Тема 6. Задачі Data Mining. Класифікація та кластеризація

План

1. Задачі та види класифікації.
2. Методи, що застосовуються для розв’язання задач класифікації.
3. Задача кластеризації.
4. Застосування кластерного аналізу.

Мета вивчення теми: вивчити задачі класифікації та кластеризації; засвоїти принцип штучної та природної класифікації.

Перелік ключових слів та понять із теми

Data Mining, класифікація, кластеризація, проста та складна класифікація, багатовимірна класифікація, штучна та природна класифікація

Теоретичні відомості з теми

1. Задачі та види класифікації.

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

Наведемо кілька визначень.

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

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

Класифікація вимагає дотримання таких правил:

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

Розрізняють:

1) **допоміжну (штучну) класифікацію**, яка виробляється за зовнішньою ознакою і служить для надання множині предметів (процесів, явищ) потрібного порядку;
2) **природну класифікацію**, яка виробляється за істотними ознаками, що характеризують внутрішню спільність предметів і явищ. Вона є результатом і

важливим засобом наукового дослідження, тому що передбачає і закріплює результати вивчення закономірностей об'єктів, що класифікуються.

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

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

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

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

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

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

Прикладом такої класифікації є періодична система хімічних елементів.

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

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

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

Маши́нне навчання — узагальнена назва штучної генерації знань із досвіду. Штучна система навчається на прикладах і після закінчення фази навчання може узагальнювати. Тобто система не просто вивчає наведені приклади, а розпізнає певні закономірності в даних для навчання.

Серед багатьох програмних продуктів варто згадати системи автоматичного діагностування, розпізнавання шахрайства з кредитними картками, аналіз ринку цінних паперів, класифікація ланцюжків ДНК, розпізнавання мовлення та тексту, автономні системи.

Практичне використання відбувається, переважно, за допомогою алгоритмів. Різноманітні алгоритми машинного навчання можна грубо поділити за такою схемою:

1. **Навчання з вчителем** (англ. Supervised learning): алгоритм вивчає функцію на основі наданих пар вхідних та вихідних даних. При цьому, в процесі навчання, «вчитель» вказує вірні вихідні дані для кожного значення вхідних даних. Одним із розділів навчання з вчителем є машинна класифікація. Такі алгоритми застосовуються для розпізнавання текстів.
2. **Навчання без вчителя** (англ. Unsupervised learning).
3. **Навчання із закріпленням** (англ. Reinforcement Learning): алгоритм навчається за допомогою тактики нагороди та покарання для максимізації вигоди для агентів (систем, до яких належить компонента, що навчається).

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

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

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

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

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

Розглянемо задачу класифікації на простому прикладі. Припустимо, є база даних про клієнтів туристичного агентства з інформацією про вік і

доходи за місяць. Є рекламний матеріал двох видів: більш дорогий і комфортний відпочинок та дешевший, молодіжний відпочинок. Відповідно, визначені два класи клієнтів: клас 1 і клас 2. База даних наведена в таблиці 6.1.

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

Таблиця 6.1 - База даних клієнтів туристичного агентства

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

Для наочності представимо нашу базу даних у двомірному просторі (вік і дохід), у вигляді множини об’єктів, що належать класам 1 (помаранчева мітка) і 2 (сіра мітка). На рис. 6.1 наведені об’єкти з двох класів.

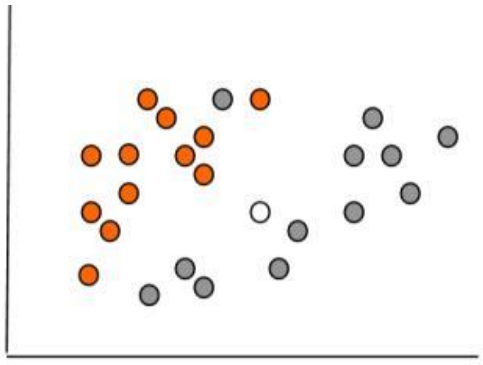


Рисунок 6.1 – Множина об’єктів бази даних у двомірному вимірі

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

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

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

Для проведення класифікації за допомогою математичних методів необхідно мати формальний опис об’єкта, яким можна оперувати,

використовуючи математичний апарат класифікації. Таким описом у нашому випадку виступає база даних. Кожен об'єкт (запис бази даних) несе інформацію про деякі властивості об'єкта.

