Тема №7 КЛАСИЧНІ ТЕХНОЛОГІЇ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

План

7.1. Класичні технології класифікації в Data Mining
7.2. Програмне забезпечення задач класифікації
7.3. Класичні технології кластеризації в Data Mining
7.4. Програмне забезпечення задач кластеризації

Література:

Основна

Черняк О. І., Захарченко П. В. Інтелектуальний аналіз даних: Підручник. Київ, 2014.

Додаткова

1. Айвазян С. А., Бухштабер В. М., Енюков И. С. Прикладная статистика: классификация и снижение размерности. Москва : Финансы и статистика, 1989.
2. Бююль А., Цефель П. SPSS: искусство обработки информации.– Санкт-Петербург : ДиаСофт, 2001.
3. Гладков Л. А., Курейчик В. В., Курейчик В. М. Генетические алгоритмы. Москва : Физматлит, 2006.
4. Елманова Н., Федоров А. Введение в OLAP-технологии Microsoft. Москва : Диалог-МИФИ, 2002.
5. Жамбю М. Иерархический кластер–анализ и соответствия. Москва : Финансы и статистика, 1988.
6. Кацко И. А., Паклин Н. Б. Практикум по анализу данных на компьютере. Москва : Колосс, 2009.
7. Мандель И. Д. Кластерный анализ. Москва : Финансы и Статистика, 1988.
8. Недосекин А. О. Нечетко-множественный анализ фондовых инвестиций. Санкт-Петербург: Сезам, 2002.
9. Орлов А. И. Нечисловая статистика. Москва : МЗ-Пресс, 2004.

7.1. Класичні технології класифікації в Data Mining

Класифікація є найбільш простою і одночасно найбільш часто вирішуваною задачею *Data Mining*. Зважаючи на поширеність задач класифікації необхідне чітке розуміння суті цього поняття. Наведемо декілька означень.

Означення 1. Класифікація – системний розподіл предметів, явищ, процесів, які вивчаються, за родами, видами, типами, за якими-небудь істотними ознаками для зручності їх дослідження; групування вихідних

понять і розташування їх у певному порядку, що відображає міру цієї схожості.

Означення 2. Класифікація – впорядкована за деяким принципом множина об'єктів, які мають схожі класифікаційні ознаки (одну або декілька властивостей), вибраних для визначення схожості або відмінності між цими об'єктами.

Класифікація вимагає дотримання наступних правил:

– у кожному акті ділення необхідно застосовувати лише одну підставу;

– ділення має бути відповідним, тобто загальний обсяг видових понять повинен дорівнювати обсягу поділеного родового поняття;

– члени ділення повинні взаємно виключати одне одного, їх обсяги не повинні перехрещуватися;

– ділення має бути послідовним.

В класичній теорії розглядають:

допоміжну (штучну) класифікацію, яка виконується за зовнішньою ознакою і служить для придання множині предметів (процесів, явищ) потрібного порядку;

*природну класифікаці*ю, яка виконується за суттєвими ознаками, що характеризують внутрішню спільність предметів і явищ. Вона є результатом і важливим засобом наукового дослідження, оскільки передбачає і закріплює результати вивчення закономірностей об'єктів, що класифікуються.

Залежно від вибраних ознак, їх поєднання і процедури ділення понять класифікація може бути:

простою – ділення родового поняття лише за ознакою і лише один раз до розкриття всіх видів. Прикладом такої класифікації є дихотомія, при якій членами ділення бувають лише два поняття, кожне з яких є таким, що суперечить іншому (тобто дотримується принцип: « і не »);

складною – застосовується для ділення одного поняття за різними підставами 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 відносно визначення характеристик конкретної групи. Таким чином, для проведення класифікації мають бути присутніми ознаки, що характеризують групу, до якої належить та або інша подія або об'єкт (зазвичай при цьому на підставі аналізу вже класифікованих подій формулюються деякі правила).

Класифікація відноситься до стратегії навчання з вчителем (*supervised learning*), яка також іменується контрольованим або керованим навчанням. Задачею класифікації часто називають передбачення категоріальної залежної змінної (тобто залежної змінної, що є категорією) на основі вибірки безперервних і 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 так далі. Цей тип задач

відноситься до задач *бінарної класифікації*, в них залежна змінна може набувати лише два значення (наприклад, «так чи ні», «0 або 1»). Інший варіант класифікації (*багатокласова класифікація*) виникає, якщо залежна змінна може приймати значення з деякої множини зумовлених 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* (по двом і більше ознакам). Багатовимірна класифікація була розроблена біологами при вирішенні проблем дискримінації для класифікації організмів. Однією з перших робіт, присвячених цьому напряму, рахують роботу *Р. Фішера*, в якій організми розділялися на підвиди залежно від результатів вимірів їх фізичних параметрів. Біологія була і залишається найбільш жаданим і зручним середовищем, для розробки багатовимірних методів класифікації.

Мета процесу класифікації полягає в тому, аби побудувати модель, яка використовує прогнозуючі атрибути як вхідні параметри і отримує значення залежного атрибуту. Процес класифікації полягає в розбитті множини об'єктів на класи по певному критерію. Класифікатором називається деяка сутність, що визначає, якому із зумовлених класів належить об'єкт по вектору ознак.

Для проведення класифікації за допомогою математичних методів необхідно мати формальний опис об'єкту, яким можна оперувати, використовуючи математичний апарат класифікації. Таким описом в нашому випадку виступає база даних. Кожен об'єкт (запис бази даних) несе інформацію про деяку властивість об'єкту. Набір 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 і вихідні (цільові) значення прикладів. Вихідні значення призначені для навчання моделі. Тестова (test set) множина також містить вхідні і вихідні значення прикладів. Тут вихідні значення використовуються для перевірки працездатності моделі.

