НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ЗАОЧНОГО ТА ДИСТАНЦІЙНОГО НАВЧАННЯ

Кафедра економічної кібернетики

МЕТОДИЧНІ РЕКОМЕНДАЦІЇ
до виконання контрольноїроботи з дисципліни
з дисципліни «Інструментальні засоби статистичного та інтелектуального аналізу даних»

для студентів заочної форми навчання

Галузь знань: 05 "Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»
Освітньо-професійна програма: «Економічна кібернетика», «Цифрова економіка»

Укладачі:
професор кафедри економічної кібернетики
д.т.н., професор Олешко Т.І.,
к.е.н., Квашук Д. М.

Методичні рекомендації
розглянуті та схвалені
на засіданні кафедри
економічної кібернетики
Протокол № 5 від 2.04.2018 р.

Завідувач кафедри_____Н. В. Касьянова

КИЇВ-2018

УДК 519.25:004.67(076.5)
ББК У.с51р
І619

Укладачі : Олешко Тамара Іванівна, Квашук Дмитро Михайлович

Рецензенти: Григорак Марія Юріївна, к.е.н., доц.
(Національний авіаційний університет)
Ластівка Іван Олексійович, д.т.н., проф.
(Національний авіаційний університет)

Інструментальні засоби статистичного та інтелектуального аналізу даних : Методичні вказівки до виконання контрольноїроботи / Уклад.: Т.І. Олешко, Д. М. Квашук - К.: НАУ, 2018. – 56 с.

Методичні вказівки містять рекомендації до виконання контрольноїроботи з дисципліни «Інструментальні засоби статистичного та інтелектуального аналізу даних» для студентів-магістрів заочної форми навчання ОПП «Економічна кібернетика».

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.

Контрольна робота виконуються у одинадцятому семестрі, відповідно до затверджених в установленому порядку методичних рекомендацій і є важливим етапом у засвоєнні навчального матеріалу.

Робота складається з одного теоретичного питання та практичного завдання.Обсяг роботи – до 40 сторінок.

Час, потрібний для виконання контрольноїроботи - до 8 годин самостійної роботи.

Варіанти завдань з контрольноїроботи визначається за двома останніми цифрами студентської книжки.

Позначимо число, складене із цих цифр через n. Тоді:

а) якщо $0 < n \leq 30$, номер варіанту дорівнює n.

б) якщо $30 < n \leq 60$, номер варіанту дорівнює n-30.

в) якщо $60 < n \leq 90$, номер варіанту дорівнює n-60.

г) якщо $90 < n \leq 99$, номер варіанту дорівнює n-90.

д) якщо $n = 0$, номер варіанту дорівнює 30.

ТЕОРЕТИЧНІ ВІДОМОСТІ

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	Письмові опитування (анкетування); Усні опитування (інтерв'ювання); Очні опитування; Заочні	Збір первинної інформації (наприклад, про використовувані забезпечують засобах, про користувачів і їх

	питаннями до певної групи людей (респондентам). Дозволяє виявити певні типові фактичні дані, а також поняття і судження з різних питань.	опитування (телефонні, поштові, опитування за допомогою Інтернету, через пресу); Інтерактивні опитування по телебаченню або по радіо; Вибіркові опитування; Суцільні опитування	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 безпосередньому контакті з респондентом усно задає питання, передбачені програмою дослідження. Є одним з основних видів опитування.	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, обассортименте і якості інформаційних продуктів і послуг.
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	Метод комісії; Метод мозкового штурму (або метод колективної генеральної ідеї); метод Дельфі; Методеврістіческого прогнозування; Метод узагальнення незалежних характеристик; Метод простий ранжування;	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

	інтуїтивно-логічного аналізу проблеми з кількісним судженням і формальною обробкою результатів. Отримане в результаті обробки узагальнена думка експертів є рішення проблеми.	Метод завдання вагових коефіцієнтів; Метод парних срвненій; Метод послідовних порівнянь	продуктів і послуг.
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 (кількісною обробкою) отриманих даних.		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.
Факторний аналіз	Метод багатовимірної математичної статистики,		Застосовують в тих випадках, коли необхідно встановити і виявити приховані для дослідника

	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.		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 отриманих результатів.		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.
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 (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.	Ієрархічний методФасетний метод (метод паралельних класифікацій)	Дозволяє встановити зв'язку між досліджуваними об'єктами; служить основою для узагальнюючих висновків і прогнозів. 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	Глобальне прогнозування; Нормативне прогнозування; Аналітичне прогнозування	Визначення перспектив розвитку інформаційних систем, мереж і технологій.

	перспектив розвитку тих чи інших процесів на основі аналізу даних про їхнє минуле і нинішній стан.		
моделювання	Один з методів пізнання (відображення) і перетворення світу, сутність якого зводиться до побудови та вивчення деякої моделі з подальшим «перенесенням» отриманих знань на досліджуваний об'єкт.	Матеріальне моделювання: фізичне, аналоговоеІдеальное моделювання: знакове (графічне, логічне, математичне), інтуїтивне	Застосовується в 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.

