Лекція 5. Моделювання СЕС на базі методів інтелектуального аналізу даних (Data Mining)

Мета: ознайомитись з сутністю та призначенням методології інтелектуального аналізу даних та основними моделями інтелектуальних обчислень.

План

5.1. Поняття інтелектуального аналізу даних.
5.2. Етапи та методи знаходження нових знань.
5.3. Основні моделі інтелектуальних обчислювань.

Перелік ключових термінів і понять: *інтелектуальний аналіз даних, Data Mining, нейронні мережі, дерева рішень, генетичні алгоритми, еволюційне програмування.*

5.1. Поняття інтелектуального аналізу даних.

Більшість організацій накопичують під час своєї діяльності величезні обсяги 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 у середовищі даних, – моделей, які не можуть бути знайдені звичайними методами.

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 у сучасному трактуванні приблизно з першої половини 1990-х років. Донині обробка й аналіз даних здійснювалися в рамках прикладної статистики, при цьому в основному вирішувалися завдання обробки невеликих баз даних.

В основу сучасної технології Data Mining покладена концепція шаблонів (паттернів), що відображають фрагменти багатоаспектних взаємин у даних. Ці шаблони є закономірностями, властивими підвибіркам даних, які можуть бути компактно виражені в зрозумілій людині формі. Пошук шаблонів проводиться

методами, які не обмежені рамками апріорних припущень про структуру вибірки і виду розподілів значень аналізованих показників.

У 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 у різних сферах людської діяльності».

Сутність та мету технології Data Mining можна охарактеризувати так: це технологія, яка призначена для пошуку у великих обсягах даних неочевидних, об'єктивних і корисних на практиці закономірностей.

Неочевидних – 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.

Data Mining – мультидисциплінарна область, що виникла і розвивалася на базі таких наукових напрямів як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних, математичні дисципліни, алгоритмізація, машинне навчання.

Data Mining – є потужним інструментом для полегшення і поліпшення роботи аналітика.

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 у порівнянні з іншими методами аналізу є очевидною. Більшість статистичних методів для виявлення взаємозв'язків у даних використовують концепцію усереднювання по вибірці, що призводить 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 майбутнє.

Якщо розглядати майбутнє Data Mining у *короткостроковій перспективі*, то очевидно, що розвиток цієї технології найбільш зумовлений розвитком сфер, пов'язаних із 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 – це може бути пошук інтелектуальними агентами як нового вигляду лікування різних захворювань, так і нового розуміння природи всесвіту.

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 будь-які дані, це може бути сфера роздрібної торгівлі, банківська справа, телекомунікації, страхування, управління виробництвом, менеджмент якості, молекулярна генетика і генна інженерія, медицина, прикладна хімія тощо.

Приклад застосування Data Mining у банківській справі.

Банківські установи сьогодні збирають докладну інформацію про кожну окрему транзакцію клієнта, його депозитні і кредитні рахунки. Досягнення технології Data Mining використовуються в банківській справі для вирішення таких поширених завдань:

– виявлення шахрайства з кредитними картками. Шляхом аналізу минулих транзакцій, які згодом виявилися шахрайськими, банк виявляє деякі стереотипи такого шахрайства;
– сегментація клієнтів. Розбиваючи клієнтів на різні категорії, банки роблять свою маркетингову політику більш цілеспрямованою і результативною, пропонуючи різні види послуг різним групам клієнтів;
– прогнозування змін клієнтури. Data Mining допомагає банкам будувати прогнозні моделі цінності своїх клієнтів, і відповідним чином обслуговувати кожну категорію.

5.2. Етапи та методи знаходження нових знань.

Побудова моделі інтелектуального аналізу даних є складовою частиною більш масштабного процесу, який включає всі етапи, починаючи з визначення базової проблеми, яку модель вирішуватиме, до розгортання моделі в робочому середовищі. Цей процес може бути заданий за допомогою таких шести базових кроків (рис. 5.1):

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.

Рис. 5.1. Етапи інтелектуального аналізу даних

Хоча процес, що ілюструється за допомогою рис. 5.1, 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.

Може виникнути необхідність в оновленні вже розгорнених моделей за рахунок нових даних, що поступили. Отже, важливо розуміти, що створення моделі інтелектуального аналізу даних – це процес, і що кожен крок такого процесу може бути повторений стільки разів, 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 проєкту 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 та ін.).

