1. Завдання Data Mining

В основу технології Data Mining покладена концепція шаблонів, що представляють собою закономірності. В результаті виявлення цих, прихованих від неозброєного ока закономірностей вирішуються завдання інтелектуального аналізу даних. Різним типам закономірностей, які можуть бути виражені у формі, зрозумілій людині, відповідають певні завдання інтелектуального аналізу даних.

Завдання (tasks) Data Mining іноді називають закономірностями (regularity) або техніками (techniques).

Єдиної думки щодо того, які завдання слід відносити до Data Mining, немає.

Більшість авторитетних джерел перераховують наступні завдання: класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз і виявлення відхилень, оцінювання, аналіз зв'язків, підведення підсумків.

Класифікація (Classification). Найбільш проста і поширена задача Data Mining. В результаті рішення задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних - класи; за цими ознаками новий об'єкт можна віднести до того чи іншого класу.

Методи рішення. Для вирішення завдання класифікації можуть використовуватися методи: найближчого сусіда (Nearest Neighbor), K-найближчого сусіда (k-Nearest Neighbor); байєсовські мережі (Bayesian Networks); індукція дерев рішень ; нейронні мережі (neural networks).

Кластеризація (Clustering) є логічним продовженням ідеї класифікації. Це завдання більш складне, особливість кластеризації полягає в тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи.

Приклад методу розв'язання задачі кластеризації: навчання "без вчителя" особливого виду нейронних мереж - самоорганізованих карт Кохонена.

Асоціація (Associations). У ході вирішення задачі пошуку асоціативних правил відшукуються закономірності між пов'язаними подіями в наборі даних.

Відмінність асоціації від двох попередніх завдань Data Mining: пошук закономірностей здійснюється не на основі властивостей аналізованого об'єкта, а між кількома подіями, які відбуваються одночасно.

Найбільш відомий алгоритм вирішення задачі пошуку асоціативних правил - алгоритм Apriori.

Послідовність (Sequence), або послідовна асоціація (sequential association). Послідовність дозволяє знайти тимчасові закономірності між транзакціями. Задача послідовності подібна асоціації, але її метою є встановлення закономірностей не між одночасно наступаючими подіями, а між подіями, пов'язаними в часі (тобто відбуваються з деяким певним інтервалом у часі). Іншими словами, послідовність визначається високою ймовірністю ланцюжка пов'язаних у часі подій.

Фактично, асоціація є окремим випадком послідовності з тимчасовим лагом, рівним нулю. Цю задачу Data Mining також називають задачею знаходження послідовних шаблонів (sequential pattern).

Правило послідовності: після події X через певний час відбудеться подія Y.

Приклад. Після покупки квартири мешканці в 60% випадків протягом двох тижнів купують холодильник, а протягом двох місяців в 50% випадків купується телевізор. Рішення даної задачі широко застосовується в маркетингу і менеджменті, наприклад, при управлінні циклом роботи з клієнтом (управління життєвим циклом клієнта).

Прогнозування (Forecasting). В результаті рішення задачі прогнозування на основі особливостей історичних даних оцінюються пропущені або ж майбутні значення цільових чисельних показників.

Для вирішення таких завдань широко застосовуються методи математичної статистики, нейронні мережі та ін.

Визначення відхилень або викидів (Deviation Detection), аналіз відхилень або викидів. Мета розв’язання даної задачі - виявлення та аналіз даних, що найбільш відрізняються від загальної множини даних, виявлення так званих нехарактерних шаблонів.

Оцінювання (оцінка). Завдання оцінювання зводиться до передбачення безперервних значень ознаки.

Аналіз зв'язків (Link Analysis) - задача знаходження залежностей в наборі даних.

Візуалізація (Visualization, Graph Mining). В результаті візуалізації створюється графічний образ аналізованих даних. Для вирішення задачі візуалізації використовуються графічні методи, що показують наявність закономірностей в даних.

