ТЕМА №4 АСОЦІАТИВНІ ПРАВИЛА ТА ДЕРЕВА РІШЕНЬ

План

4.1. Основні поняття теорії асоціативних правил.
4.2. Програмні засоби пошуку асоціативних правил
4.3. Практичний аспект застосування технології асоціативних правил
4.4. Дерева рішень – загальні принципи технології
4.5. Комп'ютерні системи та напрямки застосування дерев рішень

Література:

Основна

Черняк О. І., Захарченко П. В. Інтелектуальний аналіз даних: Підручник. Київ, 2014.

Додаткова

1. Ананий В. Левитин Е. Алгоритмы: введение в разработку и анализ. Москва :Вильямс, 2006.
2. Дюк В. А., Самойленко А. П. Data Mining: учебный курс. Санкт-Петербург : Питер, 2001.
3. МакКоннелл Дж. Основы современных алгоритмов. Москва : Техносфера, 2004.
4. Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних(дейтамайнінг). Київ : КНЕУ, 2007.
5. Черняк О. І., Ставицький А. В., Чорноус Г. О. Системи обробки економічної інформації: Підручник. Київ : Знання, 2006.
6. Чубукова И. А. Data Mining. Москва : БИНОМ, 2008.

4.1. Основні поняття теорії асоціативних правил

Пошук асоціативних правил (*Association Rules*) – ключова тема в інтелектуальному аналізі даних. Асоціація має місце в тому випадку, якщо декілька подій зв'язані одна з одною. Наприклад, дослідження, проведене в супермаркеті, може показати, що 65% відвідувачів, які купили кукурудзяні чіпси беруть також і «кока-колу», а за наявності знижки за такий комплект купують його в 85% випадків. Маючи в своєму розпорядженні відомості про подібну асоціацію, менеджерам легко оцінити, наскільки дієва знижка, що надається.

Пошук виявляє приховані зв'язки в, на перший погляд, ніяк незв'язаних даних. Ці зв'язки – правила. Ті, які перевищують певний поріг, вважаються цікавими. Такі правила дають можливість виконувати дії ґрунтуючись на певних шаблонах. Вони так само допомагають в прийнятті і поясненні рішень. Як і більшість методів data mining, даний метод дозволяє перетворити потенційно величезну кількість інформації в невеликий і зрозумілий набір статистичних показників.

Пошук асоціативних правил зовсім не тривіальна задача, як може здатися на перший погляд. Одна з проблем – алгоритмічна складність при знаходженні часто зустрічаючих наборів елементів, оскільки із зростанням числа елементів в експоненціально зростає число потенційних наборів елементів.

Існують різні типи асоціативних правил. У простій формі асоціативні правила повідомляють лише про наявність або відсутність асоціації. Логічна природа таких правил озвучена в їх назві – булеві асоціативні правила (*Boolean Association Rule*). На прикладі корзини споживача це відбувається так, «споживачі, які купують зняте молоко так само купують масло з низьким рівнем жиру», – типове булеве асоціативне правило.

Правила, які збирають декілька асоціативних правил разом, називаються ***мультирівневі*** або ***узагальнені асоціативні правила*** (*Multilevel or Generalized Association Rules*). При побудові таких правил елементи зазвичай групуються згідно ієрархії і пошук ведеться на найвищому концептуальному рівні.

Складнішим типом правил є ***кількісні асоціативні правила*** (*Quantitative Association Rules*). Цей тип правил шукається із застосуванням кількісних (наприклад, ціна) або категоріальних (наприклад, стать) атрибутів. Наприклад, «покупці, чий вік знаходиться між 30 і 35 роками з доходом більше 75000 в рік купують машини вартістю більше 20000».

Окрім описаних вище асоціативних правил існують ***непрямі асоціативні правила, асоціативні правила із запереченням*** та інші.