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

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

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

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

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

1. Конструювання моделі: опис множини визначених класів.

Кожен приклад набору даних відноситься до одного визначеного класу.

На цьому етапі використовується навчальна множина, на ньому відбувається конструювання моделі.

Отримана модель **представлена класифікаційними правилами, деревом рішень або математичною формулою.**

2. Використання моделі: класифікація нових або невідомих значень.

Оцінка правильності (точності) моделі.

Відомі значення з тестового прикладу порівнюються з результатами використання отриманої моделі.

Рівень точності – відсоток правильно класифікованих прикладів у тестовій множині.

Тестова множина, тобто множина, на якій тестується побудована модель, не повинна залежати від навчальної множини.

Якщо точність моделі допустима, можливе використання моделі для класифікації нових прикладів, клас яких невідомий.

2. Методи, що застосовуються для розв'язання задач класифікації

Для класифікації використовуються різні методи. Основні з них:

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

Схематичне розв’язок задачі класифікації деякими методами (за допомогою лінійної регресії, дерев рішень і нейронних мереж) наведені на рис. 6.2- 6.4.

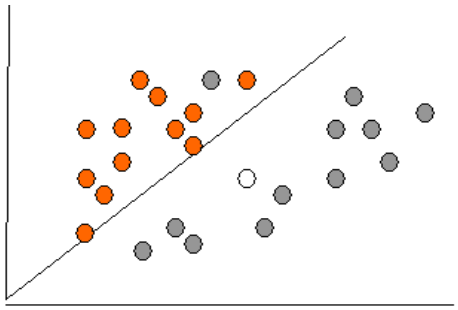


Рисунок 6.2 – Розв’язок задачі класифікації методом лінійної регресії

```
if X> 5 then grey
  else if Y> 3 then orange
          else if X> 2 then grey
                  else orange
```

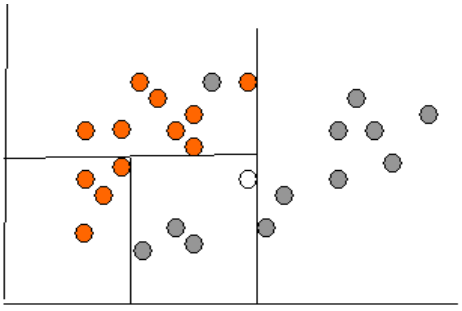


Рисунок 6.3 – Розв’язок задачі класифікації методом дерев рішень

Рисунок 6.4 – Розв’язок задачі класифікації методом нейронних мереж

Точність класифікації: оцінка рівня помилок. Оцінка точності класифікації може проводитися за допомогою крос-перевірки. Крос-перевірка (Cross-validation) – це процедура оцінки точності класифікації на даних із тестової множини, яку також називають крос-перевірочною множиною. Точність класифікації тестової множини порівнюється з точністю класифікації навчаючої множини. Якщо класифікація тестової множини дає

приблизно такі ж результати за точністю, як і класифікація навчальної множини, вважається, що дана модель пройшла крос-перевірку.

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

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

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

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

Інтерпретованість забезпечує можливість розуміння моделі аналітиком.

Властивості класифікаційних правил:

- розмір дерева рішень;
- компактність класифікаційних правил.

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

3. Задача кластеризації

Введемо поняття кластеризації, кластера, коротко розглянемо класи методів, за допомогою яких вирішується задача кластеризації, деякі моменти процесу кластеризації, а також розберемо приклади застосування кластерного аналізу.

Задача кластеризації схожа із задачею класифікації, є її логічним продовженням, але відмінність її в тому, що класи досліджуваного набору даних заздалегідь не зумовлені.

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

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

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

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

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

Характеристиками кластера можна назвати дві ознаки:

- внутрішня однорідність;
- зовнішня ізольованість.

