ТЕМА 5. ІНФОРМАЦІЙНІ СИСТЕМИ І ТЕХНОЛОГІЇ (ІСТ)

План:

1. Загальна характеристика сучасних напрямків розвитку ІСТ
2. Технології побудови ІС
3. Методи інтелектуального аналізу даних
4. Основні етапи та алгоритми інтелектуального аналізу даних
5. Огляд алгоритмів та ІС Data Mining

Література:

1. *Волькенштейн М. В.* Энтропия и информация. — М.: Наука, 1986. — 192 с.
2. *Глушков В. М.* Введение в АСУ. — К.: Тэхника, 1972. — 312 с.
3. *Кальянов Г. Н.* CASE структурный системный анализ. — М.: Лори, 1996. — 242 с.
4. *Мамиконов А. Г.* Информация и управление. — М.: Наука, 1975. — 184 с.
5. *Марка Д. А., Мак-Гоуэ К.* Методология структурного анализа и проектирования: Пер. с англ. — М.: Наука, 1993. — 240 с.
6. *Одинцов Б. Е.* Проектирование экономических экспертных систем: Учеб. пособие для студ. вузов, обучающихся по специальности «Информационные системы в экономике» — М.: Компьютер, 1996. — 166 с.

1. Загальна характеристика сучасних напрямків розвитку ІСТ

Одним із найважливіших практичних наслідків розвитку кібернетики можна вважати те, що вона стала теоретичним фундаментом для створення комп'ютерної техніки та сучасних ІСТ, які принципово змінили підходи до процесу обробки інформації та управління практично в усіх галузях людської діяльності. ІСТ настільки інтегровані у процеси обробки інформації та управління в економіці, що стали невід'ємною складовою економічної кібернетики, значно збагативши її арсенал.

У загальному випадку ***інформаційна технологія*** *— це сукупність методів і способів нагромадження*, *оброблення*, *зберігання*, *передавання*, *подання та використання інформації*.

Сучасний стан розвитку ІСТ характеризується стійкою тенденцією до зростання обсягів та інтенсивності інформаційних потоків майже в усіх галузях знань. При цьому зростання має приблизно експоненціальний характер. Діяльність будь-якої економічної системи, зокрема й підприємства (комерційного, виробничого, наукового і т. ін.) супроводжується нагромадженням, зберіганням та обробленням величезних масивів інформації. Тому без засобів продуктивної переробки потоків «сирих», первинних даних ефективне управління економічними системами практично неможливе.

Можна виокремити такі сучасні вимоги до даних і їх обробки:

- дані мають бути значного обсягу;
- характеризуватися різнорідністю (кількісною, якісною, текстовою);
- результати обробки мають бути конкретними й зрозумілими;
- інструменти для обробки первинних даних — простими в користуванні.

Усе це зумовило необхідність автоматизації аналізу даних комп'ютерною їх обробкою із застосуванням методів прикладної статистики та економетрії.

Нині існують численні інформаційні технології, спрямовані на полегшення економічної діяльності людини. Наявні системи поділяються на певні типи, передусім за безпосереднім призначенням та підходами, що використовуються в них. У галузі ІСТ умовно можна виокремити три напрямки розвитку, які доповнюють один одного, визначаючи тип ІС. Системи першого типу зорієнтовано на операційну обробку даних — *системи обробки даних*

(*СОД*). До них належать спеціалізовані пакети програм для статистичного аналізу, математичні пакети тощо. Другий тип ІС зорієнтований на задачі аналізу даних та управління — *системи підтримки та прийняття рішень* (*СППР*).

До третього, одного з найпоширеніших типів ІС, застосовуваних в управлінні, належать такі:

- *АСУ* — автоматизовані системи управління;
- *СППР* — системи підтримки прийняття рішення;
- *ЕС* — експертні системи.

Автоматизовані системи управління. АСУ мають широкий спектр застосування: від автоматизації базових функцій підприємства до автоматизації прийняття управлінських рішень. Розвиток цих систем відбувався від найпростіших систем обробки інформації до сучасних інтегрованих інформаційних комплексів. АСУ можна поділити на вузькоспеціальні та інтегровані.

Перші підтримують деякі спеціалізовані напрямки діяльності (наприклад, бухоблік, фінанси, кадри, маркетинг) та частково інших базових функціональних галузей. Інтегровані системи забезпечують повну підтримку більшості функціональних сфер діяльності підприємства, пропонуючи широкий перелік спеціалізованих рішень як для різноманітних видів діяльності, так і для всіляких аспектів управління (стратегічне планування, управління спеціальними видами активів і т. ін.).

