МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ
Факультет економіки та бізнес-адміністрування
Кафедра економічної кібернетики

Олешко Т.І., Квашук Д. М., О. М. Густера, Я. В. Крисак

КОНСПЕКТ ЛЕКЦІЙ
з дисципліни «Інструментальні засоби статистичного та інтелектуального аналізу даних»

за спеціальністю 051 «Економіка», освітньо-професійними програмами «Економічна кібернетика», «Цифрова економіка»

Укладачі:
професор кафедри економічної кібернетики
д.т.н., професор Олешко Т.І.
доцент кафедри економічної кібернетики
к.е.н Квашук Д. М.
асистент кафедри економічної кібернетики
к.е.н Густера О. М.
доцент кафедри економічної кібернетики
доцент НАУ Крисак Я. В.

Конспект лекцій розглянутий та схвалений
на засіданні кафедри
економічної кібернетики
Протокол № № __ від _____ р.

Завідувач кафедри_____Н. Іванченко

Рекомендовано Вченою радою ФЕБА (протокол №

Олешко Т.І., Квашук Д. М., О. М. Густера, Я. В. Крисак
Інструментальні засоби статистичного та інтелектуального аналізу даних.
Конспект лекцій. – К. :НАУ, 2020. – 80 с.

Містить основні положення з дисципліни «Інструментальні засоби статистичного та інтелектуального аналізу даних». Для студентів-магістрів спеціальності 051 «Економіка» спеціалізацій «Економічна кібернетика», «Цифрова економіка».

© Олешко Т.І. ,Квашук Д. М.,
О. М. Густера, Я. В. Крисак 2020 р.

Вступ

Метою викладання дисципліни є теоретична та практична підготовка студентів до вивчення систем обробки даних та принципів статистичного та інтелектуального аналізу даних на основі методів та алгоритмів Data Mining.

Завданнями вивчення навчальної дисципліни є: дослідження технологій зберігання та організації даних; оволодіння методами та алгоритмами Data Mining; дослідження процесів виявлення знань; дослідження принципів побудови сховищ даних.

У результаті вивчення даної навчальної дисципліни студент повинен:

Знати:

методи та технології статистичного та інтелектуального аналізу даних;

методи реалізації OLAP та Data Mining технологій.

Вміти:

самостійно застосовувати алгоритми Data Mining при обробці даних;

самостійно розробляти та будувати моделі сховищ даних;

самостійно проводити аналіз даних для виявлення знань;

самостійно використовувати OLAP-систему при обробці баз даних.

Навчальний матеріал дисципліни структурований за модульним принципом і складається з двох класичних навчальних модулів.

У результаті засвоєння навчального матеріалу навчального модуля №1 «Статистичний аналіз даних» студент повинен:

Знати:

методи збору даних;

методи первісної обробки даних;

методи кластерізаці та прогнозуваня.

Вміти:

самостійно організовувати сховище даних;

самостійно підготовлювати дані для їх аналізу;

самостійно застосовувати методи використання навчальної інформації.

У результаті засвоєння навчального матеріалу навчального модуля №2 «Інтелектуальний аналіз даних» студент повинен:

Знати:

структуру багатовимірної моделі даних;

методи та задачі Data Mining;

архітектуру OLAP-систем;

методи асоціативних правил.

Вміти:

самостійно застосовувати методи Data Mining;

самостійно використовувати OLAP-системи для обробки сховищ даних;

самостійно методи асоціативних правил.

Знання та вміння, отримані студентом під час вивчення даної навчальної дисципліни, використовуються для закріплення отриманих знань та є базою для написання магістерського дипломного проекту (роботи).

Модуль №1 «Статистичний аналіз даних»

Лекція № 1.1. Основи статистичного аналізу даних

Статистика — наука, яка вивчає методи кількісного охоплення і дослідження масових, зокрема суспільних, явищ і процесів. Збирання інформації про них сягає найдавніших часів. Вона мала спершу наскрізь практичний характер; з XIX ст. статистика поступово здобуває солідну наукову основу, коли почалося впорядкування і вдосконалення статистичних методів. З них розвинулися дві основні: описова (дескриптивна) — збирання інформації, перевірка її якості, її інтерпретація, зображення статистичного матеріалу; та індуктивна — застосування теорії ймовірності, закону великих чисел. Статистика поділяється за своїм змістом на демографічну, економічну, фінансову, соціальну, санітарну, судову, біологічну, технічну тощо;математична статистика вивчає математичні методи систематизації, обробки й використання статистичних даних для наукових і практичних висновків

Основні поняття (категорії) статистики

Статистична сукупність — це маса однорідних в певному відношенні елементів, які мають єдину якісну основу, але різняться між собою певними ознаками і підлягають певному закону розподілу. Статистична сукупність — це певна множина елементів, поєднана умовами існування і розвитку.

Однорідна сукупність — якщо одна чи декілька ознак, що вивчаються, є загальними для всіх одиниць.

Різнорідна сукупність об'єднує явища різного типу.

Одиниця сукупності — це первинний елемент статистичної сукупності, який є носієм ознак, що підлягають реєстрації і є основою обліку.

Ознака — властивість окремої одиниці сукупності.

Якісні ознаки (атрибутивні ознаки) виражаються в вигляді понять, визначень, які характеризують їхню суть, стан або якість. Наприклад, сорт продукції, професія, сімейний статус.

Кількісні ознаки виражають окремі значення якісних ознак у числовому виразі.

Дискретні — ознаки, виражені окремими цілими числами, без проміжних значень.

Неперервні — ознаки, що можуть набувати будь-яких значень у певних чисел.

Прямі — характеризують об'єкт дослідження безпосередньо (вік осіб, кількість присутніх в аудиторії).

Непрямі — ознаки, що не належать безпосередньо досліджуваному об'єкту (чи сукупності), а які належать іншій сукупності, що входить в дану.

Багатоваріантні — перш за все характеризуються рангами (шкалою рангів) від більшого до меншого (напр. дуже низький, низький, середній, високий, дуже високий).

Альтернативні — взаємовиключаючі значення: так-ні, позитивне-негативне.

Інтервальні — це ознаки, які характеризують результат процесів.

Моментні — характеризують об'єкт в певний момент часу.
Окремі значення кількісних ознак називаються варіантами.
Первинні варіанти характеризують одиницю сукупності в цілому: абсолютні значення, виміряні, розраховані.
Вторинні варіанти (похідні, розрахункові) — дані, що неможливо перевірити, оскільки вони взяті з певних джерел.
Адитивність — підсумовувати, складати.
Статистичні показники — це числа в сукупності з набором ознак, що характеризують обставини, до яких вони відносяться, що, де, коли, і яким чином підлягають вимірюванню. Статистичний показник — це кількісна характеристика соціально-економічних явищ і процесів в умовах якісної визначеності.
Статистичні дані — це сукупність показників, отриманих внаслідок статистичного спостереження або обробки даних.
Статистична закономірність — це закономірність, в якій необхідність пов'язана в кожному окремому явищі з випадковістю, і лише в сукупності явищ виявляє себе як закон.
Система статистичних показників — це сукупність статистичних показників, які відображають взаємозв'язки, які об'єктивно існують між явищами.
Вибірка
Вибірка — це множина об'єктів, подій, зразків або сукупність вимірів, за допомогою визначеної процедури вибраних з статистичної популяції або генеральної сукупності для участі в дослідженні. Зазвичай, розміри популяції дуже великі, що робить прийняття до уваги всіх членів популяції непрактичним або неможливим. Вибірка представляє собою множину або сукупність певного обсягу, члени якої збираються і статистичні характеристики обчислюється таким чином, що в результаті можна зробити висновки або екстраполяцію із вибірки на всю популяцію або генеральну сукупність.
Обсяг вибірки — число випадків, включених у вибіркову сукупність. Із статистичних міркувань рекомендується, щоб число випадків становило не менше 30—35.
Залежні і незалежні вибірки
При порівнянні двох (і більш) вибірок важливим параметром є їх залежність. Якщо можна ознаки), такі вибірки встановити гомоморфну пару (тобто, коли одному випадку з вибірки X відповідає один і лише один називаються залежними. Приклади залежних вибірок:
• пари близнят
• два вимірювання якої-небудь ознаки до і після експериментальної дії
• чоловіки і дружини
• тощо
У випадку, якщо такий взаємозв'язок між вибірками відсутній, то ці вибірки вважаються незалежними, наприклад:
• чоловіки і жінки
• психологи і математики.
Репрезентативність

Вибірка може розглядатися як репрезентативна або нерепрезентативна.
Довідка: РЕПРЕЗЕНТАТИВНИЙ (рос. репрезентативный, англ. representative, нім. repräsentativ) – представницький, характерний, типовий для чого-небудь. Напр., репрезентативна вибірка – сукупність випадкових чисел, в якій визначається множина елементів вибірки, що характеризує генеральну сукупність.
Якщо вибірка являє собою числову змінну, наприклад зріст або вік людей, тоді репрезентативність такої вибірки визначають залежно її наповненості і добротності.
Приклад нерепрезентативної вибірки
У США одним з найвідоміших історичних прикладів нерепрезентативної вибірки вважається випадок, що стався під час президентських виборів в 1936 році. Журнал «Літрері Дайджест», що успішно прогнозував події декількох попередніх виборів, помилився у своїх прогнозах, розіславши десять мільйонів пробних бюлетенів своїм підписчикам, людям, вибраним по телефонним книгах всієї країни, і людям з реєстраційних списків автомобілів. У 25 % бюлетенів (майже 2,5 мільйона) голосів, що повернулися, були розподілені таким чином:
• 57 % віддавали перевагу кандидату-республіканцю Альфу Лендону
• 40 % вибрали діючого на той час президента-демократа Франкліна Рузвельта
На дійсних же виборах, як відомо, переміг Рузвельт, набравши більше 60 % голосів. Помилка «Літрері Дайджест» полягала в наступному: бажаючи збільшити репрезентативність вибірки, — оскільки їм було відомо, що більшість їхніх передплатників вважають себе республіканцями, — вони розширили вибірку за рахунок людей, вибраних з телефонних книг і реєстраційних списків. Проте вони не врахували тогочасних реалій і насправді набрали ще більше республіканців: у часи Великої депресії володіти телефонами і автомобілями могли собі дозволити переважно представники середнього і верхнього класу (в більшості республіканці, а не демократи).

Види плану побудови груп з вибірок
Виділяють декілька основних видів плану побудови груп[2]:
1. Дослідження з експериментальною і контрольною групами, які ставляться в різні умови.
• Дослідження з експериментальною і контрольною групами із залученням стратегії попарного відбору
2. Дослідження з використанням тільки однієї групи — експериментальною.
3. Дослідження з використанням змішаного (чинника) плану — всі групи ставляться в різні умови.

Статистична сукупність
Під статистичною сукупністю розуміють масу однорідних у певному відношенні елементів (явищ, фактів і т. ін.), які мають єдину якісну основу, але різняться між собою за певними ознаками.

Під одноякісністю, або однорідністю, розуміють підпорядкованість елементів, що складають сукупність, загальному закону розвитку або їх закону розвитку

або їх однотиповість (наприклад, Інформація про сукупність одиниць господарств, малих підприємств; про сукупність одиниць виробленої ними продукції.
Статистична сукупність складається з окремих одиниць (наприклад, у конкретному малому підприємстві є зведення про вид, вантажопідйомність, кількість днів роботи, витрати на ремонт по кожному автомобілю). Такі окремі первинні елементи, або індивідуальні явища, які складають статистичну сукупність, називають одиницями сукупності.
Залежно від мети досліджень однорідність сукупності можна вивчати у різноманітних аспектах розвитку. Так, по молочному стаду корів у господарстві - це породний склад, продуктивність, класність, захворюваність тощо.
Повне уявлення про статистичні сукупності можна мати лише при досконалому вивченні їх ознак. У навчальній літературі найбільш вдало ці питання розкриті в навчальному посібнику І. П. Суслова1. Розглянемо це питання в запропонованій автором послідовності.
Статистичні сукупності у сфері суспільного життя можна поділити на дві групи:
1. Сукупності, створені самим життям, які утворюють єдність незалежно від того, чи підлягають вони вивченню статистикою (наприклад, вивчення у господарстві сукупності робітників за освітою, віком, участю у громадській роботі тощо);
2. Сукупності, утворені спеціально з метою статистичного аналізу (наприклад, сукупності підприємств за видами їх комерційної діяльності, за чисельністю в них кваліфікованих робітників, кількістю унікальних видів вироблюваної продукції і т. ін.).
Формування статистичної сукупності передбачає реалізацію одночасно діючих, протилежних один одному прийомів: об'єднання і роз'єднання елементів і частин статистичної сукупності.
Виникає запитання: навіщо потрібні такі операції у формуваннях сукупностей? Відповідь міститься у постановці і формулюванні таких завдань:
1. За становленими правилами па підставі локальних статистичних характеристик визначити загальні характеристики;
2. Виходячи із загальних статистичних характеристик сукупності, на підставі заданих критеріїв знайти локальні статистичні характеристики.
Це все свідчить про важливість категорії "статистична сукупність", адже особливості законів розвитку суспільних явищ вимагають статистичних методів пізнання досліджуваних сукупностей, а шлях цей досить складний і пролягає через методи, які розробляє статистична наука.

Ймовірність

Ймові́рність (лат. probabilitas, англ. probability) — числова характеристика можливості того, що випадкова подія відбудеться в умовах, які можуть бути відтворені необмежену кількість разів. Імовірність є основним поняттям розділу математики, що називається теорія імовірностей.

Випадковою подією називається подія, результат якої не може бути відомий наперед. Навіть у тому разі, коли насправді подія детермінована своїми передумовами, вплив цих передумов може бути настільки складним, що вивести з них наслідок логічно й послідовно, неможливо. Наприклад, при підкидуванні монети, сторона на яку монета впаде визначається положенням руки і монети в руці, швидкістю, обертовим моментом тощо, однак відслідкувати всі ці фактори неможливо, тому результат можна вважати випадковим.

Методи збору і підготовки вихідного набору даних

Найменування методу	характеристика методу	різновиди	Галузь застосування
Опитування	Метод збору первинної інформації за допомогою звернення з питаннями до певної групи людей (респондентам). Дозволяє виявити певні типові фактичні дані, а також поняття і судження з різних питань.	Письмові опитування (анкетування); Усні опитування (інтерв'ювання); Очні опитування; Заочні опитування (телефонні, поштові, опитування за допомогою Інтернету, через пресу); Інтерактивні опитування по телебаченню або по радіо; Вибіркові опитування; Суцільні опитування	Збір первинної інформації (наприклад, про використовувані забезпечують засобах, про користувачів і їх інформаційні потреби, про ринок інформаційних продуктів і послуг, посередників, про асортимент і якість інформаційних продуктів і послуг).
інтерв'ювання	Метод збору даних, що полягає в тому, що спеціально навчений інтерв'юер, як правило, в безпосередньом у контакті з респондентом усно задає питання, передбачені програмою дослідження. Є одним з	Індивідуальні інтерв'ю; Парні інтерв'ю; Групові інтерв'ю; Масові інтерв'ю; Глибинне інтерв'ю; Вільне інтерв'ю; Фокусування інтерв'ю; Стандартизованого інтерв'ю; Направлене інтерв'ю; ненаправленим інтерв'ю; Опосередковане інтерв'ю.	Збір даних (наприклад, про використовувані забезпечують засобах, про користувачів і їх інформаційні потреби, про ринок інформаційних продуктів і послуг, посередників, обассортіменте і якості інформаційних продуктів і послуг.

	основних видів опитування.		
спостереження	Метод збору первинної інформації шляхом безпосередньої реєстрації дослідником подій, явищ і процесів, що відбуваються в певних умовах. Здійснюється відповідно до мети і завдань конкретного дослідження.	Систематичне спостереження; Випадкове спостереження; Безпосереднє спостереження; Опосередковане спостереження; Лабораторне (експериментальне) спостереження; Польове спостереження; Короткочасне спостереження; Тривале спостереження	Збір первинної інформації (наприклад, про реалізацію технологічних процесів і операцій, етапи створення і впровадження інформаційних систем).
Метод експертних оцінок	Метод отримання інформації шляхом опитування експертів, є фахівцями в заданій предметній області. Суть методу полягає в проведенні експертами інтуїтивно-логічного аналізу проблеми з кількісним судженням і формальною обробкою результатів. Отримане в результаті обробки узагальнена	Метод комісії; Метод мозкового штурму (або метод колективної генеральної ідеї); метод Дельфі; Методеврістіческого прогнозування; Метод узагальнення незалежних характеристик; Метод простий ранжування; Метод завдання вагових коефіцієнтів; Метод парних срвненій; Метод послідовних порівнянь	Використовується при визначенні проблем, цілей, об'єктів, процедур, критеріїв дослідження. Наприклад, при виявленні та обгрунтуванні складу завдань, що підлягають автоматизації; оцінці проектних рішень; оцінці якості інформаційних продуктів і послуг.

	думка експертів є рішення проблеми.		
аналіз документів	Один з основних методів збору даних, який спрямований на отримання надійної інформації, зафіксованої в документах.	Традиційний аналіз документів (зовнішній і внутрішній); Формалізований аналіз документів (контент-аналіз)	Використовується в якості основного методу при вивченні різних видів документів; може також використовуватися какдополнітельний метод при уточненні, збагаченні або порівнянні результатів спостереження, опитування, їх перевірки.
Контент-аналіз	Формалізований метод дослідження змісту інформації за допомогою виявлення стійко повторюваних смислових одиниць тексту (назв, понять, імен, суджень тощо.). Передбачає систематичну і надійну фіксацію певних елементів змісту деякої сукупності документів (слово, словосполучення, просте речення) з подальшою квантифікації (		Вивчення масивів однорідних документів (наприклад, при аналізі первинного і вторинного документальних потоків); аналіз відповідей на відкриті запитання анкети, інтерв'ю, особистих документів і т.п.

	кількісною обробкою) отриманих даних.		
Факторний аналіз	Метод багатовимірної математичної статистики, спрямований на виявлення і специфічне математичний вираз структур в системах випадкових явищ. Використовується для вимірювання взаємозв'язків між ознаками об'єктів і класифікації ознак з урахуванням цих взаємозв'язків.		Застосовують в тих випадках, коли необхідно встановити і виявити приховані для дослідника чинники, по відношенню до яких первинні емпіричні показники гіпотетично вважаються похідними. Наприклад, при оцінці повноти інформації про вибірку; при визначенні інформативності підсумкового набору вихідних змінних; при аналізі об'єктів проектування; при проведенні пілотажного дослідження.
Порівняльний аналіз	Метод аналізу інформації, що полягає в порівнянні результатів досліджень, проведених на різних об'єктах або в різний час одним або різними дослідницькими колективами з метою узагальнення інформації і забезпечення надійності		Використовується при узагальненні результатів однотипних локальних досліджень з метою отримання висновків, що стосуються великих (масштабних) об'єктів. Наприклад, при аналізі ринку засобів забезпечення інформаційних систем (програмних, технічних, лінгвістичних і т.п.), інформаційних продуктів і послуг, пошукових засобів.

	отриманих результатів.		
ранжування	Метод оцінки змінної, коли її значенням приписується місце в послідовності величин (ранг), яке визначається за допомогою порядкової шкали. Розташування об'єктів сукупності може в порядку зростання або зменшення величини відповідних їм варіантів.		Впорядкування первинних даних. Наприклад, при аналізі контенту сайтів, дослідженні ринку інформаційних продуктів і послуг, обгрунтуванні вибору конкретних засобів забезпечення інформаційних систем. Широко використовується в експертному опитуванні.
угруповання	Метод, що полягає в об'єднанні за істотними ознаками одиниць спостережуваного об'єкта в однорідні сукупності. Угруповання здійснюється як за якісними, так і за кількісними критеріями.	Дискретна угруповання; Интервальная группіровкаГруппіровка за допомогою простого підсумовування однорідних ознак; Ранжування; Угруповання на основі логічно виділених ознак; Табулювання	Обробка матеріалів дослідження; попереднє упорядкування первинної інформації. Наприклад, при аналізі контенту сайтів, дослідженні ринку інформаційних продуктів і послуг, проектуванні інформаційного забезпечення інформаційних систем.
Класифікація	Метод, що полягає в розподілі будь-яких об'єктів за класами на основі їхніх спільних ознак	Ієрархічний методФасетний метод (метод паралельних класифікацій)	Дозволяє встановити зв'язку між досліджуваними об'єктами; служить основою для узагальнюючих висновків і прогнозів.

	(властивостей, характеристик або параметрів об'єктів), й відмінностей, що відображають зв'язки між класами об'єктів в єдиній системі даної галузі знання. Класифікація здійснюється відповідно до обраного підставою розподілу.		Наприклад, при розгляді підприємств, установ, організацій як об'єктів автоматизації, описі складних об'єктів (інформаційних систем, баз і банків даних, сайтів, автоматизованих навчальних систем і т.п.), який передбачає встановлення їх типів і видів.
прогнозування	Метод, що передбачає наукове дослідження перспектив розвитку будь-якого явища або процесу, переважно з кількісними оцінками і з зазначенням більш-менш визначених термінів їх зміни. Направлено на визначення тенденцій і перспектив розвитку тих чи інших процесів на основі аналізу даних про їхнє минуле і нинішній стан.	Глобальне прогнозування; Нормативне прогнозування; Аналітичне прогнозування	Визначення перспектив розвитку інформаційних систем, мереж і технологій.
моделювання	Один з методів	Матеріальне	Застосовується в

	пізнання (відображення) і перетворення світу, сутність якого зводиться до побудови та вивчення деякої моделі з подальшим «перенесенням» отриманих знань на досліджуваний об'єкт.	моделювання: фізичне, аналоговоеІдеальное моделювання: знакове (графічне, логічне, математичне), інтуїтивне	якості універсальної форми пізнання при дослідженні і перетворенні явищ в будь-якій сфері деятельності.Пріменяется в тих випадках, коли об'єкт пізнання недоступний безпосередньому спостереженню і вивченню. Наприклад, при моделюванні предметних областей, побудові моделей баз даних, сайту і т.п.
експеримент	Метод, в основі якого лежить спеціально поставлений досвід в певних умовах, що містять оптимальні можливості для об'єкта дослідження, відповідні задумом експерименту.	Лабораторний експеримент (експерименти, які здійснюють емпіричну перевірку гіпотези або теорії; експерименти, в ході яких відбувається збір необхідної емпіричної інформації для уточнення припущеного); Природний експеримент	Застосовують у випадках, коли стоїть завдання виявлення зв'язків і залежностей між явищами, що вивчаються. Здійснюється на проектної та після проектного стадіях створення інформаційних систем. Наприклад, в ході перевірки працездатності створеної методики, технології, бази даних, інформаційної системи, автоматизованої навчальної системи.

Лекція № 1.2. Методи первісної обробки даних

Первинна статистична обробка даних

Всі методи кількісної обробки прийнято поділяти на первинні та вторинні.

Первинна статистична обробка націлена на упорядкування інформації про об'єкт і предмет вивчення. На цій стадії «сирі» відомості групуються за тими чи іншими критеріями, заносяться в зведені таблиці. Первинно оброблені дані,

представлені в зручній формі, дають дослідникові в першому наближенні поняття про характер всієї сукупності даних в цілому: про їх однорідності - неоднорідності, компактності - розкиданості, чіткості - розмитості і т. Д. Ця інформація добре зчитується з наочних форм представлення даних і дає відомості про їх розподіл.

В ході застосування первинних методів статистичної обробки виходять показники, безпосередньо пов'язані з виробленими в дослідженні вимірами.

До основних методів первинної статистичної обробки відносяться: обчислення заходів центральної тенденції та заходів розкиду (мінливості) даних.

Первинний статистичний аналіз всієї сукупності отриманих в дослідженні даних дає можливість охарактеризувати її в гранично стислому вигляді і відповісти на два головних питання: 1) яке значення найбільш характерно для вибірки; 2) чи великий розкид даних щодо цього характерного значення, т. Е. Яка «розмитість» даних. Для вирішення першого питання обчислюються заходи центральної тенденції, для вирішення другого - заходи мінливості (або розкиду). Ці статистичні показники використовуються щодо кількісних даних, представлених в порядкової, інтервальної або пропорційною шкалою.

Заходи центральної тенденції - це величини, навколо яких групуються інші дані. Дані величини є як би узагальнюючими всю вибірку показниками, що, по-перше, дозволяє судити по ним про всю вибірці, а по-друге, дає можливість порівнювати різні вибірки, різні серії між собою. До заходів центральної тенденції в обробці результатів психологічних досліджень відносяться: вибіркове середнє, медіана, мода.