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:

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

$D = \frac{\sum d^2}{(N-1)}$ 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:

$$Q_{лев} = \frac{(Q_2 - Q_1)}{2};$$

$$Q_{прав} = \frac{(Q_3 - Q_2)}{2}.$$

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.

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:

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

де y - залежна змінна;

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

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

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

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:

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

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;

$$X=\begin{bmatrix} 1 & x_{11} & x_{12} & \_ & x_{1p} \\ 1 & x_{21} & x_{22} & \_ & x_{2p} \\ - & - & - & \_ & - \\ 1 & x_{n1} & x_{n2} & \_ & x_{np} \end{bmatrix}$$ - 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 і:
$D[\varepsilon_i]=\sigma^2$.
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

$$\begin{cases} \frac{\partial S}{\partial b_0} = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip}\right) = 0, \\ \frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n} x_{i1}\left(y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip}\right) = 0, \\ \ldots \\ \frac{\partial S}{\partial b_p} = -2\sum_{i=1}^{n} x_{ip}\left(y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip}\right) = 0, \end{cases}$$

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

$$b = \left(X^T X\right)^{-1} X^T y, \quad (5)$$

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

Оцінки b_j є незміщеними, обґрунтованими та ефективними.

Оцінка дисперсії похибок

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

є незміщеною та обґрунтованою.

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

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

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).

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: $\sigma^2 = \sigma_A^2 + \sigma_B^2 + \ldots$

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 вдається 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.

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 .

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 .

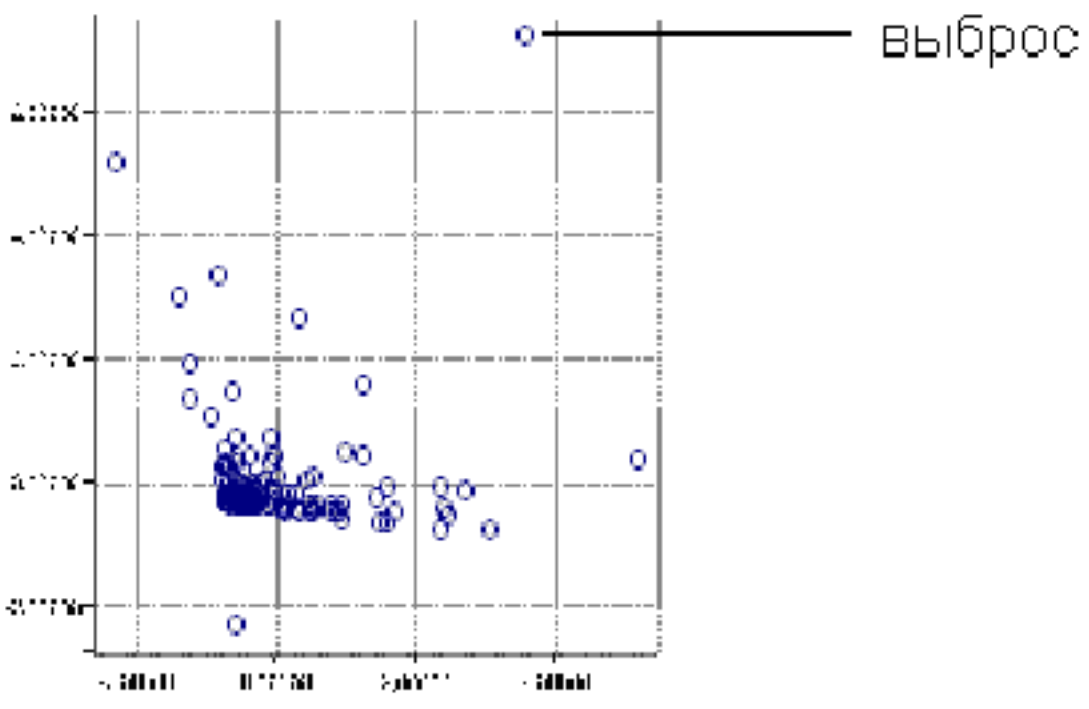

Рис. 2.1.1. Приклад набору даних з викидами

Нейронні 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.

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 .