Системи підтримки прийняття рішень. СППР призначені допомагати робити обґрунтований вибір із певного переліку альтернатив. Перш ніж набула поширення клієнт-серверна архітектура, застосовували два типи СППР: ІС для керівництва (управлінські) — Executive Information System (EIS), та системи підтримки рішень — Decision Support System (DSS). EIS створювались на великих ЕОМ і призначалися для керівництва верхнього рівня. DSS виконувались на робочих станціях і призначалися для менеджерів середньої ланки. Але останнім часом завдяки делегуванню повноважень із прийняття рішень середній та нижній ланці відмінності між цими типами СППР поступово зникають. У загальному випадку СППР складаються із СУБД, системи управління банком моделей та інтерфейсу користувача.

Експертні системи. ЕС — це ІС, що моделюють дії людини-експерта під час розв'язання задач у певній предметній галузі на основі логічного аналізу накопичених знань, що зберігаються в базі знань (БЗ) Мета досліджень з ЕС полягає передусім у розробці програм, які у процесі розв'язання задач, що виникають у слабо структурованій і важко формалізованій предметній галузі та є складними для експерта-людини, дають результати, не гірші за якістю та ефективністю рішенням, ніж експерти.

Експертні системи та системи штучного інтелекту відрізняються від систем обробки даних тим, що в них використовується символьний (а не числовий) спосіб подання інформації, символьний вивід та евристичний пошук розв'язку (а не виконання відомого алгоритму). Технологія ЕС нині використовується для розв'язання різних типів задач (інтерпретація, прогноз, діагностика, планування, конструювання, контроль, інструктування, управління) у найрізноманітніших проблемних галузях, таких як фінанси, нафтова та газова промисловість, гірнича справа, хімія, освіта, телекомунікації та зв'язок тощо. Нині спостерігається тенденція до дедалі більшої інтеграції ЕС та СППР, тому поступово ці типи ІС зближуються.

2. Технології побудови ІС

Сучасні концепції створення ІС ґрунтуються на таких підходах.

Об'єктно-орієнтований підхід дає змогу подати задачу розробки ІС як задачу побудови ієрархії об'єктів, що взаємодіють. При цьому об'єкти кожного рівня розглядаються як представники певних класів, що характеризуються наборами властивостей і методів. Функціонування ІС в об'єктно-орієнтованій методології описується за допомогою низки спеціалізованих діаграм. Однією з переваг такого підходу є наочність його засобів

(графічних) та можливість їх практичного застосування за допомогою уніфікованої мови моделювання UML.

UML (Unified modeling language) — уніфікована графічна мова моделювання призначена для візуалізації, специфікації, конструювання та документування систем, в яких провідну роль відіграє програмне забезпечення. За допомогою UML можна розробити докладний план створюваної системи, що відбиває не тільки її концептуальні елементи, такі як системні функції та бізнес-процеси, а й конкретні особливості реалізації, зокрема класи, записані спеціальними мовами програмування, схеми баз даних, а також програмні компоненти багаторазового використання.

CASE (Computer Aided System Engeneering) — технологія комп'ютерного проектування ІС, призначена для розробки складних ІС у цілому. Під CASE-технологією розуміють програмні засоби, що підтримують процеси створення та супроводження ІС (зокрема, аналіз і формулювання вимог), проектування прикладного програмного забезпечення (додатків) і баз даних, генерування коду, тестування, документування, конфігураційне керування, управління проектом та інші процеси.

CASE-технологія містить набір інструментальних засобів, що дають змогу в наочній формі моделювати будь-яку предметну область, аналізувати побудовану модель на всіх етапах розробки й супроводження ІС і створювати прикладні програми згідно з інформаційними потребами користувачів. Більшість наявних CASE-засобів ґрунтуються на методології структурного й об'єктно-орієнтованого аналізу та проектування, що передбачає використання специфікації у вигляді діаграм або текстів для описування зовнішніх вимог, зв'язків між моделями системи, динаміки поводження системи та архітектури програмних засобів.

SADT (Structure Analyse and Design Technic) — технологія структурного моделювання, призначена для побудови функціональної моделі об'єкта певної предметної області. Головна мета SADT-технології — описувати складні об'єкти як ієрархічні, багаторівневі модульні системи за допомогою невеликого набору типових елементів. До найістотніших властивостей SADT-технології належать:

- принцип побудови моделі згори вниз;
- реалізація ієрархічного, багаторівневого моделювання;
- можливість одночасно зі структуруванням проблеми розробляти структуру баз даних.

Сучасні концепції побудови СППР спрямовані на розв'язання суперечності між відсутністю корисної інформації, з одного боку, та наявністю величезних обсягів інформації — з другого. До найвідоміших підходів, спрямованих на підвищення ефективності зберігання та використання інформації, можна віднести:

- ♦ Data Warehouse — концепцію побудови сховища даних;
- ♦ Data Mart — вітрини даних;
- ♦ OLAP (On-Line Analitical Processing) — багатовимірний оперативний аналіз даних;
- ♦ Data Mining (DM) — інтелектуальний аналіз даних.

