Маріупольський державний університет

Кафедра математичних методів та системного аналізу

МАТЕМАТИЧНІ МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Навчальний посібник
для здобувачів першого рівня вищої освіти
спеціальності 124 Системний аналіз

МДУ
Маріуполь - 2021

УДК 004.4:004.05

Розглянуто та затверджено Вченою радою Маріупольського державного університету як навчальний посібник для студентів спеціальності «Системний аналіз» протокол № 8 від 24.02.2021р.

Шабельник Тетяна Володимирівна
Дяченко Оксана Федорівна

Математичні методи інтелектуального аналізу даних: [навчальний посібник для здобувачів першого рівня вищої освіти спеціальності 124 Системний аналіз] / Тетяна Шабельник, Оксана Дяченко,. – Маріуполь: МДУ, 2021. – 163 с.

Рецензенти:
Мінц О.Ю. – доктор економічних наук, доцент, в.о. завкафедри фінансів та банківської справи ДВНЗ «Приазовський державний технічний університет»,
Тонких Л.С. – кандидат фізико-математичних наук, доцент кафедри «Природничо-наукових та гуманітарних дисциплін» Азовського морського інституту Національного університету «Одеська морська академія»

Навчальний посібник з дисципліни «Математичні методи інтелектуального аналізу даних» призначений для здобувачів вищої освіти ОП 124 «Системний аналіз» відповідно з базовим курсом підготовки бакалаврів, містить теоретичний матеріал, методичні рекомендації для виконання практичних завдань, а також питання для самоконтролю. Навчальний посібник враховує сучасні тенденції кредитно-модульної системи та Болонських ініціатив. Зміст розділів посібника відповідає робочій програмі дисципліни та містить інформацію щодо певного модуля дисципліни.

Навчальний посібник буде корисний для здобувачів вищої освіти, викладачів та науковців для набуття практичних навичок в галузі організації інтелектуального аналізу даних та сучасних інформаційно-комунікаційних технологій в професійній діяльності.

ЗМІСТ

ПЕРЕДМОВА 5
ТЕМА 1 ВСТУП ДО ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ 7
1.1 Порівняння статистики, машинного навчання та інтелектуального аналізу даних 7
1.2 Основні задачі інтелектуального аналізу даних 9
1.3 Класифікація методів інтелектуального аналізу даних 11
1.4 Сфери застосування задач інтелектуального аналізу даних 18
Самостійна робота №1 28
Практична (семінарська) робота №1 29
Тест до теми 1 29
ТЕМА 2 НАБІР ДАНИХ ТА ЇХ ВЛАСТИВОСТІ 33
2.1 Атрибути даних 33
2.2 Шкали вимірювання 35
2.3 Типи наборів даних 38
2.4 Описова статистика 42
Самостійна робота №2 48
Практична (лабораторна) робота №2 49
Тест до теми 2 51
ТЕМА 3 ЕЛЕМЕНТИ ТЕОРІЇ МНОЖИН 55
3.1 Множина та її елементи 55
3.2 Множина і підмножини 59
3.3 Операції над множинами 61
3.4 Основні закони алгебри множин 62
3.5 Добуток множин 65
Самостійна робота №3 66
Практична робота №3 69
Тести до теми 3 72
ТЕМА 4 МЕТРИЧНІ ОСНОВИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ 75
4.1 Відношення 75
4.2 Метричні простори 84
4.3 Приклади метрик 86
4.4 Відкриті і замкнуті множини 88

4.5 Зменшення вимірності. Векторна репрезентація мови 92
Самостійна робота №4 97
Практична робота №4 98
Тест до теми 4 99
ТЕМА 5 СТАТИСТИЧНІ МЕТОДИ АНАЛІЗУ ДАНИХ 102
5.1 Кореляційний аналіз 102
5.2 Регресійний аналіз 105
Самостійна робота №5 112
Практична (лабораторна) робота №5 119
Тест до теми 5 127
ТЕМА 6 ЗАДАЧІ КЛАСИФІКАЦІЇ Й КЛАСТЕРИЗАЦІЇ 130
6.1 Задача класифікації 130
6.2 Метод опорних векторів 134
6.3 Постановка задачі кластеризації 137
6.4 Методи кластерного аналізу 139
Самостійна робота №6 147
Практична (лабораторна) робота №6 147
Тест до теми 6 155
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 159
ГЛОСАРІЙ 162

ПЕРЕДМОВА

Підготовка майбутніх фахівців з системного аналізу передбачає формування у студентів здатності розв'язувати складні спеціалізовані задачі та практичні проблеми системного аналізу у професійній діяльності або у процесі навчання, що передбачає застосування теоретичних положень та математичних методів і інформаційних технологій, створення і дослідження складних систем різної природи (інформаційних, економічних, фінансових, соціальних, політичних, технічних, організаційних, екологічних тощо) і характеризується комплексністю та невизначеністю умов.

