УДК 681.3, 621.3

О.М. Колодчак
Національний університет “Львівська політехніка”,
кафедра електронних обчислювальних машин

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

Розглянуто методи інтелектуального аналізу даних. Проаналізовано сферу застосування інтелектуального аналізу даних та існуючі системи. Зроблено висновки стосовно перспектив використання методів інтелектуального аналізу даних.

Ключові слова: інтелектуальний аналіз даних, прогностичне моделювання, прогноз невідомих значень, прогнозування розвитку процесів, аналіз виключень.

Overviewed discovery-driven data mining. An application of data mining domain and existent systems are analysed on their basis. Drawn conclusion in relation to the prospects of the use of methods of data mining – systems.

Key words: discovery-driven data mining, predictive modeling, outcome prediction, forecasting, forensic analysis.

Вступ

У зв'язку з удосконаленням технологій запису та зберігання даних людство отримало колосальні потоки інформаційної “руди” в усіляких областях. Діяльність будь-якого підприємства (комерційного, виробничого, медичного, наукового тощо) тепер супроводжується реєстрацією та записом усіх подробиць його діяльності. Став очевидним той факт, що без продуктивної переробки, потоки сирих даних нікому не потрібні. Специфіка сучасних вимог до такої переробки є такою:

- дані мають необмежений обсяг;
- дані є різнорідними (кількісними, якісними, текстовими);
- результати повинні бути конкретні та зрозумілі;
- інструменти для обробки сирих даних повинні бути прості у використанні.

Алгоритми традиційної математичнї статистики тривалий час, як основні, підтимували концепцію усереднення з вибірки, що зводиться до операцій над фіктивними величинами (типу середньої температури аудиторій в усіх приміщеннях університету, середньої висоти будинку міста, що складається з палаців і халуп тощо). Методи математичної статистики виявилися корисними переважно для перевірки заздалегідь сформульованих гіпотез (verification-driven data mining) і для “грубого” розвідницького аналізу, що становить основу оперативної аналітичної обробки даних (online analytical processing, OLAP).

В основу інтелектуального аналізу даних (discovery-driven data mining, Data Mining, ІАД) покладена концепція шаблонів (паттернів), що відбивають фрагменти багатоаспектних взаємин у даних. Ці шаблони являють собою закономірності, властиві підвибіркам даних, які можуть бути компактно виражені у зрозумілій людині формі. Пошук шаблонів проводиться методами, не обмеженими апріорними припущеннями про структуру вибірки, та видами розподілів значень аналізованих показників. Важливе положення ІАД – нетривіальність розшукуваних шаблонів. Це означає, що знайдені шаблони повинні відбивати неочевидні, несподівані (unexpected) регулярності в даних, що становлять так звані приховані знання (hidden knowledge). До суспільства прийшло розуміння, що сирі дані (raw data) містять глибинний шар знань, за грамотного “розкопування” якого можуть бути виявлені справжні “самородки”.

Отже, інтелектуальний аналіз даних – це процес виявлення в сирих даних раніше невідомих, нетривіальних, фактично корисних і доступних інтерпретації знань, необхідних для прийняття рішень у різних сферах людської діяльності.

Огляд літературних джерел

Розвиток методів запису і зберігання даних привів до бурхливого зростання обсягів збираної та аналізованої інформації. Обсяги даних настільки великі, що людина просто невзмозі проаналізувати їх самостійно, хоча необхідність проведення такого аналізу цілком очевидна, адже в цих "сирих даних" є знання, які можуть бути використані для прийняття рішень [1]. Традиційна математична статистика, як було сказано, ще довгий час претендувала на роль основного інструменту аналізу даних, не відповідала проблемам, що виникали. Тому виникла необхіднсть у розвитку нових сучасних методологій обробки та аналізу даних. Такою новою методологією і став інтелектуальний аналіз даних ІАД. Причини популярності ІАД такі:

- стрімке накопичення даних (рахунок йде вже на экзабайти);
- загальна комп'ютеризація бізнес-процесів;
- проникнення Інтернет в усі сфери діяльності;
- прогрес в області інформаційних технологій: вдосконалення СУБД і сховищ даних; прогрес в області виробничих технологій: стрімке зростання продуктивності комп'ютерів, обсягів накопичувачів, впровадження Grid систем.

Алгоритми, що використовуються в ІАД, вимагають великої кількості обчислень. Раніше це було стримувальним чинником широкого практичного застосування ІАД, проте сьогоднішнє зростання продуктивності сучасних процесорів зняло гостроту цієї проблеми. Тепер за прийнятний час можна провести якісний аналіз сотень тисяч і мільйонів записів. ІАД – міждисциплінарна галузь, що виникла і розвивалася на основі таких наук, як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних тощо [2].