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

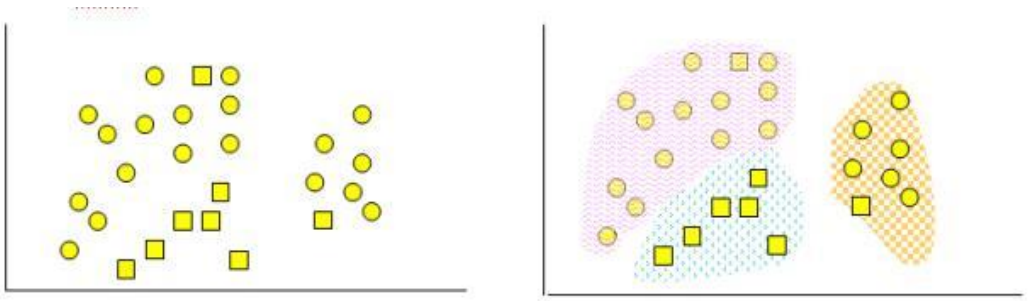
Найбільше застосування кластеризація спочатку отримала в таких науках як біологія, антропологія, психологія. Для вирішення економічних завдань кластеризація тривалий час мало використовувалася через специфіку економічних даних і явищ.

У таблиці 6.2 наведено порівняння деяких параметрів задач класифікації та кластеризації.

Таблиця 6.2 - Порівняння класифікації та кластерізаціі

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

На рисунку 6.5 схематично представлені задачі класифікації і кластеризації.



а) класифікація б) кластерізація

Рисунок 6.5 – Порівняння задач класифікації та кластеризації

Кластери можуть бути **такими, що не перетинаються**, або **ексклюзивними** (non-overlapping, exclusive), і такими, що **перетинаються** (overlapping). Схематичне зображення таких кластерів дано на рисунку 6.6.

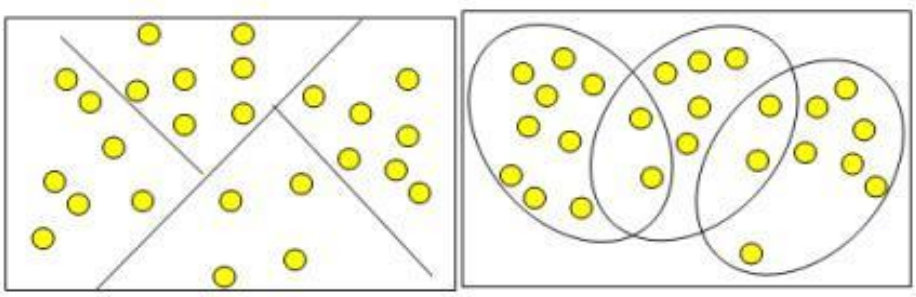


Рисунок 6.6 – Кластери, що не перетинаються і перетинаються

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

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

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

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

Ці особливості слід враховувати при виборі методу кластеризації. На цей час розроблено більше сотні різних алгоритмів кластеризації.

Наведемо коротку характеристику підходів до кластеризації.

– *Алгоритми, засновані на поділі даних (Partitioning algorithms), у тому числі ітеративні:*

- поділ об’єктів на *k* кластерів;
- ітеративний перерозподіл об’єктів для поліпшення кластеризації.

– *Ієрархічні алгоритми (Hierarchy Algorithms):*

- агломерація: кожен об’єкт спочатку є кластером, кластери, з’єднуючись один з одним, формують більший кластер і т.д.

– *Методи, засновані на концентрації об’єктів (Density-based methods):*

- засновані на можливості з’єднання об’єктів;
- ігнорують шуми, знаходження кластерів довільної форми.

– *Грід-методи (Grid-based methods):*

- квантування об’єктів в грід-структури.

– *Модельні методи (Model-based):*

- використання моделі для знаходження кластерів, найбільш відповідних даним.

Оцінка якості кластеризації може бути проведена на основі таких процедур:

- ручна перевірка;
- встановлення контрольних точок та перевірка на отриманих кластерах;

- визначення стабільності кластеризації шляхом додавання в модель нових змінних;
- створення і порівняння кластерів із використанням різних методів. Різні методи кластеризації можуть створювати різні кластери, і це є нормальним явищем. Однак створення схожих кластерів різними методами вказує на правильність кластеризації.

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

4. Застосування кластерного аналізу

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

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

Питання для самоконтролю

1. Назвіть задачі класифікації.
2. Яких правил потрібно дотримуватися при класифікації?
3. Які види класифікацій ви знаєте?
4. Які існують алгоритми машинного навчання? У чому їх відмінність?
5. Проведіть порівняння класифікації та кластерізації.
6. Як проходить оцінка якості кластеризації?