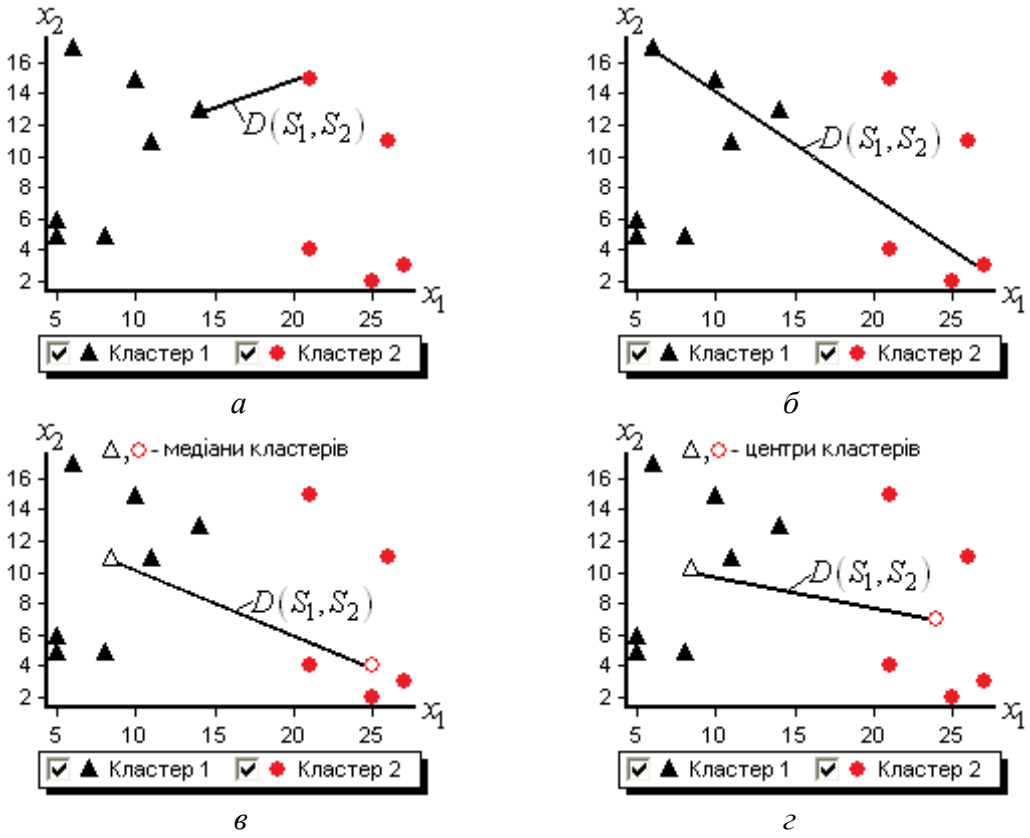


Рис. 2.1. Відстані між кластерами у разі евклідової відстані між об’єктами:
а – відстань найближчого сусіда; *б* – відстань найвіддаленішого сусіда;
в – медіанна відстань; *г* – відстань між центрами

На практиці найчастіше застосовують агломеративні методи, тому зупинимося на них детальніше. Нехай обрано метрику відстані між об’єктами й обчислено матрицю відстаней $d = \left(d_{i,q} = d\left(X_i, X_q\right); i, q = \overline{1, N}\right)$. Тоді процес розбиття об’єктів на кластери відбуватиметься послідовно за $N-1$ крок. На

першому кроці кожен об'єкт вважають окремим кластером, тобто $S_l = \{X_l\}$, $l = \overline{1,N}$, а матрицю відстаней між кластерами такою, що дорівнює d: $D = \left(D(S_l, S_h) = d_{l,h}; l,h = \overline{1,N}\right)$. У матриці D знаходять мінімальний елемент $D(S_l, S_h)$, і кластери S_l та S_h об'єднують в один кластер $S_{l+h} = S_l \cup S_h$, що складається вже з двох об'єктів. Після цього матрицю D змінюють. Із неї вилучають два рядки і два стовпчики, що містять відстані від S_l та S_h до інших кластерів, але додають один рядок і один стовпець із відстанями від кластера S_{l+h} до інших. Далі на кожному кроці процедуру повторюють, тобто знаходять мінімальний елемент у перетвореній матриці відстаней і відповідні кластери об'єднують в один, поки не буде одержаний один кластер.

При цьому відстані від нового кластера до інших можна обчислювати не за означенням, а застосовуючи вже відомі відстані за формулою Ланса–Уільямса

$$D(S_{l+h}, S_m) = \alpha_l D(S_l, S_m) + \alpha_h D(S_h, S_m) + \beta D(S_l, S_h) + \gamma \left| D(S_l, S_m) - D(S_h, S_m) \right|,$$

де α_l, α_h, β, γ – числові параметри.

Різні поєднання параметрів α_l, α_h, β, γ відповідають різним способам обрахування відстані між кластерами та породжують різні види агломеративних ієрархічних методів:

1) найближчого сусіда (або одного зв'язку):

$$\alpha_l = 0{,}5; \qquad \alpha_h = 0{,}5; \qquad \beta = 0; \qquad \gamma = -0{,}5;$$

2) найвіддаленішого сусіда (або повного зв'язку):

$$\alpha_l = 0{,}5; \qquad \alpha_h = 0{,}5; \qquad \beta = 0; \qquad \gamma = 0{,}5;$$

3) середнього зваженого зв'язку:

$$\alpha_l = \frac{N_l}{N_l + N_h}; \qquad \alpha_h = \frac{N_h}{N_l + N_h}; \qquad \beta = 0; \qquad \gamma = 0,$$

де N_l, N_h – кількість об'єктів у кластерах S_l та S_h, які об'єднують;

4) простого середнього зв'язку:

$$\alpha_l = 0{,}5; \qquad \alpha_h = 0{,}5; \qquad \beta = 0; \qquad \gamma = 0;$$

5) медіанного зв'язку:

$$\alpha_l = 0{,}5; \qquad \alpha_h = 0{,}5; \qquad \beta = -0{,}25; \qquad \gamma = 0;$$

6) центроїдний:

$$\alpha_l = \frac{N_l}{N_l + N_h}; \alpha_h = \frac{N_h}{N_l + N_h}; \beta = -\alpha_l \alpha_h = -\frac{N_l N_h}{\left(N_l + N_h\right)^2}; \gamma = 0;$$

7) Уорда:

$$\alpha_i = \frac{N_m + N_l}{N_m + N_l + N_h}; \alpha_j = \frac{N_m + N_h}{N_m + N_l + N_h}; \beta = -\frac{N_m}{N_m + N_l + N_h}; \gamma = 0,$$

де N_m – кількість об'єктів у кластері S_m, відстань до якого обчислюють.

Особливістю методу найближчого сусіда є наявність ланцюгового ефекту, він об'єднує в один кластер навіть дуже віддалені об'єкти, якщо існує ланцюг, що їх з'єднує. За рахунок цього за допомогою методу можна виділяти кластери

«стрічкової» форми. Метод найвіддаленішого сусіда, навпаки, може виявляти компактні гіперсферичні кластери.

Методи середнього зваженого та простого середнього зв'язку вважають проміжними за своїми властивостями.

Методи медіанного зв'язку й центроїдний рідко застосовують на практиці. По-перше, одержана за ними кластеризація має інверсії, тобто у процесі послідовного об'єднання кластерів відстань на кожному кроці не обов'язково збільшується. У результаті дендрограма має самоперетини. По-друге, міжкластерні відстані в них нередуктивні. Властивість редуктивності, уведена М. Брюіношем, полягає в тому, що для $\forall\delta > 0$ δ-окіл кластера, одержаного об'єднанням двох інших кластерів, повинен знаходитися всередині δ-околів початкових кластерів. Якщо ця властивість має місце, то можна прискорити процедуру пошуку кластерів-кандидатів для об'єднання на кожному кроці, виявляючи їх лише серед кластерів, що потрапили до δ-околу розглянутих на попередніх кроках алгоритму кластерів.

Метод Уорда мінімізує суму квадратів відстаней між двома гіпотетичними кластерами, які можуть бути сформовані на кожному кроці. Його недолік – можливе утворення кластерів малого розміру.

Суттєвою перевагою всіх агломеративних ієрархічних методів є те, що результати їх роботи можна подати наочно у вигляді **дендрограми**. Вона відображає ієрархічну структуру даних, а також дозволяє оцінити кількість кластерів, коли вона невідома. Дендрограма – це графік, у якому за вертикальною віссю відкладають номери об'єктів, а за горизонтальною – міжкластерні відстані, за яких відбувалося об'єднання двох кластерів (рис. 2.2, *б*). Слід зазначити, що під час побудови дендрогами об'єкти краще відкладати у порядку, в якому вони розташовані на останньому кроці роботи методу. Це дозволить одержати дендрограму без самоперетинів.

Приклад 2.1. Задано 6 об'єктів, кожен із яких описують дві ознаки: $\{(29\ \ 24);\ (29\ \ 32);\ (30\ \ 25);\ (30\ \ 35);\ (28\ \ 27);\ (31\ \ 33)\}$. Проведемо їх кластеризацію агломеративним ієрархічним методом простого середнього зв'язку. Як метрику відстані між об'єктами застосовуватимемо евклідову.

На першому кроці роботи методу кожен об'єкт розглядають як окремий кластер, матриця відстаней між кластерами збігається з матрицею евклідових відстаней між об'єктами (табл. 2.1). Найменша відстань (1,41) має місце між кластерами з об'єктів 1 і 3, отже, їх об'єднують в один кластер. На другому кроці (табл. 2.2) мінімальна відстань (2,24) спостерігається між кластерами з об'єктів 2 і 6, а також 4 і 6 (для визначеності у такому випадку будемо завжди об'єднувати першу знайдену пару; таким чином об'єднують кластери з об'єктів 2 і 6). На третьому кроці об'єднують кластер із двох об'єктів 2 і 6 із кластером, до якого входить об'єкт 4, оскільки саме між ними відстань мінімальна (табл. 2.3). На четвертому кроці кластер, що складається з об'єктів 1 і 3, об'єднують із кластером з об'єкта 5 (табл. 2.4). І на останньому п'ятому кроці два кластери (із об'єктів 1, 3, 5 та 2, 4, 6) об'єднують в один.