До методів і алгоритмів ІАД належать такі: штучні нейроні мережі, дерева рішень, символьні правила, методи найближчого сусіда і к-найближчого сусіда, метод опорних векторів, байєсові мережі, лінійна регресія, кореляційно-регресійний аналіз; ієрархічні методи кластерного аналізу, неієрархічні методи кластерного аналізу, зокрема і алгоритми к-середніх і к-медіани; методи пошуку асоціативних правил, зокрема алгоритм Apriori; метод обмеженого перебору, еволюційне програмування і генетичні алгоритми, різноманітні методи візуалізації даних і безліч інших методів. Варто зазначити, що більшість методів ІАД була розроблена у межах теорії штучного інтелекту. Єдиної думки щодо того, які задачі необхідно зараховувати до ІАД, немає. Більшість авторитетних джерел перераховує такі: класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз виявлення відхилень, оцінювання, аналіз зв'язків, підведення підсумків. Розглянемо деякі з них [3].

Класифікація (Classification). Це найпростіша і найпоширеніша задача ІАД. В результаті розв'язання задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних – класи; за цими ознаками новий об'єкт можна зарахувати до того чи іншого класу. Для розв'язання задачі класифікації можуть використовуватися методи: найближчого сусіда (Nearest Neighbor); к-ближнього сусіда (к-Nearest Neighbor); байєсових мереж (Bayesian Networks); індукції дерев рішень; нейронних мереж (neural networks).

Кластеризація (Clustering). Кластеризація є логічним продовженням ідеї класифікації. Це є складніша задача. Особливість кластеризації полягає у тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи. Прикладом методу задачі кластеризації є особливий вид нейроних мереж (карти Кохонена), що самоорганізовуються без вчителя.

Асоціація (Associations). У процесі розв'язання задачі пошуку асоціативних правил відшукуються закономірності міжзв'язаними подіями в наборі даних. Відмінність асоціації від двох попередніх задач ІАД: пошук закономірностей здійснюється не на основі властивостей об'єкта, що аналізується, а між кількома подіями, які відбуваються одночасно. Найвідоміший алгоритм розв'язку задачі пошуку асоціативних правил – алгоритм Apriori.

Послідовність (Sequence), або послідовна асоціація (sequential association). Послідовність дає змогу знайти тимчасові закономірності між транзакціями. Задача послідовності подібна до асоціації, але її метою є встановлення закономірностей не міжодночасно наступаючими подіями, а між подіями, пов'язаними в часі (тобто, що відбуваються з деяким певним інтервалом у часі. Цю задачу ІАД також називають задачею знаходження послідовних шаблонів (sequential pattern). Правило послідовності: після події X через певний час відбудеться подія У.

Прогнозування (Forecasting). В результаті розв’язання задачі прогнозування на основі особливостей існуючих даних оцінюються пропущені або ж майбутні значення цільових числових показників. Для розв’язання таких задач широко застосовуються методи математичної статистики, нейроні мережі тощо.

Візуалізація (Visualization, Graph Mining). В результаті візуалізації створюється графічний образ аналізованих даних. Для розв’язання задачі візуалізації використовуються графічні методи, що показують наявність закономірностей у даних. Приклад методів візуалізації – представлення даних в 2D- і 3D-вимірюваннях.

Підведення підсумків (Summarization) – задача, мета якої – опис конкретних груп об'єктів з аналізованого набору даних тощо [4].

Задачі ІАД, залежно від моделей, що використовуються, можуть бути дескриптивними і прогнозуючими. В результаті розв’язання описових (descriptive) задач аналітик одержує шаблони, що описують дані, які піддаються інтерпретації. Ці задачі описують загальну концепцію аналізованих даних, визначають інформативні, підсумкові, відмітні особливості даних.

Прогнозуючі (predictive) задачі ґрунтуються на аналізі даних, створенні моделі, прогнозі тенденцій або властивостей нових або невідомих даних.

ІАД може складатися з двох або трьох стадій [4]:

Стадія 1. Виявлення закономірностей (вільний пошук).

Стадія 2. Використовування виявлених закономірностей для прогнозу невідомих значень (прогностичне моделювання).

