**Лекція 5. Інтелектуальний аналіз даних (Data Mining)**

З штучним інтелектом тісно пов’язаний напрям інтелектуального аналізу даних - дейтамайнінгу, або здобування знань, що може використовуватись як один із способів одержання знань для систем штучного інтелекту. **Дейтамайнінг**  (Data Mining) - це дослідження й виявлення за допомогою спеціальних алгоритмів і засобів штучного інтелекту в «сирих» даних схованих знань, що раніше не були відомі і є нетривіальними, практично корисними і доступними для інтерпретації людиною.

Термін «Data Mining» запропонував Григорій П’ятецький-Шапіро у 1992 р.

В широкому сенсі  Data Mining - це сучасна концепція аналізу даних, яка передбачає те, що:

* дані можуть бути неточними, неповними (мати пропуски), суперечливими, різнорідними, непрямими, і при цьому мати гігантські об’єми; тому розуміння даних у конкретних застосуваннях вимагає значних інтелектуальних зусиль;
* самі алгоритми аналізу даних можуть мати «елементи інтелекту», зокрема, здатність навчатися на прецедентах, тобто робити загальні висновки на основі окремих спостережень; розроблення таких алгоритмів також потребує значних інтелектуальних зусиль;
* процеси перероблення сирих даних в інформацію, а інформації у знання вже не можуть бути виконані вручну, і потребують нетривіальної автоматизації.

**Основні методи дейтамайнінгу** включають:

* кластерний аналіз;
* нечітка логіка;
* дерева рішень;
* генетичні алгоритми;
* нейромережі та ін.

***Кластерний аналіз*** – процес розділення набору даних (чи об’єктів) у набір значущих підкласів, що називаються **кластерами**. Кластери допомагають користувачу зрозуміти природне групування чи структуру в наборі даних.

***Нечітка логіка*** застосовується для таких наборів даних, де приналежність даних до якої-небудь групи є імовірносною і може бути оцінена в інтервалі від 0 до 1.  На відміну від традиційної математики, що вимагає на кожному кроці моделювання точних і однозначних формулювань закономірностей, нечітка логіка пропонує підхід м’яких формулювань і надає математичні методи для представлення і оброблення таких даних.

***Дерево рішень*** – це схема (граф), яка відображає структуру задачі багатокрокового процесу прийняття рішень у вибраній сфері аналізу (класифікація набору даних або зразків, пошук оптимального рішення на множині альтернатив, отримання логічного висновку за допомогою евристики (бази правил)).

***Генетичні алгоритми***.  Ці алгоритми певним чином імітують процес природного добору і є потужним засобом рішення різноманітних комбінаторних задач і задач оптимізації. У цьому методі рішення подається у вигляді хромосоми, яку утворює набір генів. Пошук оптимального рішення при цьому схожий на еволюцію популяції індивідів, представлених їхніми наборами хромосом.

***Нейромережні і нейрокомп’ютерні технології***. Штучні нейронні мережі та нейрокомп’ютери значною мірою відтворюють принципи роботи людського мозку. Знання в них не відокремлені від процесора, а рівномірно розподілені й існують неявно у вигляді ваг синаптичних зв’язків. Такі знання не закладаються з самого початку, а набуваються в процесі навчання.

***Задачі Data Mining:***

1. Класифікація (Classification) - виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних - класи; за цими ознаками новий об'єкт можна віднести до того або іншого класу. Для вирішення задач класифікації можуть використовуватися методи: найближчий сусід (Nearest Neighbor); к-найближий сусід (k-Nearest Neighbor); байє-совські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks).

2. Кластеризація (Clustering) - результатом її є поділ об'єктів на групи.

3. Асоціація (Associations) - знаходять закономірності між пов'язаними подіями у наборі даних. Найбільш відомий алгоритм рішення задачі пошуку асоціативних правил - алгоритм Аргіогі.

4. Послідовність (Sequence), або послідовна асоціація (sequential association), - дає можливість знайти часові закономірності між транзакціями. Завдання послідовності подібне до асоціації, але її метою є встановлення закономірностей між подіями, пов'язаними за часом, тобто послідовність визначається високою ймовірністю ланцюжка пов'язаних за часом подій.

5. Прогнозування (Forecasting) - на основі особливостей історичних даних оцінюються майбутні значення показників. Застосовуються методи математичної статистики, нейронні мережі тощо.

6. Визначення відхилень (Deviation Detection), аналіз відхилень або викидів - виявлення й аналіз даних, що найбільше відрізняються від загальної чисельності даних, виявлення нехарактерних шаблонів.

7. Оцінювання (Estimation) - зводиться до прогнозу безперервних значень ознак.

8. Аналіз зв'язків (Link Analysis) - задача знаходження залежностей у наборі даних.

9. Візуалізація (Visualization, Graph Mining) - створюється графічний образ аналізованих даних. Для вирішення задач візуалізації використовуються графічні методи, що показують наявність закономірностей в даних.