Не дивлячись на різні типи правил, алгоритм для пошуку асоціативних правил може бути в загальному вигляді розділений на два етапи:

1. пошук найбільш часто зустрічающихся наборів елементів (*large (frequent) itemsets*). Набір, що часто зустрічається, – це набір, в якого підтримка перевищує мінімальне значення;
2. генерація правил на основі часто зустрічающихся наборів.

4.2. Програмні засоби пошуку асоціативних правил

За допомогою алгоритмів виявлення асоціативних правил можна вирішувати чималий спектр практичних завдань. Саме тому ринок програмних продуктів, що реалізовують ці технології, досить представницький та різнорідний. Практично кожна відома компанія в тому або іншому вигляді використовує технології пошуку асоціативних правил в своїх програмних продуктах. Зупинимося на деяких програмних рішеннях, що отримали найбільшу популярність .

Пакет Deductor. Розглянемо приклад рішення задачі пошуку асоціативних правил. Вважатимемо, що існує транзакційна база даних. Необхідно знайти набори товарів, що зустрічаються найбільш часто, і набір асоціативних правил з певними границями значень підтримки і довіри.

Процес побудови асоціативних правил виконаємо в аналітичному пакеті Deductor. Транзакційна база даних, яка містить в кожній записі номер чека і товар, придбаний по цьому чеку, має формат *MS Excel*. Спершу імпортуємо дані з файлу *MS Excel* в середовище *Deductor*. Для номера транзакції (зазвичай в базі даних – це поле «номер чека») вказуємо тип «ідентифікатор транзакції (ID)», а для найменувань товару - тип «елемент».

Система PolyAnalyst. Система *PolyAnalyst* призначена для автоматичного аналізу числових і текстових даних з метою виявлення в них раніше невідомих, нетривіальних, практично корисних і доступних розумінню закономірностей, необхідних для ухвалення оптимальних рішень в бізнесі і в інших областях людської діяльності. В даний час вона є однією з найпотужніших систем *Data Mining* в світі, реалізованих для Intel платформ і операційних систем *Microsoft Window*s. Аналогічні системи *Data Mining* таких провідних виробників, як *IBM (Intelligent Miner, Data Miner), Silicon Graphics (SGI Miner), Integral Solutions (Clementine), SAS Institute (SAS)* працюють на середніх і великих машинах і коштують десятки і навіть сотні тисяч доларів. Завдяки унікальній технології «Еволюційного програмування» та іншим інноваційним математичним алгоритмам, *PolyAnalyst* поєднує в собі високу продуктивність «великих систем» з низькою вартістю, властивою програмам для *Windows*.

Клієнт Data Mining для Excel. Клієнт *Data Mining* для *Excel* дозволить провести повний цикл інтелектуального аналізу даних за допомогою клієнта *Excel* з використанням даних електронних таблиць або зовнішнього джерела, доступного базі даних *Analysis Services*, зокрема, пошук асоціативних правил.

Існують також інші програмні засоби для пошуку асоціативних правил, серед яких слід виділити комерційне та вільно поширюване програмне забезпечення. Зокрема, серед комерційного програмного забезпечення найбільш популярними є:

Azmy SuperQuery – пошукова система асоціативних правил;

Clementine, набір від SPSS, що включає аналіз ринкової корзини;

IBM Intelligent Miner for Data;

The LPA Data Mining Toolkit – підтримує пошук асоціативних правил в реляційних базах даних;

Magnum Opus є швидким інструментом пошуку асоціативних правил в даних, підтримується операційними системами *Windows, Linux і Solaris*;

Nuggets – це набір, що включає пошук асоціативних правил і інші алгоритми;

Purple Insight MineSet є набором візуального *Data Mining*, що включає візуалізатор асоціативних правил.