Вибіркове середнє (M) - це результат ділення суми всіх значень (X) на їх кількість (N).

$$M = \frac{\sum X}{N}.$$

Медіана (Me) - це значення, вище і нижче якого кількість відмінних значень однаково, т. Е. Це центральне значення в послідовному ряду даних. Медіана не обов'язково повинна співпадати з конкретним значенням. Збіг відбувається в разі непарного числа значень (відповідей), розбіжність - при парному їх числі. В останньому випадку медіана обчислюється як середнє арифметичне двох центральних значень в упорядкованому ряду.

Мода (Mo) - це значення, що найчастіше зустрічається у вибірці, т. Е. Значення з найбільшою частотою. Якщо всі значення в групі зустрічаються однаково часто, то вважається, що моди немає. Якщо два сусідніх значення мають однакову частоту і більше частоти будь-якого іншого значення, мода є середнє цих двох значень. Якщо те ж саме відноситься до двох несуміжних значенням, то існує дві моди, а група оцінок є бімодальною.

Зазвичай вибіркове середнє застосовується при прагненні до найбільшої точності у визначенні центральної тенденції. Медіана обчислюється в тому випадку, коли в серії є «нетипові» дані, різко впливають на середнє. Мода використовується в ситуаціях, коли не потрібна висока точність, але важлива швидкість визначення міри центральної тенденції.

Обчислення всіх трьох показників проводиться також для оцінки розподілу даних. При нормальному розподілі значення вибіркового середнього, медіани і моди однакові або дуже близькі.

Заходи розкиду (мінливості) - це статистичні показники, що характеризують відмінності між окремими значеннями вибірки. Вони дозволяють судити про ступінь однорідності отриманого безлічі, його компактності, а побічно і про надійність отриманих даних і що випливають із них результатів. Найбільш використовувані в психологічних дослідженнях показники: середнє відхилення, дисперсія, стандартне відхилення.

Розмах (Р) - це інтервал між максимальним і мінімальним значеннями ознаки. Визначається легко і швидко, але чутливий до випадковостям, особливо при малому числі даних.

Середнє відхилення (МД) - це середньоарифметичне різниці (за абсолютною величиною) між кожним значенням у вибірці і її середнім.

$$МД = \frac{\sum d}{N},$$

де d = | X - M |, M - середнє вибірки, X - конкретне значення, N - число значень.

Безліч всіх конкретних відхилень від середнього характеризує мінливість даних, але якщо не взяти їх за абсолютною величиною, то їх сума буде дорівнює нулю і ми не отримаємо інформації про їх мінливості. Середнє відхилення показує ступінь скупченості даних навколо вибіркового середнього. До речі, іноді при визначенні цієї характеристики вибірки замість середнього (М) беруть інші заходи центральної тенденції - моду або медіану.

Дисперсія (D) характеризує відхилення від середньої величини в даній вибірці. Обчислення дисперсії позляет уникнути нульової суми конкретних різниць (d = X - M) НЕ через їх абсолютні величини, а через їх зведення в квадрат:

$$D = \frac{\sum d^2}{N} \text{ для больших выборок } (N > 30);$$

$$D = \frac{\sum d^2}{(N-1)} \text{ для малых выборок } (N < 30),$$

де d = | X - M |, M - середнє вибірки, X - конкретне значення, N - число значень.

Стандартне відхилення (б). Через зведення в квадрат окремих відхилень d при обчисленні дисперсії отримана величина виявляється далекою від початкових відхилень і тому не дає про них наочного уявлення. Щоб цього уникнути і отримати характеристику, яку можна порівняти із середнім відхиленням, проробляють зворотний математичну операцію - з дисперсії витягають квадратний корінь. Його позитивне значення і приймається за міру мінливості, іменовану среднеквадратическим, або стандартним, відхиленням:

$$\delta = \sqrt{D} = \sqrt{\frac{\sum d^2}{N}}$$ для больших выборок ($N > 30$);

$$\delta = \sqrt{D} = \sqrt{\frac{\sum d^2}{(N-1)}}$$ для малых выборок ($N < 30$),

де d = | X- M |, M - середнє вибірки, X- конкретне значення, N - число значень.

МД, D і? застосовні для інтервальних і Пропорційні даних. Для порядкових даних в якості запобіжного мінливості зазвичай беруть полуквартільное відхилення (Q), яке також називається полуквартільним коефіцієнтом. Обчислюється цей показник наступним чином. Вся область розподілу даних ділиться на чотири рівні частини. Якщо відраховувати спостереження починаючи від мінімальної величини на вимірювальній шкалі, то перша чверть шкали називається першим Квартиль, а точка, яка відокремлює його від іншої частини шкали, позначається символом Qv Другі 25% розподілу - другий квартиль, а відповідна точка на шкалі - Q2. Між третьою і четвертою чвертями розподілу розташована точка Q3. Полуквартільний коефіцієнт визначається як половина інтервалу між першим і третім квартилями:

$$Q = \frac{(Q_3 - Q_1)}{2}.$$

При симетричному розподілі точка Q2 співпаде з медіаною (а отже, і з середнім), і тоді можна обчислити коефіцієнт Q для характеристики розкиду даних щодо середини розподілу. При несиметричному розподілі цього недостатньо. Тоді додатково обчислюють коефіцієнти для лівого і правого ділянок:

$$Q_{лев} = \frac{(Q_2 - Q_1)}{2};$$

$$Q_{прав} = \frac{(Q_3 - Q_2)}{2}.$$

Вторинна статистична обробка даних

До вторинних відносять такі методи статистичної обробки, за допомогою яких на базі первинних даних виявляють приховані в них статистичні закономірності. Вторинні методи можна поділити на способи оцінки значущості відмінностей і способи встановлення статистичних взаємозв'язків.

Способи оцінки значущості відмінностей. Для порівняння вибіркових середніх величин, що належать до двох сукупностей даних, і для вирішення питання про те, чи відрізняються середні значення статистично достовірно друг від друга, використовують t-критерій Стьюдента. Його формула виглядає наступним чином:

$$t = \frac{|M_1 - M_2|}{\sqrt{|m_1^2 - m_2^2|}},$$

де M1, M2 - вибіркові середні значення порівнюваних вибірок, m1, m2 - інтегровані показники відхилень приватних значень з двох порівнюваних вибірок, обчислюються за такими формулами:

$$m_1^2 = \frac{D}{N_1};$$

$$m_2^2 = \frac{D}{N_2},$$

де D1, D2 - дисперсії першої і другої вибірок, N1, N2 - число значень в першій і другій вибірках.

Після обчислення значення показника t по таблиці критичних значень (див. Статистичне додаток 1), заданого числа ступенів свободи (N 1 + N 2 - 2) і обраної ймовірності припустимою помилки (0,05, 0,01, 0,02, 001 і т.д.) знаходять табличне значення t. Якщо обчислене значення t більше або дорівнює табличному, роблять висновок про те, що порівнювані середні значення двох вибірок статистично достовірно різняться з імовірністю допустимої помилки, меншою або рівною обраної.

Якщо в процесі дослідження постає завдання порівняти неабсолютності середні величини, частотні розподілу даних, то використовується? 2 критерій (див. Додаток 2). Його формула виглядає наступним чином:

$$\chi^2 = \sum_{k=1}^{m} \frac{(V_k - P_k)^2}{P_k},$$

де Pk - частоти розподілу в першому вимірі, Vk - частоти розподілу в другому вимірі, m - загальне число груп, на які розділилися результати вимірів.

Після обчислення значення показника? 2по таблиці критичних значень (див. Статистичне додаток 2), заданого числа ступенів свободи (m - 1) і обраної ймовірності припустимою помилки (0,05, 0,0? 2 t більше або дорівнює табличному) роблять висновок про тому, що порівнювані розподілу даних в двох вибірках статистично достовірно різняться з імовірністю допустимої помилки, меншою або рівною обраної.

Для порівняння дисперсій двох вибірок використовується F-критерій Фішера. Його формула виглядає наступним чином:

$$F(N_1 - 1, N_2 - 1) = \frac{D_1}{D_2}$$

де D 1, D 2 - дисперсії першої і другої вибірок, N 1, N 2 - число значень в першій і другій вибірках.

Після обчислення значення показника F по таблиці критичних значень (див. Статистичне додаток 3), заданого числа ступенів свободи (N 1 - 1, N2 - 1) перебуває F кр. Якщо обчислене значення F більше або дорівнює табличному, роблять висновок про те, що відмінність дисперсій в двох вибірках статистично достовірно.

Способи встановлення статистичних взаємозв'язків. Попередні показники характеризують сукупність даних по якогось одного ознакою. Цей змінюється ознака називають змінною величиною або просто змінної. Заходи зв'язку виявляють співвідношення між двома змінними або між двома вибірками. Ці зв'язки, або кореляції, визначають через обчислення коефіцієнтів кореляції. Однак наявність кореляції не означає, що між змінними існує причинний (або функціональна) зв'язок. Функціональна залежність - це окремий випадок кореляції. Навіть якщо зв'язок причинна, кореляційні показники не можуть вказати, яка з двох змінних є причиною, а яка - наслідком. Крім того, будь-яка виявлена в психологічних дослідженнях зв'язок, як правило, існує завдяки і іншим змінним, а не тільки двом розглянутим. До того ж взаємозв'язку психологічних ознак настільки складні, що їх обумовленість однією причиною навряд чи спроможна, вони детерміновані безліччю причин.

За тісноті зв'язку можна виділити наступні види кореляції: повна, висока, виражена, часткова; відсутність кореляції. Ці види кореляцій визначають залежно від значення коефіцієнта кореляції.

При повній кореляції його абсолютні значення рівні або дуже близькі до 1. У цьому випадку встановлюється обов'язкова взаємозалежність між змінними. Тут можлива функціональна залежність.

Висока кореляція встановлюється при абсолютному значенні коефіцієнта 0,8-0,9. Виражена кореляція вважається при абсолютному значенні коефіцієнта 0,6-0,7. Часткова кореляція існує при абсолютному значенні коефіцієнта 0,4-0,5.

Абсолютні значення коефіцієнта кореляції менше 0,4 свідчать про дуже слабкою кореляційної зв'язку і, як правило, до уваги не беруться. Відсутність кореляції констатується при значенні коефіцієнта 0.

Крім того, в психології при оцінці тісноти зв'язку використовують так звану «приватну» класифікацію кореляційних зв'язків. Вона орієнтована не на абсолютну величину коефіцієнтів кореляції, а на рівень значущості цієї величини при певному обсязі вибірки. Ця класифікація застосовується при статистичній оцінці гіпотез. При цьому підході передбачається, що чим більше вибірка, тим менше значення коефіцієнта кореляції може бути прийнято для визнання достовірності зв'язків, а для малих вибірок навіть абсолютно велике значення коефіцієнта може виявитися недостовірним. [86]

За спрямованості виділяють такі види кореляційних зв'язків: позитивна (пряма) і негативна (зворотна). Позитивна (пряма) кореляційний зв'язок реєструється при коефіцієнті зі знаком «плюс»: при збільшенні значення однієї змінної спостерігається збільшення іншого. Негативна (зворотна) кореляція має місце при значенні коефіцієнта зі знаком «мінус». Це означає зворотну залежність: збільшення значення однієї змінної тягне за собою зменшення іншого.

За формою розрізняють наступні види кореляційних зв'язків: прямолінійну і криволінійну. При прямолінійною зв'язку рівномірним змін однієї змінної відповідають рівномірні зміни іншої. Якщо говорити не тільки про кореляції, але і про функціональні залежності, то такі форми залежності називають

пропорційними. У психології строго прямолінійні зв'язку - явище рідкісне. При криволінійної зв'язку рівномірний зміна однієї ознаки поєднується з нерівномірним зміною іншого. Ця ситуація для психології типова.

Коефіцієнт лінійної кореляції по К. Пірсон (r) обчислюється с допомогою наступної формули:

$$r = \frac{\sum xy}{N \delta x \delta y},$$

де x - відхилення окремого значення X від середнього вибірки (MX), y - відхилення окремого значення Y від середнього вибірки (M o), Ьх - стандартне відхилення для X,? y - стандартне відхилення для Y, N - число пар значень X і Y.

Оцінка значущості коефіцієнта кореляції проводиться по таблиці (див. Статистичне додаток 4).

При порівнянні порядкових даних застосовується коефіцієнт рангової кореляції по Ч. Спирмену (R):

$$R = 1 - \frac{6 \sum d^2}{N(N^2 - 1)},$$

де d - різниця рангів (порядкових місць) двох величин, N - число порівнюваних пар величин двох змінних (X і Y).

Оцінка значущості коефіцієнта кореляції проводиться по таблиці (див. Статистичне додаток 5).

Впровадження в наукові дослідження автоматизованих засобів обробки даних дозволяє швидко і точно визначати будь-які кількісні характеристики будь-яких масивів даних. Розроблено різні програми для комп'ютерів, за якими можна проводити відповідний статистичний аналіз практично будь-яких вибірок. З маси статистичних прийомів в психології найбільшого поширення набули наступні: 1) комплексне обчислення статистик; 2) кореляційний аналіз; 3) дисперсійний аналіз; 4) регресійний аналіз; 5) факторний аналіз; 6) таксономический (кластерний) аналіз; 7) шкалювання. Познайомитися з характеристиками цих методів можна в спеціальній літературі («Статистичні методи в педагогіці і психології» Стенлі Дж., Голосу Дж. (М., 1976), «Математична психологія» Г.В. Суходольського (СПб., 1997), «Математичні методи психологічного дослідження »А.Д. Наследова (СПб., 2005) і ін.).

Лекція № 1.3. Кластерний аналіз

Кластерний аналіз (англ. Data clustering) — задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, що називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя.

Загальна характеристика

Кластерний аналіз — це багатовимірна статистична процедура, яка виконує збір даних, що містять інформацію про вибірку об'єктів і потім упорядковує об'єкти в порівняно однорідні групи — кластери (Q-кластеризація, або Q-техніка, власне кластерний аналіз).

Основна мета кластерного аналізу — знаходження груп схожих об'єктів у вибірці. Спектр застосувань кластерного аналізу дуже широкий: його використовують в археології, антропології, медицині, психології, хімії, біології, державному управлінні, філології, маркетингу, соціології та інших дисциплінах. Однак універсальність застосування привела до появи великої кількості несумісних термінів, методів і підходів, що ускладнюють однозначне використання і несуперечливу інтерпретацію кластерного аналізу.

Завдання

Кластерний аналіз виконує наступні основні завдання:

- Розробка типології або класифікації.
- Дослідження корисних концептуальних схем групування об'єктів.
- Породження гіпотез на основі дослідження даних.
- Перевірка гіпотез або дослідження для визначення, чи дійсно групи, виділені тим чи іншим способом, присутні в наявних даних.

Етапи

Незалежно від конкретної сфери, застосування кластерного аналізу передбачає наступні етапи:

- Відбір вибірки для кластеризації.
- Визначення множини характеристик, по яких будуть оцінюватися об'єкти у вибірці.
- Обчислення значень тієї чи іншої міри схожості між об'єктами.
- Застосування одного з методів кластерного аналізу для створення груп схожих об'єктів.
- Перевірка достовірності результатів кластеризації.

Якщо кластерному аналізу передує факторний аналіз, то вибірка не потребує коректування — викладені вимоги виконуються автоматично самою процедурою факторного моделювання. В іншому випадку вибірку потрібно коректувати.

Методи кластеризації

Об'єднання схожих об'єктів у групи може бути здійснене різними способами. Саме для цього етапу існує цілий ряд методів:

- K-середніх (K-means)
- Нечітка кластеризація C-середніх (C-means)
- Графові алгоритми кластеризації
- Статистичні алгоритми кластеризації
- Алгоритми сімейства FOREL
- Ієрархічна кластеризація або таксономія
- Нейронна мережа Кохонена
- Ансамбль кластеризаторів
- Алгоритми сімейства KRAB

- EM-алгоритм
- Метод просіювання

Вхідні дані

Типи вхідних даних

Вхідними даними кластерного аналізу є набір об'єктів. В залежності від способу представлення цих об'єктів розрізняють такі типи вхідних даних:

- Вектор характеристик. Кожен об'єкт описується набором своїх характеристик; ці характеристики можуть бути числовими або нечисловими.
- Матриця відстаней. Кожен об'єкт описується відстанями до всіх інших об'єктів вибірки.

Вимоги до вхідних даних

Кластерний аналіз висуває наступні вимоги до даних:

- Об'єкти не повинні корелювати між собою.
- Об'єкти мають бути безрозмірними.
- Розподіл об'єктів має бути близьким до нормального.
- Об'єкти повинні відповідати вимозі стійкості, під якою розуміється відсутність впливу на їх значення випадкових чинників.
- Вибірка повинна бути однорідна.

Результати

Причини неоднозначності

Рішення задачі кластеризації принципове неоднозначне, і цьому є декілька причин:

- Не існує однозначно якнайкращого критерію якості кластеризації. Відомий цілий ряд евристичних критеріїв, а також ряд алгоритмів, що не мають чітко вираженого критерію, але здійснюють достатньо розумну кластеризацію «по побудові». Всі вони можуть давати різні результати.
- Число кластерів, як правило, невідоме заздалегідь і встановлюється відповідно до деякого суб'єктивного критерію.
- Результат кластеризації істотно залежить від метрики, вибір якої, як правило, також суб'єктивний і визначається експертом.

Інтерпретація результатів

Результатом кластеризації є групи об'єктів, об'єднані за певною характеристикою чи характеристиками. Однак ці результати можуть бути інтерпретовані по-різному. Зокрема, при аналізі результатів соціологічних досліджень рекомендується здійснювати аналіз ієрархічними методами, наприклад методом Уорда, при якому всередині кластерів оптимізується мінімальна дисперсія і в результаті створюються кластери приблизно рівних розмірів. Як міра відмінності між кластерами використовується квадратична евклідова відстань, що сприяє збільшенню контрастності кластерів.

Тепер виникає питання стійкості знайденого кластерного рішення. По суті, перевірка стійкості кластеризації зводиться до перевірки її достовірності. Тут існує емпіричне правило — стійка типологія зберігається при зміні методів кластеризації. Результати ієрархічного кластерного аналізу можна перевіряти ітеративним кластерним аналізом методом k-середніх. Якщо при порівнянні

групи збігаються більше, ніж на 70 % (понад 2/3 збігів), то кластерне рішення приймається.
Перевірити адекватність рішення, не вдаючись до допомоги інших видів аналізу, не можна. Принаймні, в теоретичному плані ця проблема не вирішена. Деякі додаткові методи перевірки стійкості відкидаються з певних причин:

- Кофенетична кореляція — не рекомендується і обмежена у використанні.
- Тести значущості (дисперсійний аналіз) — завжди дають значущий результат.
- Метод повторних випадкових вибірок — не доводить правильність рішення.
- Тести значущості для зовнішніх ознак — придатні тільки для повторних вимірювань.
- Методи Монте-Карло — дуже складні і доступні тільки досвідченим математикам.

Ієрархічна кластеризація

Ієрархічна кластеризація (також «графові алгоритми кластеризації») — сукупність алгоритмів впорядкування даних, візуалізація яких забезпечується за допомогою графів.
Алгоритми сортування даних зазначеного типу виходять з того, що якась безліч об'єктів характеризується певним ступенем зв'язності. Передбачається наявність вкладених груп (кластерів різного порядку). Алгоритми в свою чергу поділяються на агломеративні (об'єднувальні) і дивізівні (розділяючі). По кількості ознак іноді виділяють монотетичні та політетичні методи класифікації. Як і більшість візуальних способів подання залежностей графи швидко втрачають наочність при збільшенні числа об'єктів. Існує ряд спеціалізованих програм для побудови графів.

Дискримінáнтний анáліз

Дискримінáнтний анáліз — різновид багатовимірного аналізу, призначеного для вирішення задач розпізнавання образів. Використовується для прийняття рішення про те, які змінні розділюють (тобто «дискримінують») певні масиви даних (так звані «групи»).

Дискримінантний аналіз є близьким до дисперсійного і регресійного аналізів, які також намагаються виразити одну із залежних змінних у вигляді лінійної комбінації інших показників або вимірювань. Однак, у двох інших методів залежна змінна є числовий величиною, в той час як у дискримінантному аналізі це категоріальна змінна. Більш подібними до дискримінантного аналізу є логістична і пробіт-регресія, оскільки вони також пояснюють категоріальну змінну. Ці та інші методи використовуються переважно в тих випадках, коли не припускається нормальний розподіл незалежних змінних, що є основним припущенням методу дискримінантного аналізу.

Дискримінантний аналіз широко застосовується в економіці маркетингових дослідженнях при вирішенні питань сегментації ринку, при об'єктивній оцінці ступеня новизни товарів тощо.

Лекція № 1.4. Регресійний, кореляційний та дисперсійний аналіз

Кореляційний аналіз даних

Кореляційний аналіз — це статистичне дослідження (стохастичної) залежності між випадковими величинами (англ. correlation — взаємозв'язок). У найпростішому випадку досліджують дві вибірки (набори даних), у загальному — їх багатовимірні комплекси (групи).[1]
Мета кореляційного аналізу — виявити чи існує істотна залежність однієї змінної від інших.
Головні завдання кореляційного аналізу:

1. оцінка за вибірковими даними коефіцієнтів кореляції
2. перевірка значущості вибіркових коефіцієнтів кореляції або кореляційного відношення
3. оцінка близькості виявленого зв'язку до лінійного
4. побудова довірчого інтервалу для коефіцієнтів кореляції.

Обмеження кореляційного аналізу
Кореляція відображає лише лінійну залежність величин, але не відображає їх функціональної зв'язаності. Наприклад, якщо обчислити коефіцієнт кореляції між величинами A = sin(x) та B = cos(x), він буде наближений до нуля, тобто залежність між величинами відсутня. Між тим, величини A та B очевидно зв'язані між собою за законом $\sin^2(x) + \cos^2(x) = 1$.
Використання можливе у випадку наявності достатньої кількості випадків для вивчення: для конкретного типу коефіцієнту кореляції становить від 25 до 100 пар спостережень.
Кореляція не означає причинність.

Регресійний аналіз даних

Регресі́йний ана́ліз — розділ математичної статистики, присвячений методам аналізу залежності однієї величини від іншої. На відміну від кореляційного аналізу не з'ясовує чи істотний зв'язок, а займається пошуком моделі цього зв'язку, вираженої у функції регресії.
Регресійний аналіз використовується в тому випадку, якщо відношення між змінними можуть бути виражені кількісно у виді деякої комбінації цих змінних. Отримана комбінація використовується для передбачення значення, що може приймати цільова (залежна) змінна, яка обчислюється на заданому наборі значень вхідних (незалежних) змінних. У найпростішому випадку для цього використовуються стандартні статистичні методи, такі як лінійна регресія. На жаль, більшість реальних моделей не вкладаються в рамки лінійної регресії. Наприклад, розміри продажів чи фондові ціни дуже складні для передбачення,

оскільки можуть залежати від комплексу взаємозв'язків множин змінних. Таким чином, необхідні комплексні методи для передбачення майбутніх значень.

Мета регресійного аналізу

1. Визначення ступеня детермінованості варіації критеріальної (залежної) змінної предикторами (незалежними змінними).
2. Прогнозування значення залежної змінної за допомогою незалежної.
3. Визначення внеску окремих незалежних змінних у варіацію залежної.

Регресійний аналіз не можна використовувати для визначення наявності зв'язку між змінними, оскільки наявність такого зв'язку і є передумова для застосування аналізу.

Класична нормальна лінійна модель множинної регресії

Економічні явища, як правило, визначаються більш чим одним одночасно та сукупно діючих факторів. У зв'язку з цим виникає задача дослідження залежності однієї залежної змінної У від декількох пояснюючих змінних $X_1, X_2, \ldots, X_p$. Ця задача вирішується за допомогою множинного регресійного аналізу. Множинна регресія широко використовується при рішенні питань попиту, доходності акцій, при вивченні витрат виробництва, у макроекономічних розрахунках і тощо.

Загальна множинна регресійна модель має слідуючий вигляд:

$$y = f(x_1, x_2, \ldots x_p) + \varepsilon, \qquad (1)$$

де у - залежна змінна;

$x_1, x_2, \ldots, x_p$ - фактори (незалежні змінні).

Якщо множинна регресійна модель є лінійною (ЛМР), то вона подається у вигляді:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon. \qquad (2)$$

Позначимо і-е спостереження змінної у через y_i, а факторів – $x_{i1}, x_{i2}, \ldots, x_{ip}$. Відтоді модель (2) можна подати у вигляді:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \; i = \overline{1, n}, \qquad (3)$$

або у матричній формі:

$$y = X\beta + \varepsilon,$$

де $y = [y_1, y_2, \ldots y_n]^T$ - вектор (матриця-стовпець) значень залежної змінної;

$\beta = [\beta_1, \beta_2, \ldots \beta_p]^T$ - вектор (матриця-стовпець) коефіцієнтів регресійної моделі;

$\varepsilon = [\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n]^T$ - вектор (матриця-стовпець) похибок;

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$ - матриця значень факторів.