10. Підбивання підсумків (Summarization) - опис конкретних груп об'єктів за допомогою аналізованого набору даних.

Зазначені вище задачі поділяються за призначенням на описові і предиктивні.

Описові, або дескриптивні (descriptive), задачі пов'язані з поліпшенням розуміння аналізованих даних. Ключовий момент у таких моделях - простота і прозорість результатів для сприйняття людиною. До такого типу задач належать кластеризація і пошук асоціативних правил.

Рішення предиктивних (predictive), або прогнозуючих, задач поділяється на два етапи. На першому етапі на підставі набору даних з відомими результатами будується модель. На другому етапі вона використовується для прогнозу результатів на підставі нових наборів даних. Вимагається, щоб побудовані моделі працювали максимально точно. До цього типу задач відносять задачі класифікації і регресії. Сюди можна віднести і задачу пошуку асоціативних правил, якщо результати її рішення можуть бути використані для прогнозу появи деяких подій.

За способами рішення задачі поділяють на такі, що вирішують за допомогою вчителя і без його допомоги. Категорія навчання з учителем представлена такими задачами Data Mining: класифікація, оцінка, прогнозування, категорія навчання; без учителя - задачею кластеризації.

У випадку рішення з допомогою вчителя задача аналізу даних розв'язується у кілька етапів. Спочатку за допомогою конкретного алгоритму Data Mining будується модель аналізованих даних - класифікатор. Потім класифікатор піддається навчанню. Іншими словами, перевіряється якість його роботи і, якщо вона незадовільна, відбувається додаткове навчання класифікатора. Так продовжується доти, доки не буде досягнуто необхідного рівня якості або не стане зрозуміло, що обраний алгоритм не працює коректно з даними, або дані не мають структури, здатної проявитися. До цього типу задач відносять задачі класифікації і регресії.

Рішення без допомоги вчителя об'єднує задачі, що виявляють описові моделі, наприклад, закономірності в часових рядах макропоказників. Очевидно, якщо ці закономірності існують, то модель має їх проявити. Перевагою цих задач є можливість їх рішення без будь-яких попередніх знань про дані аналізу. До них належать кластеризація і пошук асоціативних правил.

**Задача класифікації і регресії**

Під час аналізу часто необхідно визначити, до якого з відомих класів відносять досліджувані об'єкти, тобто як їх класифікувати.

Задачу класифікації розглядають як задачу визначення значення одного з параметрів аналізованого об'єкта на підставі значень інших параметрів. Досліджуваний параметр часто називають залежною змінною, а параметри, що беруть участь у його визначенні - незалежними змінними.

Задача класифікації і регресії розв'язується у два етапи. На першому виділяється навчальна вибірка. До неї входять об'єкти, для яких відомі значення як незалежних, так і залежних змінних.

На підставі навчальної вибірки будується модель визначення значення залежної змінної. її часто називають функцією класифікації або регресії. Для отримання максимально точної функції до навчальної вибірки пред'являються такі основні вимоги:

o кількість об'єктів, що входять до вибірки, має бути досить великою;

o до вибірки мають входити об'єкти, що представляють усі можливі класи у задачі класифікації або всю область значень у задачі регресії;

o для кожного класу в задачі класифікації або кожного інтервалу області значень у задачі регресії вибірка має містити достатню кількість об'єктів.

На другому етапі побудовану модель застосовують до об'єктів аналізу. Задача класифікації і регресії має геометричну інтерпретацію.

**Задача пошуку асоціативних правил**

Пошук асоціативних правил є поширеним застосуванням Data Mining. Суть задачі полягає у визначенні наборів об'єктів, що часто зустрічаються, в інформаційних масивах. Ця задача є окремим випадком задачі класифікації.

При аналізі потрібною є інформація про послідовність подій, що відбуваються. При виявленні закономірностей у таких послідовностях можна з певною часткою ймовірності передбачати появу подій у майбутньому, що дає змогу приймати правильніші рішення. Така задача є різновидом задачі пошуку асоціативних правил і називається ***секвенціональним аналізом.***

Він широко використовується, наприклад, в телекомунікаційних компаніях для аналізу даних про аварії на різних вузлах мережі.

**Задача кластеризації**

Задача кластеризації полягає в поділі об'єктів на групи подібних об'єктів, що називаються кластерами (cluster), тобто сукупності осіб, предметів. Задачі поділу множини елементів на кластери називають ***кластер-ним аналізом.***

Кластеризація може застосовуватися практично в будь-якій сфері, де необхідне дослідження експериментальних або статистичних даних.

Для задачі кластеризації характерна відсутність яких-не-будь відмінностей між змінними і об'єктами. Кластерний аналіз додає змогу розглядати досить великий обсяг інформації і різко скорочувати, стискати великі масиви інформації, робити їх компактними.

Слід зазначити деякі особливості, властиві задачі кластеризації.