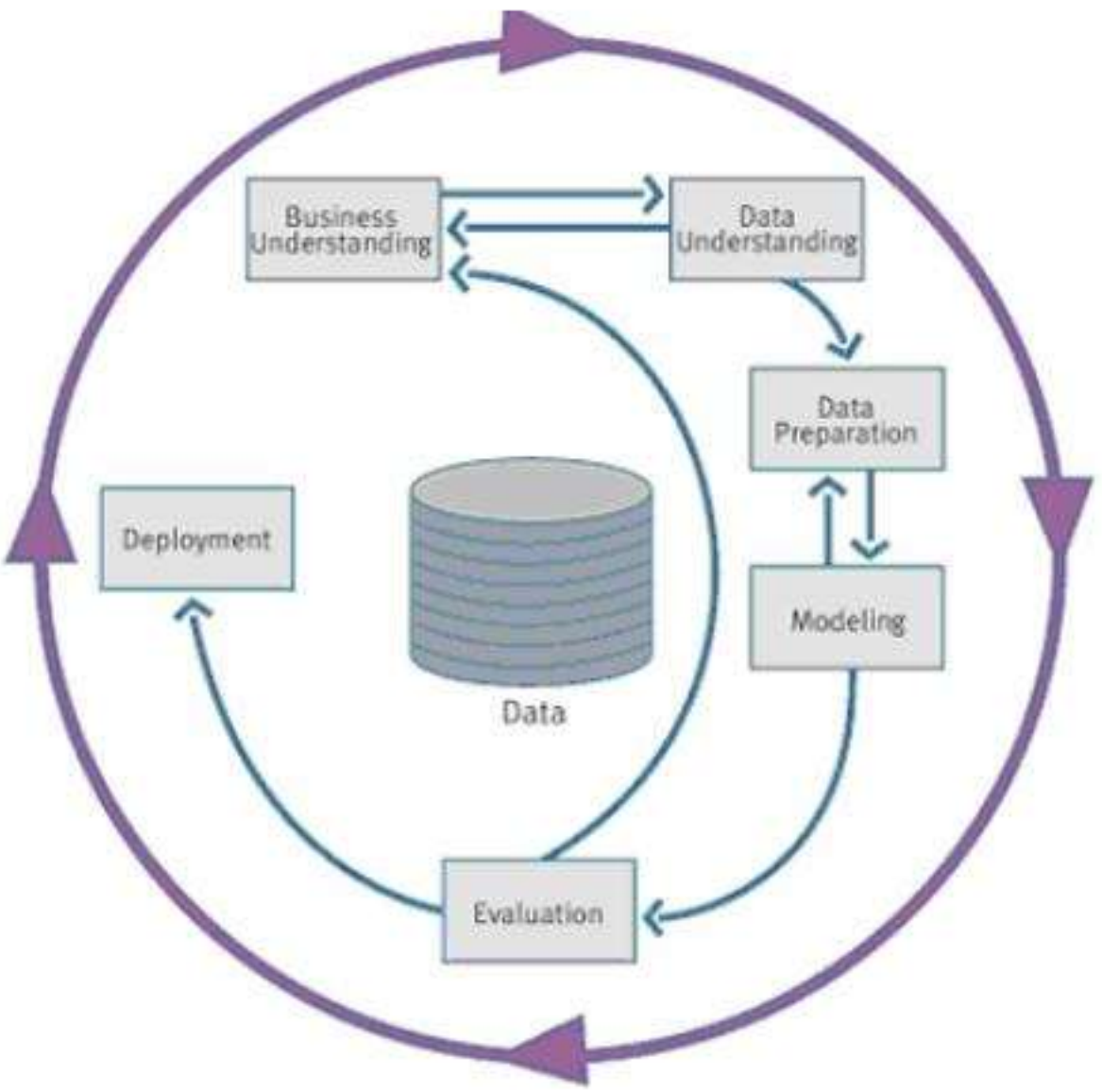

Рис. 2.2.1. Фази, рекомендовані моделлю CRISP-DM

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.