На додаток до цих стадій інколи вводять стадію оцінювання (валідації), наступну за стадією вільного пошуку [5]. Мета валідації – перевірка достовірності знайдених закономірностей. Проте вважається, що валідація здебільшого є частиною першої стадії, оскільки в реалізації багатьох методів, зокрема нейронних мереж і дерев рішень, передбачений розподіл загальної множини даних на навчальні і перевіркові, і останні уможливлюють перевіряти достовірність отриманих результатів.

Стадія 3. Аналіз виключень – стадія, призначена для виявлення і пояснення аномалій, знайдених у закономірностях.

Вільний пошук (Discovery). На стадії вільного пошуку здійснюється дослідження набору даних з метою пошуку прихованих закономірностей. Попередні гіпотези щодо виду закономірностей тут не визначаються. Закономірність (law) – істотний і такий, що постійно повторюється взаємозв'язок, що визначає етапи і форми процесу становлення, розвитку різних явищ або процесів. Система ІАД на цій стадії визначає шаблони, для отримання яких у системах OLAP, наприклад, аналітику необхідно обдумувати і створювати множину запитів. Тут же аналітик звільняється від такої роботи – шаблони шукає за нього система. Особливо корисне застосування цього підходу в надвеликих базах даних, де вловити закономірність за допомогою створення запитів доволі складно, для цього вимагається перепробувати безліч різноманітних варіантів. Вільний пошук подано такими діями [6]:

- виявлення закономірностей умовної логіки (conditional logic);
- виявлення закономірностей асоціативної логіки (associations and affinities);
- виявлення трендів і коливань (trends and variations).

Описані дії у межах стадії вільного пошуку виконуються за допомогою:

- індукції правил умовної логіки (задачі класифікації і кластеризації, опис в компактній формі близьких або подібних груп об'єктів);
- індукції правил асоціативної логіки (задачі асоціації і послідовності і витягування за їх допомогою інформації);
- визначення трендів і коливань (початковий етап задачі прогнозування).

На стадії вільного пошуку також повинна здійснюватись валідація закономірностей, тобто перевірка їх достовірності на частини даних, які не брали участі у формуванні закономірностей.

Прогностичне моделювання (Predictive Modeling). Друга стадія ІАД – прогностичне моделювання – використовує результати роботи першої стадії. Тут знайдені закономірності використовуються безпосередньо для прогнозування. Прогностичне моделювання охоплює такі дії:

- прогнозування невідомих значень (outcome prediction);
- прогнозування розвитку процесів (forecasting).

У процесі прогностичного моделювання розв'язуються задачі класифікації і прогнозування.

Під час розв’язування задачі класифікації результати роботи першої стадії (індукції правил) використовуються для зарахування нового об'єкта з певною упевненістю до одного з відомих, наперед визначених класів на підставі відомих значень.

Під час розв’язування задачі прогнозування результати першої стадії (визначення тренду або коливань) використовуються для прогнозу невідомих (пропущених або ж майбутніх) значень цільової змінної (змінних).

Закономірності, отримані на цій стадії, формуються від часткового до загального. У результаті ми одержуємо деяке загальне знання про деякий клас об'єктів на підставі дослідження окремих представників цього класу.

Прогностичне моделювання, навпаки, дедуктивне. Закономірності, отримані на цій стадії, формуються від загального до часткового. Тут ми одержуємо нове знання про деякий об'єкт або ж групу об'єктів на підставі:

- знання класу, до якого належать досліджувані об'єкти;
- знання загального правила, що діє в межах цього класу об'єктів.

Аналіз виключень (forensic analysis). На третій стадії ІАД аналізуються виключення або аномалії, виявлені у знайдених закономірностях. Дія, що виконується на цій стадії, – виявлення відхилень (deviation detection). Для виявлення відхилень необхідно визначити норму, яка розраховується на стадії вільного пошуку. Стадія аналізу виключень може бути використана як очищення даних [4].

Постановка завдання

На основі аналізу літературних джерел зробити висновки щодо особливостей, перспектив використання та можливостей розвитку інтелектуального аналізу даних у сучасних умовах розвитку комп’ютених технологій.

Результати досліджень

Сфера застосування ІАД нічим не обмежена – вона скрізь, де є якісь дані. Але насамперед методи ІАД сьогодні зацікавили комерційні підприємства, що розгортають свої проекти на основі інформаційних сховищ даних (Data Warehousing). ІАД являють собою велику цінність для керівників і аналітиків у їх повсякденній діяльності. Ділові люди усвідомили, що за допомогою методів ІАД вони можуть одержати відчутні переваги у конкурентній боротьбі. Досвід багатьох підприємств показує, що віддача від використання ІАД може сягати 1000 %. Наприклад, відомі повідомлення про економічний ефект, що в 10–70 разів перевищив первісні витрати від 350 до 750 тис. дол. Відома інформація про проект у 20 млн. дол., що окупився усього за 4 місяці. Інший приклад – річна економія 700 тис. дол. за рахунок впровадження ІАД в мережі універсамів Великобританії. Нижче розглянуто сучасні системи, в основу яких покладений ІАД.