Розв'язання завдань технічного та біомедичного діагностування, розпізнавання образів, класифікації та прогнозування пов'язано з необхідністю виконання аналізу даних.

Дисципліна «Математичні методи інтелектуального аналізу даних» є важливим елементом професійної підготовки бакалаврів за спеціальністю 124 Системний аналіз. Він формує уявлення про математичний апарат, який застосовується для вирішення задач інтелектуального аналізу даних.

Метою даного посібника є формування системи знань і практичних навичок, які можуть використовуватися при вирішенні практичних завдань розпізнавання образів, прийняття рішень, класифікації та прогнозування.

В навчальному посібнику розглядаються методи і технології, що застосовуються в інтелектуальному аналізі даних і базуються на поняттях подібності, близькості, аналогії. Ідея подібності властива людському мисленню, це породило цілий комплекс підходів для всіх фундаментальних задач інтелектуального аналізу даних, серед яких основна увага в курсі приділена класифікації, кластеризації.

Представлена теоретична основа для побудови, реалізації та аналізу широкого спектра моделей і методів інтелектуального аналізу даних. Розглянуто методи побудови і

обчислення функцій подібності, узгодження подібності на різних множинах об'єктів, синтез нових способів порівняння об'єктів на базі вже наявних. Розглянуто комплекс технологій, призначений для ефективного представлення та обробки метричної інформації обчислювальними системами.

Для спрощення самостійного опрацювання та кращого засвоєння матеріалу книги наприкінці кожного розділу наведено контрольні питання, а також практичні та тестові завдання.

Видання орієнтоване на студентів комп'ютерних спеціальностей закладів вищої освіти, а також може використовуватися педагогічними працівниками та практичними фахівцями.

ТЕМА 1 ВСТУП ДО ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Теоретичний матеріал до лекції

План

1.1 Порівняння статистики, машинного навчання та інтелектуального аналізу даних

1.2 Основні задачі інтелектуального аналізу даних

1.3 Класифікація методів інтелектуального аналізу даних

1.4 Сфери застосування задач інтелектуального аналізу даних

Основні поняття: інтелектуальний аналіз даних, статистика, машинне навчання, штучний інтелект, класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз і виявлення відхилень, оцінювання, аналіз зв'язків, логічні методи, методи крос-табуляції, статистичні методи, кібернетичні методи, лінійна регресія, нейроні мережі, дерево рішень, метод найближчого сусіда.

1.1 Порівняння статистики, машинного навчання та інтелектуального аналізу даних

Глобальна комп'ютеризація, прогрес в сфері інформаційних технологій обумовлюють збільшення об'ємів інформації. Загальний об'єм даних подвоюється кожні 2 роки. Збереження такого об'єму інформації потребує немалих ресурсів та зусиль. Але найбільша проблема полягає не в збереженні даних, а в їх обробці. Таким чином необхідні ефективні механізми для аналізу великих даних. Одним з таких напрямків є інтелектуальний аналіз даних.

Інтелектуальний аналіз даних (Data Mining) – це мультидисциплінарна область, що виникла і розвивається на базі таких наук як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних. Інтелектуальний аналіз даних вивчає процес знаходження нових, дійсних і потенційно корисних знань в базах даних.

Наведемо порівняння статистики, машинного навчання та штучного інтелекту на стику яких з'явилася технологія інтелектуального аналізу даних.

Статистика – це наука про методи збору даних, їх обробку і аналіз для виявлення закономірностей, властивих досліджуваному явищу. Статистика є сукупністю методів планування експерименту, збору даних, їх подання та узагальнення, а також аналізу і отримання висновків на підставі цих даних. Статистика оперує даними, отриманими в результаті спостережень або експериментів.

Статистика більше, ніж інтелектуальний аналіз даних, базується на теорії та більш зосереджується на перевірці гіпотез.