Таблиця 2.1

Матриця відстаней між кластерами на першому кроці (дорівнює матриці евклідових відстаней між об'єктами)

	1	**2**	**3**	**4**	**5**	**6**
1	0	8,00	**1,41**	11,05	3,16	9,22
2	8,00	0	7,07	3,16	5,10	2,24
3	**1,41**	7,07	0	10,00	2,83	8,06
4	11,05	3,16	10,00	0	8,25	2,24
5	3,16	5,10	2,83	8,25	0	6,71
6	9,22	2,24	8,06	2,24	6,71	0

Таблиця 2.2

Матриця відстаней між кластерами на другому кроці

	1, 3	2	4	5	6
1, 3	0	7,54	10,52	3,00	8,64
2	7,54	0	3,16	5,10	**2,24**
4	10,52	3,16	0	8,25	2,24
5	3,00	5,10	8,25	0	6,71
6	8,64	2,24	**2,24**	6,71	0

Таблиця 2.3

Матриця відстаней між кластерами на третьому кроці

	1, 3	2, 6	4	5
1, 3	0	8,09	10,52	3,00
2, 6	8,09	0	**2,70**	5,90
4	10,52	**2,70**	0	8,25
5	3,00	5,90	8,25	0

Таблиця 2.4

Матриця відстаней між кластерами на четвертому кроці

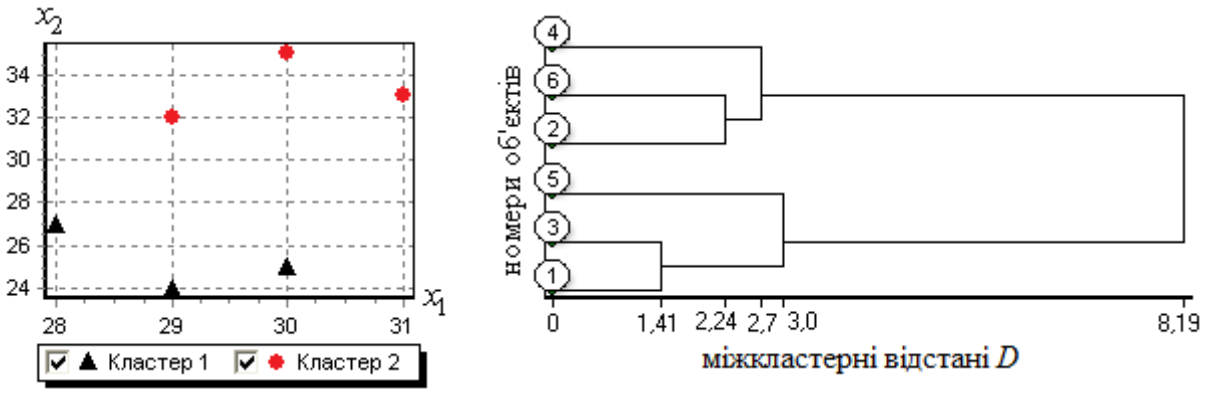
	1, 3	2, 6, 4	5
1, 3	0	9,31	**3,00**
2, 6, 4	9,31	0	7,07
5	**3,00**	7,07	0

Таблиця 2.5

Матриця відстаней між кластерами на п'ятому кроці

	1, 3, 5	2, 6, 4
1, 3, 5	0	**8,19**
2, 6, 4	**8,19**	0

Результати проведеної кластеризації подано нижче у вигляді графічного відображення об'єктів та дендрограми (рис. 2.2).



а *б*

Рис. 2.2. Результати агломеративної ієрархічної кластеризації:
а – графічне зображення об’єктів; *б* – дендрограма

Переваги розглянутих методів полягають у можливості наочного зображення ієрархічної структури даних у вигляді дендрограми і їх здатності виділяти кластери дуже різних форм за рахунок вибору відповідної метрики відстані між кластерами.

Головний недолік методів – необхідність великого обсягу пам’яті та значної кількості машинного часу, що унеможливлює їх застосування, коли обсяг даних перевищує декілька сотень. Такого недоліку не має метод *K*-середніх.

2.2. Метод *K*-середніх

Метод *K*-середніх (*K*-means) [1] – найпоширеніший серед неієрархічних методів кластеризації. Для його роботи необхідно попередньо задавати кількість кластерів *K*. Розрізняють два варіанти методу: Ллойда та Мак-Кіна.

Нижче наведено алгоритм **методу *K*-середніх у варіанті Ллойда**.

1. Обирають K об’єктів, які вважають центрами кластерів $\overline{X}^{(l)} = \left(\overline{x}_1^{(l)} \quad \overline{x}_2^{(l)} \quad \dots \quad \overline{x}_p^{(l)} \right)$, $l = \overline{1,K}$. Їх можна визначити як випадкові K об’єкти або як K найвіддаленіші об’єкти вихідної вибірки.

2. Кожен об’єкт X_i, $i = \overline{1,N}$ приєднують до кластера, центр якого ближчий. Номер кластера y_i, до якого відносять X_i, визначають так:

$$y_i = h: \qquad d\left(X_i, \overline{X}^{(h)} \right) = \min_{l=1,K} d\left(X_i, \overline{X}^{(l)} \right).$$

3. Обчислюють центри кластерів як центри мас:

$$\overline{x}_j^{(l)} = \frac{1}{N_l} \sum_{\substack{i=1,N, \\ y_i = l}} x_{i,j}, \quad j = \overline{1,p}, \quad l = \overline{1,K},$$

де N_l – кількість об’єктів у l-му кластері.

4. Кроки 2–3 повторюють для нових центрів кластерів, поки не буде виконано одну з умов:

а) центри кластерів стабілізувалися, тобто для двох послідовних ітерацій правдива умова

$$\left|\overline{x}_j^{(l,t+1)} - \overline{x}_j^{(l,t)}\right| \le \varepsilon, \ \ j = \overline{1,p}, \ l = \overline{1,K},$$

де t – номер ітерації; $\varepsilon > 0$ – будь-яке наперед задане число;

б) кількість ітерацій дорівнює заданій максимальній кількості.

У наведеному алгоритмі зазвичай застосовують евклідову відстань, що обумовлює виділення кластерів сферичної форми. Застосування відстані Махаланобіса дозволить виділяти кластери еліптичної форми.

Приклад 2.2. Проілюструємо роботу алгоритму на прикладі кластеризації 12 об'єктів, описаних двома ознаками: $\{(11\ \ 11); (27\ \ 3); (5\ \ 5); (6\ \ 17); (14\ \ 13); (25\ \ 2); (5\ \ 6); (21\ \ 15); (26\ \ 11); (21\ \ 4); (8\ \ 5); (10\ \ 15)\}$. Розіб'ємо об'єкти на два кластери, застосовуючи евклідову відстань.

Якщо на першому кроці як початкові центри кластерів обрати точки з координатами (21 4) і (25 2), то всі об'єкти розподіляться на кластери в такий спосіб: 2-й та 6-й будуть віднесені до другого кластера, усі інші – до першого (рис. 2.3, колонка «Крок 1»). Центром мас утворених сукупностей є точки (12,7 10,2) і (26 2,5) відповідно. На другому кроці об'єкти знову розподіляться за кластерами, але відносно нових центрів (рис. 2.3, колонка «Крок 2»), центрами утворених кластерів є точки (10 10,875) і (24,75 5). Після третього кроку 1-й, 3-й, 4-й, 5-й, 7-й, 11-й і 12-й об'єкти утворять перший кластер, а інші – другий; центрами їх мас будуть точки (8,43 10,29) та (24 7) відповідно (рис. 2.3, колонка «Крок 3»). На четвертому кроці буде одержано таке саме розбиття, що й на третьому, тобто центри кластерів не зміняться, і робота алгоритму завершиться.

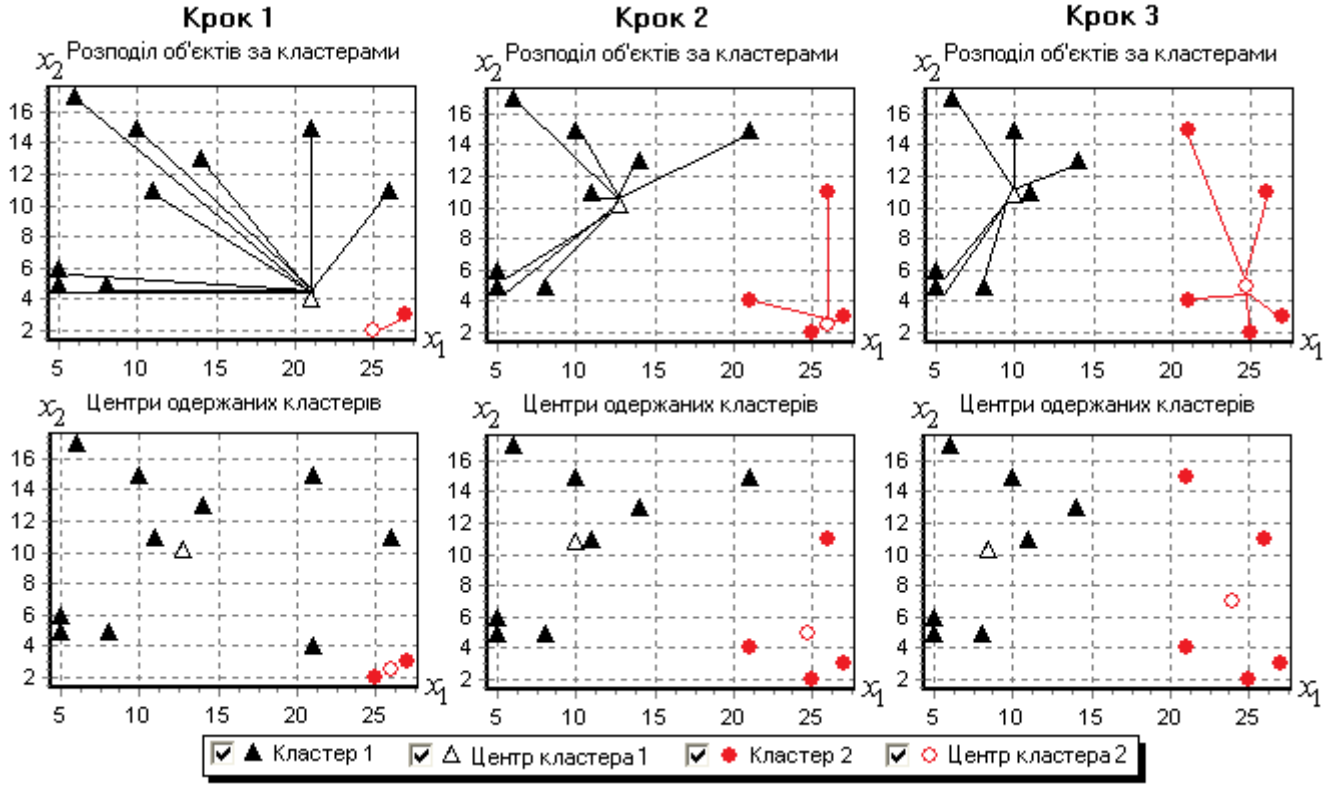


Рис. 2.3. Робота методу *K*-середніх у варіанті Ллойда

Метод *K*-середніх у варіанті Мак-Кіна відрізняється тим, що центри кластерів перераховують після віднесення кожної точки до найближчого кластера, тобто кроки 2 та 3 алгоритму виконують не послідовно, а паралельно. Мак-Кін довів, що такий алгоритм дозволяє мінімізувати суму внутрішньокластерних дисперсій

$$Q(S)=\sum_{l=1}^{K}\sum_{X_i\in S_l} d^2\left(X_i,\bar{X}^{(l)}\right). \qquad (2.1)$$

Зауваження. Назва «метод *K*-середніх» запропонована Дж. Мак-Кіном. Але фактично алгоритм у варіанті Ллойда реалізує ідею того ж методу з іншим порядком стабілізації, не послідовно для кожного кластера, а паралельно для всіх кластерів. Тому його також називають методом *K*-середніх [12].

