1. Історія виникнення та причини розвитку

Поняття Data Mining, що з'явилося в 1978 р., набуло високої популярності в сучасному трактуванні приблизно з першої половини 90-х років. До цього часу обробка та аналіз даних здійснювалися в рамках прикладної статистики, при цьому в основному вирішувалися завдання обробки невеликих баз даних.

Термін інтелектуальний аналіз даних походить від поняття Data Mining котре отримало свою назву з двох понять: пошуку цінної інформації у великій базі даних (Data) і видобутку (Mining). Обидва процеси вимагають або просіювання величезної кількості сирого матеріалу, або розумного дослідження і пошуку шуканих цінностей. Також термін Data Mining часто перекладається як видобуток даних, витягування інформації, розкопування даних, інтелектуальний аналіз даних, засоби пошуку закономірностей, вилучення знань, аналіз шаблонів, розкопування знань в базах даних.

Поняття «Виявлення знань в базах даних» (knowledge discovery in databases, KDD) можна вважати синонімом інтелектуального аналізу даних.

У зв'язку з удосконаленням технологій запису і зберігання даних на людей обвалилися колосальні потоки «інформаційного видобутку» в найрізноманітніших областях. Діяльність будь-якого підприємства (комерційного, виробничого, медичного, наукового і т.д.) тепер супроводжується реєстрацією та записом всіх подробиць його діяльності. Що робити з цією інформацією? Стало ясно, що без продуктивної переробки потоки сирих даних утворюють нікому не потрібну звалище.

Специфіка сучасних вимог до такої переробки наступна:

• дані мають необмежений обсяг;

• дані є різнорідними (кількісними, якісними, текстовими);

• результати мають бути конкретні і зрозумілі;

•інструменти для обробки сирих даних повинні бути прості у використанні.

Традиційна математична статистика, яка довгий час претендувала на роль основного інструменту аналізу даних не могла більше ефективно вирішувати дані завдання. Головна причина - концепція усереднення за вибіркою, що призводить до операцій над фіктивними величинами (типу середньої температури пацієнтів по лікарні, середньої висоти будинку на вулиці і т.п.). Методи математичної статистики виявилися корисними головним чином для перевірки заздалегідь сформульованих гіпотез (перевірка керованості інтелектуального аналізу даних) і для "грубого" розвідувального аналізу, що становить основу оперативної аналітичної обробки даних (аналітична обробка в реальному часі, online analytical processing, OLAP).

Причини популярності Data Mining:

• стрімке накопичення даних;

• загальна комп'ютеризація бізнес-процесів;

• проникнення Інтернету у всі сфери діяльності;

• прогрес в області інформаційних технологій: вдосконалення СУБД і сховищ даних;

• прогрес в області виробничих технологій: стрімке зростання продуктивності комп'ютерів, об'ємів накопичувачів, впровадження Grid систем.

Data Mining - мультидисциплінарна галузь, що виникла і розвивається на базі таких наук як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних та ін., див. рис. 1.1.

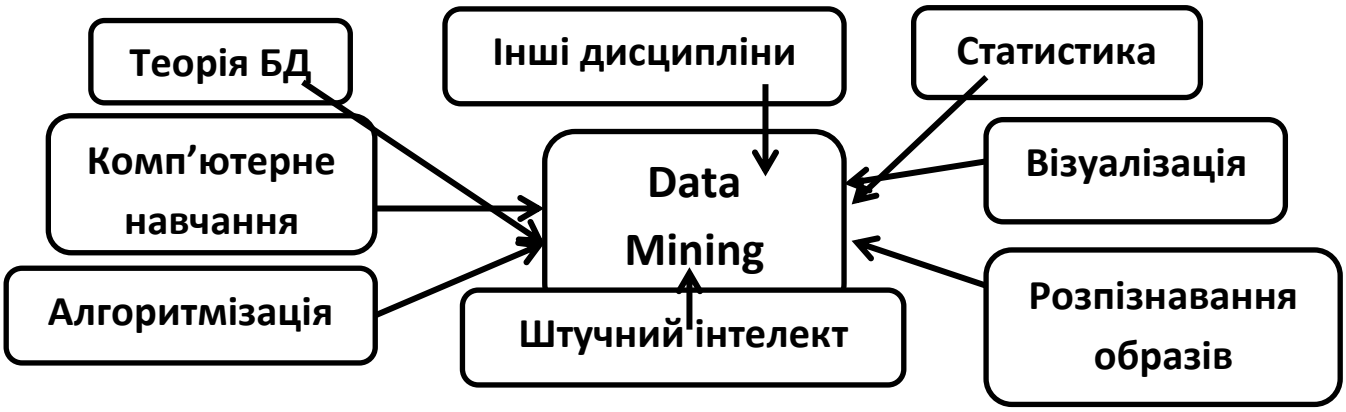

Рисунок 1.1 – Data Mining, як міждисциплінарна галузь

2. Суть, мета та сфера застосування технології Data Mining

Суть та мету технології Data Mining можна охарактеризувати так: це технологія, яка призначена для пошуку у великих обсягах даних неочевидних, об'єктивних і корисних на практиці закономірностей.

Неочевидних - це означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом.

Об'єктивних - це означає, що виявлені закономірності будуть повністю відповідати дійсності, на відміну від експертної думки, яке завжди є суб'єктивним.

Практично корисних - це означає, що висновки мають конкретне значення, котрому можна знайти практичне застосування.

Знання - сукупність відомостей, яка утворює цілісний опис, відповідне деякому рівню обізнаності про описуваному питанні, предметі, проблемі і т.д.

Використання знань означає дійсне застосування знайдених знань для досягнення конкретних переваг (наприклад, в конкурентній боротьбі за ринок).

Наведемо ще кілька визначень поняття Data Mining.

Data Mining - це процес виділення з даних неявної і неструктурованою інформації та представлення її у вигляді, придатному для використання.

Data Mining - це процес виділення, дослідження і моделювання великих обсягів даних для виявлення невідомих до цього структур (моделей) з метою досягнення переваг у бізнесі (визначення SAS Institute).

Data Mining - це процес, мета якого - виявити нові значущі кореляції, зразки і тенденції в результаті просіювання великого обсягу збережених даних з використанням методик розпізнавання зразків плюс застосування статистичних і математичних методів (визначення Gartner Group).

Data Mining становлять велику цінність для керівників та аналітиків в їх повсякденній діяльності. Ділові люди усвідомили, що за допомогою методів Data Mining вони можуть отримати відчутні переваги в конкурентній боротьбі. Коротко охарактеризуємо деякі можливі бізнес-додатки інтелектуального аналізу даних.

