Лекція 1.

ОСНОВНІ ПОНЯТТЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

1.1. Сутність аналітичних технологій.

Аналітичні технології почали застосовуватися людством досить давно. Простим прикладом аналітичної технології є теорема Піфагора, яка дозволяє визначити довжину гіпотенузи маючи відомі довжини катетів по відомій формулі $c^2 = a^2 + b^2$. Іншим прикладом аналітичної технології можна назвати алгоритм обробки інформації людським мозком. Навіть мозок дитини може виконувати задачі, непідвладні сучасним комп'ютерам, наприклад, розпізнавання знайомих облич в юрбі або ефективне управління декількома десятками м'язів при грі у футбол. Унікальністю мозку є здібність до розв’язання нових задач - гри в шахи, водінню автомобіля і т.д. Але при цьому, мозок погано пристосований до обробки великих об'ємів числової інформації - людина не може перемножити два багатозначні числа, не використовуючи калькулятора або алгоритму обчислення в стовпчик. Реальні завдання з числами, набагато складніше, ніж множення і людині для вирішення таких завдань необхідні додаткові методики і інструменти.

Під *аналітичною технологією* будемо розуміти методики, які на основі певних моделей, алгоритмів, математичних теорем дозволяють за відомими даними оцінити значення невідомих характеристик і параметрів.

Аналітичні технології потрібні в першу чергу людям, що ухвалюють важливі рішення, - керівникам, аналітикам, експертам, консультантам. Дохід компанії більшою мірою визначається якістю цих рішень - точністю прогнозів, оптимальністю вибраних стратегій. І від якості цих рішень залежить розвиток компанії. За допомогою аналітичних технологій можна вирішувати проблеми прогнозування, наприклад, курсів валют, цін на сировині, попиту, доходу

компанії, рівня безробіття і оптимізації, наприклад, плану закупівель, плану інвестицій, стратегії розвитку. Слід також зазначити, що для реальних задач бізнесу і виробництва не існує чітких алгоритмів їх розв’язання. Тому керівники і експерти знаходять рішення таких задачі тільки на основі особистого досвіду. Часто класичні методики виявляються малоефективними для багатьох практичних завдань, оскільки неможливо точно описати реальність за допомогою невеликого числа параметрів моделі, або розрахунок моделі займає дуже багато часу і обчислювальних ресурсів. Аналітичні технології дозволяють створювати моделі, що істотним чином підвищують ефективність рішень [3, 24, 30].

Серед класичних підходів до аналізу даних на практиці найбільш поширеними виявилися детерміновані технології та імовірнісні технології.

Детерміновані аналітичні технології типа теореми Піфагора використовуються людиною вже багато сторіч. За цей час було створено величезну кількість формул, теорем і алгоритмів для вирішення класичних завдань - визначення об'ємів, знаходження рішень систем лінійних рівнянь, пошуку коренів багаточленів. Розроблені складні і ефективні методи аналізу задач оптимального управління, розв’язання диференціальних рівнянь і т.д. Всі ці методи діють по одній і тій же схемі (рис. 1.1.).

Рис.1.1. Схема детермінованої аналітичної технології.

Для застосування алгоритму необхідно, щоб дане завдання цілком описувалося певною детермінованою моделлю (деяким набором відомих функцій і параметрів). У такому разі алгоритм дає точна відповідь. Наприклад, для застосування теореми Піфагора потрібно перевірити, що трикутник - прямокутний.

На практиці часто зустрічаються завдання, пов'язані із спостереженням випадкових величин, наприклад, задача прогнозування курсу акцій. Для подібних проблем не можна будувати детерміновані моделі, тому застосовується принципово інший, імовірнісний підхід. Параметри імовірнісних моделей - це розподіли випадкових величин, їх середні значення, дисперсії і т.д. Як правило, ці параметри наперед невідомі, а для їх оцінки використовуються статистичні методи, вживані до вибірок зафіксованих значень (історичних даних) (рис. 1.2.).

Рис.1.2. Схема імовірнісної аналітичної технології.

Такого роду методи припускають, що відома деяка імовірнісна модель задачі. Наприклад, в задачі прогнозування курсу можна припустити, що завтрашній курс акцій залежить тільки від курсу за останні 2 дні (авторегресійна модель). Якщо це вірно, то спостереження курсу впродовж декількох місяців дозволяють досить точно оцінити коефіцієнти цієї залежності і прогнозувати курс в майбутньому (рис. 1.3.).

Рис.1.3. Схема прогнозування курсу акцій.

На жаль, класичні методики виявляються малоефективними в багатьох практичних задачах. Це пов'язано з тим, що неможливе достатньо повно описати реальність за допомогою невеликого числа параметрів моделі, або розрахунок моделі вимагає дуже багато час і обчислювальних ресурсів. Зокрема, розглянемо проблеми, що виникають при пошуку рішення задачі оптимального розподілу інвестицій [19]:

- у реальній задачі жодна з функцій не відома точно, а відомі лише приблизні або очікувані значення прибутку. Для того, щоб позбавитися від невизначеності, ми вимушені зафіксувати функції, втрачаючи при цьому в точності опису задачі;
- детермінований алгоритм для пошуку оптимального рішення може бути застосовний тільки в тому випадку, якщо всі дані функції лінійні. У реальних задачах бізнесу ця умова не виконується. Хоча дані функції можна апроксимувати лінійними, рішення в цьому випадку буде далеким від оптимального;
- якщо одна з функцій нелінійна, то симплекс-метод непридатний, і залишається два традиційні шляхи пошуку рішення цієї задачі. Перший шлях - використовувати метод градієнтного спуску для пошуку максимуму прибутку. В випадку, коли область визначення функції прибутку має складну форму, а сама функція - декілька локальних максимумів, градієнтний метод може привести до неоптимального

рішення. Другий шлях - провести повний перебір варіантів інвестування. Якщо кожна з 10 функцій задана в 100 крапках, то доведеться перевірити близько 1020 варіантів, що зажадає не менші декілька місяців роботи сучасного комп'ютера.

Імовірнісні технології також володіють істотними недоліками при вирішенні практичних задач. Ми проілюстрували роботу імовірнісного підходу на прикладі простої лінійній авторегресійній моделі, проте залежності, що зустрічаються на практиці, часто нелінійні. Навіть якщо і існує проста залежність, то її вигляд наперед невідомий. Якщо ж ми хочемо враховувати для прогнозування курсу акцій декілька взаємозв'язаних чинників (наприклад, кількість операцій, курс долара і т.д.), то доведеться звернутися до побудови багатовимірної статистичної моделі. Проте, такі моделі або припускають гауссовський розподіл спостережень (що не виконується на практиці), або не обґрунтовані теоретично. У багатовимірній статистиці за відсутністю кращого нерідко застосовують малообгрунтовані евристичні методи, які за своєю суттю дуже близькі до технології нейронних мереж.

У останні роки відбувається бурхливий розвиток аналітичних систем нового типу. У їх основі - технології штучного інтелекту, що імітують природні процеси, наприклад, діяльність нейронів мозку або процес природного відбору. При розробці сучасних аналітичних технологій враховується їх здатність:

- розуміння задачі, загального процесу і знання можливостей інших систем і людей, що беруть участь у взаємодії;
- зв'язок з користувачами за допомогою розуміння природної мови, малюнків, зображень, і знаків;
- знання, засновані на здоровому глузді;
- координування ухвалення рішень, планування і дії;
- навчання на попередньому досвіді і адаптація поведінки.

Розуміння цих можливостей в людях і втілення їх при розробці програм є центральним в створеннях новітніх аналітичних технологій, здатних набувати і використовувати знання. Від зростання потужностей для проведення

інформаційного аналізу, ухвалення рішення, гнучкого проектування і виробництва залежить національна конкурентоспроможність. При додаванні інтелекту до комп'ютерних систем усуваються багато обмежень в рішенні реальних задач.

1.2. Поняття інтелектуального аналізу даних.

Більшість організацій накопичують під час своєї діяльності величезні об'єми даних, але головне, що вони хочуть від них отримати - це корисну інформацію. Як можна дізнатися з даних про те, що є вигіднішим для клієнтів організації, як розмістити ресурси ефективним чином або як мінімізувати втрати? Для вирішення цих проблем призначені новітні технології інтелектуального аналізу, які використовують для знаходження моделей і відносин, прихованих в середовищі даних, - моделей, які не можуть бути знайдені звичайними методами [9, 30, 45, 66].

Модель, як і карта - це абстрактне представлення реальності. Карта може указувати на шлях від аеропорту до будинку, але вона не може показати аварію, яка створила пробку, або ремонтні роботи, які ведуться в даний момент і вимагають об'їзду. До тих пір, поки модель не відповідає існуючим реальним відносинам, неможливо отримати сприятливий результат. Існують два види моделей: прогнозуючі і описові. Перші використовують один набір даних з відомими результатами для побудови моделей, які явним чином прогнозують результати для інших наборів, а другі описують залежності в існуючих даних, які у свою чергу використовуються для ухвалення рішень або дій. Звичайно ж, компанія, що давно знаходиться на ринку і знає своїх клієнтів користується безліччю моделей. Технології інтелектуального аналізу можуть не тільки підтвердити ці емпіричні спостереження, але і знайти нові, невідомі раніше моделі. Спочатку це може дати користувачеві лише невелику перевагу, але

якщо його об'єднати по кожному товару і кожному клієнтові, дає великий відрив від тих, хто не використовує таких технологій. З іншого боку, за допомогою методів інтелектуального аналізу можна знайти таку модель, яка приведе до радикального поліпшення у фінансовому і ринковому положенні компанії.

Термін Data Mining отримав свою назву з двох понять: пошуку цінної інформації у великій базі даних (Data) і видобутку гірської руди (Mining). Обидва процеси вимагають або просіювання величезної кількості “сирого” матеріалу, або розумного дослідження і пошуку корисних цінностей. Найчастіше Data Mining переводиться як видобуток даних, витягання інформації, розкопка даних, інтелектуальний аналіз даних, засоби пошуку закономірностей, витягання знань, аналіз шаблонів, "витягання зерен знань з гір даних", розкопка знань в базах даних, інформаційна проходка даних, "промивання" даних. Поняття "виявлення знань в базах даних" (Knowledge Discovery in Databases, KDD) можна вважати синонімом Data Mining.

Поняття Data Mining вперше з'явилося в 1978 році та придбало високу популярність в сучасному трактуванні приблизно з першої половини 1990-х років. Донині обробка і аналіз даних здійснювався в рамках прикладної статистики, при цьому в основному вирішувалися завдання обробки невеликих баз даних.

У основу сучасної технології Data Mining покладена концепція шаблонів (паттернів), що відображають фрагменти багатоаспектних взаємин в даних. Ці шаблони є закономірностями, властивими підвибіркам даних, які можуть бути компактно виражені в зрозумілій людині формі. Пошук шаблонів проводиться методами, не обмеженими рамками апріорних припущень про структуру вибірки і виду розподілів значень аналізованих показників.

Важливе положення Data Mining - нетривіальність розшукуваних шаблонів. Це означає, що знайдені шаблони повинні відображати неочевидні, несподівані (unexpected) регулярності в даних, такі, що становлять так звані приховані знання (hidden Knowledge). До суспільства прийшло розуміння, що

неякісні дані (raw Data) містять глибинний пласт знань, при грамотній розкопці якого можуть бути виявлені справжні самородки (рис.1.4.).

В цілому технологію Data Mining достатньо точно визначає Григорій Піатецький-Шапіро (Gregory Piatetsky-Shapiro) - один із засновників цього напряму: «*Data Mining* - це процес виявлення в “сирих” даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для ухвалення рішень в різних сферах людської діяльності» [57]**.**

Рис. 1.4. Рівні знань, які вилучаються зданих.

Суть і мету технології Data Mining можна охарактеризувати так: це технологія, яка призначена для пошуку у великих об'ємах 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 повністю відповідатимуть дійсності, на відміну від експертної думки, яка завжди є суб'єктивною. *Практично корисних* - це означає, що висновки мають конкретне значення, якому можна знайти практичне застосування [4, 9, 45, 58].

Приведемо ще декілька визначень поняття Data Mining.

Data Mining - це процес виділення, дослідження і моделювання великих об'ємів даних для виявлення невідомих до цього структур (patterns) з метою досягнення переваг в бізнесі (визначення SAS Institute) [73].

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

Data Mining - мультидисциплінарна область, що виникла і розвивалася на базі таких наук як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних та інше (рис. 1.5.).

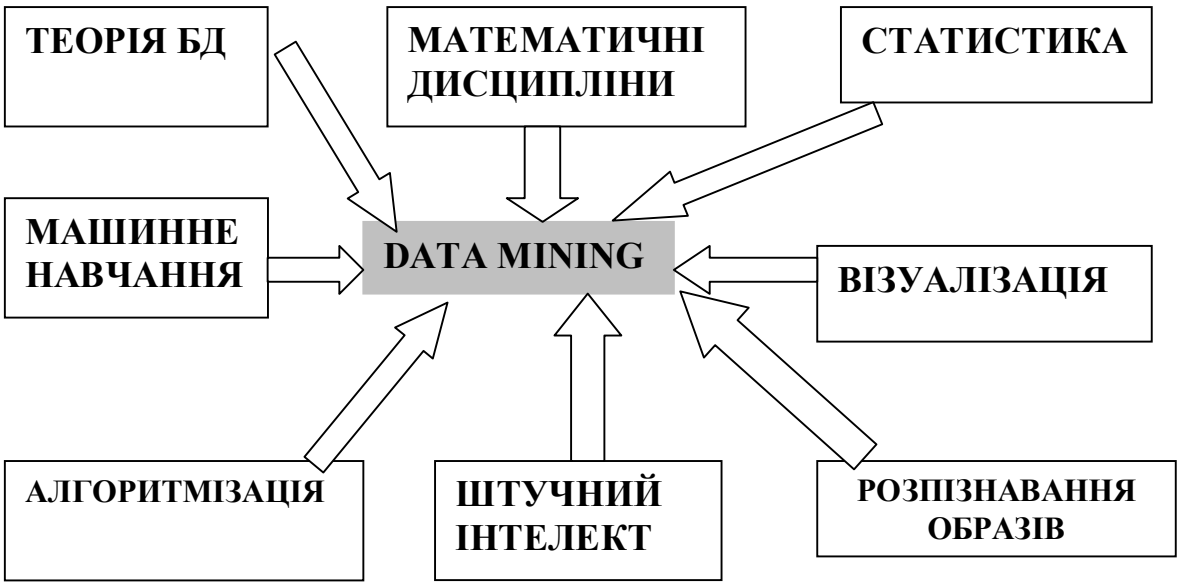

Рис. 1.5. Data Mining як мультидисциплінарна область.

Розглянемо сутність деяких дисциплін, на стику яких з'явилася технологія Data Mining:

- *статистика* - це наука про методи збору даних, їх обробки і аналізу для виявлення закономірностей, властивих явищу, що вивчається. Статистика є сукупністю методів планування експерименту, збору даних, їх уявлення і

узагальнення, а також аналізу і отримання висновків на підставі цих даних. Вона оперує даними, отриманими в результаті спостережень або експериментів;

- *машинне навчання* можна охарактеризувати як процес отримання програмою нових знань. Мітчелл в 1996 році дав таке визначення: "Машинне навчання - це наука, яка вивчає комп'ютерні алгоритми, що автоматично поліпшуються під час роботи". Одним з найбільш популярних прикладів алгоритму машинного навчання є нейронні мережі [52];
- *штучний інтелект* - науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, що традиційно вважаються інтелектуальними. Термін інтелект (intelligence) походить від латинського intellectus, що означає розум, розумові здібності людини. Відповідно, штучний інтелект (AI, Artificial Intelligence) тлумачиться як властивість автоматичних систем брати на себе окремі функції інтелекту людини. Штучним інтелектом називають властивість інтелектуальних систем виконувати творчі функції, які традиційно вважаються прерогативою людини.

Поняття Data Mining також тісно пов'язане з *технологіями баз даних* і поняттям дані. Дослідження історичного аспекту цієї проблеми дозволяє виявити основні моменти, пов’язані з появленням систем інтелектуального аналізу даних. Так у 1968 році була введена в експлуатацію перша промислова СУБД система IMS фірми IBM. Вже у 1975 році з'явився перший стандарт асоціації по мовах систем обробки даних - Conference on Data System Languages (CODASYL), що визначив ряд фундаментальних понять в теорії систем баз даних, які до цих пір є основоположними для мережевої моделі даних. Подальший розвиток теорії баз даних пов'язаний з ім’ям американського математика Е.Ф. Кодда, який є творцем реляційної моделі даних. Протягом 80-х років багато дослідників експериментували з новим підходом в напрямах структуризації баз даних і забезпечення до них доступу. Метою цих пошуків було отримання реляційних прототипів для простішого моделювання даних. В

результаті, в 1985 році була створена мова, названий SQL. На сьогоднішній день практично всі СУБД забезпечують даний інтерфейс. Приблизно у цей час з'явилися специфічні типи даних - "графічний образ", "документ", "звук", "карта", типи даних для часу, інтервалів часу, символьних рядків з двобайтовим представленням символів були додані в мову SQL. Результати цих всіх зусиль зробили можливим появу технології Data Mining, сховищ даних, мультимедійних баз даних і web-баз даних [3, 9, 58, 66].

Для успішного проведення процесу знаходження нового знання необхідною умовою є наявність сховища даних. *Сховище даних* - це предметно-орієнтований, інтегрований, прив'язаний до часу, незмінний збір даних для підтримки процесу ухвалення рішень. *Предметна орієнтація* означає, що дані об'єднані в категорії і зберігаються відповідно областям, що вони описують, а не до застосувань, що їх використовують. *Інтегрованість* означає, що дані задовольняють вимогам всього підприємства, а не одній функції бізнесу. Цим, сховище даних гарантує, що однакові звіти, що згенерували для різних аналітиків, міститимуть однакові результати. *Прив'язка до часу* означає, що сховище можна розглядати як сукупність "історичних" даних тобто можна відновити картину на будь-який момент часу. Атрибут часу завжди явно присутній в структурах сховища даних. *Незмінність* означає, що, потрапивши один раз в сховище, дані там зберігаються і не змінюються. У сховищі дані лише додаються. Для організації і експлуатації інформаційного сховища створюється спеціалізоване програмне забезпечення, яке забезпечує ефективну взаємодію з користувачем.