Рішення залежить від природи об'єктів даних (і їх атрибутів), а також від представлення кластерів і передбачуваних відношень об'єктів даних і кластерів. Так, необхідно враховувати такі властивості, як можливість/неможливість приналежності об'єктів кільком кластерам. Необхідне визначення самого поняття приналежності кластеру: однозначна ймовірність приналежності, нечітка ступінь приналежності.

2) дані деталізуються для подальшої обробки, тобто необхідним є виявлення і використання формалізованих закономірностей або дистиляція шаблонів.

При технології дистиляції шаблонів один зразок (шаблон) інформації витягується з початкових даних і перетворюється у певні формальні конструкції, вид яких залежить від методу Data Mining. Цей процес відбувається на стадії вільного пошуку, у першої групи методів ця стадія - відсутня. На стадіях прогностичного моделювання і аналізу виключень використовуються результати стадії вільного пошуку. Методи цієї групи: логічні методи; методи візуалізації; методи крос-табуляції; методи, засновані на рівняннях.

Логічні методи, або методи логічної індукції, включають нечіткі запити й аналізи, символьні правила, дерева рішень, генетичні алгоритми.

Методи цієї групи придатні для інтерпретації. Вони підтримують знайдені закономірності у прозорому вигляді з погляду користувача. Методи крос-табуляції забезпечують пошук шаблонів.

Методи на основі рівнянь виражають наявні закономірності у вигляді математичних виразів - рівнянь. Основні методи цієї групи: статистичні методи і нейронні мережі.

Статистичні методи найчастіше застосовуються для вирішення задач прогнозування. Є багато методів статистичного аналізу даних, наприклад, кореляційно-регресійний аналіз, кореляція рядів динаміки, виявлення тенденцій динамічних рядів, гармонійний аналіз.

Інша класифікація поділяє все різноманіття методів Data Mining на дві групи: статистичні і кібернетичні методи. Ця схема поділу заснована на різних підходах щодо навчання математичним моделям.

Статистичні методи Data Mining. Ці методи включають: попередній аналіз природи статистичних даних (перевірка гіпотез стаціонарності, нормальності, незалежності, однорідності, оцінка виду функції розподілу, її параметрів); виявлення зв'язків і закономірностей (лінійний і нелінійний регресійний аналіз, кореляційний аналіз); багатовимірний статистичний аналіз (лінійний і нелінійний дискримінантний аналіз, кластерний аналіз, компонентний аналіз, факторний аналіз); динамічні моделі і прогноз на основі часових рядів.

Статистичні методи Data Mining поділяються на чотири групи методів: дескриптивний аналіз і опис початкових даних; аналіз зв'язків (кореляційний і регресійний аналіз, факторний аналіз, дисперсійний аналіз); багатовимірний статистичний аналіз (компонентний аналіз, дискримінантний аналіз, багатовимірний регресійний аналіз, канонічні кореляції); аналіз часових рядів (динамічні моделі і прогнозування).

Кібернетичні методи Data Mining. До цієї групи належать такі методи: еволюційне програмування; асоціативна пам'ять (пошук аналогів, прототипів); нечітка логіка; дерева рішень; системи обробки експертних знань, штучні нейронні мережі (розпізнавання, кластеризація, прогноз); генетичні алгоритми (оптимізація).

***Нейронні мережі*** (Neural Networks) - це клас моделей, що базуються на аналогії з роботою мозку людини і призначаються для вирішення різноманітних задач аналізу даних після проходження етапу навчання на даних.

***Нейронні мережі*** - це моделі біологічних нейронних мереж мозку, в яких нейрони імітуються однотипними елементами (штучними нейронами).

Нейронна мережа може бути представлена направленим графом зі зваженими зв'язками, у якому штучні нейрони є вершинами, а синаптичні зв'язки - дугами.

Серед сфер застосування нейронних мереж - автоматизація процесів розпізнавання образів, прогнозування показників діяльності підприємства, медична діагностика, прогнозування, адаптивне управління, створення експертних систем, організація асоціативної пам'яті, оброблення аналогових і цифрових сигналів, синтез й ідентифікація електронних систем.

За допомогою нейронних мереж можна, наприклад, передбачати обсяги продажу виробів, показники фінансового ринку, розпізнавати сигнали, конструювати самонавчальні системи.

Нейронна мережа є сукупністю нейронів, з яких складаються шари. У кожному шарі нейрони пов'язані з нейронами

**Моделі Data Mining**

Мета технології Data Mining - знаходження в даних таких закономірностей, які не можуть бути знайдені традиційними методами. Є два види моделей: предик-тивні та описові.

Предиктивні моделі будуються на підставі набору даних з відомими результатами. Вони використовуються для прогнозу результатів на підставі інших наборів даних. Вимагається, щоб модель працювала максимально точно, була статистично значимою і виправданою. До них належать моделі класифікації - описують правила або набір правил, відповідно до яких можна віднести опис будь-якого нового об'єкта до одного з класів. Такі правила будуються на підставі інформації про наявні об'єкти шляхом поділу їх на класи; моделі послідовностей - описують функції, що дають змогу прогнозувати зміну параметрів. Вони будуються на підставі даних про зміну певного параметра за минулий період часу.