Роздрібна торгівля. Підприємства роздрібної торгівлі сьогодні збирають докладну інформацію про кожну окрему покупку, використовуючи кредитні картки з маркою магазина і комп'ютеризовані системи контролю. Ось типові завдання, які можна вирішувати за допомогою Data Mining у сфері роздрібної торгівлі:

• аналіз купівельної корзини (аналіз подібності) призначений для виявлення товарів, які покупці прагнуть купувати разом. Володіння купівельної корзини необхідно для поліпшення реклами, вироблення стратегії створення запасів товарів і способів їх розкладки в торгових залах.

• дослідження тимчасових шаблонів допомагає торговим підприємствам приймати рішення про створення товарних запасів. Воно дає відповіді на питання типу "Якщо сьогодні покупець придбав відеокамеру, то через який час він найімовірніше купить нові батарейки і плівку?"

• створення прогнозуючих моделей дає можливість торговельним підприємствам дізнаватися характер потреб різних категорій клієнтів з певною поведінкою, наприклад, купують товари відомих дизайнерів або відвідують розпродажі. Ці знання потрібні для розробки точно спрямованих, економічних заходів щодо просування товарів.

Банківська справа. Досягнення технології Data Mining використовуються в банківській справі для вирішення наступних поширених завдань:

• виявлення шахрайства з кредитними картками. Шляхом аналізу минулих транзакцій, які згодом виявилися шахрайськими, банк виявляє деякі стереотипи такого шахрайства.

• сегментація клієнтів. Розбиваючи клієнтів на різні категорії, банки роблять свою маркетингову політику більш цілеспрямованою і результативною , пропонуючи різні види послуг різним групам клієнтів.

• прогнозування змін клієнтури. Data Mining допомагає банкам будувати прогнозні моделі цінності своїх клієнтів, і відповідним чином обслуговувати кожну категорію.

Телекомунікації. В області телекомунікацій методи Data Mining допомагають компаніям більш енергійно просувати свої програми маркетингу і ціноутворення, щоб утримувати існуючих клієнтів і залучати нових. Серед типових заходів відзначимо наступні:

• аналіз записів про докладних характеристиках викликів. Призначення такого аналізу - виявлення категорій клієнтів з схожими стереотипами користування їх послугами та розробка привабливих наборів цін і послуг;

• виявлення лояльності клієнтів. Data Mining можна використовувати для визначення характеристик клієнтів, які, один раз скориставшись послугами даної компанії, з великою часткою ймовірності залишаться їй вірними. У підсумку кошти, що виділяються на маркетинг, можна витрачати там, де віддача найбільше.

Страхування. Страхові компанії протягом ряду років накопичують великі обсяги даних. Тут широке поле діяльності для методів Data Mining:

• виявлення шахрайства. Страхові компанії можуть знизити рівень шахрайства, відшукуючи певні стереотипи в заявах про виплату страхового відшкодування, що характеризують взаємини між юристами, лікарями та заявниками.

• аналіз ризику. Шляхом виявлення поєднань факторів, пов'язаних з оплаченими заявами, страховики можуть зменшити свої втрати за зобов'язаннями. Відомий випадок, коли в США велика страхова компанія виявила, що суми, виплачені за заявами людей, одружених, вдвічі перевищує суми за заявами самотніх людей. Компанія відреагувала на це нове знання переглядом своєї загальної політики надання знижок сімейним клієнтам.

Інші області в бізнесі. Data Mining може застосовуватися в безлічі інших областей:

• розвиток автомобільної промисловості. При складанні автомобілів виробники повинні враховувати вимоги кожного окремого клієнта, тому їм потрібні можливість прогнозування популярності певних характеристик і знання того, які характеристики зазвичай замовляються разом;

• політика гарантій. Виробникам потрібно передбачати число клієнтів, які подадуть гарантійні заявки, і середню вартість заявок;

• заохочення часто літаючих клієнтів. Авіакомпанії можуть виявити групу клієнтів, яких даними заохочувальними заходами можна спонукати літати більше. Наприклад, одна авіакомпанія виявила категорію клієнтів, які здійснювали багато польотів на короткі відстані, що не накопичуючи досить миль для вступу в їхні клуби, тому вона таким чином змінила правила прийому до клубу, щоб заохочувати число польотів так само, як і милі.

Медицина. Відомо багато експертних систем для постановки медичних діагнозів. Вони побудовані головним чином на основі правил, що описують поєднання різних симптомів різних захворювань. За допомогою таких правил дізнаються не тільки, на що хворий пацієнт, але і як потрібно його лікувати. Правила допомагають вибирати засоби медикаментозного впливу, визначати показання - протипоказання, орієнтуватися в лікувальних процедурах, створювати умови найбільш ефективного лікування, пророкувати результати призначеного курсу лікування і т. п. Технології Data Mining дозволяють виявляти в медичних даних шаблони, що становлять основу зазначених правил.

Молекулярна генетика і генна інженерія. Мабуть, найбільш гостро і водночас чітко завдання виявлення закономірностей в експериментальних даних коштує в молекулярній генетиці та генної інженерії. Тут вона формулюється як визначення так званих маркерів, під якими розуміють генетичні коди, контролюючі ті чи інші фенотипічні ознаки живого організму. Такі коди можуть містити сотні, тисячі і більше пов'язаних елементів.

На розвиток генетичних досліджень виділяються великі кошти. Останнім часом в даній області виник особливий інтерес до застосування методів Data Mining. Відомо кілька великих фірм, що спеціалізуються на застосуванні цих методів для розшифровки генома людини і рослин.

Прикладна хімія. Методи Data Mining знаходять широке застосування в прикладної хімії (органічної та неорганічної). Тут нерідко виникає питання про з'ясування особливостей хімічної будови тих чи інших сполук, що визначають їх властивості. Особливо актуальна така задача при аналізі складних хімічних сполук, опис яких включає сотні і тисячі структурних елементів та їх зв'язків.

Можна навести ще багато прикладів різних областей знання, де методи Data Mining відіграють провідну роль. Особливість цих областей полягає в їх складній системної організації. Вони відносяться головним чином до над кібернетичному рівню організації систем, закономірності якого не можуть бути достатньо точно описані на мові статистичних чи інших аналітичних математичних моделей. Дані в зазначених областях неоднорідні, гетерогенні, нестаціонарні і часто відрізняються високою розмірністю.

3. Типи закономірностей

Виділяють п'ять стандартних типів закономірностей, які дозволяють виявляти методи Data Mining: асоціація, послідовність, класифікація, кластеризація і прогнозування.