Ключовою можливістю застосування новітніх технологій стало величезне падіння ціни за останні декілька років на пристрої зберігання інформації з десятків доларів за зберігання мегабайта інформації, до десятків центів. Це істотним чином здешевило і збільшило можливості збору і зберігання великих об'ємів інформації. Падіння цін на процесори з одночасним збільшенням їх швидкодії сприяло розвитку технологій, пов'язаних з обробкою величезних масивів інформації. В результаті цього була подолана безліч

бар'єрів, що зустрічаються при знаходженні нового знання в сховищах інформації. Клієнт-серверная архітектура також є необхідним атрибутом технології інтелектуального аналізу даних. Такий підхід надає можливості виконувати найбільш трудомісткі процедури обробки даних на високопродуктивному сервері як розробникам проектів, так і користувачам. На цьому ж сервері можуть зберігатися, і по запитах клієнтів, виконуватися корпоративні проекти.

Технології інтелектуального аналізу даних є важливою частиною ринку сучасних інформаційних технологій. Агентство Gartner Group, що займається аналізом ринків інформаційних технологій, в 1980-х роках ввело термін "Business Intelligence" (BI), діловий інтелект або бізнес-інтелект. Цей термін запропонований для опису різних концепцій і методів, які покращують бізнес рішення шляхом використання систем підтримки ухвалення рішень. У 1996 році агентство уточнило визначення даного терміну: «Business Intelligence - програмні засоби, функціонуючі в рамках підприємства і забезпечуючи функції доступу і аналізу інформації, яка знаходиться в сховищі даних, а також що забезпечують ухвалення правильних і обґрунтованих управлінських рішень». Поняття BI об'єднує в собі різні засоби і технології аналізу і обробки даних масштабу підприємства. На основі цих засобів створюються BI-системи, мета яких - підвищити якість інформації для ухвалення управлінських рішень. BI-системи також відомі під назвою Систем Підтримки Ухвалення Рішень (СППР, DSS, Decision Support System). Ці системи перетворюють дані в інформацію, на основі якої можна ухвалювати рішення, тобто що підтримує ухвалення рішень. Gartner Group визначає склад ринку систем Business Intelligence як набір програмних продуктів наступних класів [36, 40, 58]:

- засоби побудови сховищ даних (Data Warehousing, ХД);
- системи оперативної аналітичної обробки (OLAP);
- інформаційно-аналітичні системи (Enterprise Information Systems, EIS);
- засоби інтелектуального аналізу даних (Data Mining);

- інструменти для виконання запитів і побудови звітів (Query and Reporting tools).

Класифікація Gartner базується на методі функціональних задач, де програмні продукти кожного класу виконують певний набір функцій або операцій з використанням спеціальних технологій.

Приведемо декілька коротких цитат найбільш впливових членів бізнес-організацій, які є експертами в цій відносно новій технології. Керівництво з придбання продуктів Data Mining (Enterprise Data Mining Buying Guide) компанії Aberdeen Group: "Data Mining - технологія здобичі корисної інформації з баз даних. Проте у зв'язку з істотними відмінностями між інструментами, досвідом і фінансовим станом постачальників продуктів, підприємствам необхідно ретельно оцінювати передбачуваних розробників Data Mining і партнерів. Щоб максимально використовувати потужність інструментів Data Mining комерційного рівня, підприємству необхідно вибрати, очистити і перетворити дані, іноді інтегрувати інформацію, здобуту із зовнішніх джерел, і встановити спеціальне середовище для роботи Data Mining алгоритмів. Результати Data Mining великою мірою залежать від рівня підготовки даних, а не від "чудових можливостей" якогось алгоритму або набору алгоритмів. Близько 75% роботи над Data Mining полягає в зборі даних, який здійснюється ще до того, як запускаються самі інструменти. Безграмотно застосувавши деякі інструменти, підприємство може безглуздо розтратити свій потенціал, а іноді і мільйони доларів" [72].

Думка Херба Едельштайна (Herb Edelstein), відомого в світі експерта в області Data Mining, сховищ даних і CRM: "Недавнє дослідження компанії Two Crows показало, що Data Mining знаходиться все ще на ранній стадії розвитку. Багато організацій цікавляться цією технологією, але лише деякі активно упроваджують такі проекти. Вдалося з'ясувати ще один важливий момент: процес реалізації Data Mining на практиці виявляється складнішим, ніж очікується. IT-команди захопилися міфом про те, що засоби Data Mining прості у використанні. Передбачається, що досить запустити такий інструмент на

терабайтной базі даних, і вмить з'явиться корисна інформація. Насправді, успішний Data Mining-проект вимагає розуміння суті діяльності, знання даних і інструментів, а також процесу аналізу даних" [61].

Перш ніж використовувати технологію Data Mining, необхідно ретельно проаналізувати її проблеми, обмеження і критичні питання, з нею зв'язані, а також зрозуміти, чого ця технологія не може. Зокрема, Data Mining не може замінити аналітика. Також технологія не може дати відповіді на ті питання, які не були задані. Вона не може замінити аналітика, а всього лише дає йому могутній інструмент для полегшення і поліпшення його роботи. Оскільки дана технологія є мультидисциплінарною областю, для розробки додатку, включаючого Data Mining, необхідно задіювати фахівців з різних областей, а також забезпечити їх якісну взаємодію.

Різні інструменти Data Mining мають різний ступінь "доброзичливості" інтерфейсу і вимагають певної кваліфікації користувача. Тому програмне забезпечення повинне відповідати рівню підготовки користувача. Використання Data Mining повинне бути нерозривно пов'язане з підвищенням кваліфікації користувача. Проте фахівців з Data Mining, які б добре розбиралися в бізнесі, поки що мало [2, 25].

Успішний аналіз вимагає якісної передобробки даних. За ствердженням аналітиків і користувачів баз даних, процес передобробки може зайняти до 80% відсотків всього Data Mining-процесу. Тому, щоб технологія працювала на себе, буде потрібно багато зусиль і часу, які йдуть на попередній аналіз даних, вибір моделі і її коректування.

За допомогою Data Mining можна відшукувати дійсно дуже цінну інформацію, яка незабаром дасть великі дивіденди у вигляді фінансової і конкурентної вигоди. Проте Data Mining достатньо часто робить безліч помилкових відкриттів, що не мають сенсу. Багато фахівців стверджують, що засоби Data Mining можуть видавати величезну кількість статистично недостовірних результатів. Щоб цього уникнути, необхідна перевірка адекватності отриманих моделей на тестових даних.

Слід також зазначити, що якісна Data Mining-програма може коштувати достатньо дорого для компанії. Варіантом служить придбання вже готового рішення з попередньою перевіркою його використання, наприклад на демо-версії з невеликою вибіркою даних. Засоби Data Mining, на відміну від статистичних, теоретично не вимагають наявності строго певної кількості ретроспективних даних. Ця особливість може стати причиною виявлення недостовірних, помилкових моделей і, як результат, ухвалення на їх основі невірних рішень. Необхідно здійснювати контроль статистичної значущості виявлених знань.

Data Mining має досить суттєві відмінності від інших методів аналізу даних. Традиційні методи аналізу даних (статистичні методи) і OLAP в основному орієнтовані на перевірку наперед сформульованих гіпотез (verification-driven Data Mining) і на "грубий" розвідувальний аналіз, що становить основу оперативної аналітичної обробки даних (Online Analytical Processing, OLAP), тоді як одне з основних положень Data Mining - пошук неочевидних закономірностей. Інструменти Data Mining можуть знаходити такі закономірності самостійно і також самостійно будувати гіпотези про взаємозв'язки. Оскільки саме формулювання гіпотези щодо залежностей є найскладнішим завданням, перевага Data Mining в порівнянні з іншими методами аналізу є очевидною. Більшість статистичних методів для виявлення взаємозв'язків в даних використовують концепцію усереднювання по вибірці, що приводить до операцій над неіснуючими величинами, тоді як Data Mining оперує реальними значеннями. OLAP більше підходить для розуміння ретроспективних даних, Data Mining спирається на ретроспективні дані для отримання відповідей на питання про майбутньому.

Потенціал Data Mining дає "зелене світло" для розширення меж застосування технології. Щодо перспектив Data Mining можливі наступні напрями розвитку:

- виділення типів предметних областей з відповідними ним евристиками, формалізація яких полегшить рішення відповідних задач Data Mining, що відносяться до цих областей;
- створення формальних мов і логічних засобів, за допомогою яких буде формалізовані міркування і автоматизація яких стане інструментом рішення задач Data Mining в конкретних наочних областях;
- створення методів Data Mining, здатних не тільки витягувати з даних закономірності, але і формувати якісь теорії, що спираються на емпіричні дані;
- подолання істотного відставання можливостей інструментальних засобів Data Mining від теоретичних досягнень в цій області.

Якщо розглядати майбутнє Data Mining в короткостроковій перспективі, то очевидно, що розвиток цієї технології найбільш направлений до областей, пов'язаних з бізнесом. Продукти Data Mining можуть стати такими ж звичайними і необхідними, як електронна пошта, і, наприклад, використовуватися користувачами для пошуку найнижчих цін на певний товар або найбільш дешевих квитків.

У довгостроковій перспективі майбутнє Data Mining є таким, що дійсно захоплює - це може бути пошук інтелектуальними агентами як нового вигляду лікування різних захворювань, так і нового розуміння природи всесвіту.

Проте Data Mining таїть в собі і потенційну небезпеку - адже все більша кількість інформації стає доступною через всесвітню мережу, у тому числі і відомості приватного характеру, і все більше знань можливо добути з неї. Не так давно найбільший онлайновий магазин "Amazon" опинився в центрі скандалу з приводу отриманого ним патенту "Методи і системи допомоги користувачам при покупці товарів", який є не що інше як черговий продукт Data Mining, призначений для збору персональних даних про відвідувачів магазина. Нова методика дозволяє прогнозувати майбутні запити на підставі фактів покупок, а також робити висновки про їх призначення. Мета даної методики - те, про що мовилося вище - отримання як можна більшої кількості

інформації про клієнтів, у тому числі і приватного характеру (стан, вік, переваги і т.д.). Таким чином, збираються дані про приватне життя покупців магазина, а також членів їх сімей, включаючи дітей. Останнє заборонене законодавством багатьох країн - збір інформації про неповнолітніх можливий там тільки з дозволу батьків.

Дослідження відзначають, що існують як успішні рішення, які використовують Data Mining, так і невдалий досвід застосування цієї технології. Напрями, де застосування технологій Data Mining, швидше за все, будуть успішними, мають такі особливості:

- вимагають рішень, заснованих на знаннях;
- мають навколишнє середовище, що змінюється;
- мають доступні, достатні і значущі дані;
- забезпечують високі дивіденди від правильних рішень.

Сфера застосування Data Mining нічим не обмежена - вона скрізь, де є які-небудь дані. Але в першу чергу методи Data Mining сьогодні, м'яко кажучи, заінтригували комерційні підприємства, що розгортають проекти на основі інформаційних сховищ даних (Data Warehousing). Досвід багато таких підприємств показує, що віддача від використання Data Mining може досягати 1000%. Наприклад, відомі повідомлення про економічний ефект, в 10-70 разів що перевищив первинні витрати від 350 до 750 тис. дол.; про проект в 20 млн. дол., який окупився всього за 4 місяці. Інший приклад - річна економія 700 тис. дол. за рахунок впровадження Data Mining в мережі універсамів у Великобританії. Data Mining представляють велику цінність для керівників і аналітиків в їх повсякденній діяльності. Ділові люди усвідомили, що за допомогою методів Data Mining вони можуть отримати відчутні переваги в конкурентній боротьбі [19, 30, 58,66].

Розглянемо деякі бізнес-приклади Data Mining.

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 задачі, які можна вирішувати за допомогою Data Mining у сфері роздрібної торгівлі:

- аналіз купівельної корзини (аналіз схожості) призначений для виявлення товарів, яких покупці прагнуть придбати разом. Знання купівельної корзини необхідне для поліпшення реклами, вироблення стратегії створення запасів товарів і способів їх розкладки в торгових залах;
- дослідження часових 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 батареї і плівку?";
- створення прогнозуючих моделей дає можливість торговим підприємствам дізнаватися характер потреб різних категорій клієнтів з певною поведінкою, наприклад, таких, що купують товари відомих дизайнерів або таких, що відвідують розпродажі. Ці знання потрібні для розробки точно направлених, 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 енергійніше просувати свої програми маркетингу і

ціноутворення, щоб утримувати існуючих клієнтів і привертати нових. Серед типових заходів відзначимо наступні:

- аналіз записів про докладні характеристики викликів. Призначення такого аналізу - виявлення категорій клієнтів з схожими стереотипами користування їх послугами і 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 вірогідність залишиться їй вірними. У результаті засоби, що виділяються на маркетинг, можна витрачати там, де віддача більше всього.

Страхування. Страхові компанії протягом ряду років накопичують великі об'єми даних. Тут обширне поле діяльності для методів Data Mining:

- виявлення шахрайства. Страхові компанії можуть понизити рівень шахрайства, відшукуючи певні стереотипи в заявах про виплату страхового відшкодування, що характеризують взаємини між юристами, лікарями і заявниками;
- аналіз ризику. Шляхом виявлення поєднань чинників, пов'язаних із сплаченими заявами, страховики можуть зменшити свої втрати за зобов'язаннями. Відомий випадок, коли в США крупна страхова компанія виявила, що суми, виплачені за заявами людей, що є в браку, удвічі перевищує суми за заявами самотніх людей. Компанія відреагувала на це нове знання переглядом своєї загальної політики надання знижок сімейним клієнтам.

Управління виробництвом, менеджмент якості. Шляхом аналізу даних автоматизованого виробництва і відхилень від нього можна ідентифікувати проблеми на етапах виробництва як з погляду якості, так і з погляду збереження темпу виробництва. На підставі такої встановленої інформації можна, наприклад, у виробничий процес ввести етап додаткового контролю, завдяки якому вже в процесі виробництва будуть виявлені розроблені вироби, які після закінчення виробничого процесу не пройдуть вихідний контроль.

Молекулярна генетика і генна інженерія. Мабуть найгостріше завдання виявлення закономірностей в експериментальних даних стоїть в молекулярній генетиці і генній інженерії. Тут вона формулюється як визначення так званих маркерів, під якими розуміють генетичні коди, контролюючі ті або інші фенотипічні ознаки живого організму. Такі коди можуть містити сотні, тисячі і більш зв'язаних елементів. На розвиток генетичних досліджень виділяються великі кошти. Останнім часом в даній області виник особливий інтерес до застосування методів Data Mining. Відомо декілька крупних фірм, що спеціалізуються на застосуванні цих методів для розшифровки генома людини і рослин.

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 медикаментозної дії, визначати свідчення - протипоказання, орієнтуватися в лікувальних процедурах, створювати умови найбільш ефективного лікування, передбачати результати призначеного курсу лікування і т.п. Технології Data Mining дозволяють виявляти в медичних даних шаблони, які складають основу вказаних правил.

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

Можна привести ще багато прикладів різних областей знання, де методи Data Mining грають провідну роль. Особливість цих областей полягає в їх складній системній організації. Вони відносяться головним чином до суперскладного рівня організації систем, закономірності якого не можуть бути достатньо точно описані на мові статистичних або інших аналітичних

математичних моделей. Дані у вказаних областях неоднорідні, гетерогенні, нестаціонарні і часто відрізняються високою розмірністю.

1.3. Етапи та методи знаходження нових знань.

Важливо розуміти, що побудова моделі інтелектуального аналізу даних є складовою частиною масштабнішого процесу, який включає всі етапи, починаючи з визначення базової проблеми, яку модель вирішуватиме, до розгортання моделі в робочому середовищі. Даний процес може бути заданий за допомогою наступних шести базових кроків (рис. 1. 6.):

1. Постановка задачі.
2. Підготовка та огляд даних:

- оцінювання даних;
- об'єднання і очищення даних;
- відбір даних;
- перетворення.

3. Побудова моделей:

- оцінка і інтерпретація;
- зовнішня перевірка.

4. Використання моделей.
5. Нагляд за моделлю.

Хоча процес, що ілюструється за допомогою рисунку 1.6, носить циклічний характер, кожен крок не обов'язково веде безпосередньо до наступного кроку. Створення моделі інтелектуального аналізу даних є динамічний ітеративний процес. Виконавши огляд даних, користувач може виявити, що існуючих даних недостатньо для створення необхідних моделей інтелектуального аналізу даних, що, відповідно, веде до необхідності пошуку додаткових даних. Можна розробити декілька моделей і зрозуміти, що вони не

вирішують сформульованої задачі. Отже, потрібна зміна характеристик завдання.

Рис 1.6. Етапи інтелектуального аналізу даних.

Може виникнути необхідність в оновленні вже розгорнених моделей за рахунок нових даних, що поступили. Таким чином, важливо розуміти, що створення моделі інтелектуального аналізу даних є процесом і що кожен крок такого процесу може бути повторений стільки раз, скільки необхідно для створення ефективної моделі. Розглянемо докладніше кожний з цих етапів:

*Визначення проблеми***.** Для того, щоб повніше використовувати всі переваги інтелектуальних технологій необхідно ясно представити мету майбутнього аналізу. Побудова моделі проводиться залежно від мети. Якщо необхідно збільшити прибуток торгової організації, то для цілей: "збільшення кількості продажів" і "збільшення ефективності реклами" необхідно будувати різні моделі. На цьому ж етапі визначаються способи оцінювання результатів майбутнього проекту і можливі витрати на його реалізацію.

Підготовка та огляд даних. Це є найтриваліший етап, який може займати від 50% до 85% часу всього процесу знаходження нового знання. На цьому етапі необхідно визначити джерела отримання даних. Це можуть бути дані, накоплені самою організацією або зовнішні дані від загальнодоступних

джерел (відомості про погоду або перепис населення) або приватних джерел (різні архівні дані, бази нотаріальних контор і ін.).

Оцінювання даних. При побудові моделі необхідно пам'ятати одне правило, що стосується коректності вхідних даних: "Якщо на вхід задачі поступає "сміття", то і результатом теж буде "сміття". Вхідні дані можуть знаходитися в одній базі або в декількох. Перед "завантаженням" даних в сховище необхідно врахувати, що різні джерела даних можуть бути спроектовані під певні задачі і, відповідно, виникають проблеми, пов'язані з об'єднанням даних: різні формати представлення числових даних (наприклад, цілі або вещественні); різне кодування даних (наприклад, різний формат дат); різні способи зберігання даних; різні одиниці вимірювання (дюйми і сантиметри); а також частота збору даних і дата останнього оновлення. Навіть, якщо дані знаходяться в одній базі, то все одно треба звертати увагу на пропущені значення і значення, нереальної величини, так звані "викиди". Аналітик винен завжди знати, як, де і за яких умов збираються дані, і бути упевненим, що всі дані, які використовуються для проведення аналізу зміряні однаковим способом.