Даний метод вирізняється простотою реалізації, наочністю та швидкістю роботи, що дозволяє його застосовувати для дуже великих за обсягом вибірок.

Головний недолік методу полягає в тому, що він дуже чутливий до початкового вибору центрів кластерів. Для його усунення рекомендовано застосовувати метод декілька раз із різними початковими центрами, а серед одержаних розбиттів обирати те, за якого функціонал (2.1) набуває найменшого значення. Крім того, метод нестійкий у разі наявності аномальних об’єктів. Для усунення даного недоліку запропоновано модифікувати метод шляхом обчислення центрів кластерів як медіан, а не середніх.

2.3. Методи FOREL

Методи групи FOREL (FORmal ELement) дозволяють виділяти кластери сферичної форми. Кожен кластер утворюють об’єкти, що потрапили у сферу з центром O і радіусом R. Центр сфери в загальному випадку не збігається з об’єктами, які слід кластеризувати, і його називають формальним елементом.

Базовий метод FOREL запропонований у 1967 р. для розв’язання однієї прикладної задачі в області палеонтології [8]. Його суть у такому.

Нехай задано радіус сфери R (параметр методу). Центр сфери O_1 поміщають у будь-яку точку множини об’єктів $\left\{X_i; i=\overline{1,N}\right\}$. Знаходять усі об’єкти, що потрапили у цю сферу. Вони утворюють множину

$$S_1=\left\{X_i : d\left(X_i,O_1\right)\le R\right\}.$$

За точками з множини S_1 обчислюють центр ваги, а центр сфери переміщують у нього, тобто новим центром сфери стає точка

$$O_1=\frac{1}{N_1}\sum_{X_i\in S_1} X_i,$$

де N_1 – кількість точок у множині S_1.

Відносно нового положення сфери знову формують множину об’єктів S_1, розраховують їх центр ваги і визначають новий центр сфери. Процедуру

повторюють, поки координати центра сфери O_1 не перестануть змінюватися. По закінченні цього ітераційного процесу сфера зупиняється в області локального скупчення точок. Об'єкти, що потрапили до неї, утворюють перший кластер S_1 і їх вилучають із подальшої обробки.

В одну з точок, що залишились, поміщають центр другої сфери O_2. Описану вище процедуру повторюють до стабілізації O_2 і виділення другого кластера S_2. І так до тих пір, поки не залишиться «вільних» об'єктів.

Доведено, що метод дозволяє одержати розбиття за скінченну кількість кроків, проте розв'язок може бути різний залежно від вибору початкових центрів сфер-кластерів (рис. 2.4).

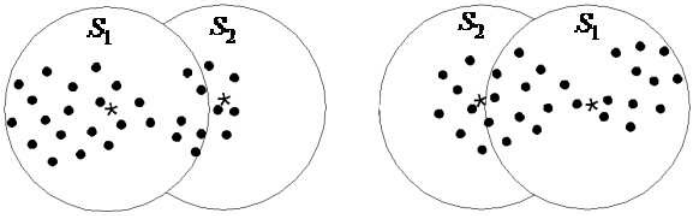


Рис. 2.4. Залежність одержаного розбиття від вибору початкового центра [8, с. 39]

Вибір найкращого розбиття S можна здійснити на основі функціонала

$$Q(S)=\sum_{l=1}^{K}\sum_{X_i\in S_l} d(X_i,O_l). \qquad (2.2)$$

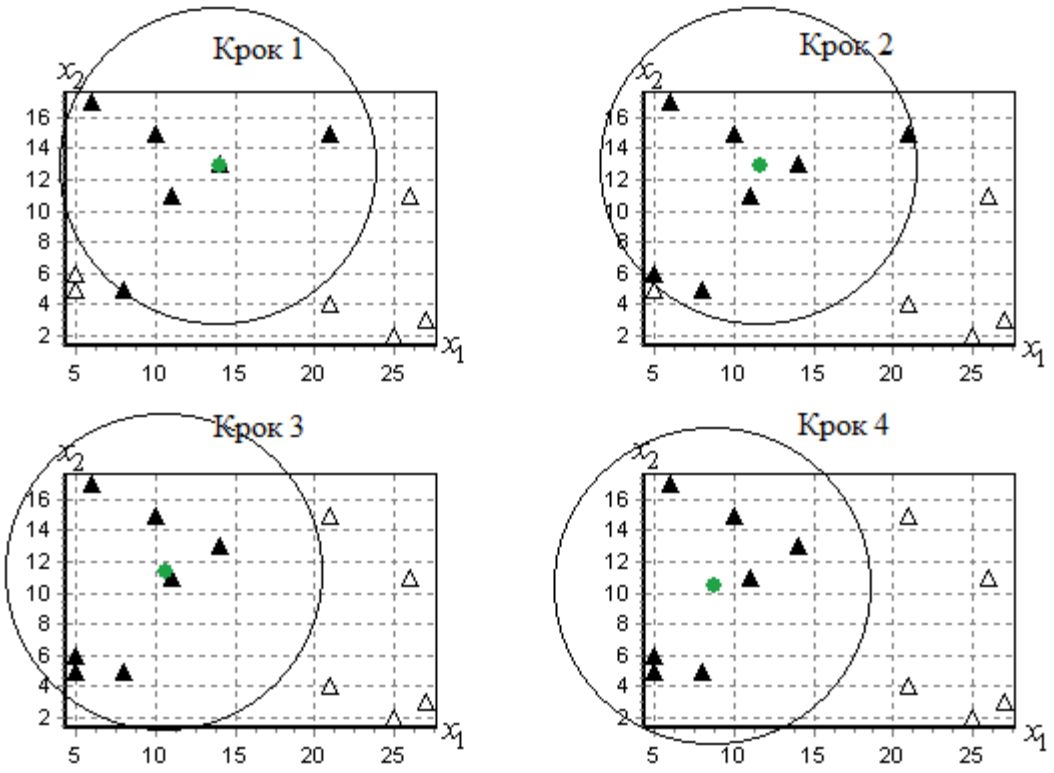
Кращому розбиттю відповідає найменше значення функціонала (2.2).

Приклад 2.3. Розіб'ємо об'єкти, подані в прикладі 2.2, на кластери за допомогою базового FOREL, поклавши $R=10$. Відстань між об'єктами розраховуватимемо як евклідову.

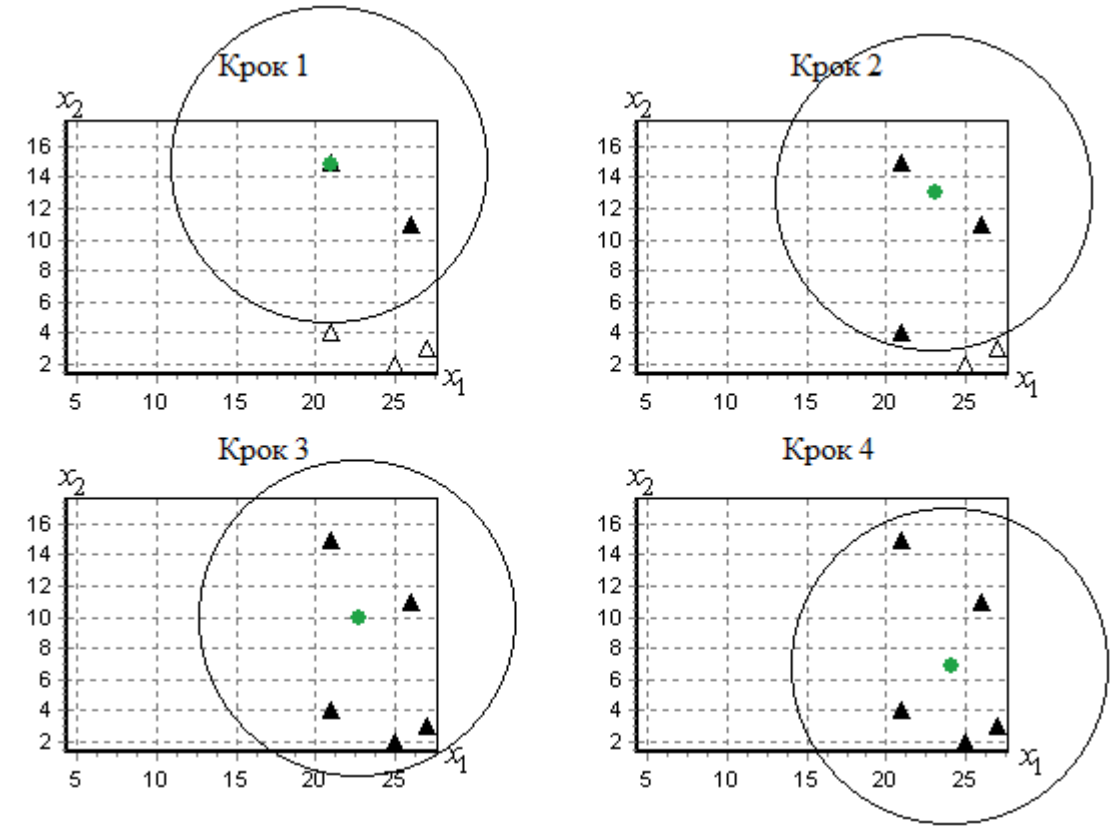
Якщо як початкове положення сфери обрати 5-й об'єкт із набору даних, тобто точку (14 13), то за 4 кроки центр сфери стабілізується (рис. 2.5, *а*). Об'єкти, що потраплять до сфери на 4-му кроці, утворять перший кластер. Це 1-й, 3-й, 4-й, 5-й, 7-й, 11-й і 12-й об'єкти. Вилучивши їх із подальшої обробки, оберемо як початкове значення центра сфери для другого кластера точку (21 15) – 8-й об'єкт. Стабілізація цієї сфери також відбудеться за 4 кроки (рис 2.5, *б*). У цю сферу потраплять 2-й, 6-й, 8-й, 9-й і 10-й об'єкти, вони утворять другий кластер. Ураховуючи, що «вільних» об'єктів не залишилося, метод завершить свою роботу. Таким чином, буде виділено 2 кластери, таких самих, як і у методі *K*-середніх. Хоча, обираючи інші об'єкти як початкові центри сфер, можна одержати інші розбиття, у тому числі й на 3 кластери.