Машинне навчання можна охарактеризувати як процес отримання програмою нових знань. Мітчелл в 1996 році дав таке визначення: "Машинне навчання – це наука, яка вивчає комп'ютерні алгоритми, що автоматично покращуються під час роботи". Одним з найбільш популярних прикладів алгоритму машинного навчання є нейронні мережі.

Машинне навчання порівняно з інтелектуальним аналізом даних більш евристичне, концентрується на поліпшенні роботи агентів навчання.

Штучний інтелект – науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, що традиційно вважаються інтелектуальними. Термін інтелект (intelligence) походить від латинського intellectus, що означає розум, розумові здібності людини. Відповідно, штучний інтелект (AI, Artificial Intelligence) тлумачиться як властивість автоматичних систем брати на себе окремі функції інтелекту людини. Штучним інтелектом називають властивість інтелектуальних систем виконувати творчі функції, які традиційно вважаються прерогативою людини.

Інтелектуальний аналіз даних характеризує інтеграція теорії і евристики, сконцентрованість на єдиному процесі

аналізу даних, включає очищення даних, навчання, інтеграцію і візуалізацію результатів.

Інтелектуальний аналіз даних може складатися з двох або трьох стадій.

Стадія 1. Виявлення закономірностей (вільний пошук).

Іноді вводять стадію валідації, наступну за стадією вільного пошуку. Мета валідації – перевірка достовірності знайдених закономірностей. Однак, ми будемо вважати валідацію частиною першою стадії, оскільки в реалізації багатьох методів, зокрема, нейронних мереж і дерев рішень, передбачено поділ загальної множині даних на навчальне і перевірочне, що дозволяє перевіряти достовірність отриманих результатів.

Стадія 2. Використання виявлених закономірностей для передбачення невідомих значень (прогностичне моделювання).

Стадія 3. Аналіз винятків – стадія призначена для виявлення і пояснення аномалій, знайдених в закономірностях.

1.2 Основні задачі інтелектуального аналізу даних

Задачі (tasks) інтелектуального аналізу даних іноді називають закономірностями (regularity) або техніками (techniques).

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

Наведемо загальний опис найбільш поширених задач інтелектуального аналізу даних.

Класифікація (Classification)

В результаті рішення задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних – класи; за цими ознаками новий об'єкт можна віднести до того чи іншого класу.

Методи рішення. Для вирішення задачі класифікації можуть використовуватися методи: найближчого сусіда (Nearest Neighbor); k-найближчого сусіда (k-Nearest Neighbor); баєсівські

мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks).

Кластеризація (Clustering)

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

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

Асоціація (Associations)

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

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

Найбільш відомий алгоритм рішення задачі пошуку асоціативних правил – алгоритм Apriori.

Послідовна асоціація (sequential association)

Послідовність дозволяє знайти тимчасові закономірності між транзакціями. Завдання послідовності подібна асоціації, але її метою є встановлення закономірностей не між одночасно наступаючими подіями, а між подіями, пов'язаними в часі (тобто відбуваються з деяким інтервалом у часі). Іншими словами, послідовність визначається високою ймовірністю ланцюжка пов'язаних у часі подій. Фактично, асоціація є окремим випадком послідовності з тимчасовим кроком, рівним нулю. Цю задачу також називають задачею знаходження послідовних шаблонів (sequential pattern).

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

Приклад. Після покупки квартири мешканці в 60% випадків протягом двох тижнів купують холодильник, а

протягом двох місяців в 50% випадків купується телевізор. Рішення даного завдання широко застосовується в маркетингу та менеджменті, наприклад, при управлінні циклом роботи з клієнтом (Customer Lifecycle Management).

Прогнозування (Forecasting)

В результаті рішення задачі прогнозування на основі особливостей історичних даних оцінюються пропущені або ж майбутні значення цільових чисельних показників. Для вирішення таких завдань широко застосовуються методи математичної статистики, нейронні мережі та ін.

Визначення відхилень або викидів (Deviation Detection)

Мета рішення даного завдання – виявлення та аналіз даних, які найбільш відрізняються від загальної множини даних, виявлення так званих нехарактерних шаблонів.

Оцінювання (Estimation)

Завдання оцінювання зводиться до передбачення неперервних значень ознаки.

Аналіз зв'язків (Link Analysis)

Задача знаходження залежностей в наборі даних.

Візуалізація (Visualization)

В результаті візуалізації створюється графічний образ аналізованих даних. Для вирішення завдання візуалізації використовуються графічні методи, що показують наявність закономірностей в даних.