Відмітимо основні припущення регресійного аналізу:

1. В моделі (3) похибка ε_i (або залежна змінна y_i) є випадковою величиною, а фактори x_{ip} невипадкові величини ($i=\overline{1,n}$).

2. Математичне сподівання похибки ε_i дорівнює нулю:

$M[\varepsilon_i]=0$, $i=\overline{1,n}$.

3. Дисперсія похибки ε_i (або залежної змінної y_i) постійна дія будь-якої i:

$D[\varepsilon_i]=\sigma^2$.

тобто виконується умова гомоскедастичності.

4. Похибки ε_i та ε_j не корельовані:

$M[\varepsilon_i\varepsilon_j]=0$, $i\neq j$.

5. Похибка ε_i (або залежна змінна y_i) є нормально розподіленою випадковою величиною.

6. Матриця значень факторів невироджена, тобто її ранг дорівнює $p+1$:

$rangX=p+1<n$.

Модель (4.20), для якої виконуються припущення 1-6, називається класичною нормальною лінійною моделлю множинної регресії (CNLMR-model).

Оцінкою цієї моделі за вибіркою є рівняння регресії:

$$\hat{y}=b_0+b_1x_1+b_2x_2+\_+b_px_p, \qquad (4)$$

де $\hat{y}$ - оцінка математичного сподівання залежної змінної $M_x[y]$;

$b_i\left(i=\overline{0,p}\right)$ - оцінка коефіцієнтів $\beta_i\left(i=\overline{0,p}\right)$ регресійної моделі (або коефіцієнти регресії).

Як і раніше, для оцінки коефіцієнтів CNLMR-model використовують МНК:

$$S(b_0,b_1,\_,b_p)=\sum_{i=1}^{n}\left(y_i-b_0-b_1x_{i1}-b_2x_{i2}-\_-b_px_{ip}\right)^2\to\min.$$

Після розв'язання системи нормальних рівнянь

$$\begin{cases}\dfrac{\partial S}{\partial b_0}=-2\sum\limits_{i=1}^{n}\left(y_i-b_0-b_1x_{i1}-\_-b_px_{ip}\right)=0,\\ \dfrac{\partial S}{\partial b_1}=-2\sum\limits_{i=1}^{n}x_{i1}\left(y_i-b_0-b_1x_{i1}-\_-b_px_{ip}\right)=0,\\ \text{- - - - - - -}\\ \dfrac{\partial S}{\partial b_p}=-2\sum\limits_{i=1}^{n}x_{ip}\left(y_i-b_0-b_1x_{i1}-\_-b_px_{ip}\right)=0,\end{cases}$$

отримаємо значення коефіцієнтів рівняння регресії, які в матричній формі мають вигляд:

$$b=\left(X^TX\right)^{-1}X^Ty, \qquad (5)$$

де $b=\left[b_1,b_2,\_,b_p\right]^T$ - вектор (матриця-стовпець) коефіцієнтів рівняння регресії.

Оцінки b_j є незміщеними, обґрунтованими та ефективними.

Оцінка дисперсії похибок

$$S^2 = \frac{\sum_{i}^{n}\left(y_i - \hat{y}_i\right)^2}{n-p-1} \qquad (6)$$

є незміщеною та обґрунтованою.

Коефіцієнт (індекс) множинної кореляції R використовується для оцінки тісноти спільного впливу факторів на залежну змінну:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} . \qquad (7)$$

Властивості коефіцієнта множинної кореляції R:

1. Коефіцієнт множинної кореляції приймає значення на відрізку $[0;1]$, тобто $0 \le R \le 1$.

Чим ближче R до одиниці, тим тісніше зв'язок між залежною у та факторами $x_1, x_2, \ldots, x_p$.

2. При R=1 кореляційний зв'язок є лінійною функціональною залежністю.

3. При R=0 лінійний кореляційний зв'язок відсутній.

Щодо оцінки ступеня взаємозв'язку, можна керуватись аналогічними емпіричними правилами, як і для випадку ЛПР (лекція 3.1).

Оцінка значущості множинної регресії. Коефіцієнт детермінації

Оцінка значущості ЛМР.

Значущість рівня ЛМР у цілому оцінюється за допомогою F-критерія Фішера

$$F_{факт} = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p} \qquad (8)$$

із зрівнянням його з табличним значенням

$$F_{табл} = F_{\alpha;p;n-p-1}. \qquad (9)$$

F-тест.

Якщо $F_{факт} > F_{\alpha;p;n-p-1}$, то рівняння ЛМР признається статистично значущим на рівні значущості α (зазвичай, $\alpha = 0{,}05$).

Якщо $F_{факт} < F_{\alpha;p;n-p-1}$, то рівняння ЛМР признається статистична незначущість ЛМР на рівні значущості α.

Другий варіант F- тесту: якщо рівень значущості фактичного F-критерію $\alpha_F < \alpha$, то рівняння ЛМР – статистичного значуще на рівні значущості α.

Якщо $\alpha_F > \alpha$, то ЛМР – статистичного незначуще на рівні значущості α.

Оцінка значущості коефіцієнтів рівняння ЛМР

Оцінка значущості коефіцієнтів рівняння ЛМР здійснюються за допомогою t-критерію Ст'юдента:

$$t_{b_j} = \frac{|b_j|}{S_{b_j}} \qquad (10)$$

із зрівнянням його з табличним значенням

$$t_{табл} = t_{1-\alpha;n-p-1}, \qquad (11)$$

де $S_{b_j} = S\sqrt{(X^T X)^{-1}_{jj}}\ (j = \overline{0,p})$ - середньоквадратичне відхилення (стандартна похибка) коефіцієнт а регресії b_j;

$S = \frac{\sum_i^n \left(y_i - \hat{y}_i\right)^2}{n-p-1}$ - оцінка середньоквадратичного відхилення похибок;

$(X^T X)^{-1}_{jj}\ (j = \overline{0,p})$ - відповідний діагональний елемент матриці $(X^T X)^{-1}$.

t-тест.

Якщо $t_{b_j} > t_{1-\alpha;n-p-1}$, то коефіцієнт b_j признається статистично значущим; якщо $t_{b_j} < t_{1-\alpha;n-p-1}$, то b_j – статистично незначущий на рівні значущості α.

Другий варіант (див. t-тест для ЛПР у лекції 3.1): при $\alpha_{b_j} < \alpha\ (\alpha_{b_j} > \alpha)$ _ b_j – статистично значущий (незначущий) на рівні значущості α.

Коефіцієнт (індекс) множинної детермінації R^2

Для оцінки адекватності регресії моделі, мірою якості рівняння регресії використовують коефіцієнт детермінації, який визначається, як і раніше, за формулою:

$$R^2 = \sqrt{\frac{\sum_{i=1}^n \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2}}. \qquad (12)$$

Нагадаємо, що R^2 характеризує частку варіації залежної змінної, що обумовлена варіаціями факторів.

Властивості коефіцієнта множинної детермінації R^2:

1. Коефіцієнт множинної детермінації приймає значення на відрізку $[0;1]$, тобто $0 \le R^2 \le 1$.

Чим ближче R^2 до одиниці, тим краще регресія апроксимує емпіричні дані.

2. Якщо $R^2=1$, між змінними y та $x_1, x_2,\ldots, x_p$ існує лінійна функціональна залежність.

3. Якщо $R^2=0$, то варіація залежної змінної повністю обумовлена виливом випадкових та неврахованих факторів.

Для оцінки ступеня апроксимації емпіричних даних рівнянням ЛМР можна керуватись аналогічними емпіричними правилами, як і для випадку ЛПР (лекція 3.1).

Зауваження

Недоліком коефіцієнта множинної детермінації R^2 являється те, що він, взагалі, збільшується при додаванні нових факторів, хоча це не обов'язково означає поліпшення якості регресійної моделі. Тому має сенс використовувати скоригований (адаптований, виправлений) коефіцієнт детермінації $\bar{R}^2$, який визначається за формулою:

$$\bar{R}^2 = 1 - \frac{n-1}{n-p-1}\left(1-R^2\right). \qquad (13)$$

На відміну від R^2 скоригований коефіцієнт $\bar{R}^2$ може зменшуватись при введенні у модель нових факторів, які не чинять істотного впливу на залежну змінну.

Визначення довірчих інтервалів для функції регресії та її параметрів

Прогнозне значення y_H визначається за шляхом підстановки у рівняння регресії (4) відповідних значень факторів $x_{H1}, x_{H2}, \ldots, x_{Hp}$:

$$\hat{y}_H = \hat{y}\left(x_{H1}, x_{H2}, \ldots, x_{Hp}\right) = b_0 + b_1 x_{H1} + b_2 x_{H2} + \ldots + b_p x_{Hp}. \qquad (14)$$

Довірчий інтервал прогнозу обчислюється за слідуючими формулами:

$$M_{x_{H1}\ldots x_{Hp}}[y] = \hat{y}_H \pm t_{1-\alpha;n-p-1} \cdot S_{\hat{y}}, \qquad (15)$$

де $M_{x_{H1}\ldots x_{Hp}}[y]$ - умовне математичне сподівання залежної змінної в точці прогнозу;

$S_{\hat{y}}$ - оцінка стандартної похибки прогнозу, яка обчислюється за формулою

$$S_{\hat{y}} = S\sqrt{X_H^T\left(X^T X\right)^{-1} X_H}; \qquad (16)$$

X - матриця значень факторів;

$X_H = \left[1\, x_{H1}\, x_{H2} \ldots x_{Hp}\right]^T$ - вектор (матриця-стовпець) значень факторів для прогнозу.

Довірчі інтервали для коефіцієнтів регресійної моделі:

$$\beta_j = b_j \pm t_{1-\alpha;n-p-1} \cdot S_{b_j},\ j = \overline{0,p}. \qquad (17)$$

Обчислення коефіцієнтів еластичності.

Коефіцієнт еластичності E_j показує – на скільки відсотків (від середньої) змінюється у середньому y при зміненні тільки x_j на 1% та обчислюється за формулою:

$$E_j = b_j \frac{\overline{x_j}}{\overline{y}}, \quad j = \overline{1,p}. \qquad (18)$$

Важливою економічною характеристикою моделі є сумарна еластичність:

$$E = \sum_{j=1}^{p} E_j. \qquad (19)$$

Сумарна еластичність показує на скільки відсотків (від середньої) змінюється у середньому y при зміненні всіх факторів на 1%.

Дисперсійний аналіз

Дисперсійний аналіз (англ. analysis of variance (ANOVA)) являє собою статистичний метод аналізу результатів, які залежать від якісних ознак. Кожен фактор може бути дискретною чи неперервною випадковою змінною, яку розділяють на декілька сталих рівнів (градацій, інтервалів). Якщо кількість вимірювань (проб, даних) на всіх рівнях кожного з факторів однакова, то дисперсійний аналіз називають рівномірним, інакше — нерівномірним. В основі дисперсійного аналізу є такий принцип (факт з математичної статистики): якщо на випадкову величину діють взаємно незалежні фактори A, B, …, то загальна дисперсія дорівнює сумі дисперсій, зумовлених дією окремо кожного з факторів: $\sigma^2 = \sigma_A^2 + \sigma_B^2 + \dots$

Задачі дисперсійного аналізу

В будь-якому експерименті середні значення досліджуваних величин змінюються у зв'язку зі зміною основних факторів (кількісних та якісних), що визначають умови досліду, а також і випадкових факторів. Дослідження впливу тих чи інших факторів на мінливість середніх є задачею дисперсійного аналізу. Дисперсійний аналіз використовує властивість адитивності дисперсії випадкової величини, що обумовлено дією незалежних факторів. В залежності від числа джерел дисперсії розрізняють однофакторний та багатофакторний дисперсійний аналіз.

Дисперсійний аналіз особливо ефективний при вивченні кількох факторів. При класичному методі вивчення змінюють тільки один фактор, а решту залишають постійними. При цьому для кожного фактору проводиться своя серія спостережень, що не використовується при вивченні інших факторів. Крім того, при такому методі досліджень не вдасться визначити взаємодію факторів при одночасній їх зміні. При дисперсійному аналізі кожне спостереження служить для одночасної оцінки всіх факторів та їх взаємодії.

Дисперсійний аналіз полягає у виділенні й оцінюванні окремих факторів, що викликають зміну досліджуваної випадкової величини. При цьому проводиться

розклад сумарної вибіркової дисперсії на складові, обумовлені незалежними факторами. Кожна з цих складових є оцінкою дисперсії генеральної сукупності. Щоб дати оцінку дієвості впливу даного фактору, необхідно оцінити значимість відповідної вибіркової дисперсії у порівнянні з дисперсією відтворення, обумовленою випадковими факторами. Перевірка значимості оцінок дисперсії проводять з допомогою критерію Фішера.
Коли розрахункове значення критерію Фішера виявиться меншим табличного, то вплив досліджуваного фактору немає підстав вважати значимим. Коли ж розрахункове значення критерію Фішера виявиться більшим табличного, то цей фактор впливає на зміни середніх. В подальшому ми вважаємо, що виконуються наступні припущення:

1. Випадкові помилки спостережень мають нормальний розподіл.
2. Фактори впливають тільки на зміну середніх значень, а дисперсія спостережень залишається постійною.

Фактори, що розглядаються в дисперсійному аналізі, бувають трьох родів:

- з випадковими рівнями, коли вибір рівнів проходить з безмежної сукупності можливих рівнів та супроводжується рандомізацією і рівні вибираються випадковим чином;
- з фіксованими рівнями;
- змішаного типу — частина факторів розглядається на фіксованих рівнях, але рівні решти вибираються випадковим чином.

Дисперсійний аналіз застосовується в різних формах в залежності від структури об'єкту, що досліджується; вибір відповідної форми є однією з головних трудностей в практичному застосуванні аналізу.
Дисперсійний аналіз використовує властивість адитивності дисперсії випадкової величини, що обумовлено дією незалежних факторів. В залежності від числа джерел дисперсії розрізняють однофакторний та багатофакторний дисперсійний аналіз.

Прогнозування

Прогнозування — процес передбачення майбутнього стану предмета чи явища на основі аналізу його минулого і сучасного, систематично оцінювана інформація про якісні й кількісні характеристики розвитку обраного предмета чи явища в перспективі. Результатом прогнозування є прогноз — знання про майбутнє і про ймовірний розвиток сьогочасних тенденцій конкретного явища-об'єкту в подальшому існуванні.

Застосування прогнозування

Прогнозування застосовується в наступних ситуаціях:

- Ланцюг постачання — прогнозування застосовується в ланцюгу постачання для забезпечення клієнтів компанії правильним продуктом у правильний час. Є складовою частиною процесів управління попитом та планування продажів і операцій, який є складовою частиною процесів в алгоритмі MRP II
- Бізнес-планування — частина підготовки та розробки бізнес-планів
- Прогноз погоди, метеорологія
- Планування транспорту

- Економічне прогнозування
- Технологічне прогнозування
- Передбачення землетрусів
- Політичне прогнозування
- Прогнозування педагогічне

Методи прогнозування

Існують два підходи до прогнозування: **якісний** та **кількісний**.

Кількісний підхід базується на математичних моделях й історичних данних. Якісний підхід покладається на освічену думку, інтуїцію й досвід професіоналів. Серед його різновидів є консенсус керівництва, Делфі-метод, оцінка торговими працівниками — кожного за своїм регіоном, опитування клієнтів.

Кількісні методи діляться на два види: **причинно-наслідкові** й **моделі часових рядів**. Часові ряди діляться на:

- моделі з декомпозицією: виділення сезонності й тренду;
- моделі згладжування:
  - середнє арифметичне;
  - ковзне середнє арифметичне;
  - середнє зважене;
  - ковзне середнє зважене;
  - експоненційне згладжування

Оцінка прогнозу

Надзвичайно важливо оцінювати прогноз. Серед безлічі надійних методів є декілька, які найпростішими, але, поєднані разом, дають безпомилкову оцінку якості поданого плану продажу та його фактичного виконання. Це:

- середнє абсолютне відхилення / Mean Average Deviation (MAD);
- сума помилок прогнозу зростаючим підсумком / Running Sum of Forecast Error (RSFE);
- сигнал відслідковування / Tracking Signal (TS)

Процедури, процеси, відповідальність у прогнозуванні

Система прогнозування вимагає певного рівня формалізації, тобто описана процедурами й інструкціями. В бізнес-прогнозуванні процесом займаються та за нього відповідають підрозділи продажу та маркетингу, які відповідають за збут у компаніях: відділ збуту, відділ маркетингу, відділ з роботи з клієнтами, товарознавців. Проте, досить поширеною є практика коли короткострокове прогнозування здійснюється працівниками планових відділів логістики.

Економічне прогнозування

Є найважливішим фактором успішного розвитку усіх держав світу. Саме від визначення раціональних прогностичних аспектів залежить вдалий соціально-економічний розвиток будь-якого суспільства. Найавторитетнішою науковою установою в питаннях економічного прогнозування в Україні є Інститут економіки та прогнозування НАН України, який наразі очолює видатний вчений-економіст академік Геєць Валерій Михайлович.

Модуль №2 «Інтелектуальний аналіз даних»

Лекція № 2.1. Методи інтелектуального аналізу даних (Data Mining)

Технології аналізу даних, що базуються на застосуванні класичних статистичних підходів, мають низку недоліків. Відповідні методи ґрунтуються на використанні усереднених показників, на підставі яких важко з'ясувати справжній стан справ у досліджуваній сфері (наприклад, середня зарплата по країні не відбиває її розміру у великих містах та в селах). Методи математичної статистики виявилися корисними насамперед для перевірки заздалегідь сформульованих гіпотез та «грубого» розвідницького аналізу, що становить основу оперативної аналітичної обробки даних (OLAP).

Наприклад, дослідження спеціалістів Гарвардського інституту показують, що на основі наявної інформації за допомогою стандартних статистичних методів не можна було передбачити великої депресії кінця 1920-х років.

Окрім того, стандартні статистичні методи відкидають (нехтують) нетипові спостереження — так звані піки та сплески. Проте окремі нетипові значення можуть становити самостійний інтерес для дослідження, характеризуючи деякі виняткові, але важливі явища. Навіть сама ідентифікація цих спостережень, не говорячи про їх подальший аналіз і докладний розгляд, може бути корисною для розуміння сутності досліджуваних об'єктів чи явищ. Як показують сучасні дослідження, саме такі події можуть стати вирішальними щодо майбутнього поводження та розвитку складних систем.

Ці недоліки статистичних методів спонукали до розвитку нових методів дослідження складних систем, що базуються на нелінійній динаміці, теорії катастроф, фрактальній геометрії тощо.

Водночас постала нагальна потреба в такій технології, яка автоматично видобувала б із даних нові нетривіальні знання у формі моделей, залежностей, законів тощо, гарантуючи при цьому їхню статистичну значущість. Новітні підходи, спрямовані на розв'язання цих проблем, дістали назву технологій інтелектуаль-ного аналізу даних.

В основу цих технологій покладено концепцію шаблонів (патернів), що відбивають певні фрагменти багатоаспектних зв'язків у множині даних, характеризуючи закономірності, притаманні під-вибіркам даних, які можна компактно подати у зрозумілій людині формі. Шаблони відшукують методами, що виходять за межі апрі-орних припущень стосовно структури вибірки та вигляду розподілів значень аналізованих показників. Важлива особливість цієї технології полягає в нетривіальності відшукуваних шаблонів. Це означає, що вони мають відбивати неочевидні, несподівані регулярності у множині даних, складові так званого прихованого знання. Адже сукупність первинних («сирих») даних може містити й глибинні шари знань.

Knowledge Discovery in Databases (дослівно: «виявлення знань у базах даних» — KDD) — аналітичний процес дослідження значних обсягів інформації із залученням засобів автоматизації, що має на меті виявити приховані у множині даних структури, залежності й взаємозв'язки. При цьому передбачається повна чи часткова відсутність апріорних уявлень про характер прихованих структур

та залежностей. KDD передбачає, що людина попередньо осмислює задачу й подає неповне (у термінах цільових змінних) її формулювання, перетворює дані до формату придатного для їх автоматизованого аналізу й попередньої обробки, виявляє засобами автоматичного дослідження даних приховані структури й залежності, апробовує виявлені моделі на нових даних, не використовуваних для побудови моделей, та інтерпретує виявлені моделі й результати.

Отже, KDD — це синтетична технологія, що поєднує в собі останні досягнення штучного інтелекту, чисельних математичних методів, статистики й евристичних підходів. Методи KDD особливо стрімко розвиваються протягом останніх 20 років, а раніше задачі комп'ютерного аналізу баз даних виконувалися переважно за допомогою різного роду стандартних статистичних методів.

Data Mining (дослівно: «Розробка, добування даних» — DM) — дослідження «сирих» даних і виявлення в них за допомогою «машини» (алгоритмів, засобів штучного інтелекту) прихованих нетривіальних структур і залежностей, які раніше не були відомі й мають практичну цінність та придатні для того, щоб їх інтерпретувала людина.

Розглянемо відмінності між засобами Data Mining і OLAP. Технологія OLAP спрямована на підтримання процесу прийняття управлінських рішень і використовується з метою пошуку відповіді на запитання: чому деякі речі є такими, якими вони є насправді? При цьому користувач сам формує модель-гіпотезу про дані чи відношення між даними, а далі, застосовуючи серію запитів до бази даних, підтверджує чи відхиляє висунуті гіпотези. Засоби Data Mining відрізняються від засобів OLAP тим, що замість перевірки передбачуваних користувачем взаємозалежностей вони на основі наявних даних самі можуть будувати моделі, які дають змогу кіль-кісно та якісно оцінювати ступінь впливу різних досліджуваних факторів на задану властивість об'єкта. Крім того, засоби DM дають змогу формулювати нові гіпотези про характер досі невідомих, але таких, що реально існують, залежностей між даними.

Засоби OLAP застосовуються на ранніх стадіях процесу KDD, оскільки вони дають змогу краще зрозуміти дані, що, у свою чергу, забезпечує ефективніший результат процесу KDD.

Головна мета технології KDD — побудова моделей і відношень, прихованих у базі даних, тобто таких, які не можна знайти звичайними методами. Варто зазначити, що на комп'ютери перекладаються не лише рутинні операції (скажімо, перевірка статистичної значущості гіпотез), а й операції, що донедавна були аж ніяк не рутинними (вироблення нових гіпотез). KDD дає змогу побачити такі відношення між даними, що залишалися поза увагою дослідників.

Будуючи моделі, ми встановлюємо кількісні зв'язки між характеристиками досліджуваного явища. Щодо призначення можна виокремити моделі двох типів: прогнозні та описові (дескриптивні). Моделі першого типу використовують набори даних із відомими результатами для побудови моделей, що явно прогнозують результати для інших наборів даних, а моделі другого

типу описують залежності в наявних даних. Обидва типи моделей використовуються для прийняття управлінських рішень.
Технологія KDD дає змогу не лише підтверджувати (відкидати) емпіричні висновки, а й будувати нові, невідомі раніше моделі. Знайдена модель не зможе здебільшого претендувати на абсолютне знання, але вона надає аналітикові деякі переваги вже завдяки самому факту виявлення альтернативної статистично значущої моделі, а також, можливо, стає приводом для пошуку відповіді на запитання: чи справді існує виявлений взаємозв'язок і чи є він причинним? А це, у свою чергу, стимулює поглиблені дослідження, сприяючи глибшому розумінню досліджуваного явища.
Отже, найважливіша мета застосування технології KDD до дослідження реальних систем — це поліпшення розуміння суті їх функціонування.
Відзначимо, що процес виявлення знань не є цілком автоматизованим — він вимагає участі користувача (експерта, особи що приймає рішення). Користувач має чітко усвідомлювати, що він шукає, ґрунтуючись на власних гіпотезах. Зрештою замість того, щоб підтверджувати наявну гіпотезу, процес пошуку часто сприяє появі ряду нових гіпотез. Усе це позначається терміном «discovery-driven data mining» (DDDM), і терміни Data Mining, Knowledge Discovery у загальному випадку стосуються до технології DDDM.