У цілому базовий метод FOREL розбиває об'єкти на заздалегідь невідому кількість кластерів, яка буде тим більша, чим менший параметр методу R.

Для одержання розбиття на задану кількість кластерів K застосовують **метод FOREL 2**. Він передбачає ітераційне проведення кластеризації за допомогою базового методу FOREL із радіусом сфери R, який змінюють на кожному кроці.



а



б

Рис. 2.5. Робота методу FOREL:
а – виділення першого кластера;
б – виділення другого кластера

На початку радіус задають таким, щоб за допомогою базового FOREL одержати один кластер

$$R_0 = R_{\max},$$

де

$$R_{\max} = \max_{X_i, X_q \in X} d\left(X_i, X_q\right).$$

На кожному кроці t радіус зменшують або збільшують на величину ΔR:

$$R_t = R_{t-1} \pm \Delta R.$$

Знак «+» або «–» обирають залежно від того, скільки кластерів було одержано на попередньому кроці. Коли $K^{(t-1)} \le K$, обирають знак «–», інакше – знак «+». Якщо на $(t-1)$-му кроці одержано розбиття на потрібну кількість кластерів, тобто $K^{(t-1)} = K$, його запам'ятовують і робота методу продовжується.

За нового значення радіуса R_t проводять кластеризацію із застосуванням базового FOREL.

Величина ΔR на кожному кроці повинна зменшуватися, наприклад, удвічі:

$$\Delta R = \frac{R_{\max}}{2^t},\ t = 1, 2, \ldots$$

Ітераційний процес завершують, якщо ΔR стане менше наперед заданої похибки. Як остаточне розбиття обирають таке, що відповідає найменшому значенню радіуса сфери або функціонала (2.2).

2.4. Графові методи кластеризації

Реальні дані можуть утворювати кластери, форма яких значно відрізняється від сферичної або еліптичної, що виділяють методи *K*-середніх та FOREL. У такому разі можуть бути корисними графові методи. Вони засновані на поданні вихідних даних у вигляді зваженого графа, вершинами якого є об'єкти, а вагами ребер – відстані між об'єктами.

Одним із методів цієї групи є **метод на основі виділення компонент зв'язності графа**. Компонентою зв'язності називають таку підмножину вершин графа, що для будь-яких двох вершин із цієї підмножини існує шлях із однієї в іншу, і не існує шляху з вершини цієї підмножини до вершини іншої підмножини. Ідея методу полягає в розбитті графа на компоненти зв'язності, які й будуть являти собою кластери. Для цього задають параметр *R* і з графа видаляють усі ребра, вага яких більша за *R*. З'єднаними залишаються лише найбільш близькі пари об'єктів. Складність полягає в тому, щоб підібрати таке значення *R*, за якого граф розпадеться на декілька компонент зв'язності. Для підбору параметра *R* можна побудувати гістограму за вибіркою попарних відстаней. Якщо дані мають виражену кластерну структуру, ця гістограма буде двомодальна, одна мода буде відповідати внутрішньокластерним відстаням, інша – міжкластерним. Як *R* можна взяти точку мінімуму між модами.

Ще один представник цієї групи – **метод на основі коротшого незамкненого шляху** (КНШ) [8]. КНШ називають граф, усі вершини якого з'єднано таким чином, що немає петель, а сума ваг усіх ребер мінімальна. Побудувати КНШ дозволяє, наприклад, алгоритм Пріма. Знаходять пару об'єктів із найменшою відстанню і з'єднують їх ребром. Серед об'єктів, не доданих до графа, шукають найближчий до об'єкта графа і з'єднують із ним ребром. Коли КНШ побудовано, із нього видаляють $K-1$ ребро із найбільшою вагою. У результаті граф розпадається на K підграфів, і дані розбиваються на K кластерів.

Приклад 2.4. Розіб'ємо об'єкти із прикладу 2.2 на 4 кластери за допомогою методу на основі КНШ. Як відстань між об'єктами застосуємо евклідову.

КНШ, побудований за цими даними, подано нижче (рис. 2.6). Якщо видалити три ребра з найбільшою вагою (позначені тонкими лініями), утворяться чотири кластери. До першого увійдуть 3-й, 7-й та 11-й об'єкти, до другого – 1-й, 4-й, 5-й і 12-й, до третього – 8-й та 9-й, до четвертого – 2-й, 6-й і 10-й об'єкти.

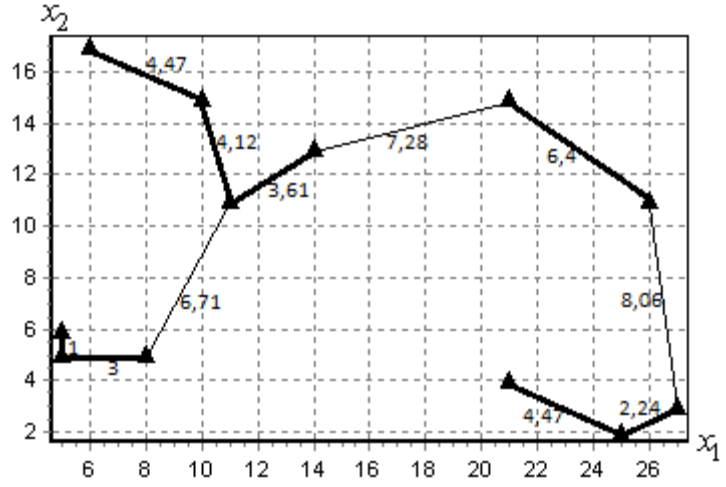


Рис. 2.6. Коротший незамкнений шлях

Модифікацією попереднього є **метод KRAB** [8]. Він також передбачає побудову КНШ і видалення з нього $K-1$ ребра. Але видаляють таку комбінацію ребер, за якої функціонал методу набуває максимального значення

$$Q = \ln\left(\frac{dh}{\lambda\rho}\right),$$

де d – середня вага ребер, які видаляють,

$$d = \frac{1}{K-1}\sum_{l=1}^{K-1} d_l;$$

d_l – l-те ребро, яке потрібно видалити; ρ – середня вага всіх внутрішніх ребер

$$\rho = \frac{1}{K}\sum_{l=1}^{K}\left(\frac{1}{N_l-1}\sum_{q=1}^{N_l-1}\alpha_q^l\right);$$

α_q^l – вага q-го ребра в l-му кластері; λ – міра неоднорідності на межах кластерів

$$\lambda = \frac{1}{K-1}\sum_{l=1}^{K-1}\lambda_l; \qquad \lambda_l = \frac{\alpha_{\min}^l}{d_l};$$

$\alpha_{\min}^{l}$ – довжина ребра з мінімальною вагою, що межує з l-м ребром, яке видаляють;
h – міра, що характеризує рівномірність розподілу об'єктів за кластерами

$$h = K^K \prod_{l=1}^{K} \frac{N_l}{N}.$$

Якщо рівномірність кластерів за кількістю об'єктів не важлива, тоді застосовують функціонал

$$Q = \ln\left(\frac{d}{\lambda\rho}\right).$$

Комбінацію ребер, що надає максимуму функціоналу Q, шукають шляхом перебору. Це робить метод KRAB досить трудомістким. Для скорочення перебору потрібну комбінацію можна шукати лише серед ребер, для яких величина λ_l менша порога, наприклад, одиниці.

Перевагою всіх графових методів є їх наочність, відносна простота реалізації і наявність безлічі можливостей для удосконалення. Їх недолік полягає у високій чутливості до шуму, зокрема, наявність розрідженого фону або вузьких перемичок між кластерами стає причиною неадекватної кластеризації на основі цих методів [3].

2.5. Функціонали якості кластеризації

Розбиття вихідної вибірки на кластери, одержані різними методами або за різних значень параметрів, можуть відрізнятися. Кількісним критерієм, що дозволяє надати перевагу одному із розбиттів, є **функціонал якості**.

Нехай у результаті застосування методу кластеризації одержано розбиття $S = \{S_1, S_2, \ldots, S_K\}$ вихідної вибірки на K кластерів. Для оцінки якості розбиття S відомо близько 50 функціоналів якості. Найпоширеніші з них такі:

1. Функціонал, заснований на аналізі міжкластрених та внутрішньокластерних коваріацій, для найкращого розбиття має бути мінімальним:

$$Q_1(S) = \operatorname{trace}\left(\Sigma_{\text{В}}^{-1}\Sigma_{\text{М}}\right), \text{або } Q_1'(S) = \det\left(\Sigma_{\text{В}}^{-1}\Sigma_{\text{М}}\right),$$

де trace – слід; det – визначник матриці; $\Sigma_{\text{М}}$ – матриця міжкластерних коваріацій

$$\Sigma_{\text{М}} = \frac{1}{N}\sum_{l=1}^{K} N_l \left(\bar{X}^{(l)} - \bar{X}\right)^{\text{T}} \left(\bar{X}^{(l)} - \bar{X}\right);$$

$\bar{X}^{(l)}$ – центр кластера S_l

$$\bar{X}^{(l)} = \frac{1}{N_l}\sum_{i=1}^{N_l} X_i^{(l)};$$

$\bar{X}$ – центр усієї сукупності об'єктів

$$\bar{X} = \frac{1}{N}\sum_{l=1}^{K} N_l \bar{X}^{(l)} = \frac{1}{N}\sum_{i=1}^{N} X_i;$$

$\Sigma_{\text{В}}$ – матриця внутрішньокластерних коваріацій

$$\Sigma_{\text{В}} = \frac{1}{N}\sum_{l=1}^{K} N_l \Sigma_l ;$$

Σ_l – матриця коваріацій кластера S_l

$$\Sigma_l = \frac{1}{N_l}\sum_{i=1}^{N_l}\left(X_i^{(l)} - \overline{X}^{(l)}\right)^{\text{T}}\left(X_i^{(l)} - \overline{X}^{(l)}\right).$$

2. Сума внутрішньокластерних дисперсій, яка повинна бути мінімальна:

$$Q_2(S) = \text{trace}(\Sigma_{\text{В}}).$$

Замість $Q_2(S)$ можна застосовувати функціонал, що враховує і внутрішньокластерні коваріації:

$$Q_2'(S) = \det(\Sigma_{\text{В}}).$$

3. Сума попарних внутрішньокластерних відстаней, яка також повинна бути мінімальна:

$$Q_3(S) = \sum_{l=1}^{K}\sum_{i=1}^{N_l-1}\sum_{q=i+1}^{N_l} d\left(X_i^{(l)}, X_q^{(l)}\right).$$

Зручність даного функціонала в тому, що його мінімізація забезпечує максимізацію суми міжкластерних відстаней.