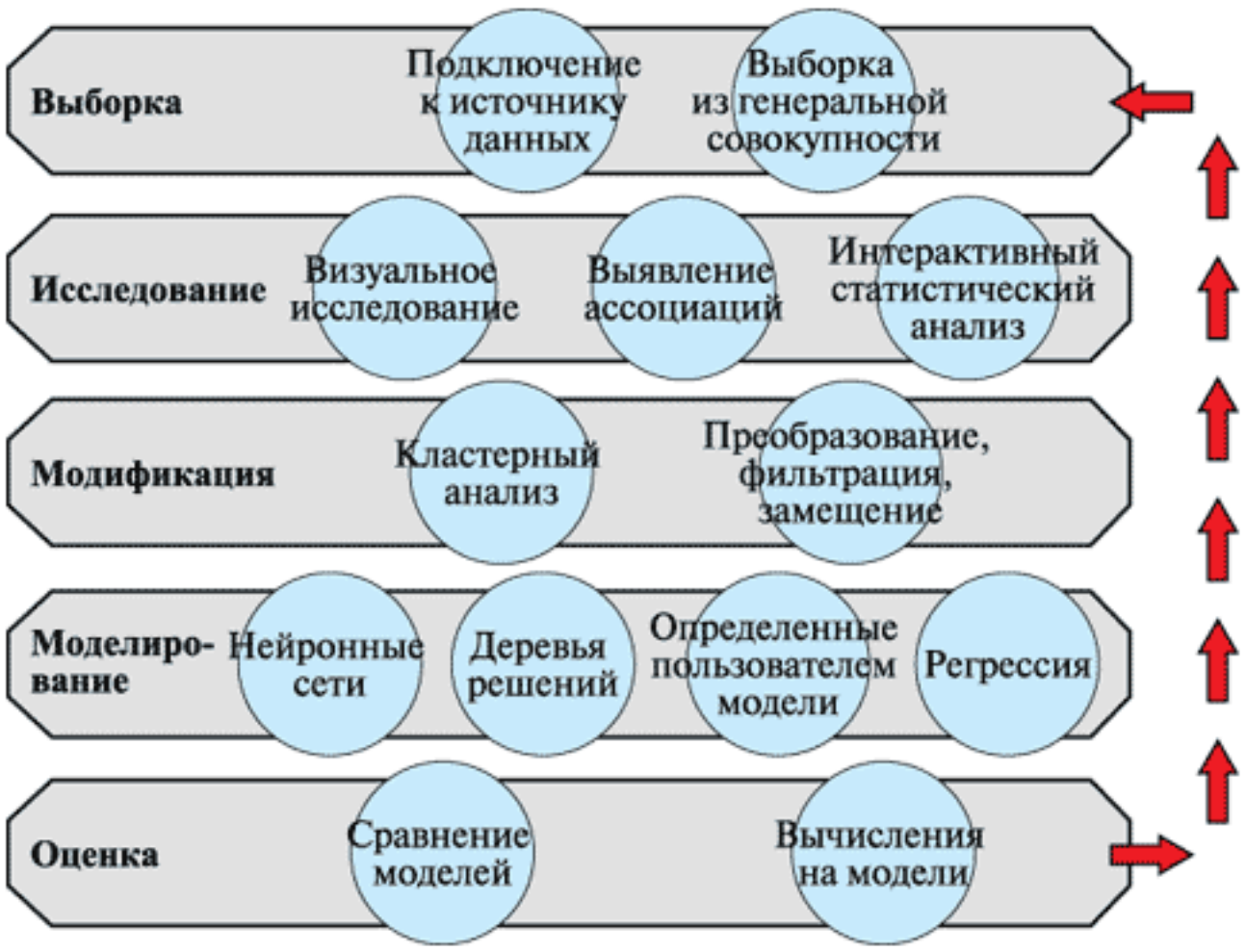

Рис.2.2.2. Методологія розробки проекту Data Mining відповідно до методології SEMMA

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.

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 і сховища даних .

Характеристика та основні відмінності OLAP і OLTP

OLAP	OLTP
Сховище даних має включати як внутрішні корпоративні дані, так і зовнішні дані	Основним джерелом інформації, що надходить в оперативну БД, є діяльність корпорації, а для проведення аналізу даних потрібне залучення зовнішніх джерел інформації (наприклад, статистичних звітів)
Обсяг аналітичних БД як мінімум на порядок більше обсягу оперативних. для проведення достовірних аналізу і прогнозування в сховище даних потрібно мати інформацію про діяльність корпорації та стан ринку протягом декількох років	Для оперативної обробки потрібні дані за кілька останніх місяців
Сховище даних має містити одноманітно представлену і узгоджену інформацію, максимально відповідає змісту оперативних БД. Необхідна компонента для вилучення і "очищення" інформації з різних джерел. У багатьох великих корпораціях одночасно існують кілька оперативних ІС з власними БД (з історичних причин).	Оперативні БД можуть містити семантично еквівалентну інформацію, представлену в різних форматах, з різними зазначенням часу її надходження, іноді навіть суперечливу
Набір запитів до аналітичної бази даних передбачити неможливо. сховища даних існують, щоб	Системи обробки даних створюються в розрахунку на рішення конкретних завдань. Інформація з БД вибирається

відповідати на нерегламентовані запити аналітиків. Можна розраховувати тільки на те, що запити будуть надходити не надто часто і зачіпати великі обсяги інформації. Розміри аналітичної БД стимулюють використання запитів з агрегатами (сума, мінімальне, максимальне, середнє значення і т.д.)	часто і невеликими порціями. Зазвичай набір запитів до оперативної БД відомий вже при проектуванні
При малої мінливості аналітичних БД (тільки при завантаженні даних) виявляються розумними впорядкованість масивів, більш швидкі методи індексації при масовій вибірці, зберігання заздалегідь агрегованих даних	Системи обробки даних за своєю природою є сильно мінливими, що враховується в використовуваних СУБД (нормалізована структура БД, рядки зберігаються неупорядоченно, B-дерева для індексації, транзакційність)
Інформація аналітичних БД настільки критична для корпорації, що потрібні велика грануляція захисту (індивідуальні права доступу до певних рядках і / або стовпцями таблиці)	Для систем обробки даних зазвичай вистачає захисту інформації на рівні 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 - сервернимі 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 .

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.

КОНТРОЛЬНА РОБОТА

ТЕОРЕТИЧНІ ПИТАННЯ

1.Статистичний аналіз даних.
2. Методи первісної обробки даних.
3. Кластерний аналіз.
4. Регресія. Кореляційний і дисперсійний аналіз.
5. Методи інтелектуального аналізу даних (Data Mining).
6. Стандарти та інструменти Data Mining.
7. OLAP-системи.
8. Пошук асоціаційних правил. Метод Apriori.