Об'єднання і очищення даних. На цьому етапі відбувається побудова сховища даних для подальшої обробки, тобто, відбувається наповнення сховища або додавання до нього даних, відібраних на попередніх етапах. В цей же час відбувається очищення, тобто виправлення всіх виявлених помилок. Існують різні аспекти очищення даних. Всі вони направлені на знаходження і виправлення помилок, допущених на етапі збору інформації. Помилкою в даних можуть вважатися: пропущене значення, неможлива подія (невірно набране значення - "викид"). Корекція відбувається на основі здорового глузду, використання правил або із залученням експерта, добре обізнаного з предметною областю. Запис в базі даних, в якому є помилка, повинен бути виправлений або, в спірних випадках, виключений з подальшого розгляду. Після перевірки даних, вони перетворюються і форматуються відповідно результатам оцінювання. Це робиться для більшої зручності спостереження за

даними. Дані дискретних подій перетворюються на спеціально розроблену або стандартну форму, в якій відбивається час і опис подій. Якщо користувачі легко розбиратимуться в цій формі, вони зможуть швидко вивчити події, які були в основі побудови цієї форми. Може здатися, що цей крок дублює етап збору даних, але насправді це два зовсім різних етапа. На першому з них відбувається відбір даних для прискорення машинної обробки інформації без втрати якості, на другому дані доводяться до вигляду, зручному для візуального контролю користувача. Людина, яка проводить аналіз, може повніше уявити собі вхідні дані. Це необхідно для різного роду звітів, коли потрібно коротко охарактеризувати вхідні дані, вживані для аналізу.

Відбір даних. Якщо сховище сформоване і визначені типи моделей, які будуть побудовані для вирішення задач, відбувається відбір даних необхідних саме для цих моделей. Мається на увазі не тільки зменшення кількості записів в базі по певній умові, але також і зміну кількості полів, злиття різних таблиць в одну, або, навпаки, створення на основі однієї таблиці декілька. Тобто, перетворення відбувається в "трьох вимірюваннях": по кількості записів, по кількості полів і по структурі.

Перетворення даних. Служить для збагачення отриманої бази, тобто додавання різних відносин на основі існуючих полів (не просто "ціна" і "кількість", а їх твір - "загальна сума", не борг і дохід, а відношення борга до доходу), додавання інтервалів (по номеру місяця можна поставити номер кварталу, а відсоток виконання плану можна доповнити характеристиками "добре", "задовільно"), додавання критичних значень (максимум, середнє, мінімум).

Побудова моделі є ітераційний процес, тобто, необхідно побудувати ряд моделей для знаходження однієї, що найбільш задовольняє поставленим цілям. Моделі можна розділити на дві групи:

- контрольовані (моделі класифікації, регресії, прогнозування часових послідовностей);
- неконтрольовані (кластеризація, асоціація і послідовність).

Після того, як визначений тип моделі, необхідно вибрати алгоритм побудови моделі або технологію знаходження знання.

Суть процесу побудови контрольованої моделі зводиться до знаходження залежностей на одній частині даних ("навчання моделі") і перевірки цих залежностей на іншій частині даних (оцінка точності). Модель вважається побудованою, якщо завершується цикл "навчання" і перевірок. Якщо точність моделі при чергових ітераціях не поліпшується, то це говорить про завершення побудови моделі. Оскільки "навчальні" і тестові дані знаходяться в одній базі даних, то часто виникає необхідність в третьому наборі даних - контрольному, який вибирається з таких даних, що не перетинаються з "навчальними" і тестовими. Він потрібний для незалежного оцінювання точності моделі. Як правило, всі три набори даних належать множині даних, необхідній для реалізації певного проекту[2, 9, 45].

Найбільш відомий тестовий метод - називається простою оцінкою. В цьому випадку розподіл даних на два набори відбувається випадковим чином. Відношення кількості тестових даних до кількості даних, на яких відбувається побудова моделі повинно бути в межах від 5% до 33%. Після побудови моделі, її використовують для передбачення значень на тестовому наборі. Мірою точності моделі вважають відношення кількості вдалих результатів до загальної кількості прикладів в тестовому наборі (можна використовувати таку змінну, як міра неточності, яка дорівнює 1, - "міра точності") .

Якщо для побудови моделі використовується не дуже велика база даних, то застосовується так звана перехресна оцінка точності. В цьому випадку дані випадковим чином діляться на дві приблизно рівні частини. Після цього модель будуватиметься на одній з них, а інша використовується для визначення точності. Потім частини бази міняються ролями. Отримані дві незалежні оцінки точності об'єднуються (як середнє арифметичне або іншим способом) для якнайкращої оцінки точності моделі, побудованої на всій базі.

Для ще менших баз, в декілька тисяч записів, використовується n-перехресна оцінка точності. В цьому випадку база ділиться на n приблизно

рівних непересічних груп. Далі перша з цих груп стає тестовим набором, а інші групи об'єднуються, і на їх основі відбувається побудова моделі. Отримана модель використовується для передбачення значень для тестового набору і таким чином виходить перше значення точності. Аналогічним чином набувають всі n незалежних значень точності. Середнє з них є точністю всієї моделі.

Ще один спосіб використовується для знаходження точності в маленьких базах даних. В цьому випадку модель будуватиметься на основі даних всієї бази. Після цього випадковим чином із записів бази створюється безліч тестових наборів (мінімум 200, а іноді навіть більше 1000). Один запис може бути присутнім в різних тестових наборах. Для будь-якої з них визначається точність. Середнє з них є точністю всієї моделі.

Після того, як побудова моделі завершена, можна корегувати модель, використовуючи інші параметри або навіть змінити алгоритм побудови моделі, оскільки ніколи не можна сказати, який алгоритм, яка технологія знаходження знання дасть кращі результати. Не можна бути упевненим, що певна технологія працюватиме краще всього. Часто доводиться будувати велику кількість моделей і оцінювати кожну для знаходження кращої. Окрім цього, для різних моделей необхідна різна підготовка даних і неминуче повторення кроків. Все це збільшує час знаходження кращої моделі, тому необхідно застосовувати технології паралельних обчислень.

*Оцінка і інтерпретація***.** Після побудови моделі необхідно оцінити результати і пояснити (інтерпретувати) їх значущість. При оцінці моделі обчислюється точність, але треба пам'ятати, що це значення вірно лише для даних, на яких модель побудована і бути готовим, що нові дані, до яких надалі застосовуватиметься модель, можуть відрізнятися від результатних невідомим чином.

Зовнішня перевірка. Висока точність моделі не є гарантією того, що модель правильно відображає реальне середовище. Однією з причин, є існування так званих неявних припущень в моделі. Тобто, сам по собі

коефіцієнт інфляції не може бути частиною моделі, що пояснює схильність покупців до покупки чи того іншого товару, але різка зміна цього коефіцієнта з 3% до 20% вже, напевно, може пояснити таку поведінку. Інша причина - це існування неминучих проблем з даними, що приводять до некоректності моделі, тому дуже важливо перевірити модель в реальному середовищі. Наприклад, якщо модель використовується для відбору кандидатів для цільової реклами, то можна зробити тестову розсилку для перевірки моделі на невеликому об'ємі даних. Якщо модель використовується для передбачення ризику неповернення кредиту, то слід піддати випробуванню цю модель на невеликій кількості претендентів на позику. Чим більше ризик, пов'язаний з некоректністю моделі, тим більше важливо провести попередні експерименти для перевірки моделі перед її експлуатацією.

Використання моделі. Після побудови і оцінки моделі, її можна використовувати різними способами. Наприклад, аналітик може подивитися групи, яка визначила модель кластеризації, графіки ефективності моделі або отримані правила. Іноді аналітик може використовувати модель для вибору деяких записів з бази даних для проведення додаткового аналізу. Базуючись на результатах використання моделі, аналітик може рекомендувати дії, які можна починати в діловій сфері. Проте, часто технології інтелектуальних обчислень - це частина автоматизованої системи (наприклад, знаходження кредитних ризиків, визначення можливості втрати клієнтів і ін.), тобто модель вбудовується в систему, яку аналітик або менеджер можуть застосовувати для ухвалення рішення. З іншого боку, модель можна включати в систему, що генерує деяку дію (наказ), коли прогнозована величина починає виходити за межі якихось значень. Загалом, методи інтелектуальних обчислень, це невелика, хоч і важлива частина кінцевого програмного продукту. Процедура знаходження знання за допомогою цих методів може об'єднуватися із знаннями експертів і застосовуватися до даних в базах.

Спостереження за моделлю. Коли модель починає працювати в реальному середовищі, то необхідно вимірювати точність моделі на реальних

даних. Проте, навіть якщо модель працює добре, і можна вважати, що робота на цьому закінчується, то все одно необхідно продовжувати спостереження за моделлю. Всі системи мають властивість розвиватися, і отримані дані (їх структура, точність, періодичність) теж міняються. Зовнішня змінна, така як коефіцієнт інфляції, своєю зміною теж можуть впливати на поведінку людей і на чинники, що впливають на цю зміну. Час від часу модель необхідно піддавати повторному тестуванню, і навіть перебудові.

Простим способом спостереження діяльності моделі є графіки розбіжностей між величинами, що передбачаються, і реальними значеннями. Вони прості для побудови і розуміння і можуть вбудовуватися в програмні продукти. Така автоматизована система може стежити сама за собою і сповіщати користувача, коли величина цих розбіжностей починає виходити за певний граничний рівень [9, 45, 58, 61].

Можна виділити принаймні шість методів виявлення і аналізу знань:

- класифікація;
- регресія;
- прогнозування часових послідовностей (рядів);
- кластеризація;
- асоціація;
- послідовність.

Перші три використовуються головним чином для передбачення, тоді як останні зручні для опису існуючих закономірностей в даних.

Класифікація - найпоширеніша модель інтелектуального аналізу даних. З її допомогою виявляються ознаки, що характеризують групу, до якої належить той або інший об'єкт. Це робиться за допомогою аналізу вже класифікованих об'єктів і формулювання деякого набору правил. Наприклад, в багатьох видах бізнесу проблемою є втрата постійних клієнтів. Класифікація допомагає виявити характеристики "нестійких" покупців і створити модель, яка передбачає, хто саме схильний піти до іншого постачальника. Використовуючи її, можна визначити ефективні види знижок і інші вигідні пропозиції, що діють

для разних покупців. Завдяки цьому, можна утримати клієнтів, витративши стільки грошей, скільки необхідно.

Певний ефективний класифікатор може використовуватися для класифікації нових записів в базі даних у вже існуючі класи і в цьому випадку він набуває характеру прогнозу. Наприклад, класифікатор, що уміє ідентифікувати ризик віддачі позики, може бути використаний для ухвалення рішення, чи великий ризик надання позики певному клієнтові. Тобто, класифікатор використовується для прогнозування вірогідності повернення позики.

Регресійний аналіз використовується, коли стосунки між змінними можуть бути виражені кількісно у вигляді деякої комбінації цих змінних. Отримана комбінація використовується для передбачення значення, яке може приймати цільова (залежна) змінна, що обчислюється на заданому наборі значень вхідних (незалежних) змінних. У простому випадку, для цього використовуються стандартні статистичні методи, такі як лінійна регресія, але більшість реальних моделей не укладаються в її рамки. Наприклад, розміри продажів або фондові ціни складні для передбачення, оскільки можуть залежати від комплексу взаємин змінних.

Прогнозування часових послідовностей. Основою для будь-яких систем прогнозування служить історична інформація, що зберігається в інформаційних сховищах у вигляді часових рядів. Якщо можна побудувати математичну модель і знайти шаблони, що адекватно відображують цю динаміку, є вірогідність, що з їх допомогою можна передбачати і поведінку системи в майбутньому. Прогнозування часових послідовностей дозволяє на основі аналізу поведінки часових рядів оцінювати майбутні значення прогнозованих змінних. Ці моделі повинні включати особливі ознаки часу: ієрархію періодів (місяць-квартал-рік), особливі відрізки часу (пяти- шести або семиденний робочий тиждень), сезонність, свята та ін.

Кластеризація відрізняється від класифікації тим, що класи заздалегідь не задані і за допомогою моделі кластеризації засоби інтелектуальних обчислень самостійно створюють однорідні групи даних.

Асоціація відноситься до аналізу структури і застосовується, коли декілька подій зв'язано між собою. Класичний приклад аналізу структури покупок відноситься до представлення придбання якої-небудь кількості товарів як одиночної економічної операції (транзакції). Оскільки велика кількість покупок відбувається в супермаркетах, а покупці для зручності використовують корзини, куди і складається весь товар, то для знаходження асоціацій служить аналіз вмісту корзини. Метою підходу є знаходження трендів (однакових ділянок) серед великого числа транзакцій, які можна використовувати для пояснення поведінки покупців. Така інформація може бути використана для регулювання запасів, зміни розміщення товарів на території магазина і ухвалення рішення по проведенню рекламної кампанії для збільшення продажів або для просування певного вигляду продукції. Наприклад, дослідження, проведене в супермаркеті, може показати, що 65% людей купують картопляні чипси, беруть також і "кока-колу", а за наявності знижки за такий комплект "кока-колу" купують в 85% випадків. Маючи такі дані, менеджерам легко оцінити, наскільки дієва надана знижка.

Хоча цей підхід прийшов виключно з роздрібної торгівлі, він може також застосовуватися у фінансовій сфері для аналізу портфеля коштовних паперів і знаходження наборів фінансових послуг, які клієнти часто купують разом. Це може використовуватися для створення деякого набору послуг, як частини кампанії по стимулюванню продажів.

Послідовність має місце, якщо існує ланцюжок зв'язаних в часі подій. Традиційний аналіз структури покупок має справу з набором товарів, які представляють одну транзакцію. Варіант такого аналізу зустрічається, якщо існує додаткова інформація (номер кредитної карти клієнта або номер його банківського рахунку) для скріплення різних покупок в єдину тимчасову серію. У такій ситуації важливе не лише співіснування даних усередині однієї

транзакції, але і порядок, в якому ці дані з'являються в різних транзакціях і час між цими транзакціями. Правила, що встановлюють ці стосунки, можуть бути використані для визначення типового набору попередніх продажів, які можуть повести за собою наступні продажі певного товару. Після покупки будинку в 45% випадків протягом місяця купується і нова кухонна плита, а в перебігу наступних двох тижнів 60% новоселів обзаводяться холодильником.

1.4. Основні моделі інтелектуальних обчислювань.

Розглянемо основні види моделей, які використовуються для знаходження нового знання на основі даних інформаційного сховища. Метою інтелектуальних технологій є знаходження нового знання, яке користувач може надалі застосувати для поліпшення результатів своєї діяльності. Результат моделювання - це виявлення відносин в даних [4, 30, 57, 70].

На практиці широке застосування знайшли такі види (алгоритми) інтелектуальних обчислень:

- нейронні мережі;
- дерева рішень;
- системи роздумів на основі аналогічних випадків;
- алгоритми визначення асоціацій і послідовностей;
- нечітка логіка;
- генетичні алгоритми;
- еволюційне програмування;
- візуалізація даних;
- комбіновані методи.

Нейронні мережі це системи з архітектурою, що умовно імітують роботу нейронів. Математична модель нейрона є деяким універсальним нелінійним елементом з можливістю широкої зміни і настроювання його характеристик. Нейронні мережі є сукупністю зв'язаних між собою прошарків

нейронів, які отримують вхідні дані, здійснюють їх обробку і генерують на виході результат. Між вузлами видимих вхідного і вихідного прошарків може знаходитися певне число прихованих прошарків. Нейронні мережі реалізують непрозорий процес. Це означає, що побудована модель, як правило, не має чіткої інтерпретації. Багато пакетів, які реалізують алгоритми нейронних мереж, застосовуються у сфері обробки комерційної інформації, при розпізнаванні образів, розшифровки рукописного тексту, інтерпретації кардіограм. Апаратні або програмні реалізації алгоритмів нейромереж називаються нейрокомп'ютером. Його основними особливостями є:

- нейрокомп'ютери дають стандартний спосіб рішення багатьох нестандартних задач. І неважливо, що спеціалізована машина краще вирішить один клас завдань. Важливіше, що один нейрокомп'ютер вирішить і цю задачу, і іншу, і третю і не треба кожного разу проектувати спеціалізовану ЕОМ, нейрокомп'ютер зробить все сам і майже не гірше;
- замість програмування застосовується навчання. Нейрокомп'ютер навчається, потрібно лише формувати навчальну множину. Таким чином, робота програміста замінюється новою роботою вчителя. Краще це чи гірше? Не те, ні друге. Програміст наказує машині виконати всі деталі роботи, вчитель створює "навчальне середовище", до якого пристосовується нейрокомп'ютер. З'являються нові можливості для роботи;
- нейрокомп'ютери ефективні там, де потрібний аналог людської інтуїції, зокрема, для розпізнавання образів, читання рукописних текстів, підготовки аналітичних прогнозів, перекладу з однієї мови на іншій і т.п. Саме для таких завдань зазвичай важко скласти явний алгоритм;
- нейронні мережі дозволяють створити ефективне програмне і математичне забезпечення для комп'ютерів з високим ступенем розпаралелювання обробки;
- нейрокомп'ютери "демократичні", вони прості, як текстові процесори, тому з ними може працювати будь-якій, навіть зовсім недосвідчений користувач (рис 1.7.).

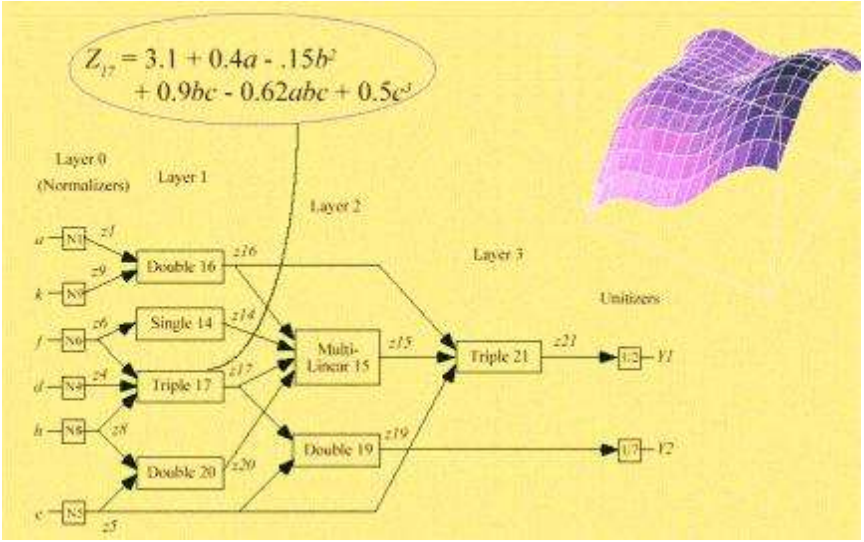

Рис.1.7. Поліномінальна нейромережа.