3. Методи інтелектуального аналізу даних

Технології аналізу даних, що базуються на застосуванні класичних статистичних підходів, мають низку недоліків. Відповідні методи ґрунтуються на використанні усереднених показників, на підставі яких важко з'ясувати справжній стан справ у досліджуваній сфері (наприклад, середня зарплата по країні не відбиває її розміру у великих містах та в селах). Методи математичної статистики виявилися корисними насамперед для перевірки заздалегідь сформульованих гіпотез та «грубого» розвідницького аналізу, що становить основу оперативної аналітичної обробки даних (OLAP).

Наприклад, дослідження спеціалістів Гарвардського інституту показують, що на основі наявної інформації за допомогою стандартних статистичних методів не можна було передбачити великої депресії кінця 1920-х років.

Окрім того, стандартні статистичні методи відкидають (нехтують) нетипові спостереження — так звані піки та сплески. Проте окремі нетипові значення можуть становити самостійний інтерес для дослідження, характеризуючи деякі виняткові, але важливі явища. Навіть сама ідентифікація цих спостережень, не говорячи про їх подальший аналіз і докладний розгляд, може бути корисною для розуміння сутності досліджуваних об'єктів чи явищ. Як показують сучасні дослідження, саме такі події можуть стати вирішальними щодо майбутнього поводження та розвитку складних систем.

Ці недоліки статистичних методів спонукали до розвитку нових методів дослідження складних систем, що базуються на нелінійній динаміці, теорії катастроф, фрактальній геометрії тощо (див. розд. 5).

Водночас постала нагальна потреба в такій технології, яка автоматично видобувала б із даних нові нетривіальні знання у формі моделей, залежностей, законів тощо, гарантуючи при цьому їхню статистичну значущість. Новітні підходи, спрямовані на розв'язання цих проблем, дістали назву *технологій інтелектуального аналізу даних*.

В основу цих технологій покладено концепцію шаблонів (патернів), що відбивають певні фрагменти багатоаспектних зв'язків у множині даних, характеризуючи закономірності, притаманні підвибіркам даних, які можна компактно подати у зрозумілій людині формі. Шаблони відшукують методами, що виходять за межі апріорних припущень стосовно структури вибірки та вигляду розподілів значень аналізованих показників. Важлива особливість цієї технології полягає в нетривіальності відшукуваних шаблонів. Це означає, що вони мають відбивати неочевидні, несподівані регулярності у множині даних, складові так званого прихованого знання. Адже сукупність первинних («сирих») даних може містити й глибинні шари знань.

Knowledge Discovery in Databases (дослівно: «виявлення знань у базах даних» — KDD) — аналітичний процес дослідження значних обсягів інформації із залученням засобів автоматизації, що має на меті виявити приховані у множині даних структури, залежності й взаємозв'язки. При цьому передбачається повна чи часткова відсутність апріорних уявлень про характер прихованих структур та залежностей. KDD передбачає, що людина попередньо осмислює задачу й подає неповне (у термінах цільових змінних) її формулювання, перетворює дані до формату придатного для їх автоматизованого аналізу й попередньої обробки, виявляє засобами автоматичного дослідження даних приховані структури й залежності, апробовує виявлені моделі на нових даних, не використовуваних для побудови моделей, та інтерпретує виявлені моделі й результати.

Отже, KDD — це синтетична технологія, що поєднує в собі останні досягнення штучного інтелекту, чисельних математичних методів, статистики й евристичних підходів. Методи KDD особливо стрімко розвиваються протягом останніх 20 років, а раніше задачі комп'ютерного аналізу баз даних виконувалися переважно за допомогою різного роду стандартних статистичних методів.

Data Mining (дослівно: «Розробка, добування даних» — DM) — дослідження «сирих» даних і виявлення в них за допомогою «машини» (алгоритмів, засобів штучного інтелекту) прихованих нетривіальних структур і залежностей, які раніше не були відомі й мають практичну цінність та придатні для того, щоб їх інтерпретувала людина.

Розглянемо відмінності між засобами Data Mining і OLAP. Технологія OLAP спрямована на підтримання процесу прийняття управлінських рішень і використовується з метою пошуку відповіді на запитання: чому деякі речі є такими, якими вони є насправді? При цьому користувач сам формує модель-гіпотезу про дані чи відношення між даними, а далі, застосовуючи серію запитів до бази даних, підтверджує чи відхиляє висунуті гіпотези. Засоби Data Mining відрізняються від засобів OLAP тим, що замість перевірки передбачуваних користувачем взаємозалежностей вони на основі наявних даних самі можуть будувати моделі, які дають змогу кількісно та якісно оцінювати ступінь впливу різних досліджуваних факторів на задану властивість об'єкта. Крім того, засоби DM дають змогу формулювати нові гіпотези про характер досі невідомих, але таких, що реально існують, залежностей між даними.

Засоби OLAP застосовуються на ранніх стадіях процесу KDD, оскільки вони дають змогу краще зрозуміти дані, що, у свою чергу, забезпечує ефективніший результат процесу KDD.