Асоціація має місце в тому випадку, якщо кілька подій зв'язані один з одним. Наприклад, дослідження, проведене в супермаркеті, може показати, що 65 % купили кукурудзяні чіпси беруть також і "кока - колу", а за наявності знижки за такий комплект "колу" здобувають у 85% випадків. Маючи в своєму розпорядженні відомостями про подібну асоціацію, менеджерам легко оцінити, наскільки дієва надається знижка.

Якщо існує ланцюжок пов'язаних у часі подій, то говорять про послідовність. Так, наприклад, після покупки будинку в 45 % випадків протягом місяця купується і нова кухонна плита, а в межах двох тижнів 60 % новоселів обзаводяться холодильником.

За допомогою класифікації виявляються ознаки, що характеризують групу, до якої належить той чи інший об'єкт. Це робиться за допомогою аналізу вже класифікованих об'єктів і формулювання деякого набору правил.

Кластеризація відрізняється від класифікації тим, що самі групи заздалегідь не задані. За допомогою кластеризації засобів Data Mining самостійно виділяють різні однорідні групи даних.

Основою для всіляких систем прогнозування служить історична інформація, що зберігається в БД у вигляді часових рядів. Якщо вдається побудувати знайти шаблони, які адекватно відображають динаміку поведінки цільових показників, є ймовірність, що з їх допомогою можна передбачити і поведінку системи в майбутньому.

4. Класи систем Data Mining

Data Mining є мультидисциплінарною областю, яка виникла і розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних та ін. (рис. 1.2). Звідси велика кількість методів і алгоритмів, реалізованих у різних діючих системах Data Mining. Багато з таких систем інтегрують в собі відразу кілька підходів. Проте, як правило, в кожній системі є якась ключова компонента, на яку робиться головна ставка. Нижче наводиться класифікація зазначених ключових компонент на основі роботи. Виділеним класам дається коротка характеристика.

Рисунок 1.2 – Data Mining - мультидисциплінарна область

Предметно-орієнтовані аналітичні системи дуже різноманітні. Найбільш широкий підклас таких систем, що одержав поширення в галузі дослідження фінансових ринків, носить назву "технічний аналіз". Він являє собою сукупність декількох десятків методів прогнозу динаміки цін і вибору оптимальної структури інвестиційного портфеля, заснованих на різних емпіричних моделях динаміки ринку. Ці методи часто використовують нескладний статистичний апарат, але максимально враховують сформовану своїй області специфіку (професійна мова, системи різних індексів і пр.). На

ринку є безліч програм цього класу. Як правило, вони досить дешеві (зазвичай $ 300-1000).

Статистичні пакети. Останні версії майже всіх відомих статистичних пакетів включають поряд з традиційними статистичними методами також елементи Data Mining. Але основна увага в них приділяється все ж класичним методикам - кореляційному, регресійному, факторному аналізу і іншим. Недоліком систем цього класу вважають вимогу до спеціальної підготовки користувача. Також відзначають, що потужні сучасні статистичні пакети є занадто "ваговитими" для масового застосування у фінансах і бізнесі. До того ж часто ці системи досить дорогі - від $ 1000 до $ 15000.

Є ще більш серйозний принциповий недолік статистичних пакетів, що обмежує їх застосування в Data Mining. Більшість методів, що входять до складу пакетів спираються на статистичну парадигму, в якій головними фігурантами служать усереднені характеристики вибірки. А ці характеристики, як зазначалося вище, при дослідженні реальних складних життєвих феноменів часто є фіктивними величинами.

В якості прикладів найбільш потужних і поширених статистичних пакетів можна назвати SAS (компанія SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA та інші.

Нейронні мережі. Це великий клас систем, архітектура яких має аналогію (як тепер відомо, досить слабку) з побудовою нервової тканини з нейронів. В одній з найбільш поширених архітектурі зі зворотним поширенням помилки, імітується робота нейронів у складі ієрархічної мережі, де кожен нейрон більш високого рівня з'єднаний своїми входами з виходами нейронів нижчого шару. На нейрони самого нижнього шару подаються значення вхідних параметрів, на основі яких потрібно приймати якісь рішення, прогнозувати розвиток ситуації і т. д. Ці значення розглядаються як сигнали, що передаються в наступний шар, ослабляючись або посилюючись в залежності від числових значень (ваг), приписуваних між нейронних зв'язків. У результаті на виході нейрона самого верхнього шару виробляється деяке значення, яке розглядається як відповідь - реакція всієї мережі на введені значення вхідних параметрів. Для того щоб мережу можна було застосовувати надалі, її колись треба "натренувати" на отриманих раніше даних, для яких відомі і значення вхідних параметрів, і правильні відповіді на них. Тренування полягає в підборі ваг між нейронних зв'язків, що забезпечують найбільшу близькість відповідей мережі до відомих правильних відповідей.

Основним недоліком нейронно-мережевої парадигми є необхідність мати дуже великий обсяг навчальної вибірки. Інший суттєвий недолік полягає в тому, що навіть натренована нейронна мережа являє собою чорний ящик. Знання, зафіксовані як ваги кількох сотень між нейронних зв'язків, абсолютно не піддаються аналізу та інтерпретації людиною (відомі спроби дати інтерпретацію структурі налаштованої нейронної мережі виглядають непереконливими - система " KINOsuite - PR").

Приклади нейромережевих систем - BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic). Вартість їх досить значна: $ 1500-8000.

