Тема 1. Інтелектуальний аналіз даних (Data Mining). Особливості технології Data Mining та її відмінності від інших методів аналізу даних

План

1. Історія виникнення та причини розвитку.
2. Суть, мета та сфера застосування технології Data Mining.
3. Типи закономірностей.
4. Класи систем Data Mining.

Мета вивчення теми: засвоїти основні концептуальні поняття з курсу «Інтелектуальний аналіз даних»; засвоїти відмінності Data Mining від класичних статистичних методів аналізу й OLAP-систем, вивчити типи закономірностей, що виявляють Data Mining та класи систем інтелектуального аналізу даних.

Перелік ключових слів та понять із теми

Data Mining, асоціація, класифікація, послідовність, кластерізація, прогнозування

Теоретичні відомості з теми

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 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 у базах даних.

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

Розвиток технології баз даних:

- 1960-і рр. У 1968 році була введена в експлуатацію перша промислова СУБД система IMS фірми IBM.
- 1970-і рр. У 1975 році з'явився перший стандарт асоціації по мовах систем обробки даних – Conference on Data System Languages (CODASYL), який визначив низку фундаментальних понять у теорії систем баз даних, які досі є основоположними для мережевої моделі даних. У подальший розвиток

теорії баз даних великий внесок був зроблений американським математиком Е.Ф. Коддом, який є творцем реляційної моделі даних.

- 1980-і рр. Протягом цього періоду багато дослідників експериментували з новим підходом у напрямках структуризації баз даних і забезпечення до них доступу. Метою цих пошуків було отримання реляційних прототипів для простішого моделювання даних. У результаті, в 1985 році була створена мова, названа SQL. На сьогоднішній день практично всі СУБД забезпечують цей інтерфейс.
- 1990-і рр. З'явилися специфічні типи даних – «графічний образ», «документ», «звук», «карта», типи даних для часу, інтервалів часу, символьних рядків із двобайтовим поданням символів були додані в мову SQL. З'явилися технології **Data Mining**, сховища даних, мультимедійні бази даних і веб-бази 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 зрозумілим, що без продуктивної переробки потоки сирих даних утворюють нікому не потрібне звалище.

Специфіка сучасних вимог до такої переробки така:

• 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 ці 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.

Про популярність Data Mining говорить і той факт, що результат пошуку терміну «Data Mining» у пошуковій система Google (на вересень 2017 року) – становить більше 18 мільярдів сторінок, на вересень 2013 – 198 мільйонів сторінок.

Data Mining – мультидисциплінарна галузь, що виникла і розвивається на базі таких наук як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних тощо.

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 – означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом.

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

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

Знання – сукупність відомостей, яка утворює цілісний опис, відповідне деякому рівню обізнаності про описуване питання, предмет, проблему тощо.

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).

Сфера застосування Data Mining нічим не обмежена – вона скрізь, де є будь-які дані. Але в першу чергу методи Data Mining сьогодні, м'яко кажучи, заінтригували комерційні підприємства, що розгортають проекти на основі інформаційних сховищ даних (сховища даних). Досвід багатьох таких підприємств показує, що віддача від використання Data Mining може досягати 1000%. Наприклад, відомі повідомлення про економічний ефект, що в 10-70 разів перевищив початкові витрати від 350 до 750 тис. дол. Є

відомості про проект в 20 млн. дол., який окупився всього за 4 місяці. Інший приклад – річна економія 700 тис. дол. за рахунок впровадження Data Mining у мережі універсамів у Великобританії.

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. Подібний аналіз потрібен 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 про характер потреб різних категорій клієнтів із 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 таких 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 такі:

• *аналіз записів про докладних характеристиках викликів.* Призначення такого аналізу – виявлення категорій клієнтів із 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 найбільша.