Предметно-предметно-орієнтовані аналітичні системи

Предметно-предметно-орієнтовані аналітичні системи дуже різноманітні. Найширший підклас таких систем, що одержав поширення у галузі дослідження фінансових ринків, називається “технічний аналіз”. Він являє собою сукупність кількох десятків методів прогнозу динаміки цін і вибору оптимальної структури інвестиційного портфеля, що ґрунтуються на різних емпіричних моделях динаміки ринку. Ці методи часто використовують нескладний статистичний апарат, але максимально враховують сформованоу своєю областю специфіку (професійна мова, системи різних індексів тощо). На ринку є багато програм цього класу. Як правило, вони доволі дешеві.

Статистичні пакети

Останні версії майже усіх відомих статистичних пакетів включають поряд із традиційними статистичними методами також елементи ІАД. Але основна увага в них приділяється все таки класичним методикам – кореляційному, регресійному, факторному аналізу тощо. Доволі свіжий детальний огляд пакетів для статистичного аналізу наведений на сторінках ЦЕМІ http://is1.cemi.rssi.ru/ruswin/index.htm. Недоліком систем цього класу вважається вимога до спеціальної підготовки користувача. Також відзначають, що потужні сучасні статистичні пакети є занадто "великоваговими" для масового застосування у фінансах і бізнесі. До того ж часто ці системи доволі дорогі.

Є ще серйозніший принциповий недолік статистичних пакетів, що обмежує їх застосування в ІАД. Більшість методів, що належать до складу пакетів, спираються на статистичну парадигму, у якій головними фігурантами слугують усереднені характеристики вибірки. А ці характеристики, як вказувалося вище, під час дослідження реальних складних життєвих феноменів часто є фіктивними величинами.

Як приклади найпотужніших і найрозповсюдженіших статистичних пакетів можна назвати SAS (компанія SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA та ін.

Нейронні мережі

Це великий клас систем, архітектура яких має аналогію (як тепер відомо, доволі слабку) з побудовою нервової тканини з нейронів. В одній з найпоширеніших архітектур – багатошаровому перцептроні зі зворотним поширенням помилки – імітується робота нейронів у складі ієрархічної мережі, де кожен нейрон вищого рівня з'єднаний своїми входами з виходами нейронів нижчого шару. На нейрони найнижчого шару подаються значення вхідних параметрів, на основі яких потрібно приймати якісь рішення, прогнозувати розвиток ситуації тощо. Ці значення розглядаються як сигнали, що передаються у наступний шар, послаблюючись або підсилюючись, залежно від числових значень (ваг), що приписуються міжнейроним зв'язкам. У результаті на виході нейрона найвищого верхнього шару виробляється деяке значення, яке розглядається як відповідь – реакція усієї мережі на введені значення вхідних параметрів. Для того, щоб мережу можна було застосовувати й далі, її колись треба "натренувати" на отриманих раніше даних, для яких відомі й значення вхідних параметрів, і правильні відповіді на них. Тренування полягає у підборі ваг міжнейроних зв'язків, відповідей, що забезпечують найбільшу близькість мережі до відомих правильних відповідей.

Основним недоліком нейромережевої парадигми є необхідність у дуже великому обсязі навчальної вибірки. Інший істотний недолік полягає у тому, що навіть натренована нейронна мережа являє собою "чорну скриньку". Знання, що зафіксовані як ваги кількох сотень міжнейронних зв'язків, зовсім не піддаються аналізу й інтерпретації людиною (відомі спроби дати інтерпретацію структурі налаштованої нейромережі виглядають непереконливими – система " KINOsuite-PR").

Приклади нейромережевих систем – BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic). Вартість їх доволі висока.

Системи міркувань на основі аналогічних випадків

Ідея систем міркувань на основі аналогічних випадків (case based reasoning, CBR) на перший погляд украй проста. Для того, щоб зробити прогноз на майбутнє або вибрати правильне рішення, ці системи знаходять у минулому близькі аналоги наявної ситуації й вибирають ту саму відповідь, що була для них правильною. Тому цей метод ще називають методом "найближчого сусіда" (nearest neighbour). Останнім часом поширення одержав також термін memory based reasoning, що акцентує увагу на тому, що рішення приймається на підставі усієї інформації, накопиченої у пам'яті.