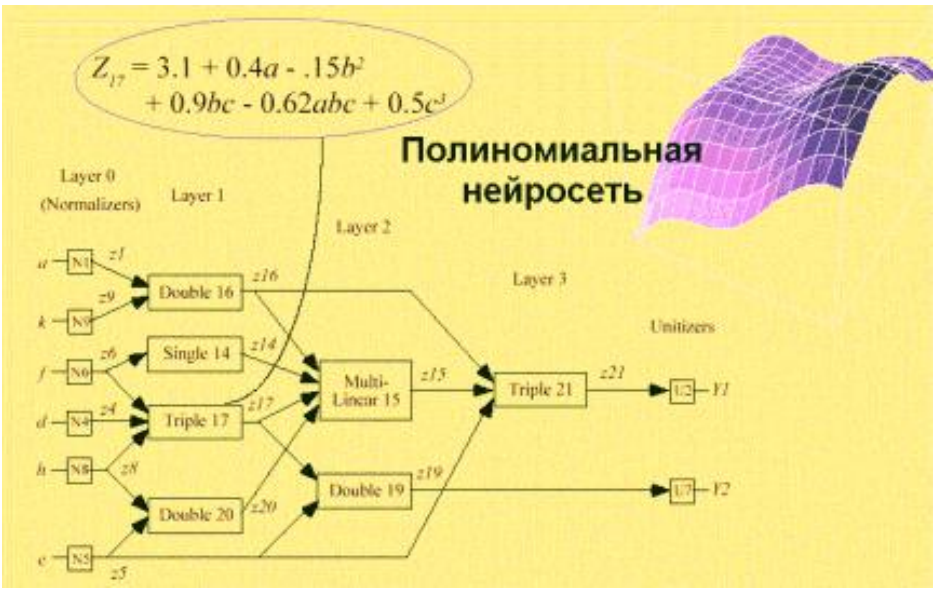

Рисунок 1.4 – Поліноміальна нейронна мережа

Системи міркувань на основі аналогічних випадків. Ідея систем case based reasoning - CBR - на перший погляд вкрай проста. Для того щоб зробити прогноз на майбутнє чи вибрати правильне рішення, ці системи знаходять у минулому близькі аналоги наявної ситуації і вибирають ту ж відповідь, який був для них правильним. Тому цей метод ще називають методом "найближчого сусіда" (nearest neighbour). Останнім часом поширення отримав також термін memory based reasoning, який акцентує увагу, що рішення приймається на підставі всієї інформації, накопиченої в пам'яті.

Системи CBR показують непогані результати в найрізноманітніших задачах. Головним їх мінусом вважають те, що вони взагалі не створюють будь-яких моделей або правил, узагальнюючих попередній досвід у виборі рішення вони ґрунтуються на всьому масиві доступних історичних даних, тому неможливо сказати, на основі яких конкретно факторів CBR системи будують свої відповіді.

Інший мінус полягає в свавіллі, який допускають системи CBR при виборі міри "близькості". Від цієї міри найрішучішим чином залежить обсяг безлічі прецедентів, які потрібно зберігати в пам'яті для досягнення задовільною класифікації або прогнозу.

Приклади систем, що використовують CBR, - KATE tools (Acknosoft, Франція), Pattern Recognition Workbench (Unica, США).

Дерева рішень (decision trees) є одним з найбільш популярних підходів до вирішення завдань Data Mining. Вони створюють ієрархічну структуру класифікуючи правил типу "ЯКЩО ... ТО ..." (if - then), що має вигляд дерева. Для прийняття рішення, до якого класу віднести деякий об'єкт або ситуацію,

потрібно відповісти на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Запитання мають вигляд "значення параметра А більше х?". Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня, якщо негативний - то до лівого вузла; потім знову слід питання, пов'язане з відповідним вузлом.

Популярність підходу пов'язана як би з наочністю і зрозумілістю. Але дерева рішень принципово не здатні знаходити "кращі" (найбільш повні і точні) правила в даних. Вони реалізують наївний принцип послідовного перегляду ознак і "чіпляють" фактично оскільки справжніх закономірностей, створюючи лише ілюзію логічного висновку.

Разом з тим, більшість сислтем використовують саме цей метод. Найвідомішими є See5/C5.0 (RuleQuest, Австралія), Clementine (Integral Solutions, Великобританія), SIPINA (University of Lyon, Франція), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS , Канада). Вартість цих систем варіюється від 1 до 10 тис. дол.

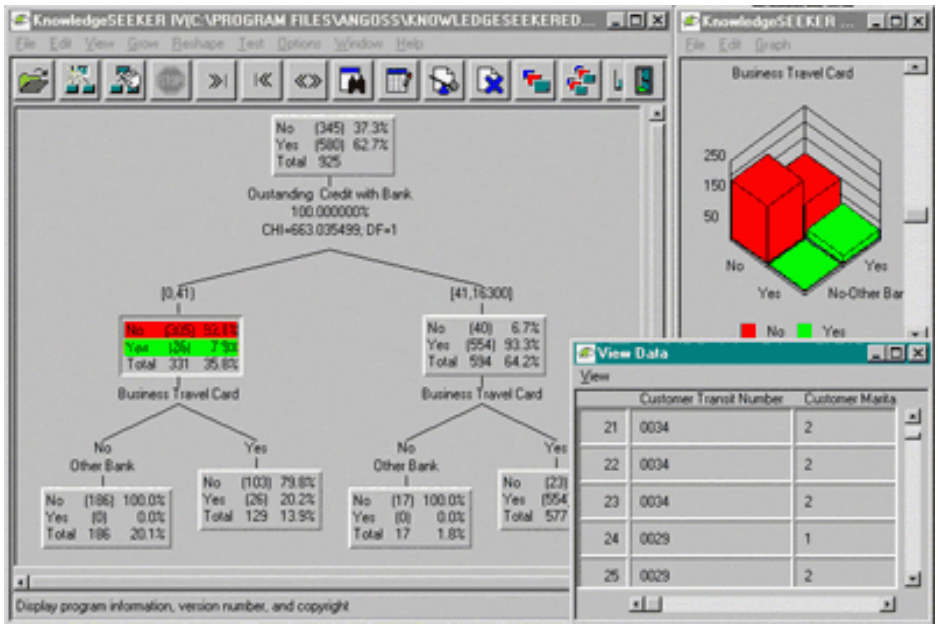

Рисунок 1.5 – Система KnowledgeSeeker обробляє банківську інформацію

Еволюційне програмування. Проілюструємо сучасний стан даного підходу на прикладі системи PolyAnalyst - вітчизняної розробки, що отримала сьогодні загальне визнання на ринку Data Mining. У даній системі гіпотези про вид залежності цільової змінної від інших змінних формулюються у вигляді програм на деякій внутрішній мові програмування. Процес побудови програм будується як еволюція у світі програм (цим підхід трохи схожий на генетичні алгоритми). Коли система знаходить програму, більш або менш задовільно виражає шукану залежність, вона починає вносити до неї невеликі модифікації і відбирає серед побудованих дочірніх програм ті, які підвищують точність. Таким чином система "вирощує" кілька генетичних ліній програм, які конкурують між собою в точності

висловлювання шуканої залежності. Спеціальний модуль системи PolyAnalyst переводить знайдені залежності з внутрішньої мови системи на зрозумілу користувачеві мову (математичні формули, таблиці та ін.).

Генетичні алгоритми. Data Mining не основна область застосування генетичних алгоритмів. Їх потрібно розглядати скоріше як потужний засіб вирішення різноманітних комбінаторних завдань і завдань оптимізації. Проте генетичні алгоритми увійшли зараз в стандартний інструментарій методів Data Mining, тому вони і включені в даний огляд.