Підготовка початкових даних

Процес Data Mining є свого роду дослідженням. Як будь-яке дослідження, цей процес складається з певних етапів, що включають елементи порівняння, типізації, класифікації, узагальнення, абстрагування, повторення.
процес Data Mining нерозривно пов'язаний з процесомприйняття рішень .
процес Data Mining будує модель, а в процесіприйняття рішень ця модель експлуатується.
Розглянемо традиційний процес Data Mining . Він включає наступні етапи:
аналіз предметної області ;
постановка задачі;
підготовка данних;
побудова моделей;
перевірка і оцінка моделей ;
вибір моделі;
застосування моделі;
корекція і оновлення моделі.
У цій лекції ми докладно розглянемо перші три етапи процесу Data Mining , інші етапи будуть розглянуті в наступній лекції.
Етап 1. Аналіз предметної області
Дослідження - це процес пізнання певної предметної області , об'єкта чи явища з певною метою.
Процес дослідження полягає в спостереженні властивостей об'єктів з метою виявлення і оцінки важливих, з точки зору суб'єкта-дослідника, закономірних відносин між показниками даних властивостей.

Рішення будь-якої задачі в сфері розробки програмного забезпечення повинно починатися з вивчення предметної області .
Предметна область - це подумки обмежена область реальної дійсності, що підлягає опису або моделювання та дослідження.
Предметна область складається з об'єктів, що розрізняються за властивостями і знаходяться в певних відносинах між собою або взаємодіючих яким-небудь чином.
Предметна область - це частина реального світу, вона нескінченна і містить як істотні, так і не значущі дані, з точки зору проведеного дослідження.
Досліднику необхідно вміти виділити істотну їх частину. Наприклад, при вирішенні завдання "Видавати чикредит ? "важливими є всі дані про приватне життя клієнта, аж до того, чи має роботу чоловік, чи є у клієнта неповнолітні діти, яким є рівень його освіти і т.д. Для вирішення іншої задачі банківської діяльності ці дані будуть абсолютно неважливі. Суттєвість даних, таким чином, залежить від виборупредметної області .
В процесі вивчення предметної області повинна бути створена її модель. Знання з різних джерел повинні бути формалізовані за допомогою будь-яких засобів.
Це можуть бути текстові описи предметної області або спеціалізовані графічні нотації. Існує велика кількість методик описупредметної області : наприклад, методикаструктурного аналізу SADT і заснована на ньомуIDEF0 ,діаграми потоків даних Гейне-Сарсона, методика об'єктно-орієнтованого аналізуUML та інші. Модельпредметної області описує процеси, що відбуваються впредметної області , і дані, які в цих процесах використовуються.
Це перший етап процесу Data Mining . Але від того, наскільки вірно змодельованапредметна область , залежить успіх подальшої розробки програмиData Mining .
Етап 2. Постановка завдання
Постановка задачі Data Mining включає наступні кроки:
формулювання завдання;
формалізація завдання.
Постановка завдання включає також опис статичного і динамічного поведінки досліджуваних об'єктів.
Приклад завдання. При просуванні нового товару на ринок необхідно визначити, якагрупа клієнтів фірми буде найбільш зацікавлена в даному товарі.
Опис статики на увазі опис об'єктів і їх властивостей.
Приклад. Клієнт є об'єктом. Властивості об'єкта "клієнт": сімейний стан, дохід за попередній рік,місце проживання.
При описі динаміки описується поведінка об'єктів і ті причини, які впливають на їх поведінку.
Приклад. клієнт купуєтовар А. При появі нового товару В клієнт вже не купуєтовар А, а купує тількитовар В. Поява товару В змінило поведінку клієнта. Динаміка поведінки об'єктів часто описується разом зі статикою.
технологія Data Mining не може замінити аналітика і відповісти на ті питання, які не були задані. Томупостановка задачі є необхідним етапом процесуData Mining , оскільки саме на цьому етапі ми визначаємо, яку ж завдання необхідно

вирішити. Іноді етапи аналізупредметної області і постановки завдання об'єднують в один етап.

3. Підготовка даних

Мета етапу: розробка бази даних дляData Mining .

Поняття даних було розглянуто в лекції № 2 цього курсу лекцій.

Підготовка даних є найважливішим етапом, від якості виконання якого залежить можливість отримання якісних результатів усього процесу Data Mining . Крім того, слід пам'ятати, що на етап підготовки даних, за деякими оцінками, може бути витрачено до 80% всього часу, відведеного на проект.

Розглянемо докладно, що ж являє собою цей етап.

1. Визначення та аналіз вимог до даних

На цьому етапі здійснюється так зване моделювання даних, тобто визначення та аналіз вимог до даних, які необхідні для здійснення Data Mining. При цьому вивчаються питання розподілу користувачів (географічне, організаційне, функціональне); питання доступу до даних, які необхідні для аналізу, необхідність у зовнішніх і / або внутрішніх джерелах даних; а також аналітичні характеристики системи (вимірювання даних, основні види вихідних документів, послідовність перетворення інформації та ін.).

2. Збір даних

Наявність в організації сховища даних робить аналіз простіше і ефективніше, його використання, з точки зору вкладень, обходиться дешевше, ніж використання окремих баз даних або вітрин даних. Однак далеко не всі підприємства оснащені сховищами даних. У цьому випадку джерелом для вихідних даних є оперативні, довідкові та архівні БД, тобто дані з існуючих інформаційних систем.

Також для Data Mining може знадобитися інформація з інформаційних систем керівників, зовнішніх джерел, паперових носіїв, а також знання експертів або результати опитувань.

Слід пам'ятати, що в процесі підготовки даних аналітики і розробники не повинні прив'язуватися до показників, які є в наявності, і описати максимальну кількість факторів і ознак, що впливають на аналізований процес.

На цьому етапі здійснюється кодування деяких даних. Припустимо, одним з атрибутів клієнта є рівень доходу, який повинен бути представлений в системі одним зі значень: дуже низьким, низьким, середнім, високим, дуже високим. Необхідно визначити градації рівня доходу, в цьому процесі буде потрібно співпрацю аналітика з експертом впредметної області . Можливо, для таких перетворень даних буде потрібно написання спеціальних процедур.

Визначення необхідної кількості даних

При визначенні необхідної кількості даних слід враховувати, чи є дані впорядкованими чи ні.

Якщо дані впорядковані і ми маємо справу з тимчасовими рядами, бажано знати, чи включає такий набір даних сезонну / циклічну компоненту. У разі присутності в наборі даних сезонної / циклової компоненти, необхідно мати дані як мінімум за один сезон / цикл.

Якщо дані не впорядковані, тобто події з набору даних не пов'язані за часом, в ході збору даних слід дотримуватися таких правил.

Кількість записів в наборі. Недостатня кількість записів в наборі даних може стати причиною побудови некоректною моделі. З точки зору статистики, точність моделі збільшується зі збільшенням кількості досліджуваних даних. Можливо, деякі дані є застарілими або описують якусь нетипову ситуацію, і їх потрібно виключити з бази даних. Алгоритми, що використовуються для побудови моделей на надвеликих базах даних, повинні бути масштабованими.
Співвідношення кількості записів в наборі і кількості вхідних змінних. При використанні багатьох алгоритмів необхідна певна (бажане) співвідношення вхідних змінних і кількості спостережень. Кількість записів (прикладів) в наборі даних має бути значно більше кількості чинників (змінних).
Набір даних повинен бути репрезентативним і представлятиме якомога більше можливих ситуацій. Пропорції представлення різних прикладів в наборі даних повинні відповідати реальній ситуації.

Попередня обробка даних

Аналізувати можна як якісні, так і неякісні дані. Результат буде досягнутий і в тому, і в іншому випадку. Для забезпечення якісного аналізу необхідно проведення попередньої обробки даних, яка є необхідним етапом процесу Data Mining.
оцінювання якості даних . Дані, отримані в результаті збору, повинні відповідати певним критеріям якості. Таким чином, можна виділити важливий підетапів процесу Data Mining - оцінюванняякості даних .
Якість даних (Data quality) - це критерій, який визначає повноту, точність, своєчасність і можливість інтерпретації даних.
Дані можуть бути високої якості і низької якості, останні - це так званібрудні або "погані" дані.
Дані високої якості - це повні, точні, своєчасні дані, які піддаються інтерпретації.
Такі дані забезпечують отримання якісного результату: знань, які зможуть підтримувати процес прийняття рішень.
Про важливість обговорюваної проблеми говорить той факт, що "серйозне ставлення до якості даних" займає перше місце серед десяти основних тенденцій, прогнозується на початку 2005 року в області Business Intelligence і Сховищ даних компанією Knightsbridge Solutions. Цей прогноз був зроблений в січні 2005 року, а в червні 2005 року Даффі Брансон (Duffie Brunson), один з керівників компанії Knightsbridge Solutions, проаналізував спроможність даних раніше прогнозів.
Скорочений виклад його аналізу представлено в [90]. Нижче викладено прогноз і його аналіз півроку тому.
Прогноз. Багато компаній стали звертати більше уваги наякість даних , оскільки низька якість коштує грошей в тому сенсі, що веде до зниження продуктивності, прийняття неправильних бізнес-рішень і неможливості отримати бажаний результат, а також ускладнює виконання вимог законодавства. Тому компанії дійсно мають наміру робити конкретні дії для вирішення проблемякості даних .

Реальність. Дана тенденція зберігається, особливо в індустрії фінансових послуг. В першу чергу це відноситься до фірм, які намагається виконувати угоду Basel II. Неякісні дані не можуть використовуватися в системах оцінки ризиків, які застосовуються для установки цін на кредити і обчислення потреб організації в капіталі. Цікаво відзначити, що істотно змінилися погляди на способи вирішення проблемиякості даних . Спочатку менеджери звертали основну увагу на інструменти оцінки якості, вважаючи, що "власник" даних повинен вирішувати проблему на рівні джерела, наприклад, очищаючи дані та перепідготовці співробітників. Але зараз їх погляди суттєво змінилися. поняттяякості даних набагато ширше, ніж просто їх акуратне введення в систему на першому етапі. Сьогодні вже багато хто розуміє, щоякість даних повинне забезпечуватися процесами вилучення, перетворення і завантаження (Extraction, Transformation, Loading -ETL), а також отримання даних з джерел, які готують дані для аналізу.

Розглянемо поняття якості даних більш детально.

Дані низької якості, або **брудні дані** - це відсутні, неточні або непотрібні дані з точки зору практичного застосування (наприклад, представлені в невірному форматі, який не відповідає стандарту).Брудні дані з'явилися не сьогодні, вони виникли одночасно з системами введення даних.

Брудні дані можуть з'явитися з різних причин, таким як помилка при введенні даних, використання інших форматів представлення або одиниць вимірювання, невідповідність стандартам, відсутність своєчасного оновлення, невдале оновлення всіх копій даних, невдале видалення записів-дублікатів і т.д. Необхідно оцінити вартість наявностібрудних даних ; іншими словами, наявністьбрудних даних може дійсно привести до фінансових втрат і юридичної відповідальності, якщо їх присутність не запобігає або вони не виявляються і не очищаються [91].

Для більш детального знайомства з брудними даними можна рекомендувати [92], де представлена таксономія 33 типівбрудних даних і також розроблена таксономія методів запобігання або розпізнавання іочищення даних . Описано різні типибрудних даних , серед них виділено такі групи:

- брудні дані , які можуть бути автоматично виявлені і очищені;
- дані, поява яких може бути припинено;
- дані, які непридатні для автоматичного виявлення і очищення ;
- дані, поява яких неможливо запобігти.

Тому важливо розуміти, що спеціальні засоби очищення можуть впоратися не з усіма видамибрудних даних .

Розглянемо найбільш поширені види брудних даних :

- пропущені значення ;
- дублікати даних;
- шуми і викиди .

Пропущені значення (Missing Values).

Деякі значення даних можуть бути пропущені у зв'язку з тим, що:

- дані взагалі не були зібрані (наприклад, при анкетуванні прихований вік);
- деякі атрибути можуть бути незастосовні для деяких об'єктів (наприклад, атрибут "річний дохід" непридатний до дитини).

Як ми можемо бути з пропущеними даними?

- Виключити об'єкти з пропущеними значеннями з обробки.
- Розрахувати нові значення для пропущених даних.
- нехтувати пропущені значення в процесі аналізу.
- замінити пропущені значення на можливі значення.

Дублювання даних (Duplicate Data).

Набір даних може включати продубльовані дані, тобто дублікати.

Дублікатами називаються записи з однаковими значеннями всіх атрибутів.

Наявність дублікатів в наборі даних може бути способом підвищення значущості деяких записів. Така необхідність іноді виникає для особливого виділення певних записів з набору даних. Однак в більшості випадків, продубльовані дані є результатом помилок при підготовці даних.

Як ми можемо бути з продубльованими даними?

Існує два варіанти обробки дублікатів. При першому варіанті видаляється всягрупа записів , що містить дублікати. Цей варіант використовується в тому випадку, якщо наявність дублікатів викликає недовіру до інформації, повністю її знецінює.

Другий варіант полягає в заміні групи дублікатів на одну унікальну запис.

шуми і викиди .

Викиди - різко відрізняються об'єкти або спостереження в наборі даних.

шуми і викиди є досить загальною проблемою в аналізі даних.Викиди можуть як являти собою окремі спостереження, так і бути об'єднаними в якісь групи. Завдання аналітика - не тільки їх виявити, але і оцінити ступінь їх впливу на результати подальшого аналізу. якщовикиди є інформативною частиною аналізованого набору даних, використовують робастні методи і процедури.

Досить поширена практика проведення двоетапного аналізу - з викидами і з їх відсутністю - і порівняння отриманих результатів.

Різні методи Data Mining мають різну чутливість до викидів , цей факт необхідно враховувати при виборі методу аналізу даних. Також деякі інструменти Data Mining мають вбудовані процедуриочищення від шумів івикидів .

Візуалізація даних дозволяє представити дані, в тому числі і викиди , в графічному вигляді. приклад наявностівикидів зображений на діаграмі розсіювання на рис.2.1.1 . Ми бачимо кілька спостережень, різко відрізняються від інших (які перебувають на великій відстані від більшості спостережень).

Очевидно, що результати Data Mining на основі брудних даних не можуть вважатися надійними і корисними. Однак наявність таких даних не обов'язково означає необхідність їх очищення або ж запобігання появи. Завжди повинен бути розумний вибір між наявністю брудних даних і вартістю і / або часом, необхідним для їх очищення .

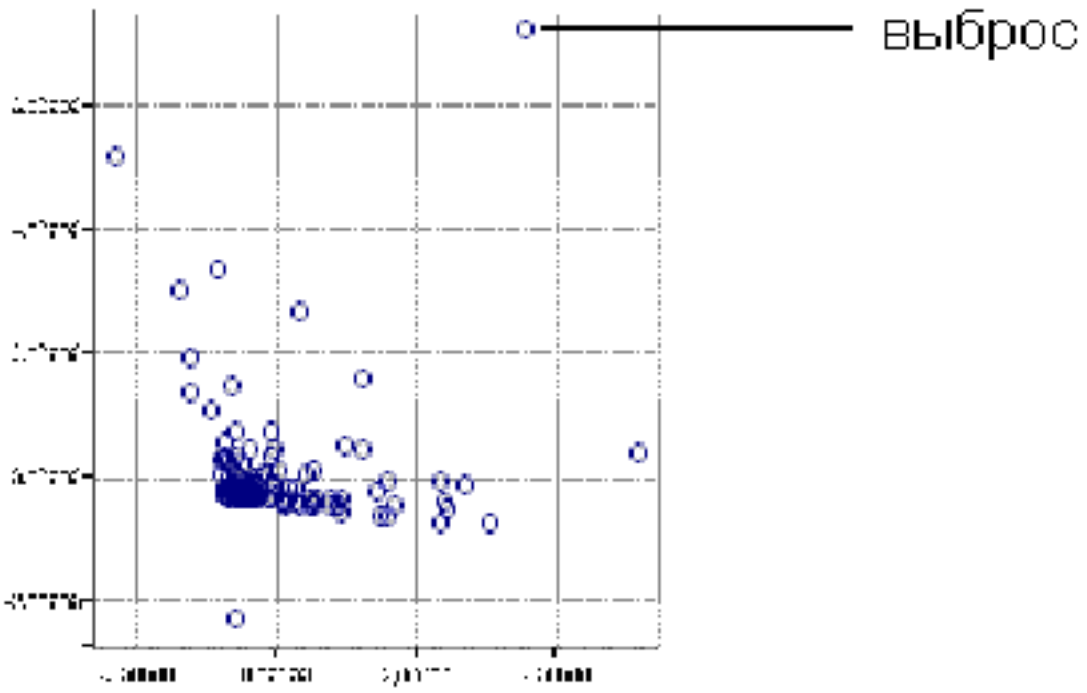


Рис. 2.1.1. Приклад набору даних з викидами

Очищення даних

Очищення даних (data cleaning, data cleansing або scrubbing) займається виявленням і видаленням помилок і невідповідностей в даних з метою поліпшенняякості даних .

Проблеми з якістю зустрічаються в окремих наборах даних - таких як файли і бази даних. Коли інтеграції підлягає безліч джерел даних (наприклад в Сховищах, інтегрованих системах баз даних або глобальних інформаційних Інтернет-системах), необхідність в очищенні даних істотно зростає. Це відбувається тому, що джерела часто містять розрізнені дані в різному поданні. Для забезпечення доступу до точних і узгодженим даними необхідна консолідація різних уявлень даних і виключення дублюється інформації. спеціальні засоби очищення зазвичай мають справу з конкретними областями - в основному це імена і адреси - або ж з виключенням дублікатів. Перетворення забезпечуються або у формі бібліотеки правил, або користувачем в інтерактивному режимі. Перетворення даних можуть бути автоматично отримані за допомогою засобів узгодження схеми .

метод очищення даних повинен задовольняти ряду критеріїв .

1. Він повинен виявляти і видаляти всі основні помилки і невідповідності, як в окремих джерелах даних, так і при інтеграції декількох джерел.
2. Метод повинен підтримуватися певними інструментами, щоб скоротити обсяги ручної перевірки і програмування, і бути гнучким в плані роботи з додатковими джерелами.
3. Очищення даних не повинно проводитися у відриві від пов'язаних зі схемою перетворення даних, які виконуються на основі складних метаданих.
4. Функції маппірования для очищення та інших перетворень даних повинні бути визначені декларативним чином і підходити для використання в інших джерелах даних і в обробці запитів.

5. Інфраструктура технологічного процесу повинна особливо інтенсивно підтримуватися для сховищ даних, забезпечуючи ефективне і надійне виконання всіх етапів перетворення для безлічі джерел і великих наборів даних.

На сьогоднішній день інтерес до очищення даних зростає. Цілий ряд дослідницьких груп займається загальними проблемами, пов'язаними з очищенням даних , в тому числі, зі специфічними підходами до Data Mining і перетворенню даних на підставі зіставлення схеми. Останнім часом деякі дослідження торкнулися єдиного, більш складного підходу до очищенні даних , що включає ряд аспектів перетворення даних, специфічних операторів і їх реалізації.

Етапи очищення даних

В цілому, очищення даних включає наступні етапи:

1. Аналіз даних.
2. Визначення порядку і правил перетворення даних.
3. Підтвердження.
4. Перетворення.
5. Противоток очищених даних.

Етап № 1. Аналіз даних .

Докладний аналіз даних необхідний для виявлення що підлягають видаленню видів помилок і невідповідностей. Тут можна використовувати як ручну перевірку даних або їх шаблонів, так і спеціальні програми для отримання метаданих про властивості даних і визначення проблем якості.

Етап № 2. Визначення порядку і правил перетворення даних .

Залежно від числа джерел даних, ступеня їх неоднорідності і забрудненості, дані можуть вимагати досить великого перетворення і очищення . Іноді для відображення джерел загальноїмоделі даних використовується трансляція схеми; для сховищ даних зазвичай використовується реляционное уявлення. Перші кроки поочищенні можуть уточнити або змінити опис проблем окремих джерел даних, а також підготувати дані для інтеграції. Подальші кроки повинні бути спрямовані на інтеграцію схеми / даних і усунення проблем множинних елементів, наприклад, дублікатів. Для сховищ в процесі роботи по визначеннюETL повинні бути визначені методи контролю і потік даних, що підлягає перетворенню іочищенні .

Перетворення даних, пов'язані зі схемою, так само як і етапи очищення , повинні, наскільки можливо, визначатися за допомогою декларативного запиту і мови маппірованія, забезпечуючи, таким чином, автоматичну генерацію коду перетворення. До того ж, в процесі перетворення повинна існувати можливість запуску написаного користувачем кодуочищення і спеціальних засобів. Етапи перетворення можуть вимагати зворотного зв'язку з користувачем за тими елементами даних, для яких відсутня вбудована логікаочищення .

Етап № 3. Підтвердження .

На цьому етапі визначається правильність і ефективність процесу і визначень перетворення. Це здійснюється шляхом тестування і оцінювання, наприклад, на прикладі або на копії даних джерела, - щоб з'ясувати, чи потрібно якось поліпшити ці визначення. При аналізі, проектуванні і підтвердженні може

знадобитися безліч ітерацій, наприклад, в зв'язку з тим, що деякі помилки стають помітні тільки після проведення певних перетворень.

Етап № 4. Перетворення .

На цьому етапі здійснюється виконання перетворень або в процесі ETL для завантаження і оновлення Сховища даних, або при відповіді на запити по безлічі джерел.

Етап № 5. Противоток очищених даних .

Після того як помилки окремого джерела видалені, забруднені дані у вихідних джерелах повинні замінитися на очищені, для того щоб поліпшені дані потрапили також в успадковані додатки і надалі при добуванні не вимагали додаткової очищення . Для сховищ очищені дані знаходяться в області зберігання даних.

Такий процес перетворення вимагає великих обсягів метаданих (схем, характеристик даних рівня схеми, визначень технологічного процесу та ін.). Для узгодженості, гнучкості та спрощення використання в інших випадках, ці метадані повинні зберігатися в репозиторії на основі СУБД. Для підтримки якості даних детальна інформація про процес перетворення повинна записуватися як в репозиторій, так і в трансформовані елементи даних, особливо інформація про повноту і свіжості вихідних даних і походження інформації про першоджерело трансформованих об'єктів і вироблених з ними зміни.

Далі детально описуються можливі методи аналізу даних (виявлення конфліктів), визначення перетворень і вирішення конфліктів. Конфлікти найменувань зазвичай вирішуються шляхом перейменування; структурні конфлікти вимагають часткового перешикування та уніфікації вихідних схем.

Таким чином, ми почали розглядати етапи процесу Data Mining , зокрема, приділили багато уваги етапу підготовки даних і їх попередній обробці, детально зупинилися на понятті брудних даних і етапах очищення даних .

Увага, відведена обговоренню цієї проблеми, викликана необхідністю використання при безпосередньому проведенні Data Mining максимально повних, точних, своєчасних даних, що піддаються інтерпретації, тобто даних високої якості .

У наступній лекції ми розглянемо інструменти очищення даних , їх сильні сторони і проблеми.

Задачі Data Mining.

Нагадаємо, що в основу технології Data Mining покладена концепція шаблонів, що представляють собою закономірності. В результаті виявлення цих, прихованих від неозброєного ока закономірностей вирішуютьсязавдання Data Mining . Різним типам закономірностей, які можуть бути виражені в формі, зрозумілою людині, відповідають певнізавдання Data Mining .

Завдання (tasks) Data Mining іноді називають закономірностями (regularity) або техніками (techniques).

Єдиної думки щодо того, які завдання слід відносити до Data Mining, немає. Більшість авторитетних джерел перераховують наступні: класифікація ,кластеризація , прогнозування , асоціація , візуалізація ,аналіз і виявлення відхилень, оцінювання , аналіз зв'язків , підведення підсумків .

Мета опису, яке слід нижче, - дати загальне уявлення прозавданнях Data Mining , порівняти деякі з них, а також представити деякі методи, за допомогою яких ці завдання вирішуються. Найбільш поширені завдання Data Mining - класифікація ,кластеризація ,асоціація ,прогнозування і візуалізація - будуть детально розглянуті в наступних лекціях.

Завдання Data Mining

Класифікація (Classification)

Короткий опис. Найбільш проста і поширена задача Data Mining. В результаті рішення задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних - класи; по цих ознаках новий об'єкт можна віднести до того чи іншого класу.

Методи рішення. Для вирішення завдання класифікації можуть використовуватися методи: найближчого сусіда (Nearest Neighbor); k-найближчого сусіда (k-Nearest Neighbor); байєсовські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks).

Кластеризація (Clustering)

Короткий опис. Кластеризація є логічним продовженням ідеї класифікації. Це завдання більш складна, особливість кластеризації полягає в тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи.

Приклад методу розв'язання задачі кластеризації: навчання "без вчителя" особливого виду нейронних мереж - самоорганізованих карт Кохонена.

Асоціація (Associations)

Короткий опис. В результаті виконання завдання пошуку асоціативних правил відшукуються закономірності між пов'язаними подіями в наборі даних.

відмінність асоціації від двох попередніх задач Data Mining : пошук закономірностей здійснюється не на основі властивостей аналізованого об'єкта, а між кількома подіями, які відбуваються одночасно.