Дерева рішень - метод, широко вживаний в області фінансів і бізнесу, де частіше зустрічаються задачі числового прогнозу. В результаті застосування цього методу, для навчальної вибірки даних створюється ієрархічна структура правил класифікації типу, "ЯКЩО... ТОДІ...", що мають вид дерева. Для того, щоб вирішити, до якого класу віднести деякий об'єкт або ситуацію, треба відповісти на питання, що стоїть у вузлах цього дерева, починаючи з його кореня. Питання можуть мати вигляд "Значення параметра А більше Х ? " або вигляду "Значення змінної В належить підмножині ознак З ? ". Якщо відповідь позитивна, перехід до правого вузла наступного рівня, якщо негативний - то до лівого вузла; потім знову відповідь на питання, пов'язане з відповідним вузлом. Таким чином, врешті-решт, можна дійти до одного з кінцевих вузлів, де визначений клас об'єкту. Цей метод гарантує предметне уявлення правил і його легко зрозуміти (рис. 1.8.).

Сьогодні спостерігається підйом інтересу до продуктів, що застосовують дерева рішень. В основному це пояснюється тим, що більшість комерційних проблем вирішуються ними швидше, ніж алгоритмами нейтронних мереж, вони простіше і зрозуміліше для користувачів. В той же час не можна сказати, що дерева рішень завжди діють безвідмовно: для певних типів даних вони можуть виявитися неприйнятними.

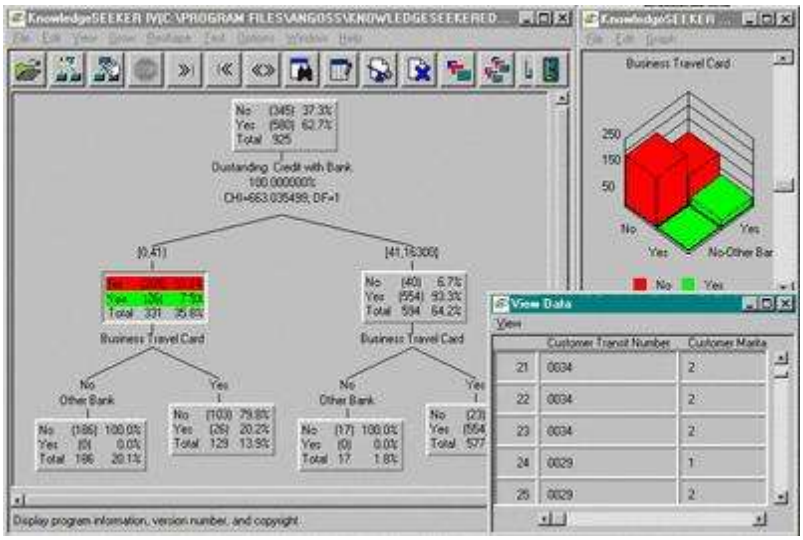

Рис.1.8. Система веде обробку банківської інформації.

Річ у тому, що окремим вузлам на кожній гілці відводиться менше число записів даних - дерево може сегментувати дані на велику кількість окремих випадків. Чим більше таких окремих випадків, тим менше навчальних прикладів потрапляє в кожен такий окремий випадок, і їх класифікація стає менш надійною. Якщо дерево дуже "гіллясте" - складається з невиправдано великого числа дрібних гілок - воно не даватиме статистично обґрунтованих відповідей. Як показує практика, у більшості систем, що використовують дерева рішень, ця проблема не знаходить задовільного рішення.

Системи роздумів на основі аналогічних випадків. Ідея алгоритму проста. Для того, щоб зробити прогноз майбутнього або вибрати правильне рішення, системи знаходять у минулому близькі аналоги наявної ситуації і вибирають ту ж відповідь, що була для них правильною. Тому, цей метод ще називають методом "найближчого сусіда". Системи роздумів на основі аналогічних випадків дають добрі результати в різних завданнях. Головний їх мінус в тому, що вони взагалі не створюють яких-небудь моделей або правил, узагальнювальних попередній досвід. У виборі рішення вони базуються на всьому масиві доступних історичних даних, тому неможливо сказати, на основі яких конкретно чинників ці системи будують свої відповіді.

Алгоритми виявлення асоціацій знаходять правила про окремі предмети, які з'являються разом в одній транзакції, наприклад в одній покупці. Послідовність - ця теж асоціація, але залежна від часу. Асоціація записується як А->Б, де А називається передумовою, Б - наслідком. Частота появи кожного окремого предмету або групи предметів, визначається дуже просто - підраховується кількість появи цього предмету у всіх подіях (покупках) і ділиться на загальну кількість подій. Ця величина вимірюється у відсотках і носить назву "поширеність". Низький рівень поширеності (менш одного тисячною відсотка) говорить про неістотність асоціації.

Для визначення важливості кожного отриманого асоціативного правила необхідно отримати величину, яка носить назву "довірчість А до Б" (взаємозв'язок А і Б). Ця величина показує, як часто з появою А з'являється Б і розраховується як відношення частоти появи (поширеності) А і Б разом до поширеності А. Тобто, якщо довірчість А до Б дорівнює 20%, то це означає, що при покупці товару А в кожному п'ятому випадку купують і товар Б. Якщо поширеність А не рівна поширеності Б, то і довірчість А до Б не дорівнює довірчості Б до А. Насправді, покупка комп'ютера частіше веде до покупки дискет, чим покупка дискети до покупки комп'ютера.

Ще однією важливою характеристикою асоціації є потужність асоціації. Чим більше потужність, то сильніше вплив, який поява А робить на появу Б. Потужність розраховується по формулі: (довірчість А до Б) / (поширеність Б).

Деякі алгоритми пошуку асоціацій спочатку сортують дані і лише після цього визначають взаємозв'язок і поширеність. Єдиною розбіжністю таких алгоритмів є швидкість або ефективність знаходження асоціацій. Це важливо, у зв'язку з величезною кількістю комбінацій, що необхідно перебрати для знаходження більш значущих правил. Алгоритми пошуку асоціацій можуть створювати свої бази даних поширеності, довірчості і потужності, до яких можна звертатися при запиті. Наприклад: "Знайти всі асоціації, в яких для товару Х довірчість більше 50% і поширеність не меншого 2,5%". При знаходженні послідовностей додається змінна часу, що дозволяє працювати з

серією подій для знаходження послідовних асоціацій впродовж деякого періоду часу (рис.1.9.).

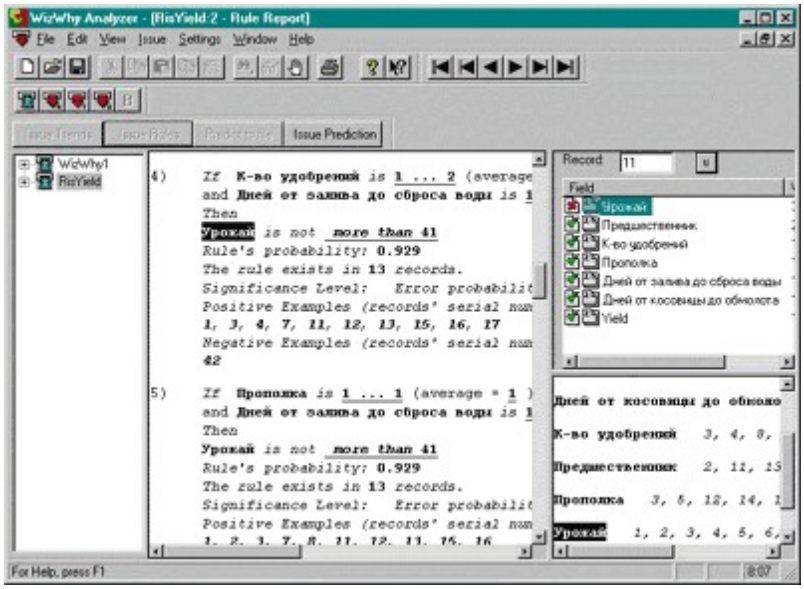

Рис.1.9. Система знайшла правила, які пояснюють низьку врожайність деяких сільськогосподарських культур.

Підводячи підсумки цьому методу аналізу, необхідно сказати, що випадково може виникнути така ситуація, коли товари в супермаркеті будуть згруповані за допомогою знайдених моделей, але це, замість очікуваного прибутку, дасть зворотний ефект. Це може відбутися через те, що клієнт довго не ходитиме по магазину у пошуках бажаного товару, купуючи при цьому ще щось, що попадається на очі, і те, що він ніколи не планував купувати.

Нечітка логіка застосовується для наборів даних, де приналежність даних до якої-небудь групи є вірогідністю в інтервалі від 0 до 1. Чітка логіка маніпулює результатами, які можуть бути або істиною, або ложью. Нечітка логіка застосовується в тих випадках, коли існує "може бути" в доповненні до "та" чи ні". Областю впровадження алгоритмів нечіткої логіки є будь-які аналітичні системи, зокрема :

- нелінійний контроль за процесами (виробництво);
- удосконалення стратегій управління і координації дій, наприклад складне промислове виробництво;
- самонавчальні системи (або класифікатори);

- дослідження ризикованих і критичних ситуацій;
- розпізнавання образів;
- фінансовий аналіз (ринки цінних паперів);
- дослідження даних (корпоративні сховища).

У Японії цей напрям переживає бум. Тут функціонує спеціально створена лабораторія Laboratory for International Fuzzy Engineering Research (LIFE). Програмою організації є створення ближчих до людини обчислювальних пристроїв. LIFE об'єднує 48 компаній в числі яких знаходяться: Hitachi, Mitsubishi, NEC, Sharp, Sony, Honda, Mazda, Toyota. З іноземних учасників LIFE можна виділити: IBM, Fuji Xerox, до діяльності LIFE також виявляє цікавість NASA [75].

Потужність і інтуїтивна простота нечіткої логіки як методології вирішення проблем гарантує її успішне використання у вбудованих системах контролю і аналізу інформації. При цьому відбувається підключення людської інтуїції і досвіду оператора. На відміну від традиційної математики, яка вимагає на кожному кроці моделювання точних і однозначних формулювань закономірностей, нечітка логіка пропонує зовсім інший рівень мислення, завдяки чому творчий процес моделювання відбувається на високому рівні абстракцій, при якому постулірується лише мінімальний набір закономірностей.

Недоліками нечітких систем є:

- відсутність стандартної методики конструювання нечітких систем;
- неможливість математичного аналізу нечітких систем існуючими методами.

Генетичні алгоритми є могутнім засобом рішення різних комбінаторних задач і проблем оптимізації. Проте, генетичні алгоритми увійшли зараз до стандартного інструментарію методів інтелектуальних обчислень. Цей метод названий так тому, що якоюсь мірою імітує процес природного відбору в природі. Хай нам треба знайти рішення задач, найбільш оптимальні з погляду деякого критерію, де кожне рішення цілком описується певним набором чисел або величин нечислової природи. Скажімо, якщо нам

треба вибрати сукупність фіксованого числа параметрів ринку, що істотно впливають на його динаміку, це буде набір імен цих параметрів. Про цей набір можна говорити як про сукупність хромосом, що визначають якості індивіда, - даного рішення поставленої задачі. Значення параметрів, що визначають рішення, називаються генами. Пошук оптимального рішення при цьому схожий на еволюцію популяції індивідів, представлених наборами хромосом.

У еволюції діють три механізми: по-перше, відбір найсильніших - наборів хромосом, яким відповідають найбільш оптимальні рішення; по-друге, схрещування - виробництво нових індивідів за допомогою змішування хромосомних наборів відібраних індивідів; і, по-третє, мутації - випадкові зміни генів у деяких індивідів популяції. В результаті зміни поколінь виробляється рішення поставленої задачі, яке вже не може бути далі покращене.

Генетичні алгоритми мають два слабкі місця. По-перше, постановка задачі не дає можливості проаналізувати статистичну значущість отримуваного з їх допомогою рішення і, по-друге, ефективно сформулювати завдання, визначити критерій відбору хромосом під силу тільки фахівцеві. Через ці чинники, генетичні алгоритми треба розглядати скоріше як інструмент наукового дослідження, чим засіб аналізу даних для практичного застосування в бізнесі і фінансах.

Еволюційне програмування наймолодша область інтелектуальних обчислень. Гіпотези про вид залежності цільової змінної від інших змінних формулюються системою у вигляді програм на деякій внутрішній мові програмування. Якщо це універсальна мова, то теоретично на ній можна виразити залежність будь-якого вигляду. Процес побудови таких програм будується як еволюція в світі програм (цим метод трохи схожий на генетичні алгоритми). Якщо система знаходить програму, яка точно виражає залежність, яка шукається, вона починає вносити до неї невеликі модифікації і відбирає серед побудованих таким чином дочірніх програм ті, які підвищують точність. Система "вирощує" декілька генетичних ліній програм, що конкурують між собою в точності знаходження шуканої залежності. Спеціальний транслюючий

модуль, перекладає знайдені залежності з внутрішньої мови системи на зрозумілий користувачеві мова (математичні формули, таблиці і ін.), роблячи їх досяжними. Для того, щоб зробити отримані результати зрозумілішими для користувача-нематематіка, існує великий арсенал різноманітних засобів візуалізації виявлених залежностей.

Пошук залежності цільових змінних від інших проводиться у формі функцій якого-небудь певного вигляду. Наприклад, в одному з найбільш вдалих алгоритмів цього типу - методі групового обліку аргументів (МГУА) залежність шукають у формі поліномів. Причому складні поліноми замінюються декількома простими, що враховують лише деякі ознаки (групи аргументів). Зазвичай використовуються попарні об'єднання ознак. Цей метод не має великих переваг в порівнянні з нейронними мережами з готовим набором стандартних нелінійних функцій, але, отримані формули залежності, в принципі, піддаються аналізу і інтерпретації (хоча на практиці це все-таки складно).

Програми візуалізації даних в певному значенні не є засобом аналізу інформації, оскільки вони тільки представляють її користувачеві. Проте, візуальне уявлення, скажімо, відразу чотирьох змінних наочно узагальнює величезні об'єми даних (рис. 1.10.).

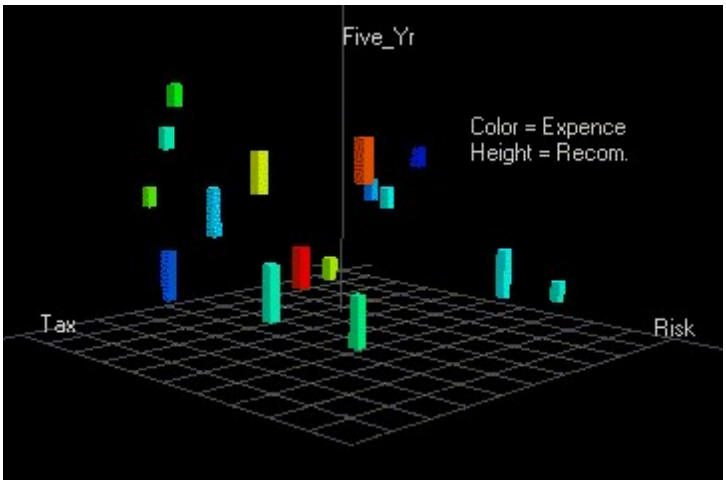

Рис. 1.10. Візуалізація показників діяльності інвестиційних фондів.

Комбіновані методи. Часто виробники об’єднують вказані підходи. Об'єднання алгоритмів нейронних мереж і технології дерев рішень сприяє побудові точнішої моделі і підвищенню швидкості. Для вирішення кожної проблеми слід шукати свій оптимальний метод.

1.5. Засоби програмної підтримки інтелектуального аналізу даних.

Сьогодні ми є свідками активного розвитку технологій інтелектуального аналізу даних, поява яких зв'язана, в першу чергу, з необхідністю аналітичної обробки надвеликих об'ємів інформації, що накопичується в сучасних сховищах даних. Можливість використання добре відомих методів математичної статистики і машинного навчання для вирішення завдань подібного роду відкрило нові можливості перед аналітиками, дослідниками, а також тими, хто ухвалює рішення - менеджерами і керівниками компаній. Аналітики передбачають, що з 2005 року такі технології стали дуже привабливою областю для IT-інвестицій. Згідно з результатами компанії Forrester Research, майже 30% європейських фірм розглядатимуть можливість впровадження цього програмного забезпечення в 2007 році. За даними іншої компанії, International Data Corporation (IDC), в країнах східної і південно-східної Азії (за винятком Японії) в 2005-2010 рр. середньорічний темп зростання рішень складе 21% [2, 17, 40, 61, 71].

На ринку програмного забезпечення Data Mining існує величезна різноманітність продуктів, що відносяться до цієї категорії. І не розгубитися в них достатньо складно. Для вибору продукту слід ретельно вивчити поставлені задачі і позначити ті результати, які необхідно отримати.

На початку 90-х років минулого сторіччя ринок Data Mining налічував близько десяти постачальників. У середині 90-х число постачальників, представлених компаніями малого, середнього і великого розміру, налічувало більше 50 фірм. Зараз до аналітичних технологій, зокрема до Data Mining,

виявляється величезний інтерес. На цьому ринку працює безліч фірм, орієнтованих на створення інструментів Data Mining, а також комплексного впровадження Data Mining, OLAP і сховищ даних. Інструменти Data Mining у багатьох випадках розглядаються як складова частина BI-платформ, до складу яких також входять засоби побудови сховищ і вітрин даних, засоби обробки несподіваних запитів (ad-hoc query), засоби звітності (reporting), а також інструменти OLAP. Розробкою в секторі Data Mining всесвітнього ринку програмного забезпечення зайняті як всесвітньо відомі лідери (IBM, Microsoft), так і нові компанії, що розвиваються.