Приклад методів візуалізації – представлення даних в 2D і 3D вимірах.

Підведення підсумків (Summarization)

Задача, мета якої – опис конкретних груп об'єктів з аналізованого набору даних.

1.3 Класифікація методів інтелектуального аналізу даних

Розглянемо кілька відомих класифікацій методів інтелектуального аналізу даних за різними ознаками.

1.3.1 Класифікація технологічних методів. Всі методи інтелектуального аналізу даних підрозділяються на дві великі

групи за принципом роботи з вихідними навчальними даними. У цій класифікації верхній рівень визначається на підставі того, чи зберігаються дані після інтелектуального аналізу даних або вони дистилюється для подальшого використання.

1.3.1.1 **Безпосереднє використання даних, або збереження даних.** У цьому випадку вихідні дані зберігаються в явному детальному вигляді і безпосередньо використовуються на стадіях прогностичного моделювання та/або аналізу винятків. Проблема цієї групи методів – при їх використанні можуть виникнути складності аналізу надвеликих баз даних. Методи цієї групи: кластерний аналіз, метод найближчого сусіда, метод k-найближчого сусіда, міркування за аналогією.

1.3.1.2 **Виявлення і використання формалізованих закономірностей, або дистиляція шаблонів**. При технології дистиляції шаблонів один зразок (шаблон) інформації витягується з вихідних даних і перетворюється в якісь формальні конструкції, вид яких залежить від використовуваного методу інтелектуального аналізу даних. Цей процес виконується на стадії вільного пошуку, у першій же групи методів дана стадія в принципі відсутня. На стадіях прогностичного моделювання та аналізу винятків використовуються результати стадії вільного пошуку, вони значно компактніше самих баз даних. Нагадаємо, що конструкції цих моделей можуть бути по трактованим аналітиком або нетрактованим ("чорними ящиками"). Методи цієї групи: логічні методи; методи візуалізації; методи крос-табуляції; методи, засновані на рівняннях.

1.Логічні методи, або методи логічної індукції, включають: нечіткі запити і аналізи; символьні правила; дерева рішень; генетичні алгоритми.

Методи цієї групи є такими, що найбільш інтерпретуються – вони оформлюють знайдені закономірності, в досить

прозорому вигляді для користувача. Отримані правила можуть включати неперервні і дискретні змінні. Слід зауважити, що дерева рішень можуть бути легко перетворені в набори символьних правил шляхом генерації одного правила по шляху від кореня дерева до його термінальній вершини. Дерева рішень і правила фактично є різними способами вирішення однієї задачі і відрізняються лише за своїми можливостями. Крім того, реалізація правил здійснюється більш повільними алгоритмами, ніж індукція дерев рішень.

2. Методи крос-табуляції: агенти, баєсівські мережі, крос-таблична візуалізація. Останній метод не зовсім відповідає одному з властивостей інтелектуального аналізу даних – самостійного пошуку закономірностей аналітичною системою. Однак, надання інформації у вигляді крос-таблиць забезпечує реалізацію основного завдання інтелектуального аналізу даних – пошук шаблонів, тому цей метод можна також вважати одним з методів інтелектуального аналізу даних.

3. Методи на основі рівнянь. Методи цієї групи висловлюють виявлені закономірності у вигляді математичних виразів – рівнянь. Отже, вони можуть працювати лише з чисельними змінними, і змінні інших типів повинні бути закодовані відповідним чином. Це дещо обмежує застосування методів даної групи, проте вони широко використовуються при вирішенні різних завдань, особливо завдань прогнозування. Основні методи даної групи: статистичні методи та нейронні мережі.

Статистичні методи найбільш часто застосовуються для вирішення завдань прогнозування. Існує безліч методів статистичного аналізу даних, серед них, наприклад, кореляційно-регресійний аналіз, кореляція рядів динаміки, виявлення тенденцій динамічних рядів.

1.3.2 Класифікація за підходами до навчання математичних моделей.

Ця схема поділу заснована на різних підходах до навчання математичних моделей.

У цій класифікації розрізняють дві групи методів:

– *статистичні методи*, засновані на використанні усередненого накопиченого досвіду, який відображений в ретроспективних даних;
– *кібернетичні методи*, що включають безліч різнорідних математичних підходів.