Головна мета технології KDD — побудова моделей і відношень, *прихованих у базі даних*, *тобто таких*, *які не можна знайти звичайними методами*. Варто зазначити, що на комп'ютери перекладаються не лише рутинні операції (скажімо, перевірка статистичної значущості гіпотез), а й операції, що донедавна були аж ніяк не рутинними (вироблення нових гіпотез). *KDD дає змогу побачити такі відношення між даними*, *що залишалися поза увагою дослідників*.

Будуючи моделі, ми встановлюємо кількісні зв'язки між характеристиками досліджуваного явища. Щодо призначення можна виокремити моделі двох типів: прогнозні та описові (дескриптивні). Моделі першого типу використовують набори даних із відомими результатами для побудови моделей, що явно прогнозують результати для інших наборів даних, а моделі другого типу описують залежності в наявних даних. Обидва типи моделей використовуються для прийняття управлінських рішень.

Технологія KDD дає змогу не лише підтверджувати (відкидати) емпіричні висновки, *а й будувати нові*, *невідомі раніше моделі*. Знайдена модель не зможе здебільшого претендувати на абсолютне знання, але вона надає аналітикові деякі переваги вже завдяки самому факту виявлення альтернативної статистично значущої моделі, а також, можливо, стає приводом для пошуку відповіді на запитання: чи справді існує виявлений взаємозв'язок і чи є він причинним? А це, у свою чергу, стимулює поглиблені дослідження, сприяючи глибшому розумінню досліджуваного явища.

Отже, *найважливіша мета застосування технології KDD до дослідження реальних систем — це поліпшення розуміння суті їх функціонування*.

Відзначимо, що процес виявлення знань не є цілком автоматизованим — він вимагає участі користувача (експерта, особи що приймає рішення). Користувач має чітко усвідомлювати, що він шукає, ґрунтуючись на власних гіпотезах. Зрештою замість того, щоб підтверджувати наявну гіпотезу, процес пошуку часто сприяє появі ряду нових гіпотез. Усе це позначається терміном «discovery-driven data mining» (DDDM), і терміни Data Mining, Knowledge Discovery у загальному випадку стосуються до технології DDDM.

4. Основні етапи та алгоритми інтелектуального аналізу даних

Виокремимо два типи задач, розв'язуваних із різною ефективністю різними методами KDD (хоча, втім, реальні задачі дослідження даних можуть охоплювати обидва типи).

Задачі першого типу полягають у побудові на підставі наявних даних різних моделей, якими можна скористатися з метою прогнозування та ухвалення рішення в майбутньому, за схожої ситуації.

Задачі другого типу характерні тим, що наголос у них робиться на з'ясуванні сутності залежностей у множині даних, а також взаємовпливу, тобто на побудові емпіричних моделей різних систем, які легко може сприймати людина. При цьому не так уже й важливо, щоб система добре передбачала і працювала в майбутньому, а важливо зрозуміти взаємні впливи досліджуваних закономірностей (що і чим визначається в наявному масиві даних). І навіть якщо встановлені закономірності належатимуть до специфічних особливостей саме конкретних досліджуваних даних і більше ніде не траплятимуться, але нам усе одно потрібно їх з'ясувати.

Розглянемо головні етапи (кроки), характерні для будь-якого дослідження даних за допомогою методів KDD і становлять основний цикл пошуку нового знання та його

оцінювання (рис. 5.1). Залежно від задачі кількість етапів, а також обсяг виконуваних на кожному з них дій можуть змінюватися, але загалом усі вони необхідні і так чи інакше мають належати процесу інтелектуального аналізу даних.