Серед вільно поширюваного програмного забезпечення слід виділити:

Apriori – інструмент для знаходження асоціативних правил за допомогою алгоритму Apriori;

ARtool – інструмент, що містить набір алгоритмів для пошуку асоціативних правил в бінарних базах даних (binary databases);

DM-II system – інструмент включає алгоритм *CBA* для виконання класифікації на основі асоціативних правил і деяких інших характеристик;

FIMI, Frequent Itemset Mining Implementations – є репозиторієм, що включає програмне забезпечення і бази даних.

4.3. Практичний аспект застосування технології асоціативних правил

Практичні додатки систем інтелектуального аналізу даних на основі технологій асоціативних правил воістину безмежні: виробництво, торгівля, фінанси, медицина, соціологія, наукові дослідження, освіта та ін. Такий інструментарій використовується для задач технічної і медичної діагностики, проектування, управління процесами, контролю якості, прогнозування, оцінки кредитоспроможності, аналізу стану ринків, маркетингових досліджень, роботи з клієнтами, соціологічних опитів, моделювання і вивчення складних систем на основі історії їх еволюції.

У великих масивах корпоративних даних часто зберігаються відповіді на багато питань, які цікавлять керівництво і співробітників організацій. Розглянемо деякі з них.

Маркетингові задачі. Одним з прикладів аналізу механізмів стимулювання продажів на основі великих масивів накопичених даних про поведінку споживачів виступають узагальнені асоціативні правила. Ставиться завдання знайти приховані закономірності і типові шаблони поведінки покупців.

Задачі митниці. Технологія асоціативних правил може успішно застосовуватися для виявлення прихованих тенденцій в зовнішньоторговельній діяльності. Одна з основних проблем, що стоїть перед митними органами, полягає у виявленні навмисного спотворення вантажних митних декларацій. Через обмежені ресурси повна перевірка всіх переміщуваних через кордон вантажів неможлива. Проте митниця збирає детальні бази даних по вантажних митних деклараціях. Аналіз цих даних може бути використаний для виявлення тенденцій в зовнішній торгівлі України по групах товарів, найбільш схильних до фальсифікації при проходженні митниці – «товарів ризику». Маючи дані про такі товари, митні пости могли б ретельніше перевіряти проходження відповідних вантажів і зменшити втрати від фальсифікації митних документів. Однією з особливостей задачі стала відсутність «тренувального» набору даних – даних, для яких було б апріорі відомо, які з них є спробою фальсифікації вантажної митної декларації, а які є сумлінно задекларованими товарами.

Задачі медицини. Сучасний рівень розвитку медицини характеризується тим, що при здійсненні практично будь-якого різновиду лікувального процесу збирається велика кількість супутньої інформації: результати аналізів та обстежень, протоколи оперативних втручань та оглядів. Особливостями цієї інформації є:

Велика кількість атрибутів (у тому числі числових).

Зашумленість. Наявність помилкових значень, викликаних ручним введенням даних або погрішностями апаратури.

Часткова заповненість. Наявність незаповнених полів в дослідженнях або відсутність цілих досліджень у деяких пацієнтів.

Погана структурованість. Вихідна інформація зберігається в різних типах СУБД, файлах, на паперових носіях і, як правило, складається з наборів моніторингів та аналізів різної структури.

Таким чином, актуальним завданням є розробка ефективних методів і алгоритмів виявлення залежностей у вигляді, доступному інтерпретації людині, між значеннями атрибутів даних, що володіють перерахованими властивостями, та реалізація вказаних методів і алгоритмів в програмних системах, що забезпечують всі фази обробки даних, включаючи: автоматизацію збору даних, первинну обробку і нормалізацію, інтерактивну взаємодію з експертом, візуалізацію результатів.

Інтернет - торгівля. Для ефективного управління бізнесом у сфері електронної комерції великого поширення набувають методи бізнес - аналітики. У сферу їх застосування входять задачі по прогнозуванню об'ємів продажів, управлінню кількістю товарних запасів, визначенню оптимальних торгівельних націнок, виявленню типових паттернів купівельної поведінки, оптимізації навігації по сайту, поліпшенню рубрикації і тому подібне.