Страхування. Страхові компанії протягом декількох 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 відносини 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 одружених людей, вдвічі перевищують 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, зокрема таких, як:

• *розвиток автомобільної промисловості.* При виготовленні 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 певними заохочувальними заходами можна спонукати літати більше. Наприклад, одна авіакомпанія виявила категорію клієнтів, які здійснювали багато перельотів на короткі відстані, що не накопичували достатню відстань для вступу в їхній дисконтний клуб, тому вона змінила правила прийому до клубу, щоб заохочувати число перельотів так само, як і накопичену відстань.

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 й 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 тощо. 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 постає в молекулярній генетиці та генній інженерії. Тут воно 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.

Прикладна хімія. Методи Data Mining знаходять широке застосування в прикладній хімії (органічній та неорганічній). 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 системній організації. Вони відносяться головним чином до надкібернетичного рівня організації систем, закономірності якого не можуть бути достатньо точно описані на мові статистичних чи інших аналітичних математичних моделей. Дані в зазначених сферах 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 одна з одною. Наприклад, дослідження, проведене в супермаркеті, може показати, що 65% тих, хто купив кукурудзяні чіпси, беруть також і «Кока-колу», а за наявності знижки за такий комплект «Колу» придбають у 85% випадків. Маючи в своєму розпорядженні відомості про подібну асоціацію, менеджерам легко оцінити, наскільки дієво 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 вирішують придбати холодильник.

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 шаблони, які адекватно відображають динаміку поведінки

цільових показників, є ймовірність, що за їх допомогою можна передбачити і поведінку системи в майбутньому.

4. Класи систем Data Mining

Data Mining є мультидисциплінарною галуззю, яка виникла і розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних тощо (рис. 1.1). Звідси велика кількість **методів і алгоритмів**, реалізованих у різних діючих системах Data Mining. Багато з таких систем інтегрують у собі відразу кілька підходів. Проте, як правило, в кожній системі є якась ключова компонента (рис. 1.2).

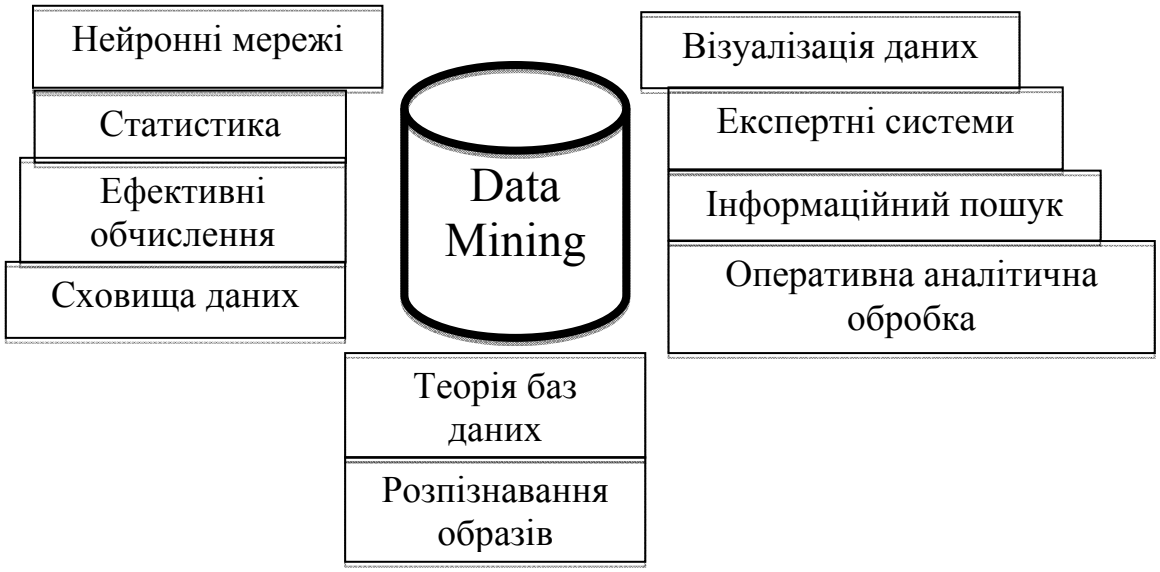

Рисунок 1.1 – Data Mining – мультидисциплінарна галузь

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 у своїй сфері специфіку (професійна мова, системи різних індексів). На ринку є безліч програм цього класу. Як правило, вони досить дешеві (зазвичай коштують близько $300-1000).