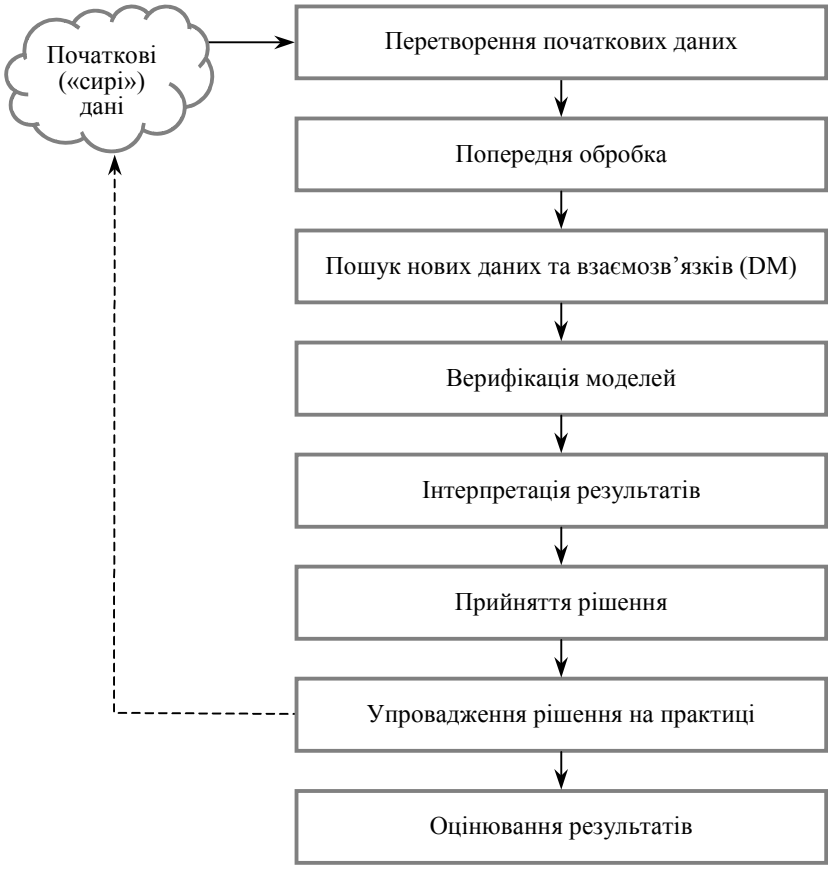

Рис. 5.1. Схема інтелектуального аналізу даних і оцінювання виявленого нового знання

Перший етап полягає у зведенні даних до форми, придатної для застосування конкретних реалізацій систем KDD. Нехай, скажімо, інформацію подано у вигляді текстів і потрібно побудувати автоматичний рубрикатор, класифікатор якихось анотацій, описів тощо. Вхідна інформація являє собою тексти в електронному вигляді, але практично жодна з наявних систем KDD не здатна працювати безпосередньо з текстами. Щоб працювати з певним текстом, ми маємо з вихідної текстової інформації заздалегідь дістати деякі похідні параметри (наприклад, частоту появи ключових слів, середню довжину речень, параметри, що характеризують сполучуваність тих чи інших слів у реченні тощо), тобто побудувати чіткий набір кількісних або якісних параметрів даного тексту. Ця задача найменш автоматизована в тому сенсі, що систему шуканих параметрів формує людина, хоча значення параметрів можуть обчислюватися автоматично в рамках відповідної технології первинної обробки даних. Вибравши параметри, дані можна подати у вигляді прямокутної таблиці, де кожний рядок характеризує окрему ознаку (стан, властивість) досліджуваного об'єкта, а кожний стовпець — ознаки (стани, властивості) всіх досліджуваних об'єктів. Рядки такої таблиці в теорії KDD, як і в теорії баз даних, називають *записами*, а стовпці — *полями*.

Практично всі наявні системи KDD працюють тільки зі щойно описаними прямокутними таблицями.

Здобута прямокутна таблиця — це лише «сировинний» матеріал для застосування методів KDD, і дані, що входять до неї, необхідно передусім обробити. По-перше, таблиця може містити параметри (ознаки об'єктів), що мають однакові значення в якомусь зі стовпців. Коли б досліджувані об'єкти мали тільки такі ознаки, усі вони були б абсолютно ідентичними. Звідси випливає, що відповідні ознаки жодним чином не характеризували б

досліджуваних об’єктів, а отже, їх потрібно вилучити з аналізу. Можлива й така ситуація, що деяка категоріальна ознака в усіх її записах має різні значення, через що відповідне поле не придатне для аналізу даних і його також доведеться вилучити. Нарешті може статися так, що полів буде дуже багато, і якщо ми всі їх намагатимемося досліджувати, то надто відчутно збільшиться час розрахунків, оскільки практично для всіх методів KDD характерна сильна (не менш ніж квадратична, а нерідко й експоненціальна) залежність часу розрахунків від кількості параметрів, тоді як залежність часу розрахунків від кількості записів лінійна або близька до неї.

Тому у процесі попередньої обробки даних необхідно, по-перше, розглянути множину всіх ознак, що стосуються шуканої залежності, вилучити з неї ті, які явно не придатні для подальшого дослідження, та виокремити ті, що найімовірніше ввійдуть у шукану залежність. Для цього, як правило, застосовують статистичні методи, що ґрунтуються на застосуванні кореляційного аналізу, лінійних регресій, тобто методи, що дають змогу швидко, хоча й наближено оцінити вплив одного параметра на інші.

Третій етап — безпосереднє застосування методів KDD за різними сценаріями, що містять складні комбінації тих методів, які допомагають аналізувати дані з різних поглядів. Власне, цей етап дослідження і називають Data Mining (добування даних).

Четвертий етап — верифікація та перевірка результатів, найчастіше здійснювані в такий спосіб. Усі наявні дані, що мають бути проаналізовані, розбивають на дві (як правило, не однакові за розміром) групи. У більшій групі даних за допомогою тих чи інших методів KDD дістають моделі й залежності, а в меншій виконують їх перевірку. Далі за різницею в точності між результатами, здобутими для обох груп, доходять висновку щодо адекватності й статистичної значущості побудованої моделі. Існує багато інших, складніших способів верифікації (перехресна перевірка, бутстреп-аналіз тощо), які дають змогу оцінити значущість побудованих моделей без розбиття даних на дві групи.