4. Сума міжкластерних відстаней

$$Q_4(S) = \sum_{l=1}^{K-1}\sum_{i=1}^{N_l}\left(\sum_{h=l+1}^{K}\sum_{q=1}^{N_h} d\left(X_i^{(l)}, X_q^{(l)}\right)\right)$$

має бути максимальна.

5. Відношення функціоналів повинно бути мінімальне:

$$Q_5(S) = \frac{\overline{Q_3}(S)}{\overline{Q_4}(S)},$$

де $\overline{Q_3}(S)$, $\overline{Q_4}(S)$ – середня внутрішньокластерна та міжкластерна відстані

$$\overline{Q_3}(S) = \frac{Q_3(S)}{\frac{1}{2}\sum_{l=1}^{K} N_l(N_l - 1)}; \qquad \overline{Q_4}(S) = \frac{Q_4}{\sum_{l=1}^{K-1}\sum_{h=l+1}^{K} N_l N_h}.$$

Оскільки мінімізація $\overline{Q_3}(S)$ не гарантує максимізації $\overline{Q_4}(S)$, щоб урахувати як внутрішньокластерну, так і міжкластерну відстань застосовують їх відношення.

Якість багатьох методів кластеризації суттєво залежить від **вибору кількості кластерів** K. Задача вибору оптимальної кількості кластерів поки не має однозначного розв’язку.

3. ФОРМУВАННЯ НАБОРУ ІНФОРМАТИВНИХ ОЗНАК

Застосування методів розпізнавання образів тісно пов’язане із виявленням та використанням у вирішальних правилах наборів інформативних ознак.

Дуже часто на етапі збирання даних дослідники не мають чітких уявлень щодо корисності тих чи інших ознак і намагаються описати об’єкт якомога більшою їх кількістю. Це призводить до перевантаженості початкового опису через включення ознак, які не важливі для розділення класів, а інколи за рахунок випадкового впливу навіть зашумляють це розділення і погіршують якість розпізнавання, та ознак, що дублюють інформацію, закладену в інших показниках. Наприклад, під час обстеження пацієнтів, у кожного з них можна вимірювати рівень систолічного та діастолічного тиску, пульсу, тромбоцитів, лейкоцитів, цукру, параметри серця і ще деякі ознаки. Але для діагностики артеріальної гіпертензії достатньо величин тиску, пульсу і параметрів серця; якщо інші відкинути, це не погіршить якість діагностики, тобто інші ознаки можна вважати неінформативними для розпізнавання діагнозу артеріальної гіпертензії. Тому виникає потреба у відшуканні інформативних ознак, що найкраще виявляють відмінності між класами. Їх знаходження також дозволяє скоротити кількість ознак без суттєвого зменшення якості розпізнавання і за рахунок цього зменшити економічні й часові витрати на розпізнавання та, можливо, візуалізувати об’єкти у дво- або тривимірному просторі.

Методи, застосовні для формування набору інформативних ознак, можна розділити на дві групи:

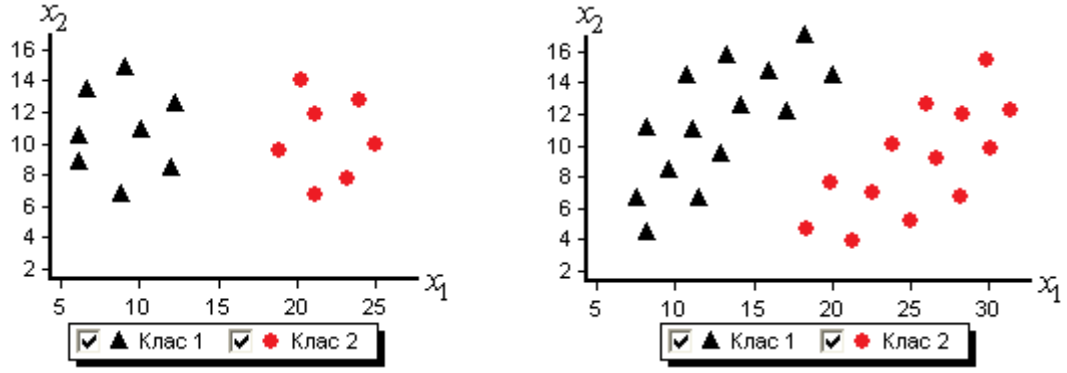
1) такі, у яких інформативні ознаки обирають із сукупності вихідних (саме вони – предмет вивчення в даному посібнику);

2) методи, у яких інформативні ознаки формують як лінійні або нелінійні комбінації вихідних, у результаті одержують нові ознаки, які безпосередньо не заміряли на об’єктах (методи головних компонент, факторного аналізу, багатовимірного шкалювання).

У межах першої групи виділяють методи, у яких інформативний набір:

1) формують із ознак, найкращих за своєю індивідуальною цінністю. Кожну ознаку розглядають окремо, незалежно від інших, їй приписують вагу, що визначає цінність ознаки для розпізнавання; ознаки впорядковують за зменшенням або збільшенням ваги (залежно від методу), перші ознаки в упорядкуванні складають інформативний набір;

2) відшукують серед різноманітних наборів, здебільшого шляхом перебору. Це ефективніший підхід, оскільки часто цінність ознаки залежить від інформативності всього набору, за яким проводять розпізнавання. Таку ситуацію подано на рис. 3.1. В обох випадках за ознакою x_2 два класи зовсім не розділяються, тобто стосовно індивідуальної цінності ознака x_2 зовсім неінформативна для класифікації. Проте в першому випадку (рис. 3.1, *а*) її дійсно можна відкинути як неінформативну, а в іншому (рис. 3.1, *б*) – ні, оскільки класифікація за сукупністю ознак x_1, x_2 набагато краща, ніж за однією x_1.



а *б*

Рис. 3.1. Ознака x_2 без індивідуальної цінності для розпізнавання:
а – дійсно неінформативна; *б* – інформативна в наборі з x_1

3.1. Формування набору інформативних ознак за їх індивідуальною цінністю

Для формування набору інформативних ознак у випадку, коли вони кількісні, можна застосувати методи апроксимації матриці відстаней та Кендалла.

Апроксимація матриці відстаней між об'єктами. Нехай у p-вимірному просторі задано множину $X = \left\{x_{i,j}; i = \overline{1,N},\ j = \overline{1,p}\right\}$, де $x_{i,j}$ – значення j-ї ознаки для i-го об'єкта (дійсні числа).

На основі X будують матрицю відстаней між об'єктами $d = \left(d_{i,q}; i,q = \overline{1,N}\right)$, наприклад, евклідових:

$$d_{i,l} = \sqrt{\sum_{h=1}^{p} \left|x_{i,h} - x_{l,h}\right|^2} = \sqrt{\left|x_{i,1} - x_{l,1}\right|^2 + \left|x_{i,2} - x_{l,2}\right|^2 + \ldots + \left|x_{i,p} - x_{l,p}\right|^2}\,.$$

Обчислюють за кожною ознакою матриці відстаней між об'єктами $d_1,\ d_2,\ \ldots,\ d_p$: $d_j = \left(d_{i,q}^{j}; i,q = \overline{1,N}\right)$, де $d_{i,q}^{j} = \left|x_{i,j} - x_{q,j}\right|,\ j = \overline{1,p}$.

Необхідно апроксимувати матрицю d із заданою точністю деяким набором матриць d_j, де $j \in J$, J – множина номерів ознак потужності r. Для цього обраховують відстані між d та всіма матрицями d_j, $j = \overline{1,p}$ за формулою

$$\left|d - d_j\right| = \sum_{i=1}^{N} \sum_{q=i+1}^{N} \left|d_{i,q} - d_{i,d}^{j}\right|,\ j = \overline{1,p}\,.$$

Найкраще матрицю d апроксимують перші r матриць d_j, найближчі до d. Тому матриці d_j, $j = \overline{1,p}$ розташовують за зростанням величин $\left|d - d_j\right|$. Відповідний їм набір ознак вважають інформативним. При цьому величину $\left|d - d_j\right|$ можна розглядати як вагу ознаки x_j – чим вона менша, тим інформативніша ознака.

Приклад 3.1. Задано результати спостережень над трьома 4-вимірними об'єктами

$$X = \begin{pmatrix} 56 & 62 & 52 & 66 \\ 69 & 77 & 70 & 81 \\ 60 & 67 & 55 & 79 \end{pmatrix}.$$

Матриця відстаней між цими об'єктами має вигляд

$$d = \begin{pmatrix} 0 & 31 & 15 \\ 31 & 0 & 20 \\ 15 & 20 & 0 \end{pmatrix},$$

у ній для простоти подальших обчислень округлили значення до цілих.

Матриці відстаней між об'єктами за кожною ознакою відповідно такі:

$$d_1 = \begin{pmatrix} 0 & 13 & 4 \\ 13 & 0 & 9 \\ 4 & 9 & 0 \end{pmatrix}, \quad d_2 = \begin{pmatrix} 0 & 15 & 5 \\ 15 & 0 & 10 \\ 5 & 10 & 0 \end{pmatrix}, \quad d_3 = \begin{pmatrix} 0 & 18 & 3 \\ 18 & 0 & 15 \\ 3 & 15 & 0 \end{pmatrix}, \quad d_4 = \begin{pmatrix} 0 & 15 & 13 \\ 15 & 0 & 2 \\ 13 & 2 & 0 \end{pmatrix}.$$

Обрахуємо відстані між d і всіма матрицями d_j, $j = \overline{1,4}$:

$$|d - d_1| = 40, \quad |d - d_2| = 36, \quad |d - d_3| = 30, \quad |d - d_4| = 36.$$

За віддаленістю від d матриці d_j розташовані

$$d_3,\ d_2,\ d_4,\ d_1,$$

де матриці d_2 та d_4 однаково віддалені від d.