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

Рисунок 1.2 – Популярні продукти для Data Mining

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.

Як приклади найбільш потужних і поширених статистичних пакетів можна назвати SAS (компанія SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), Statistica, STADIA та ін.

Нейронні мережі. Це великий клас систем, архітектура яких має аналогію (як тепер відомо, досить слабку) з побудовою нервової тканини з нейронів. В одній із 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 тощо. Ці значення розглядаються як сигнали, що передаються в наступний шар, ослабляючись або посилюючись в залежності від числових значень (ваг), приписуваних дугами між нейронними зв'язками. У результаті на виході нейрона 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 треба «натренувати» на отриманих раніше даних, для яких відомі і значення вхідних параметрів, і правильні відповіді на них. Тренування полягає в підборі ваг, 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, абсолютно не піддаються аналізу та інтерпретації людиною (відомі спроби дати інтерпретацію структурі

налаштованої нейронної мережі виглядають непереконливими – система «KINOsuite – PR»).

Приклади нейромережевих систем – це системи BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic) (рис. 1.3). Вартість їх досить значна: $1500-8000.

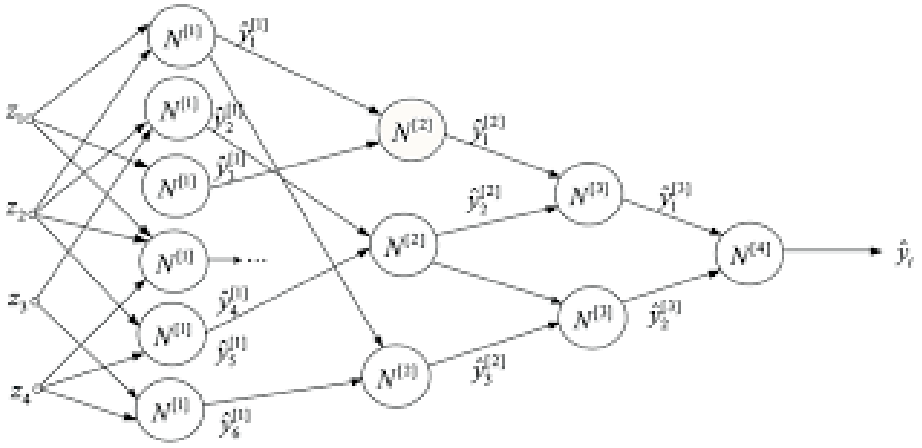

Рисунок 1.3 – Поліноміальна нейронна мережа

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 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 негативна – 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, створюючи лише ілюзію логічного висновку.

Разом із тим, більшість систем 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) (рис. 1.4). Вартість цих систем варіюється від 1 до 10 тис. дол.