Інструменти Data Mining можуть бути представлені або як самостійне рішення, або як доповнення до основного продукту. Останній варіант реалізується багатьма лідерами ринку програмного забезпечення. Так, вже стало традицією, що розробники універсальних статистичних пакетів, на додаток до традиційних методів статистичного аналізу, включають в пакет певний набір методів Data Mining. Це такі пакети як SPSS (SPSS, Clementine), Statistica (Statsoft), SAS Institute (SAS Enterprise Miner). Деякі розробники OLAP-рішень також пропонують набір методів Data Mining, наприклад, сімейство продуктів Cognos. Є постачальники, які включають Data Mining рішення у функціональність СУБД: це Microsoft (Microsoft SQL Server), Oracle, IBM (IBM Intelligent Miner for Data) [70, 73, 74, 76, 77].

Інструменти Data Mining можна оцінювати по різних критеріях. Оцінка програмних засобів Data Mining з погляду кінцевого користувача визначається шляхом оцінки набору його характеристик. Їх можна поділити на дві групи: бізнес-характеристики і технічні характеристики. Цей розподіл є достатньо умовним, і деякі характеристики можуть потрапляти одночасно в обидві категорії.

Інтуїтивний інтерфейс. Інтерфейс - середовище передачі інформації між програмним середовищем і користувачем, діалогова система, яка дозволяє передати людині всі необхідні дані, отримані на етапі формалізації і обчислення. Він припускає розташування різних елементів, в т.ч. блоків меню,

інформаційних полів, графічних блоків, блоків форм, на екранних формах. Для зручності роботи користувача необхідно, щоб інтерфейс був інтуїтивним.

Інтуїтивний інтерфейс дозволяє користувачеві легко і швидко сприймати елементи інтерфейсу, завдяки чому діалог "програмне середовище-користувач" простіше і доступніше. Поняття інтуїтивного інтерфейсу включає також поняття знайомого навколишнього середовища і наявність виразної нетехнічної термінології (наприклад, для повідомлення користувача про скоєну помилку).

Зручність експорту - імпорту даних. При роботі з інструментом Data Mining користувач часто застосовує різноманітні набори даних, працює з різними джерелами даних. Це можуть бути текстові файли, файли електронних таблиць, файли баз даних. Інструмент Data Mining повинен мати зручний спосіб завантаження (імпорту) даних. Після закінчення роботи користувач також повинен мати зручний спосіб вивантаження (експорту) даних в зручне для нього середовище. Програма повинна підтримувати найбільш поширені формати даних. Додаткова зручність для користувача створюється при нагоді завантаження і вивантаження певної частини (по вибору користувача) полів, що імпортуються або експортуються.

Наочність і різноманітність отримуваної звітності. Ця характеристика припускає отримання звітності в термінах предметної області, а також в якісно спроектованих вихідних формах в тій кількості, яка може надати користувачеві всю необхідну результативну інформацію.

Зручність і простота використання. Істотно полегшує роботу користувача можливість використовувати програми Майстер.

Можливості візуалізації. Наявність графічного представлення інформації істотно полегшує інтерпритованість отриманих результатів.

Наявність значень параметрів, заданих за умовчанням. Для користувачів, що починають, - це достатньо істотна характеристика, оскільки при виконанні багатьох алгоритмів від користувача потрібне завдання або вибір великого числа параметрів. Особливо багато їх в інструментах, що

реалізовують метод нейронних мереж. У нейросимуляторах найчастіше наперед задані значення основних параметрів, інший раз недосвідченим користувачам навіть не рекомендується змінювати ці значення. Якщо ж такі значення відсутні, користувачеві доводиться перепробувати безліч варіантів, перш ніж отримати прийнятний результат.

Кількість методів і алгоритмів. У багатьох інструментах Data Mining реалізоване відразу декілька методів, що дозволяють вирішувати одну або декілька задач. Якщо для вирішення одного завдання (класифікації) передбачена можливість використання декількох методів (дерев рішень і нейронних мереж), користувач дістає можливість порівнювати характеристики моделей, побудованих за допомогою цих методів.

Можливості пошуку, сортування, фільтрації. Така можливість корисна як для вхідних даних, так і для вихідної інформації. Застосовується сортування по різних критеріях (полям), з можливістю накладення умов. За умови фільтрації вхідних даних з'являється можливість побудови моделі Data Mining на одній з вибірок набору даних. Фільтрація вихідної інформації корисна з погляду інтерпретації результатів. Так, наприклад, іноді при побудові дерев рішень результати виходять дуже громіздкими, і тут можуть виявитися корисними функція як фільтрації, так і пошуку і сортування. Додаткова зручність для користувача - колірне підсвічування деяких категорій записів.

Захист, пароль. Дуже часто за допомогою Data Mining аналізується конфіденційна інформація, тому наявність пароля доступу в систему є бажаною характеристикою для інструменту.

Описані характеристики є критеріями функціональності, зручності, безпеки інструменту Data Mining. При виборі інструменту слід керуватися потребами, а також задачами, які необхідно вирішити. Так, наприклад, якщо точно відомо, що фірмі необхідно вирішувати виключно задачі класифікації, то можливість рішення інструментом інших завдань зовсім не є критичною. Проте, слід враховувати, що впровадження Data Mining при серйозному підході

вимагає серйозних фінансових вкладень, тому необхідно враховувати всі можливі задачі, які можуть виникнути в перспективі.

Ринок інструментів Data Mining визначається широтою цієї технології і внаслідок цього - величезним різноманіттям програмного забезпечення. Найбільш популярна група інструментів містить наступні категорії:

- набори інструментів;
- класифікація даних;
- кластеризація і сегментація;
- інструменти статистичного аналізу;
- аналіз текстів (Text Mining), витягання відхилень (Information Retrieval);
- інструменти візуалізації.

Набори інструментів. До цієї категорії відносяться універсальні інструменти, які включають методи класифікації, кластеризації і попередньої підготовки даних. До цієї групи відносяться такі відомі комерційні інструменти як:

- Clementine. Data Mining з використанням Clementine є бізнес-процесом, розробленим для мінімізації часу рішення задач. Clementine підтримує процес Data Mining: доступ до даним, перетворення, моделювання, оцінювання і впровадження. За допомогою Clementine Data Mining виконується з методологією CRISP-DM.
- IBM Intelligent Miner for Data. Інструмент пропонує останні Data Mining-методи, підтримує повний Data Mining процес: від підготовки даних до презентації результатів. Підтримка мов XML і PMML.
- KXEN (Knowledge extraction Engines). Інструмент, що працює на основі теорії Вапника SVM. Вирішує задачі підготовки даних, сегментації, тимчасових рядів і SVM-класифікації.
- Oracle Data Mining (ODM). Інструмент забезпечує GUI, PL/SQL-інтерфейси, Java-інтерфейс. Використовувані методи: байесовська класифікація, алгоритми пошуку асоціативних правил, кластерні методи, SVM та інші.

- Polyanalyst. Набір, що забезпечує вселякий Data Mining, також включає аналіз текстів, ліс рішень, аналіз зв'язків. Підтримує OLE DB for Data Mining і DCOM-технологію.
- SPSS. Один з найбільш популярних інструментів, підтримується безліч методів Data Mining.
- Statistica Data Miner. Інструмент забезпечує вселякий, інтегрований статистичний аналіз даних, має могутні графічні можливості, управління базами даних, а також додатки розробки систем.

Друга група задач представлена інструментами, що реалізовують наступні рішення:

- Magnum Opus є швидким інструментом пошуку асоціативних правил в даних, підтримується операційними системами Windows, Linux і Solaris.
- Nuggets - це набір, що включає пошук асоціативних правил і інші алгоритми.
- Artool, інструмент містить набір алгоритмів для пошуку асоціативних правил в бінарних базах даних (binary databases).
- DM-II System, інструмент включає алгоритм CBA для виконання класифікації на основі асоціативних правил і деяких інших характеристик.

Для розв’язання задач *кластеризації та сегментації* широке застосування отримали:

- ClustanGraphics 3 - ієрархічний кластерний аналіз «зверху вниз», в якому підтримуються могутні графічні можливості.
- Starprobe, заснований на Web кросс-платформенній системі, включає методи кластеризації, нейронні мережі, дерева рішень, візуалізацію і т.д.
- Visipoint - кластеризація методом карт Кохонена (Self-organizing Map clustering) і візуалізація.
- Autoclass C - "навчання без вчителя" за допомогою байесовських мереж від NASA, працює на основі операційних систем Unix і Windows.
- Reckless є набором кластерних алгоритмів, заснованих на концепції к-ближніх сусідів. Інструмент перед проведенням кластеризації виконує пошук і

ідентифікацію шумів і викидів для зменшення їх впливу на результати кластеризації.

Для вирішення задач оцінювання і прогнозування застосовуються:

- Alyuda Forecaster XL - інструмент реалізований у вигляді Excel-надбудови і призначений для вирішення задач прогнозування і оцінювання з використанням нейронних мереж.
- Excelneuralpackage, в якому реалізовано дві базові парадигми нейронних мереж - багатопрошарковий персептрон і мережі Кохонена.

Як ми бачимо, ринок програмного забезпечення Data Mining представлений безліччю інструментів і на ньому йде постійна конкурентна боротьба за споживача. Така конкуренція породжує нові якісні рішення. Все більше число постачальників прагнуть об'єднати в своїх інструментах як можна більше число сучасних методів і технологій. Data Mining-інструменти найчастіше розглядаються як складова частина ринку Business Intelligence, який, не дивлячись на деякий загальний спад в індустрії інформаційних технологій, упевнено і постійно розвивається. В той же час деякі фахівці відзначають відставання існуючого програмного забезпечення від теоретичних розробок у зв'язку з складністю програмної реалізації деяких нових теоретичних розробок методів і алгоритмів Data Mining. В цілому, можна резюмувати, що ринок Business Intelligence, зокрема ринок інструментів Data Mining, настільки широкий і різноманітний, що будь-яка компанія може вибрати для себе інструмент, який підійде їй по функціональності і по можливостях бюджету.

Більш докладніше розглянемо деякі комп’ютерні системи Data Mining, які знайшли значне поширення в практиці діяльності підприємств та організацій.

Програмний продукт SAS Enterprise Miner (розробник SAS Institute Inc.,) є інтегрованою компонентою системи SAS, створеною спеціально для виявлення у величезних масивах даних інформації, яка необхідна для ухвалення рішень. Розроблений для пошуку і аналізу глибоко прихованих

закономірностей в даних SAS, Enterprise Miner включає методи статистичного аналізу, відповідну методологію виконання проектів Data Mining (SEMMA) і графічний інтерфейс користувача. Важливою особливістю SAS Enterprise Miner є його повна інтеграція з програмним продуктом SAS Warehouse Administrator, призначеним для розробки і експлуатації інформаційних сховищ, і іншими компонентами системи SAS. Розробка проектів Data Mining може виконуватися як локально, так і в архітектурі клієнт-сервер.

Пакет підтримує виконання всіх необхідних процедур в рамках єдиного інтегрованого рішення з можливостями колективної роботи і поставляється як розподілений клієнт-серверний додаток, що особливо зручно для здійснення аналізу даних в масштабах крупних організацій. Пакет SAS Enterprise Miner призначений для фахівців з аналізу даних, маркетингових аналітиків, маркетологів, фахівців з аналізу ризиків, фахівців з виявлення шахрайських дій, а також інженерів і учених, відповідальних за ухвалення ключових рішень в бізнесі або дослідницькій діяльності (рис. 1.11).

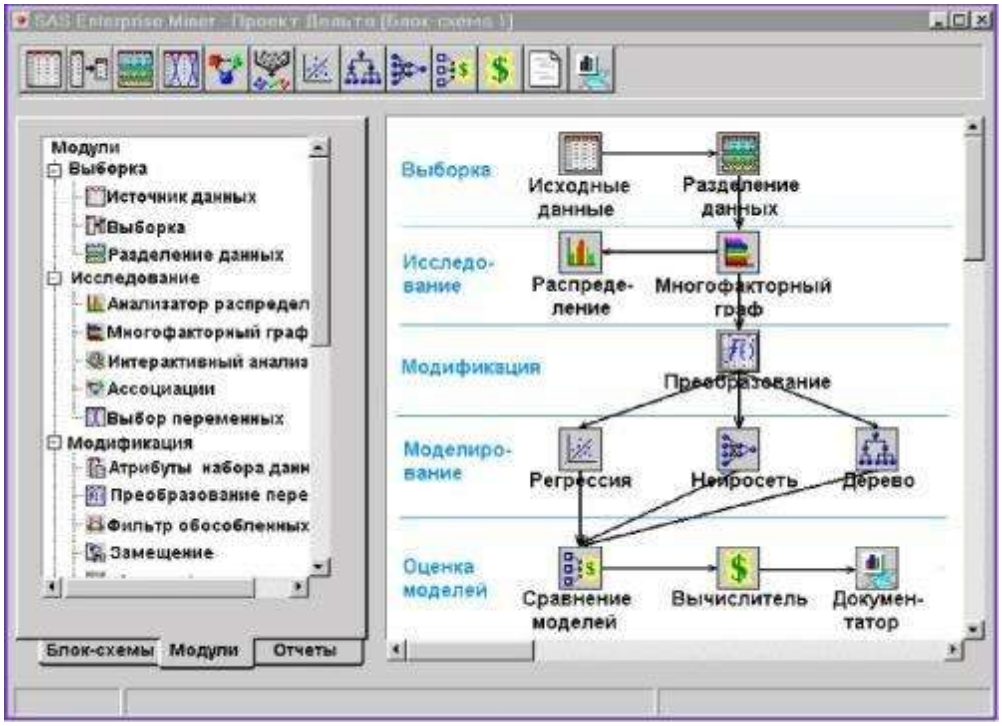

Рис. 1.11. Головне вікно SAS Enterprise Miner.

У пакеті SAS Enterprise Miner реалізований підхід, що заснований на створенні діаграм процесів обробки даних і дозволяє усунути необхідність ручного кодування і прискорити розробку моделей завдяки методиці Data Mining SEMMA. Графічний інтерфейс пакету є інтерфейсом типу "вказати і клацнути". З його допомогою користувачі можуть виконати всі стадії процесу Data Mining від вибору джерел даних, їх дослідження і модифікації до моделювання і оцінки якості моделей з подальшим застосуванням отриманих моделей, як для обробки нових даних, так і для підтримки процесів ухвалення рішень.

Пакет SAS Enterprise Miner пропонує різні інструменти для здійснення підготовки даних, які дають можливість, наприклад, зробити вибірку або розбиття даних, здійснити вставку неотриманих значень, провести кластеризацію, об'єднати джерела даних, усунути зайві змінні, виконати обробку на мові SAS за допомогою спеціалізованого вузла SAS code, здійснити перетворення змінних і фільтрацію недостовірних даних. Пакет оснащений функціями описової статистики, а також розширеними засобами візуалізація, яка дозволяє досліджувати надвеликі об'єми даних, представлених у вигляді багатовимірних графіків, і проводити графічне порівняння результатів моделювання. Також він надає набір інструментів і алгоритмів прогностичного і описового моделювання, що включають дерева рішень, нейронні мережі, нейронні мережі, що самоорганізовуються, методи міркування, засновані на механізмах пошуку в пам'яті (memorybased reasoning), лінійну і логістична регресії, кластеризацію, асоціації, тимчасові ряди і багато що інше. Інтеграція різних моделей і алгоритмів в пакеті Enterprise Miner дозволяє проводити послідовне порівняння моделей, створених на основі різних методів, і залишатися при цьому в рамках єдиного графічного інтерфейсу. Вбудовані засоби оцінки формують єдине середовище для порівняння різних методів моделювання, як з погляду статистики, так і з погляду бізнесу, дозволяючи виявити найбільш відповідні методи для наявних даних. Результатом є якісний

аналіз даних, виконаний з урахуванням специфічних проблем конкретного бізнесу.

Отримані моделі можна публікувати для сумісного використання в рамках підприємства за допомогою репозітарію моделей, що є першою на ринку системою управління моделями. Пакет надає ряд вбудованих оціночних функцій, що дозволяють порівняти результати різних методів моделювання, як в термінах бізнесу, так і з використанням статистичної діагностики. Це дає можливість вимірювати ефективність моделі в термінах її прибутковості. Підсумком робіт по інтелектуальному аналізу даних є розгортання створеної моделі - це завершальна стадія, на якій реалізується економічна віддача від проведених досліджень. Процес застосування моделі до нових даних, відомий як скорінг, часто вимагає ручного написання або перетворення програмного коду. Пакет SAS Enterprise Miner автоматизує процес підбору коефіцієнтів і надає готовий програмний код для скорінга на всіх стадіях створення моделі, підтримує створення різних програмних середовищ для розгортання моделі на мовах SAS, C, Java і PMML.

Система Polyanalyst призначена для автоматичного і напівавтоматичного аналізу числових баз даних і витягання з сирих даних практично корисних знань. Polyanalyst знаходить багатофакторні залежності між змінними в базі даних, автоматично будує і тестує багатомірні нелінійні моделі, що виражають знайдені залежності, виводить класифікаційні правила по навчальних прикладах, знаходить в даних багатомірні кластери, будує алгоритми рішень. Розробник системи Polyanalyst - російська компанія Megaputer Intelligence ("Мегапьютер"). За своєю природою Polyanalyst є клієнт-серверним додатком. Користувач працює з клієнтською програмою Polyanalyst Workplace. Математичні модулі виділені в серверну частину - Polyanalyst Knowledge Server. Така архітектура надає природну можливість для масштабування системи (рис. 1.12.).

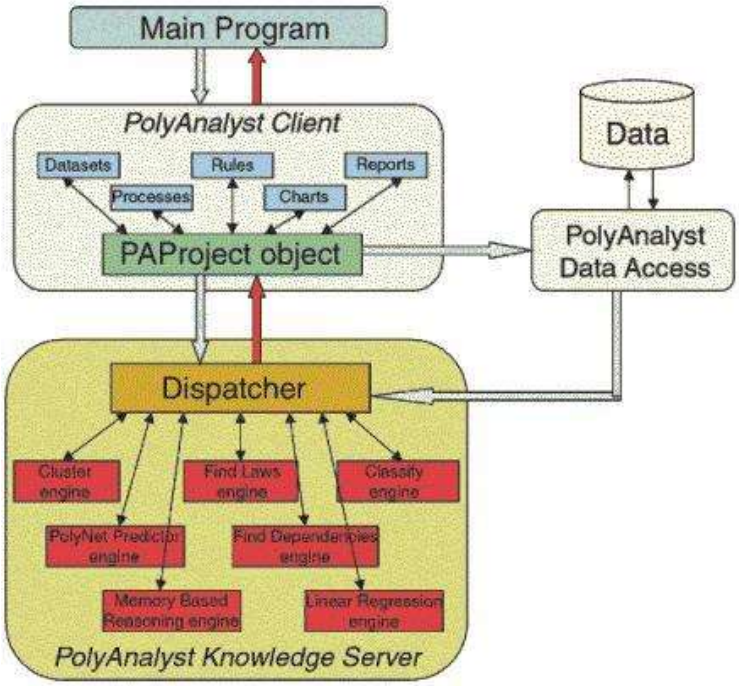

Рис. 1.12. Архітектура системи Polyanalyst.

