Математичне моделювання електротехнічних систем

Лекція 5

3.2 Історичний екскурс Data Mining

Область Data Mining почалася з семінару (англ. workshop), проведеного Григорієм Патецьким-Шапіро в 1989 році.
http://ru.wikipedia.org/wiki/Data_mining – cite_note-comp-0
Раніше, працюючи в компанії GTE Labs, Григорій Патецький-Шапіро зацікавився питанням: чи можна автоматично знаходити певні правила, щоб прискорити деякі запити до крупних баз даних. Тоді ж було запропоновано два терміни — Data Mining (який слід переводити як «розкопка даних») і Knowledge Discovery In Data (який слід переводити як «відкриття знань в базах даних»).

У 1993 році вийшла перша розсилка «Knowledge Discovery Nuggets», а в 1994 році був створений один з перших сайтів по Data Mining.

Постановка задачі

Спочатку, завдання ставиться наступним чином:

• є достатньо крупна база даних;

• передбачається, що в базі даних знаходяться якісь «приховані знання».

Необхідно розробити методи виявлення знань, прихованих у великих об'ємах початкових «сирих» даних.

Що означає «приховані знання»? Це повинні бути обов'язково знання:

• раніше не відомі — тобто, такі знання, які повинні бути новими (а не підтверджуючими якісь раніше отримані відомості);

• нетривіальні — тобто, такі, які не можна просто так побачити (при безпосередньому візуальному аналізі даних або при обчисленні простих статистичних характеристик);

• практично корисні — тобто, такі знання, які представляють цінність для дослідника або споживача;

• доступні для інтерпретації — тобто, такі знання, які легко представити в наочній для користувача формі і легко пояснити в термінах наочної області.

Цими вимоги, багато в чому, визначають суть методів Data mining і те, в якому вигляді і в якому співвідношенні в технології Data mining використовуються системи управління базами даних, статистичні методи аналізу і методи штучного інтелекту.

Data mining і бази даних

Методи Data mining має сенс застосовувати тільки для достатньо великих баз даних. У кожній конкретній області досліджень існує свій критерій «великости» бази даних.

Розвиток технологій баз даних спочатку привів до створення спеціалізованої мови — мови запитів до баз даних. Для реляційних баз даних — це мова SQL, яка надала широкі можливості для створення, зміни і витягання даних, що зберігаються. Потім виникла необхідність в отриманні аналітичної інформації (наприклад, інформації про діяльність підприємства за певний період), і тут виявилось, що традиційні реляційні бази даних, добре пристосовані, наприклад, для ведення оперативного обліку (на підприємстві), погано пристосовані для проведення аналізу. Це привело, у свою чергу, до створення т.з. «сховищ даних», сама структура яких якнай кращим способом відповідає проведенню всебічного математичного аналізу.

Data mining і статистика. У основі методів Data mining лежать математичні методи обробки даних, включаючи і статистичні методи. У промислових рішеннях, нерідко, такі методи безпосередньо включаються в пакети Data mining. Проте, слід враховувати, що статистичні методи, по-перше, ґрунтуються на статистичній природі аналізованих явищ (наприклад, зазвичай постулювали форму розподілу випадкової величини), а, по-друге, результати статистичних методів, як правило, є тривіальними (легко розраховуються), практично даремними (наприклад, всілякі середні) і такими, що важко інтерпретуються (ті ж середні), що повністю розходиться з цілями і завданнями Data mining. Проте, статистичні методи використовуються, але їх застосування обмежується виконанням тільки певних етапів дослідження.

Data mining і штучний інтелект

Знання, що здобуваються методами Data mining прийнято представляти у вигляді моделей. Як такі моделі виступають:

- асоціативні правила;
- дерева рішень;
- кластери;
- математичні функції.

Методи побудови таких моделей прийнято відносити до області т.з. «штучного інтелекту».

Класи систем Data Mining

Data Mining є мультидисциплінарною областю, виниклої прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних і ін., що розвивається на базі досягнень Звідси велика кількість методів і алгоритмів, реалізованих в різних системах Data Mining, що діють. Багато хто з таких систем інтегрує в собі відразу декілька підходів. Проте, як правило, в кожній системі є якась ключова компоненту, на яку робиться головна ставка.