Описові (descriptive) моделі пов'язані із залежностями в наборі даних, взаємного впливу різних чинників, тобто на побудові емпіричних моделей різних систем. Ключовий момент у таких моделях - легкість і прозорість для сприйняття людиною. Можливо, знайдені закономірності будуть специфічною межею саме конкретних досліджуваних даних і більше ніде не зустрінуться, але це все може бути корисним. До них належать такі види моделей:

o кластеризації - описують групи (кластери), на які можна поділити об'єкти, дані про яких піддаються аналізу. Групуються об'єкти (спостереження, події) на основі даних (властивостей), що описують суть об'єктів. Об'єкти усередині кластера мають бути подібними один до одного і відрізнятися від об'єктів, що ввійшли до складу інших кластерів;

o виключень - описують виняткові ситуації в записах, які різко відрізняються від основної множини записів;

o підсумкові (результатні) - виявлення обмежень на даних масиву аналізу. Подібні обмеження важливі для розуміння даних масиву, тобто це нове знання, здобуте в результаті аналізу. Таким чином, Data Summarization - це знаходження яких-небудь фактів, які істинні для всіх або майже всіх записів у вибірці даних, що вивчається, але які досить рідко зустрічалися в усьому різноманітті записів такого самого формату;

o асоціації - виявлення закономірностей між пов'язаними подіями.

Для побудови розглянутих моделей використовуються різні методи й алгоритми Data Mining.

Більшість аналітичних методів, що використовуться у технології Data Mining - це математичні алгоритми і методи. Зокрема до методів і алгоритмів Data Mining відносять штучні нейронні мережі, дерева рішень, символьні правила, метод найближчого сусіда і к-найближчого сусіда, метод опорних векторів, байєсовські мережі, лінійну регресію, кореляційно-регре-сійний аналіз, ієрархічні методи кластерного аналізу, неієрар-хічні методи кластерного аналізу, зокрема алгоритми /с-серед-ніх і к-медіа ми, методи пошуку асоціативних правил, метод обмеженого перебору, еволюційне програмування і генетичні алгоритми, різноманітні методи візуалізації даних тощо.

До базових методів Data Mining належать також підходи, що використовують елементи теорії статистики. Основна їх ідея зводиться до кореляційного, регресійного та інших видів статистичного аналізу. Основним недоліком їх є усереднювання значень, що призводить до втрати інформативності даних. Це у свою чергу спричинює зменшення кількості знань, що здобуваються.

Основним способом дослідження задач аналізу даних є їх відображення формалізованою мовою та їх подальший аналіз за допомогою моделі.

***Класифікація стадій Data Mining,*** Data Mining складається з кількох фаз: 1) виявлення закономірностей (вільний пошук); 2) використання виявлених закономірностей для прогнозу невідомих значень (прогностичне моделювання); 3) аналіз виключень - стадія призначена для виявлення і пояснення аномалій, знайдених у закономірностях.

***Класифікація методів Data Mining.*** Усі методи Data Mining поділяються на дві групи за принципом роботи з початковими навчальними даними: 1) дані після Data Mining зберігаються - кластерний аналіз, метод найближчого сусіда, метод -найближчого сусіда, міркування за аналогією (традукцією);

2) дані деталізуються для подальшої обробки, тобто необхідним є виявлення і використання формалізованих закономірностей або дистиляція шаблонів.

При технології дистиляції шаблонів один зразок (шаблон) інформації витягується з початкових даних і перетворюється у певні формальні конструкції, вид яких залежить від методу Data Mining. Цей процес відбувається на стадії вільного пошуку, у першої групи методів ця стадія - відсутня. На стадіях прогностичного моделювання і аналізу виключень використовуються результати стадії вільного пошуку. Методи цієї групи: логічні методи; методи візуалізації; методи крос-табуляції; методи, засновані на рівняннях.

Логічні методи, або методи логічної індукції, включають нечіткі запити й аналізи, символьні правила, дерева рішень, генетичні алгоритми.

Методи цієї групи придатні для інтерпретації. Вони підтримують знайдені закономірності у прозорому вигляді з погляду користувача. Методи крос-табуляції забезпечують пошук шаблонів.

Методи на основі рівнянь виражають наявні закономірності у вигляді математичних виразів - рівнянь. Основні методи цієї групи: статистичні методи і нейронні мережі.

Статистичні методи найчастіше застосовуються для вирішення задач прогнозування. Є багато методів статистичного аналізу даних, наприклад, кореляційно-регресійний аналіз, кореляція рядів динаміки, виявлення тенденцій динамічних рядів, гармонійний аналіз.

Інша класифікація поділяє все різноманіття методів Data Mining на дві групи: статистичні і кібернетичні методи. Ця схема поділу заснована на різних підходах щодо навчання математичним моделям.