Математичні модулі (Exploration Engines) і багато інших компонентів Polyanalyst виділені в окремі динамічні бібліотеки і доступні з інших додатків. Це дає можливість інтегрувати математику Polyanalyst в ті програмні ресурси, що існують в інформаційних системах, наприклад, в CRM- або ERP- системах.

Основні риси, призначеного для користувача, інтерфейсу програми: розвинені можливості маніпулювання з даними, графіка для представлення даних і візуалізації результатів, майстри створення об'єктів, сквозний логічний зв'язок між об'єктами, мова символьних правил, інтуїтивне управління через drop-down і pop-up меню, докладна контекстна довідка. Одиницею Data Mining дослідження в Polyanalyst є "проект". Проект об'єднує в собі всі об'єкти дослідження, дерево проекту, графіки, правила, звіти і т.д. Проект зберігається у файлі внутрішнього формату системи. Звіти досліджень представляються у форматі HTML і доступні через Інтернет (рис.1.13.).

Пакет Polyanalyst включає 18 математичних модулів, заснованих на різних алгоритмах Data і Text Mining. Більшість з цих алгоритмів є Know-How компанії Мегапьютер і не мають аналогів в інших системах. До них

відносяться: моделювання, прогнозування, кластеризація, класифікація, текстовий аналіз.

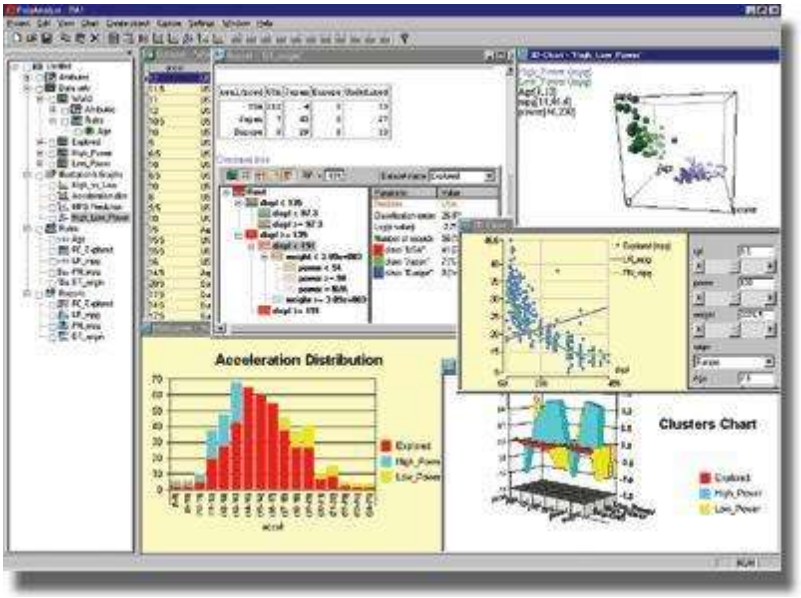

Рис.1.13. Головне меню пакету Polyanalyst.

Програмні продукти Cognos (розробник - компанія Cognos) - це інструменти інтелектуального або ділового аналізу даних (Business Intelligence Tools) або BI-інструменти. За їх допомогою вирішуються слідуючи задачі:

Робота із запитами і звітами. Рішення в області роботи із звітами орієнтовані на різні типи користувачів. Продукти відрізняються вимогами до рівня складності звітів і рівня навиків кінцевих користувачів:

- Decision Stream - засіб для створення вітрин даних (Data marts), оптимізованих на формування запитів і побудову звітів;
- Impromptu - засіб для роботи із запитами, а також із статичними звітами, що настроюються;
- Powerplay - засіб побудови багатомірних звітів;
- Impromptu Web Reports - засоби для роботи із статичними звітами через Web;
- Cognos Query - засіб для створення запитів, навігації і дослідження даних в т.ч. через Web;

- Visualizer - засіб для роботи з могутніми візуальними звітами.

Аналіз даних. Засоби аналізу даних призначені для аналізу критичної інформації і виявлення значущих чинників. Цей процес охоплює повний набір аналітичних задач і завдань по побудові звітів, включаючи роботу із звітами бізнес-уровня, можливість переходу до даним нижнього рівня, створення і проглядання уявлень з метою виявлення пріоритетів. Інтеграція засобів дозволяє зручно переходити від дослідження і аналізу даних за допомогою звітів бізнес-уровня до дослідження і аналізу даних по звітах нижнього рівня:

- Powerplay - засіб багатовимірного (OLAP) аналізу і побудови бізнес-звітів;
- Impromptu - засіб для проглядання звітів з детальною інформацією нижнього рівня (для Windows);
- Impromptu Web Reports - засіб для проглядання звітів з детальною інформацією нижнього рівня (для Web);
- Visualizer - засіб візуального представлення даних.

Візуалізація і виявлення пріоритетів. До розділу візуалізації інформації і виявлення пріоритетів можна віднести цілий спектр продуктів. З їх допомогою користувачеві стає доступна візуалізована інформація, представлена в зручному вигляді для виявлення критичних чинників на великих масивах даних. У цих продуктах за основу береться можливість аналізу ключових чинників, що впливають на дану галузь знань (бізнесу) за допомогою широких можливостей по візуалізації даних. Правильно виявлені пріоритети є основою для ухвалення ефективних рішень:

- Visualizer - засіб для представлення інформації у формі візуальних вистав з використанням візуальних елементів для виявлення пріоритетів;
- PowerPlay - засіб багатовимірного представлення інформації;
- Impromptu - засіб для роботи із звітами, що будуються;
- Cognos Query - засіб Web-користувачів для побудови запитів.

Розвідка даних (data mining). Засоби розвідки і добування даних пропонують цілий ряд можливостей по автоматизованому перегляду даних,

дозволяючи розкривати приховані тенденції, виявляти пріоритетні рішення і дії шляхом відображення тих чинників, які більш за інших впливають на досліджувані показники:

- Scenario - засіб сегментації і класифікації;
- 4Thought - засіб прогнозування (рис. 1.14.);
- Visualazer як засіб візуалізації.

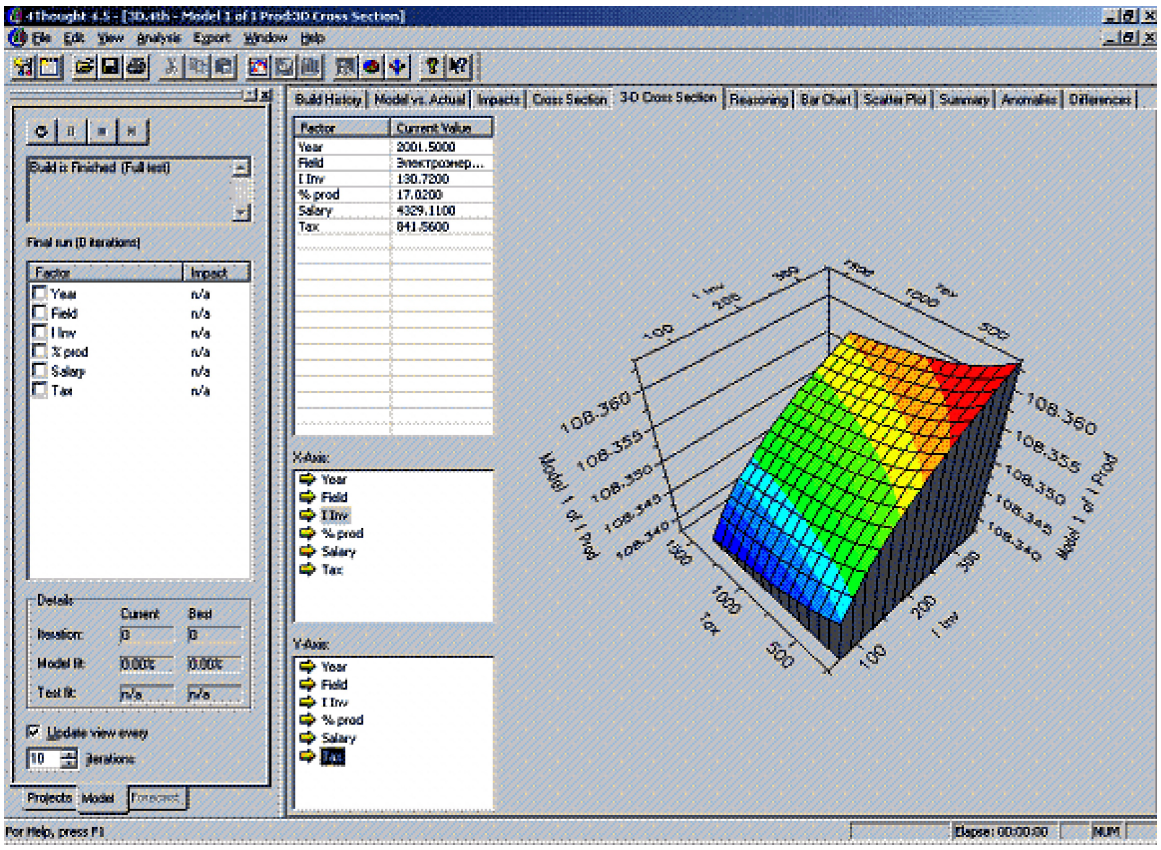

Рис. 1.14.. Інтерфейс програмного комплексу Cognos 4Thought.

Захист інформації. Захист інформації досягається за рахунок використання єдиного для всіх застосувань компонента Access Manager, що дозволяє описувати класи користувачів і управляти ними для всіх типів аналітичних додатків Cognos. На додаток до Access Manager, можуть бути використані також звичайні можливості забезпечення безпеки на рівні бази даних і операційної системи. На практиці можливе одночасне використання всіх трьох рівнів захисту інформації.

Опис метаданих. Як засіб опису метаданих може бути використаний єдиний для всіх Cognos BI продуктів компонент Cognos Architect. Привабливість використання єдиного для всіх засобів модуля полягає в можливості одноманітного представлення бізнес-інформації. Одного разу сформульовані метадані стають доступними в будь-якому аналітичному застосуванні Cognos.

Система STATISTICA Data Miner (розробник - компанія StatSoft) спроектована і реалізована як універсальний і всесторонній засіб аналізу даних - від взаємодії з різними базами даних до створення готових звітів, що реалізовують так званий графічно-орієнтований підхід. Серцем STATISTICA Data Miner є браузер процедур Data Mining, який містить більше 300 основних процедур, спеціально оптимізованих під задачі Data Mining, засоби логічного зв'язку між ними і управління потоками даних, що дозволить конструювати власні аналітичні методи (рис. 1.15.).

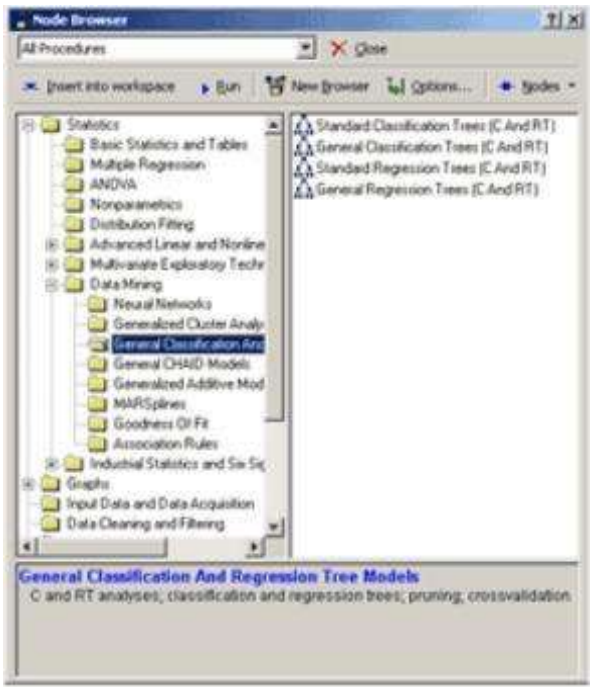

Рис. 1.15. Браузер процедур Data Mining.

Робочий простір STATISTICA Data Miner складається з чотирьох основних частин:

- Data Acquisition - збір даних. У даній частині користувач ідентифікує джерело даних для аналізу, будь то файл даних або запит з бази даних.
- Data Preparation, Cleaning, Transformation - підготовка, перетворення і очищення даних. Тут дані перетворяться, фільтруються, групуються.
- Data Analysis, Modeling, Classification, Forecasting - аналіз даних, моделювання, класифікація, прогнозування. Користувач може за допомогою браузеру або готових моделей задати необхідні види аналізу даних, таких як прогнозування, класифікація, моделювання.
- Reports - результати. У даній частині користувач може проглянути, задати вигляд і побудувати результати аналізу (наприклад, робоча книга, звіт або електронна таблиця).

Засоби аналізу STATISTICA Data Miner можна розділити на п'ять основних класів:

- General Slicer/Dicer and Drill-Down Explorer – розмітка - розподіл і поглиблений аналіз. Набор процедур, що дозволяє розбивати, групувати змінні, обчислювати описові статистики, будувати дослідницькі графіки.
- General Classifier - класифікація. STATISTICA Data Miner включає повний пакет процедур класифікації: узагальнені лінійні моделі, дерева класифікації, регресійні дерева, кластерний аналіз.
- General Modeler/Multivariate Explorer - узагальнені лінійні, нелінійні і регресійні моделі. Даний елемент містить лінійні, нелінійні, узагальнені регресійні моделі і елементи аналізу дерев класифікації.
- General Forecaster - прогнозування. Включає моделі АРПСС, сезонні моделі АРПСС, експоненціальне згладжування, спектральний аналіз Фур'є, сезонна декомпозиція, прогнозування за допомогою нейронних мереж.
- General Neural Networks Explorer - нейромережевий аналіз. У даній частині міститься якнайповніший пакет процедур нейромережевого аналізу.

Приведені вище елементи є комбінацією модулів інших продуктів StatSoft. Окрім них, STATISTICA Data Miner містить набір спеціалізованих процедур Data Mining, які доповнюють лінійку інструментів Data Mining:

- Feature Selection and Variable Filtering (for very large data sets) - спеціальна вибірка і фільтрація даних (для великих об'ємів даних). Даний модуль автоматично вибирає підмножини змінних із заданого файлу даних для подальшого аналізу.
- Association Rules - правила асоціації. Модуль є реалізацією так званого апріорного алгоритму виявлення правил асоціації. Наприклад, результат роботи цього алгоритму міг би бути наступним: клієнт після покупки продукт "A", в 95 випадках з 100 протягом наступних двох тижнів після цього замовляє продукт "B" або "C".
- Interactive Drill-Down Explorer - інтерактивний поглиблений аналіз. Є набором засобів для гнучкого дослідження великих наборів даних. На першому кроці ви задаєте набір змінних для поглибленого аналізу даних, на кожному подальшому кроці вибираєте необхідну підгрупу даних для подальшого аналізу.
- Generalized EM & к-Means Cluster Analysis - узагальнений метод максимуму середнього і кластеризація методом середніх. Даний модуль - це розширення методів кластерного аналізу. Він призначений для обробки великих наборів даних і дозволяє кластеризувати як безперервні, так і категоріальні змінні, забезпечує всі необхідні функціональні можливості для розпізнавання образів.
- Generalized Additive Models (GAM) - узагальнені аддитивні моделі (GAM). Набор методів, розроблених і популяризованных Hastie і Tibshirani.
- General Classification and Regression Trees (GTrees) - узагальнені класифікаційні і регресійні дерева (GTrees). Модуль є повною реалізацією методів, розроблених Breiman, Friedman, Olshen і Stone (1984). Окрім цього, модуль містить різного роду доопрацювання і доповнення, такі як

оптимізації алгоритмів для великих об'ємів даних. Модуль є набором методів узагальненої класифікації і регресійних дерев.

- General CHAID (Chi-square Automatic Interaction Detection) Models - узагальнені CHAID-модели (Хі-квадрат автоматичне виявлення взаємодії). Подібно до попереднього елементу, цей модуль є оптимізацією даної математичної моделі для великих об'ємів даних.
- Boosted Trees - розширювані прості дерева. Останні дослідження аналітичних алгоритмів показують, що для деяких завдань побудови "складних" оцінок, прогнозів і класифікацій використання послідовно збільшуваних простих дерев дає точніші результати, ніж нейронні мережі або складні цілісні дерева. Даний модуль реалізує алгоритм побудови простих збільшуваних (розширюваних) дерев.
- Multivariate Adaptive Regression Splines (Mar Splines) - багатовимірні адаптивні регресійні сплайни (Mar Splines). Даний модуль заснований на реалізації методики запропонованої Friedman (1991).

Нескладно відмітити, що система STATISTICA включає величезний набір різних аналітичних процедур, і це робить її недоступною для звичайних користувачів, які слабо знаються на методах аналізу даних. Компанією StatSoft запропонований варіант роботи для звичайних користувачів, що володіють невеликими досвідом і знаннями в аналізі даних і математичній статистиці. Для цього, окрім загальних методів аналізу, були вбудовані готові закінчені модулі аналізу даних, призначені для вирішення найбільш важливих і популярних завдань: прогнозування, класифікації, створення правил асоціації і т. д.

Oracle Data Mining. У березні 1998 компанія Oracle оголосила про спільну діяльність з 7 партнерами, які є постачальниками інструментів Data Mining. Далі послідувало включення в Oracle8i засобів підтримки алгоритмів Data mining. У червні 1999 року Oracle купує Darwin (Thinking Machines Corp.) і у 2000-2001 році виходять нові версії Darwin, Oracle Data Mining Suite. У червні 2001 року виходить Oracle9i Data Mining.