ПРАКТИЧНЕ ЗАВДАННЯ
Надбудови інтелектуального аналізу даних для Microsoft Office

Завдання 1. *Встановіть надбудови інтелектуального аналізу даних для Microsoft Office. Виконайте необхідну конфігурацію MS SQL Server для роботи з надбудовами. Створіть і протестуйте підключення.*

Завдання 2. *Підготовлений набір даних (для прикладу, можна взяти наведений на рис. 14) відформатуйте як таблицю. Переконайтеся, що ви можете отримати доступ до вкладки з інструментами інтелектуального аналізу таблиць.*

Один з можливий варіантів проведення інтелектуального аналізу даних засобами Microsoft SQL Server - використання надбудов для пакета Microsoft Office. У цьому випадку, джерелом даних для аналізу може служити електронна таблиця Excel. Дані передаються на SQL Server, там обробляються, а результати повертаються Excel для відображення.

Для використання подібної «зв'язки», вам повинен бути доступний MS SQL Server в одній з версій, що підтримують інструменти Data Mining (Enterprise, Developer або з деякими обмеженнями - Standard), MS Office в версії Professional або більш старшої. Самі надбудови інтелектуального аналізу даних для MS Office вільно доступні на сайті Microsoft.

Особливих складнощів процес встановлення додаткових програм не викликає. Єдине, що хочеться відзначити, за замовчуванням пропонується встановлювати не всі компоненти. Але для виконання подальших практичних, краще зробити повну установку (рис. 1).

Наступний крок - конфігурація MS SQL Server для роботи з надбудовами. Для цього використовується майстер «Приступаючи до роботи» (Getting Started), що запускається з головного меню (рис. 2).

Для того, щоб виконати конфігурацію MS SQL Server треба мати права адміністратора.

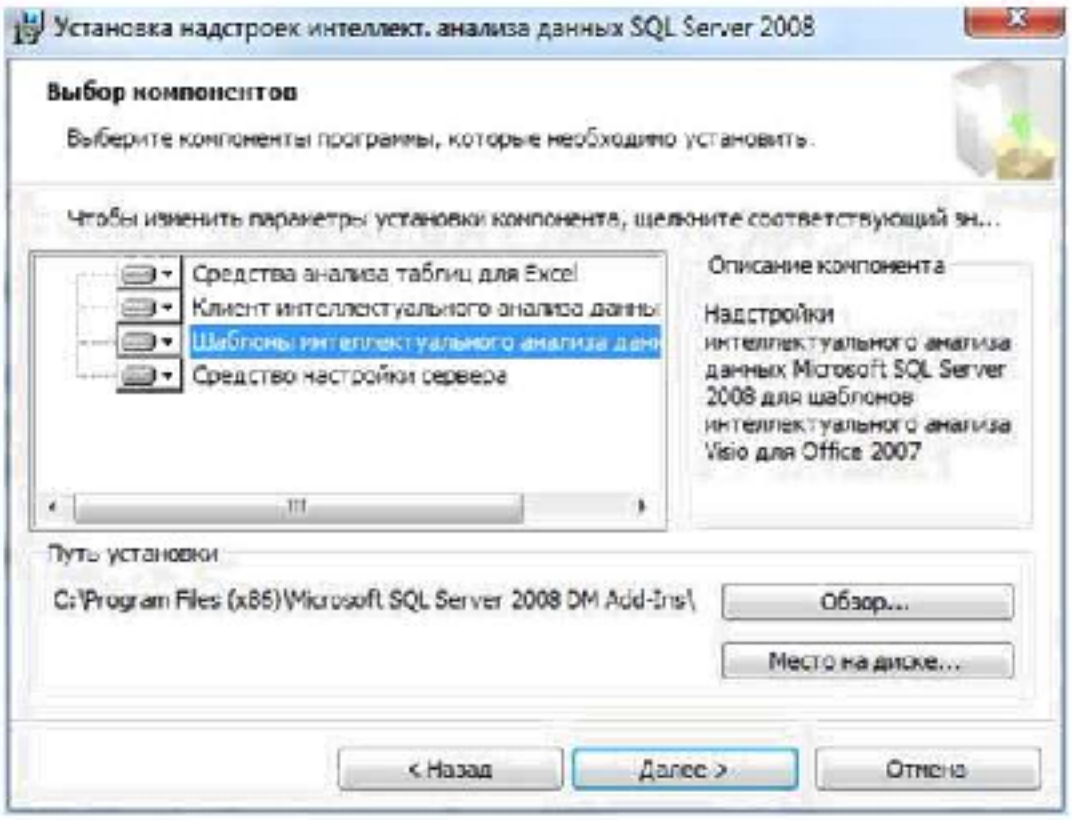

Рис. 1 Вибір компонент для встановлення

Рис. 2. Запуск майстра «Приступая к работе»

На першому кроці майстер пропонує вибрати, завантажити пробну версію MS SQL Server, конфігурувати існуючий екземпляр сервера, де у користувача адміністраторські права, або використовувати сервер, на якому користувач не є адміністратором (рис. 3). Розглянемо варіант 2, при виборі якого майстер покаже вікно з посиланням на інструмент «Засіб налаштування сервера». Його також можна запустити з меню Пуск > Надбудови інтелектуального аналізу даних > Засіб налаштування сервера (рис. 4).