найбільш відомий алгоритм вирішення задачі пошуку асоціативних правил - алгоритм Apriori.

Послідовність (Sequence), або послідовнаасоціація (sequential association)

Короткий опис. Послідовність дозволяє знайти тимчасові закономірності між транзакціями. завданняпослідовності подібнаасоціації , але її метою є встановлення закономірностей не між одночасно наступаючими подіями, а між подіями, пов'язаними в часі (тобто відбуваються з деяким певним інтервалом у часі). Іншими словами,послідовність визначається високою ймовірністю ланцюжка пов'язаних у часі подій. фактично,асоціація є окремим випадкомпослідовності з тимчасовим кроком, рівним нулю. цюзадачу Data Mining також називають завданням знаходження послідовних шаблонів (sequential pattern).

правило послідовності : після події X через певний час відбудеться подія Y.

Приклад. Після покупки квартири мешканці в 60% випадків протягом двох тижнів набувають холодильник, а протягом двох місяців в 50% випадків купуєтьсятелевізор . Рішення даного завдання широко застосовується в маркетингу та менеджменті, наприклад, при управлінні циклом роботи з клієнтом (Customer LifecycleManagement).

прогнозування (Forecasting)
Короткий опис. В результаті рішення задачі прогнозування на основі особливостей історичних даних оцінюються пропущені або ж майбутні значення цільових чисельних показників.
Для вирішення таких завдань широко застосовуються методи математичної статистики, нейронні мережі та ін.
Визначення відхилень або викидів (Deviation Detection),аналіз відхилень або викидів
Короткий опис. Мета рішення даного завдання - виявлення іаналіз даних , найбільш відрізняються від загальногобезлічі даних, виявлення так званих нехарактерних шаблонів.
оцінювання (Estimation)
завдання оцінювання зводиться до передбачення безперервних значень ознаки.
Аналіз зв'язків (LinkAnalysis) - задача знаходження залежностей в наборі даних.
візуалізація (Visualization ,Graph Mining)
В результаті візуалізації створюється графічний образ аналізованих даних. Для вирішення завданнявізуалізації використовуються графічні методи, що показують наявність закономірностей в даних.
приклад методів візуалізації -уявлення даних в 2D і3D вимірах.
Підведення підсумків (Summarization) - задача, мета якої - опис конкретних груп об'єктів з аналізованого набору даних.

Нечітка логіка

Нечітка логіка (англ. fuzzy logic) — розділ математики, який є узагальненням класичної логіки і теорії множин. Уперше введений Лотфі Заде в 1965 році[1][2][3] як розділ, що вивчає об'єкти з функцією належності елемента до множини, яка приймає значення у інтервалі [0, 1], а не тільки 0 або 1. На основі цього поняття вводяться логічні операції над нечіткими множинами, і формулюється поняття лінгвістичної змінної, якою виступають нечіткі множини.
Предметом нечіткої логіки вважається дослідження суджень в умовах нечіткості, які схожі з судженнями у звичайному сенсі, та їх застосування у обчислювальних системах.

Нейроні мережі
Штучна нейронна мережа (**ШНМ**, англ. artificial neural network, ANN, рос. искусственная нейронная сеть, ИНС) — це математична модель, а також її програмна та апаратна реалізація, побудовані за принципом функціювання біологічних нейронних мереж — мереж нервових клітин живого організму. Це поняття виникло при вивченні процесів, які відбуваються в мозку, та при намаганні змоделювати ці процеси. Першою такою спробою були нейронні мережі У. Маккалока та У. Піттса[en]. Після розробки алгоритмів навчання отримувані моделі стали використовуватися в

практичних цілях: в задачах прогнозування, для розпізнавання образів, в задачах керування тощо.

ШНМ являють собою систему з'єднаних між собою простих обробників (штучних нейронів), які взаємодіють. Такі обробники зазвичай є доволі простими (особливо в порівнянні з процесорами, що застосовуються в персональних комп'ютерах). Кожен обробник подібної мережі має справу лише з сигналами, які він періодично отримує, і сигналами, які він періодично надсилає іншим обробникам. І тим не менш, будучи з'єднаними в достатньо велику мережу з керованою взаємодією, такі локально прості обробники разом здатні виконувати доволі складні завдання.

- З точки зору машинного навчання, нейронна мережа є окремим випадком методів розпізнавання образів, дискримінантного аналізу, методів кластерування тощо.
- З математичної точки зору, навчання нейронних мереж — це багатопараметрична задача нелінійної оптимізації.
- З точки зору кібернетики, нейронна мережа використовується в задачах адаптивного керування, і як алгоритми для робототехніки.
- З точки зору розвитку обчислювальної техніки та програмування, нейронна мережа — спосіб розв'язання задачі ефективного паралелізму.
- А з точки зору штучного інтелекту, ШНМ є основою філософської течії коннективізму[en] й основним напрямком в структурному підході до вивчення можливості побудови (моделювання) природного інтелекту за допомогою комп'ютерних алгоритмів.

Нейронні мережі не програмуються в звичайному розумінні цього слова, вони **навчаються**. Можливість навчання — одна з головних переваг нейронних мереж перед традиційними алгоритмами. Технічно, навчання полягає в знаходженні коефіцієнтів зв'язків між нейронами. В процесі навчання нейронна мережа здатна виявляти складні залежності між вхідними даними й вихідними, а також здійснювати узагальнення. Це означає, що в разі успішного навчання мережа зможе повернути правильний результат на підставі даних, які були відсутні в навчальній вибірці, а також неповних та/або «зашумлених», частково спотворених даних.

Лекція № 2.2. Стандарти та інструменти Data Mining

CRISP-DM методологія

Ми розглянули процес Data Mining з двох сторін: як послідовність етапів і як послідовністьробіт , виконуваних виконавцями ролейData Mining .

Існує ще одна сторона - це стандарти, що описують методологію Data Mining . Останні розглядають організацію процесуData Mining і розробкуData Mining -систем.

CRISP-DM (The CrossIndustrie Standard Process forData Mining - Стандартний міжгалузевої процесData Mining) є найбільш популярною і поширеною методологією. членами консорціумуCRISP-DM є NCR, SPSS та DaimlerChrysler.

Відповідно до стандарту CRISP , **Data Mining є безперервним процесом з багатьма циклами і зворотними зв'язками** .
Data Mining по стандартуCRISP-DM включає наступні фази:

1. Осмислення бізнесу (Business understanding).
2. Осмислення даних (Data understanding).
3. Підготовка даних (Data preparation).
4. Моделювання (Modeling).
5. Оцінка результатів (Evaluation).
6. Впровадження (Deployment).

До цього набору фаз іноді додають сьомий крок - Контроль , він закінчує коло. фазиData Mining по стандартуCRISP-DM зображені на рис. 2.2.1 .

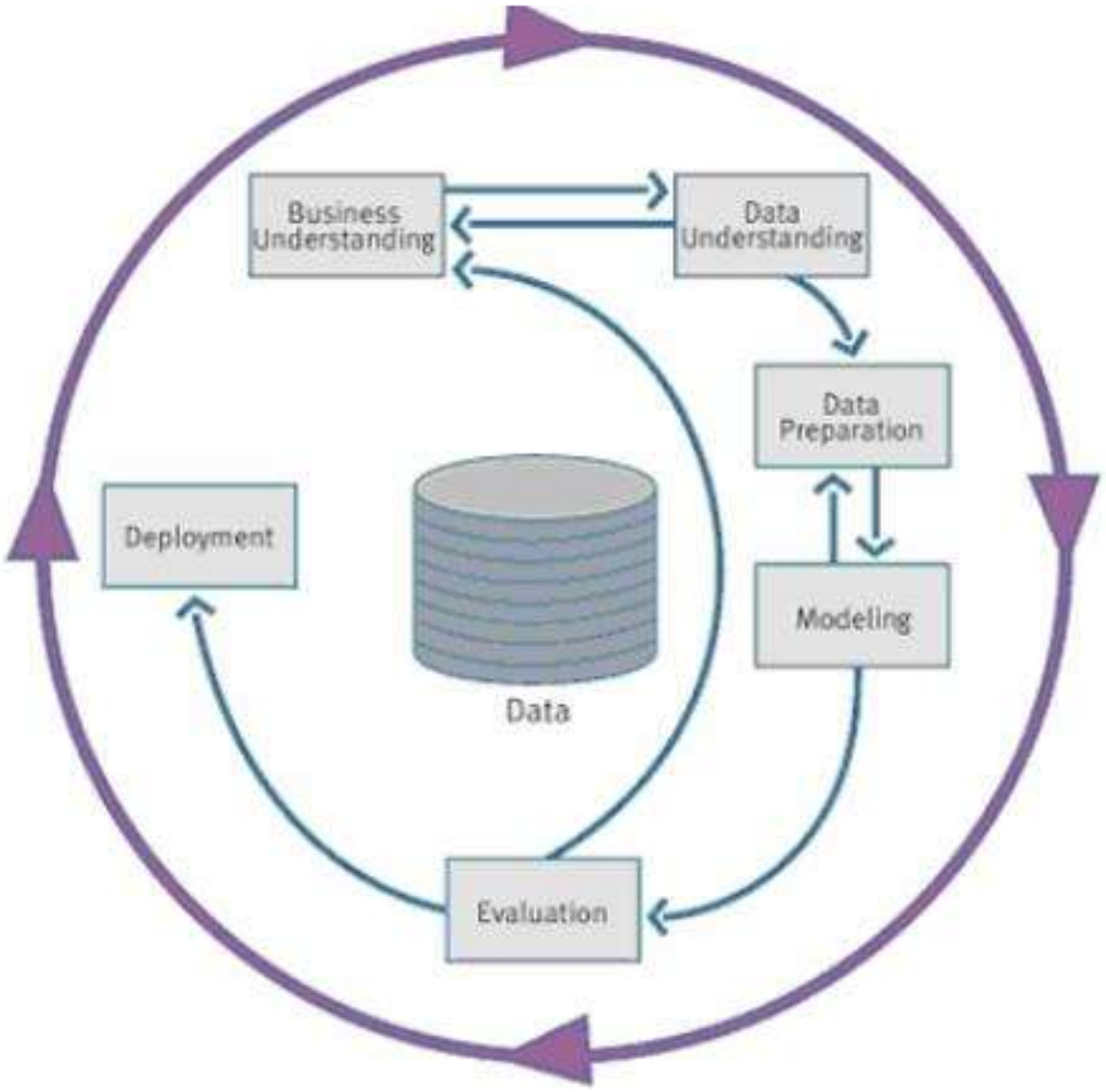


Рис. 2.2.1. Фази, рекомендовані моделлю CRISP-DM

За допомогою методології CRISP-DM Data Mining перетворюється в бізнес-процес , в ході якого технологія Data Mining фокусується на вирішенні конкретних проблем бізнесу. Методологія CRISP-DM , яка розроблена експертами в індустрії Data Mining , являє собою покрокове керівництво, де визначені завдання і цілі для кожного етапу процесуData Mining .
Методологія CRISP-DM описується в термінах ієрархічного моделювання процесу , який складається з набору завдань, описаних чотирма рівнями узагальнення (від загальних до специфічних): фази, спільні завдання, спеціалізовані завдання і запити.
На верхньому рівні процес Data Mining організовується в певну кількість фаз , на другому рівні кожна фаза розділяється на кілька загальних

завдань . Завдання другого рівня називаються загальними, тому що вони є позначенням (плануванням) досить широких завдань, які охоплюють всі можливіData Mining -ситуація. Третій рівень є рівнем спеціалізації завдання , тобто тим місцем, де дії загальних завдань переносяться на конкретні специфічні ситуації. Четвертий рівень є звітом по дій, рішень і результатів фактичного використанняData Mining .

CRISP-DM - це не єдиний стандарт, що описує методологіюData Mining . Крім нього, можна застосовувати такі відомі методології, що є світовими стандартами, як Two Crows,SEMMA , а також методології організації або свої власні.

SEMMA методологія

SEMMA методологія реалізована в середовищіSAS Data Mining Solution (SAS) . Її абревіатура утворена від слів Sample ("Відбір даних", тобто створення вибірки),Explore ("Дослідження відносин в даних"),Modify ("Модифікація даних"),Model ("Моделювання взаємозалежностей "),Assess ("Оцінка отриманих моделей і результатів").Методологія розробки проектуData Mining відповідно до методологіїSEMMA зображена на рис.2.2.2.

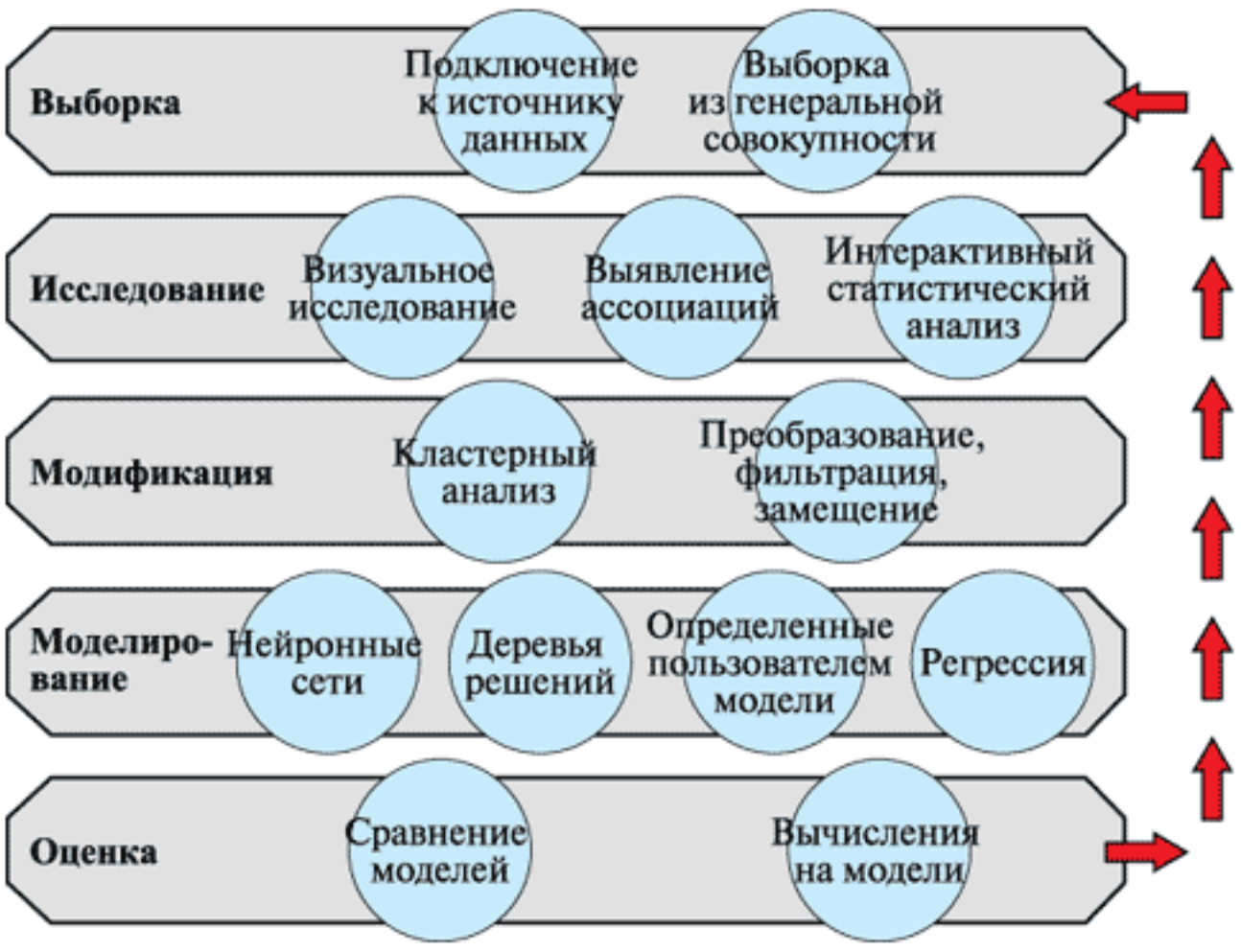


Рис.2.2.2. Методологія розробки проекту Data Mining відповідно до методології SEMMA

Підхід SEMMA має на увазі, що всі процеси виконуються в рамках гнучкої оболонки, що підтримує виконання всіх необхіднихробіт по обробці та аналізу даних. підхідSEMMA поєднує структурованість процесу і логічну організацію інструментальних засобів, що підтримують виконання кожного з кроків. завдякидіаграм процесів обробки даних , підхідSEMMA спрощує застосування методів статистичного дослідження і візуалізації, дозволяє вибирати і перетворювати найбільш значущі змінні, створювати моделі з цими

змінними, щоб передбачити результати, підтвердититочність моделі і підготувати модель до розгортання.
Ця методологія не нав'язує жодних жорстких правил. В результаті використання методологіїSEMMA розробник може мати у своєму розпорядженні науковими методами побудови концепції проекту, його реалізації, а також оцінки результатів проектування.
За результатами останніх опитувань KDnuggets (2004), 42% опитаних осіб використовує методологію CRISP-DM , 10% - методологію SEMMA , 6% - власну методологію організації, 28% - свою власну методологію, іншими методологіями користується 6% опитаних, не користуються ніякою методологією 7% опитаних.

Інші стандарти Data Mining

Як уже зазначалося, описані стандарти є методологіями Data Mining , тобто розглядають організацію процесу і розробку системData Mining . Крім цієї групи, в останні роки з'явився ряд стандартів, мета яких - узгодити досягнення вData Mining , спростити управління моделюванням процесів і подальше використання створених моделей. Ці стандарти умовно можна поділити на дві категорії:

1. Стандарти, які стосуються вироблення єдиної угоди зі зберігання і передачі моделей Data Mining .
2. Стандарти, які стосуються уніфікації інтерфейсів.

стандарт PMML

У попередніх лекціях ми вже згадували про стандарт PMML (Predictive Modeling mark-upLanguage) - мовою опису предікторних (або прогнозних) моделей або мовою розмітки для прогнозного моделювання.
PMML відноситься до групи стандартів по зберіганню і передачі моделей Data Mining .
Розробка і впровадження цього стандарту ведеться IT-консорціумом DMG (Data Mining Group). DMG [103] -група , в яку входять всі лідируючі компанії, які розробляютьпрограмне забезпечення в області аналізу даних.
Основа цього стандарту - мова XML . Прикладом іншого стандарту, також заснованого на мовіXML , є стандарт обміну статистичними даними та метаданими. стандартPMML використовується для опису моделейData Mining і статистичних моделей.
Основна мета стандарту PMML - забезпечення можливості обміну моделями даних між програмним забезпеченням різних розробників.
За допомогою стандарту PMML сумісні додатки можуть легко обмінюватися моделями даних з іншимиPMML -Інструменти. Таким чином, модель, створена в одному програмному продукті, може використовуватися для прогнозного моделювання в іншому.
За словами прихильниківPMML , цей стандарт "робитьData Mining демократичнішим ", дозволяє все великій кількості користувачів користуватися продуктамиData Mining . Це досягається за рахунок можливості використання раніше створених моделей даних.PMML дозволяє

використовувати моделі даних як завгодно часто і суттєво допомагає в практичній роботі з ними.

стандарт PMML включає:

- опис аналізованих даних (структура і типи даних);
- опис схеми аналізу (використовувані поля даних);
- опис трансформацій даних (наприклад, перетворення типів даних);
- опис статистик, прогнозованих полів і самих прогнозних моделей.

стандарт PMML забезпечує підтримку найбільш поширених прогнозних моделей, створених за допомогою алгоритмів і методів аналізу даних, зокрема - нейронних мереж, дерев рішень, алгоритмів асоціативних правил, кластерного аналізу, логічних правил і ін.

Стандарти, які стосуються уніфікації інтерфейсів

За допомогою стандартів цієї групи будь додаток може отриматидоступ до функціональностіData Mining . Тут можна виділити стандарти, спрямовані на стандартизацію інтерфейсів дляоб'єктних мов програмування, і стандарти, спрямовані на розробку надбудови над мовоюSQL .

До стандартам, спрямованим на стандартизацію інтерфейсів для об'єктних мов програмування, можна віднести: **CWM Data Mining** , **JDM** .

У 2000 році організації MDC (MetaData Coalition, www.mdcinfo.com) iOMG (Object Management Group , www.omg .org), які розробляють два конкуруючих стандарту - в області інтелектуальних технологій для бізнесу - **OIM** (Open Information Model) і **CWM** (Common Warehouse Metamodel) - загальнуметамодель сховищ даних вирішили об'єднати свої досягнення і зусилля під управліннямOMG . Стандарт CWM включає опис базових елементів об'єктної моделі, реляційних відносин, мовиXML , структури семантикипредметної області , архітектуриOLAP , видобутку даних, технології перевантаження даних і деяких розширень.

JDM (TheJava Data Mining standard -Java Specification Request 73,JSR - 73). Стандарт, розроблений групоюJSR 73,Java Data Mining API (JDM) - це перша спроба створити стандартнийJava API (програмнийінтерфейс програми) для отримання доступу до інструментівData Mining зJava -додатків.

друга група стандартів спрямована на розробку надбудови над мовоюSQL , яка дозволяла б звертатися до інструментаріюData Mining , вбудованому безпосередньо в реляційну базу даних. До цієї групи можна віднести такі стандарти:SQL / MM ,OLE DB forData Mining .

Стандарт SQL / MM являє собою набір певних користувачемSQL процедур для можливостей обчислень і використань моделейData Mining .

The OLE DB for Data Mining standard of Microsoft . Цей стандарт дозволяє, подібноSQL / MM , застосовувати методиData Mining в структурі реляційних баз даних. Цей стандарт є розширеннямOLE DB.

Стандарти, що мають пряме або опосередковане ставлення доData Mining , можна об'єднати в групи:

- стандарти, які базуються на послуги Data Mining (послуги створення моделі управління, скорингові послуги, послуги аналізу даних, послуги дослідження даних, статистичні послуги моделювання);

- стандарти web-служби (SOAP / XML, WSRF, і т.д), Grid-Послуги (OGSA, OGSA / DAI, і т.д.), Семантичні Стандарти Web (RDF, OWL, і т.д.);
- стандарти, які повинні з'явитися найближчим часом: стандарти для технологічного процесу, стандарти для перетворень даних, стандарти для оперативного (real time) Data Mining , стандарти для мереж даних (data webs).

Як ми бачимо, стандарти Data Mining розвиваються, з'являються також нові, що мають як пряме, так і опосередкованеставлення до цієї технології. Це свідчить про достатню "зрілості"Data Mining і вступ її в новий етап розвитку.

Лекція № 2.3. OLAP-системи

OLAP (англ. online analytical processing, аналітична обробка в реальному часі) — це технологія обробки інформації, що дозволяє швидко отримувати відповіді на багатовимірні аналітичні запити. OLAP є частиною такого ширшого поняття, як бізнес-аналітика, що також включає такі дисципліни як реляційна звітність та добування данних (спосіб аналізу інформації в базі даних з метою відшукання аномалій та трендів без з'ясування смислового значення записів). Служить для підготовки бізнес-звітів з продажів, маркетингу, для потреб управління, для прогнозування, фінансової звітності та в схожих областях.

Бази даних, сконфігуровані для OLAP, використовують багатовимірні моделі даних, що дозволяє виконувати складні аналітичні та спеціалізовані запити за короткий проміжок часу. Вони запозичують окремі аспекти навігаційних та ієрархічних баз даних, які є швидшими за реляційні БД.

Зазвичай результати OLAP-запитів представляють у формі матриць, де виміри складають рядки та колонки, а значеннями матриці є розміри.

Головна причина використання OLAP для обробки запитів — це швидкість. Реляційні БД зберігають сутності в окремих таблицях, які зазвичай добре нормалізовані. Ця структура зручна для операційних БД (системи OLTP), але складні багатотабличні запити в ній виконуються відносно повільно. Зручнішою моделлю для виконання запитів (але не для внесення змін) є просторова БД. OLAP робить миттєвий знімок реляційної БД і структурує її в просторову модель для запитів. Заявлений час обробки запитів в OLAP становить близько 0,1% від аналогічних запитів до реляційної БД.