Перший крок при побудові генетичних алгоритмів - це кодування вихідних логічних закономірностей в базі даних, які іменують хромосомами, а весь набір таких закономірностей називають популяцією хромосом. Далі для реалізації концепції відбору вводиться спосіб зіставлення різних хромосом. Населення обробляється за допомогою процедур репродукції, мінливості (мутацій), генетичної композиції. Ці процедури імітують біологічні процеси. Найбільш важливі серед них: випадкові мутації даних в індивідуальних хромосомах, переходи (кросинговер) і рекомбінація генетичного матеріалу, що міститься в індивідуальних батьківських, та міграції генів. У ході роботи процедур на кожній стадії еволюції виходять популяції з усе більш досконалими індивідуумами.

Генетичні алгоритми зручні тим, що їх легко розпаралелювати. Наприклад, можна розбити покоління на кілька груп і працювати з кожною з них незалежно, обмінюючись час від часу кількома хромосомами. Існують також і інші методи розпаралелювання генетичних алгоритмів.

Генетичні алгоритми мають ряд недоліків. Критерій відбору хромосом і використовувані процедури є евристичними і далеко не гарантують знаходження "кращого" рішення. Як і в реальному житті, еволюцію може "заклинити" на який-небудь непродуктивною гілки. І, навпаки, можна навести приклади, як два неперспективних батька, які будуть виключені з еволюції генетичним алгоритмом, виявляються здатними призвести високоефективного нащадка. Це особливо стає помітно при вирішенні високо розмірних завдань зі складними внутрішніми зв'язками.

Прикладом може служити система GeneHunter фірми Ward Systems Group. Його вартість - близько $ 1000.

Алгоритми обмеженого перебору були запропоновані в середині 60- х років М.М. Бонгард для пошуку логічних закономірностей в даних. З тих пір вони продемонстрували свою ефективність при вирішенні безлічі завдань із всіляких областей.

Ці алгоритми обчислюють частоти комбінацій простих логічних подій у підгрупах даних. Приклади простих логічних подій: X = a ; X < a ; X a ; a < X < b та ін, де X - який або параметр, " a " і " b " - константи. Обмеженням служить довжина комбінації простих логічних подій (у М. Бонгард вона дорівнювала 3). На підставі аналізу обчислених частот робиться висновок про корисність тієї чи іншої комбінації для встановлення асоціації в даних, для класифікації, прогнозування.

Найбільш яскравим сучасним представником цього підходу є система WizWhy підприємства WizSoft. Хоча автор системи Абрахам Мейдан не розкриває специфіку алгоритму, покладеного в основу роботи WizWhy, за результатами ретельного тестування системи були зроблені висновки про наявність тут обмеженого перебору (вивчалися результати, залежно часу їх отримання від числа аналізованих параметрів та ін.)

Автор WizWhy стверджує, що його система виявляє ВСЕ логічні if - then правила в даних. Насправді це, звичайно, не так. По-перше, максимальна довжина комбінації в if - then правилі в системі WizWhy дорівнює 6, і, по-друге, з самого початку роботи алгоритму виробляється евристичний пошук простих логічних подій, на яких потім будується весь подальший аналіз. Зрозумівши ці особливості WizWhy, неважко було запропонувати найпростішу тестову задачу, яку система не змогла взагалі вирішити. Інший момент - система видає рішення за прийнятний час тільки для порівняно невеликої розмірності даних.

Проте, система WizWhy є на сьогоднішній день одним з лідерів на ринку продуктів Data Mining. Це не позбавлене підстав. Система постійно демонструє більш високі показники при вирішенні практичних завдань, ніж всі інші алгоритми. Вартість системи близько $4000, кількість продажів - 30000.

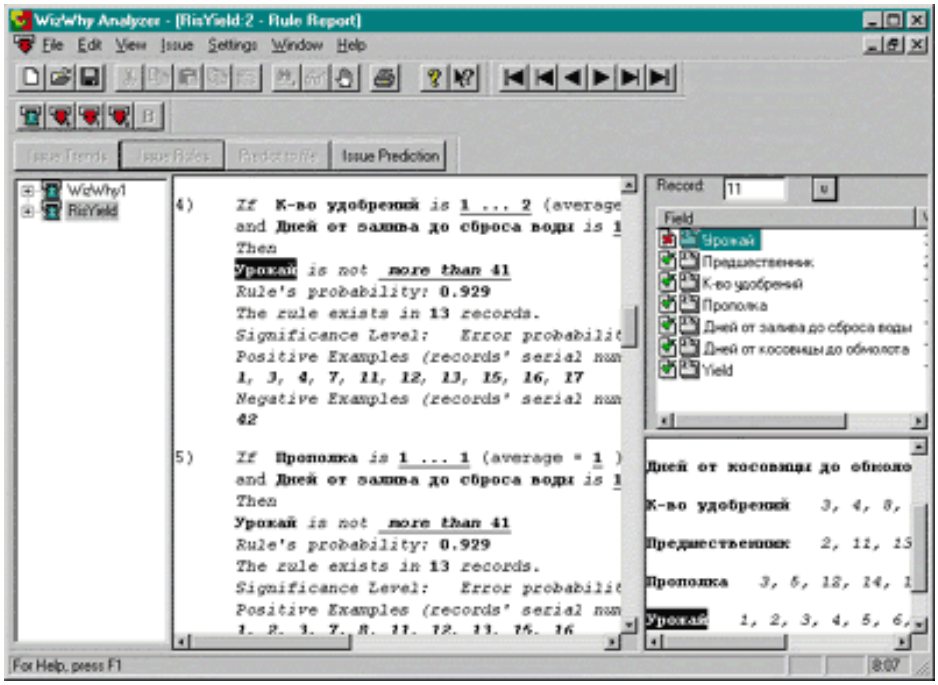

Рисунок 1.6 – Система WizWhy виявила правила, що пояснюють низьку врожайність деяких сільськогосподарських ділянок

Системи для візуалізації багатовимірних даних. В тій чи іншій мірі засоби для графічного відображення даних підтримуються всіма системами Data Mining. Разом з тим, досить значну частку ринку займають системи, що спеціалізуються виключно на цій функції. Прикладом тут може служити програма DataMiner 3D словацької фірми Dimension5 (5-й вимір).

У подібних системах основну увагу сконцентровано на доброзичливості користувацького інтерфейсу, що дозволяє асоціювати з аналізованими показниками різні параметри діаграми розсіювання об'єктів (записів) бази даних. До таких параметрів належать колір, форма, орієнтація щодо власної осі, розміри та інші властивості графічних елементів зображення. Крім того, системи візуалізації даних забезпечені зручними засобами для масштабування і обертання зображень. Вартість систем візуалізації може досягати декількох сотень доларів.