Приклад методів візуалізації - подання даних у 2-D і 3-D вимірах.

Підведення підсумків (Summarization) - завдання, мета якого - опис конкретних груп об'єктів з аналізованого набору даних.

2. Задачі інтелектуального аналізу даних

Згідно класифікації за стратегіями, задачі Data Mining поділяються на такі групи:

- навчання з учителем;
- навчання без вчителя;
- інші.

Категорія навчання з учителем представлена наступними задачами Data Mining: класифікація, оцінка, прогнозування.

Категорія навчання без вчителя представлена задачею кластеризації.

У категорію «інші» входять задачі, не включені в попередні дві стратегії.

Задачі інтелектуального аналізу даних, залежно від використовуваних моделей, можуть бути **дескриптивними** і **прогнозуючими.**

Відповідно до цієї класифікації, **задачі Data Mining** представлені **групами описових і прогнозуючих завдань.**

У результаті рішення описових (descriptive) задач аналітик отримує шаблони, що описують дані, які піддаються інтерпретації.

Ці задачі описують загальну концепцію аналізованих даних, визначають інформативні, підсумкові, відмінні особливості даних. Концепція описових завдань передбачає характеристику і порівняння наборів даних.

Характеристика набору даних забезпечує короткий і стислий опис деякого набору даних.

Порівняння забезпечує порівняльний опис двох або більше наборів даних.

Прогнозуючі (predictive) ґрунтуються на аналізі даних, створенні моделі, передбаченні тенденцій або властивостей нових або невідомих даних.

Досить близьким до вищезгаданої класифікації є розділення задач Data Mining на наступні:

a. дослідження та відкриття;

b. прогнозування та класифікація;

c. пояснення і опис.

Автоматичне дослідження і відкриття (вільний пошук). *Приклад задачі: виявлення нових сегментів ринку.*

Для вирішення даного класу задач використовуються методи кластерного аналізу прогнозування та класифікація.

Приклад задачі: передбачення зростання обсягів продажів на основі поточних значень.

Методи: регресія, нейронні мережі, генетичні алгоритми, дерева рішень.

Задачі класифікації та прогнозування становлять групу так званого індуктивного моделювання, в результаті якого забезпечується вивчення аналізованого об'єкта або системи. У процесі вирішення цих завдань на основі набору даних розробляється загальна модель або гіпотеза.

Пояснення й опис. *Приклад задачі: характеристика клієнтів за демографічними даними і історіями покупок.*

Методи: дерева рішень, системи правил, правила асоціації, аналіз зв'язків.

Якщо дохід клієнта більше, ніж 50 умовних одиниць, і його вік - понад 30 років, тоді клас клієнта - перший.

В інтерпретації узагальненої моделі аналітик отримує нове знання. Групування об'єктів відбувається на основі їх подібності.

Отже, у попередній лекції нами були розглянуті методи Data Mining і дії, що виконуються в рамках стадій інтелектуального аналізу даних. Щойно ми розглянули основні завдання інтелектуального аналізу даних.

Нагадаємо, що головна цінність Data Mining - це практична спрямованість даної технології, шлях від сирих даних до конкретного знання, від постановки завдання до готового додатку, за підтримки якого можна приймати рішення.

Велика кількість понять, які об'єдналися в Data Mining, а також різноманітність методів, що підтримують дану технологію, починаючому аналітику можуть нагадати мозаїку, частини якої мало пов'язані між собою.

Як же ми можемо зв'язати в одне ціле задачі, методи, дії, закономірності, додатки, дані, інформацію, рішення?

Розглянемо два потоки:

1. Дані - інформація - знання і рішення
2. Завдання - дії і методи рішення - програми

Ці потоки є "двома сторонами однієї медалі", відображенням одного процесу, результатом якого має бути знання і прийняття рішення.

Від даних до рішень. Для початку розглянемо перший потік. На рис. 3.1. показаний зв’язок понять «дані», «інформація» і «рішення», яка виникає в процесі прийняття рішень.