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

Конструювання моделі передбачає опис множини зумовлених класів. Зокрема, на цьому етапі виконуються наступні дії:

кожен приклад набору даних відноситься до одного зумовленого класу;

використовується навчальна множина, на якій відбувається конструювання моделі;

отримана модель представляється класифікаційними правилами, деревом рішень або математичною моделлю.

Використання моделі здійснює класифікацію нових або невідомих значень. Зокрема, на цьому етапі виконуються такі дії:

1. Оцінюється правильність та точність моделі, тобто:

відомі значення з тестового прикладу порівнюються з результатами використання отриманої моделі;

рівень точності визначається як відсоток правильно класифікованих прикладів в тестовій множині;

тестова множина не повинна залежати від навчальної множини.

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

При виборі методів класифікації слід проводити їх оцінювання, виходячи з таких характеристик: швидкість, робастність, інтерпретуємість, надійність. *Швидкість* характеризує час, який потрібний на створення моделі і її використання. *Робастність*, тобто стійкість до яких-небудь порушень деяких передумов, означає можливість роботи із зашумленими даними і пропущеними значеннями в даних. *Інтерпретуємість* забезпечує можливість розуміння моделі аналітиком. *Надійність* методів класифікації передбачає можливість роботи цих методів за наявності в наборі даних шумів і викидів.

Розглянемо деякі прикладні задачі, які ефективно вирішуються методами класифікації.

Задачі медичної діагностики. В ролі об'єктів виступають пацієнти. Ознаки характеризують результати обстежень, симптоми захворювання і методи лікування, що застосовувалися. Приклади бінарних ознак: стать, наявність головного болю, слабкості. Порядкова ознака – тяжкість стану (задовільний, середньої тяжкості, важкий, вкрай важкий). Кількісні ознаки – вік, пульс, артеріальний тиск, вміст гемоглобіну в крові, доза препарату. Ознаковий опис пацієнта є, по суті справи, формалізованою історією хвороби.

Передбачення родовищ корисних копалин. Ознаками є дані геологічної розвідки. Наявність або відсутність тих або інших порід на території району кодується бінарними ознаками. Фізико - хімічні властивості цих порід можуть описуватися як кількісними, так і якісними ознаками. Навчальна вибірка складається з прецедентів двох класів: районів відомих родовищ і схожих районів, в яких копалина, що цікавить, виявлена не була. При пошуку рідких корисних копалин кількість об'єктів може виявитися набагато менше, ніж кількість ознак. У цій ситуації погано працюють класичні статистичні методи. Задача вирішується шляхом пошуку закономірностей в наявному масиві даних.

Оцінювання кредитоспроможності позичальників. Ця задача вирішується банками при видачі кредитів. Потреба в автоматизації процедури видачі кредитів вперше виникла в період буму кредитних карт 60-70-х років в США і інших європейських країнах. Об'єктами в даному випадку є фізичні або юридичні особи, що претендують на здобуття кредиту. В разі фізичних осіб ознаковий опис складається з анкети, яку заповнює сам позичальник, і, можливо, додаткової інформації, яку банк збирає про нього з власних джерел.

Серед важливих задач, які також вирішуються методами класифікації, слід відзначити задачу передбачення відтоку клієнтів, оптичне розпізнавання символів, розпізнавання мови, виявлення спаму, класифікація документів та інше.

Існує велика різноманітність методів класифікації. Найбільш поширеними з них є:

– класифікація методом опорних векторів;
– байєсівська класифікація;
– статистичні методи, зокрема, лінійна регресія;
– класифікація за допомогою методу найближчого сусіда;
– класифікація CBR - методом;
– класифікація за допомогою штучних нейронних мереж;
– класифікація за допомогою дерев рішень;
– класифікація за допомогою генетичних алгоритмів.

Метод опорних векторів. У 60–70-і роки колективом математиків під керівництвом В. Н. Вапника був розроблений метод узагальненого портрета, заснований на побудові оптимальної розділяючої гіперплощини. Вимога оптимальності полягало в тому, що навчальні об'єкти мають бути віддалені від розділяючої поверхні настільки далеко, наскільки це можливо. У 90-і роки метод здобув широку світову популярність і після деякої переробки і серії узагальнень став називатися машиною опорних векторів (*Support Vector Machine – SVM*). В даний час він вважається одним з кращих методів класифікації.

Метод опорних векторів відноситься до групи граничних методів. Він визначає класи за допомогою границь областей. За допомогою даного методу вирішуються задачі бінарної класифікації.

Метод *SVM* володіє декількома чудовими властивостями. По-перше, навчання *SVM* зводиться до задачі квадратичного програмування, яка має єдине рішення, яке обчислюється досить ефективно навіть на вибірках в сотні тисяч об'єктів. По-друге, рішення володіє властивістю розрідженості: положення оптимальної розділяючої гіперплощини залежить лише від невеликої долі навчальних об'єктів. Вони і називаються *опорними векторами*; останні об'єкти фактично не задіюються. Нарешті, за допомогою введення *функції ядра* метод узагальнюється на випадок нелінійних розділяючих поверхонь.

Байєсівська класифікація. Байєсівські процедури класифікації розроблені на основі теореми Байєса і спеціально призначені для роботи з вхідними даними високої розмірності. Не дивлячись на простоту таких процедур, результати їх роботи по своїх характеристиках можуть перевершити результати роботи досить складних алгоритмів класифікації. Спочатку байєсівська класифікація використовувалася для формалізації знань експертів в експертних системах, зараз байєсівська класифікація широко застосовується як один з методів *Data Mining*.

Результатом роботи методу є так звані «прозорі» моделі. Властивостями наївної класифікації є:

1. Використання всіх змінних і визначення всіх залежностей між ними.