Оцінювання даних. При побудові моделі доцільно пам'ятати одне правило, що стосується коректності вхідних даних: «Якщо на вхід задачі поступає «сміття», то і результатом теж буде «сміття». Перед «завантаженням» даних у сховище необхідно врахувати, що різні джерела даних можуть бути спроєктовані під певні задачі і, відповідно, виникають проблеми, пов'язані з об'єднанням даних: різні формати представлення числових даних (наприклад, цілі або дійсні); 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. Аналітик повинен 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. У цей же час відбувається очищення, тобто виправлення всіх виявлених помилок. 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.

Відбір даних. Якщо сховище сформоване і визначені типи моделей, які будуть побудовані для розв'язання задач, відбувається відбір даних необхідних саме для цих моделей.

Перетворення даних. Служить для збагачення отриманої бази, тобто додавання різних відносин на основі існуючих полів (не борг і дохід, а відношення боргу 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 низку моделей для знаходження однієї, що найбільше задовольняє поставленим цілям. Моделі можна розділити на дві групи:

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

Після того, як визначено 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.

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 надає кращі результати. Не можна бути впевненим, що певна технологія працюватиме краще за інші. 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. Усе 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 у 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 із даними, що призводять до некоректності моделі, тому дуже важливо перевірити модель у реальному середовищі.

Використання моделі. Після побудови і оцінки моделі, її можна використовувати різними способами. Наприклад, базуючись на результатах використання моделі, аналітик може рекомендувати дії, які можна починати у 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. Процедура знаходження знання за допомогою цих методів може об'єднуватися із знаннями експертів і застосовуватися до даних у 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. Усі 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.

Можна виділити принаймні ***шість методів виявлення й 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 у 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. Певний ефективний класифікатор може використовуватися для класифікації нових записів у базі даних в уже існуючі класи, і в цьому випадку він набуває характеру прогнозу. Наприклад, класифікатор, що вміє 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 набувати цільова (залежна) змінна, що обчислюється на заданому наборі значень вхідних (незалежних) 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 за їх допомогою можна передбачати і поведінку системи в майбутньому.

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 пов'язано між собою. Класичний приклад аналізу структури покупок відноситься до представлення придбання якої-небудь кількості товарів як окремої 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. 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 у 85% випадків. Маючи такі дані, менеджерам легко оцінити, наскільки дієва надана знижка.

Послідовність має місце, якщо існує ланцюжок зв'язаних у 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 й порядок, у 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 придбають ще й холодильник.

5.3. 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 у даних.

На практиці широке застосування знайшли такі **інструменти (моделі та алгоритми) інтелектуальних обчислень**:

- нейронні мережі;
- дерева рішень;
- системи міркувань 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 із 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 розв'язання багатьох нестандартних задач. І неважливо, що спеціалізована машина краще вирішить один клас завдань. Важливіше, що один нейрокомп'ютер розв'яже і цю задачу, і іншу, і третю, і не треба кожного разу проєктувати спеціалізовану ЕОМ, нейрокомп'ютер зробить все сам і майже не гірше;

- замість програмування застосовується навчання. Нейрокомп'ютер навчається, потрібно лише формувати навчальну множину. Отже, робота програміста замінюється новою роботою вчителя. 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 іншу і т.п. Саме для таких завдань зазвичай важко скласти явний алгоритм;

- нейронні мережі дозволяють створити ефективне програмне і математичне забезпечення для комп'ютерів із високим ступенем розпаралелювання обробки;

- нейрокомп'ютери «демократичні», вони прості, як текстові процесори, тому з ними може працювати будь-якій, навіть зовсім недосвідчений користувач.

Дерева рішень – метод, широко вживаний в сфері фінансів і бізнесу, де частіше зустрічаються задачі числового прогнозу. У 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 за Х?» або вигляду «Чи належить значення змінної В підмножині ознак С?». 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 представлення правил і його легко зрозуміти.

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

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 міркувань 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 міркувань на основі аналогічних випадків дають гарні результати в різних завданнях. 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 $A \to B$, де A називається передумовою, B – 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 одної тисячною відсотка) говорить про неістотність асоціації.

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

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

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 X довірчість більше 50% і поширеність не менше 2,5%». При знаходженні

послідовностей додається змінна часу, що дозволяє працювати із серією подій для знаходження послідовних асоціацій впродовж 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 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 неможливо оцінити чітко («так» або «ні»), проте, можна оцінити мірою, що визначена на інтервалі від 0 до 1.