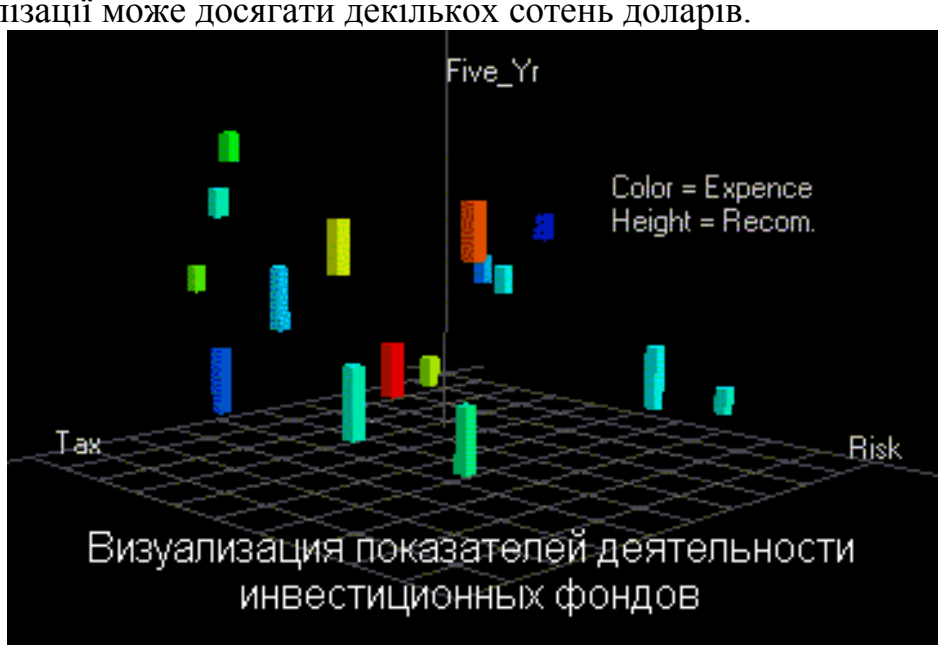

Рисунок 1.7 – Візуалізація даних системою DataMiner 3D

Загальний алгоритм аналізу Data Mining. Методика аналізу з використанням Data Mining базується на різних алгоритмах видобутку закономірностей у вхідних даних. Таких алгоритмів є багато, але вони не можуть гарантувати якісного кінцевого результату, бо існує багато чинників, що можуть вплинути на сам хід аналізу.

5. Якісний аналіз даних з використанням DM.

Для якісного аналізу будь-яких даних слід дотримуватися загальної схеми використання DM:

Висування гіпотез;

Збір та систематизація даних;

Підбір адекватної моделі;

Тестування та інтерпретація отриманих даних;

Використання у реальних умовах.

Ця схема не залежить від предметної області та сфери діяльності. Вона є універсальною.

1. Висування гіпотез.

Гіпотезою тут будемо вважати припущення про вплив певних факторів на процес, що досліджується.

Автоматизувати процес висування гіпотез є вкрай складно, тому, цю задачу мають вирішувати експерти – фахівці в предметній області.

Слід довіритися їх досвіду та здоровому глузду, максимально використати ці знання про предмет досліджень і зібрати як найбільше гіпотез/припущень.

Зазвичай, добрі результати надають тактики «круглого столу» або «мозкової атаки». На початку слід зібрати та систематизувати всі ідеї, а оцінювати їх пізніше. В результаті повинен бути складений перелік з описів всіх факторів досліджуваного об'єкту.

Для задачі прогнозування потрібно скласти перелік факторів, що впливатимуть на об'єкт і експертно-оцінити суттєвість кожного з них. Така оцінка не є вирішальною, але від неї починають відштовхуватися.

Згодом, під час аналіз, може з'ясуватися, що фактор, який експерти оцінили як важливий, буде мати незначний вплив на процес і навпаки.

2. Збір та систематизація даних

2.1. Збір даних.

Для аналізу потрібно як найбільше даних, бо це надає можливість оцінити вплив максимальної кількості показників. Згодом, простіше відхилити певну частину даних, аніж розпочинати новий збір.

Методи збору:

1. Отримання даних з внутрішніх джерел.

Це не складно, бо така інформація зазвичай зберігається в облікових системах у табличній формі, де існують різні механізми отримання звітів та експортування даних.

2. Отримання відомостей з непрямих даних.

Наприклад, потрібно оцінити реальний фінансовий стан мешканців певного регіону. Існує кілька категорій товару (авто), що різняться за ціною – для незаможних, середнього класу, заможних. Якщо отримати звіт про продажі товару в цьому районі і проаналізувати пропорції, то робиться висновок: чим більшим є відсоток продажів дорогого товару, тим заможнішими є мешканці.

3. Використання відкритих джерел.

До широкого загалу надаються статистичні збірники, звіти корпорацій, результати маркетингових досліджень, соціологічні опитування.

4. Влаштування власних маркетингових досліджень та подібних заходів по збору даних.

Це зазвичай є дорогим заходом, але доволі ефективним.

5. Наповнення даних згідно експертних оцінок співробітниками організації.

Слід оцінити вартість збору даних, що потрібні для аналізу. Одні дані беруться з публічних інформаційних джерел, інші мають бути оплачені, дані про діяльність конкурентів можуть бути доволі дорогими.

Вартість збору інформації різними методами суттєво різниться за ціною та витраченим часом, тому, слід вважати на співвідношення теперішніх витрат з майбутніми результатами.

Від даних, які експерти вважають несуттєвими, певна річ, можна відмовитися, але від значущих даних не можна, бо аналіз буде базуватися у цьому випадку на другорядних факторах і відповідно, отримана модель буде надавати нестабільні та невірні результати.

6. Дані, набір даних та їх атрибутів

У широкому розумінні дані - це факти, текст, графіки, картинки, звуки, аналогові або цифрові відео-сегменти. Дані можуть бути отримані в результаті вимірювань, експериментів, арифметичних і логічних операцій. Дані повинні бути представлені у формі, придатній для зберігання, передачі і обробки. Іншими словами, дані - це необроблений матеріал, що надається постачальниками даних і використовується споживачами для формування інформації на основі даних.

Таблиця 1.1. Двовимірна таблиця "об’єкт-атрибут"

	Атрибути				
Об’єкти	Код клієнта	Вік	Сімейний статус	Прибуток	Клас
	1	19	неодр.	1234	1
	2	23	одр.	1222	1
	3	34	одр.	2700	1
	4	24	неодр.	2343	1
	5	26	одр.	1765	2
	6	32	розл.	2652	1
	7	19	неодр.	1200	2
	8	22	неодр.	1765	2
	9	40	одр.	1998	1
	10	43	розл.	4332	1

По горизонталі таблиці розташовуються атрибути об'єкта або його ознаки. По вертикалі таблиці - об'єкти. Об'єкт описується як набір атрибутів. Об'єкт також відомий як запис, випадок, приклад, рядок таблиці і т.д.