Системи CBR показують непогані результати у найрізноманітніших задачах. Головним їх мінусом є те, що вони взагалі не створюють якихось моделей або правил, що узагальнюють попередній досвід, – у виборі рішення вони ґрунтуються на усьому масиві доступних історичних даних, тому неможливо сказати, на основі яких конкретно чинників CBR системи будують свої відповіді.

Інший мінус полягає у сваволі, яку допускають системи CBR під час вибору міри "близькості". Від цієї міри у найвирішальніший спосіб залежить обсяг множини 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 найпопулярніших підходів до розв'язання задач ІАД. Вони створюють ієрархічну структуру правил типу "ЯКЩО... ТО..." (if-then), що має вид дерева. Для ухвалення рішення, до якого класу зарахувати 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 негативна, – то до лівого вузла; потім знову треба поставити питання, пов'язане з відповідним вузлом.

Популярність підходу пов'язана ніби з наочністю та зрозумілістю. Але дерева рішень принципово нездатні знаходити "кращі" (найповніші й найточніші) правила в даних. Вони реалізують наївний принцип послідовного перегляду ознак і "зачіпають" фактично осколки справжніх закономірностей, створюючи лише ілюзію логічного висновку.

До того ж більшість систем використовують саме цей метод. Найвідомішими є See5/35.0 (RuleQuest, Австралія), Clementine (Integral Solutions, Великобританія), SIPINA (University of Lyon, Франція), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада). Вартість цих систем варіюється від дешевих до доволі дорогих.

Еволюційне програмування

Проілюструємо сучасний стан цього підходу на прикладі системи PolyAnalyst – російської розробки, що одержала сьогодні загальне визнання на ринку ІАД. У цій системі гіпотези про вид залежності цільової змінної від інших змінних формулюються у вигляді програм на деякій внутрішній мові програмування. Процес побудови програм будується як еволюція у світі програм (цим підхід дуже подібний до генетичних алгоритмів). Коли система знаходить програму, що більш-менш задовільно виражає шукану залежність, вона починає вносити в неї невеликі модифікації й відбирає серед побудованих дочірніх програм ті, які підвищують її точність. У такий спосіб система "вирощує" кілька генетичних ліній програм, які конкурують між собою в точності вираження шуканої залежності. Спеціальний модуль системи PolyAnalyst переводить знайдені залежності із внутрішньої мови системи на зрозумілу користувачеві мову (математичні формули, таблиці тощо).

Інший напрямок еволюційного програмування пов'язане з пошуком залежності цільових змінних від інших у формі функцій якогось певного виду. Наприклад, в одному з найудаліших алгоритмів цього типу – методі групового урахування аргументів (МГУА) – залежність шукають у формі поліномів. У цей час із систем, що продаються в Росії, МГУА реалізована у системі NeuroShell компанії Ward Systems Group. Вартість систем варіюється до середніх цінових меж.

Генетичні алгоритми

ІАД не основна область застосування генетичних алгоритмів. Їх потрібно розглядати скоріше як потужний засіб розв'язання різноманітних комбінаторних задач і задач оптимізації. Проте генетичні алгоритми увійшли сьогодні до стандартного інструментарію методів ІАД, тому вони й включені для розгляду.

Перший крок під час побудови генетичних алгоритмів – це кодування вихідних логічних закономірностей у базі даних, які називаються хромосомами, а увесь набір таких закономірностей називають популяцією хромосом. Далі для реалізації концепції відбору вводиться спосіб зіставлення різних хромосом. Популяція обробляється за допомогою процедур репродукції, мінливості (мутацій), генетичної композиції. Ці процедури імітують біологічні процеси. Найважливіші серед них: випадкові мутації даних в індивідуальних хромосомах, переходи (кросинговер) і рекомбінація генетичного матеріалу, що міститься в індивідуальних батьківських хромосомах, і міграції генів.

У ході роботи процедур на кожній стадії еволюції виходять популяції з усе досконалішими індивідуумами.

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

Генетичні алгоритми мають багато недоліків. Критерій відбору хромосом і використовуваних процедур є евристичними й далеко не гарантують знаходження "кращого" рішення. Як і в реальному житті, еволюцію може "заклинити" на будь-якій непродуктивній галузі. І, навпаки, можна навести приклади, як два безперспективні батьки, які будуть виключені з еволюції генетичним алгоритмом, виявляються здатними зробити високоефективного нащадка. Це особливо стає помітно під час розв'язання високорозмірних задач зі складними внутрішніми зв'язками.