Концепція OLAP

Ядром будь-якої OLAP-системи є ідея OLAP-куба (багатовимірний куб, або гіперкуб). OLAP-структура, створена з робочих даних, називається OLAP-кубом. Він складається з чисельних фактів (розмірів), розподіленних за вимірами. Зазвичай куб створюється за допомогою з'єднання таблиць із застосуванням схеми «зірка», або схеми «сніжинка». В центрі «зірки» знаходиться таблиця, яка містить ключові факти, за якими робляться запити. Множинні таблиці з вимірами приєднані до таблиці фактів. Ці таблиці показують, як можуть аналізуватися агреговані реляційні дані. Кількість можливих агрегацій визначається кількістю способів, якими первинні дані можуть бути ієрархічно відображені. Наприклад, всі клієнти можуть бути

згруповані за містами, або за регіонами країни (Захід, Схід, Північ і т. д.), таким чином, 50 міст, 8 регіонів і 2 країни складуть 3 рівні ієрархії з 60 членами. Також клієнти можуть бути об'єднані за відношенням до продукції; якщо існують 250 продуктів у двох категоріях, 3 групи продукції і 3 виробничих підрозділи, то кількість агрегатів складе 16560. При додаванні вимірів в схему, кількість можливих варіантів швидко досягає десятків мільйонів і більше.

OLAP-куб містить в собі базові дані і інформацію про вимірювання (агрегати). Куб потенційно містить всю інформацію, яка може виявитися необхідною для відповідей на будь-які запити. Через величезну кількість агрегатів, часто повний розрахунок відбувається тільки для деяких вимірювань, для останніх же проводиться «на вимогу».

Типи

Традиційно OLAP-системи поділяють на такі види:

- OLAP з багатьма вимірюваннями (Multidimensional OLAP), MOLAP;
- реляційна OLAP (Relational OLAP), ROLAP;
- гібридна OLAP (Hybrid OLAP), HOLAP.

MOLAP це класична форма OLAP, так що її часто називають просто OLAP. Вона використовує підсумовуючу БД, спеціальний варіант процесора просторових БД і створює необхідну просторову схему даних зі збереженням як базових даних, так і агрегатів. ROLAP працює безпосередньо з реляційним сховищем, факти і таблиці з вимірюваннями зберігаються в реляційних таблицях, і для зберігання агрегатів створюються додаткові реляційні таблиці. HOLAP використовує реляційні таблиці для зберігання базових даних і багатовимірні таблиці для агрегатів. Особливим випадком ROLAP є ROLAP реального часу (Real-time ROLAP, або R-ROLAP). На відміну від ROLAP, в R-ROLAP для зберігання агрегатів не створюються додаткові реляційні таблиці, а агрегати розраховуються у момент запиту. При цьому багатовимірний запит до OLAP-системи автоматично перетвориться в SQL-запит до реляційних даних.

Кожен тип зберігання має певні переваги, хоча є розбіжності в їх оцінці у різних виробників. MOLAP краще всього підходить для невеликих наборів даних, він швидко розраховує агрегати і дає відповіді, але при цьому генеруються величезні обсяги даних. ROLAP оцінюється як більш масштабоване рішення, яке до того ж використовує найменший можливий простір. При цьому швидкість обробки значно знижується. HOLAP знаходиться між цими двома підходами, він досить добре масштабується і швидко обробляється. Архітектура R-ROLAP дозволяє проводити багатовимірний аналіз OLTP-даних в режимі реального часу.

Складність в застосуванні OLAP полягає в створенні запитів, виборі базових даних і розробці схеми, внаслідок чого більшість сучасних продуктів OLAP поставляються разом з величезною кількістю заздалегідь сконфігурованих запитів. Інша проблема полягає в базових даних. Вони повинні бути повними і несуперечливими.

Реалізації OLAP

Першим продуктом, що виконував OLAP-запити, був Express (компанія IRI). Проте сам термін OLAP був запропонований «батьком реляційних БД» Едгаром Коддом. А робота Кодда фінансувалася Arbor, компанією, що випустила свій власний OLAP-продукт Essbase роком раніше (пізніше куплений Hyperion, яка в 2007 р. була поглинена компанією Oracle). Як результат, «OLAP» Кодда з'явився в їх описі Essbase.

Інші добре відомі OLAP-продукти включають Microsoft Analysis Services (що раніше називалися OLAP Services, частина SQL Server), DB2 OLAP Server від IBM (фактично, EssBase з доповненнями від IBM), продукти MicroStrategy і інших виробників.

З технічної точки зору, представлені на ринку продукти діляться на «**фізичний OLAP**» і «**віртуальний**».

У першому випадку наявна програма, що виконує попередній розрахунок агрегатів, які потім зберігаються в спеціальній багатовимірній БД, що забезпечує швидкий доступ. Приклади таких продуктів: Microsoft Analysis Services, Oracle OLAP Option, Oracle/Hyperion EssBase, Cognos PowerPlay.

У другому випадку дані зберігаються у реляційних СУБД, а агрегати можуть не існувати взагалі або створюватися за першим запитом у СУБД або кеші аналітичного ПО. Приклади таких продуктів: SAP BW, BusinessObjects, Microstrategy.

Системи, що мають в своїй основі «фізичний OLAP» забезпечують стабільно кращий час відгуку на запити, ніж системи «віртуальний OLAP». Постачальники систем «віртуальний OLAP» заявляють про більшу масштабованість їх продуктів в плані підтримки дуже великих обсягів даних.

З погляду користувача обидва варіанти виглядають схожими за можливостями.

Найбільше застосування OLAP знаходить в продуктах для бізнес-планування і сховищах даних.

Сховище даних

Сховище даних (Data Warehouse) - предметно - орієнтований,інтегрований , прив'язаний до часу і незмінний набір даних, призначений дляпідтримки прийняття рішень .

Сховище даних містить несуперечливі консолідовані історичні дані і надає інструментальні засоби для їханалізу з метоюпідтримки прийняття стратегічних рішень. Інформаційні ресурсисховища даних формуються на основі фіксованих протягом тривалого періоду часу моментальних знімків баз даних оперативноїінформаційної системи і, можливо, різних зовнішніх джерел. У сховищах даних застосовуються технології баз даних,OLAP , глибинногоаналізу даних ,візуалізації даних.

Основні характеристики сховищ даних.

- містить історичні дані;
- зберігає докладні відомості, а також частково і повністю узагальнені дані;
- дані в основному є статичними;
- нерегламентований, неструктурований і евристичний спосіб обробки даних;

- середня і низька інтенсивність обробки транзакцій ;
- непередбачуваний спосіб використання даних;
- призначене для проведення аналізу ;
- орієнтоване на предметні області ;
- підтримка прийняття стратегічних рішень;
- обслуговує відносно мала кількість працівників керівної ланки.

термін OLAP (On-LineAnalytical Processing) служить для опису моделі представлення даних і відповідно технології їх обробки в сховищах даних. BOLAP застосовується багатовимірнеуявлення агрегованих даних для забезпечення швидкого доступу до стратегічно важливої інформації з метою поглибленогоаналізу . додаткиOLAP повинні володіти наступними основними властивостями:

- багатовимірне представлення даних ;
- підтримка складних розрахунків;
- правильний облік фактора часу.

переваги OLAP :

- підвищення продуктивності виробничого персоналу, розробниківприкладних програм . Своєчасний доступ до стратегічної інформації.
- надання користувачам достатніх можливостей для внесення власних змін в схему.
- додатки OLAP спираються насховища даних і системи OLTP , отримуючи від них актуальні дані, що дає збереженняконтролю цілісності корпоративних даних.
- зменшення навантаження на системи OLTP ісховища даних .
-

Характеристика та основні відмінності OLAP і OLTP

Сховище даних має включати як внутрішні корпоративні дані, так і зовнішні дані

обсяг аналітичних БД якмінімум на порядок більше обсягу оперативних. для проведення достовірниханалізу і прогнозування всховище даних потрібно мати інформацію про діяльністькорпорації і стан ринку протягом декількох років

Сховище даних має містити одноманітно представлену і узгоджену інформацію, максимально відповідає змісту оперативнихБД . Необхідна компонента для вилучення і "очищення" інформації з різних джерел. У багатьох великих корпораціях одночасно існують кілька оперативних ІС з власнимиБД (за історичнимипричин).

Набір запитів до аналітичної бази даних передбачити неможливо. сховища даних існують, щоб відповідати на нерегламентовані запити аналітиків. Можна розраховувати тільки на те, що запити будуть надходити не надто часто і зачіпати великі обсяги інформації. розміри аналітичноїБД стимулюють використання запитів з агрегатами (сума, мінімальне, максимальне,середнє значення і т.д.)
При малої мінливості аналітичних БД (тільки призавантаженні даних) виявляються розумними впорядкованість масивів, більш швидкі методиіндексації при масової вибірці, зберігання заздалегідь агрегованих даних
інформація аналітичних БД настільки критична для корпорації , що потрібні велика грануляція захисту (індивідуальні права доступу до певних рядках і / або стовпцями таблиці)

основним джерелом інформації, що надходить в оперативну БД , єдіяльність корпорації , а для проведенняаналізу даних потрібне залучення зовнішніх джерел інформації (наприклад, статистичнихзвітів)
Для оперативної обробки потрібні дані за кілька останніх місяців
оперативні БД можуть містити семантичноеквівалентну інформацію, представлену в різних форматах, з різними зазначенням часу її надходження, іноді навіть суперечливу
Системи обробки даних створюються в розрахунку на рішення конкретних завдань. інформація зБД вибирається часто і невеликими порціями. Зазвичай набір запитів до оперативноїБД відомий вже при проектуванні
Системи обробки даних за своєю природою є сильно мінливими, що враховується в використовуваних СУБД (нормалізована структураБД , рядки зберігаються неупорядоченно, B-дерева дляіндексації ,транзакційність)
Для систем обробки даних зазвичай вистачає захи таблиць

Правила Кодда для OLAP систем

У 1993 році Кодд опублікував працю під назвою "OLAP для користувачів-аналітиків: яким він повинен бути ". У ньому він виклав основні концепції оперативної аналітичної обробки і визначив 12 правил, яким повинні задовольняти продукти, що надають можливість виконання оперативної аналітичної обробки.

1. Концептуальне багатовимірне уявлення. OLAP -модель повинна бути багатовимірної в своїй основі. Багатовимірна концептуальна схема або призначене для користувача подання полегшують моделювання тааналіз так само, втім, як іобчислення .
2. Прозорість. Користувач здатний отримати всі необхідні дані зOLAP -машини, навіть не підозрюючи, звідки вони беруться. Незалежно від того, єOLAP -продукт частиною коштів користувача чи ні, цей факт повинен бути непомітний для користувача. якщоOLAP надаєтьсяклієнт -серверними обчисленнями, то цей факт також, по можливості, повинен бути невидимий для користувача.OLAP повинен надаватися вконтексті істинно відкритої архітектури, дозволяючи користувачеві, де б він не знаходився, зв'язуватися за допомогою аналітичного інструменту з сервером. На додаток до цього прозорість повинна досягатися і при взаємодії аналітичного інструмента з гомогенної ігетерогенної середовищамиБД .
3. Доступність. OLAP повинен надавати свою власнулогічну схему для доступу вгетерогенної середовищіБД і виконувати відповідні перетворення для надання даних користувачеві. Більш того, необхідно заздалегідь подбати про те, де і як, і які типи фізичної організації даних дійсно будуть використовуватися.OLAP -система повинна виконувати доступ тільки до дійсно потрібними даними, а не застосовувати загальний принцип "кухонної воронки", який тягне непотрібний введення.
4. Постійна продуктивність при
розробцізвітів .продуктивність формуваннязвітів не повинна істотно падати з ростом кількості вимірювань і розмірів бази даних.
5. Клієнт -серверная архітектура. Потрібно, щоб продукт був не тількиклієнт -серверним, але і щоб серверний компонент був би досить інтелектуальним для того, щоб різні клієнти могли підключатися змінімумом зусиль і програмування.
6. Загальна багатовимірність. Всі вимірювання повинні бути рівноправні, кожний вимір має бутиеквівалентно і в структурі, і в операційних можливостях. Правда, допускаються додаткові операційні можливості для окремих вимірів (мабуть, мається на увазі час), але такі додаткові функції повинні бути надані будь-якому вимірюванню. Не повинно бути так, щоб базовіструктури даних , обчислювальні або звітні формати були більш властиві якомусь одному вимірюванню.
7. динамічне управління розрідженими матрицями .OLAP системи повинні автоматично налаштовувати свою фізичну схему в залежності відтипу моделі , обсягів даних і розрідженості бази даних.

8. Розрахована на багато користувачів підтримка .OLAP -Інструмент повинен надавати можливостіспільного доступу (запиту і доповнення),цілісності і безпеки.
9. Необмежені перехресні операції. Всі види операцій повинні бути дозволені для будь-яких вимірювань.
10. Інтуїтивна маніпуляція даними. Маніпулювання даними здійснювалося за допомогою прямих дій над осередками в режимі перегляду без використання меню і множинних операцій.
11. Гнучкі можливості отримання звітів . Виміри повинні бути розміщені в звіті так, як це потрібно користувачеві.
12. необмежена розмірність і число рівнівагрегації . Дослідження про можливе числі необхідних вимірювань, потрібних в аналітичній моделі, показало, що одночасно може використовуватися до 19 вимірювань. Звідси випливає нагальна рекомендація, щоб аналітичний інструмент був здатний одночасно надати якмінімум 15 вимірювань, а краще 20. Більш того, кожне з загальних вимірювань не повинно бути обмежене за кількістю визначених користувачем-аналітиком рівнів агрегації і шляхівконсолідації .

Основні елементи і операції OLAP

В основі OLAP лежить поняттягиперкуба , або багатовимірногокуба даних , восередках якого зберігаються аналізовані дані.

Факт - це числова величина яка розташовується восередках гиперкуба . ОдинOLAP -куб може мати одну або декілька показниками.

Вимірювання (dimension) - цебезліч об'єктів одного або декількох типів, організованих у вигляді ієрархічної структури і забезпечують інформаційнийконтекст числового показника. Вимірювання прийнято візуалізувати у виглядіребра багатовимірного куба.

Об'єкти, сукупність яких і утворює вимір, називаються членами вимірювань (members). Члени вимірювань візуалізують як точки або долі, що відкладаються на осяхгиперкуба .

Осередок (cell) -атомарна структура куба, відповідна повного набору конкретний значень вимірів.

Ієрархія - групування об'єктів одного виміру в об'єкти більш високого рівня. Наприклад - день-місяць-рік. Ієрархії в вимірах необхідні для можливості агрегації ідеталізації значень показників згідно їх ієрархічній структурі.Ієрархія цілком грунтується на одному вимірі і формується з рівнів.

В OLAP -системи підтримуються наступні базовіоперації :

- поворот;
- проекція . припроекції значення восередках , що лежать на осіпроекції , підсумовуються по деякому зумовленої законом;
- розкриття (drill -down). Одне зі значень вимірювання замінюється сукупністю значень з наступного рівняієрархії вимірювання ; відповідно замінюються значення восередках гиперкуба ;
- згортка (roll-up /drill -up). Операція, зворотна розкриттю;
- перетин (slice -and-dice).

Типи OLAP. Переваги і недоліки

Вибір способу зберігання даних залежить від обсягу і структури детальних даних, вимог до швидкості виконання запитів і частоти поновленняOLAP - куб. В даний час застосовуються три способи зберігання даних :

MOLAP (Multidimensional OLAP)

детальні і агреговані дані зберігаються в багатовимірної базі даних.Зберігання даних в багатовимірних структурах дозволяє маніпулювати даними як багатовимірним масивом, завдяки чому швидкістьобчислення агрегатних значень однакова для будь-якого з вимірів. Однак в цьому випадку багатовимірна база даних виявляєтьсянадлишкової , так як багатовимірні дані повністю містять детальні реляційні дані.

переваги MOLAP .

- висока продуктивність . Пошук івибірка даних здійснюється значно швидше, ніж при багатовимірному концептуальному погляді нареляційну базу даних .
- Структура і інтерфейси найкращим чином відповідають структурі аналітичних запитів.
- багатовимірні СУБД легко справляються з завданнями включення в інформаційну модель різноманітнихвбудованих функцій .

недоліки MOLAP .

- MOLAP можуть працювати тільки зі своїми власними багатовимірнимиБД і ґрунтуються на патентованих технологіях для багатовимірнихСУБД , тому є найбільш дорогими. Ці системи забезпечують повний циклOLAP -оброблення та або включають в себе, крім серверного компонента, власнийінтегрований клієнтський інтерфейс, або використовують для зв'язку з користувачем зовнішні програми роботи з електронними таблицями.
- У порівнянні з реляційними, дуже неефективно використовують зовнішню пам'ять , мають гірші в порівнянні з реляційнимиБД механізмамитранзакцій .
- Відсутні єдині стандарти на інтерфейс, мови опису і маніпулювання даними.
- Не підтримують реплікацію даних, часто використовується в якості механізму завантаження.

ROLAP (Relational OLAP)

ROLAP -системи дозволяють представляти дані, що зберігаються в класичній реляційної базі, в багатовимірної формі або в плоских локальних таблицях нафайл-сервері , забезпечуючи перетворення інформації вбагатовимірну модель через проміжний шар метаданих . Агрегати зберігаються в тій же БД в спеціально створених службових таблицях. В цьому випадку гіперкуб емулюєтьсяСУБД на логічному рівні.

Переваги ROLAP .

- реляційні СУБД мають реальний досвід роботи з дуже великимиБД і розвинені засоби адміністрування . При використанніROLAP розмір сховища не є таким критичним параметром, як у випадкуMOLAP .

- При оперативної аналітичної обробки вмісту сховища даних інструменти ROLAP дозволяють виробляти аналіз безпосередньо над сховищем (бо в переважній більшості випадків корпоративнісховища даних реалізуються засобами реляційнихСУБД).
- У разі змінної розмірності задачі, коли зміни в структуру вимірювань доводиться вносити досить часто,ROLAP системи здинамічним поданням розмірності єоптимальним рішенням, так як в них такі модифікації не вимагають фізичної реорганізаціїБД , як у випадкуMOLAP .
- системи ROLAP можуть функціонувати на набагато менш потужних клієнтських станціях, ніж системиMOLAP , оскільки основна обчислювальна навантаження в них лягає на сервер, де виконуються складні аналітичні SQL-запити , що формуються системою.
- реляційні СУБД забезпечують значно вищий рівень захисту даних і хороші можливості розмежування прав доступу.

Недоліки ROLAP .

- Обмежені можливості з точки зору розрахунку значень функціонального типу.
- менша продуктивність , ніж уMOLAP . Для забезпечення порівнянної зMOLAP продуктивності реляційні системи вимагають ретельного опрацювання схемиБД та спеціальної настройки індексів. Але в результаті цих операційпродуктивність добре налаштованих реляційних систем при використанні схеми "зірка "можна порівняти з продуктивністю систем на основі багатовимірнихБД .

HOLAP (Hybrid OLAP)

Детальні дані залишаються в тій же реляційної базі даних, де вони спочатку знаходилися, а агрегатні дані зберігаються в багатовимірної базі даних.

Моделювання багатовимірних кубів на реляційної моделі даних

Схема зірка. Переваги і недоліки

Схема типу зірки (Star Schema) -схема реляційної бази даних , що служить дляпідтримки багатовимірного уявлення містяться в ній даних.

Особливості ROLAP -схеми типу "зірка "

1. Одна таблиця фактів (fact table), яка сильно денормалізована. Є центральною в схемі, може складатися з мільйонів рядків і містить підсумовувані або фактичні дані, з допомогою яких можна відповісти на різні питання.
2. Кілька денормалізованних таблиць вимірів (dimensional table). Мають меншу кількість рядків, ніж таблиці фактів, і містять описову інформацію. Ці таблиці дозволяють користувачеві швидко переходити від таблиці фактів до додаткової інформації.
3. Таблиця фактів і таблиці розмірності пов'язані ідентифікують зв'язками, при цьому первинні ключі таблиці розмірності мігрують в таблицю фактів як зовнішніх ключів .Первинний ключ таблиці факту цілком складається зпервинних ключів всіх таблиць розмірності .
4. Агреговані дані зберігаютьсяспільно з вихідними.

- переваги

завдяки денормалізації таблиць вимірів спрощується сприйняття структури даних користувачем і формулювання запитів, зменшується кількість операцій з'єднання таблиць при обробці запитів. деякі промисловіСУБД і інструментикласу OLAP /Reporting вміють використовувати переваги схеми "зірка "для скорочення часу виконання запитів.

- недоліки

Денормализация таблиць вимірів вносить надмірність даних, зростає необхідний для їх зберігання обсяг пам'яті. Якщо агрегати зберігаютьсяспільно з вихідними даними, то в вимірах необхідно використовувати додатковий параметр – рівень ієрархії .

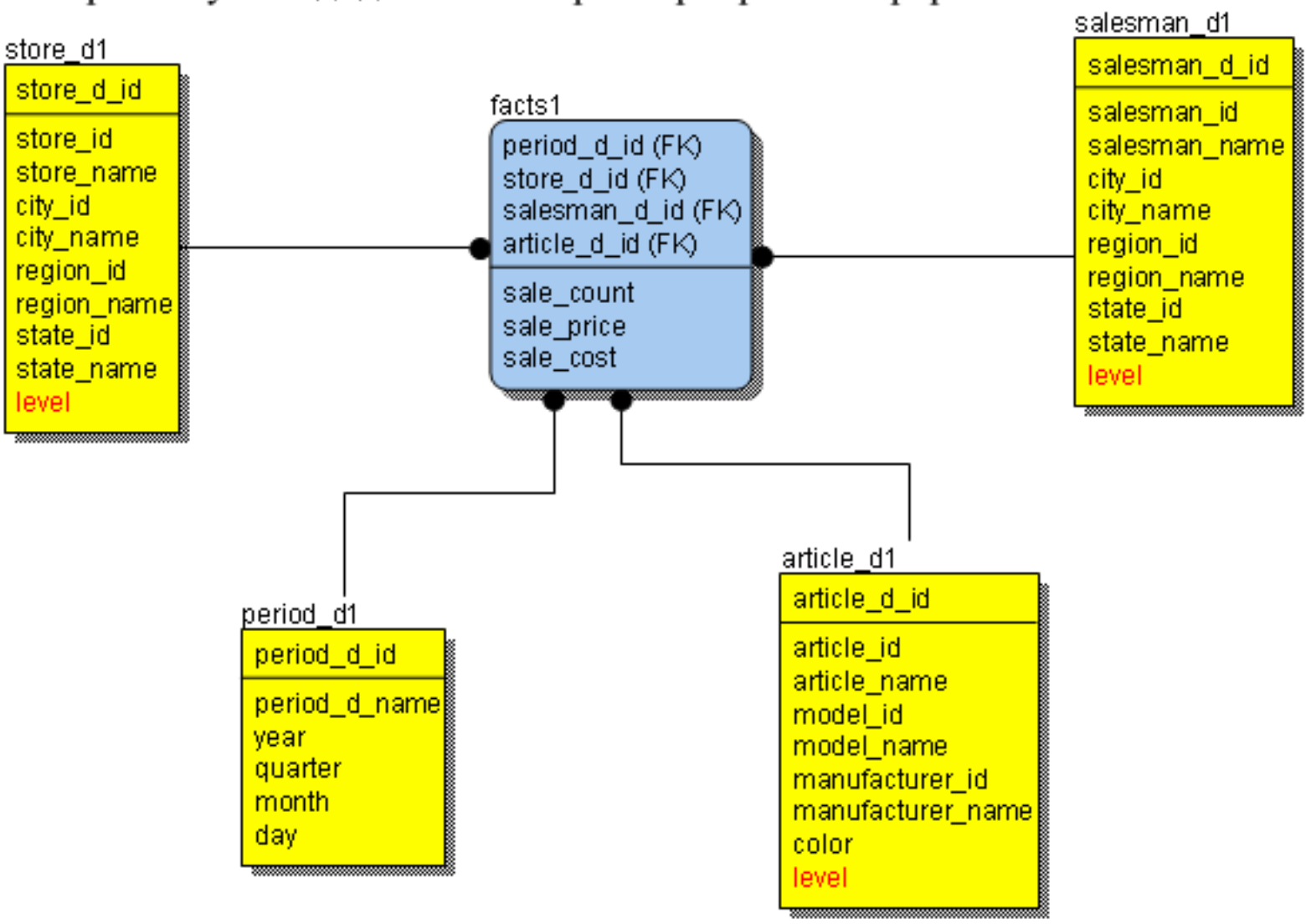


Схема типу сніжинки (Snowflake Schema) -схема реляційної бази даних , що служить дляпідтримки багатовимірного уявлення містяться в ній даних, є різновидом схеми типу "зірка "(Star Schema).

*Особливості ROLAP -схеми типу "сніжинка" *

1. Одна таблиця фактів (fact table), яка сильно денормалізована. Є центральною в схемі, може складатися з мільйонів рядків і містити підсумовувані або фактичні дані, з допомогою яких можна відповісти на різні питання.
2. Кілька таблиць вимірів (dimensional table), які нормалізовані на відміну від схеми "зірка ". Мають меншу кількість рядків, ніж таблиці фактів, і містять описову інформацію. Ці таблиці дозволяють користувачеві швидко переходити від таблиці фактів до додаткової інформації. Первинніключі в них складаються з єдиного атрибута (відповідають єдиному елементу вимірювання).