Нарешті, *на п’ятому етапі* знання, що їх здобула людина, автоматично інтерпретуються з метою їх використання для прийняття рішень та внесення сформульованих правил і залежностей до баз знань тощо. Цей етап часто передбачає застосування методів, що є проміжними між технологією KDD і технологією експертних систем. Від того, наскільки ефективним він буде, значною мірою залежить успіх розв’язання поставленої задачі.

Цим етапом і закінчується цикл KDD. Остаточне оцінювання вагомості здобутого нового знання виходить за рамки аналізу, автоматизованого чи традиційного, і стає можливим тільки після впровадження на практиці рішення, прийнятого на основі такого знання. Дослідженням практичних результатів, досягнутих за допомогою здобутого засобами KDD нового знання, завершується його оцінювання (див. рис. 5.1).

5. Огляд алгоритмів та ІС Data Mining

Data Mining — це сукупність багатьох різних методів здобування знань. Вибір методу часто залежить від типу наявних даних і від того, яку інформацію потрібно дістати.

До найпоширеніших методів можна віднести такі:

- *об’єднання* (association; іноді вживають термін affinity, що означає подібність, структурну близькість) — виокремлення структур, що повторюються в часовій послідовності. Цей метод визначає правила, за якими можна встановити, що один набір елементів корелює з іншим. Користуючись ним, аналізують ринковий кошик пакетів продуктів, розробляють каталоги, здійснюють перехресний маркетинг тощо;
- *аналіз часових рядів* (sequence-based analysis, або sequential association) дає змогу відшукувати часові закономірності між даними (трансакціями). Наприклад, можна відповісти на запитання: купівля яких товарів передує купівлі даного виду продукції? Метод

застосовується, коли йдеться про аналіз цільових ринків, керування гнучкістю цін або циклом роботи із замовником (Customer Lifecycle Management);

- *кластеризація* (clustering) — групування записів, що мають однакові характеристики, наприклад за близькістю значень полів у БД. Використовується для сегментування ринку та замовників. Можуть залучатися статистичні методи або нейромережі. Кластеризація часто розглядається як перший необхідний крок для подальшого аналізу даних;
- *класифікація* (classification) — віднесення запису до одного із заздалегідь визначених класів, наприклад під час оцінюваня ризиків, пов'язаних із видачею кредиту;
- *оцінювання* (estimation);
- *нечітка логіка* (fuzzy logic);
- *статистичні методи*, що дають змогу знаходити криву, найближче розміщену до набору точок даних;
- *генетичні алгоритми* (genetic algorithms);
- *фрактальні перетворення* (fractal-based transforms);
- *нейронні мережі* (neural networks) — дані пропускаються через шари вузлів, «навчених» розпізнавати ті чи інші структури — використовуються для аналізу переваг і цільових ринків,

а також для приваблювання замовників.

До DM можна віднести ще візуалізацію даних — побудову графічного образу даних, що допомагає у процесі загального аналізу даних вбачати аномалії, структури, тренди. Частково до DM примикають дерева рішень і паралельні бази даних.

DM тісно пов'язана (інтегрована) зі сховищами даних (Data Warehousing, DW), які, можна сказати, забезпечують роботу Data Mining.

Data Mining — міждисциплінарна технологія, що виникла й розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних тощо (рис. 5.2). Звідси й численні методи та алгоритми, реалізовані в різних дійових системах Data Mining. Багато з таких систем інтегрують у собі відразу кілька підходів. Проте, як правило, у кожній системі присутній певний ключовий компонент, на який робиться головна ставка.

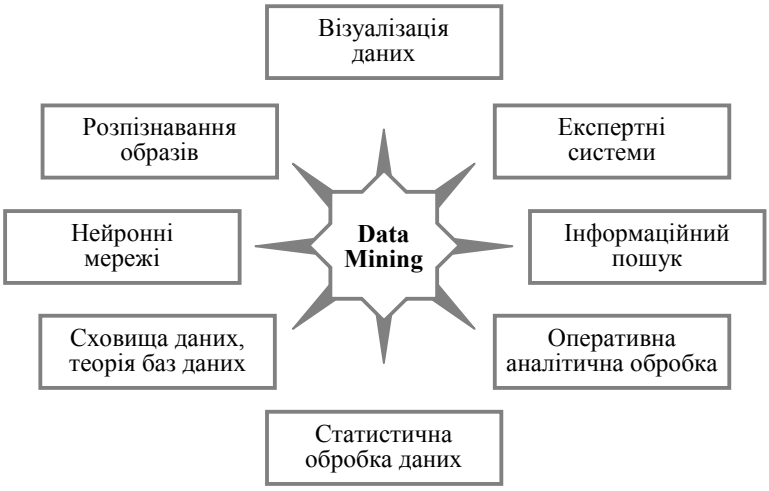

Рис. 5.2. Data Mining — міждисциплінарна галузь

Предметно-орієнтовані аналітичні системи. Такі системи дуже різноманітні. Найширший їх підклас, що набув поширення у сфері дослідження фінансових ринків, дістав назву «технічний аналіз». Він містить кілька десятків методів прогнозування динаміки цін і вибору оптимальної структури інвестиційного портфеля, які ґрунтуються на різних емпіричних моделях динаміки ринку. Зазначені методи, застосовуючи здебільшого нескладний статистичний апарат, максимально враховують специфіку своєї предметної галузі (професійна мова, системи різних індексів тощо). На ринку пропонується багато відповідних програм.