**5.3. Технологія аналізу сховищ даних (Data Mining)**

Data Mining (добування знань, даних) - технологія аналізу сховищ даних, що грунтується на методах штучного інтелекту та інструментах підтримки прийняття рішень. Зокрема сюди входить знаходження трендів і комерційно корисних залежностей. Деколи використовують термін "knowledge discovery"

(виявлення знань) - виявлення прихованих структур (patterns) у сховищах даних, щоб перетворити їх на знання або термін "інтелектуальний аналіз даних". Всі ці терміни є синонімами.

Класичне визначення технології "видобування даних" (Data Mining) звучить таким чином: це виявлення в початкових ("сирих") даних - раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань. Тобто інформація, знайдена в процесі застосування методів Data Mining, повинна бути нетривіальною і раніше невідомою, наприклад, отриманий показник середніх продажів не є таким. Знання повинні описувати нові зв'язки між властивостями, передбачати значення та характеристики одних параметрів на основі інших.

Основна мета Data Mining полягає у виявленні прихованих правил і закономірностей у великих масивах даних. На відміну від оперативної аналітичної обробки даних (OLAP), у Data Mining основну задачу формулювання гіпотез і виявлення незвичайних шаблонів (закономірностей) перекладено з людини на комп'ютер.

Фірми Oracle, Microsoft, IBM т. ін. випустили ряд продуктів (DarWin, Microsoft SQL Server 200, IBM Intelligent for Data, відповідно), що реалізують алгоритми Data Mining і дають змогу автоматизувати процес аналізу даних.

Методи Data Mining дозволяють виявляти стандартні закономірності:

o асоціація (кілька подій пов'язані одна з одною, наприклад при купівлі пива дуже часто купують і чіпси чи горішки);

o послідовність (ланцюжок пов'язаних у часі подій, наприклад: нова квартира - нові меблі);

o кластеризація (відрізняється від класифікації тим, що групи заздалегідь не створені. Використовується для сегментації ринку і замовників);

o прогнозування (базою служить історична інформація. Ґрунтується на побудові математичних моделей).

Існують два способи впровадження нової інформаційної технології в локальні інформаційні структури:

1. пристосування її до організаційної структури підприємства;

2. модернізування організаційної структури з метою найбільш ефективного використання нової інформаційної технології.

Перший спосіб є дешевшим і не вимагає великих змін в організації діяльності підприємства. Проте ефект від його впровадження може бути незначним. Другий спосіб вимагає більших капіталовкладень, але забезпечує якісно новий рівень діяльності підприємства чи організації.

Наведені нижче приклади з різних областей економіки демонструють основну перевагу методів Data Mining - здатність виявлення нових знань, які неможливо отримати методами статистичного, регресивного аналізу або економетрики.

1. Клієнти компанії за допомогою одного з інструментів Data Mining були об'єднані в сегменти з схожими ознаками. Це дозволило проводити різну маркетингову політику і будувати окремі моделі поведінки для кожного сегменту. Найважливішими чинниками для розподілу були: віддаленість регіону клієнта, сфера діяльності, середньорічні суми операцій, кількість операцій за тиждень.

2. Автоматичний аналіз банківської бази даних кредитних операцій фізичних осіб виявив правила, за якими позичальникам відмовляли у видачі кредиту. Вирішальними чинниками, виявились: термін кредиту, середньомісячний дохід і витрати позичальника. Надалі це враховувалося при експрес-кредитуванні.

3. При аналізі бази даних клієнтів страхової компанії був встановлений соціальний портрет людини, що страхує життя - це виявився чоловік 35-50 років, що має двох і більше дітей і середньомісячний дохід вище $2000.

Висунення гіпотез. Під гіпотезою в даному випадку будемо розуміти припущення про вплив певних чинників на досліджувану задачу. При цьому форма цієї залежності в значення не має. Тобто можна припустити, що на продаж впливає відхилення ціни на товар від середньоринкової, але при цьому не зазначати, як саме цей чинник впливає на продажі. Для вирішення цієї задачі і використовується Data Mining. Наприклад, для обробки даних про обсяги продажів певного товару висувається гіпотеза про вплив чинника його відсутності у торговій точці.

Автоматизувати процес висунення гіпотез не представляється можливим, принаймні, на сьогоднішньому рівні розвитку технологій. Цю задачу повинні вирішувати експерти - фахівці в даній області. З використанням їх знань про предмет, методом опитування накопичується максимальна кількість гіпотез/припущень.

Результатом цього кроку буде список з описом всіх чинників. Наприклад, для задачі прогнозування попиту це може бути список наступного вигляду: сезон, день тижня, обсяги продажів за попередні тижні, обсяги продажів за аналогічний період минулого року, рекламна компанія, маркетингові заходи, якість продукції, бренд, відхилення ціни від середньоринкової, наявність даного товару у конкурентів, тощо.

При розв'язку певної задачі необхідно створювати спеціалізований набір даних, причому їх велика кількість не обумовлює якість рішення.