Отже, найінформативніша – третя ознака, однаково менш інформативні друга та четверта ознаки, а найменш інформативна перша ознака.

Метод Кендалла. Нехай вхідні дані подано у вигляді матриці $X = \left\{ \left(x_{i,1} \quad x_{i,2} \quad \ldots \quad x_{i,p} \right), y_i; i = \overline{1,N} \right\}$, де N – кількість об'єктів; p – кількість ознак; $x_{i,j}$ – значення j-ї ознаки для i-го об'єкта, дійсні числа; y_i – номер класу, до якого належить i-й об'єкт, $y_i \in \{1,2,\ldots,K\}$; K – кількість класів.

Застосування методу Кендалла дозволяє впорядкувати ознаки за спаданням їх інформативності. Якщо $K = 2$, ранжування ознак проводять таким чином. Для кожної ознаки x_j, $j = \overline{1,p}$ виконують нижченаведені кроки:

1. Будують матрицю частот із трьох стовпців; перший стовпець містить упорядковані значення ознаки, другий – частоту появи цього значення для першого класу, третій – для другого класу.

2. Знаходять інтервал зміни значень ознаки, спільний для двох класів.

3. Обчислюють N_j як кількість об'єктів обох класів, значення яких за ознакою x_j не потрапили у спільний інтервал.

Ознаки впорядковують за спаданням N_j, $j = \overline{1,p}$, це відповідає їх впорядкуванню за спаданням індивідуальної цінності. Величини N_j, $j = \overline{1,p}$

можна розглядати як ваги ознак. Чим більше значення N_j, тим більшу кількість об'єктів можна розпізнати за ознакою x_j, тобто тим інформативніша ця ознака.

Якщо кількість класів $K > 2$, то метод Кендалла застосовують до кожної пари класів, і обраховані ваги усереднюють за кількістю пар.

Приклад 3.2. Задано результати спостережень над чотирма 3-вимірними об'єктами у вигляді матриці (останній стовпець містить номер класу об'єкта):

$$X = \begin{pmatrix} 56 & 62 & 52 & 1 \\ 69 & 77 & 70 & 2 \\ 60 & 67 & 55 & 1 \\ 59 & 77 & 52 & 2 \end{pmatrix}.$$

Матриця частот для першої ознаки має вигляд

$$\begin{pmatrix} 56 & 1 & 0 \\ 59 & 0 & 1 \\ 60 & 1 & 0 \\ 69 & 0 & 1 \end{pmatrix}.$$

Для об'єктів першого класу значення першої ознаки лежать в інтервалі $[56;60]$. Для об'єктів другого класу значення першої ознаки змінюються в діапазоні $[59;69]$. Спільний для двох класів інтервал – $[59;60]$. Тому $N_1 = 2$.

Матриця частот для другої ознаки така:

$$\begin{pmatrix} 62 & 1 & 0 \\ 67 & 1 & 0 \\ 77 & 0 & 2 \end{pmatrix}.$$

Для об'єктів першого класу значення другої ознаки лежать в інтервалі $[62;67]$, для об'єктів другого класу – у діапазоні $[77;77]$. Спільний для двох класів інтервал порожній. Отже, $N_2 = 4$.

Матриця частот для третьої ознаки має вигляд

$$\begin{pmatrix} 52 & 1 & 1 \\ 55 & 1 & 0 \\ 70 & 0 & 1 \end{pmatrix}.$$

Для об'єктів першого класу значення третьої ознаки лежать в інтервалі $[52;55]$, для об'єктів другого класу – у діапазоні $[52;70]$. Спільний для двох класів інтервал – $[52;55]$. Тому $N_3 = 1$.

Таким чином,

$$N_2 = 4,\ N_1 = 2,\ N_3 = 1$$

і найінформативніша друга ознака, менш інформативні перша та третя.

Для формування простору інформативних ознак, якщо ознаки якісні, можна застосовувати методи «гойдалки» та голосування.

Метод «гойдалки». Позитивні результати дало застосування методу в медичних дослідженнях. Він дозволяє впорядковувати не лише ознаки за їх інформативністю, але й об'єкти.

Вхідні дані подають у вигляді матриці з нулів та одиниць $X = \left\{ x_{i,j}; i = \overline{1,N},\ j = \overline{1,p} \right\}$, де $x_{i,j}$ – значення *j*-ї ознаки для *i*-го об'єкта, яке може набувати значення 0 чи 1.

Матриця *X* повинна задовольняти такі вимоги:

1) $p \geq 2,\ N \geq 2$;

2) матриця не повинна мати однакових рядків і стовпців;

3) матриця не повинна мати рядків і стовпців, що повністю складаються із нулів або одиниць.

Метод «гойдалки» ітераційний, його суть така. Кожній ознаці й кожному об'єкту приписують вагу. Ваги ознак утворюють вектор $\vec{z}$, а ваги об'єктів – вектор $\vec{y}$. Початкові значення цих векторів задають як

$$\vec{z}_1 = \begin{pmatrix} z_1 & z_2 & \dots & z_p \end{pmatrix}, \qquad \vec{y}_1 = \begin{pmatrix} y_1 & y_2 & \dots & y_N \end{pmatrix},$$

де

$$z_j = \sum_{i=1}^{N} x_{i,j}, \qquad y_i = \sum_{j=1}^{p} x_{i,j}.$$

Далі визначають пари векторів

$$\vec{W} = \begin{pmatrix} w_1 & w_2 & \dots & w_p \end{pmatrix}, \qquad \vec{V} = \begin{pmatrix} v_1 & v_2 & \dots & v_N \end{pmatrix},$$

де

$$w_j = \sum_{i=1}^{N} x_{i,j} y_i,\ j = \overline{1,p}, \qquad v_i = \sum_{j=1}^{p} x_{i,j} z_j,\ i = \overline{1,N}.$$

Тоді на другому кроці одержують вектори

$$\vec{z}_2 = \frac{\vec{W}}{\max\limits_{j=1,p} \left| w_j \right|}, \qquad \vec{y}_2 = \frac{\vec{V}}{\max\limits_{i=1,N} \left| v_i \right|}.$$

Таким чином, одержують послідовність векторів

$$\vec{z}_1, \vec{z}_2, \dots, \vec{z}_k, \dots,$$

що містять ваги ознак, та послідовність векторів

$$\vec{y}_1, \vec{y}_2, \dots, \vec{y}_k, \dots,$$

які містять ваги об'єктів.

Кожен наступний вектор із вагами визначають за допомогою попереднього. Ітераційний процес закінчують, коли вектори з вагами ознак та об'єктів перестають змінюватися. Його збіжність була доведена авторами методу Ю.Л. Васильєвим та А.Н. Дмитрієвим.

Розміщенню ознак за спаданням ваги відповідає їх впорядкування за важливістю для класифікації. Розташування об'єктів за спаданням ваги дозволяє

виділяти класичних представників класів і розв'язувати задачу щодо оптимізації обсягу вибірки.

Приклад 3.3. Задано початкові дані у вигляді матриці

$$X = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Виконаємо декілька ітерацій методу «гойдалки»:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \vec{y}_1 \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{pmatrix} \quad \vec{V} \begin{pmatrix} 3 \\ 3 \\ 5 \\ 6 \\ 5 \end{pmatrix} \quad \vec{y}_2 \begin{pmatrix} 1/2 \\ 1/2 \\ 5/6 \\ 1 \\ 5/6 \end{pmatrix} \quad \vec{V} \begin{pmatrix} 1 \\ 1 \\ 1{,}8 \\ 2 \\ 1{,}8 \end{pmatrix} \quad \vec{y}_3 \begin{pmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}9 \\ 1 \\ 0{,}9 \end{pmatrix}$$

$$(\max v_i = 6) \qquad (\max v_i = 2)$$

$$\vec{z}_1 = (2 \quad 3 \quad 3)$$

$$\vec{W} = (4 \quad 5 \quad 5) \ (\max w_j = 5)$$

$$\vec{z}_2 = (0{,}8 \quad 1 \quad 1)$$

$$\vec{W} = (10/6 \quad 14/6 \quad 14/6) \ (\max w_j = 14/6)$$

$$\vec{z}_3 = (5/7 \quad 1 \quad 1)$$

На третьому кроці ітерацій одержано вектор $\vec{z}_3 = (5/7 \quad 1 \quad 1)$, згідно з яким друга та третя ознаки однаково інформативні, а найменш інформативна – перша ознака.

Метод голосування. Вхідні дані для методу подають у вигляді матриці $X = \left\{ \left(x_{i,1} \quad x_{i,2} \quad \dots \quad x_{i,p} \right), y_i; i = \overline{1,N} \right\}$, де $x_{ij} = 0 \vee 1$ – значення j-ї ознаки для i-го об'єкта; $y_i \in \{1,2,\dots,K\}$ – номер класу i-го об'єкта; K – кількість класів.

Для зручності подають вихідні дані у вигляді K матриць (навчальних вибірок) $X^{(1)}, X^{(2)}, ..., X^{(K)}$, таких, що кожна матриця $X^{(l)}$ містить об'єкти лише l-го класу

$$X^{(l)} = \begin{pmatrix} X_1^{(l)} \\ X_2^{(l)} \\ \dots \\ X_{N_j}^{(l)} \end{pmatrix} = \begin{pmatrix} x_{1,1}^{(l)} & x_{1,2}^{(l)} & \dots & x_{1,p}^{(l)} \\ x_{2,1}^{(l)} & x_{2,2}^{(l)} & \dots & x_{2,p}^{(l)} \\ \dots & \dots & \ddots & \dots \\ x_{N_l,1}^{(l)} & x_{N_l,2}^{(l)} & \dots & x_{N_l,p}^{(l)} \end{pmatrix}, \ l = \overline{1,K}.$$

Обчислюють кількість голосів, поданих об’єктами *l*-го класу (*l*-ї навчальної вибірки) за свій клас, за формулою

$$\Gamma_l = \sum_{i=1}^{N_l-1} \sum_{q=i+1}^{N_l} \rho\left(X_i^{(l)}, X_q^{(l)} \right),$$

де $\rho\left(X_i^{(l)}, X_q^{(l)} \right)$ – кількість ознак, що збігаються в i-го і q-го об’єктів,

$$\rho\left(X_i^{(l)}, X_q^{(l)} \right) = \sum_{h=1}^{p} \rho'\left(x_{i,h}^{(l)}, x_{q,h}^{(l)} \right);$$

$$\rho'\left(x_{i,h}^{(l)}, x_{q,h}^{(l)} \right) = \begin{cases} 1, & \text{якщо } h\text{-та ознака в } i\text{-го й } q\text{-го об’єктів збігається;} \\ 0 & \text{в іншому випадку.} \end{cases}$$