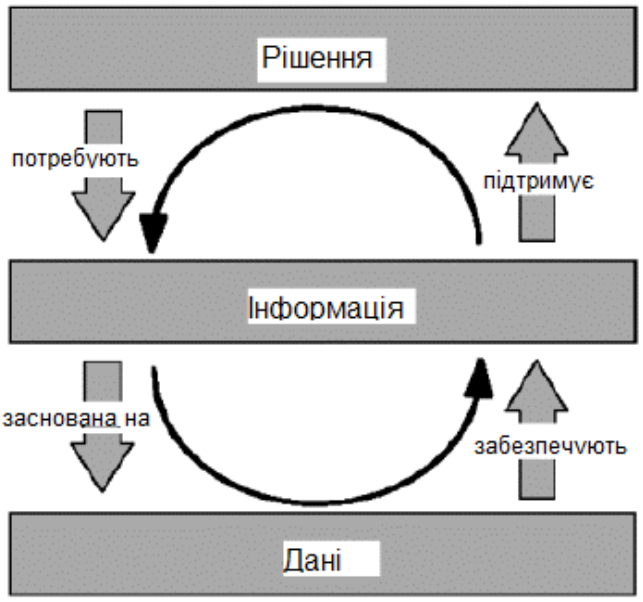

Рисунок 3.1 – Рішення, інформація і дані

Як видно з рисунку, даний процес являється циклічним. Прийняття рішень потребує інформації, яка заснована на даних. Дані забезпечують інформацію, яка підтримує рішення і т.д.

Розглянуті поняття є складовою частиною так званої інформаційної піраміди, в основі якої знаходяться дані, наступний рівень - це інформація, потім йде рішення, завершує піраміду рівень знання. У міру просування вгору по інформаційній піраміді обсяги даних переходять у цінність рішень, тобто цінність для бізнесу. А, як відомо, метою Business Intelligence є перетворення обсягів даних у цінність бізнесу.

Тепер підійдемо до цього ж процесу з іншого боку. Розглянемо рис. 3.2.

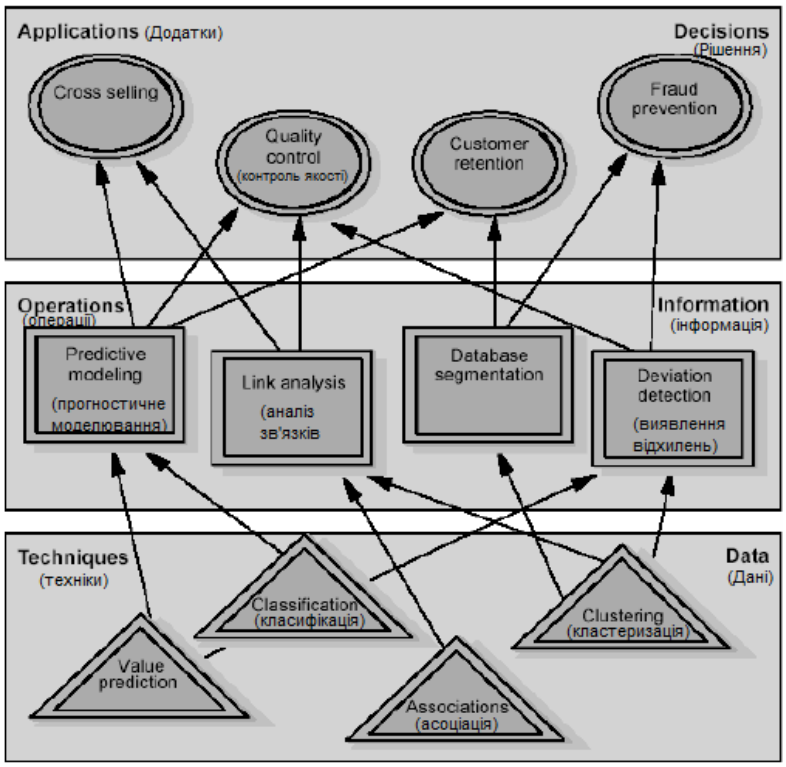

Рисунок 3.2 – Задачі, дії, додатки

Слід зазначити, що рівні аналізу (дані, інформація, знання) практично відповідають етапам еволюції аналізу даних, яка відбувалася протягом останніх років.