2. Наявність двох припущень відносно змінних:

всі змінні є однаково важливими;

всі змінні є статистично незалежними, тобто значення однієї змінної нічого не говорить про значення іншої.

Навчання байєсівських мереж стало одним з актуальних напрямів обчислювальної математики і до цих пір є предметом активних досліджень. Проте, до цих пір визначення структури байєсівської мережі в загальному вигляді є складним завданням як з теоретичної, так і з обчислювальної точки зору. Підхід в загальному вигляді володіє наступними *недоліками*:

обчислювальна складність;

при спробі врахувати велику кількість залежностей між змінними, оцінки умовної ймовірності набувають великої дисперсії, оскільки їх спільна поява в даних є маловірогідною подією. Таким чином, оцінки параметрів можуть стати недостовірними, що у результаті може приводити до погіршення якості класифікації навіть в порівнянні з «наївним» алгоритмом Байєса;

через велику кількість параметрів, модель виходить дуже орієнтованою на навчальні дані. Це приводить до дуже добрих результатів класифікації на навчальних даних і незадовільних результатів на тестових даних. Тобто модель описує не загальні закономірності в структурі даних, а швидше набір окремих випадків в навчальній вибірці.

Відзначають також такі достоїнства байєсівських мереж як методу *Data Mining*:

у моделі визначаються залежності між всіма змінними, це дозволяє легко обробляти ситуації, в яких значення деяких змінних невідомі;

байєсівські мережі досить просто інтерпретуються і дозволяють на етапі прогностичного моделювання легко проводити аналіз сценарію «що, … якщо»;

байєсівський метод дозволяє природним чином поєднувати закономірності, виведені з даних, і, наприклад, експертні знання, отримані в явному вигляді;

використання байєсівських мереж дозволяє уникнути проблеми перенавчання (overfitting), тобто надлишкового ускладнення моделі, що є слабкою стороною багатьох методів (наприклад, дерев рішень і нейронних мереж).

Метод «найближчого сусіда» або системи міркувань на основі аналогічних випадків.

У багатьох прикладних задачах вимірювати міру схожості об'єктів істотно простіше, ніж формувати ознакові описи. Наприклад, набагато легко порівняти дві фотографії і сказати, що вони належать одній людині, чим зрозуміти, на підставі яких ознак вони схожі. Такі ситуації часто виникають

при розпізнаванні часових рядів або символьних послідовностей. Вони характеризуються тим, що «сирі» вихідні дані не годяться як ознакові описи, але в той же час, існують ефективні і змістовно обґрунтовані способи оцінити міру схожості будь-якої пари «сирих» описів.

Є ще одна характерна особливість цих задач. Якщо міра схожості введена досить вдало, то виявляється, що схожим об'єктам, як правило, відповідають схожі відповіді. У задачах класифікації це означає, що схожі об'єкти набагато частіше лежать в одному класі, чим в різних. Якщо задача в принципі піддається рішенню, то границя між класами не може «проходить всюди»; класи утворюють компактно локалізовані підмножини в просторі об'єктів. Це припущення прийнято називати *гіпотезою компактності*.

7.2. Програмне забезпечення задач класифікації

Розробка програмних систем інтелектуального аналізу даних і, зокрема, комп'ютерною підтримки рішення задач класифікації активно ведуться в провідних зарубіжних країнах. Перш за все, це статистичні пакети обробки даних і візуалізації, в основі яких лежать методи різних розділів математичної статистики – перевірка статистичних гіпотез, регресійний аналіз, дисперсійний аналіз, аналіз часових рядів, і ін. Використання статистичних програмних продуктів стало стандартним і ефективним інструментом рішення задач класифікації, і, перш за все, початкового етапу досліджень, коли знаходяться значення різних усереднених показників, перевіряється статистична достовірність різних гіпотез, знаходяться регресійні залежності. В той же час статистичні підходи мають і істотні недоліки. Вони дозволяють оцінити статистичну достовірність значення параметра, гіпотези або залежності, проте самі методи обчислення величин, висунення гіпотез або знаходження залежностей мають очевидні обмеження. Перш за все, знаходяться усереднені по вибірці величини, що може бути досить грубим уявленням про аналізуємі або класифікуємі параметри. Будь-яка статистична модель використовує поняття «випадкових подій», «функцій розподілу випадкових величин» і тому подібне, тоді як взаємозв'язок між різними параметрами досліджуваних об'єктів, ситуацій або явищ є детермінованим. Саме використання статистичних методів передбачає наявність певного числа спостережень для обґрунтованості кінцевого результату, тоді як дане число може бути істотно більше можливого. Таким чином, при аналізі в принципі непредставимих даних, або на етапах початку накопичення даних, статистичні підходи стають неефективними як засіб аналізу і класифікації.

Останніми роками з'явилися спеціалізовані пакети інтелектуального аналізу даних. Для даних пакетів характерна орієнтація на широкий круг практичних задач, а їх алгоритмічною основою є сукупність альтернативних моделей. Таким чином, на сьогоднішньому рівні розвитку методів рішення задач інтелектуального аналізу даних і класифікації, переважною

представляється дорога застосування програмних засобів, що включають основні існуючі підходи. В даному випадку підвищуються шанси підбору з наявних алгоритмів такого алгоритму, який забезпечить найбільш точне вирішення задач користувача на нових даних. Іншим важливим атрибутом систем аналізу і класифікації має бути наявність засобів автоматичного вирішення задач класифікації колективами алгоритмів. Дійсно, стандартною ситуацією є наявність декількох альтернативних алгоритмів або рішень, рівнозначних для користувача. Для вибору з них одного найбільш ефективного не вистачає інформації. Тоді природною альтернативою вибору є створення на базі наявних алгоритмів або рішень нових, більш перспективніших.