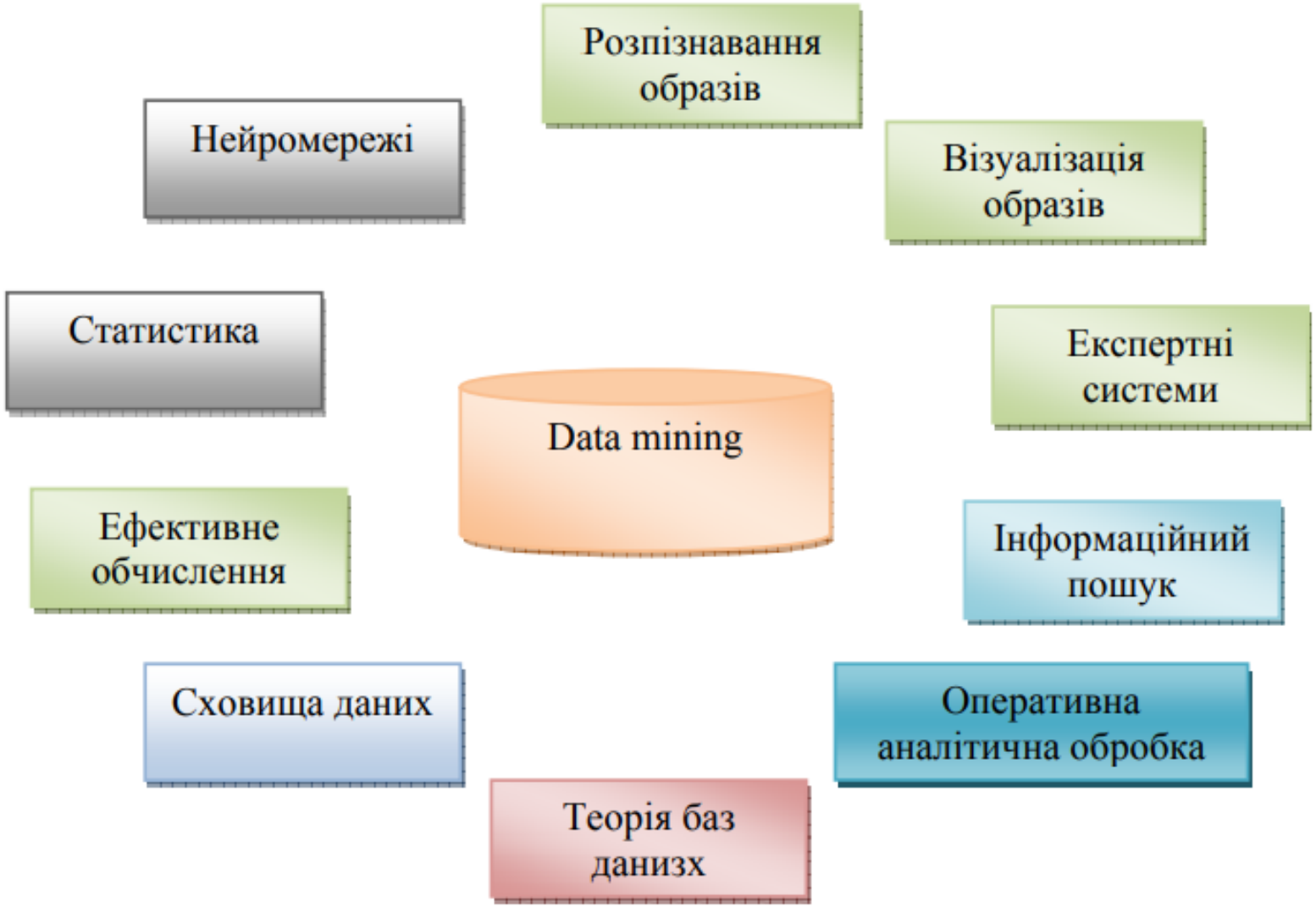

Рисунок 4.1 – Data Mining — мультидисциплінарня область

Предметно-орієнтовані аналітичні системи

Предметно-орієнтовані аналітичні системи дуже різноманітні. Найбільш широкий підклас таких систем, що набув поширення в області дослідження фінансових ринків, носить назву "Технічний аналіз". Він є сукупністю декількох десятків методів прогнозу динаміки цін і вибору оптимальної структури інвестиційного портфеля, заснованих на різних емпіричних моделях динаміки ринку. Ці методи часто використовують нескладний статистичний апарат, але максимально враховують ту, що склалася своїй області специфіку Розпізнавання образів, Нейромережі, Статистика, Візуалізація образів, Експертні системи Data mining, Ефективне обчислення, Інформаційний пошук, Оперативна аналітична обробка, Сховища даних, Теорія баз даних, (професійна мова, системи різних індексів і ін.). На ринку є безліч програм цього класу. Як правило, вони досить дешеві (зазвичай $300–1000)

Статистичні пакети. Останні версії майже всіх відомих статистичних пакетів включають разом з традиційними статистичними методами також елементи Data Mining. Але основна увага в них приділяється все ж таки класичним методикам — кореляційному, регресійному, факторному аналізу і іншим. Недоліком систем цього класу вважають вимогу до спеціальної підготовки користувача. Також відзначають, що могутні сучасні статистичні пакети є дуже "ваговитими" для масового застосування у фінансах і бізнесі. До того ж часто ці системи вельми дорогі — від $1000 до $15000. Є ще серйозніший принциповий недолік статистичних пакетів, що обмежує їх застосування в Data Mining. Більшість методів, що входять до складу пакетів спираються на статистичну парадигму, в якій головними фігурантами служать усереднені характеристики вибірки. Як приклади найбільш потужних і поширених статистичних пакетів можна назвати SAS (компанія SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA та інші.