Атрибут - властивість, що характеризує об'єкт. Наприклад : колір очей людини, температура води і т.д. Атрибут також називають змінною, полем таблиці, виміром, характеристикою.

Змінна (variable) - властивість або характеристика, загальна для всіх досліджуваних об'єктів, прояв якої може змінюватися від об'єкта до об'єкта.

Значення (value) змінної є проявом ознаки.

При аналізі даних, як правило, немає можливості розглянути всю сукупність об'єктів, що нас цікавить. Вивчення дуже великих обсягів даних є дорогим процесом, що вимагає великих затрат часу, а також неминуче призводить до помилок, пов'язаних з людським фактором.

Цілком достатньо розглянути деяку частину всієї сукупності, тобто вибірку, і отримати цікаву для нас інформацію на її підставі.

Однак розмір вибірки повинен залежати від різноманітності об'єктів, представлених у генеральній сукупності. У вибірці повинні бути представлені різні комбінації і елементи генеральної сукупності.

Генеральна сукупність (population) - вся сукупність досліджуваних об'єктів, що цікавить дослідника.

Вибірка (sample) - частина генеральної сукупності, певним способом відібрана з метою дослідження та отримання висновків про властивості та характеристики генеральної сукупності.

Параметри - числові характеристики генеральної сукупності.

Статистики - числові характеристики вибірки. Часто дослідження ґрунтуються на гіпотезах. Гіпотези перевіряються за допомогою даних.

Гіпотеза - припущення щодо параметрів сукупності об'єктів, яке має бути перевірено на її частині.

Гіпотеза - частково обґрунтована закономірність знань, що служить або для зв'язку між різними емпіричними фактами, або для пояснення факту групи фактів.

Приклад гіпотези: між показниками тривалості життя та якістю харчування є зв'язок. У цьому випадку метою дослідження може бути пояснення змін конкретної змінної, в даному випадку - тривалості життя. Припустимо, існує гіпотеза, що залежна змінна (тривалість життя) змінюється залежно від деяких причин (якість харчування, спосіб життя, місце проживання і т.д.), які і є незалежними змінними.

Однак змінна першопочатково не є залежною або незалежною, вона стає такою після формулювання конкретної гіпотези. Залежна змінна в одній гіпотезі може бути незалежною в іншій.

Вимірювання - процес присвоєння чисел характеристикам досліджуваних об'єктів згідно певного правила.

У процесі підготовки даних вимірюється не сам об'єкт, а його характеристики.

Шкала - правило, відповідно до якого об'єктам присвоюються числа.

Багато інструментів Data Mining при імпорті даних з інших джерел пропонують вибрати тип шкали для кожної змінної та/або вибрати тип даних для вхідних і вихідних змінних (символьні, числові, дискретні і безперервні).

Користувачеві такого інструменту необхідно володіти цими поняттями.

Змінні можуть бути числовими даними або символьними.

Числові дані, у свою чергу, можуть бути дискретними і безперервними.

Дискретні дані являються значеннями ознаки, загальне число яких скінченне або нескінченне, але може бути підраховане за допомогою натуральних чисел від одного до нескінченності.

Приклад дискретних даних. Тривалість маршруту тролейбуса (кількість варіантів тривалості скінченне): 10, 15, 25 хв.

Безперервні дані - дані, значення яких можуть набувати якого-завгодно значення в деякому інтервалі. Вимірювання безперервних даних передбачає велику точність.

Приклад безперервних даних: температура, висота, вага, довжина і т.д.

Шкали. Існує п'ять типів шкал вимірювань: номінальна, порядкова, інтервальна, відносна і дихотомічна.

Номінальна шкала (nominal scale) - шкала, яка містить тільки категорії; дані в ній не можуть упорядковуватися, з ними не можуть бути зроблені ніякі арифметичні дії.

Номінальна шкала складається з назв, категорій, імен для класифікації і сортування об'єктів або спостережень за деякою ознакою.

Приклад такої шкали: професії, місто проживання, сімейний стан.

Для цієї шкали застосовні тільки такі операції: дорівнює (=), не дорівнює (≠).

Порядкова шкала (ordinal scale) - шкала, в якій числа присвоюють об'єктам для позначення відносної позиції об'єктів, але не величини відмінностей між ними.

Шкала вимірювань дає можливість ранжувати значення змінних. Вимірювання ж у порядковій шкалі містять інформацію лише про порядок проходження величин, але не дозволяють сказати "наскільки одна величина більше іншої, або" наскільки вона менше інший".

Приклад такої шкали: місце (1-ше, 2-ге, 3-є), яке команда отримала на змаганнях, номер студента в рейтингу успішності (1-й, 23-й, і т.д.), при цьому невідомо, наскільки один студент успішніше іншого, відомий лише його номер у рейтингу.

Для цієї шкали застосовні тільки такі операції: рівно (=), не дорівнює (≠), більше (>), менше (<).

Інтервальна шкала (interval scale) - шкала, різниці між значеннями якої можуть бути обчислені, проте їх відношення не мають сенсу.

Ця шкала дозволяє знаходити різницю між двома величинами, має властивості номінальної та порядкової шкал, а також дозволяє визначити кількісну зміну ознаки.

Приклад такої шкали: температура води в морі вранці - 19 градусів, ввечері - 24, тобто вечірня на 5 градусів вище, але не можна сказати, що вона в 1,26 разів вище.

Номінальна і порядкова шкали є дискретними, а інтервальна шкала - безперервною, вона дозволяє здійснювати точні вимірювання ознаки і виробляти арифметичні операції додавання, віднімання, множення, ділення.

Для цієї шкали застосовні тільки такі операції: одно (=), не дорівнює ( ), більше (>), менше (<), операції додавання (+) і віднімання (-).

Відносна шкала (ratio scale) - шкала, в якій є певна точка відліку і можливі відносини між значеннями шкали.

Приклад такої шкали : вага новонародженої дитини (4 кг і 3 кг). Перша в 1,33 рази важче.

Ціна на картоплю в супермаркеті вище в 1,2 рази, ніж ціна на ринку.

Відносні та інтервальні шкали є числовими .

Для цієї шкали можуть бути застосовані тільки такі операції: рівно (=), не дорівнює (≠), більше (>), менше (<), операції додавання (+) і віднімання (-), множення (*) і ділення (/).

Дихотомічна шкала (dichotomous scale) - шкала, яка містить тільки дві категорії.