Прикладом може бути система GeneHunter фірми Ward Systems Group. Її вартість зарахована до середньої цінової категорії.

Алгоритми обмеженого перебору

Алгоритми обмеженого перебору минулого запропоновані у середині 60-х років XX ст. М.М. Бонгардом для пошуку логічних закономірностей у даних. З того часу вони продемонстрували свою ефективність під час розв'язання множини задач із усіляких областей.

Ці алгоритми обчислюють частоти комбінацій простих логічних подій у підгрупах даних. Приклади простих логічних подій: X = a; X < a; X a; a < X < b та ін., де X – параметр, "a" і "b" – константи. Обмеженням слугує довжина комбінації простих логічних подій (у М. Бонгарда вона дорівнює 3). На підставі аналізу обчислених частот робиться висновок про корисність тієї чи іншої комбінації для встановлення асоціації у даних, для класифікації, прогнозування тощо.

Найяскравішим сучасним представником цього підходу є система WizWhy підприємства WizSoft. Хоча автор системи Абрахам Мейдан не розкриває специфіку алгоритму, покладеного в основу роботи WizWhy, за результатами ретельного тестування системи були зроблені висновки про наявність тут обмеженого перебору (вивчалися результати, залежності часу їхнього одержання від кількості аналізованих параметрів тощо).

Автор WizWhy стверджує, що його система виявляє усі логічні if-then правила в даних. Насправді це не так. По-перше, максимальна довжина комбінації в if-then-правилі в системі WizWhy дорівнює 6, і, по-друге, із самого початку роботи алгоритму виробляється евристичний пошук простих логічних подій, на яких потім будується увесь подальший аналіз. Зрозумівши ці особливості WizWhy, не важко було запропонувати найпростішу тестову задачу, яку система не змогла взагалі розв'язати. Інший момент – система видає розв'язок за прийнятний час, тільки для порівняно невеликої розмірності даних.

Проте система WizWhy є сьогодні одним з лідерів на ринку продуктів ІАД. Це не позбавлено підстав. Система постійно демонструє вищі показники під час розв'язання практичних задач, ніж усі інші алгоритми. Вартість системи трохи більша за середню.

Системи для візуалізації багатовимірних даних

Тією чи іншою мірою засіб для графічного зображення даних підтримується усіма системами SFL. До того ж доволі значну частку ринку займають системи, що спеціалізуються винятково на цій функції. Прикладом тут може слугувати програма DataMiner 3D словацької фірми "Dimension 5" (5-й вимір).

У подібних системах основна увага сконцентрована на толерантності користувацького інтерфейсу, що дає змогу асоціювати з аналізованими показниками різні параметри діаграми розсіювання об'єктів (записів) бази даних. До таких параметрів належить колір, форма, орієнтація щодо власної осі, розміри та інші властивості графічних елементів зображення. Крім того, системи візуалізації даних позначені зручними засобами для масштабування й обертання зображень. Вартість систем візуалізації може сягати кілька сотень доларів.

Вище були проаналізовані існуючі системи ІАД. А тепер коротко охарактеризуємо можливі напрямки використання ІАД.

Роздрібна торгівля

Підприємства роздрібної торгівлі сьогодні збирають докладну інформацію про кожну окрему закупівлю, використовуючи кредитні картки з маркою магазину та автоматизовану систему контролю. Ось типові завдання, які можна виконувати за допомогою ІАД у сфері роздрібної торгівлі:

- аналіз купівельного кошика (аналіз подібності) призначений для виявлення товарів, які покупці прагнуть купувати разом. Знання купівельного кошика необхідне для поліпшення реклами, вироблення стратегії створення запасів товарів і способів їх розкладки у торговельних залах;
- дослідження часових шаблонів допомагає торговельним підприємствам ухвалювати рішення щодо створення товарних запасів. Воно дає відповіді на питання: “Якщо сьогодні покупець придбав відеокамеру, то через якись час він найімовірніше купить нові батарейки до неї та плівку?";
- створення прогнозних моделей дає можливість торговельним підприємствам дізнаватися про характер потреб різних категорій клієнтів з визначеною поведінкою, наприклад, про купівлю товарів відомих дизайнерів або про відвідування розпродажів. Ці знання потрібні для розроблення точно спрямованих, економічних заходів щодо просування товарів.

