Тема 4. Етапи ІАД. Класифікація методів ІАД

План

1. Класифікація стадій Data Mining.
2. Класифікація технологічних методів Data Mining.
3. Властивості методів Data Mining.

Мета вивчення теми: вивчити стадії та методи Data Mining; засвоїти властивості методів інтелектуального аналізу даних.

Перелік ключових слів та понять із теми

Data Mining, метод, штучні нейронні мережі, дерева рішень, символьні правила, методи найближчого сусіда і k-найближчого сусіда, метод опорних векторів, байєсовські мережі, лінійна регресія, кореляційно-регресійний аналіз, ієрархічні методи кластерного аналізу, неієрархічні методи кластерного аналізу

Теоретичні відомості з теми

1. Класифікація стадій Data Mining

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

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

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

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

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

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

Data Mining може складатися з двох або трьох стадій:

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

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

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

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

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

ВІЛЬНИЙ ПОШУК (у тому числі ВАЛІДАЦІЯ) ->
-> **ПРОГНОСТИЧНЕ МОДЕЛЮВАННЯ** ->
-> **АНАЛІЗ ВИНЯТКІВ**

1. Вільний пошук (Discovery).

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

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

Вільний пошук представлений такими **діями**:

• виявлення закономірностей умовної логіки (conditional logic);

• виявлення закономірностей асоціативної логіки (associations and affinities);

• виявлення трендів і коливань (trends and variations).

Припустимо, є база даних кадрового агенства з даними про професії, стаж, вік і бажаний рівень винагороди. У разі самостійного задавання

запитів аналітик може отримати приблизно такі результати: середній бажаний рівень винагороди фахівців у віці від 25 до 35 років дорівнює 1200 умовних одиниць. У разі вільного пошуку система сама шукає закономірності, необхідно лише задати цільову змінну. У результаті пошуку закономірностей система сформує набір логічних правил " якщо..., то...".

Можуть бути знайдені, наприклад, такі закономірності

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

або

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

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

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

• індукції правил умовної логіки (задачі класифікації та кластеризації, опис у компактній формі близьких або схожих груп об'єктів);

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

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

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

2. Прогностичне моделювання (Predictive Modeling).

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

Прогностичне моделювання включає такі дії:

• передбачення невідомих значень (outcome prediction);

• прогнозування розвитку процесів (forecasting).

У процесі прогностичного моделювання розв'язуються задачі класифікації та прогнозування.

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

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

Продовжуючи розглянутий приклад першої стадії, можемо зробити такий висновок.

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

Порівняємо вільний пошук і прогностичне моделювання з точки зору логіки.

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

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

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

• знання класу, до якого належать досліджувані об'єкти;

• знання загального правила, діючого в межах даного класу об'єктів.

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

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

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

3. Аналіз винятків (forensic analysis).

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

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

Повернемося до одного з прикладів, розглянутих вище.

Знайдено правило "Якщо вік >35 років і бажаний рівень винагороди >1200 умовних одиниць, то в 90% випадків здобувач шукає керівну роботу". Виникає питання – до чого віднести решту 10% випадків? Тут можливі два варіанти. Перший з них – існує деяке логічне пояснення, яке також може бути оформлено у вигляді правила. Другий варіант для решти 10% – це

помилки вихідних даних. У цьому випадку стадія аналізу виняткiв може бути використана для очищення даних.

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

2. Класифікація технологічних методів Data Mining

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

1. Безпосереднє використання даних, або збереження даних.

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

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

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

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

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

Логічні методи, або методи логічної індукції, включають:

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

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

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

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

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

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

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

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

У цій класифікації розрізняють дві групи методів:

• *статистичні методи*, засновані на використанні усередненого накопиченого досвіду, який відображений у ретроспективних даних;

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

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

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

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

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

• виявлення зв'язків і закономірностей (лінійний і нелінійний регресійний аналіз, кореляційний аналіз та ін.);

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

• динамічні моделі і прогноз на основі часових рядів.