Розглянемо можливості та практику застосування найбільш відомих програмних пакетів при вирішенні задач класифікації.

Система PolyAnalyst. В пакеті реалізован багатий інструментарій для вирішення задач класифікації та для знаходження правил віднесення записів до одного з двох або до одного з декількох класів.

Одним з таких інструментів є модуль *Stepwise Linear Regression (LR)* – покрокова багатопараметрична лінійна регресія.

Засоби аналізу STATISTICA Data Miner. Програмний комплекс STATISTICA включає величезний набір різних аналітичних процедур. Для спрощення роботи користувача в пакет були вбудовані готові закінчені модулі аналізу даних, призначені для вирішення найбільш важливих і популярних задач: прогнозування, класифікації і так далі. Зокрема, це такі модулі як:

General Classifier – класифікація. *STATISTICA Data Miner* включає повний пакет процедур класифікації: узагальнені лінійні моделі, дерева класифікації, регресійні дерева та інші.

General Modeler/Multivariate Explorer – узагальнені лінійні, нелінійні і регресійні моделі. Даний елемент містить лінійні, нелінійні, узагальнені регресійні моделі і елементи аналізу дерев класифікації.

Окрім них, *STATISTICA Data Miner* містить набір спеціалізованих процедур *Data Mining*, які доповнюють лінійку інструментів *Data Mining*:

General Classification and Regression Trees (GTrees) – узагальнені класифікаційні і регресійні дерева (*GTrees*). Модуль є повною реалізацією методів, розроблених *Breiman, Friedman, Olshen і Stone*. Окрім цього, модуль містить різного роду доопрацювання і доповнення, такі як оптимізації алгоритмів для великих об'ємів даних і так далі. Модуль є набором методів узагальненої класифікації і регресійних дерев.

Interactive Classification and Regression Trees – інтерактивна класифікація і регресійні дерева. На додаток до модулів автоматичної побудови різного роду дерев, *STATISTICA Data Miner* також включає засоби для формування таких дерев в інтерактивному режимі.

Boosted Trees – розширювані прості дерева. Останні дослідження аналітичних алгоритмів показують, що для деяких задач побудови

«складних» оцінок, прогнозів і класифікацій використання послідовно збільшуваних простих дерев дає точніші результати, ніж нейроні мережі або складні цілісні дерева. Даний модуль реалізує алгоритм побудови простих збільшуваних (розширюваних) дерев.

Oracle Data Mining. Oracle Data Mining є модулем в Oracle Enterprise Edition. ODM підтримує всі етапи технології інтелектуального аналізу даних, включаючи постановку задачі, підготовку даних, автоматичну побудову моделей, аналіз і тестування результатів, використання моделей в реальних застосуваннях. Важливо, що моделі будуються автоматично на основі аналізу наявних даних про об'єкти, спостереження і ситуації за допомогою спеціальних алгоритмів. Основу модуля *ODM* складають процедури, що реалізовують різні алгоритми побудови моделей класифікації, регресії, прогнозування.

Засоби бізнес-аналізу в SQL Server. В програмному комплексі реалізовані засоби Data Mining, які доступні користувачам цієї СУБД. В якості прикладів алгоритмів розглянемо *Microsoft Decision Trees* і байєсівський алгоритм.

Алгоритм *Microsoft Decision Tree* є алгоритмом класифікації, що дозволяє прогнозувати як безперервні, так і дискретні атрибути на основі оцінки в процесі навчання моделі міри впливу вхідних атрибутів на прогнозований атрибут і побудови ієрархічної структури, яка базується на відповіді «так чи ні» на набір питань. Алгоритми побудови дерев рішень дозволяють передбачити значення якого-небудь параметра для заданого випадку (наприклад, чи поверне вчасно чоловік виданий йому кредит) на основі великої кількості даних про інші подібні випадки (зокрема, на основі відомостей про інших осіб, яким видавалися кредити).

7.3. Класичні технології кластеризації в Data Mining

Задача кластеризації схожа із задачею класифікації, є її логічним продовженням, але її відмінність в тому, що класи вивчаємого 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 ознаковому просторі, то задача кластеризації зводиться до визначення «згущувань точок». Мета кластеризації – пошук існуючих структур.

Кластеризація є описовою процедурою, вона не робить жодних статистичних висновків, але дає можливість провести розвідувальний аналіз і вивчити «структуру даних». Само поняття «кластер» визначене неоднозначно: у кожному дослідженні свої «кластери». Переводиться поняття кластер (*cluster*) як «скупчення», «гроздь». Кластер можна

охарактеризувати як групу об'єктів, що мають загальні властивості. Характеристиками кластера можна назвати дві ознаки:

- внутрішня однорідність;
- зовнішня ізольованість.

Потреба в обробці великих масивів даних в *Data Mining* привела до формулювання вимог, яким, по можливості, повинен задовольняти алгоритм кластеризації. Розглянемо їх:

– мінімальна можлива кількість проходів по базі даних;

– робота в обмеженому об'ємі оперативної пам'яті комп'ютера;

– роботу алгоритму можна перервати із збереженням проміжних результатів, аби продовжити обчислення пізніше;

– алгоритм повинен працювати, коли об'єкти з бази даних можуть витягуватися лише в режимі однонаправленого курсора (тобто в режимі навігації по записах).

Рішення задачі кластеризації принципове неоднозначно, і тому є декілька причин. *По-перше*, не існує однозначно найкращого критерію якості кластеризації. Відомий цілий ряд досить розумних критеріїв, а також ряд алгоритмів, що не мають чітко вираженого критерію, але що здійснюють досить розумну кластеризацію «по побудові». Всі вони можуть давати різні результати. *По-друге*, число кластерів, як правило, невідоме заздалегідь і встановлюється відповідно до деякого суб'єктивного критерію. *По-третє*, результат кластеризації істотно залежить від метрики , вибір якої, як правило, також суб'єктивний і визначається експертом.