Слід зазначити, що більшість авторитетних джерел вважають статистичні методи аналізу частиною математичного інструментарію інтелектуального аналізу даних.

Недолік такої класифікації: і статистичні, і кібернетичні алгоритми тим чи іншим чином спираються на зіставлення статистичного досвіду з результатами моніторингу поточної ситуації.

Перевагою такої класифікації є її зручність для інтерпретації – вона використовується при описі математичних засобів сучасного підходу до вилучення знань з масивів спостережень (оперативних і ретроспективних), тобто в задачах інтелектуального аналізу даних.

Розглянемо докладніше представлені вище групи.

Статистичні методи інтелектуального аналізу даних.

В цих методах є чотири взаємопов'язані розділи:

– попередній аналіз природи статистичних даних;
– виявлення зв'язків і закономірностей;
– багатовимірний статистичний аналіз;
– динамічні моделі і прогноз на основі часових рядів.

Арсенал статистичних методів інтелектуального аналізу даних класифікований на чотири групи методів:

1. дескриптивний аналіз і опис вхідних даних;
2. аналіз зв'язків (кореляційний і регресійний аналіз, факторний аналіз, дисперсійний аналіз);

3. багатовимірний статистичний аналіз (компонентний аналіз, дискримінантний аналіз, багатовимірний регресійний аналіз, канонічні кореляції та ін.);

4. аналіз часових рядів (динамічні моделі і прогнозування).

Кібернетичні методи інтелектуального аналізу даних

Другий напрямок інтелектуального аналізу даних – це безліч підходів, об'єднаних ідеєю комп'ютерної математики та використання теорії штучного інтелекту.

До цієї групи належать такі методи:

– штучні нейронні мережі (розпізнавання, кластеризація, прогноз);
– еволюційне програмування (алгоритми методу групового обліку аргументів);
– генетичні алгоритми (оптимізація);
– асоціативна пам'ять (пошук аналогів, прототипів);
– нечітка логіка;
– дерева рішень;
– системи обробки експертних знань.

1.3.3 Класифікація методів за завданнями інтелектуального аналізу даних. Відповідно до такої класифікації виділяємо дві групи.

Перша – методи інтелектуального аналізу даних спрямовані на вирішення задач сегментації (тобто задач класифікації і кластеризації) і задач прогнозування.

Прогнозуючі методи використовують значення одних змінних для передбачення/прогнозування невідомих значень інших (цільових) змінних.

До методів, спрямованих на отримання прогнозів, відносяться такі методи: нейронні мережі, дерева рішень, лінійна регресія, метод найближчого сусіда, метод опорних векторів і ін.

Друга – методи інтелектуального аналізу даних, спрямовані на отримання описових і прогностичних результатів.

Описові методи служать для знаходження шаблонів або зразків, що описують дані, які піддаються інтерпретації з точки зору аналітика.

До методів, спрямованих на отримання описових результатів, відносяться ітеративні методи кластерного аналізу, в тому числі: алгоритм k-середніх, k-медіани, ієрархічні методи кластерного аналізу, карти Кохонена, методи крос-табличної візуалізації, різні методи візуалізації та інші.

1.3.4 Властивості методів інтелектуального аналізу даних

Різні методи інтелектуального аналізу даних характеризуються певними властивостями, які можуть бути визначальними при виборі методу аналізу даних. Методи можна порівнювати між собою, оцінюючи характеристики їх властивостей.

Серед основних властивостей і характеристик методів інтелектуального аналізу даних розглянемо наступні: точність, масштабованість, інтерпретованість, перевірюваність, трудомісткість, гнучкість, швидкість і популярність.

Масштабованість – властивість обчислювальної системи, що забезпечує передбачуваний ріст системних характеристик, наприклад, швидкості реакції, загальної продуктивності та ін. при додаванні до неї обчислювальних ресурсів.

У таблиці 1.1 наведено порівняльну характеристику деяких поширених методів. Оцінка кожної з характеристик проведена наступними категоріями, в порядку зростання: надзвичайно низька, дуже низька, низька / нейтральна, нейтральна / низька, нейтральна, нейтральна / висока, висока, дуже висока.

Таблиця 1.1.