Рис. 3. Вибір сервера баз даних

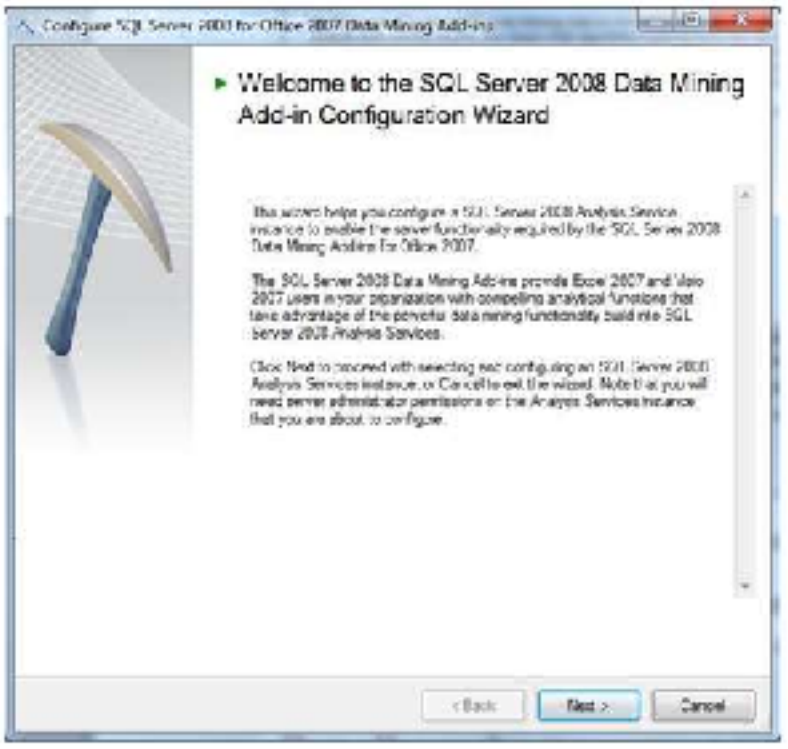

Рис. 4. Засіб налаштування сервера

Наступне вікно пропонує вибрати конфігурацію сервера (рис. 5). За замовчуванням варто обрати «localhost», що відповідає примірнику MS SQL Server, встановленому на той же комп'ютер, на якому запущено «засіб налаштування». Якщо це не так, треба вказати ім'я сервера або для іменованого примірника <ім'я сервера> \ <ім'я екземпляра>.

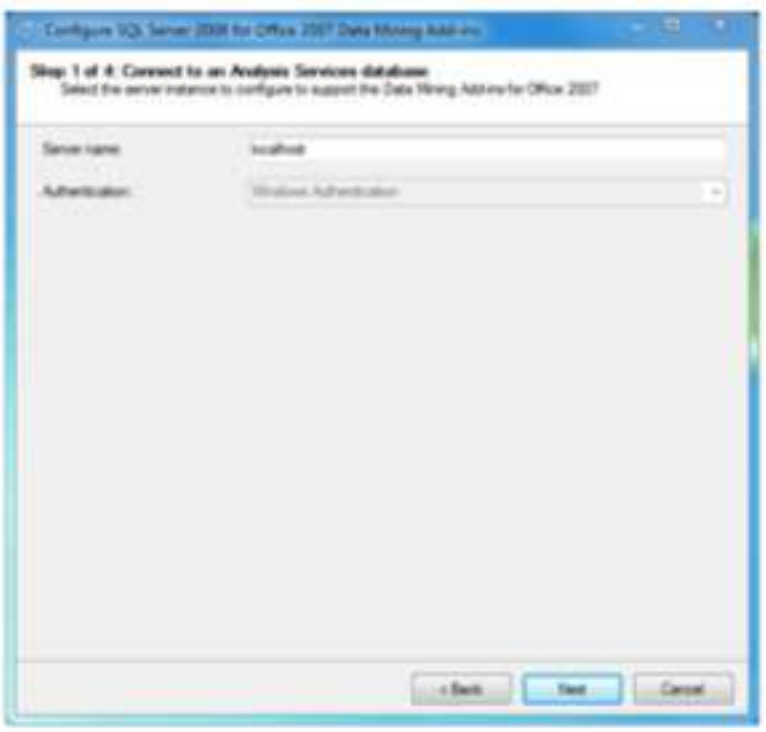

Рис. 5. Вибір екземпляра MS SQL Server

У вікні, представленому на рис. 6, дається дозвіл на створення тимчасових моделей інтелектуального аналізу (Allow creating temporary mining models). Тимчасова модель відрізняється від постійної тим, що створюється тільки на час сеансу користувача. Коли користувач, який проводить аналіз за допомогою надбудов, завершить сесію (закриє Excel), модель буде видалена, але результати аналізу збережуться в електронній таблиці. Постійна модель автоматично не видаляється, зберігається на сервері, і до роботи з нею можна повернутися. Після цього пропонується створити нову базу даних аналітичних служб (рис. 7) або вибрати для роботи існуючу.