Вибір відстані між об'єктами є вузловим моментом дослідження, від нього багато в чому залежить остаточний варіант розбиття об'єктів на класи при даному алгоритмі розбиття. Існує декілька методів визначення функції відстані.

Відстань Евкліда. Найбільш пряма дорога обчислення відстаней між об'єктами полягає в обчисленні відстаней Евкліда. Вона є геометричною відстанню в багатовимірному просторі і обчислюється таким чином:

$$d(X_j, X_i) = \left[\sum_{k=1}^{N} (x_{ki} - x_{kj})^2 \right]^{1/2}.$$

Відмітимо, що відстань Евкліда обчислюється по початкових, а не за стандартизованими даними. Це звичайний спосіб її обчислення, який має певні переваги (наприклад, відстань між двома об'єктами не змінюється при введенні в аналіз нового об'єкту, який може виявитися викидом). Проте, на відстані можуть сильно впливати відмінності між осями, по координатах яких обчислюються ці відстані.

Відстань міських кварталів (*манхэттенська відстань*). Ця відстань є просто середньою різниць по координатах. В більшості випадків ця міра відстані приводить до таких же результатів, як і звичайна відстань Евкліда. Проте відзначимо, що для цієї міри вплив окремих великих різниць (викидів)

зменшується (оскільки вони не зводяться в квадрат). Манхеттенська відстань обчислюється за формулою:

$$d(X_j, X_i) = \sum_{k=1}^{N} | x_{ki} - x_{kj} |.$$

Відстань Чебишева. Ця відстань може виявитися корисною, коли бажають визначити два об'єкти як «різні», якщо вони розрізняються по якій-небудь одній координаті (яким-небудь одним виміром). Відстань Чебишева обчислюється за формулою:

$$d(X_j, X_i) = \max | x_{ki} - x_{kj} |$$

Степенна відстань (*відстань Мінковського*). Інколи виникає необхідність прогресивно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються. Це може бути досягнуто з використанням степенної відстані, яка обчислюється за формулою:

$$d(X_j, X_i) = (\sum_{k=1}^{N} | x_{ki} - x_{kj} |^p)^{1/r}$$

де r і p – параметри, що визначаються користувачем.

Параметр p відповідальний за поступове зважування різниць по окремих координатах, параметр r відповідальний за прогресивне зважування великих відстаней між об'єктами. Якщо обоє параметра рівні двом, то ця відстань збігається з відстанню Евкліда.

Відсоток незгоди. Ця міра використовується в тих випадках, коли дані є категоріальними. Ця відстань обчислюється за формулою:

$$d(X_j, X_i) = (Кількість\_x_{ki} \neq x_{kj}) / k.$$

Коли кожен об'єкт є окремим кластером, відстані між цими об'єктами визначаються вибраною мірою. Виникає наступне питання - як визначити відстані між кластерами? Існують різні правила, названі методами об'єднання або зв'язки для двох кластерів.

Метод ближнього сусіда або одиночний зв'язок. Тут відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) в різних кластерах. Цей метод дозволяє виділяти кластери скільки завгодно складної форми за умови, що різні частини таких кластерів сполучені ланцюжками близьких один до одного елементів. В результаті роботи цього методу кластери представляються довгими «ланцюжками» або «волокнистими» кластерами, «зчепленими разом» лише окремими елементами, які випадково виявилися ближчими за останніх один до одного.

Метод найбільш віддалених сусідів або повний зв'язок. Тут відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (тобто «найбільш віддаленими сусідами»).

Метод добре використовувати, коли об'єкти дійсно походять з різних «гаїв». Якщо ж кластери мають в деякому роді подовжену форму або їх природний тип є «ланцюжковим», то цей метод не слід використовувати.

Метод Варда (*Ward's method*). В якості відстані між кластерами береться приріст суми квадратів відстаней об'єктів до центрів кластерів, отримуваний в результаті їх об'єднання. На відміну від інших методів кластерного аналізу для оцінки відстаней між кластерами, тут використовуються методи дисперсійного аналізу. На кожному кроці алгоритму об'єднуються такі два кластери, які приводять до мінімального збільшення цільової функції, тобто внутрішньогрупової суми квадратів. Цей метод направлений на об'єднання близько розташованих кластерів і «прагне» створювати кластери малого розміру.

Метод незваженого попарного середнього (*unweighted pair-group method using arithmetic averages, UPGMA*). В якості відстані між двома кластерами береться середня відстань між всіма парами об'єктів в них. Цей метод слід використовувати, якщо об'єкти дійсно походять з різних «гаїв», у випадках присутності кластерів типа «ланцюжка», при припущенні нерівних розмірів кластерів.

Метод зваженого попарного середнього (*weighted pair-group method using arithmetic averages, WPGMA*). Цей метод схожий на метод незваженого попарного середнього, різниця полягає лише в тому, що тут в якості вагового коефіцієнту використовується розмір кластера (число об'єктів, що містяться в кластері). Цей метод рекомендується використовувати саме за наявності припущення про кластери різних розмірів.

На сьогоднішній день розроблена більше сотні різних алгоритмів кластеризації. Приведемо коротку характеристику підходів до кластеризації.

1. Алгоритми, засновані на розділенні даних *(Partitioning Algorithms*), в т.ч. ітеративні: розділення об'єктів на кластерів; k ітеративний перерозподіл об'єктів для поліпшення кластеризації.