Порівняльна характеристика методів
Інтелектуального аналізу даних

Алгоритм	**Точність**	**Маштабованість**	**Інтерпретованість**	**Придатність до використання**	**Трудомісткість**	**Різнобічність**	**Бистрота**	**Популярність широта використання**
класичні методи (лінійна регресія)	нейтральна	висока	висока нейтральна	висока	нейтральна	нейтральна	висока	низька
нейронні мережі	висока	низька	низька	низька	нейтральна	низька	дуже низька	низька
методи візуалізації	висока	дуже низька	висока	висока	дуже висока	низька	надзвичайно низька	висока /нейтральна
дерева рішень	низька	висока	висока	висока / нейтральна	висока	висока	висока /нейтральна	висока /нейтральна
поліноміальні нейронні мережі	висока	нейтральна	низька	висока / нейтральна	нейтральна / низька	нейтральна	низька / нейтральна	нейтральна
k-найближчого сусіда	низька	дуже низька	висока /нейтральна	нейтральна	нейтральна/ низька	низька	висока	низька

Як видно з таблиці, кожен з методів має свої сильні і слабкі сторони. Але жоден метод не може забезпечити вирішення всього спектру завдань інтелектуального аналізу даних.

Більшість інструментів інтелектуального аналізу даних, реалізують відразу кілька методів, наприклад, дерева рішень, індукцію правил і візуалізацію, або ж нейронні мережі і візуалізацію.

В універсальних прикладних статистичних пакетах (SPSS, SAS, Statistica) реалізується широкий спектр найрізноманітніших методів (як статистичних, так і кібернетичних). Слід враховувати, що для їх використання, а також для інтерпретації результатів роботи статистичних методів потрібні спеціальні знання в галузі статистики.

1.4 Сфери застосування задач інтелектуального аналізу даних

Технологія інтелектуального аналізу даних використовується практично у всіх сферах діяльності людини, де накопичені ретроспективні дані.

Будемо розглядати чотири основні сфери застосування технології інтелектуального аналізу даних детально: наука, бізнес, дослідження для уряду і Web-напрямок.

Основні напрямки застосування інтелектуального аналізу даних для вирішення бізнес-завдань: банківська справа, фінанси, страхування, виробництво, телекомунікації, електронна комерція, маркетинг, фондовий ринок та інші.

Основні напрямки застосування інтелектуального аналізу даних для вирішення завдань державного рівня: пошук осіб, які ухиляються від податків; засіб боротьби з тероризмом.

Основні напрямки застосування інтелектуального аналізу даних для наукових досліджень: медицина, біологія, молекулярна генетика і генна інженерія, біоінформатика, астрономія, прикладна хімія, дослідження, що стосуються наркотичної залежності, і інші.

Основні напрямки застосування інтелектуального аналізу даних для вирішення Web-завдань: пошукові машини (search engines), лічильники та інші.

1.4.1 Застосування інтелектуального аналізу даних для вирішення бізнес-задач.

Банківська справа

Технологія інтелектуального аналізу даних використовується в банківській сфері для вирішення ряду типових задач.

1. Задача "Чи видавати кредит клієнту?"

Класичний приклад застосування інтелектуального аналізу даних в банківській справі – рішення задачі визначення можливої некредитоспроможності клієнта банку. Це завдання також називають аналізом кредитоспроможності клієнта. Без застосування технології інтелектуального аналізу даних завдання вирішується співробітниками банківської установи на основі їх досвіду, інтуїції і суб'єктивних уявлень про те, наскільки клієнт є благонадійним. За схожою схемою працюють системи підтримки прийняття рішень. Такі системи на основі історичної (ретроспективної) інформації і за допомогою методів класифікації виявляють клієнтів, які в минулому не повернули кредит.

Завдання "Чи видавати кредит клієнту?" за допомогою методів інтелектуального аналізу даних вирішується таким чином. Сукупність клієнтів банку розбивається на два класи (які повернули і не повернули кредит); на основі групи клієнтів, що не повернули кредит, визначаються основні "риси" потенційного неплатника; при надходженні інформації про нового клієнта визначається його клас ("поверне кредит", "не поверне кредит").

2. Завдання залучення нових клієнтів банку.

За допомогою інструментів інтелектуального аналізу даних можливо провести класифікацію на "більш вигідних" і "менш вигідних" клієнтів. Після визначення найбільш вигідного сегменту клієнтів банку є сенс проводити більш активну маркетингову політику по залученню клієнтів саме серед знайденої групи.

3. Інші завдання сегментації клієнтів.