Oracle Data Mining є опцією або модулем в Oracle Enterprise Edition. Опція Oracle Data Mining (ODM) призначена для аналізу даних методами, що відносяться до технології витягання знань, або Data Mining. ODM підтримує всі етапи технології витягання знань, 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 опції ODM складають процедури, що реалізовують різні алгоритми побудови моделей класифікації, регресії, кластеризації. На етапі підготовки даних забезпечується доступ до будь-яких реляційних баз, текстових файлів, файлів формату SAS. Додаткові засоби перетворення і очищення даних дозволяють змінювати вигляд сценарію, проводити нормалізацію значень, виявляти невизначені або відсутні значення. На основі підготовлених даних спеціальні процедури автоматично будують моделі для подальшого прогнозування, класифікації нових ситуацій, виявлення аналогій. ODM підтримує побудову п'яти різних типів моделей. Графічні засоби надають широкі можливості для аналізу отриманих результатів, верифікації моделей на тестових наборах даних, оцінки точності і стійкості результатів. Уточнені і перевірені моделі можна включати в існуючі додатки шляхом генерації їх описів на C++, Java, а також розробляти нові спеціалізовані застосування за допомогою того, що входить до складу середовища ODM засобу розробки Software Development Kit (SDK).

Аналітична платформа Deductor. Склад і призначення аналітичної платформи Deductor (розробник - компанія BaseGroup Labs). Deductor складається з двох компонентів: аналітичного додатка Deductor Studio і багатомірного сховища даних Deductor Warehouse. Архітектура системи Deductor представлена на рис. 1.16.

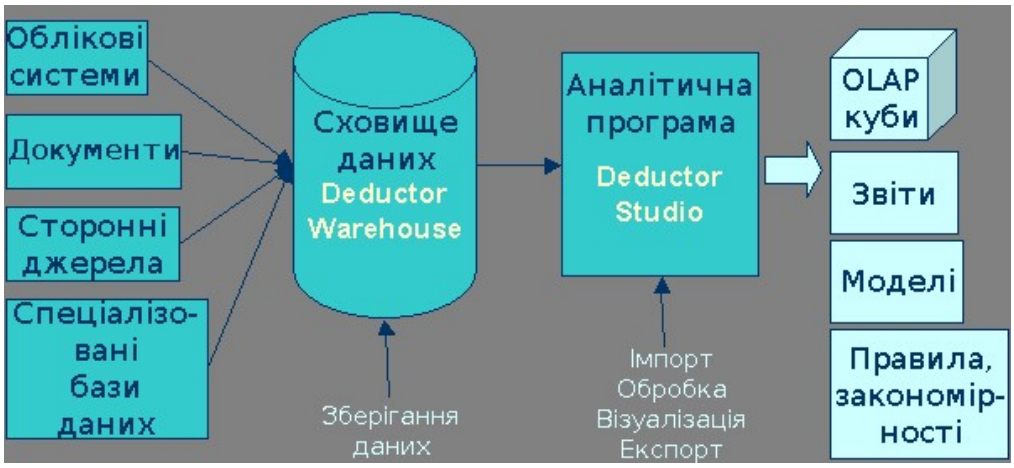

Рис. 1.16. Архітектура системи Deductor.

Deductor Warehouse - багатомірне сховище даних, що акумулює всю необхідну для аналізу наочної області інформацію. Використання єдиного сховища дозволяє забезпечити несуперечність даних, їх централізоване зберігання і автоматично створює всю необхідну підтримку процесу аналізу даних. Deductor Warehouse оптимізований для вирішення саме аналітичних задач, що позитивно позначається на швидкості доступу до даних.

Deductor Studio - це програма, призначена для аналізу інформації з різних джерел даних. Вона реалізує функції імпорту, обробки, візуалізації і експорту даних. Deductor Studio може функціонувати і без сховища даних, отримуючи інформацію з будь-яких інших джерел, але найбільш оптимальним є їх спільне використання.

Розглянемо цей процес детальніше. На початковому етапі в програму завантажуються або імпортуються дані з якого-небудь довільного джерела. Сховище даних Deductor Warehouse є одним з джерел даних. Підтримуються також інші сторонні джерела. Зазвичай в програму завантажуються не всі дані, а якась вибірка, необхідна для подальшого аналізу. Після здобуття вибірки можна отримати детальну статистику по ній, проглянути, як виглядають дані на діаграмах і гістограмах. Такий розвідувальний аналіз дає можливість приймати рішення про необхідність передобробки даних. Наприклад, якщо статистика показує, що у вибірці є порожні значення (пропуски даних), можна застосувати

фільтрацію для їх усунення. Передоброблені дані далі піддаються трансформації. Наприклад, нечислові дані перетворяться в числові, що необхідне для деяких алгоритмів. Безперервні дані можуть бути розбиті на інтервали, тобто виробляється їх дискретизація. До трансформованих даних застосовуються методи глибшого аналізу. На цьому етапі виявляються приховані залежності і закономірності в даних, на підставі яких будуються різні моделі. Модель є шаблоном, який містить формалізовані знання. Останній етап - інтерпретація – призначений для того, щоб з формалізованих знань отримати знання на мові наочної області.

Вся робота по аналізу даних в Deductor Studio базується на виконанні наступних дій: імпорт даних, обробка даних, візуалізація, експорт даних. Відправною точкою для аналізу завжди є процедура імпорту даних. Отриманий набір даних може бути оброблений будь-яким з доступних способів. Результатом обробки також є набір даних, який, у свою чергу, знову може бути оброблений. Імпортований набір даних, а також дані, отримані на кожному етапі обробки, можуть бути експортовані для подальшого використання в інших, наприклад, в облікових системах. Результати кожної дії можна відобразити різними способами: OLAP-кубы (крос-таблиця, крос-діаграма), плоска таблиця, діаграма, гістограма, статистика, аналіз за принципом "що-якщо", граф нейромережі, дерево - ієрархічна система правил, інше.

Послідовність дій, які необхідно провести для аналізу даних, називається сценарієм. Сценарій можна автоматично виконувати на будь-яких даних. Типовий сценарій змальований на рис. 1.17.

Deductor Warehouse - багатовимірне сховище даних, що акумулює всю необхідну для аналізу наочної області інформацію. Вся інформація в сховищі міститься в структурах типа "зірка", де в центрі розташовані таблиці фактів, а "променями" є виміри. Така архітектура сховища найбільш адекватна завданням аналізу даних. Кожна "зірка" називається процесом і описує певну дію. У Deductor Warehouse може одночасно зберігатися безліч процесів, що мають загальні виміри.

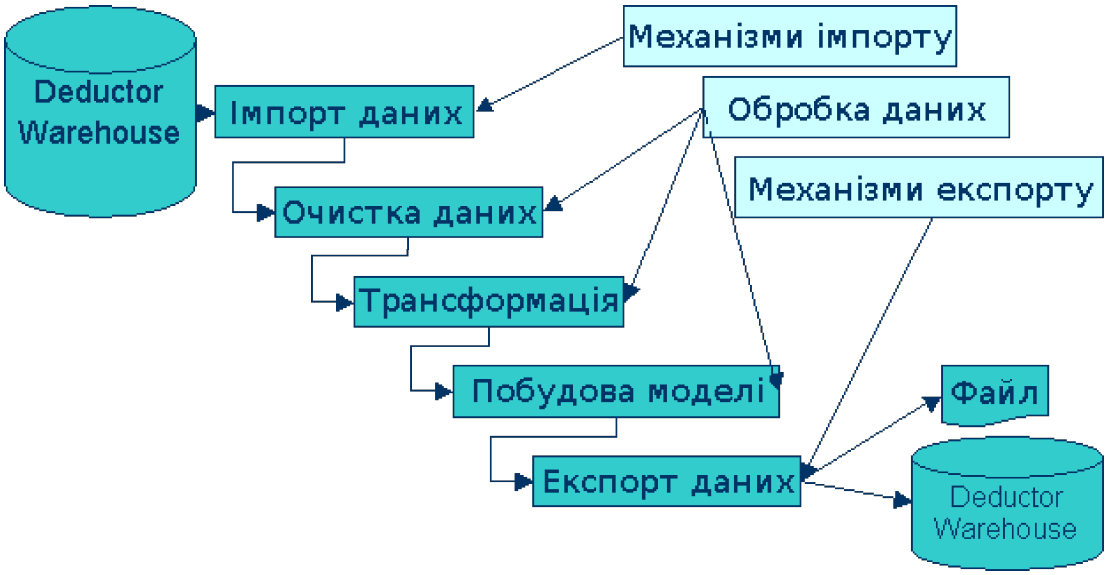

Рис. 1.17. Типовий сценарій Deductor Studio.

Що є сховищем Deductor Warehouse ? Фізично - це реляційна база даних, яка містить таблиці для зберігання інформації і таблиці зв'язків, що забезпечують цілісне зберігання відомостей. Поверх реляційної бази даних реалізований спеціальний прошарок, який перетворить реляційну систему до багатомірної. Багатомірна ідеологія використовується тому, що воно набагато краще реляційної відповідає ідеології аналізу даних. Завдяки цьому прошарку користувач оперує багатомірними поняттями, такими як "вимір" або "факт", а система автоматично виробляє всі необхідні маніпуляції, необхідні для роботи з реляційною СУБД.

Окрім консолідації даних, робота із створення закінченого аналітичного рішення містить декілька етапів.

Очищення даних. На цьому етапі проводиться редагування аномалій, заповнення пропусків, згладжування, очищення від шумів, виявлення дублікатів і протиріч.

Трансформація даних. Виробляється заміна порожніх значень, квантування, таблична заміна значень, перетворення до ковзаючого вікна, зміна формату набору даних.

Data Mining. Будуються моделі з використанням нейронних мереж, дерев рішень, самоорганізующихся карт , асоціативних правил.

Програмний комплекс KXEN. Програмне забезпечення KXEN є розробкою однойменної французько-американської компанії, яка працює на ринку з 1998 року. Абревіатура KXEN означає "Knowledge eXtraction Engines" - "движки" для витягання знань. Відразу слід сказати, що розробка KXEN має особливий підхід до аналізу даних. У KXEN немає дерев рішень, нейронних мереж і іншої популярної техніки. KXEN - це інструмент для моделювання, який дозволяє говорити про еволюцію Data Mining і реінжинірінге аналітичного процесу в організації в цілому. У основі цих тверджень лежать досягнення сучасної математики і принципово інший підхід до вивчення явищ в бізнесі. Слід зазначити, що все те, що відбувається усередині KXEN сильно відрізняється (принаймні, по своїй філософії) від того, що ми звиклися рахувати традиційним Data Mining.

Бізнес-моделювання KXEN - це аналіз діяльності компанії і її оточення шляхом побудови математичних моделей. Він використовується в тих випадках, коли необхідно зрозуміти взаємозв'язок між різними подіями і виявити ключові рушійні сили і закономірності в поведінці об'єктів, що цікавлять нас, або процесів. KXEN охоплює чотирьох основних типів аналітичних задач:

- Задачі регрессії - классифікації (в т.ч. визначення вкладів змінних).
- Задачі сегментації – кластеризації.
- Аналіз часових рядів.
- Пошук асоціативних правил (аналіз споживчої корзини).

Побудована модель в результаті стає механізмом аналізу, тобто частиною бізнес-процеса організації. Головна ідея тут - на основі побудованих моделей створити систему "крізного" аналізу процесів, що відбуваються. Це дозволяє автоматично виробляти їх оцінку і будувати прогнози в режимі реального часу (у міру того, як ті або інші операції фіксуються обліковими системами організації).

У 1990 році були отримані важливі результати в математиці і машинному навчанні. Ініціатором досліджень в цій області став Володимир

Вапник, що опублікував свою Статистичну Теорію Навчання. Він був першим, хто відчинив двері до нових доріг декомпозиції помилки, що отримується в процесі вживання методів машинного навчання. Він виявив і описав структуру цієї помилки і на основі зроблених виводів відшукав спосіб структурувати методи моделювання. Математичний апарат, який закладений в KXEN, в ході аналізу будує декілька конкуруючих моделей. Але цей процес здійснюється не випадковим чином (перебором різних методів моделювання), а шляхом вивчення різних наборів моделей з опорою на Теорію мінімізації структурних ризиків В. Вапника (Structured Risk Minimization). Творці KXEN розробили механізм порівняння моделей, з тим аби добитися найкращого співвідношення між їх точністю і надійністю, і вже цю оптимальну модель представити як результат аналізу користувачеві.

KXEN Analytic Framework за своєю суттю не є монолітним застосуванням, а виконує роль компонента, який вбудовується в існуюче програмне середовище. Цей "движок" може бути підключений до DBMS-систем (наприклад, Oracle або MS SQL-Server) через протоколи ODBS.

KXEN Analytic Framework є набором модулів для проведення описового і передбачаючого аналізу. Зважаючи на специфіку задач конкретної організації, конструюється оптимальний варіант програмного забезпечення KXEN. Завдяки відкритим програмним інтерфейсам, KXEN легко вбудовується в існуючі системи організації. Тому форма представлення результатів аналізу, з якою працюватимуть співробітники на місцях, може визначатися побажаннями замовника і особливостями його бізнес-процеса. На рис. 1.18 представлена структура KXEN Analytic Framework Version 3.0. Розглянемо ключові компоненты системи KXEN.

Компонент Агрегації Подій (KXEN Event Log - KEL) призначений для агрегації подій, подій за певні періоди часу. Вживання KEL дозволяє з'єднати транзакційні дані з демографічними даними про клієнта.

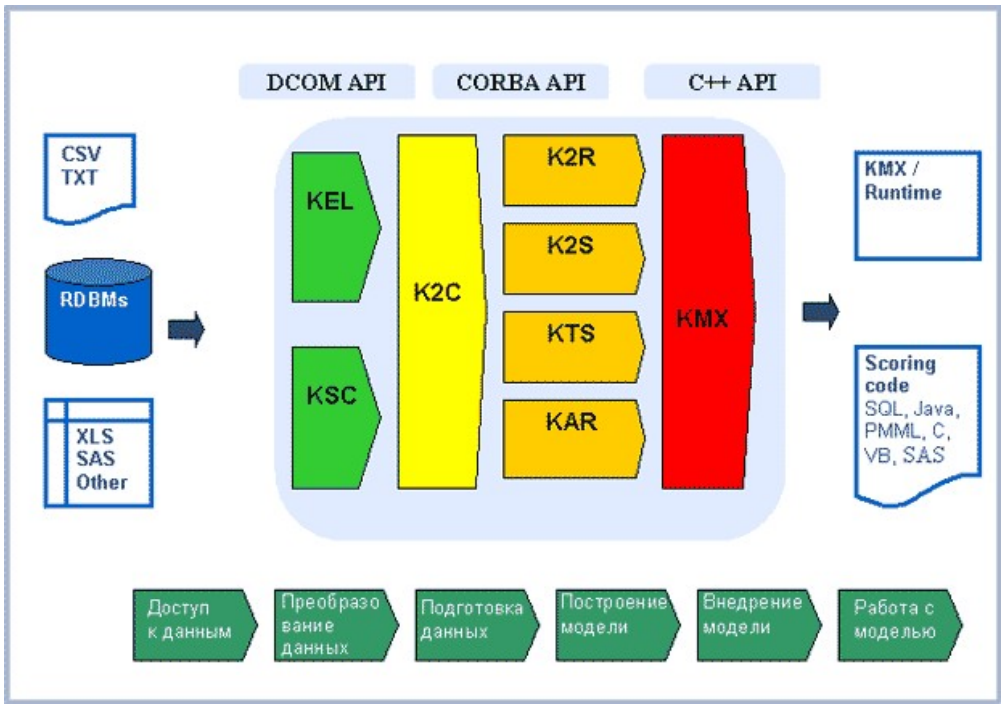

Рис. 1.18. Структура KXEN Analytic Framework Version 3.0.

Компонент використовується у випадках, коли "сирі" дані містять одночасно статичну інформацію (наприклад, вік, пів або професія індивіда) і динамічні змінні (наприклад, шаблони покупок або транзакції по кредитній карті). Дані автоматично агрегуються усередині визначених користувачем інтервалів без програмування на SQL або внесеннях змін в схему бази даних. Компонент KEL комбінує і стискує ці дані для того, щоб зробити їх доступними для інших компонентів KXEN. Перевагою використання даного компонента є можливість інтегрувати додаткові джерела інформації "на льоту" для того, щоб поліпшити якість моделі.

Компонент Кодування Послідовностей (KXEN Sequence Coder - KSC) дозволяє агрегувати події в серії транзакцій. Наприклад, потік "кліків" клієнта, що фіксується на Web-сайті, може трансформуватися в ряди даних для кожної сесії. Кожна колонка відображає конкретний перехід з однієї сторінки на іншу. Як і у випадку з KEL, нові колонки даних можуть додаватися до існуючих даних про клієнтів і доступні для обробки іншими компонентами KXEN. Перевагою використання даного компонента є можливість застосовувати

незадіяні раніше джерела інформації для того, щоб поліпшити якість прогнозуючих моделей.

Компонент Погодженого Кодування (KXEN Consistent Coder - K2C) дозволяє автоматично підготувати дані і трансформувати їх у формат, відповідний для використання аналітичними додатками KXEN. Використання K2C дозволяє трансформувати номінальні і порядкові змінні, автоматично заповнювати відсутні значення і виявляти викиди. Перевагою використання даного компонента є можливість автоматизації підготовки даних, яка дозволяє звільнити час безпосередньо досліджень і моделювання.

Компонент Робастной Регресії (KXEN Robust Regression - K2R) використовує відповідний регресійний алгоритм для того, щоб побудувати моделі, що описують існуючі залежності, і згенерувати прогнозуючі моделі. Ці моделі можуть потім застосовуватися для скоринга, регресії і класифікації. На відміну від традиційних регресійних алгоритмів, використання K2R дозволяє безпечно справлятися з великою кількістю змінних (більше 10 000). Модуль K2R будує індикатори і графіки, які дозволяють легко переконатися в якості і надійності побудованої моделі. Перевагою використання даного компонента є автоматизація процесу інтелектуального аналізу даних. Моделі дозволяють деталізувати індивідуальні вклади змінних.

Компонент Інтелектуальної Сегментації (KXEN Smart Segmenter - K2S) дозволяє виявити природні групи (кластери) в наборі даних. Модуль оптимізований для того, щоб знаходити кластери, які відносяться до конкретного поставленого завдання. Він описує властивості кожної групи і вказує на її відмінності від всієї вибірки. Як і у випадку з іншими модулями, цей модуль також будує індикатори якості і надійності моделі. Перевагою використання даного компонента є автоматичне виявлення груп, значимих для того конкретного завдання, яке необхідно вирішити.