2. Ієрархічні алгоритми (*Hierarchy Algorithms*): агломерація (*Agglomerative Nesting*): кожен об'єкт спочатку є кластером, кластери, з'єднуючись один з одним, формують більший кластер і так далі; дивізімні методи (*Divisive Analysis*): характеризуються послідовним розділенням кластера, що складається зі всіх об'єктів, і відповідним збільшенням числа кластерів, що в результаті приводить до створення послідовність розщеплюючих груп.

3. Методи, засновані на концентрації об'єктів (*Density-based methods*): засновані на можливості з'єднання об'єктів; ігнорують шуми, знаходження кластерів довільної форми.

4. Грід-методи (*Grid-based methods*): квантування об'єктів в грід-структури.

5. Модельні методи (*Model-based*): використання моделі для знаходження кластерів, найбільш відповідних даним.

6. Методи за способом аналізу даних:

- чіткі;
- нечіткі.

7. Методи по кількості застосування алгоритмів кластеризації:
   - з одноетапною кластеризацією;
   - з багатоетапною кластеризацією.
8. Методи по можливості розширення об'єму оброблюваних даних:
   - масштабовані;
   - не масштабовані.
9. Методи за часом виконання кластеризації:
   - потокові (on-line);
   - не потокові (off-line).

Оцінка якості кластеризації може бути проведена на основі наступних процедур:

– ручна перевірка;

– встановлення контрольних точок і 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. Проте створення схожих кластерів різними методами вказує на правильність кластеризації.

Кластерний аналіз застосовується в різних областях. Він корисний, коли потрібно класифікувати велику кількість інформації. Так, в медицині використовується кластеризація захворювань, лікування захворювань або їх симптомів, а також таксономія пацієнтів, препаратів і так далі. У археології встановлюються таксономії кам'яних споруд і древніх об'єктів. У менеджменті прикладом задачі кластеризації буде розбиття персоналу на різні групи, класифікація споживачів і постачальників, виявлення схожих виробничих ситуацій, при яких виникає брак. У соціології задача кластеризації – розбиття респондентів на однорідні групи. Загалом, всякий раз, коли необхідно класифікувати «гори» інформації до придатних для подальшої обробки груп, кластерний аналіз виявляється вельми корисним і ефективним.

Досить широко кластерний аналіз застосовується в маркетингових дослідженнях – як в теоретичних дослідженнях, так і практиці вирішення проблем групування різних об'єктів. При цьому вирішуються питання про групи клієнтів, продуктів і так далі.

Задачі кластерного аналізу можна об'єднати в наступні групи:

– розробка типології або класифікації;

– дослідження корисних концептуальних схем групування об'єктів;

– представлення гіпотез на основі дослідження даних;

– перевірка гіпотез або досліджень для визначення, чи дійсно типи (групи), виділені тим або іншим способом, присутні в наявних даних.

Як правило, при практичному використанні кластерного аналізу одночасно вирішується декілька з вказаних задач.

У загальному випадку всі етапи кластерного аналізу взаємозв'язані, і рішення, прийняті на одному з них, визначають дії на подальших етапах. Основними етапами кластерного аналізу є:

1. Вибір метрики і методу стандартизації вихідних даних.
2. Визначення кількості кластерів (для ітеративного кластерного аналізу).
3. Визначення методу кластеризації (правила об'єднання або зв'язку). На думку багатьох фахівців, вибір методу кластеризації є вирішальним при визначенні форми і специфіки кластерів.
4. Аналіз результатів кластеризації. Цей етап передбачає вирішення таких питань: чи не є отримане розбиття на кластери випадковим; чи є розбиття надійним і стабільним на підвибірках даних; чи існує взаємозв'язок між результатами кластеризації і змінними, які не брали участь в процесі кластеризації; чи можна інтерпретувати отримані результати кластеризації.
5. Перевірка результатів кластеризації. Результати кластеризації також мають бути перевірені формальними і неформальними методами. Формальні методи залежать від того методу, який використовувався для кластеризації. Неформальні включають наступні процедури перевірки якості кластеризації:

- аналіз результатів кластеризації, отриманих на певних вибірках набору даних;
- крос-перевірка;
- проведення кластеризації при зміні порядку спостережень в наборі даних;
- проведення кластеризації при видаленні деяких спостережень;
- проведення кластеризації на невеликих вибірках.

Один з варіантів перевірки якості кластеризації – використання декількох методів і порівняння отриманих результатів. Відсутність подібності не означатиме некоректність результатів, але присутність схожих груп вважається ознакою якісної кластеризації.

▪ ***Ієрархічні методи кластерного аналізу***.

При ієрархічній кластеризації виконується послідовне об'єднання менших кластерів у великі або розділення великих кластерів на менші.

Агломеративні методи AGNES (Agglomerative Nesting). Ця група методів характеризується послідовним об'єднанням елементів і відповідним зменшенням числа кластерів. На початку роботи алгоритму всі об'єкти є окремими кластерами. На першому кроці найбільш схожі об'єкти об'єднуються в кластер. На подальших кроках об'єднання продовжується до тих пір, поки всі об'єкти не складатимуть один кластер.

Single Link, Complete Link, Group Average. Одні із перших алгоритмів кластеризації даних. Особливістю цих методів, є те, що вони розбивають об'єкти на кластери шляхом розбиття їх на ієрархічні групи. Основна суть цих методів полягає у виконанні наступних кроків:

– обчислення значень близькості між елементами і здобуття матриці близькості;

– визначення кожного елементу в свій окремий кластер;

– злиття в один кластер найбільш близьких пар елементів;

– оновлення матриці близькості шляхом видалення колонок і рядків для кластерів, які злилися з іншими і подальшого перерахунку матриці;

– перехід на крок 3 до тих пір, поки не спрацює зупинний критерій.