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

До цієї групи відносяться такі методи:

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

• еволюційне програмування (у т.ч. алгоритми методу групового обліку аргументів);

• генетичні алгоритми (оптимізація);

• асоціативна пам'ять (пошук аналогів, прототипів);

• нечітка логіка;

• дерева рішень;

• системи обробки експертних знань.

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

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

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

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

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

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

3. Властивості методів Data Mining

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

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

Масштабованість – властивість обчислювальної системи, яка забезпечує передбачуваний зріст системних характеристик, наприклад, швидкості реакції, загальної продуктивності тощо, при додаванні до неї обчислювальних ресурсів.

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

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

Основні методи. Кілька основних методів, які використовуються для інтелектуального аналізу даних, описують тип аналізу й операцію з відновлення даних.

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

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

Створити інструменти інтелектуального аналізу даних на базі асоціацій або відносин неважко. Наприклад, в InfoSphere Warehouse є майстер, який видає конфігурації інформаційних потоків для створення асоціацій, досліджуючи джерело вхідної інформації, базис прийняття рішень і вихідну інформацію. На рис 4.1 наведено відповідний приклад бази даних.

Класифікацію можна використовувати для отримання уявлення про тип покупців, товарів або об’єктів, описуючи кілька атрибутів для ідентифікації певного класу. Наприклад, автомобілі легко класифікувати за типом (седан,

позашляховик, кабріолет), визначивши різні атрибути (кількість місць, форма кузова, ведучі колеса). Вивчаючи новий автомобіль, можна віднести його до певного класу, порівнюючи атрибути з відомим визначенням. Ті ж принципи можна застосувати і до покупців, наприклад, класифікуючи їх за віком та соціальною групою.

Крім того, класифікацію можна використовувати як вхідні дані для інших методів. Наприклад, для визначення класифікації можна застосовувати дерева прийняття рішень. Кластеризація дозволяє використовувати загальні атрибути різних класифікацій з метою виявлення кластерів.

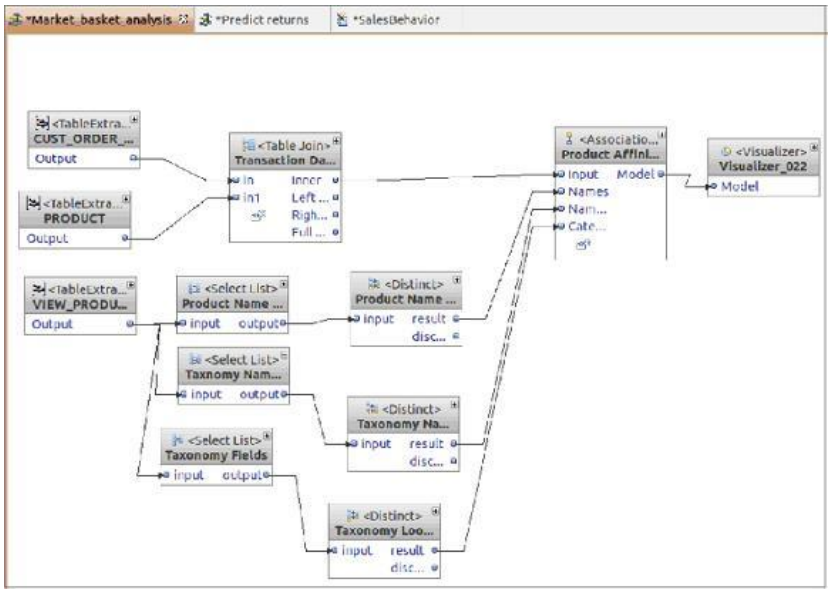

Рисунок 4.1 – Інформаційний потік, який використовується при застосуванні методу асоціації

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

Метод кластеризації працює в обидві сторони. Можна припустити, що в певній точці існує кластер, а потім використовувати свої критерії ідентифікації, щоб перевірити це. Графік, зображений на рис. 4.2, – наочний приклад. Тут вік покупця порівнюється з вартістю покупки. Розумно очікувати, що люди у віці від двадцяти до тридцяти років (до вступу в шлюб і появи дітей), а також в 50-60 років (коли діти покинули будинок) мають більш високий наявний дохід.