Рис. 6.Встановлення дозволу для створення тимчасових моделей інтелектуального аналізу

Рис. 7 Створення або вибір бази даних аналітичних служб

У вікні, представленому на рис. 8, можна додати користувачів до списку адміністраторів створеної бази даних. Це потрібно для створення на сервері постійних моделей. Якщо використовувати лише тимчасові моделі, права адміністратора користувачеві необов'язкові.

Рис. 8. Додавання користувачів в список адміністраторів обраної бази

Після закінчення налаштувань, можна відкрити Excel (а при використанні майстра «Приступаючи до роботи», він буде запущений

автоматично з документом «Зразки даних ...») і протестувати підключення до сервера. Для цього треба перейти на вкладку DataMining і в розділі Connection (рис. 9) натиснути кнопку DMAddinsDB. З'явиться вікно, яке відображає налаштовані з'єднання. Кнопка Test Connection дозволяє перевірити підключення. Якщо налаштованого з'єднання немає, і кнопка DMAddinsDB виглядає як на рис. 11, то потрібно створити нове з'єднання, вибравши у вікні Analysis Services Connection (рис. 10) кнопку New.

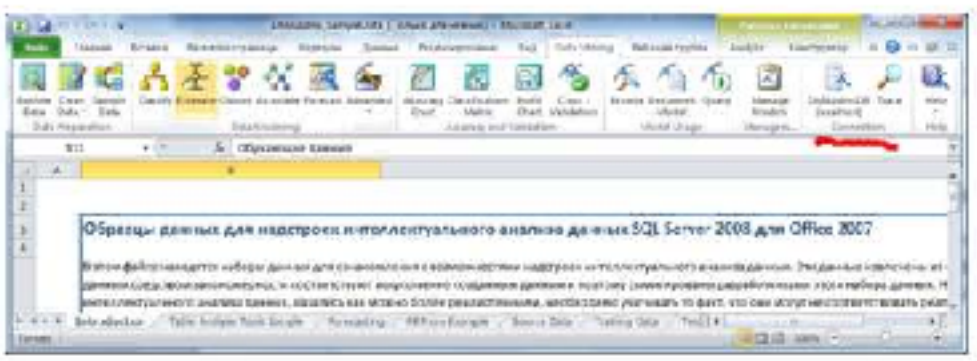

Рис. 9. Вкладка DataMining

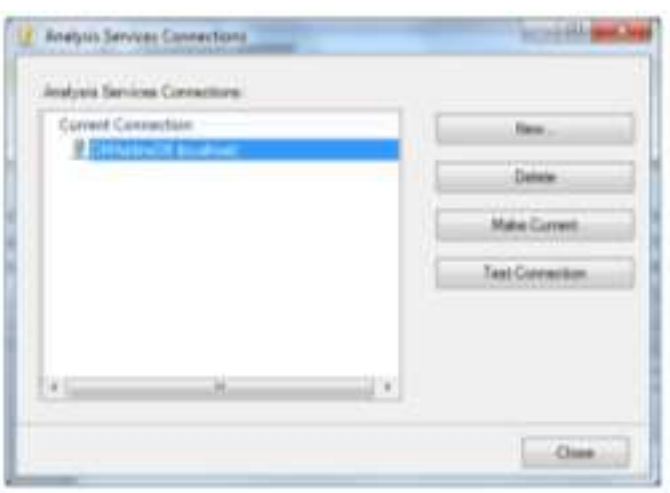

Рис. 10.Налаштовані з'єднання

Рис. 11. Налаштованих з'єднання немає

При створенні нового підключення (рис. 12) треба вказати сервер, до якого плануєте підключатися, і в розділі Catalog name рекомендується явно вказати базу даних, з якої будуть працювати надбудови. Коли з'єднання створено і перевірено, можна починати роботу.