Після підготовки таблиці з описом чинників експертно оцінюється значущість кожного з чинників. Ця оцінка не є остаточною, вона служить відправним пунктом. В процесі аналізу може виявитися, що чинник, який експерти вважали вкрай важливим, таким по суті не є і, навпаки, незначущий із їхньої точки зору чинник може мати значний вплив. У будь-якому випадку, всі варіанти проаналізувати відразу неможливо, потрібно від чогось відштовхуватися, цією крапкою і є оцінка експертів. До того ж, досить часто реальні дані підтверджують їх оцінку. Результатом цього кроку може бути таблиця наступного вигляду (табл.5.2).

Коли гіпотезу висуває один експерт, задача оцінки значущості істотно спрощується. Проте, із зростанням складності системи, зростає і складність отримання адекватної оцінки експертів.

**Методи інтелектуальних обчислень**

Засоби інтелектуальних обчислень використовують наступні основні методи:

* нейронні мережі;
* дерева рішень;
* системи міркування на основі аналогічних випадків;
* алгоритми визначення асоціацій і послідовностей;
* нечітка логіка;
* генетичні алгоритми;
* еволюційне програмування;
* візуалізація даних.

Іноді застосовується комбінація перерахованих методів.  
  
**Нейронні мережі** відносяться до класу нелінійних адаптивних систем з архітектурою, що умовно імітує нервову тканину з нейронів. Математична модель нейрона являє собою деякий універсальний нелінійний елемент із можливістю широкої зміни і налаштування його характеристик. В одній з найбільш розповсюджених нейромережевих архитектур - багатошаровому перцептроні зі зворотним поширенням похибки - моделюється робота нейронів у складі ієрархічної мережі, де кожен нейрон прошарку з'єднаний своїми виходами з входами нейронів наступного прошарку. На нейрони вхідного прошарку подаються значення вхідних параметрів, на основі яких виробляються обчислення, необхідні для прийняття рішень, прогнозування розвитку ситуації і т.п. Ці значення розглядаються як сигнали, що передаються в наступний прошарок. Величина послаблення або підсилення сигналу залежить від числових значень (ваг), приписуваних міжнейронним зв'язкам. У результаті цього на виході нейрона вихідного прошарку продукується значення, що розглядається як відповідь, реакція всієї мережі на введені значення вхідних параметрів. Для того щоб мережу можна було застосовувати надалі, її треба "навчити" на прикладах, для яких відомо і значення вхідних параметрів, і правильні відповіді на них.

Процес "навчання" складається в підборі ваг межнейронних зв'язків і модифікації внутрішніх параметрів передатної функції нейронів. Для кожного сполучення навчальних даних на виході мережі вихідні значення порівнюються з відомим результатом. Якщо вони розрізняються, то обчислюється похибка, що враховується при обробці у вузлах мережі. Зазначені кроки повторюються, поки не виконається умова останову, наприклад необхідна похибка не буде перевищувати заданої величини.

Отже, нейронні мережі уявляють собою сукупність зв'язаних між собою вузлів, що отримують вхідні дані, здійснюють їх обробку і генерують на виході результат. Між вузлами видимих вхідного і вихідного прошарків може знаходитися певне число прихованих прошарків обробки. Нейронні мережі реалізують непрозорий процес. Це означає, що побудована модель, як правило, не має чіткої інтерпретації. Багато пакетів, що реалізують алгоритми нейронних мереж, застосовуються не лише в сфері обробки комерційної інформації, без них важко обійтися при рішенні більш загальних задач розпізнавання образів, скажемо розшифровки рукописного тексту чи інтерпретації кардіограм.

Апаратні або програмні реалізації алгоритмів нейромереж називаються нейрокомп'ютером.

* Нейрокомп'ютери дають стандартний спосіб рішення багатьох нестандартних задач. І неважливо, що спеціалізована машина краще вирішує один клас задач. Важливіше, що один нейрокомп'ютер вирішить і цю задачу, і другу, і третю і не треба щораз проектувати спеціалізовану ЕОМ, нейрокомп'ютер зробить все сам і майже не гірше.
* Замість програмування навчання. Нейрокомп'ютер вчиться, потрібно лише формувати навчальні множини. Праця програміста заміняється новою працею вчителя. Краще це чи гірше? Ні те, ні інше. Програміст наказує машині 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 природної мови на іншу і т.п. Саме для таких задач звичайно важко скласти явний алгоритм.
* Нейронні мережі дозволяють створити ефективне програмне та математичне забезпечення для комп'ютерів з високим ступенем розпаралелювання обробки.
* Нейрокомп'ютери "демократичні", вони також дружні, як текстові процесори, тому з ними може працювати будь-який, навіть зовсім недосвідчений користувач.