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).

В 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, серед яких: Hitachi, Mitsubishi, NEC, Sharp, Sony, Honda, Mazda, Toyota. З іноземних учасників LIFE можна виділити: IBM, Fuji Xerox, до діяльності LIFE також виявляє цікавість 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 та 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 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 розв'язання різних комбінаторних задач і задач оптимізації. Проте генетичні алгоритми увійшли зараз до стандартного інструментарію методів інтелектуальних обчислень. Цей метод названий так тому, що якоюсь мірою імітує процес природного

(еволюційного) відбору в природі. Нехай потрібно знайти розв'язки задач, оптимальні з погляду деякого критерію, де кожний розв'язок цілком описується певним набором чисел або величин нечислової природи. Наприклад, якщо 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 розв'язку поставленої задачі. Значення параметрів, що визначають розв'язок, називаються генами. Пошук оптимального розв'язку при цьому схожий на еволюцію популяції індивідів, представлених наборами хромосом.

В еволюції діють три механізми:

по-перше, *відбір найсильніших* – наборів хромосом, яким відповідають найбільш оптимальні розв'язки;

по-друге, *схрещування* – виробництво нових індивідів за допомогою змішування хромосомних наборів відібраних індивідів;

по-третє, *мутації* – випадкові зміни генів у деяких індивідів популяції. У результаті зміни поколінь виробляється розв'язок поставленої задачі, який вже далі не може бути покращеним.

Генетичні алгоритми мають два слабкі місця. По-перше, постановка задачі не дає можливості проаналізувати статистичну значущість отримуваного з їх допомогою розв'язку 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, ніж 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 зрозумілу користувачеві мову (математичні формули, таблиці та ін.), роблячи їх досяжними. Для того, щоб зробити отримані результати зрозумілішими для користувача-нематематика, 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 (МГОА) 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 у порівнянні з нейронними мережами з готовим набором стандартних нелінійних функцій, але отримані формули залежності, в принципі, піддаються аналізу й інтерпретації (хоча на практиці це все-таки складно).

Програми візуалізації даних у певному значенні не є засобом аналізу інформації, оскільки вони тільки представляють її користувачеві. Проте візуальне представлення, наприклад, відразу чотирьох змінних наочно узагальнює величезні обсяги даних (рис. 5.2).

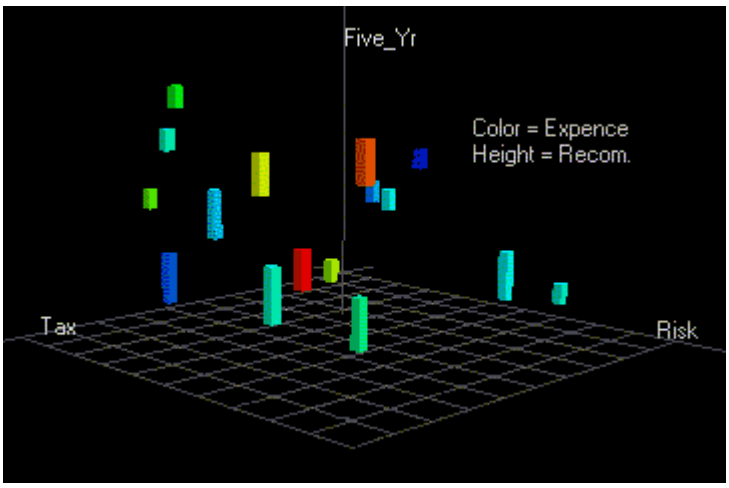

Рис. 5.2. Приклад візуалізації показників діяльності інвестиційних фондів

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 «найкращий» метод.

Питання для самоконтролю

1. У чому полягають відмінності інтелектуального аналізу даних від інших методів аналізу даних?
2. Що є передумовами для успішного застосування інтелектуального аналізу даних?
3. Назвіть основні методи виявлення та аналізу знань.
4. Що є основними особливостями моделі на основі нейронної мережі?
5. У чому полягає основна проблема при побудові моделі дерева рішень?
6. Як ще називають алгоритм системи роздумів на основі аналогічних випадків?
7. Які три механізми покладено в основу генетичних алгоритмів?
8. У чому полягає сутність еволюційного програмування?