4.4. Дерева рішень - загальні принципи технології

Стрімкий розвиток інформаційних технологій, зокрема, прогрес в методах збору, зберігання і обробки даних дозволив багатьом організаціям збирати величезні масиви даних, які необхідно аналізувати. Об'єми цих даних настільки великі, що можливостей експертів вже не вистачає, що породило попит на методи автоматичного дослідження (аналізу) даних, який з кожним роком постійно збільшується. Дерева рішень (*decision trees*) – один з таких методів автоматичного аналізу даних. Ця технологія є однією з найбільш популярних методів вирішення задач класифікації і прогнозування. Інколи цей метод інтелектуального аналізу даних також називають деревами рішаючих правил або деревами класифікації і регресії. Як видно з останньої назви, за допомогою даного методу вирішуються задачі класифікації і прогнозування. Якщо залежна, тобто цільова змінна набуває дискретних значень, при допомозі методу дерева рішень вирішується задача класифікації. Якщо ж залежна змінна набуває безперервних значень, то дерево рішень встановлює залежність цієї змінної від незалежних змінних, тобто вирішує задачу чисельного прогнозування.

Сфера застосування дерев рішень в даний час широка, але всі задачі, що вирішуються цією технологією можуть бути об'єднані в наступні три класи:

1. *Опис даних*: Дерева рішень дозволяють зберігати інформацію про дані в компактній формі. Замість самих даних ми можемо зберігати дерево рішень, яке містить точний опис об'єктів.

2. *Класифікація*: Дерева рішень відмінно справляються із задачами класифікації, тобто віднесення об'єктів до одного із заздалегідь відомих класів.

3. *Регресія*: Якщо цільова змінна має безперервні значення, дерева рішень дозволяють встановити залежність цільової змінної від незалежних (вхідних) змінних.

Дерева рішень є прекрасним інструментом в системах підтримки прийняття рішень та інтелектуального аналізу даних. В склад багатьох пакетів, призначених для інтелектуального аналізу даних, вже включені методи побудови дерев рішень. У областях, де висока ціна помилки, вони служать відмінною підмогою аналітика або керівника. Дерева рішень успішно застосовуються для вирішення практичних задач в наступних областях:

Банківська справа. Оцінка кредитоспроможності клієнтів банку при видачі кредитів.

Промисловість. Контроль за якістю продукції (виявлення дефектів), випробування без руйнувань (наприклад перевірка якості зварки) і так далі**.**

Медицина. Діагностика різних захворювань.

Молекулярна біологія. Аналіз будови амінокислот.

Автоматична класифікація тексту. Внутрішні вузли є термами, гілки, що відходять від них, характеризують вагу терма в аналізованому документі, а листя – категорії. Такий класифікатор категоризує випробовуваний документ, рекурсивно перевіряючи ваги вектора ознак по відношенню до порогів, виставлених для кожної з ваг, поки не досягне листа дерева (категорії). До цієї категорії (листа якого досяг класифікатор) і приписується аналізований документ.

Машинне навчання. Мета методів машинного навчання – здобуття простих класифікуючих виразів, які були б легко зрозумілі для людини. Достоїнством таких методів є те, що під час роботи того або іншого методу не потрібна участь людини.

Самонавчальні системи прийняття рішень. Сьогодні актуальною проблемою є створення систем, що самостійно навчаються для роботи на фінансових ринках. Здібна до самонавчання система автоматичного управління (самоналагоджувальна система), в якій пристосування до умов, які випадково змінюються, забезпечується автоматичною зміною параметрів налаштування або шляхом автоматичного пошуку оптимального налаштування. У будь-якій іншій автоматичній системі управління є параметри, які впливають на стійкість і якості процесів управління і можуть бути змінені при регуляції (налаштуванні) системи. Якщо ці параметри залишаються незмінними, а умови функціонування (характеристики

керованого об'єкту, збуджуючі дії) істотно змінюються, то процес управління може погіршати або навіть стати нестійким.