Статистичні пакети. Новітні версії майже всіх відомих статистичних пакетів поряд із традиційними статистичними методами містять також елементи Data Mining. Проте основна увага приділяється в них класичним підходам — кореляційному, регресійному, факторному аналізу та іншим. Недоліком відповідних систем можна вважати вимоги щодо спеціальної підготовки користувача. Існує, однак, і принциповий недолік статистичних пакетів, що обмежує їх застосування в Data Mining: більшість методів, що входять до складу пакетів, спираються на усереднені характеристики вибірки, які в разі дослідження складних життєвих явищ часто є фіктивними. І все ж деякі сучасні пакети пропонують модулі для інтелектуального аналізу. Наприклад, STATISTICA містить модуль Data Miner, що дає змогу будувати дерева рішень, нейронні мережі, виявляти IF THEN правила тощо.

До найпотужніших і найчастіше застосовуваних статистичних пакетів належать SAS (компанія SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA, EVIEWS тощо.

Нейронні мережі. Це великий клас систем, архітектура яких певною мірою аналогічна побудові нервової тканини з нейронів. В одній із найпоширеніших архітектур — багатошаровому перцептроні зі зворотним зв'язком помилки, імітується робота нейронів у складі ієрархічної мережі, де кожний нейрон вищого рівня з'єднаний своїми входами з виходами нейронів нижчого шару. На нейрони найнижчого шару подаються значення вхідних параметрів, на підставі яких потрібно приймати якісь рішення, прогнозувати розвиток ситуації тощо. Ці значення розглядаються як сигнали, що передаються в наступний шар, послаблюючи чи підсилюючи його залежно від числових значень (ваг), приписуваних міжнейронним зв'язкам.

У результаті на виході нейрона найвищого шару виробляється деяке значення, що розглядається як відповідь (реакція) всієї мережі на значення вхідних параметрів. Для того щоб мережу можна було використовувати надалі, її потрібно «навчити» на базі здобутих раніше даних, для яких відомі значення вхідних параметрів і правильні відповіді на них. Тренування полягає в доборі ваг міжнейронних зв'язків, що забезпечують найбільшу близькість відповідей мережі до відомих правильних відповідей.

Основним недоліком нейромережної технології є те, що вона потребує дуже великого обсягу навчальної вибірки. Ще один істотний недолік такий: навіть натренована нейронна мережа — це «чорна скринька». Знання, зафіксовані як ваги кількох сотень міжнейронних зв'язків, людина не в змозі проаналізувати й інтерпретувати.

До нейромережних систем належить, скажімо, BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic).

Системи міркувань на основі аналогічних випадків. Для того щоб зробити деякий прогноз або вибрати правильне рішення, зазначені системи (case based reasoning — CBR) відшукують у минулому близькі аналоги наявної ситуації, вибираючи ті самі відповіді, що були для них правильними. Тому цей метод ще називають методом «найближчого сусіда» (nearest neighbour). Останнім часом набув поширення також термін «memory based reasoning», який акцентує увагу на тому, що рішення приймається на підставі всієї інформації, нагромадженої в пам'яті.

Системи CBR забезпечують добрі результати в найрізноманітніших задачах. Головний їхній недолік полягає в тому, що вони взагалі не створюють будь-яких моделей чи правил, які узагальнюють попередній досвід, а ґрунтуються у виборі рішення на всьому масиві доступних історичних даних. Саме через це не можна встановити, на яких конкретно засадах системи CBR будують свої відповіді.

Інший недолік — певне «свавілля», що його припускаються такі системи, вибираючи міру «близькості», від якої залежить обсяг множини прецедентів, збережуваних у пам'яті з метою досягнення задовільної класифікації або прогнозу.

З-поміж систем CBR назвемо, наприклад, KATE tools (Acknosoft, Франція), Pattern Recognition Workbench (Unica, США).

Дерева рішень (decision trees). Дерева рішень є одним із найпопулярніших підходів до розв'язання задач Data Mining. Вони створюють ієрархічну структуру правил, класифікованих за схемою «ЯКЩО... ТО...» (if-then), яка має вигляд дерева. Для ухвалення рішення про те, до якого класу варто віднести деякий об'єкт (ситуацію, потрібно відповісти на запитання, що містяться у вузлах цього дерева, починаючи з його кореня. Запитання можуть бути, наприклад, такі: «Значення параметра *a* більше за *x*?». Якщо відповідь ствердна, відбувається перехід до правого вузла наступного рівня, якщо заперечна — до лівого вузла. Далі знову ставиться запитання, пов'язане з відповідним вузлом.