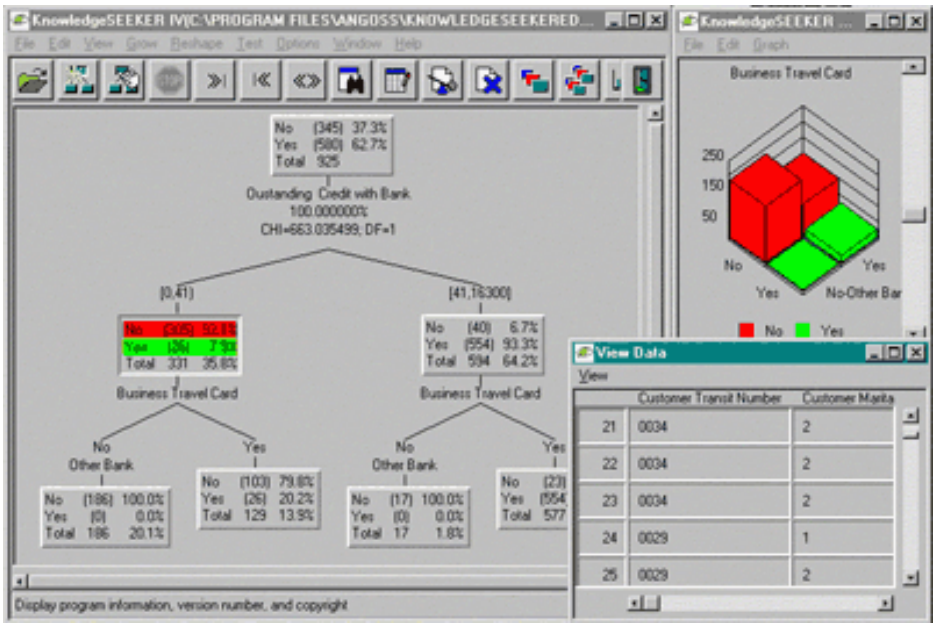

Рисунок 1.4 – Система KnowledgeSeeker обробляє банківську інформацію

Еволюційне програмування. Проілюструємо сучасний стан цього 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 виглядає як еволюція у світі програм (цим підхід трохи схожий на генетичні алгоритми). Коли система знаходить програму, яка більш або менш задовільно відображає шукану залежність, вона починає вносити до неї невеликі модифікації та 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 тощо).

Інший напрям еволюційного програмування пов'язаний з пошуком залежності цільових змінних від інших у формі функцій якогось певного виду. Наприклад, в одному з найбільш вдалих алгоритмів цього типу – метод групового урахування аргументів (МГУА) – залежність шукають у формі поліномів. У цей час системи МГУА, що реалізовані в системі NeuroShell компанії Ward Systems Group, коштують до $500.

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

Перший крок при побудові генетичних алгоритмів – це кодування вихідних логічних закономірностей у базі даних, які іменують хромосомами, а весь набір таких закономірностей називають популяцією хромосом. Далі для реалізації концепції відбору вводиться спосіб зіставлення різних хромосом, який здійснюється 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 яку-небудь непродуктивну гілку. 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 реалізації генетичного алгоритму може служити система GeneHunter фірми Ward Systems Group. Її вартість складає близько $1000.

Алгоритми обмеженого перебору були запропоновані в середині 60-х років М.М. Бонгардом для пошуку логічних закономірностей у даних. Із того часу вони продемонстрували свою ефективність при вирішенні безлічі завдань із різноманітних сфер.

Ці алгоритми обчислюють частоти комбінацій простих логічних подій у підгрупах даних. Наведемо приклад простої логічної події:

$$X = a \; ; \; X < a \; ; \; X \; a \; ; \; a < X < b, \qquad (1.1)$$

де X – параметр,

«a» і «b» – константи.

Обмеженням служить довжина комбінації простих логічних подій (у М. Бонгарда 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 (рис. 1.5). 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, залежність часу їх отримання від кількості аналізованих параметрів тощо).