**Дерева рішень** - це метод, придатний не тільки для рішення задач класифікації, але і для обчислень і тому досить широко застосовується в області фінансів і бізнесу, де частіше зустрічаються задачі чисельного прогнозу. В результаті застосування цього методу для навчальної вибірки даних створюється ієрархічна структура правил класифікації типу, "ЯКЩО... ТОДІ...", що має вид дерева. Для того щоб вирішити, до якого класу віднести деякий об'єкт або ситуацію, ми відповідаємо на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Питання можуть мати вид "Значення параметра A більше Х ? " або виду "Значення змінної В належить підмножині ознак С ? ". Якщо відповідь позитивна, ми переходимо до правого вузла наступного рівня, якщо негативна - то до лівого вузла; потім знову відповідаємо на запитання, зв'язані з відповідним вузлом. Таким чином ми, зрештою, доходимо до одного з кінцевих вузлів - листів, де знаходиться вказівка, до якого класу треба віднести розглянутий об'єкт. Цей метод добрий тим, що таке представлення правил наочно і його легко зрозуміти.

Сьогодні спостерігається підйом інтересу до продуктів, що застосовують дерева рішень. В основному це пояснюється тим, що багато комерційних проблем розв'язується ними швидше, ніж алгоритмами нейронних мереж. До того ж вони більш прості і зрозумілі для користувачів.

У той же час не можна сказати, що дерева рішень завжди діють безвідмовно: для визначених типів даних вони можуть виявитися неприйнятними. Для дерев рішень дуже гостро постає проблема значимості. Справа в тому, що окремим вузлам на кожному новому побудованому рівні дерева відповідає все менше і менше число записів даних - дерево може сегментувати дані на велику кількість окремих випадків. Чим більше цих окремих випадків, чим менше навчальних прикладів попадає в кожен такий окремий випадок, тим менш надійної стає їх класифікація. Якщо побудоване дерево занадто "рунисте" - складається з невиправдано великого числа дрібних гілочок - воно не буде давати статистично обґрунтованих відповідей. Як показує практика, у більшості систем, що використовують дерева рішень, ця проблема не знаходить задовільного рішення.

**Системи міркування на основі аналогічних випадків**. Ідея алгоритму вкрай проста. Для того щоб зробити прогноз на майбутнє або вибрати правильне рішення, ці 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 міркування на основі аналогічних випадків показують добрі результати в найрізноманітніших задачах. Головний їх мінус полягає в тому, що вони взагалі не створюють яких-небудь моделей або правил, що узагальнюють попередній досвід, - у виборі рішення вони ґрунтуються на всьому масиві доступних історичних даних, тому неможливо сказати, на основі яких конкретно факторів ці системи будують свої відповіді.

**Алгоритми виявлення асоціацій** знаходять правила про окремі предмети, що з'являються разом в одній економічній операції, наприклад в одній покупці. Послідовність - це теж асоціація, але залежна від часу.

Асоціація записується як А->Б, де А називається лівою частиною або передумовою, Б - правою частиною або наслідком.

Частота появи кожного окремого предмета або групи предметів, визначається дуже просто - підраховується кількість появи цього предмета у всіх подіях (покупках) і ділиться на загальну кількість подій. Ця величина вимірюється у відсотках і зветься "поширеність". Низький рівень поширеності (менш одного тисячної відсотка) говорить про те, що така асоціація не істотна.

Для визначення важливості кожного отриманого асоціативного правила необхідно одержати величину, що зветься "довірчість А до Б" (взаємозв'язок А і Б). Ця величина показує як часто з появою А з'являється Б и 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 придбають і товар Б.

Необхідно відзначити, що якщо поширеність А не дорівнює поширеності Б, то і довірчість А до Б не дорівнює довірчості Б к А. Справді, покупка комп'ютера частіше веде до покупки дискет, ніж покупка дискети до покупки комп'ютера.

Ще однією важливою характеристикою асоціації є потужність асоціації. Чим більше потужність, тим 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 тільки після цього визначають взаємозв'язок і поширеність. Єдиним розходженням таких алгоритмів є швидкість або ефективність знаходження асоціацій. Це важливо через величезну кількість комбінацій, які необхідно перебрати для знаходження найбільш значимих 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 більш 50% і поширеність не менш 2,5%"

При знаходженні послідовностей додається змінна часу, що дозволяє працювати із серією подій для знаходження послідовних асоціацій протягом деякого періоду часу.

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 відбутись через те, що клієнт не буде довго ходити по магазині в пошуках бажаного товару, купуючи при цьому ще щось, що потрапляється на очі, і те, що він ніколи не планував придбати.

**Нечітка логіка** застосовується для таких наборів даних, де належність даних до якої-небудь групи є імовірністю в інтервалі від 0 до 1. Чітка логіка маніпулює результатами, що можуть бути або істиною, або не істиною. Нечітка логіка застосовується в тих випадках, коли необхідно маніпулювати ступенем "може бути" у доповненні до "так" або "ні".  
Областю впровадження алгоритмів нечіткої логіки є всілякі аналітичні системи, у тому числі:

* нелінійний контроль за процесами (виробництво);
* вдосконалення стратегій керування і координації дій, наприклад складне промислове виробництво;
* самонавчаючі системи (або класифікатори);
* дослідження ризикових і критичних ситуацій;
* розпізнавання образів;
* фінансовий аналіз (ринки цінних паперів);
* дослідження даних (корпоративні сховища).