3. Рівні аналізу

Верхній - рівень додатків - є рівнем бізнесу (якщо ми маємо справу із завданням бізнесу), на ньому менеджери приймають рішення. Наведені приклади додатків: перехресні продажі, контроль якості, утримування клієнтів.

Середній - рівень дій - за своєю суттю є рівнем інформації, саме на ньому виконуються дії Data Mining; на рисунку наведені такі дії: прогностичне моделювання, аналіз зв'язків, сегментація даних та інші.

Нижній - рівень визначення задачі інтелектуального аналізу даних, яку необхідно розв’язати стосовно даних, що є в наявності, на малюнку наведені завдання передбачення числових значень, класифікація, кластеризація, асоціація.

Розглянемо таблицю, що демонструє зв'язок цих понять.

Табл.3.1. Рівні Data Mining

рівень 3	Додатки	утримання клієнтів	знання DM	результат
рівень 2	Дії	прогностичне моделювання	інформація	метод аналізу
рівень 1	Задачі	Класифікація	дані	запити

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

Розглянемо задачі утримання клієнтів (визначення надійності клієнтів фірми).

Дані - база даних по клієнтах. Є дані про клієнта (вік, стать, професія, дохід). Певна частина клієнтів, скориставшись продуктом фірми, залишилася їй вірна; інші клієнти більше не купували продукти фірми. На цьому рівні ми визначаємо тип завдання - це завдання класифікації.

На другому рівні визначаємо дію - прогностичне моделювання. За допомогою прогностичного моделювання ми з певною частиною впевненості можемо віднести новий об'єкт, в даному випадку, нового клієнта, до одного з відомих класів - постійний клієнт, або це, швидше за все, його разова покупка.

На третьому рівні ми можемо скористатися додатком для прийняття рішення. У результаті придбання знань, фірма може істотно знизити витрати, наприклад, на рекламу, знаючи заздалегідь, яким із клієнтів слід активно розсилати рекламні матеріали.

Таким чином, протягом кількох лекцій ми визначилися з поняттями "дані", "завдання", " методи", "дії ".

Зараз зупинимося на ще не розглянутому понятті інформації. Незважаючи на поширеність даного поняття, ми не завжди можемо точно його визначити і відрізнити від поняття даних. Інформація, за своєю суттю, має багатогранну природу. З розвитком людства, в тому числі, з розвитком комп'ютерних технологій, інформація набуває все нових і нових властивостей.

4. Задачі класифікації

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

Наведемо кілька визначень.

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

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

Класифікація вимагає дотримання наступних правил:

a. в кожному акті ділення необхідно застосовувати тільки одну основу;

b. ділення повинне бути пропорційним, тобто загальний обсяг видових понять повинен дорівнювати об’єму діленого родового поняття;

c. члени ділення повинні взаємно виключати один одного, їх об’єми не повинні перехрещуватися;

d. ділення повинне бути послідовним.

Розрізняють:

a. **допоміжну (штучну) класифікацію**, яка виробляється за зовнішньою ознакою і служить для надання множині предметів (процесів, явищ) потрібного порядку;

b. **природну класифікацію**, яка виробляється за істотними ознаками, що характеризують внутрішню спільність предметів і явищ. Вона є результатом і важливим засобом наукового дослідження, тому що передбачає і закріплює результати вивчення закономірностей об'єктів, що класифікуються.

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

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

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

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

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

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

Прикладом такої класифікації є періодична система хімічних елементів.