Розбиваючи клієнтів за допомогою інструментів інтелектуального аналізу даних на різні групи, банк має можливість зробити свою маркетингову політику більш цілеспрямованою, а тому – ефективною, пропонуючи різним групам клієнтів саме ті види послуг, в яких вони потребують.

4. Управління ліквідністю банку. Прогнозування залишку на рахунках клієнтів.

Проводячи прогнозування часового ряду з інформацією про залишки на рахунках клієнтів за попередні періоди, застосовуючи методи інтелектуального аналізу даних, можна отримати прогноз залишку на рахунках в певний момент в майбутньому. Отримані результати можуть бути використані для оцінки і управління ліквідністю банку.

5. Виявлення випадків шахрайства з кредитними картками.

Для виявлення підозрілих операцій з кредитними картками застосовуються так звані "підозрілі стереотипи поведінки", які визначаються в результаті аналізу банківських транзакцій, які згодом виявилися шахрайськими. Для визначення підозрілих випадків використовується сукупність послідовних операцій на певному часовому інтервалі. Якщо система інтелектуального аналізу даних вважає чергову операцію підозрілою, банківський працівник може, орієнтуючись на цю інформацію, заблокувати операції з певною карткою.

Страхування

Страховий бізнес пов'язаний з певним ризиком. Тут задачі, які вирішуються за допомогою інтелектуального аналізу даних , схожі з завданнями в банківській справі.

Інформація, отримана в результаті сегментації клієнтів на групи, використовується для визначення груп клієнтів. В результаті страхова компанія може з найбільшою вигодою і

найменшим ризиком пропонувати певні групи послуг конкретним групам клієнтів.

Завдання виявлення шахрайства вирішується шляхом знаходження якогось загального стереотипу поведінки клієнтів-шахраїв.

Телекомунікації

У сфері телекомунікацій досягнення інтелектуального аналізу даних можуть використовуватися для вирішення задач, які є типовими для будь-якої компанії, яка працює з метою залучення постійних клієнтів – визначення лояльності цих клієнтів. Необхідність розв'язання таких задач обумовлена жорсткою конкуренцією на ринку телекомунікацій і постійної міграцією клієнтів від однієї компанії в іншу. Як відомо, утримання клієнта набагато дешевше його повернення. Тому виникає необхідність виявлення певних груп клієнтів і розробка наборів послуг, найбільш привабливих саме для них. У цій сфері, так само як і в багатьох інших, важливим завданням є виявлення фактів шахрайства.

Електронна комерція

У сфері електронної комерції інтелектуальний аналіз даних застосовується для формування рекомендаційних систем і класифікації відвідувачів Web-сайтів. Така класифікація дозволяє компаніям виявляти певні групи клієнтів і проводити маркетингову політику відповідно до виявлених інтересів і потреб клієнтів. Технологія інтелектуального аналізу даних для електронної комерції тісно пов'язана з технологією Web Mining.

Промислове виробництво

Особливості промислового виробництва і технологічних процесів створюють хороші передумови для можливості використання технології інтелектуального аналізу даних в ході вирішення різних виробничих завдань. Технічний процес за своєю природою повинен бути контрольованим, а всі його відхилення знаходяться в заздалегідь відомих межах, є певна

стабільність, яка зазвичай не властива більшості задач інтелектуального аналізу даних .

Основні задачі інтелектуального аналізу даних в промисловому виробництві:

– комплексний системний аналіз виробничих ситуацій;

– короткостроковий і довгостроковий прогноз розвитку виробничих ситуацій;

– вироблення варіантів оптимізаційних рішень;

– прогнозування якості виробу в залежності від деяких параметрів технологічного процесу;

– виявлення прихованих тенденцій і закономірностей розвитку виробничих процесів;

– прогнозування закономірностей розвитку виробничих процесів;

– виявлення прихованих чинників впливу;

– виявлення та ідентифікація раніше невідомих взаємозв'язків між виробничими параметрами і факторами впливу;

– аналіз середовища взаємодії виробничих процесів і прогнозування зміни її характеристик;

– вироблення оптимізаційних рекомендацій з управління виробничими процесами;

– візуалізацію результатів аналізу, підготовку попередніх звітів і проектів допустимих рішень з оцінками достовірності та ефективності можливих реалізацій.

Маркетинг

У сфері маркетингу інтелектуальний аналіз даних знаходить дуже широке застосування. Основні питання маркетингу "Що продається?", "Як продається?", "Хто є споживачем?". Кластерний аналіз використовується для вирішення задач сегментації споживачів. Інший поширений набір методів для вирішення задач маркетингу – методи і

алгоритми пошуку асоціативних правил. Також успішно використовується пошук тимчасових закономірностей.