Машина Опорних Векторов KXEN (Support Vector Machine - KSVM) дозволяє виробляти бінарну класифікацію. Використання компонента потрібно для вирішення задач, заснованих на наборах даних з невеликою кількістю

спостережень і великою кількістю змінних. Це робить модуль ідеальним для вирішення задач в областях з дуже великою кількістю розмірності, таких як медицина і біологія. Перевагою використання даного компонента є можливість вирішення завдань, які раніше вимагали написання спеціальних програм, за допомогою промислового програмного забезпечення.

Компонент Аналізу Часових Рядів (KXEN Time Series - KTS) дозволяє прогнозувати значимі шаблони і тренди у часових рядах. Використовує хронологічні дані, що накопичилися, для того, щоб спрогнозувати результати наступних періодів. Модуль KTS виявляє тренди, періодичність і сезонність для того, щоб отримати точні і достовірні прогнози. З'являється можливість підстроїтися під шаблони бізнесу, що повторюються, і передбачати скорочення постачань до того як вони стануться.

Компонент Експорту Моделей (KXEN Model Export - KMX) дозволяє створювати коди різного типа: SQL, C, VB, SAS, PMML і багато інших для вбудовування в існуючі застосування і бізнес-процеси. Побудована модель у вигляді кода може бути передана на іншу машину для подальшого аналізу даних в пакетному або інтерактивному режимі. Використання даного модуля дає можливість виробляти аналіз знов поступаючої інформації за допомогою моделі автономно, поза самою системою моделювання. Це істотно прискорює впровадження моделей у виробничий процес і не вимагає створення спеціальних умов і програм для аналізу всієї бази даних за допомогою моделі. Користувач також може перекласти модель тією мовою, яка підтримує його комп'ютер.

1.6. Новітні напрямки застосування Data Mining.

Як відмічалось раніше, область використання Data Mining нічим не обмежена - вона скрізь, де є які-небудь дані. Сучасна практика інтелектуального аналізу даних виділяє два основні напрями вживання систем

Data Mining: як масового продукту і як інструменту для проведення унікальних досліджень [19, 57, 70].

Слід зазначити, що на сьогоднішній день найбільшого поширення технологія Data Mining набула при вирішенні бізнес-задач. Можливо, причина в тому, що саме в цьому напрямі віддача від використання інструментів Data Mining може складати, за деякими джерелами, до 1000% і витрати на її впровадження можуть досить швидко окупитися. Зараз технологія Data Mining використовується практично у всіх сферах діяльності людини, де накопичені ретроспективні дані.

Серед новітних напрямків застосування технологій інтелектуального аналізу даних слід виділити Web-задачі, практика вирішення яких викликає поширений інтерес і набуває популярності.

Web Mining. Web Mining можна перевести як "видобуток даних в Web". Web Intelligence готовий "відкрити нову главу" в стрімкому розвитку електронного бізнесу. Здатність визначати інтереси і переваги кожного відвідувача, спостерігаючи за його поведінкою, є серйозною і критичною перевагою конкурентної боротьби на ринку електронної комерції. Системи Web Mining можуть відповісти на багато питань, наприклад, хто з відвідувачів є потенційним клієнтом Web-магазина, яка група клієнтів Web-магазина приносить найбільший дохід, які інтереси певного відвідувача або групи відвідувачів [3, 72].

Технологія Web Mining охоплює методи, які здатні на основі даних сайту виявити нові, раніше невідомі знання і які надалі можна буде використовувати на практиці. Іншими словами, технологія Web Mining застосовує технологію Data Mining для аналізу неструктурованої, неоднорідної, розподіленої і значної за об'ємом інформації, що міститься на Web-узлах. Згідно таксономії Web Mining, тут можна виділити два основні напрями: Web Content Mining і Web Usage Mining.

Web Content Mining має на увазі автоматичний пошук і витягання якісної інформації зі всіляких джерел Інтернету, переобтяжених

"інформаційним шумом". Тут також йде мова про різні засоби кластеризації і анотування документів. У цьому напрямі, у свою чергу, виділяють два підходи: підхід, заснований на агентах, і підхід, заснований на базах даних.

Підхід, заснований на агентах (Agent Based Approach), включає такі системи:

- інтелектуальні пошукові агенти (Intelligent Search Agents);
- фільтрація інформації - класифікація;
- персоніфіковані агенти мережі.

Приклади систем інтелектуальних агентів пошуку: Harvest (Brown і ін., 1994), FAQ-Finder (Hammond і ін., 1995), Information Manifold (Kirk і ін., 1995), OCCAM (Kwok and Weld, 1996) and ParaSite (Spertus, 1997), ILA (Information Learning Agent) (Perkowitz and Etzioni, 1995), ShopBot (Doorenbos і ін., 1996).

Підхід, заснований на базах даних (Database Approach), включає системи:

- багаторівневі бази даних;
- системи web-запросов (Web Query Systems). Наприклад, W3QL (Konopnicki і Shmueli, 1995), WebLog (Lakshmanan і ін., 1996), Lorel (Quass і ін., 1995), UNQL (Buneman і ін., 1995 and 1996), TSIMMIS (Chawathe і ін.., 1994).

Другий напрям Web Usage Mining має на увазі виявлення закономірностей в діях користувача Web-узла або їх групи. Аналізується наступна інформація: які сторінки переглядав користувач, яка послідовність перегляду сторінок. Також аналізується, які групи користувачів можна виділити серед загального їх числа на основі історії перегляду Web-узла.

Web Usage Mining включає наступні складові:

- попередня обробка;
- операційна ідентифікація;
- інструменти виявлення шаблонів;
- інструменти аналізу шаблонів.

При використанні Web Mining перед розробниками виникає два типи завдань. Перша стосується збору даних, друга - використання методів персоніфікації. В

результаті збору деякого об'єму персоніфікованих ретроспективних даних про конкретного клієнта, система нагромаджує певні знання про нього і може рекомендувати йому, наприклад, певні набори товарів або послуг. На основі інформації про всіх відвідувачів сайту Web-система може виявити певні групи відвідувачів і також рекомендувати їм товари або ж пропонувати товари в розсилках.

Завдання Web Mining можна підрозділити на такі категорії:

- попередня обробка даних для Web Mining;
- виявлення шаблонів і відкриття знань з використанням асоціативних правил, тимчасових послідовностей, класифікації і кластеризації;
- аналіз отриманого знання.

Text Mining. Text Mining охоплює нові методи для виконання семантичного аналізу текстів, інформаційного пошуку і управління. Синонімом поняття Text Mining є KDT (Knowledge Discovering in Text - пошук або виявлення знань в тексті). На відміну від технології Data Mining, яка передбачає аналіз впорядкованої в деякі структури інформації, технологія Text Mining аналізує великі і надвеликі масиви неструктурованої інформації. Програми, що реалізовують це завдання, повинні деяким чином оперувати природною людською мовою і при цьому розуміти семантику аналізованого тексту. Один з методів, на якому засновані деякі Text Mining системи, - пошук так званого підрядка в рядку [3, 45].

Call Mining. За словами аналітиків "видобуток дзвінків" може стати популярним інструментом корпоративних інформаційних систем. Технологія Call Mining об'єднує в собі розпізнавання мови, її аналіз і Data Mining. Її мета - спрощення пошуку в аудіо-архівах, що містять записи переговорів між операторами і клієнтами. За допомогою цієї технології оператори можуть виявляти недоліки в системі обслуговування клієнтів, знаходити можливості збільшення продажів, а також виявляти тенденції в зверненнях клієнтів. Серед розробників нової технології Call Mining ("видобуток" і аналіз дзвінків) - компанії CallMiner, Nexidia, ScanSoft, Witness Systems [9, 58].

У технології Call Mining розроблено два підходи - на основі перетворення мови в текст і на базі фонетичного аналізу. Прикладом реалізації першого підходу, заснованого на перетворенні мови, є система CallMiner. В процесі Call Mining спочатку використовується система перетворення мови, потім слідує її аналіз, в ході якого залежно від змісту розмов формується статистика телефонних викликів. Отримана інформація зберігається в базі даних, в якій можливий пошук, витягання і обробка. Приклад реалізації другого підходу - фонетичного аналізу - продукція компанії Nexidia. При цьому підході мова розбивається на фонеми, що є звуками або їх поєднаннями. Такі елементи утворюють розпізнавані фрагменти. При пошуку певних слів і їх поєднань система ідентифікує їх з фонемами.

Аналітики відзначають, що за останні роки інтерес до систем на основі Call Mining значно зріс. Це пояснюється тим фактом, що менеджери вищої ланки компаній, що працюють в різних сферах, в т.ч. в області фінансів, мобільного зв'язку, авіабізнесу, не хочуть витрачати багато часу на прослухування дзвінків з метою узагальнення інформації або ж виявлення яких-небудь фактів порушень. По словах Деніела Хонг, аналітика компанії Datamonitor: "Використання цих технологій підвищує оперативність і знижує вартість обробки інформації". Типова інсталяція продукції від розробника Nexidia обходиться в суму від 100 до 300 тис. дол. Вартість впровадження системи CallMiner по перетворенню мови і набору аналітичних застосувань складає близько 450 тис. дол.

На думку Шоллера, додатки Audio Mining і Video Mining знайдуть з часом набагато ширше вживання, наприклад, при індексації учбових відеофільмів і презентацій в медіабібліотеках компаній. Проте технології Audio Mining і Video Mining знаходяться зараз на рівні становлення, а практичне їх вживання - на самій початковій стадії.

На завершення розділу хотілося б трохи зупинитися на можливих варіантах впровадження систем інтелектуального аналізу даних. Різні варіанти впровадження Data Mining мають свої сильні і слабкі сторони. Так, перевагами

готового програмного забезпечення є готові алгоритми, технічна підтримка виробника, повна конфіденційність інформації, також не потрібно дописувати програмний код, існує можливість придбання різних модулів і надбудов до використовуваного пакету, спілкування з іншими користувачами пакету і інше. Проте, таке рішення має і слабкі сторони. Залежно від інструменту, це може бути достатньо висока вартість ліцензій на програмне забезпечення, неможливість додавати свої функції, складність підготовки даних, практична відсутність в інтерфейсі термінів наочної області та інше. Таке рішення вимагає наявність висококваліфікованих кадрів, які зможуть якісно підготувати дані до аналізу, знають, які алгоритми слід застосовувати для вирішення яких завдань, зуміють проінтерпретувати отримані результати в термінах вирішуваних бізнес-задач. Далеко не кожна компанія може тримати штат таких фахівців, а часто їх утримання навіть неефективно.

Представимо ситуацію, коли менеджер стикається "наодинці" з одним з продуктів, в якому реалізовані методи технології Data Mining (від найпростіших, таких, що включають 1-2 алгоритми, до повнофункціональних програмних комплексів, що пропонують десятки різних алгоритмів). Перед ним стоїть задача - виявити найбільш перспективних потенційних клієнтів, а він бачить перед собою всього лише набір математичних алгоритмів. Це і є "зворотна сторона" використання готових інструментів. Таким чином, покупці готового інструменту повинна передувати серйозна підготовка до впровадження Data Mining.

Розглянемо інший варіант впровадження, який припускає використання Data Mining консалтингу або так званої адаптації програмного забезпечення під конкретне завдання. За даними консалтингової компанії Meta Group, в світі не менше 85% ринку Data Mining займають саме послуги, тобто консультації по ефективному впровадженню цієї технології для вирішення актуальних бізнес-задач. В Інтернеті можна знайти перелік більше ста відомих компаній, що займаються консалтингом у сфері Data Mining. Наприклад, одна з всесвітньо відомих консалтингових компаній у сфері Data Mining - компанія Two Crows.

Вона спеціалізується на публікації звітності Data Mining, проводить освітні семінари, консультує користувачів і розробників Data Mining у всьому світі. Одна з відомих методологій Data Mining розроблена компанією Two Crows.

До консалтингових Data Mining-компаній відносять і деяких виробників готового програмного забезпечення: IBM Global Business Intelligence Solutions, SAS Institute, SPSS, Statsoft та інші. Деякі консалтингові компанії надають свої послуги на певних територіях. Це, наприклад, компанія Arvato Business Intelligence, яка забезпечує Data Mining консультування і моделювання у Франції, Німеччині, Іспанії і деяких інших європейських країнах. Деякі консалтингові компанії спеціалізуються на наданні послуг в певних наочних областях. Наприклад, компанія Blue Hawk LLC, здійснює Data Mining і надає консультаційні послуги в сферах Direct Marketing і CRM. Деякі компанії надають послуги з використанням певних методів Data Mining. Компанія Bayesia, надає консультування і "настройку рішення" під клієнта на основі байєсовськой класифікації. Компанія Visual Analytics забезпечує послуги з бізнес - консультування для знаходження шаблонів з використанням візуального Data Mining.

Розглянемо переваги, які має цей варіант впровадження Data Mining в порівнянні з готовими програмними продуктами і їх самостійним використанням.

Висококваліфіковані фахівці. Для ефективного застосування технології Data Mining потрібні кваліфіковані фахівці, які зуміють якісно провести весь цикл аналізу. Поки що таких грамотних фахівців на просторах СНД та України дуже небагато, і тому вони досить дорогі. Навчання ж власних, по-перше, достатньо ризиковано (його із задоволенням переманить конкурент), по-друге, потрібні чималі витрати бо стоїть це дорого. Клієнти, скориставшись послугами консалтингової компанії, дістають доступ до висококласних професіоналів компанії, економлячи при цьому значні кошти на пошуку або навчанні власних фахівців.

Адаптованість. Готові продукти призначені для вирішення хоч і широкого, але все таки стандартного і обмеженого круга задач, тому адаптація продукту до умов конкретного бізнесу лягає на плечі співробітників компанії. Тут перед замовником знову встане згадана проблема кваліфікованих фахівців. Консалтингова компанія надає послуги, повністю адаптовані під бізнес замовника і його задач.

Гнучкість інструменту. Його можливість швидко підстроїти програмне забезпечення під потреби бізнесу:

- можливість вибору найбільш зручних понять, в термінах яких повинні бути сформульовані знання або терміни наочної області. Так, аналізуючи конкурентів, що діють на ринку, їх можна поділити на "сильних" і "слабких" або ж на "агресивних", "спокійних" і "пасивних" - залежно від того, що цікавить аналітика в певний момент. Відповідно, знання будуть сформульовані у вибраних замовником термінах, і у результаті він отримує рішення саме в тих термінах, які йому цікаві і зрозумілі;
- отримання осмислених і зрозумілих замовникові знань в природній формі. Використання адаптованого під конкретний бізнес програмного забезпечення позбавить користувача від необхідності вивчення формул або залежностей в математичній формі, а надасть знання в найбільш інтуїтивному вигляді.

Розглянемо приклад практичного використання розглянутого підходу, який був застосован компанією Snow Cactus для оцінки кредитоспроможності позичальника банку. Реалізація задачі "Видавати кредит?" відбувалася в системі dm-Score, адаптованої під конкретну бізнес-задачу програмного забезпечення кредитного скорінга.

Система кредитового скорінга dm-Score (dm - від Data Mining) впроваджена в банку для аналізу кредитних історій і виявлення прихованих впливів параметрів позичальників на їх кредитоспроможність. Така система повинна вписуватися в інформаційний простір банку, тобто безпосередньо взаємодіяти з базами даних, де зберігається інформація про позичальників і

кредити, з автоматизованою банківською системою (АБС), іншим програмним забезпеченням, і працювати з ними як єдине ціле.

В процесі впровадження фахівці знайомляться з використовуваємою в банку АБС системою автоматизації рітейла, базами даних і т.д., погоджують з фахівцями вимоги до системи скорінга - як функціональні, так і не функціональні, а також вивчають, які дані накопичені банком і які задачі вони дозволяють вирішувати, здійснюють адаптацію системи відповідно до них. Однією з важливих переваг впровадження системи dm-Score є те, що в процесі впровадження враховуються всі індивідуальні вимоги і побажання до неї з боку банку. Важливо також відзначити, що в цьому випадку відбувається інтеграція системи dm-Score в інформаційний простір банку, а не навпаки, тобто впровадження не зажадає яких-небудь змін в існуючих бізнес-процесах. Таким чином, в результаті впровадження банк отримує систему скорінга, яка враховує всі специфічні особливості і потреби клієнта.

Описувана система dm-Score дозволяє вирішувати наступні задачі:

- оцінка кредитоспроможності позичальника (скорінг позичальника);
- ухвалення рішення про видачу кредиту або відмові в ньому. При цьому система може пояснити фахівцеві банку, чому було ухвалено саме таке рішення;
- визначення максимального розміру кредиту (ліміту кредиту по кредитній карті) на основі скорінга позичальника;
- винесення професійної думки про кредитний ризик по позиках;
- вироблення індивідуальних умов кредитування для кожного позичальника з урахуванням ризику для банку;
- прогнозування поведінки позичальника, тобто наявність і частоту прострочень конкретного позичальника, середній розмір використовуваного кредиту по кредитній карті і т.д.;
- оптимізація анкети позичальника (виключення не значущих питань без погіршення якості анкети);

- перевірка анкети конкретного позичальника на повноту і внутрішню несуперечність;
- рішення інших задач, специфічних для конкретного банку.

Ця система робить свої висновки на основі даних, вже накопичених банком в процесі роботи на ринку роздрібного кредитування. При цьому в процесі впровадження система настроюється саме на той набір даних, на який орієнтований конкретний банк. Іншими словами, система dm-Score готова працювати з тими даними, які є в наявності, і не вимагає фіксації на якій-небудь конкретній жорстко заданій анкеті. В процесі аналізу даних про позичальників і кредити застосовуються різні математичні методи, які виявляють в них чинники і їх комбінації, що впливають на кредитоспроможність позичальників, і силу їх впливу. Виявлені залежності складають основу для ухвалення рішень у відповідному блоці. Блок аналізу повинен періодично використовуватися для аналізу нових даних банку (приходять нові позичальники, поточні проводять виплати), для забезпечення актуальності системи і адекватності ухвалюваних нею рішень. Блок ухвалення рішень використовується безпосередньо для отримання висновку системи dm-Score про кредитоспроможність позичальника, про можливість видачі йому кредиту, про максимально допустимий розмір кредиту і т.д. З цим блоком працює співробітник банку, який або вводить в нього анкету нового позичальника, або отримує її з торгової точки, де банк здійснює програму споживчого кредитування. Завдяки тісній інтеграції системи з інформаційним простором банку, результати роботи цього блоку передаються безпосередньо АБС і системі автоматизації рітейла, які вже формують всі необхідні документи, ведуть історію кредиту і т.д. Таким чином, і система dm-Score, і всі банківські системи працюють як одне ціле, підвищуючи продуктивність праці співробітників банку.