Приклад такої шкали: стать (чоловіча і жіноча).

Приклад використання різних шкал для вимірювань властивостей різних об'єктів, в даному випадку температурних умов, наведено в таблиці даних, зображеної в табл. 1.2.

Таблиця 1.2. Множина вимірювань властивостей різних об'єктів

Номер об’єкту	Професія (номінальна шкала)	Середній бал (інтервальна шкала)	Освіта (порядкова шкала)
1	Слюсар	22	середня
2	Вчений	55	вища
3	вчитель	47	вища

Приклад використання різних шкал для вимірювань властивостей однієї системи, в даному випадку температурних умов, наведено в таблиці даних, зображеній в табл. 1.3.

Таблиця 1.3. Множина вимірювань властивостей однієї системи

Дата змінення	Хмарність (номінальна шкала)	Температура о 7 годині (інтервальна шкала)	Сила вітру (порядкова шкала)
3 жовтня	Хмарно	22ºС	Сильний вітер
4 жовтня	напівхмарно	17ºС	Слабий вітер
5 жовтня	Ясно	23ºС	Дуже сильний вітер

Типи наборів даних. Найбільш часто зустрічаються дані, що складаються з записів (record data).

Приклади таких наборів даних: табличні дані, матричні дані, документальні дані, транзакційні або операційні.

Табличні дані - дані, що складаються з записів, кожен з яких складається з фіксованого набору атрибутів.

Транзакційні дані представляють собою особливий тип даних, де кожен запис, що являється транзакцією, включає набір значень.

Приклад транзакційної бази даних, що містить перелік покупок клієнтів магазину, наведено на рис. 1.8.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Рисунок 1.8 – Приклад транзакційних даних

Графічні дані. Приклади графічних даних: молекулярні структури; графи (рис. 1.9); карти.

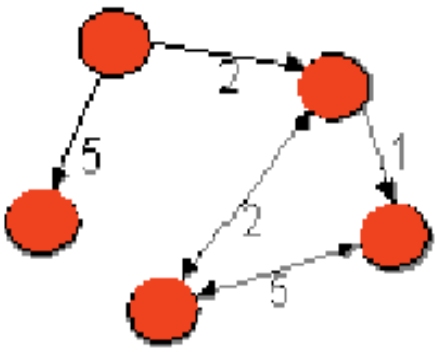

Рисунок 1.9 – Приклад графу

За допомогою карт, наприклад, можна відстежити зміни об'єктів в часі і просторі, визначити характер їх розподілу на площині або в просторі.

Перевагою графічного представлення даних є велика простота їх сприйняття, ніж, наприклад, табличних даних.

Приклад карти, що є картою Кохонена (моделлю нейронних мереж), представлений на рис. 1.10.

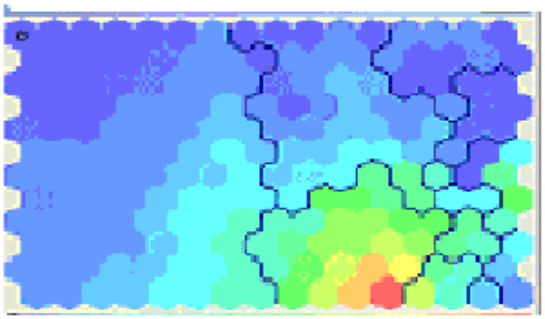

Рисунок 1.10 – Приклад даних типу "Карта Кохонена" хімічні дані

Хімічні дані представляють собою особливий тип даних. Приклад таких даних: Молекула бензолу C6H6

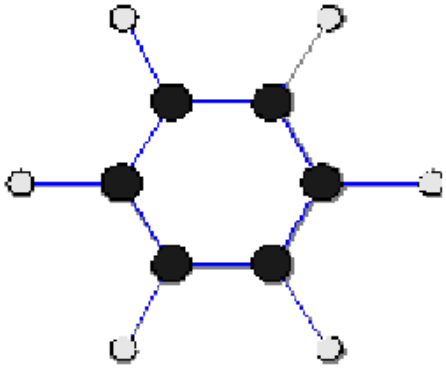

Рисунок 1.11 – Приклад хімічних даних

7. Формати зберігання даних

Одна з основних особливостей даних сучасного світу полягає в тому, що їх стає дуже багато.

Можливі чотири аспекти роботи з даними:

визначення даних;

обчислення;

маніпулювання;

обробка (збір, передача тощо).

При маніпулюванні даними використовується структура даних типу "файл". Файли можуть мати різні формати.

Більшість інструментів Data Mining дозволяють імпортувати дані з різних джерел, а також експортувати результуючі дані в різні формати.

Дані для експериментів зручно зберігати в якомусь одному форматі.

У деяких інструментах Data Mining ці процедури називаються імпорт / експорт даних, інші дозволяють напряму відкривати різні джерела даних і зберігати результати Data Mining в одному із запропонованих форматів.

Найбільш поширені формати, згідно з опитуванням "Формати зберігання даних", представлені на рис..1.12.

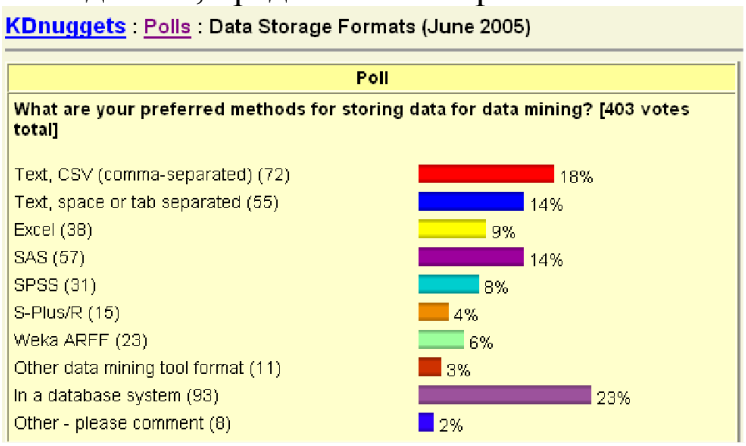

Рисунок 1.12 – Найбільш поширені формати зберігання даних

Найбільше число опитаних (23%) вважають за краще зберігати дані у форматі тієї бази даних, яку вони використовують. У форматі Text, CSV - 18%, по 14% опитаних зберігають дані у форматі Text, space or tab separated і SAS; в форматі Excel - 9%, SPSS - 8%, S-Plus/R - 4%, Weka ARFF - 6%, в інших форматах інструментів Data Mining - 2%.

Як бачимо з результатів опитування, найбільш поширеним форматом зберігання даних для Data Mining виступають бази даних.