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

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

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

Маши́нне навчання — узагальнена назва штучної генерації знань з досвіду. Штучна система 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:

- **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 з розділів навчання з вчителем є машинна класифікація. Такі алгоритми застосовуються для розпізнавання текстів.
- **Навчання без вчител**я (англ. Unsupervised learning).
- **Навчання з закріпленням** (англ. Reinforcement Learning): алгоритм навчається за допомогою тактики нагороди та покарання для максимізації вигоди для агентів (систем до яких належить компонента, що навчається)

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

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

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

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

Багатовимірна класифікація була розроблена біологами при вирішенні проблем дискримінації для класифікування організмів. Однією з

перших робіт, присвячених цьому напрямку, вважають роботу Р. Фішера (1930 р.), в якій організми поділялися на підвиди залежно від результатів вимірювань їх фізичних параметрів. Біологія була і залишається найбільш затребуваним і зручним середовищем для розробки багатовимірних методів класифікації.

Розглянемо задачу класифікації на простому прикладі. Припустимо, є база даних про клієнтів туристичного агентства з інформацією про вік і доході за місяць. Є рекламний матеріал двох видів: більш дорогий і комфортний відпочинок і дешевший, молодіжний відпочинок. Відповідно, визначені два класи клієнтів:клас 1 і клас 2. База даних наведена в таблиці 3.2.

Таблиця 3.2. База даних клієнтів туристичного агентства.

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

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

Для наочності представимо нашу базу даних у двомірному просторі (вік і дохід), у вигляді множини об'єктів, що належать класам 1 (помаранчева мітка) і 2 (сіра мітка). На рис. 5.1 наведені об'єкти з двох класів.

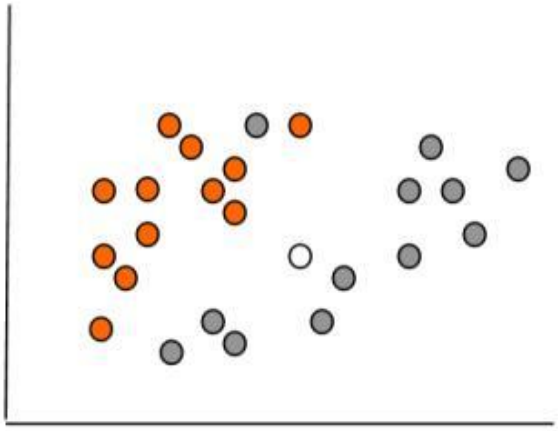

Рисунок 3.3. Множина об'єктів бази даних у двомірному вимірі

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

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

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

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

Набір вихідних даних (або вибірку даних) розбивають на **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.

Процес класифікації, а саме, конструювання моделі та її використання, представлений на рис. 3.4. – рис.3.5.