Популярність цього підходу зумовлюється наочністю та зрозумілістю. Але дерева рішень принципово не здатні знаходити «кращі» (найбільш повні і точні) правила в даних. Вони реалізують принцип послідовного перегляду ознак і збирають фактично уламки наявних закономірностей, створюючи лише ілюзію логічного висновку.

Проте більшість систем діють саме за цим методом. До таких належать, наприклад, See5/35.0 (RuleQuest, Австралія), Clementine (Integral Solutions, Великобританія), SIPINA (University of Lyon, Франція), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада).

Еволюційне програмування. Сучасний його стан схарактеризуємо, розглянувши систему PolyAnalyst, в якій гіпотези про вигляд залежності цільової змінної від інших змінних формулюються у вигляді програм, що подаються деякою внутрішньою мовою програмування. Процес побудови програм розгортається еволюційно в комплексі програм (на кшталт генетичних алгоритмів). Коли система відшукує програму, що більш-менш задовільно виражає шукану залежність, вона починає вносити до неї невеликі модифікації і добирає серед побудованих дочірніх програм ті, які підвищують точність. У такий спосіб система «вирощує» кілька генетичних ліній програм, що конкурують між собою стосовно точності вираження шуканої залежності. Спеціальний модуль системи PolyAnalyst перекладає знайдені залежності з внутрішньої мови системи зрозумілою користувачеві мовою (математичні формули, таблиці тощо).

Інший напрямок еволюційного програмування пов'язаний із пошуком залежності цільових змінних від решти у формі функцій певного вигляду. Наприклад, один із найбільш вдалих алгоритмів цього типу — метод групового врахування аргументів (МГВА) передбачає відшукання залежності у формі поліномів.

Генетичні алгоритми. Data Mining не є головною сферою застосування генетичних алгоритмів. Їх варто розглядати радше як могутній засіб розв'язання різноманітних комбінаторних задач та задач оптимізації. Проте генетичні алгоритми становлять нині стандартний інструментарій методів Data Mining.

Перший крок під час побудови генетичних алгоритмів — це кодування вихідних логічних закономірностей у базі даних, що їх іменують хромосомами, а весь набір таких закономірностей називають популяцією хромосом. Далі для реалізації концепції вибору вводиться спосіб зіставлення різних хромосом. Популяція обробляється за допомогою процедур репродукції, мінливості (мутацій), генетичної композиції. Ці процедури імітують біологічні процеси. Найважливіші з них такі:

- випадкові мутації даних в індивідуальних хромосомах, переходи і рекомбінації генетичного матеріалу, що міститься в індивідуальних батьківських хромосомах, і міграцію генів;
- у процесі виконання процедур на кожній стадії еволюції виходять популяції з дедалі досконалішими індивідами.

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

Генетичні алгоритми мають і низку недоліків. Критерій добору хромосом і використовуваних процедур є евристичним і зовсім не гарантує відшукання «найкращого»

рішення. Як і в реальному житті, еволюцію може «заклинити» на якій-небудь непродуктивній галузці. І, навпаки, може статися, що безперспективні батьки, яких вилучить з еволюції генетичний алгоритм, будуть здатні породити високоефективного нащадка. Це особливо стає помітним під час розв'язування багатовимірних задач зі складними внутрішніми зв'язками.

Як приклад можна згадати систему GeneHunter фірми Ward Systems Group.

Алгоритми обмеженого перебору. Ці алгоритми обчислюють частоти комбінацій простих логічних подій у підгрупах даних. Приклади простих логічних подій: $x = a$; $x < a$; $x > a$; $a < x < b$ тощо, де x — деякий параметр, a та b — константи. На підставі аналізу обчислених частот робиться висновок про корисність тієї чи іншої комбінації стосовно встановлення асоціацій у даних, класифікації, прогнозування і т. ін.

Найбільш виразним сучасним представником цього підходу є система WizWhy підприємства WizSoft.

Системи для візуалізації багатовимірних даних. Засоби графічного відображення даних тією чи іншою мірою підтримуються всіма системами Data Mining. Проте дуже значну частку ринку становлять системи, що спеціалізуються винятково на цій функції. Одна з них — програма DataMiner 3D словацької фірми Dimension 5.

У таких системах основна увага сконцентрована на дружньому користувальницькому інтерфейсі, що дає змогу асоціювати з аналізованими показниками різні параметри діаграми розсіювання об'єктів (записів) бази даних. До зазначених параметрів належать колір, форма, орієнтація щодо власної осі, розміри й інші властивості графічних елементів зображення. Крім того, системи візуалізації даних є зручними засобами для масштабування зображень.