У цьому прикладі видно два кластери, один в діапазоні $ 2000/20-30 років та інший в діапазоні $7000-8000/50-65 років. У цьому випадку ми висунули гіпотезу і перевірили її на простому графіку, який можна

побудувати за допомогою будь-якого відповідного програмного забезпечення для побудови графіків. Для більш складних комбінацій потрібен повний аналітичний пакет, особливо якщо потрібно автоматично засновувати рішення на інформації про найближчого сусіда.

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

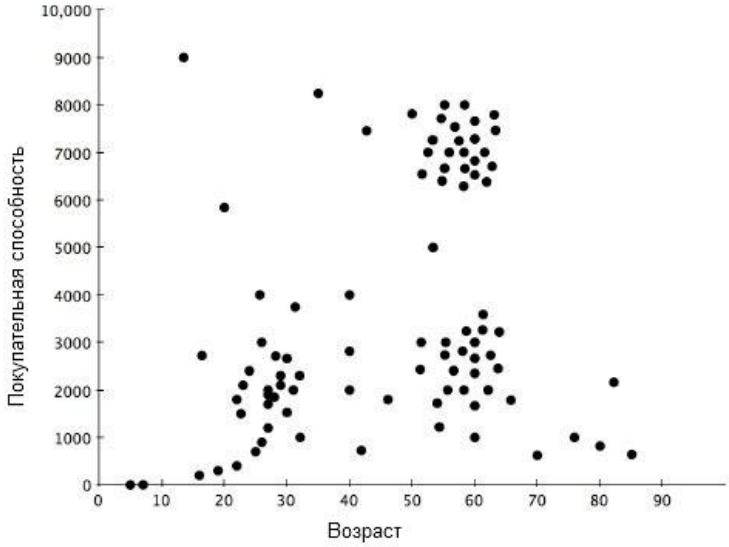

Рисунок 4.2 – Кластеризація

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

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

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

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

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

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

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

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

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

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

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

Підготовка даних і очищення даних – надзвичайно важливий крок у процесі «видобутку даних». У типових проектах «видобутку даних» великі набори даних, зібрані за допомогою деяких автоматичних методів (наприклад, за допомогою Web), служать вхідними даними аналізу. Часто

метод, за допомогою якого були зібрані дані, не був жорстко регульованим, внаслідок чого дані можуть містити значення, що виходять за допустимі межі (наприклад, Дохід: -100), неможливі комбінації даних (наприклад, Пол: Чоловік, Вагітність: Так) тощо.

При видобутку даних вхідні дані часто «зачумлені» – містять багато помилок та, іноді, інформацію в неструктурованою формі. Припустимо, що ви хочете проаналізувати велику базу даних, зібраних за допомогою Web у режимі он-лайн, ґрунтуючись на добровільних відповідях людей, які відвідують ваш Web-сайт (наприклад, потенційних клієнтів Web-продавця, який заповнив запропоновані анкети). У цьому прикладі дуже важливо спочатку перевірити і «очистити» дані на стадії підготовки даних, перед тим як застосовувати аналітичні процедури. Наприклад, деякі індивідууми можуть ввести завідомо помилкову інформацію (наприклад, вік = 300). У таких типах даних помилки не виявляються до стадії аналізу. Вони можуть сильно зміщувати результат і приводити до невиправданих висновків. зазвичай протягом стадії підготовки даних аналітик застосовує «фільтри» до даних для перевірки правильності їх діапазонів і виключення неможливих значень (наприклад, Вік = 5; Пенсіонер = Так).

Питання для самоконтролю

1. Дайте визначення методу та алгоритму.
2. Які ви знаєте методи і алгоритми Data Mining?
3. Із яких стадій складається процес Data Mining?
4. Які ви знаєте види класифікацій технологічних методів Data Mining?
5. Приведіть приклади застосування інструменту Data Mining асоціації (або відношення)?
6. Для чого застосовуються інструментарій Data Mining дерева рішень (наведіть приклад)?