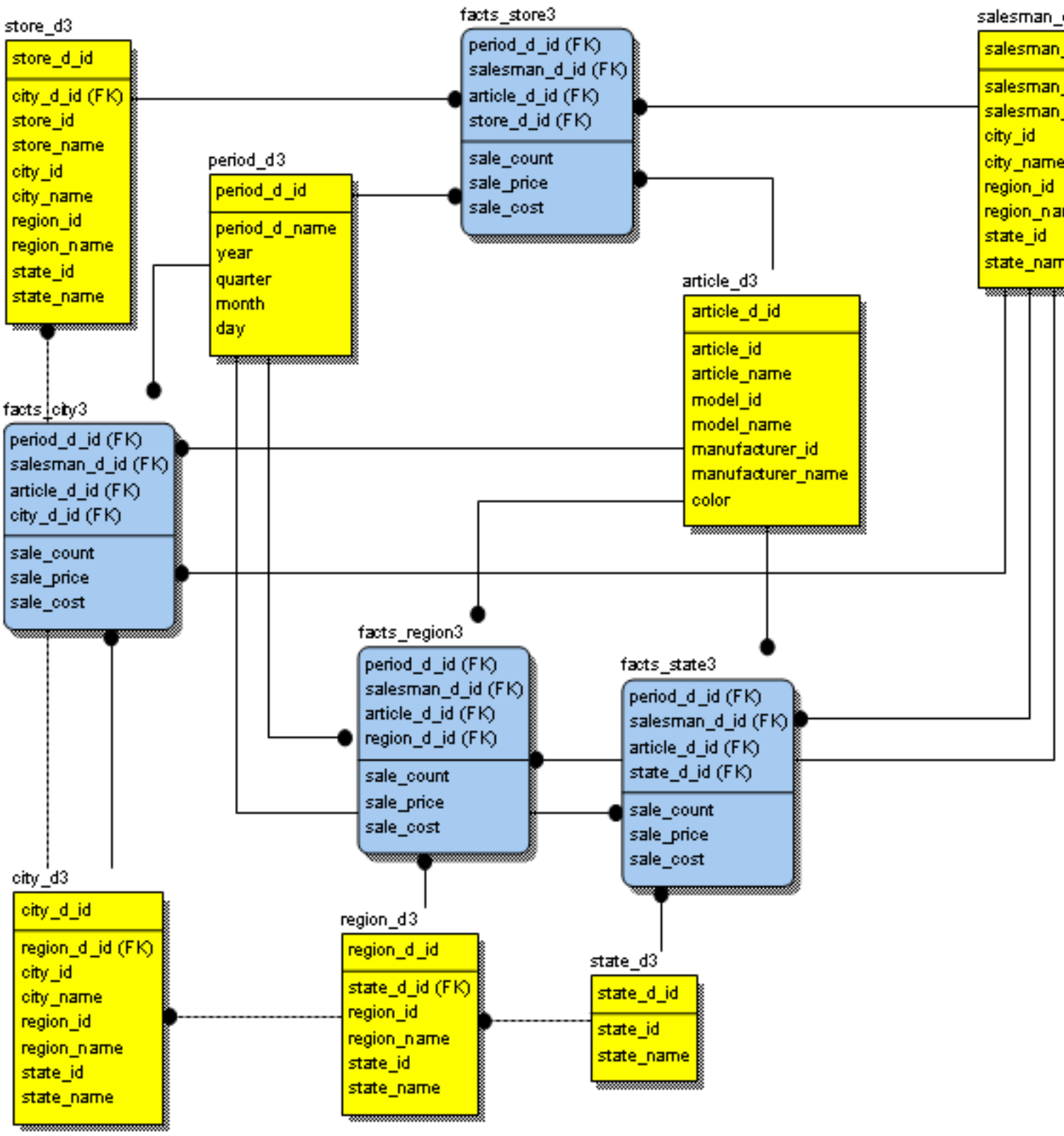
3. Таблиця фактів і таблиці розмірності пов'язаніідентифікують зв'язками, при цьому первинніключі таблицірозмірності мігрують в таблицю фактів якзовнішніх ключів .Первинний ключ таблиці факту цілком складається зпервинних ключів всіх таблицьрозмірності .
4. У схемі "сніжинка" агреговані дані можуть зберігатися окремо від вихідних.

- переваги

Нормалізація таблиць вимірів на відміну від схеми "зірка "дозволяє мінімізуватинадмірність даних і більш ефективно виконувати запити, пов'язані зі структурою значень вимірів.

- недоліки

За нормалізацію таблиць вимірів іноді доводиться платити часом виконання запитів.



Лекція № 2.4. Пошук асоціаційних правил. Метод Apriori

Асоціація - одне із завдань Data Mining . Метою пошуку асоціативних правил (associationrule) є знаходження закономірностей між пов'язаними подіями в базах даних.

У цій лекції ми докладно розглянемо такі питання:

- Що таке асоціативні правила ?
- Які існують алгоритми пошуку асоціативних правил ?
- Що таке часто зустрічаються набори товарів?
- Застосування завдання пошуку асоціативних правил ?

Дуже часто покупці купують не один товар , а кілька. У більшості випадків між цими товарами існує взаємозв'язок. Так наприклад,покупець , що придбає макаронні вироби, швидше за все, захоче придбати також кетчуп. цяінформація може бути використана длярозміщення товару на прилавках.

Часто зустрічаються додатки із застосуванням асоціативних правил:

- роздрібна торгівля: визначення товарів, які варто просувати спільно; вибір місця розташування товару в магазині; аналіз споживчого кошика; прогнозування попиту;
- перехресні продажі: якщо є інформація про те, що клієнти придбали продукти А, Б і В, то які з них найімовірніше куплять продукт Г?
- маркетинг: пошук ринкових сегментів, тенденцій купівельного поведінки;
- сегментація клієнтів: виявлення загальних характеристик клієнтів компанії, виявлення груп покупців;
- оформлення каталогів, аналіз збутових кампаній фірми, визначення послідовностей покупок клієнтів (яка покупка піде за покупкою товару А);
- аналіз Web-логів.

Наведемо простий приклад асоціативного правила :покупець , що придбає банку фарби, придбає пензлик для фарби з імовірністю 50%.

Введення в асоціативні правила

Вперше завдання пошуку асоціативних правил (association rule mining) була запропонована для знаходження типових шаблонів покупок, що здійснюються в супермаркетах, тому іноді її ще називають аналізомринкового кошика (market basketanalysis).

Ринкова кошик - це набір товарів, придбаних покупцем в рамках однієї окремо взятоїтранзакції .

Транзакції є досить характерними операціями, ними, наприклад, можуть описуватися результати відвідувань різних магазинів.

Транзакція - цебезліч подій , які сталися одночасно.

Реєструючи всі бізнес-операції протягом усього часу своєї діяльності, торговельні компанії накопичують величезні збори транзакцій . кожна така транзакція є набором товарів, куплених покупцем за один візит.

Отримані в результаті аналізу шаблони включають перелік товарів і число транзакцій , які містять дані набори.

Транзакційна або **операційна база даних** (Transaction database) являє собою двовимірну таблицю, яка складається з номератранзакції (TID) і переліку покупок, придбаних під час цієїтранзакції .
TID - унікальнийідентифікатор , що визначає кожну угоду аботранзакцію .
приклад транзакционной бази даних , що складається з купівельнихтранзакцій , наведено в таблиці 2.4.1 . У таблиці перша колонка (TID) визначає номертранзакції , у другій колонці таблиці наведені товари, придбані під час певної транзакції .

Таблиця 2.4.1. Транзакційна база даних

TID	**придбані покупки**
100	хліб, молоко, печиво
200	Молоко, сметана
300	Молоко, хліб, сметана, печиво
400	Ковбаса, сметана
500	Хліб, молоко, печиво, сметана

На основі наявної бази даних нам потрібно знайти закономірності між подіями, тобто покупками.
Часто зустрічаються шаблони або зразки.
Припустимо, є транзакційна база даних D. Привласнимо значенням товарів змінні (таблиця 2.4.2).
Хліб = a
Молоко = b
Печиво = c
Сметана = d
Ковбаса = e
Цукерки = f

Таблиця 2.4.2. Набори товарів, що Часто зустрічаються

TID **придбані покупки**	⟶ TID **придбані покупки**
100 Хліб, молоко, печиво	100 a, b, c
200 Молоко, сметана	200 b, d
300 Молоко, хліб, сметана, печиво	300 b, a, d, c
400 Ковбаса, сметана	400 e, d
500 Хліб, молоко, печиво, сметана	500 a, b, c, d
600 цукерки	600 f

Розглянемо набір товарів (Itemset), що включає, наприклад, {Хліб, молоко, печиво}. Висловимо цей набір за допомогою змінних:
abc = {a, b, c}
підтримка
Цей набір товарів зустрічається в нашій базі даних три рази, тобто підтримка цього набору товарів дорівнює 3:
SUP (abc) = 3.

При мінімальному рівні підтримки , яка дорівнює трьом, набір товарів abc єчасто зустрічається шаблоном .
min_sup = 3, {Хліб, молоко, печиво} - найпоширенішийшаблон .
Підтримкою називають кількість або відсоток транзакцій , що містять певний набір даних.
Для даного набору товарів підтримка , виражена в процентному відношенні, дорівнює 50%.
SUP (abc) = (3/6) * 100% = 50%
Підтримку іноді також називають забезпеченням набору.
Таким чином, набір становить інтерес, якщо його підтримка вище певного користувачем мінімального значення (min support). Ці набори називають часто зустрічаються (frequent).

Характеристики асоціативних правил

Асоціативне правило має вигляд: "З події А слідує подія В".
В результаті такого виду аналізу ми встановлюємо закономірність такого вигляду: "Якщо в транзакції зустрівся набір товарів (або набір елементів) А, то можна зробити висновок , що в цій же транзакції повинен з'явитися набір елементів В) "Встановлення таких закономірностей дає нам можливість знаходити дуже прості і зрозумілі правила, звані асоціативними .
Основними характеристиками асоціативного правила є підтримка і достовірність правила.
Розглянемо правило "з покупки молока слід покупка печива" для бази даних , яка була приведена вище в таблиці 2.4.1 . поняття підтримки набору ми вже розглянули.
Існує поняття підтримки правила:
правило має підтримку s, якщо s% транзакцій з усього набору містять одночасно набори елементів А і В або, іншими словами, містять обидва товари.
Молоко - це товар А, печиво - цетовар В. Підтримка правила "з покупки молока слід покупка печива" дорівнює 3, або 50%.
Достовірність правила показує, яка ймовірність того, що з події А слідує подія В.
Правило "З А слід В" справедливо з достовірністю с, якщо с% транзакцій з усієї безлічі , містять набір елементів А, також містять набір елементів В.
число транзакцій , що містять молоко, дорівнює чотирьом, число транзакцій , що містять печиво, дорівнює трьом, достовірність правила дорівнює (3/4) * 100%, тобто 75%.
Достовірність правила "з покупки молока слід покупка печива" дорівнює 75%, тобто 75%транзакцій , що містять товар А, також містять товар В.

Межі підтримки і достовірності асоціативного правила

За допомогою використання алгоритмів пошуку асоціативних правил аналітик може отримати всі можливі правила виду "З А слід В", з різними значеннямипідтримки ідостовірності . Однак в більшості випадків, кількість правил необхідно обмежувати заздалегідь встановленими мінімальними і максимальними значеннями підтримки і достовірності .

Якщо значення підтримки правила занадто велике, то в результаті роботи алгоритму будуть знайдені правила очевидні і добре відомі. занадто низьказначення підтримки призведе до знаходження дуже великої кількості правил, які, можливо, будуть в більшій частині необґрунтованими, але не відомими і не очевидними для аналітика. Таким чином, необхідно визначити такийінтервал , "золоту середину", який з одного боку забезпечить знаходження неочевидних правил, а з іншого - їх обґрунтованість.

Якщо рівень достовірності занадто малий, то цінність правила викликає серйозні сумніви. Наприклад, правило з достовірністю в 3% тільки умовно можна назвати правилом.

Методи пошуку асоціативних правил

Алгоритм AIS . першийалгоритм пошукуасоціативних правил , що називався AIS , (запропонований Agrawal, Imielinskiand Swami) був розроблений співробітниками дослідницького центруIBM Almaden в 1993 році. З цієї роботи почався інтерес до асоціативних правил ; на середину 90-х років минулого століття припав пік дослідницьких робіт в цій області, і з тих пір кожен рік з'являється кілька нових алгоритмів.

В алгоритмі AIS кандидатибезлічі наборів генеруються і підраховуються "на льоту", під час скануваннябази даних .

Алгоритм SETM . Створення цього алгоритму було мотивоване бажанням використовувати мову SQL для обчислення наборів товарів.які часто зустрічаються.Як і алгоритм AIS , SETM також формує кандидатів "на льоту", грунтуючись на перетворенняхбази даних . Щоб використовувати стандартну операцію об'єднання мови SQL для формування кандидата, SETM відокремлює формування кандидата від їх підрахунку.

Незручність алгоритмів AIS і SETM - зайве генерування і підрахунок занадто багатьох кандидатів, які в результаті не надаються часто зустрічаються. Для поліпшення їх роботи був запропонований алгоритм Apriori .

Робота даного алгоритму складається з декількох етапів, кожен з етапів складається з наступних кроків:

- формування кандидатів;
- підрахунок кандидатів.

Формування кандидатів (candidategeneration) - етап, на якому алгоритм , скануючи базу даних, створює безліч і-елементних кандидатів (і - номер етапу). На цьому етапі підтримка кандидатів не розраховується.

Підрахунок кандидатів (candidate counting) - етап, на якому обчислюєтьсяпідтримка кожного і-елементного кандидата. Тут же здійснюється відсікання кандидатів, підтримка яких менше мінімуму, встановленого користувачем (min_sup). Решта і-елементні набори називаємо часто зустрічаються.

Розглянемо роботу алгоритму Apriori на прикладі бази даних D. Ілюстрація роботи алгоритму приведена на рис. 2.4.1 мінімальний рівень підтримки дорівнює 3.

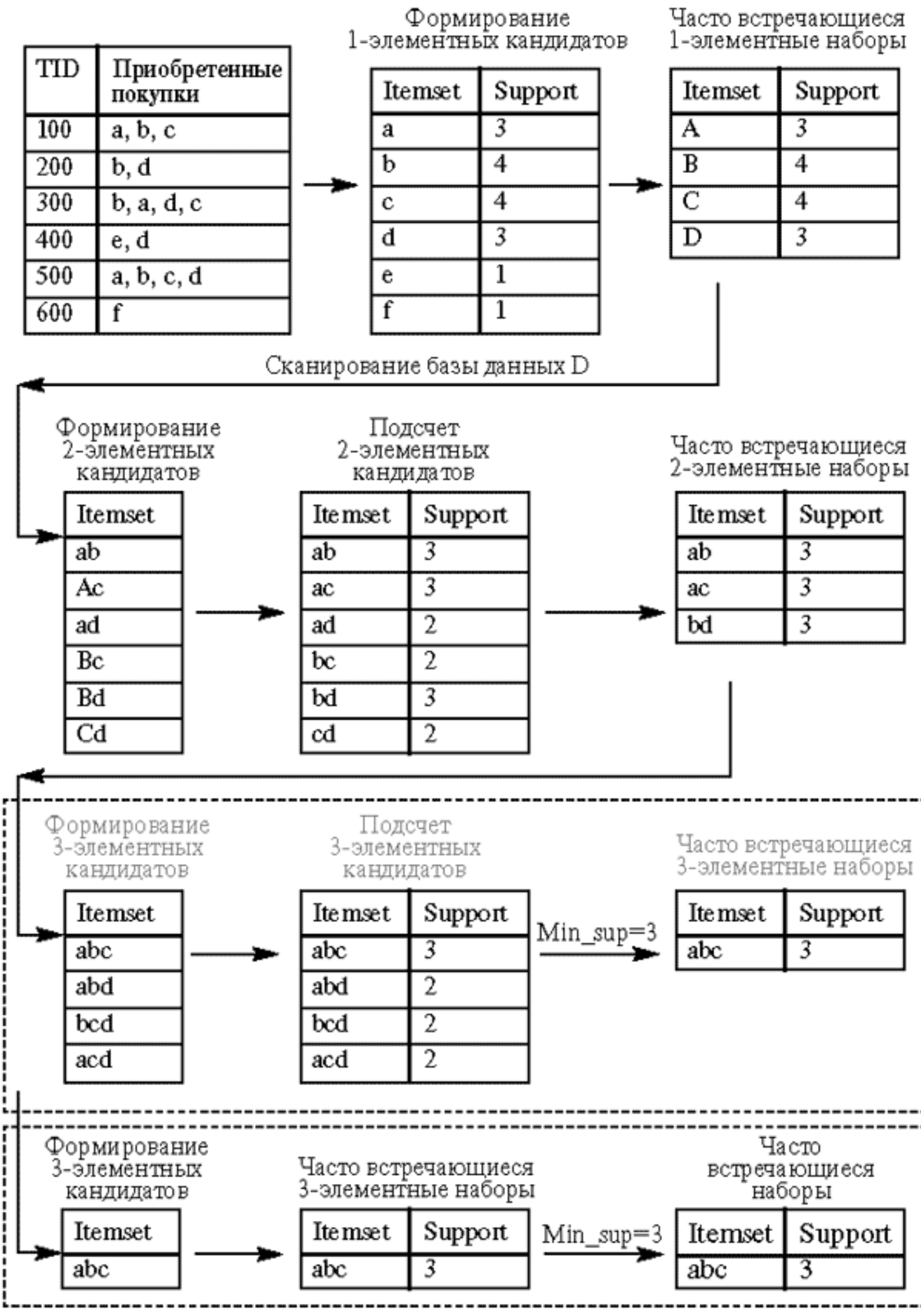


Рис.2.4.1. Алгоритм Apriori

На першому етапі відбувається формування одноелементні кандидатів. Далі алгоритм підраховує підтримку одноелементних наборів. Набори з рівнем підтримки менше встановленого, тобто 3, відсікаються. У нашому прикладі це

набори e і f, які мають підтримку , рівну 1. Решта набори товарів вважаються такими, що часто зустрічаються ,одноелементними наборами товарів: це набори a, b, c, d.

Далі відбувається формування двоелементних кандидатів, підрахунок їх підтримки і відсікання наборів з рівнем підтримки , меншим 3. Решта двоелементних наборів товарів, які вважаються такими, що часто зустрічаються двоелементними наборами ab, ac, bd, беруть участь в подальшій роботі алгоритму.

Якщо дивитися на роботу алгоритму прямолінійно, на останньому етапі алгоритм формує трьохелементні набори товарів: abc, abd,bcd ,acd , підраховує їх підтримку і відсікає набори з рівнем підтримки , меншим 3. Набір товарів abc може бути названий таким, що часто зустрічається.

Однак алгоритм Apriori зменшує кількість кандидатів, відсікаючи - апріорі - тих, які свідомо не можуть стати часто зустрічаються, на основі інформації про відсічених кандидатах на попередніх етапах роботи алгоритму.

Відсікання кандидатів відбувається на основі припущення про те, що у наборів товарів, які часто зустрічаються, все підмножини повинні бути такими, що часто зустрічаються. Якщо в наборі знаходиться підмножина , яка на попередньому етапі було визначена як така, що нечасто зустрічається, цей кандидат вже не включається у формування і підрахунок кандидатів.

Так набори товарів ad, bc, cd були відкинуті як такі, що нечасто зустрічаються ,алгоритм не розглядав набір товарів abd,bcd ,acd .

При розгляді цих наборів формування трьохелементних кандидатів відбувалося б за схемою, наведеною в верхньому пунктирному прямокутнику. Оскільки алгоритм апріорі відкинув свідомо набори,що нечасто зустрічаються, останній етап алгоритму відразу визначив набір abc як єдиний трьохелементний набір, що часто зустрічається (етап наведено в нижньому пунктирному прямокутнику).

Алгоритм Apriori розраховує також підтримку наборів, які не можуть бути відсічені апріорі. Це так звана негативна область (negative border), до неї належать набори-кандидати, які зустрічаються рідко, їх самих не можна віднести до таких, що часто зустрічаються, але все підмножини даних наборів є такими, що часто зустрічаються.

Різновиди алгоритму Apriori

Залежно від розміру найдовшого часто зустрічається набору алгоритм Apriori сканує базу даних певну кількість разів. Різновиди алгоритму Apriori, є його оптимізацією, запропоновані для скорочення кількості скануваньбази даних , кількості наборів-кандидатів або того й іншого]. Були запропоновані наступні різновиди алгоритму Apriori: AprioriTID і AprioriHybrid.

AprioriTid

Цікава особливість цього алгоритму - то, що база даних D не використовується для підрахунку підтримки кандидатів набору товарів після першого проходу.

З цією метою використовується кодування кандидатів, виконане на попередніх проходах. У наступних проходах розмір закодованих наборів може бути набагато менше, ніж база даних, і таким чином економляться значні ресурси.

AprioriHybrid

Аналіз часу роботи алгоритмів Apriori і AprioriTid показує, що в більш ранніх проходах Apriori домагається більшого успіху, ніж AprioriTid; проте AprioriTid працює краще Apriori в більш пізніх проходах. Крім того, вони використовують одну і ту ж процедуру формування наборів-кандидатів. Заснований на цьому спостереженні, алгоритм AprioriHybrid запропонований, щоб об'єднати кращі властивості алгоритмів Apriori і AprioriTid. AprioriHybrid використовує алгоритм Apriori в початкових проходах і переходить до алгоритму AprioriTid, коли очікується, що закодований набір початкового безлічі в кінці проходу буде відповідати можливостям пам'яті. Однак, перемикання від Apriori до AprioriTid вимагає залучення додаткових ресурсів.

Деякими авторами були запропоновані інші алгоритми пошуку асоціативних правил , метою яких також було удосконалення алгоритму Apriori. Коротко викладемо суть декількох, для більш докладної інформації можна рекомендувати .

Один з них - **алгоритм DHP** , також званий алгоритмом хешування (J.Park , M. Chen and P. Yu, 1995 рік). В основі його роботи - імовірнісний підрахунок наборів-кандидатів, здійснюваний для скорочення числа підраховуваних кандидатів на кожному етапі виконання алгоритму Apriori. Скорочення забезпечується за рахунок того, що кожен з k-елементних наборів-кандидатів крім кроку скорочення проходить крок хеширования. В алгоритмі на k-1 етапі під час вибору кандидата створюється так звана хеш-таблиця. Кожен запис хеш-таблиці є лічильником всіх підтримок k-елементних наборів, які відповідають цій записи в хеш-таблиці. Алгоритм використовує цю інформацію на етапі k для скорочення безлічі k-елементних наборів-кандидатів. Після скорочення підмножини, як це відбувається в Apriori, алгоритм може видалити набір-кандидат, якщо його значення в хеш-таблиці менше порогового значення, встановленого для забезпечення.

До інших вдосконалених алгоритмам відносяться: PARTITION, DIC , алгоритм "вибіркового аналізу".

PARTITION алгоритм (A. Savasere, E. Omiecinski and S. Navathe, 1995 рік). Цей алгоритм розбиття (поділу) полягає в скануванні транзакционной бази даних шляхом поділу її на непересічні розділи, кожен з яких може вміститися в оперативній пам'яті . На першому кроці в кожному з розділів за допомогою алгоритму Apriori визначаються "локальні" часто зустрічаються набори даних. На другому підраховується підтримка кожного такого набору щодо всієї бази даних. Таким чином, на другому етапі визначається безліч всіх потенційно зустрічаються наборів даних.