У Японії цей напрямок переживає дійсний бум. Тут функціонує спеціально створена лабораторія Laboratory for International Fuzzy Engineering Research ( LIFE ). Програмою цієї організації є створення більш близьких людині обчислювальних пристроїв. LIFE поєднує 48 компаній у числі яких знаходяться: Hitachi, Mitsubishi, NEC, Sharp, Sony, Honda, Mazda, Toyota. З закордонних (не Японських) учасників LIFE можна виділити: IBM, Fuji Xerox, до діяльності LIFE також виявляє інтерес NASA.

Потужність і інтуїтивна простота нечіткої логіки як методології вирішення проблем гарантує її успішне використання у вбудованих системах контролю й аналізу інформації. При цьому відбувається підключення людської інтуїції і досвіду оператора.

На відміну від традиційної математики, що вимагає на кожному кроці моделювання точних і однозначних формулювань закономірностей, нечітка логіка пропонує зовсім інший рівень мислення, завдяки якому творчий процес моделювання відбувається на найвищому рівні абстракцій, при якому постулюється 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 різноманітних комбінаторних задач і задач оптимізації. Проте, генетичні алгоритми ввійшли зараз у стандартний інструментарій методів інтелектуальних обчислень. Цей метод названий так тому, що в якійсь степені імітує процес природного відбору в природі. Нехай нам треба знайти рішення задачі, найбільш оптимальне з погляду деякого критерію, де кожне рішення цілком описується певним набором чисел чи величин нечислової природи. Скажемо, якщо нам треба вибрати сукупність фіксованого числа параметрів ринку, найбільш виражено впливаючих 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, будуть тоді називатися генами. Пошук оптимального рішення при цьому схожий на еволюцію популяції індивідів, представлених їхніми наборами хромосом. У цій 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. У результаті зміни поколінь виробляється таке рішення поставленої задачі, що вже не може бути далі поліпшено.

Генетичні алгоритми мають два слабких місця. По-перше, сама постановка задачі в їхніх термінах не дає можливості проаналізувати статистичну значимість одержуваного з їх допомогою рішення і, по-друге, ефективно сформулювати задачу, визначити критерій відбору хромосом під силу тільки фахівцю. У силу цих факторів сьогодні генетичні алгоритми треба розглядати скоріше як інструмент наукового дослідження, ніж засіб аналізу даних для практичного застосування в бізнесі і фінансах.

**Еволюційне програмування** сьогодні наймолодша ділянка інтелектуальних обчислень. Суть методу в тому, що гіпотези про вид залежності цільової змінної від інших змінних формулюються системою у виді програм на деякій внутрішній мові програмування. Якщо це універсальна мова, то теоретично на ній можна виразити залежність будь-якого виду. Процес побудови цих програм будується як еволюція у світі програм (цим метод небагато схожий на генетичні алгоритми). Коли система знаходить програму, що досить точно виражає шукану залежність, вона починає вносити в неї невеликі модифікації і відбирає серед побудованих таким чином дочірніх програм ті, які підвищують точність. Отже, система "вирощує" кілька генетичних ліній програм, що конкурують між собою в точності вираження шуканої залежності. Спеціальний транслюючий модуль, переводить знайдені залежності з внутрішньої мови системи на зрозумілу користувачу мову (математичні формули, таблиці й ін.), роблячи їх легкодоступними. Для того щоб зробити отримані результати ще зрозумілішими для користувача-нематематика, є великий арсенал різноманітних засобів візуалізації виявлених залежностей.

Пошук залежності цільових змінних від інших проводиться у формі функцій якогось визначеного виду. Наприклад, в одному з найбільш вдалих алгоритмів цього типу - методі групового урахування аргументів (МГУА) залежність шукають у формі поліномів. Причому складні поліноми заміняються декількома більш простими, враховуючі тільки деякі ознаки (групи аргументів). Звичайно використовуються попарні об'єднання ознак. Цей метод не має істотних переваг у порівнянні з нейронними мережами з їх готовим набором стандартних нелінійних функцій, але, отримані формули залежності, у принципі, піддаються аналізу й інтерпретації (хоча на практиці все-таки буває занадто складна для цього).

**Комбіновані методи.** Часто виробники з'єднують зазначені підходи. Об'єднання в собі алгоритми нейронних мереж і технології дерев рішень повинно сприяти побудові більш точної моделі і підвищенню її швидкодії. Програми візуалізації даних у певному сенсі не є засобом аналізу інформації, оскільки вони тільки представляють її користувачу. Проте, візуальне представлення, скажемо, відразу чотирьох змінних досить виразно узагальнює надвеликі обсяги даних. Отже виробники розуміють, що для рішення кожної проблеми варто застосовувати оптимальний метод.