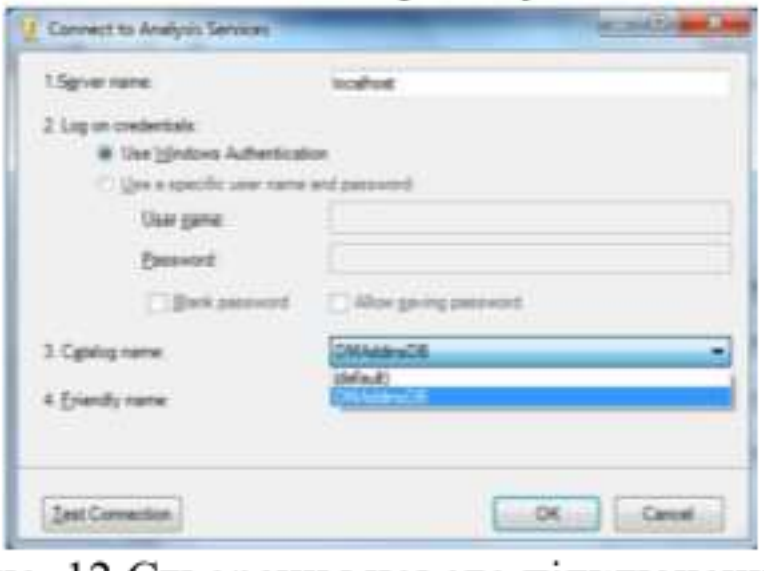

Рис. 12.Створення нового підключення

У наступних декількох практичних завданнях буде використовуватися вже підготовлений набір даних для аналізу. Якщо ви плануєте працювати з власними даними, необхідно враховувати, що інструменти інтелектуального аналізу таблиць працюють з даними, відформатовані у вигляді таблиці. Тому Ваші дані в Excel потрібно виділити і вибрати «Форматувати як таблицю» (рис. 13). Після цього треба вибрати стиль таблиці і вказати заголовок. Вкладка Analyze з інструментами Table Analysis Tools з'явиться при натисканні в області таблиці (рис. 14).

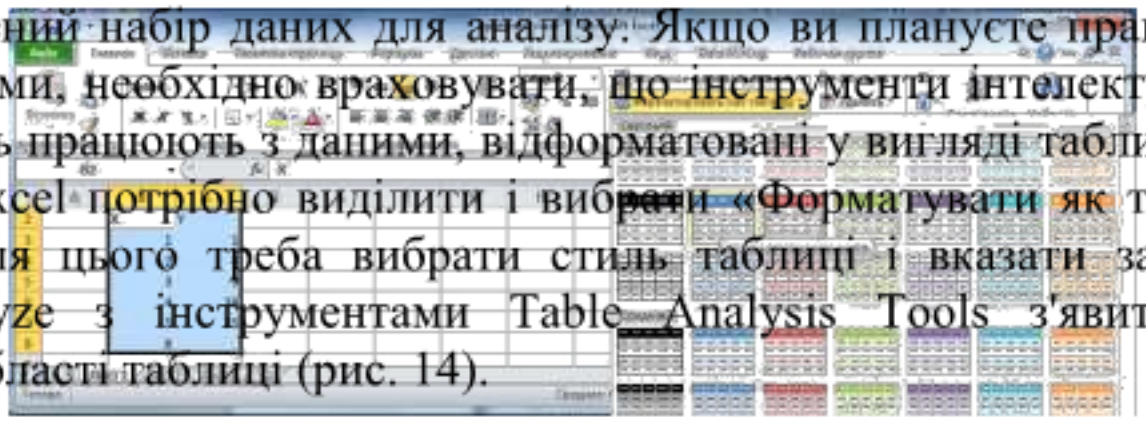

Рис. 13. Форматування підготовлених даних

Рис. 14. Вкладка з інструментами інтелектуального аналізу даних

Перелік питань на залік

1. Основи статистичного аналізу даних..
2. Методи збору і підготовки вихідного набору даних.
3. Методи первісної обробки даних.
4. Методи дослідження структури даних: візуалізація та автоматичне групування даних.
5. Кластерний аналіз. Ієрархічна та секційна кластеризації.

6. Методи кластеризації. Растрова кластеризація об'єктів. Лінійний дискримінантний аналіз.
7. Побудова канонічних та класифікаційних функцій.
8. Регресія. Кореляційний і дисперсійний аналіз.
9. Кореляційний і регресійний аналіз даних.
10. Множинний регресійний аналіз. Лінійна множинна регресійна модель. Дисперсійний аналіз.
11. Статистична обробка тимчасових рядів і прогнозування.
12. Методи інтелектуального аналізу даних (Data Mining).
13. Базові методи. Підготовка початкових даних.
14. Нечітка логіка.
15. Нейроні мережі.
16. Задачі Data Mining. Процес Data Mining.
17. Стандарти та інструменти Data Mining.
18. Стандарт CWM. Стандарт CRISP. Стандарт PMML.
19. Структури та задачі стандартів. Інші стандарти Data Mining.
20. OLAP-системи.
21. Багатовимірна модель даних.
22. Визначення OLAP-системи. Архітектура OLAP-системи.
23. Концептуальні багатовимірні представлення. Правила Кодда.
24. Пошук асоціаційних правил.
25. Метод Apriori.
26. Асоціаційні правила.
27. Сіквенціальний аналіз.
28. Різновиди задач пошуку асоціаційних правил.
29. Метод Apriori.
30. Різновиди методу Apriori.

СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

Основні рекомендовані джерела

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.

Додаткові рекомендовані джерела

8. Спирли Эрик – Корпоративные хранилища данных. Планирование, разработка и реализация. Том 1 – .: Пер с англ. – М.: Издательский дом «Вильямс», 2009. – 400с.: ил.

Олешко Тамара Іванівна

МЕТОДИЧНІ РЕКОМЕНДАЦІЇ
до виконання контрольноїроботи з дисципліни
з дисципліни «Інструментальні засоби статистичного та інтелектуального аналізу даних»
для студентів заочної форми навчання

Галузь знань: 05 "Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»
Освітня професійна програма «Економічна кібернетика»