4.5. Комп’ютерні системи та напрямки застосування дерев рішень

Дерева рішень корисні для бізнес-користувачів, оскільки надають логічний результат, який можна обговорювати в бізнес-термінах. Вони просто створюються із застосуванням різних програмних рішень, і дуже ефективні і точні при забезпеченні хорошим набором даних. Практично всі відомі виробники комп'ютерних систем включають до складу своїх програмних продуктів алгоритми побудови і аналізу дерев рішень. Розглянемо деякі з них.

Скорингові системи. Підвищення прибутковості кредитного портфеля банку безпосередньо залежить від грамотного управління кредитними ризиками. І саме скорингові системи дозволяють понизити ризики без втрати прибутковості, запропонувавши відповідь на ключові питання: наскільки проблематичною буде робота банку з конкретним позичальником, яке значення кредитного ліміту встановити, і поверне клієнт кредит чи ні

Застосування технології дерев рішень для оцінки кредитоспроможності фізичних осіб на основі пакету Deductor. При кредитуванні фізичних осіб характерні невеликі розміри позик, що породжує великий об'єм роботи по їх оформленню і досить дорогу процедуру оцінки кредитоспроможності відносно отримуваного в результаті прибутку. Для оцінки кредитоспроможності фізичних осіб банку необхідно оцінити як фінансове положення позичальника, так і його особисті якості. При цьому кредитний ризик складається з ризику неповернення основної суми боргу і відсотків по цій сумі. Зараз для оцінки ризику кредитування позичальника використовується скоринг кредитування. Сутність цієї методики полягає в тому, що кожен чинник, що характеризує позичальника, має свою кількісну оцінку. Підсумовуючи отримані бали, можна отримати оцінку кредитоспроможності фізичної особи.

Програмний комплекс Oracle Data Miner. Використання дерева рішень – це спосіб класифікації існуючих даних, визначення чинників або правил, які мають відношення до цільового результату (*target result*), і їх вживання для прогнозування результату, що означає:

1. Бізнес-користувачі можуть визначати чинники, які найбільшою мірою впливають на рішення про покупки.
2. Департаменти маркетингу можуть «цілитися» в «правильні» групи потенційних клієнтів, виключаючи тих, хто з малою вірогідністю купуватиме.
3. Аналітики даних і фінансові аналітики можуть прогнозувати продажу завдяки аналізу атрибутів потенційних клієнтів, про яких є дані.
4. Бізнес-аналітики можуть коректувати цілі і стратегії при змінах тенденцій

5. Компанії можуть реорганізовувати підтримку (support, enhancements, and desupport) для забезпечення максимального задоволення клієнтів.

Система PolyAnalyst. У системі *PolyAnalyst*, реалізований алгоритм, заснований на критерії максимізації взаємної інформації (information gain). Тобто для розщеплювання вибирається незалежна змінна, що несе максимальну (у сенсі Шенона) інформацію про залежну змінну. Цей критерій на відміну від багатьох критеріїв, вживаних в інших системах *Data Mining*, має ясну інтерпретацію і дає розумні результати при найрізноманітніших статистичних параметрах даних, що вивчаються.

Інформаційна система «1С: Підприємство 8.0». Даний алгоритм набув найбільшого поширення при виявленні причинно-наслідкових зв'язків в даних і описі поведінкових моделей. Типова зона застосовності дерев рішень – оцінка різних ризиків, наприклад, закриття замовлення клієнтом або його переходу до конкурента, невчасного постачання товару постачальником або прострочення оплати товарного кредиту. Як типові вхідні чинники моделі виступають сума і склад замовлення, поточне сальдо взаєморозрахунків, кредитний ліміт, відсоток передоплати, умови постачання і інші параметри, що характеризують об'єкт прогнозу.