Якщо виключити із розгляду ознаку x_j, то кількість голосів, поданих об’єктами *l*-го класу за свій клас за відсутності ознаки x_v, можна визначити як

$$\Gamma_{l,j} = \sum_{i=1}^{N_l-1} \sum_{q=i+1}^{N_l} \rho_{x_j}\left(X_i^{(l)}, X_q^{(l)} \right),$$

де

$$\rho_{x_j}\left(X_i^{(l)}, X_q^{(l)} \right) = \sum_{\substack{h=1 \\ h \neq j}}^{p} \rho'\left(X_{q,h}^{(l)}, X_{q,h}^{(l)} \right).$$

Тоді цінність ознаки x_j для даної класифікації становитиме

$$C_j = \sum_{l=1}^{K} \left(\Gamma_l - \Gamma_{l,j} \right).$$

Приклад 3.4. Задано вхідну матрицю X (останній стовпець у ній містить номери класів об’єктів), яку подамо у вигляді двох матриць $X^{(1)}$ та $X^{(2)}$, що містять об’єкти першого та другого класів відповідно

$$X = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix}, \qquad X^{(1)} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad X^{(2)} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Матимемо

$$\Gamma_1 = \sum_{i=1}^{2} \sum_{q=i+1}^{3} \rho\left(X_i^{(1)}, X_q^{(1)} \right) = \rho\left(X_1^{(1)}, X_2^{(1)} \right) + \rho\left(X_1^{(1)}, X_3^{(1)} \right) + \rho\left(X_2^{(1)}, X_3^{(1)} \right) = 2 + 2 + 1 = 5,$$

$$\Gamma_2 = \sum_{i=1}^{2} \sum_{q=i+1}^{2} \rho\left(X_i^{(2)}, X_q^{(2)} \right) = \rho\left(X_1^{(2)}, X_2^{(2)} \right) = 2,$$

$$\Gamma_{1,1} = \rho_{x_1}\left(X_1^{(1)}, X_2^{(1)} \right) + \rho_{x_1}\left(X_1^{(1)}, X_3^{(1)} \right) + \rho_{x_1}\left(X_2^{(1)}, X_3^{(1)} \right) = 2 + 1 + 1 = 4,$$

$$\Gamma_{2,1} = \rho_{x_1}\left(X_1^{(2)}, X_2^{(2)} \right) = 2.$$

Тоді цінність першої ознаки становитиме

$$C_1 = \left(\Gamma_1 - \Gamma_{1,1}\right) + \left(\Gamma_2 - \Gamma_{2,1}\right) = (5-4) + (2-2) = 1.$$

Для другої ознаки

$$\Gamma_{1,2} = 2, \qquad \Gamma_{2,2} = 1, \qquad C_2 = 4.$$

Для третьої ознаки

$$\Gamma_{1,3} = 4, \qquad \Gamma_{2,3} = 1, \qquad C_3 = 2.$$

За важливістю для класифікації друга ознака найінформативна, її цінність найвища і дорівнює 4; потім іде третя ознака із цінністю 2; найменш інформативна перша ознака, цінність якої 1.

Набір із індивідуально найкращих ознак, який дозволяють сформувати розглянуті вище методи, не обов'язково буде забезпечувати найякісніше розв'язання задачі розпізнавання. Щоб досягти останнього, потрібно розглядати і порівнювати набори ознак. Порівняння та вибір найкращого набору здійснюють на основі певного критерію. У задачі класифікації таким критерієм є показник якості класифікації, оцінений, наприклад, ковзним контролем (підрозд. 1.4), а в задачі кластеризації – один із функціоналів якості (підрозд. 2.5).

3.2. Формування набору інформативних ознак під час класифікації

Розглянуті нижче методи призначені для знаходження набору ознак, за яким якість класифікації даних буде найвища.

Насправді для знаходження оптимального набору з r ознак існує єдиний метод – повний перебір, який потребує перегляду C_p^r можливих наборів. Наприклад, щоб знайти оптимальний набір із 25 ознак, маючи 50 вихідних, потрібно переглянути близько 10^{15} наборів. Як правило, такий перебір занадто трудомісткий. Тому застосовують евристичні методи направленого перебору, які дозволяють за прийнятний час знаходити розв'язок, близький до оптимального. Розглянемо деякі з таких методів.

Скорочений пошук у ширину реалізує пошук інформативного набору з r ознак за r кроків. На першому кроці оцінюють якість розпізнавання за кожною ознакою окремо й обирають B ознак із найвищою якістю. Таким чином, формують B наборів, кожен із яких містить одну ознаку. Нехай на t-му кроці сформовано B наборів $V_1^{(1)},\ldots V_t^{(B)}$, до кожного з яких входить t ознак. Тоді на $(t+1)$-му кроці від кожного набору $V_1^{(1)},\ldots V_t^{(B)}$ одержують $(p-t)$ нових наборів шляхом приєднання однієї з ознак, яка набору ще не належить. Усього одержують $B(p-t)$ наборів, кожен із $(t+1)$-ї ознаки. Із них відбирають B кращих за якістю розпізнавання.

Число B є параметр методу, його називають шириною смуги. В окремих випадках на деяких кроках одержують менше ніж B наборів, наприклад, така

ситуація характерна для першого кроку, коли $p < B$. Тоді обирають усі одержані набори.

Окремий випадок попереднього методу, за якого $B=1$, відповідає **послідовному додаванню ознак** (Add). На початку шляхом перебору знаходять ознаку, якість розпізнавання за якою найвища. Далі до обраної ознаки додають по одній із тих, що залишилися. Серед усіх сформованих пар обирають таку, що забезпечує найвищу якість розпізнавання. До неї по черзі приєднують одну ознаку (із тих, що не входять у цю пару) і серед трійок обирають одну з найвищою якістю розпізнавання. Процедуру продовжують до формування набору з r ознак.

Логічною протилежністю попереднього методу є **метод послідовного вилучення ознак** (Del). Із вихідного набору по черзі вилучають по одній ознаці, тобто перебирають усі набори з $(p-1)$-ї ознаки. Серед них знаходять такий, що забезпечує найвищу якість розпізнавання. Далі почергово вилучають одну ознаку з обраного набору і визначають набір із $(p-2)$-х ознак, який має найкращу якість розпізнавання. Цю процедуру продовжують до одержання набору з r ознак.

Результати порівняння різних варіантів комбінацій методів Add та Del продемонстрували переваги методу AddDel перед Add, Del і DelAdd [8].

Послідовне додавання і вилучення ознак (AddDel). Спочатку методом Add формують набір із n_1 ознаки, потім n_2 $(n_2 < n_1)$ із них вилучають методом Del. Таким чином, одержують набір, до якого входять $n_1 - n_2$ ознак. Цей набір нарощують на n_1 ознаку методом Add, після чого знову запускають метод Del для вилучення n_2 найменш цінних ознак. У результаті формують набір із $2(n_1 - n_2)$ ознак. Чергування методів Add та Del продовжують до одержання набору з r ознак.

Інший підхід до вибору інформативного набору ознак полягає у застосуванні **випадкового пошуку**, який збільшує імовірність знаходження оптимального набору. Припускають, що імовірність входження кожної ознаки у шуканий набір однакова і становить $1/p$. Це означає, що кожній ознаці на одиничному відрізку відповідає частина довжиною $1/p$. Так, ознаці x_j відповідає відрізок $\left[(j-1)/p;\, j/p\right]$.

Формують L наборів з ir ознак. Щоб одержати один набір, r разів генерують псевдовипадкове число $\alpha \in [0;1]$ і перевіряють, до якого відрізка воно потрапило; якщо до відрізка $\left[(j-1)/p;\, j/p\right]$, то ознаку x_j включають до набору (при цьому, якщо ознака вже є в наборі, то генерують нове α).

Серед L генерованих наборів обирають той, що має найвищу якість, він і буде шуканим набором інформативних ознак. Чим більша величина L, тим більша імовірність знаходження оптимального набору, але тим ближчий буде метод до неефективного повного перебору.

Випадковий пошук можна вдосконалити, додавши процедури «заохочення» та «покарання» ознак. У такому разі буде мати місце **випадковий пошук із адаптацією**. На першому кроці всі ознаки вважають рівноймовірними, тобто за

кожною закріпляють частину відрізка $[0;1]$ довжиною $1/p$. Випадково формують L наборів із r ознак, як у методі випадкового пошуку. Серед них обирають один найкращий і один найгірший набори за якістю розпізнавання. Ознаки з найкращого набору «заохочують»: їх імовірності збільшують на величину f. Ознаки з найгіршого – «карають»: їх імовірності зменшують на f. Якщо якась ознака потрапила в обидва набори, то її імовірність у результаті одночасного застосування «заохочення» та «покарання» не зміниться.

На другому кроці знову формують L наборів із r ознак, але з урахуванням нових імовірностей ознак. Зміна імовірностей означає, що на одиничному відрізку ознакам відповідають частини різної довжини. Якщо $c_1, c_2, \ldots, c_p$ – це нові імовірності, то ознаці x_j відповідає відрізок $\left[\sum_{h=1}^{j-1} c_h; \sum_{h=1}^{j} c_h\right]$. Щоб згенерувати набір із r ознак, r разів виконують такі дії: генерують псевдовипадкове число $\alpha \in [0;1]$, якщо воно потрапляє до відрізка $\left[\sum_{h=1}^{j-1} c_h; \sum_{h=1}^{j} c_h\right]$, ознаку x_j включають до набору (якщо ця ознака вже є в наборі, генерують нове α). Із L одержаних наборів, як і на першому кроці, обирають найкращий та найгірший, ознаки з обраних наборів «заохочують» і «карають» відповідно.

Після певної кількості кроків сумарна імовірність деяких r ознак становитиме майже 1 (ці ознаки можна вважати інформативними). Імовірності ж інших $p-r$ ознак будуть близькими до нуля. Це призведе до того, що на кожному наступному кроці формуватимуться однакові набори ознак. У такому разі роботу методу можна завершити.

Чим більше L і менше f, тим вища імовірність знаходження оптимального набору, але в той же час вищі й часові витрати на пошук цього набору.