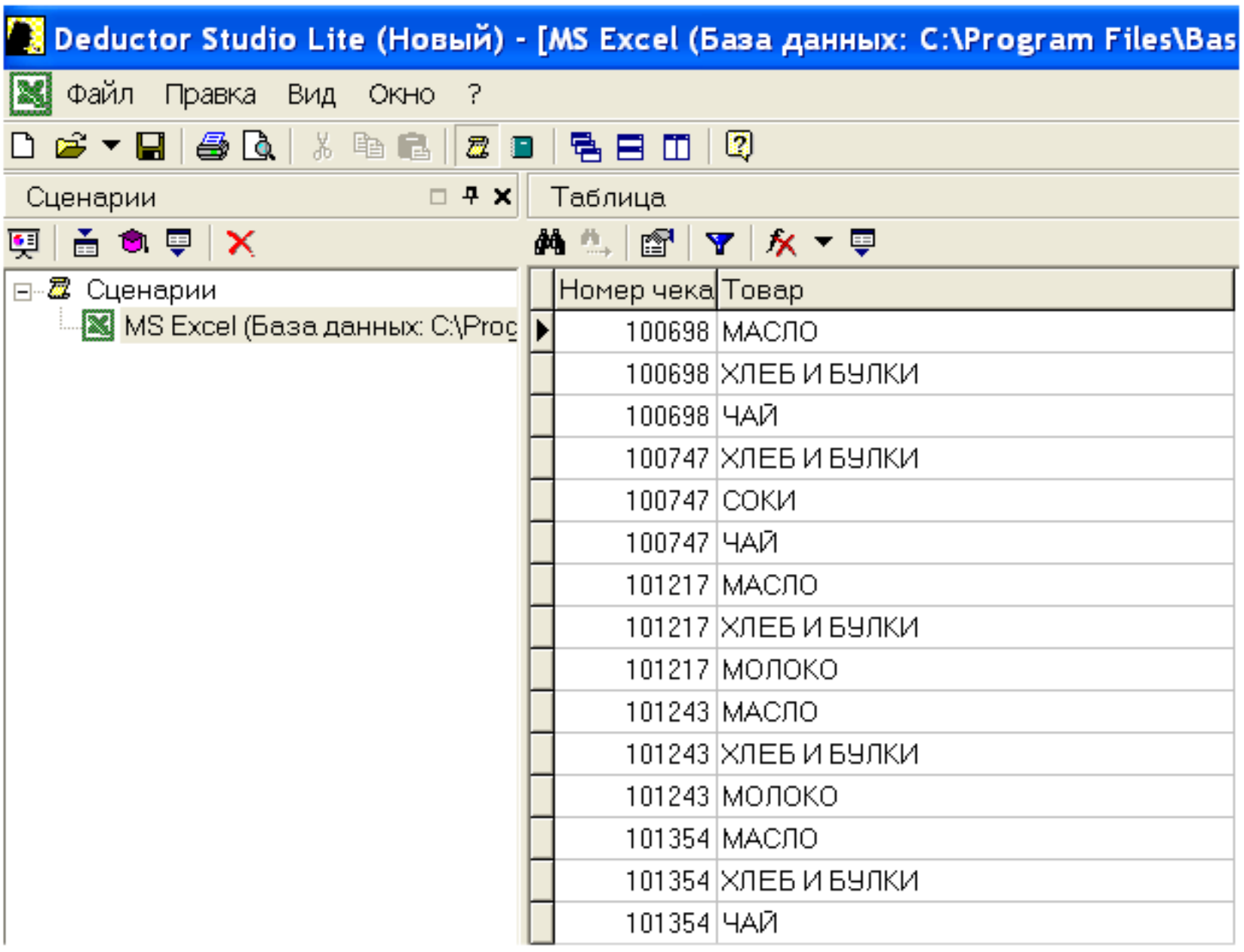
Алгоритм DIC , Dynamic Itemset Counting (S. Brin R. Motwani, J. Ullman and S. Tsur, 1997 рік). Алгоритм розбиває базу даних на кілька блоків, кожен з яких відзначається так званими "початковими точками" (start point), і потім циклічно сканує базу даних .

Приклад рішення задачі пошуку асоціативних правил

дана транзакційна база даних , необхідно знайти найбільш часто зустрічаються набори товарів і набірасоціативних правил з визначеними межами значеньпідтримки і довіри.

Розглянемо процес побудови асоціативних правил в аналітичному пакеті Deductor.

Транзакційна база даних , яка містить в кожному записі номер чека ітовар , придбаний за цим чеком, має формат MSExcel . Для початку імпортуємо дані з файлу MSExcel в середу Deductor, цей процес аналогічний тому, що був розглянутий в лекції про нейронних мережах. Єдина відмінність - в призначенні стовпців. для номератранзакції (зазвичай в базі даних - цеполе "номер чека") вказуємо тип "ідентифікатор транзакції (ID) ", а для найменувань товару - тип" елемент ". Результат імпортубази даних з файлу MSExcel в середу Deductor бачимо на рис. 2.4.2 . На рисунку наведено фрагмент бази даних , яка містить більше 140 записів.



Номер чека	Товар
100698	МАСЛО
100698	ХЛЕБ И БУЛКИ
100698	ЧАЙ
100747	ХЛЕБ И БУЛКИ
100747	СОКИ
100747	ЧАЙ
101217	МАСЛО
101217	ХЛЕБ И БУЛКИ
101217	МОЛОКО
101243	МАСЛО
101243	ХЛЕБ И БУЛКИ
101243	МОЛОКО
101354	МАСЛО
101354	ХЛЕБ И БУЛКИ
101354	ЧАЙ

Рис.. 2.4.2. Транзакційна база даних, імпортована в Deductor з файлу MS Excel

Далі викликаємо майстер обробки і вибираємо метод " Асоціативні правила ". На другому кроці майстра перевіряємо призначення вихідних стовпців даних, вони повинні мати тип"ID "і" елемент ".

На третьому кроці, проілюстрованому на рис. 2.4.3 , необхідно налаштувати параметри пошуку правил, тобто встановити мінімальні і максимальні характеристики підтримки і достовірності . Це найбільш "відповідальний" момент формування набору правил, про важливість вибору меж значень підтримки і достовірності вже говорилося на початку лекції. Вибір можна

зробити на основі будь-яких міркувань, наявного досвіду аналізу подібних даних, інтуїції або ж визначити в ході експериментів.

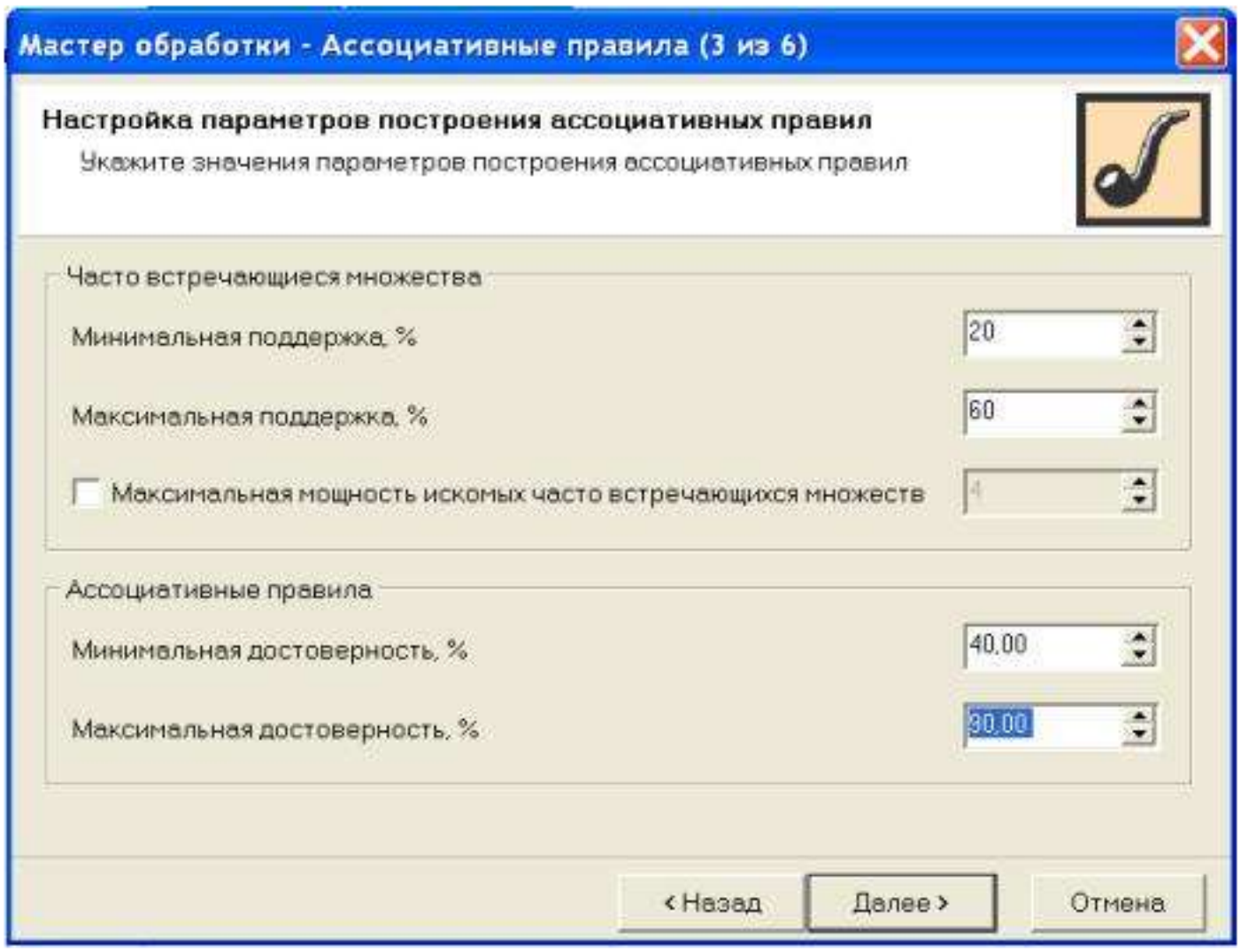


Рис. 2.4.3. Налаштування параметрів побудови асоціативних правил

Ми встановимо такі кордону для параметрів пошуку: мінімальний і максимальний рівень підтримки на рівні 20% і 60% відповідно, мінімальний і максимальний рівень значення достовірності рівні 40% і 90% відповідно. Ці значення були виявлені в ході проведення декількох експериментів, і виявилося, що саме при таких значеннях формується необхідний набір правил. При вказівці деяких значень, наприклад, рівняпідтримки від 30% до 50%, набір правил не формується, оскільки жодне правило за параметрамипідтримки не входить в цей інтервал .

На наступному кроці майстра запускається процес пошуку асоціативних правил . В результаті бачимо інформацію про кількість множин і знайдених правил у вигляді гістограми розподілу множин, що часто зустрічаються по їх потужності. Даний процес проілюстрований на рис. 2.4.4 .

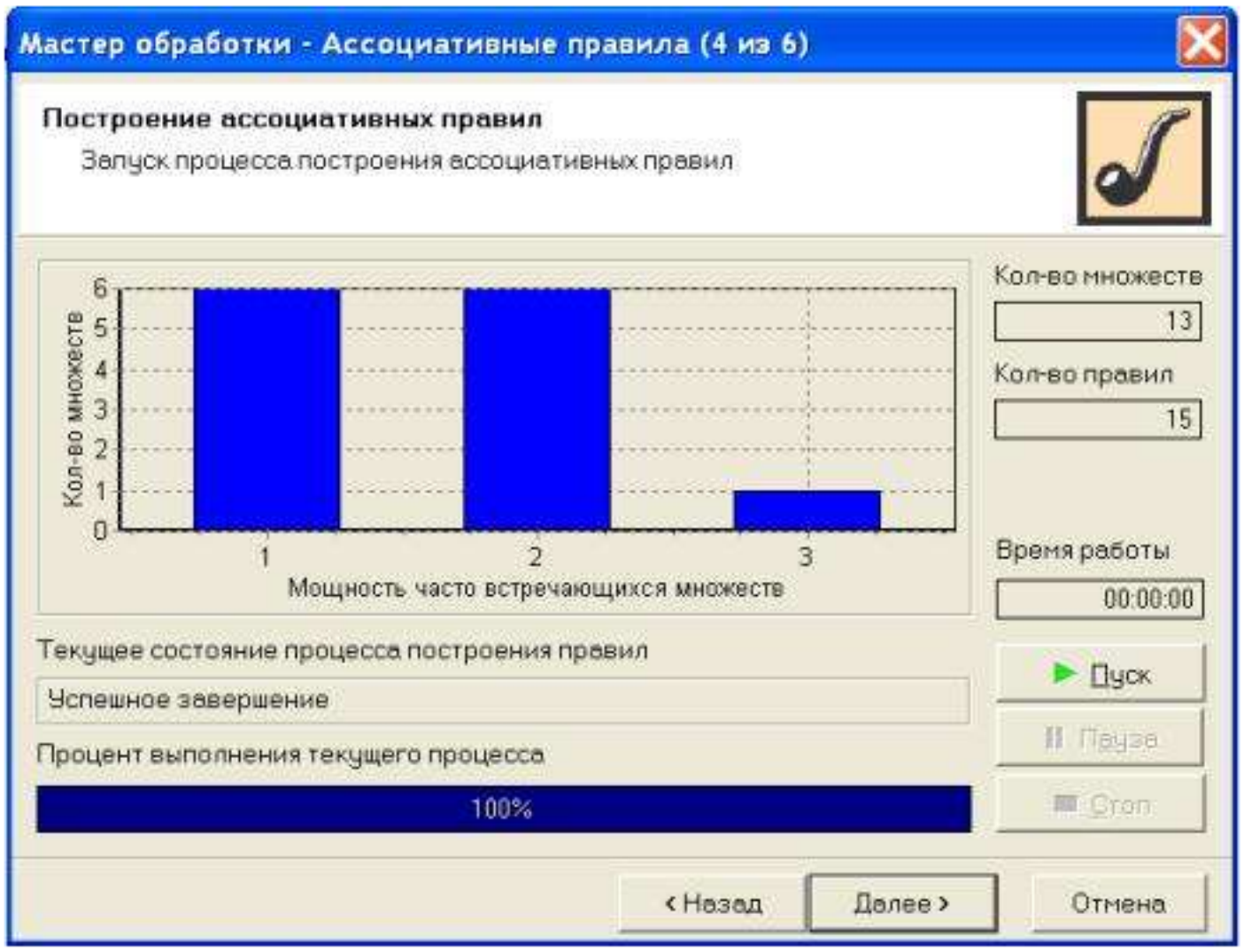


Рис. 2.4..4. Процес побудови асоціативних правил

Тут ми бачимо, що кількість сформованих множин тринадцять - це популярні набори, кількість сформованих правил - п'ятнадцять.

На наступному кроці для перегляду отриманих результатів пропонується вибрати візуалізатори зі списку; ми виберемо такі: "Популярні набори", "Правила", "Дерево правил "," Що-якщо ". Розглянемо, що вони з себе представляють.

Візуализатор "Популярні набори". Популярні набори або такі, що часто зустрічаються набори - це набори, що складаються з одного або декількох товарів, які в транзакціях найбільш часто зустрічаються одночасно. Характеристикою, наскільки часто набір зустрічається в аналізованому наборі даних, є підтримка .

Популярні набори нашого набору даних, знайдені при заданих параметрах, наведені в таблиці 2.4.3 . Є можливість відсортувати цю таблицю за різними її характеристиками. Для визначення найбільш популярних товарів і їх наборів зручно впорядкувати її за рівнем підтримки . Таким чином, ми бачимо, що найбільшою популярністю користуються такі товари: хліб і булки, масло, соки.

Таблиця 2.4.3. Візуалізатор "Популярні набори"

N	**безліч**	↑ **підтримка**	
		%	**Кількість**
6	ХЛІБ І булки	54,5524	
3	МАСЛО	52,2723	
5	СОКИ	50,0022	
10	МАСЛО **І** ХЛІБ І булки	45,4520	

4	МОЛОКО	43,1819	
2	КЕФІР	31,8214	
1	йогурт	31,8214	
12	СОКИ **І** ХЛІБ І булки	22,7310	
11	МОЛОКО **І** ХЛІБ І булки	22,7310	
8	МАСЛО **І** МОЛОКО	22,7310	
7	Йогурт **І** КЕФІР	22,7310	
13	МАСЛО **І** МОЛОКО **І** ХЛІБ І булки	20,459	
9	МАСЛО **І** СОКИ	20,459	

Візуалізатор "Правила"

Правила в даному візуалізаторі розміщені у вигляді списку. Кожне правило, представлене як "умова-наслідок", характеризується значенням підтримки в абсолютному і процентному вираженні, а також достовірністю . Таким чином, аналітик бачить поведінку покупців, описану у вигляді набору правил. Набір правил для розв'язуваної нами задачі наведено в таблиці 2.4.4 . Наприклад, перше правило говорить про те, що якщо покупець купив йогурт, то з достовірністю або ймовірністю 71% він купить також кефір. Ця інформація корисна з різних точок зору. Вона, наприклад, допомагає вирішити задачу розташування товарів у магазині.

Таблиця 2.4.4. Візуалізатор "Правила"

N	**Умова**	**слідство**	**підтримка**		**Достовірність,%**
			%	**Кількість**	
1	йогурт	КЕФІР	22,7310		71,43
2	КЕФІР	йогурт	22,7310		71,43
3	МАСЛО	МОЛОКО	22,7310		43,48
4	МОЛОКО	МАСЛО	22,7310		52,63
5	СОКИ	МАСЛО	20,459		40,91
6	МАСЛО	ХЛІБ І булки	45,4520		86,96
7	ХЛІБ І булки	МАСЛО	45,4520		83,33
8	МОЛОКО	ХЛІБ І булки	22,7310		52,63
9	ХЛІБ І булки	МОЛОКО	22,7310		41,67
10	СОКИ	ХЛІБ І булки	22,7310		45,45
11	ХЛІБ І булки	СОКИ	22,7310		41,67
12	МАСЛО **І** МОЛОКО	ХЛІБ І булки	20,459		90,00
13	МАСЛО **І** ХЛІБ булки	ІМОЛОКО	20,459		45,00
14	МОЛОКО **І** ХЛІБ булки	ІМАСЛО	20,459		90,00
15	МОЛОКО	МАСЛО **І** ХЛІБ булки	І20,459		47,37

При великій кількості знайдених правил і широкому асортименті товарів аналізувати отримані правила досить складно. Для зручності аналізу таких наборів правил пропонуються візуалізатори "Дерево правил" і "Що-якщо".

Візуализатор "Дерево правил" - дворівневе дерево, яке може бути побудовано за двома критеріями: за умовою і за наслідком. Якщо дерево побудовано за умовою, то у верхній частині списку відображаються умови правила, а список, що додається до даної умові, складається з його наслідків. При виборі певної умови, в правій частині візуализатора відображаються наслідки умови, рівень підтримки і достовірності .

У разі побудови дерева за наслідком, у верхній частині списку відображаються наслідки правила, а список складається з його умов. При виборі певного наслідку, в правій частині візуализатора ми бачимо умови цього правила із зазначенням рівня підтримки і достовірності .

Візуализатор "що-якщо" зручний, якщо нам необхідно відповісти на питання, які наслідки можуть вийти з даної умови.

Наприклад, вибравши умову "МОЛОКО", в лівій частині екрана отримуємо три наслідки "МАСЛО", "ХЛІБ І булки", "МАСЛО І ХЛІБ І булки", для яких вказані рівень підтримки і достовірності . Цей візуалізатор представлений на рис. 2.4..5 .

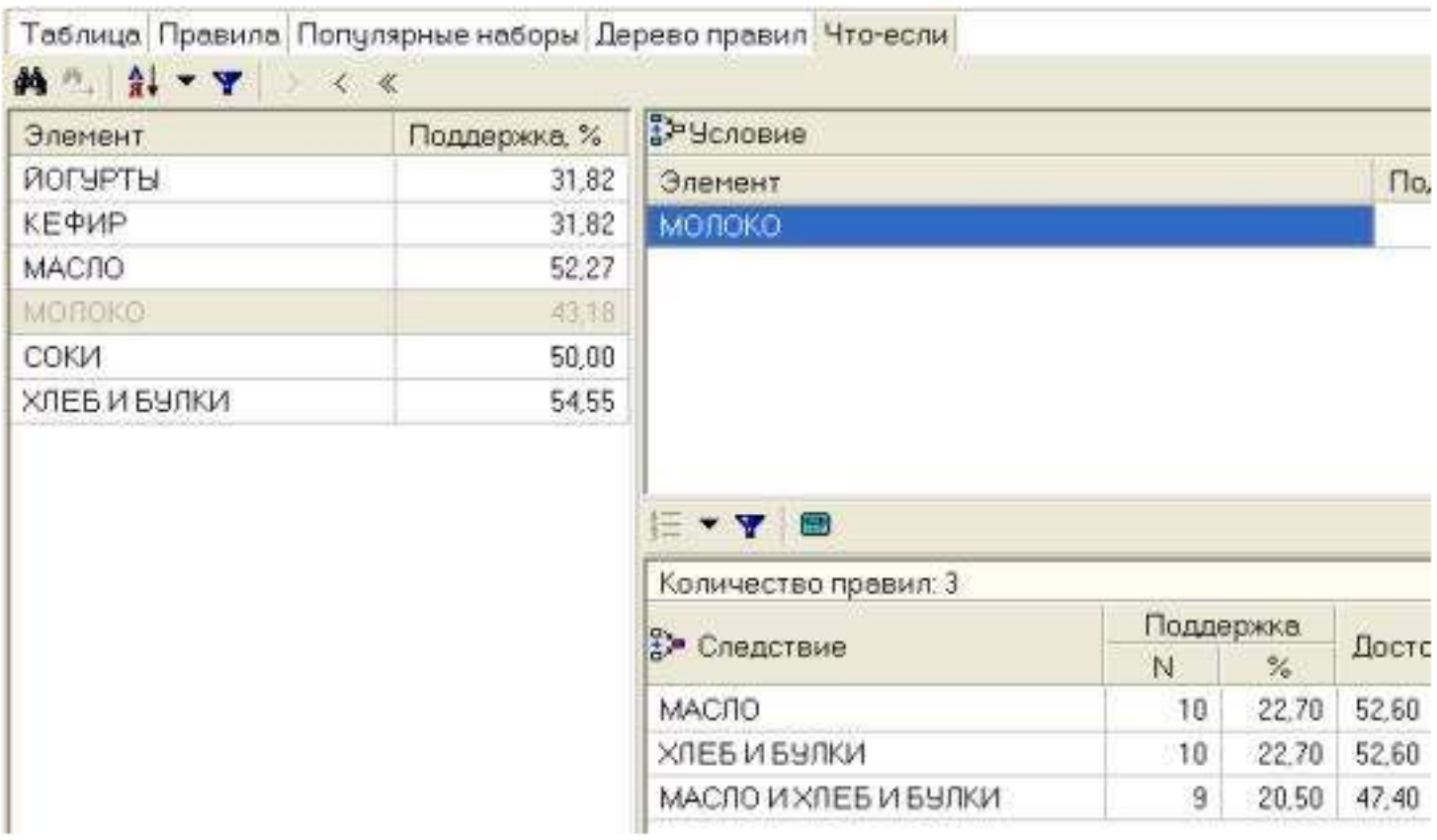


Рис. 2.4.5. Візуалізатор "Що-якщо"

Розглянутий приклад пошуку асоціативних правил є типовою ілюстрацією завдання аналізу купівельної корзини. В результаті її рішення визначаються набори товарів, що часто зустрічаються , а також набори товарів, спільно придбані покупцями. Знайдені правила можуть бути використані для вирішення різних завдань, зокрема для розміщення товарів на прилавках магазинів, надання знижок на пари товарів для підвищення обсягу продажів і, отже, прибутку та інших завдань.

Список використаної літератури

1. Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И. – Методы и модели анализа данных OLAP и Data Mining – СПб.: БВХ–Петербург, 2011, – 336с.: ил.
2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И. И. – Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP – СПб.: БВХ–Петербург, 2009, – 384с.: ил.
3. Паклин Н. Б., Орешков В. И. Бизнес-аналитика: от данных к знаниям (+ CD). — СПб.: Изд. Питер, 2009. — 624 с.
4. Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. — М.: Изд. «Фазис», 2006. — 176 с. — ISBN 5-7036-0108-8.
5. Чубукова И. А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. — 382 с. — ISBN 5-9556-0064-7.
6. Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних (дейтамайнінг): Навч. посібник. — К.: КНЕУ, 2007. — 376 с.
7. Ian H. Witten, Eibe Frank and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition. — Morgan Kaufmann, 2011. — P. 664. — ISBN 9780123748560.
8. Спирли Эрик – Корпоративные хранилища данных. Планирование, разработка и реализация. Том 1 – .: Пер с англ. – М.: Издательский дом «Вильямс», 2009. – 400с.: ил.
9. Статистика. Підручник / С.С. Герасименко та ін. - К.: КНЕУ, 2000. - 467 с.
10. Тарасенко Т.О. Статистика: Навчальний посібник. - К.: Центр навчальної літератури, 2006. - 344 с.
11. Ткач Є.І. Загальна теорія статистики: Підручник. - Тернопіль.: Лідер, 2004. - 388 с.

ЗМІСТ

Навчальне видання

Олешко Т.І., Квашук Д. М.

КОНСПЕКТ ЛЕКЦІЙ

з дисципліни «Інструментальні засоби статистичного та інтелектуального аналізу даних»
за спеціальністю 051 «Економіка», освітньо-професійною програмою:
«Економічна кібернетика»