Алгоритм CURE *(Clustering Using REpresentatives).* Виконує ієрархічну кластеризацію з використанням набору визначальних точок для визначення об'єкту в кластер. Призначення: кластеризація дуже великих наборів числових даних. Обмеження: ефективний для даних низької розмірності, працює лише на числових даних. Достоїнства: виконує кластеризацію на високому рівні навіть за наявності викидів, виділяє кластери складної форми і різних розмірів, володіє лінійно залежними вимогами до місця зберігання даних і часову складність для даних високої розмірності. Недоліки: є необхідність в задані порогових значень і кількості кластерів.

Дивізімні методи DIANA *(Divisive Analysis).* Ці методи є логічною протилежністю агломеративним методам. На початку роботи алгоритму всі об'єкти належать одному кластеру, який на подальших кроках ділиться на менші кластери, в результаті утворюється послідовність розщеплюючих груп.

Алгоритм BIRCH *(Balanced Iterative Reducing and Clustering using Hierarchies).* У цьому алгоритмі передбачений двох етапний процес кластеризації. Призначення: кластеризація дуже великих наборів числових даних. Обмеження: робота з лише числовими даними. Достоїнства: двоступінчата кластеризація, кластеризація великих об'ємів даних, працює на обмеженому об'ємі пам'яті, є локальним алгоритмом, може працювати при одному скануванні вхідного набору даних, використовує той факт, що дані неоднаково розподілені по простору, і обробляє області з великою щільністю як єдиний кластер. Недоліки: робота з лише числовими даними, добре виділяє лише кластери сферичної форми, є необхідність в задані порогових значень.

Алгоритм MST *(Algorithm based on Minimum Spanning Trees).* Призначення: кластеризація великих наборів довільних даних. Достоїнства: виділяє кластери довільної форми, вибирає з декількох оптимальних рішень найоптимальніше.

▪ ***Неієрархічні методи кластерного аналізу.***

При великій кількості спостережень ієрархічні методи кластерного аналізу не придатні. У таких випадках використовують неієрархічні методи, засновані на розділенні, які є ітеративними методами дроблення вхідної сукупності. В процесі ділення нові кластери формуються до тих пір, поки не буде виконано правило зупинки. Така неієрархічна кластеризація полягає в розділенні набору даних на певну кількість окремих кластерів. Існує два

підходи. Перший полягає у визначенні кордонів кластерів як найбільш щільних ділянок в багатовимірному просторі даних, тобто визначення кластера там, де є велике «згущення точок». Другий підхід полягає в мінімізації міри відмінності об'єктів.

Ітераційні алгоритми. Такі алгоритми засновані на оптимізації деякої цільової функції, що визначає оптимальне в певному значенні розбиття множини об'єктів на кластери. Вони носять ітеративний характер, і на кожній ітерації потрібно розраховувати матрицю відстаней між об'єктами. При великому числі об'єктів це неефективно і вимагає серйозних обчислювальних ресурсів.

Алгоритм k-середніх *(k-means).* Найбільш поширений серед неієрархічних методів алгоритм *k* -середніх, також названий швидким кластерним аналізом. На відміну від ієрархічних методів, які не вимагають попередніх припущень відносно числа кластерів, для можливості використання цього методу необхідно мати гіпотезу про найбільш вірогідну кількість кластерів.

Перевагами алгоритму *k*-середніх є: простота та швидкість використання, зрозумілість і прозорість алгоритму. Недоліки алгоритму *k* -середніх: алгоритм дуже чутливий до викидів, які можуть спотворювати середнє.

Можливим вирішенням цієї проблеми є використання модифікації алгоритму – алгоритм *k* -медіани; алгоритм може повільно працювати на великих базах даних. Можливим вирішенням даної проблеми є використання вибірки даних.

Алгоритм PAM *(partitioning around Medoids). PAM* є модифікацією алгоритму *k* -середніх алгоритмом *k* -медіани (*k* -medoids). Алгоритм менш чутливий до шумів і викидів даних, чим алгоритм *k* -means, оскільки медіана менше схильна до впливів викидів. PAM ефективний для невеликих баз даних, але його не слід використовувати для великих наборів даних.

Алгоритм EM *(Expectation - Maximization).* В основі ідеї *EM* алгоритму лежить припущення, що досліджувана множина даних може бути змодельована за допомогою лінійної комбінації багатовимірних нормальних розподілів, а метою є оцінка параметрів розподілів, які максимізували логарифмічну функцію правдоподібності, використовувану як міра якості моделі. Іншими словами, передбачається, що дані в кожному кластері підкоряються певному закону розподілу, а саме, нормальному розподілу. З врахуванням цього припущення можна визначити параметри - математичне сподівання і дисперсію, які відповідають закону розподілу елементів в кластері, щонайкраще «відповідному» до спостережуваних даних. Таким чином, ми передбачаємо, що будь-яке спостереження належить до всіх кластерів, але з різною ймовірністю. Тоді задача полягатиме в «підгонці» розподілів суміші до даних, а потім у визначенні ймовірності приналежності спостереження до кожного кластера. Вочевидь, що спостереження має бути віднесене до того кластера, для якого дана вірогідність вища.

Кластеризація категорійних даних: алгоритм CLOPE. Категорійні дані зустрічаються в будь-яких областях: виробництво, комерція, маркетинг, медицина. Вони включають і так звані транзакційні дані: чеки в супермаркетах, логи відвідин веб-ресурсів. Взагалі під категорійними даними розуміють якісні характеристики об'єктів, виміряні в шкалі найменувань.