Ще один підхід полягає в **кластеризації ознак**. За допомогою методу кластеризації набір ознак розбивають на r кластерів. Із кожного кластера обирають по одному найтиповішому представнику, із яких і формують інформативний набір. Як функцію відстані між ознаками можна застосовувати коефіцієнт кореляції. Один із недоліків такого підходу пов’язаний із тим, що вихідних набір може містити шумові ознаки, які утворять окремий кластер і увійдуть до інформативного набору. Для усунення цього недоліку слід провести кластеризацію на $r' > r$ груп, сформувати набір із r' ознак, а потім шляхом перебору відібрати з них r. Інший недолік полягає в тому, що враховують лише парну схожість ознак, якої в загальному випадку недостатньо для виділення оптимального набору. Незважаючи на це, попередню кластеризацію доцільно застосовувати для скорочення процесу перебору в інших методах.

3.3. Формування набору інформативних ознак у ході кластеризації

Набір інформативних ознак, який найкраще відображає кластерну структуру даних, можна знайти на основі методів, розглянутих у попередньому підрозділі. Є лише декілька нюансів, які слід ураховувати.

По-перше, порівняння і вибір найкращого набору здійснюють на основі іншого критерію, яким є функціонал якості кластеризації (підрозд. 2.5).

По-друге, функціонал якості кластеризації не дає можливості порівнювати розбиття на наборах ознак різної розмірності. Наприклад, результати дослідження Дж. Ди та К.Бродлі [16] показали, що значення функціонала Q_1 (підрозд. 2.5) монотонно зростає зі збільшенням розмірності простору ознак, тобто для наборів ознак $X^{(1)}$ і $X^{(2)}$, таких, що $X^{(1)} \subset X^{(2)}$, функціонал Q_1 завжди буде надавати перевагу набору $X^{(2)}$. Для усунення цього недоліку автори дослідження запропонували таке. Нехай $X^{(1)}$ і $X^{(2)}$ – два набори ознак різної розмірності, $S^{(1)}$ і $S^{(2)}$ – розбиття, одержані відповідно на цих наборах, $Q(X,S)$ – функціонал якості. У такому разі рекомендовано обчислювати нормалізовані значення функціонала

$$\text{Normalized}Q\left(X^{(1)},S^{(1)}\right)=Q\left(X^{(1)},S^{(1)}\right)\cdot Q\left(X^{(2)},S^{(1)}\right),$$

$$\text{Normalized}Q\left(X^{(2)},S^{(2)}\right)=\text{Q}\left(X^{(2)},S^{(2)}\right)\cdot Q\left(X^{(1)},S^{(2)}\right).$$

Якщо $\text{Normalized}Q\left(X^{(1)},S^{(1)}\right) > \text{Normalized}Q\left(X^{(2)},S^{(2)}\right)$, то перевагу слід надавати набору ознак $X^{(1)}$, у разі тотожності нормалізованих значень функціоналів обирають набір меншої розмірності.

По-третє, оптимальний набір ознак залежить від кількості кластерів, яку попередньо необхідно оцінити. На рис. 3.2 можна бачити, що набір із двох ознак x_1, x_2 відображає наявність трьох кластерів, а якщо кількість кластерів обрати такою, що дорівнює двом, то оптимальний набір міститиме всього одну ознаку x_1.

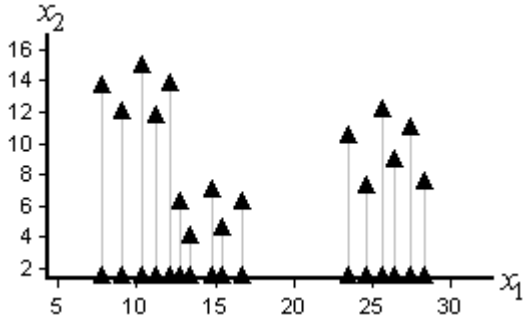


Рис. 3.2. Залежність оптимального набору ознак від кількості кластерів

Слід зазначити, що існують модифікації відомих методів кластеризації, які одночасно виконують розбиття об’єктів на кластери й формування оптимального набору ознак. Таким, наприклад, є метод *W*-*K*-середніх [17].

КОНТРОЛЬНІ ЗАПИТАННЯ ТА ЗАВДАННЯ

1. Що називають розпізнаванням образів?
2. Які задачі розв’язують у теорії розпізнавання образів?
3. Чим відрізняються задачі класифікації і кластеризації?
4. Із яких етапів складається розв’язання задачі розпізнавання образів?
5. Сформулюйте гіпотезу компактності.
6. У чому особливість метричних методів класифікації?
7. Дайте визначення поняття «функція відстані».
8. У чому суть навчання в методі ближнього сусіда та його модифікаціях?
9. У який спосіб обирають параметр k у методі k ближніх сусідів?
10. Що називають еталоном класу? Що може бути еталонами класів?
11. Дайте визначення поняття «дерево рішень». Як його будують?
12. Перелічіть критерії зупинки побудови дерева рішень.
13. Як будують вузол перевірки на основі кількісної (якісної) ознаки?
14. Назвіть критерій вибору ознаки для побудови чергового вузла, що застосовують в алгоритмі C4.5. Який функціонал покладено в основу цього критерію?
15. Як будують дерево рішень алгоритмом C4.5 за наявності пропусків у даних?
16. Яку гіперплощину вважають оптимальною в методі опорних векторів?
17. Які об’єкти називають опорними векторами?
18. За результатами розв’язання якої задачі знаходять оптимальну гіперплощину у випадку лінійно (не)відокремлювальних класів у SVM?
19. Які ядра можна застосовувати в методі опорних векторів для побудови нелінійних відокремлювальних поверхонь?
20. Що розуміють під якістю класифікації і як її оцінюють? У чому суть надійності класифікації?
21. У чому суть перенавчання? Як його можна виявити?
22. У чому полягає задача кластеризації?
23. Як і з якою метою стандартизують ознаки?
24. Перелічіть види ієрархічних методів. У чому їх суть?
25. Назвіть переваги і недоліки алгомеративних методів кластерного аналізу.
26. Дайте визначення поняття «дендрограма».
27. Які метрики мають властивість редуктивності? У чому її суть?
28. Назвіть різницю між методами K-середніх Мак-Кіна та Ллойда?
29. Які за формою кластери здатні виділяти методи FOREL?
30. У чому суть графових алгоритмів кластеризації? Які їх переваги і недоліки?
31. Дайте визначення поняття «коротший незамкнений шлях». Як його будують?
32. Як оцінюють якість кластеризації?
33. Які ознаки вважають інформативними?
34. Назвіть групи методів, застосовних для формування набору інформативних ознак.
35. Перелічіть проблеми, які можуть виникнути під час формування набору інформативних ознак для кластеризації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Айвазян, С.А. Классификация многомерных наблюдений [Текст] / С.А. Айвазян, З.И. Бежаева, О.В. Староверов. – М.: Статистика, 1974. – 240 с.

2. Айзерман, М.А. Метод потенциальных функций в теории обучения машин [Текст] / М.А. Айзерман, Э.М. Браверманн, Л.И Розоноэр. – М.: Наука, 1970. – 320 с.

3. Воронцов, К.В. Машинное обучение (курс лекций) [Электронный ресурс] / К.В. Воронцов. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций%2C_К.В.Воронцов). – Загл. с экрана.

4. Вятченин, Д.А. Нечеткие методы автоматическое классификации [Текст] / Д.А. Вятченин. – Мн.: УП «Технопринт», 2004. – 219 с.

5. Дуда, Р. Распознавание образов и анализ сцен [Текст] / Р. Дуда, П. Харт. – М.: Мир, 1976. – 512 с.

6. Дюран, Б. Кластерный анализ [Текст] / Б. Дюран, П. Оделл. – М.: Статистика, 1975. – 128 с.

7. Жамбю, М. Иерархический кластер-анализ и соответствия [Текст] / М. Жамбю. – М.: Финансы и статистика, 1988. – 342 с.

8. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний [Текст] / Н.Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. – 270 с.

9. Факторный, дискриминантный и кластерный анализ [Текст] / Дж.-О. Ким [и др.]. – М.: Финансы и статистика, 1989. – 215 с.

10. Кісельова, О.М. Системи та методи розпізнавання образів [Текст]: навч. посіб. / О.М. Кісельова, К.А. Кузнєцов, Л.С. Коряшкіна. – Д.: РВВ ДНУ, 2003. – 76 с.

11. Королев, В.Ю. ЕМ-алгоритм, его модификации и применение к задаче разделения смеси вероятностных распределений. Теоретический обзор [Текст]/ В.Ю. Королев. – М.: ИПИ РАН, 2007. – 94 с.

12. Мандель, И.Д. Кластерный анализ [Текст] / И.Д. Мандель. – М.: Финансы и статистика, 1988. – 176 с.

13. Миленький, А.В. Классификация сигналов в условиях неопределенности [Текст] / А.В. Миленький. – М.: Сов. радио, 1975. – 328 с.

14. Паклин, Н.Б. Бизнес-аналитика: от данных к знаниям [Текст]: учеб. пособие / Н.Б. Паклин, В.И. Орешков. – СПб: Питер, 2013. – 704 с.

15. Хайкин, С. Нейронные сети: полный курс [Текст] / С. Хайкин. – 2-е изд. – М.: Вильямс, 2008. – 1104 с.

16. Dy, J.G. Feature Selection for Unsupervised Learning [Text] / J.G. Dy, C.E. Brodley // J. of Machine Learning Research. – 2004. – Vol. 5. – P. 845–889.

17. Automated Variable Weighting in k-Means Type Clustering [Text] / J.Z. Huang, M.K. Ng, H. Rong, Z. Li // IEEE Transactions on Pattern Analysis and Machine Int. – 2005. – Vol. 27. – P. 657–668.

ЗМІСТ

Вступ 3
1. Класифікація даних 5
1.1. Метричні методи класифікації 6
1.1.1. Метрики, або функції, відстаней 6
1.1.2. Метод найближчого сусіда та його модифікації 7
1.1.3. Методи на основі порівняння з еталоном 10
1.1.4. Метод потенціальних функцій 13
1.2. Метод на основі дерева рішень 18
1.2.1. Алгоритм C4.5 20
1.2.2. Алгоритм CART 22
1.3. Метод опорних векторів 24
1.4. Оцінка якості класифікації 31
2. Кластерний аналіз даних 32
2.1. Ієрархічні методи кластеризації 33
2.2. Метод *K*-середніх 38
2.3. Методи FOREL 40
2.4. Графові методи кластеризації 43
2.5. Функціонали якості кластеризації 45
3. Формування набору інформативних ознак 47
3.1. Формування набору інформативних ознак за їх індивідуальною цінністю 48
3.2. Формування набору інформативних ознак під час класифікації 54
3.3. Формування набору інформативних ознак у ході кластеризації 57
Контрольні запитання та завдання 58
Список використаної літератури 59