Банківська справа

Досягнення технології ІАД використовуються у банківській справі для виконання таких завдань:

- виявлення шахрайства із кредитними картками. Аналізуючи минулі транзакції, які згодом виявилися шахрайськими, банк виявляє деякі стереотипи такого шахрайства;
- сегментація клієнтів. Розбиваючи клієнтів на різні категорії, банки роблять свою маркетингову політику цілеспрямованішою і результативнішою, пропонуючи різні види послуг різним групам клієнтів;
- прогнозування змін клієнтури. ІАД допомагає банкам будувати прогнозні моделі цінності своїх клієнтів і у відповідний спосіб обслуговувати кожну категорію покупців.

Телекомунікації

В області телекомунікацій методи ІАД допомагають компаніям енергійніше просувати свої програми маркетингу й ціноутворення, щоб утримувати існуючих клієнтів і залучати нових. Серед типових заходів відзначимо такі:

- аналіз записів про докладні характеристики викликів. Призначення такого аналізу – виявлення категорій клієнтів з подібними стереотипами користування їх послугами та розроблення привабливих наборів цін і послуг;
- виявлення лояльності клієнтів. ІАД можна використовувати для визначення характеристик клієнтів, які одого разу скориставшись послугами цієї компанії, з великою часткою ймовірності залишаться їй вірними. У підсумку засоби, які виділяються на маркетинг, можна витрачати там, де віддача буде найбільша.

Страхування

Страхові компанії протягом багатьох років накопичують великі обсяги даних. Тут велике поле для спрацьовування методів ІАД:

- виявлення шахрайства. Страхові компанії можуть знизити рівень шахрайства, відшукуючи певні стереотипи у заявах про виплату страхового відшкодування, що характеризують взаємини між юристами, лікарями та заявниками;
- аналіз ризику. Шляхом виявлення сполучень чинників, пов'язаних з оплаченими заявами, страховики можуть зменшити свої втрати по зобов'язаннях. Відомий випадок, коли у США велика страхова компанія виявила, що суми, виплачені за заявами людей одружених, удвічі перевищує суми по заявах самотніх людей. Компанія відреагувала на це нове знання переглядом своєї загальної політики надання знижок сімейним клієнтам.

Інші додатки

ІАД може застосовуватися у багатьох галузях діяльності, таких як:

- розвиток автомобільної промисловості. При складанні автомобілів виробники повинні враховувати вимоги кожного окремого клієнта, тому їм потрібна можливість прогнозування популярності певних характеристик і знання того, які характеристики переважно замовляються разом;
- політика гарантій. Виробникам потрібно вираховувати кількість клієнтів, які подадуть гарантійні заявки, і середню вартість заявок;
- заохочення клієнтів, що часто літають літаками. Авіакомпанії можуть виявити групу клієнтів, яких заохочувальними заходами можна спонукати літати більше. Наприклад, одна авіакомпанія виявила категорію клієнтів, які літали на короткі відстані, не накопичуючи достатньої суми відстаней для вступу до їхніх клубів, тому вона в такий спосіб змінила правила прийому до клубу, щоб заохочувати кількість польотів, так само, як і суму відстаней.

Медицина

Відомо багато експертних систем для постановки медичних діагнозів. Вони побудовані переважно на основі правил, що описують сполучення різних симптомів різних захворювань. За допомогою таких правил розпізнають не тільки те, чим хворіє пацієнт, але і як потрібно його лікувати. Правила допомагають вибирати засоби медикаментозного впливу, визначати показання – протипоказання, орієнтуватися у лікувальних процедурах, створювати умови найефективнішого лікування, передбачати результати призначеного курсу лікування тощо. Технології ІАД дають можливість виявляти ці шаблони, складові зазначених правил.

Молекулярна генетика та генна інженерія

Мабуть, найгостріше поставлено завдання виявлення закономірностей в експериментальних даних у молекулярній генетиці та генній інженерії. Тут воно формулюється як визначення маркерів, під якими розуміють генетичні коди, що контролюють ті чи інші фенотипічні ознаки живого організму. Такі коди можуть містити сотні, тисячі й більше пов'язаних елементів.

На розвиток генетичних досліджень виділяються великі кошти. Останнім часом у цій галузі виник особливий інтерес до застосування методів ІАД. Відомо кілька великих фірм, що спеціалізуються на застосуванні цих методів для розшифрування генома людини й рослин.

Прикладна хімія

Методи ІАД широко застосовуються у прикладній хімії (органічній і неорганічній). Тут нерідко виникає питання про з'ясуванні особливостей хімічної будови тих чи інших з'єднань, їхніх визначальних властивостей. Особливо актуальним таке завдання виявляється при аналізуванні складних хімічних сполук, опис яких включає сотні й тисячі структурних елементів та їхніх зв'язків.