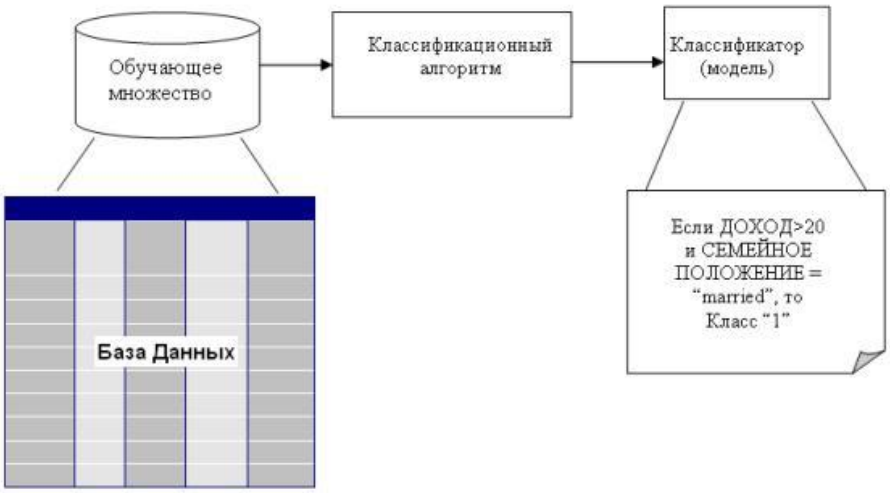

Рисунок 3.4 – Процес класифікації. Конструювання моделі

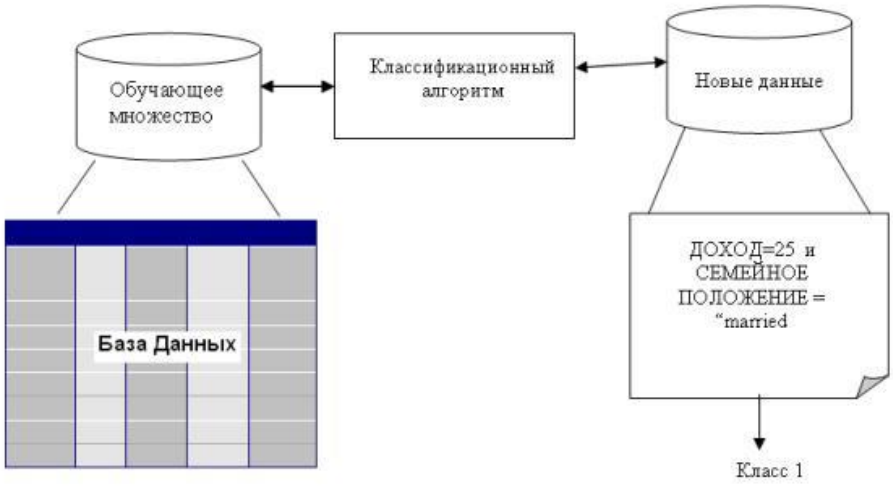

Рисунок 3.5 – Процес класифікації. Використання моделі

5. Методи, що застосовуються для вирішення задач класифікації

Для класифікації використовуються різні методи. Основні з них:

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

Схематичне рішення задачі класифікації деякими методами (за допомогою лінійної регресії, дерев рішень і нейронних мереж) наведені на рис.3.6- 3.7.

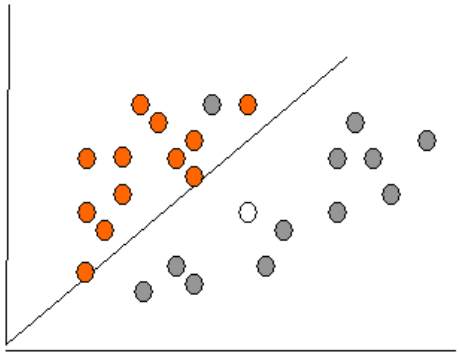

Рисунок 3.6 – Рішення задачі класифікації методом лінійної регресії

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

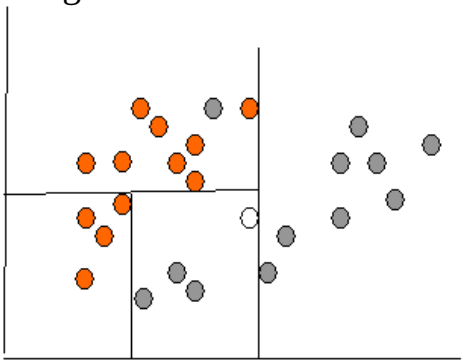

Рисунок 3.7 – Рішення задачі класифікації методом дерев рішень

Рисунок 3.8 – Рішення задачі класифікації методом нейронних мереж

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

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

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

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

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

Інтерпретованість забезпечує можливість розуміння моделі аналітиком.

Властивості класифікаційних правил:

- розмір дерева рішень;
- компактність класифікаційних правил.