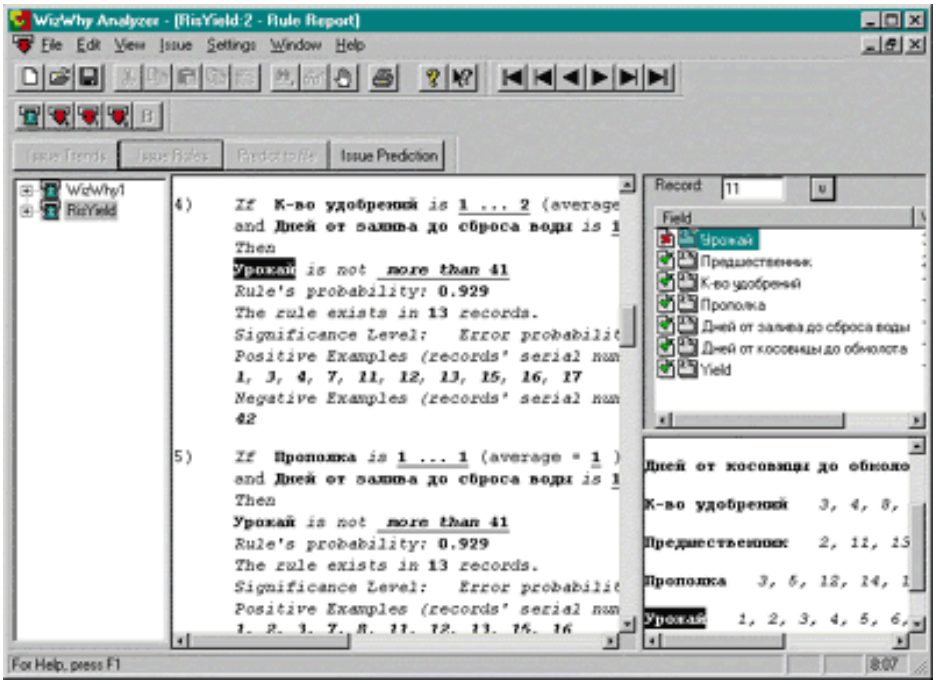

Рисунок 1.5 – Система WizWhy виявила правила, що пояснюють низьку врожайність деяких сільськогосподарських ділянок

Автор WizWhy стверджує, що його система виявляє всі 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 розв’язати. Інший недолік – система видає розв’язок за прийнятний час тільки для порівняно невеликої розмірності даних.

Проте, система WizWhy є на сьогодні одним із 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 складає близько $4000.

Системи для візуалізації багатовимірних даних. Тою чи іншою мірою засоби для графічного відображення даних підтримуються всіма системами Data Mining. Разом із тим, досить значну частку ринку займають системи, що спеціалізуються виключно на цій функції. Прикладом тут може служити програма DataMiner 3D (рис. 1.6) словацької фірми Dimension5 (5-й вимір).

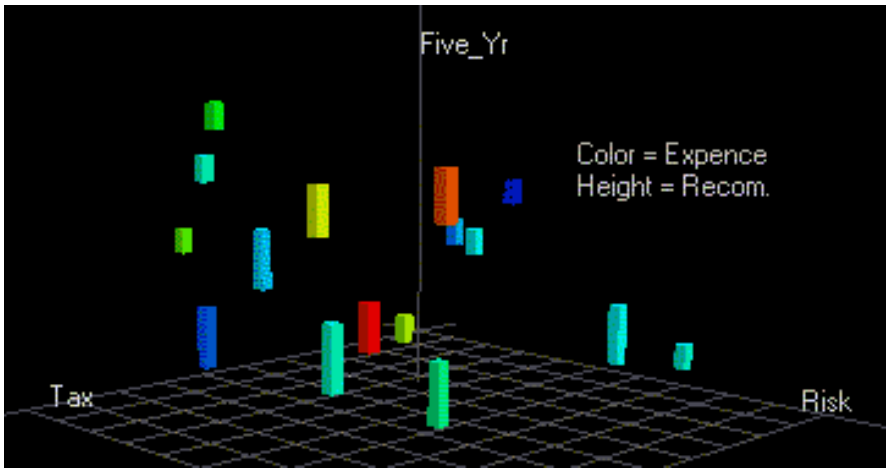

Рисунок 1.6 – Візуалізація даних системою DataMiner 3D

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.

Питання для самоконтролю

1. Навести кілька визначень поняття Data Mining.
2. Описати приклади застосування Data Mining в різних сферах економіки.
3. Які п’ять стандартних типів закономірностей, котрі реалізовані в методах Data Mining?
4. Який принцип дії алгоритма дерева рішень (decision trees).
5. Пояснити принцип дії алгоритму обмеженого перебору.
6. Пояснити різницю між методами еволюційного програмування та генетичних алгоритмів.