Можна навести ще багато прикладів, де методи ІАД відіграють велику роль. Особливість цього полягає в їх складній системній організації. Вони належать переважно до надкібернетичного рівня організації систем, закономірності якого не можуть бути точно описані мовою статистичних чи інших аналітичних математичних моделей. Дані у зазначених областях неоднорідні, гетерогенні, нестаціонарні й часто відрізняються високою розмірністю.

На основі розглянутого вище матеріалу можна стверджувати, що потенціал ІАД дає поштовх до розширення меж застосування цієї технології в сучасному світі комп’ютерних технологій. Щодо перспектив ІАД можливі такі напрямки розвитку:

- виділення типів предметних галузей з відповідними їм евристиками, формалізація яких полегшить розв’язання відповідних задач ІАД, що належать до цих галузей;
- створення формальних мов і логічних засобів, за допомогою яких будуть формалізовані міркування і автоматизація яких стане інструментом розв’язання задач ІАД у конкретних предметних галузях;
- створення методів ІАД, здатних не тільки витягувати з даних закономірності, але й формувати деякі теорії, що спираються на емпіричні дані;
- подолання істотного відставання можливостей інструментальних засобів ІАД від теоретичних досягнень у цій області.

Основна особливість ІАД полягає в поєднанні широкого математичного інструментарію (від класичного статистичного аналізу до нових кібернетичних методів) і останніх досягнень у сфері інформаційних технологій. У технології ІАД гармонійно поєднані чітко формалізовані методи і методи неформального аналізу, тобто кількісний і якісний аналіз даних.

Більшість аналітичних методів, що використовуються в ІАД, – це відомі математичні алгоритми і методи. Новим в їх застосуванні є можливість їх використання під час вирішення тих чи інших конкретних проблем, зумовлена новими технічними і програмними засобами, що з'явилися.

Висновки

Системи інтелектуального аналізу даних застосовуються як масовий продукт для бізнес-додатків і як інструменти для проведення унікальних досліджень (генетика, хімія, медицина тощо). Лідери ІАД пов'язують майбутнє цих систем з використанням їх як інтелектуальних додатків, вбудованих у корпоративні сховища даних.

Незважаючи на достатню кількість методів ІАД, пріоритет поступово зміщується у бік логічних алгоритмів пошуку в даних причинно-наслідкових правил. За їх допомогою розв’язуються задачі прогнозування, класифікації, розпізнавання образів, сегментації БД, здобування з даних “схованих” знань, інтерпретації даних, установлення асоціацій в БД тощо. Результати таких алгоритмів ефективні й легко інтерпретуються.

До того ж головною проблемою логічних методів виявлення закономірностей є проблема перебору варіантів за прийнятний час. Відомі методи або штучно обмежують такий перебір (алгоритми KOPA, WizWhy), або будують дерева рішень (алгоритми CART, CHAID, ID3, See5, Sipina та ін.), що мають принципові обмеження ефективності пошуку причинно-наслідкових правил. Інші проблеми пов'язані з тим, що відомі методи пошуку логічних правил не підтримують функцію узагальнення вищерозглянутих правил і функцію пошуку оптимальної композиції таких правил. Вдале рішення зазначених проблем може бути покладене в основу нових методик ІАД та відповідних розробок.

1. Чубукова И.А. Data Mining: учебн. пособ. – М.: Интернет-университет информационных технологий БИНОМ: Лаборатория знаний, 2006. – 382 с. 2. Дюк В. Data Mining: учеб. курс (+CD)/.Дюк В., Самойленко А. – СПб.: Изд-во Питер, 2001. – 368 с. 3. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? – Tandem Computers Inc., 1996 – 253 s. 4. Кречетов Н. Продукты для интеллектуального анализа данных // Рынок программных средств, N14-15_97. – 1997. – С. 32–39. 5. Киселев М. Средства добычи знаний в бизнесе и финансах / М. Киселев, Е. Соломатин // Открытые системы. – 1997. – № 4. – С. 41–44. 6. Data Mining and Image Processing Toolkits. – [Електронний ресурс]. – Режим доступу http://datamining.itsc.uah.edu/adam/. 7. Барсегян Ф. Методы и модели анализа данных OLAP и DataMining / Ф. Барсегян, М. Куприянов, В. Степанеенко, И. Холод. – СПб.: БХВ-Петербург, 2008. – 354 с.