▪***Нові алгоритми кластерного аналізу.***

Методи, які ми розглянули, є «класикою» кластерного аналізу. До останнього часу основним критерієм, по якому оцінювався алгоритм кластеризації, була якість кластеризації: вважалося, аби весь набір даних уміщався в оперативній пам'яті. Проте зараз, у зв'язку з появою надвеликих баз даних, з'явилися нові вимоги, яким повинен задовольняти алгоритм кластеризації. Основне з них – це масштабованість алгоритму. Відзначимо також інші властивості, яким повинен задовольняти алгоритм кластеризації: незалежність результатів від порядку вхідних даних; незалежність параметрів алгоритму від вхідних даних. Останнім часом ведуться активні розробки нових алгоритмів кластеризації, здатних обробляти надвеликі бази даних. У них основна увага приділяється масштабованості. До таких алгоритмів відноситься узагальнене представлення кластерів (*summarized cluster representation*), а також вибірка і використання структур даних, підтримуваних нижче лежачих СУБД.

Алгоритми нечіткої кластеризації *– алгоритм Fuzzy C-means.* Призначення: кластеризація великих наборів числових даних. Достоїнства: нечіткість при визначенні об'єкту в кластер дозволяє визначати об'єкти, які знаходяться на кордоні, в кластери. Недоліки: обчислювальна складність, задання кількості кластерів, виникає невизначеність з об'єктами, які віддалені від центрів всіх кластерів.

Алгоритм WaveCluster*. WaveCluster* є алгоритмом кластеризації на основі хвилевих перетворень. На початку роботи алгоритму дані узагальнюються шляхом накладання на простір даних багатовимірних грат. На подальших кроках алгоритму аналізуються не окремі точки, а узагальнені характеристики точок, що попали в одну чарунку грат. В результаті такого узагальнення необхідна інформація уміщається в оперативній пам'яті. На подальших кроках для визначення кластерів алгоритм застосовує хвилеве перетворення до узагальнених даних. Головні особливості *Wave Cluster*: складність реалізації, алгоритм може виявляти кластери довільних форм, алгоритм не чутливий до шумів, алгоритм застосовний лише до даних низької розмірності.

Алгоритм CLARA *(Clustering LARge Applications).* Алгоритм *CLARA* був розроблений Kaufmann і Rousseeuw в 1990 році для кластеризації даних у великих базах даних. Даний алгоритм виконується в статистичних аналітичних пакетах, наприклад, таких як S+.

Викладемо коротко суть алгоритму. Алгоритм *CLARA* витягує множину зразків з бази даних. Кластеризація застосовується до кожного із зразків, на виході алгоритму пропонується краща кластеризація. Для великих

баз даних цей алгоритм ефективніший, ніж алгоритм *PAM*. Ефективність алгоритму залежить від вибраного як зразок набору даних. Хороша кластеризація на вибраному наборі може не дати хорошу кластеризацію на всій множині даних.

Алгоритми Clarans, CURE, DBScan. Алгоритм *Clarans (Clustering Large Applications based upon RANdomized Search*) формулює задачу кластеризації як випадковий пошук в графі. В результаті роботи цього алгоритму сукупність вузлів графа є розбиттям множини даних на число кластерів, визначеним користувачем. «Якість» отриманих кластерів визначається за допомогою критеріальної функції. Алгоритм Clarans сортує все можливе розбиття множини даних у пошуках прийнятного рішення. Пошук рішення зупиняється в тому вузлі, де досягається мінімум серед зумовленого числа локальних мінімумів.

7.4. Програмне забезпечення задач кластеризації

Одним з основних підходів в «виявленні знань в даних» (*Data Mining*) є кластеризація. Кластерний аналіз дозволяє відкрити в даних раніше невідомі закономірності, які практично неможливо досліджувати іншими способами і представити їх в зручній для користувача формі. Методи кластерного аналізу використовуються як самостійні інструменти досліджень, так і у складі інших засобів *Data Mining*. До теперішнього часу розроблена велика кількість програмних продуктів, що застосовуються до даних різного типа. Розглянемо основні з них.

Система PolyAnalyst. В програмному комплексі реалізовано два модулі, які відповідають за проведення кластерного аналізу: *Find Dependecies (FD)* – *N*-мірний аналіз розподілів та *Find Clusters (FC)* – N-мірний кластеризатор.

Система «Багатовимірні розвідувальні технології аналізу STATISTICA».

У модулі «*Кластерний аналіз*» реалізований повний набір методів кластерного аналізу даних, включаючи методи -середніх, ієрархічної кластеризації і двовходового об'єднання. Дані можуть поступати як у ісходному вигляді, так і у вигляді матриці відстаней між об'єктами. Спостереження і змінні можна кластеризувати, використовуючи різні міри відстані (евклідову, квадрат евклідова, міських кварталів (манхэттенське), Чебишева, степенне, відсоток незгоди і коефіцієнта кореляції Пірсона), а також різні правила об'єднання кластерів.

Пакет програм SPSS (Statistical Package for Social Science). Пакет програм є одним з поширених, потужних і зручних інструментів статистичного аналізу. Пакет SPSS користується популярністю у економістів, соціологів, маркетологів, надає користувачеві широкі можливості по статистичній обробці емпіричних даних, по формуванню і модифікації баз даних, а також по створенню звітів, надаючи широкі можливості по

представленню результатів статистичної обробки в текстовій, табличній і графічній формах. Пакет орієнтований, головним чином, на аналіз просторових даних і на кластерний аналіз. Інтерфейс програми інтуїтивно зрозумілий користувачеві і дозволяє застосувати різні варіанти статистичного аналізу.

***Кластеризація в програмному комплексі «1С**: Предприятие 8.0».* Метою кластеризації в програмному комплексі є виділення з множини об'єктів однієї природи деякої кількості відносно однорідних груп – сегментів або кластерів. Об'єкти розподіляються по групах так, щоб внутрішньогрупові відмінності були мінімальними, а міжгрупові – максимальними. Методи кластеризації дозволяють перейти від пооб’єктного до групового представлення сукупності довільних об'єктів, що істотно спрощує операцію ними.