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

6.Завдання кластеризації

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

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

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

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

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

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

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

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

Характеристиками кластера можна назвати дві ознаки:

- внутрішня однорідність;
- зовнішня ізольованість.

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

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

У таблиці 3.3 наведено порівняння деяких параметрів задач класифікації та кластеризації.

Таблиця 3.3. Порівняння класифікації та кластерзації

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 3.10 схематично представлені завдання класифікації і кластеризації.

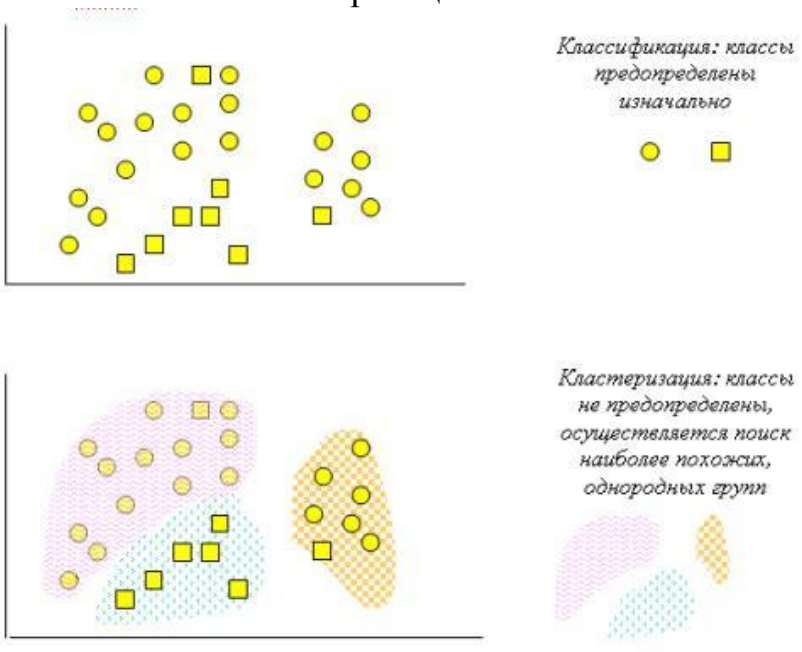

Рисунок 3.10 – Порівняння задач класифікації та кластеризації

Кластери можуть бути **такими, що не перетинаються**, або **ексклюзивними** (non-overlapping, exclusive), і такими, що **перетинаються** (overlapping). Схематичне зображення таких кластерів дано на рисунку 3.11.

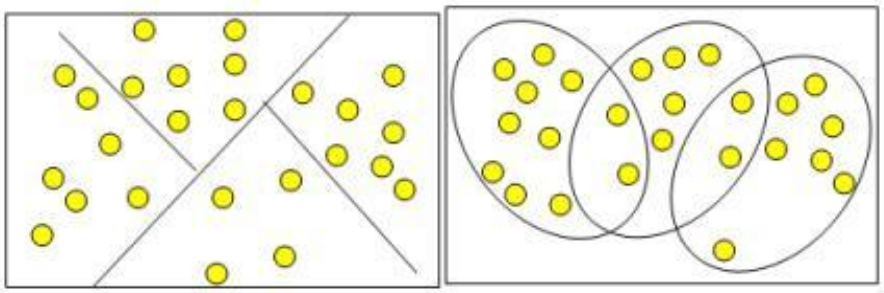

Рисунок 3.11– Кластери, що не перетинаються і перетинаються

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

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

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

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

Дані особливості слід враховувати при виборі методу кластеризації.

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

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

Наведемо коротку характеристику підходів до кластеризації.

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

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

– *Ієрархічні алгоритми (Hierarchy Algorithms):*

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

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

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

– *Грід-методи (Grid-based methods):*

- квантування об'єктів в грід-структури.

– *Модельні методи (Model-based):*

- використання моделі для знаходження кластерів, найбільш відповідних даним.

Оцінка якості кластеризації може бути проведена на основі таких процедур:

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

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