Нейронні мережі Це великий клас систем, архітектура яких має аналогію (як тепер відомо, досить слабку) з побудовою нервової тканини з нейронів. У одній з найбільш поширеної архітектури, багатошаровому перцептроні із зворотним розповсюдженням помилки, імітується робота нейронів у складі ієрархічної мережі, де кожен нейрон більш високого рівня сполучений своїми входами з виходами нейронів шару, що пролягає нижче. На нейрони самого нижнього шару подаються значення вхідних параметрів, на основі яких потрібно ухвалювати якісь рішення, прогнозувати розвиток ситуації і так далі Ці значення розглядаються як сигнали, що передаються в наступний шар, ослабляючись або посилюючись залежно від числових значень (вагів), що приписуються міжнейронним зв'язкам. В результаті на виході нейрона самого верхнього шару виробляється деяке значення, яке розглядається як відповідь — реакція всієї мережі на введені значення вхідних параметрів.

Для того, щоб мережу можна було застосовувати надалі, її раніше треба "натренувати" на отриманих раніше даних, для яких відомі і значення вхідних параметрів, і правильні відповіді на них. Тренування полягає в підборі вагів міжнейронних зв'язків, що забезпечують найбільшу близькість відповідей мережі до відомих правильних відповідей. Основним недоліком нейромережевої парадигми є необхідність мати дуже великий об'єм повчальної вибірки. Інший істотний недолік полягає в тому, що навіть натренована нейронна мережа є чорним ящиком.

Знання, зафіксовані як ваги декількох сотень міжнейронних зв'язків, абсолютно не піддаються аналізу і інтерпретації людиною (відомі спроби дати інтерпретацію структурі настроєної нейромережі виглядають непереконливими – система “KINOsuitePR”). Приклади нейромережевих систем — BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic). Вартість їх досить значна: $1500–8000.

Системи міркувань на основі аналогічних випадків

Ідея систем case based reasoning — CBR — на перший погляд украй проста. Для того, щоб зробити прогноз майбутнє або вибрати правильне рішення, ці системи знаходять у минулому близькі аналоги наявної ситуації і вибирають ту ж відповідь, яка була для них правильним. Тому цей метод ще називають методом "найближчого сусіда" (nearest neighbour). Останнім часом поширення набув також термін memory based reasoning, який акцентує увагу, що рішення ухвалюється на підставі всієї інформації, накопиченої в пам'яті.

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

Іншою мінус полягає в свавіллі, яке допускають системи CBR при виборі міри "близькості". Від цього заходу найрішучішим чином залежить об'єм безлічі прецедентів, які потрібно зберігати в пам'яті для досягнення задовільної класифікації або прогнозу. Приклади систем, що використовують CBR, — KATE tools (Acknosoft, Франція), Pattern Recognition Workbench (Unica, США).

Дерева рішень (decision trees)

Дерева рішення є одним з найбільш популярних підходів до вирішення завдань Data Mining. Вони створюють ієрархічну структуру класифікуючих правил типу "ЯКЩО... ТО..." (if-then), що має вид дерева. Для ухвалення 40 рішення, до якого класу віднести деякий об'єкт або ситуацію, потрібно відповісти на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Питання мають вигляд "значення параметра А більше х?". Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня, якщо негативний — те до лівого вузла; потім знову слідує питання, пов'язане з відповідним вузлом.

Популярність підходу пов'язана як би з наочністю і зрозумілістю. Але дерева рішень принципово не здатні знаходити “кращі” (якнайповніші і точніші) правила даних. Вони реалізують наївний принцип послідовного перегляду ознак і “чіпляють” фактично осколки справжніх закономірностей, створюючи лише ілюзію логічного виводу. Разом з тим, більшість систем використовують саме цей метод. Найвідомішими є See5/C5.0 (RuleQuest, Австралія), Clementine (Integral Solutions, Великобританія), SIPINA (University of Lyon, Франція), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада). Вартість цих систем варіюється від 1 до 10 тис. дол.

4.1.6 Еволюційне програмування