Роздрібна торгівля

У сфері роздрібної торгівлі, як і в маркетингу, застосовуються:

– алгоритми пошуку асоціативних правил (для визначення наборів товарів, які покупці купують одночасно). Виявлення таких правил допомагає розміщувати товари на прилавках торгових залів, виробляти стратегії закупівлі товарів і їх розміщення на складах і т.д.

– використання часових послідовностей, наприклад, для визначення необхідних обсягів запасів товарів на складі.

– методи класифікації та кластеризації для визначення груп або категорій клієнтів, знання про яких сприяє успішному просуванню товарів.

Фондовий ринок

Ось список завдань фондового ринку, які можна вирішувати за допомогою технології інтелектуального аналізу даних :

– прогнозування майбутніх значень фінансових інструментів та індикаторів за їх минулими значеннями;

– прогноз тренда (майбутнього напрямку руху – зростання, падіння) фінансового інструменту і його сили;

– виділення кластерної структури ринку, галузі, сектора по деякому набору характеристик;

– динамічне управління портфелем;

– оцінка ризиків;

– передбачення кризи і прогноз її розвитку;

– вибір активів і ін.

Крім описаних вище сфер діяльності, технологія інтелектуального аналізу даних може застосовуватися в найрізноманітніших областях бізнесу, де є необхідність в аналізі даних і накопичений певний обсяг ретроспективної інформації.

1.4.2 Застосування інтелектуального аналізу даних для досліджень уряду

Створення системи, яка дозволить відстежувати всіх іноземців, які приїжджають в країну. Завдання цього комплексу: починаючи з прикордонного терміналу, на основі технології біометричної ідентифікації особистості і різних інших баз даних контролювати, наскільки реальні плани іноземців відповідають заявленим раніше (включаючи переміщення по країні, терміни від'їзду та ін.).

За даними аналітичного звіту Головного контрольного управління американського Конгресу, урядові відомства США беруть участь приблизно в двохстах проектах на основі аналізу даних (Data Mining), які збирають різноманітну інформацію про населення. Понад сто з цих проектів спрямовані на збір персональної інформації (імена, прізвища, адреси e-mail, номери соціального страхування і посвідчень водійських прав), і на основі цієї інформації здійснюють передбачення можливої поведінки людей. Оскільки в згаданому звіті не наведено інформацію про секретні звіти, треба вважати, що загальне число таких систем значно більше.

Незважаючи на користь, яку приносять системи відстеження, експерти згаданого управління, так само як і незалежні експерти, попереджають про значний ризик, з яким пов'язані подібні проекти. Причина побоювань – проблеми, які можуть виникнути при управлінні і нагляд за такими базами.

1.4.3 Застосування інтелектуального аналізу даних для наукових досліджень

Біоінформатика

Біоінформатика – напрямок, метою якого є розробка алгоритмів для аналізу і систематизації генетичної інформації. Отримані алгоритми використовуються для визначення

структур макромолекул, а також їх функцій, з метою пояснення різних біологічних явищ.

Медицина

Незважаючи на консервативність медицини в багатьох її аспектах, технологія інтелектуального аналізу даних активно застосовується для різних досліджень і в цій сфері людської діяльності. Традиційно для постановки медичних діагнозів використовуються експертні системи, які побудовані на основі символьних правил, що поєднують, наприклад, симптоми пацієнта і його захворювання. З використанням інтелектуального аналізу даних за допомогою шаблонів можна розробити базу знань для експертної системи.

Фармацевтика

В області фармацевтики методи інтелектуального аналізу даних також мають досить широке застосування. Дослідження ефективності клінічного застосування певних препаратів, визначення груп препаратів, які будуть ефективні для конкретних груп пацієнтів. Актуальними також є завдання просування лікарських препаратів на ринок.

Молекулярна генетика і генна інженерія

В молекулярної генетики та генної інженерії виділяють окремий напрямок інтелектуального аналізу даних, який має назву аналіз даних в мікро-масивах (Microarray Data Analysis, MDA). Основні поняття, якими оперує інтелектуальний аналіз даних в областях "Молекулярна генетика і генна інженерія" – маркери, тобто генетичні коди, які контролюють різні ознаки живого організму. На фінансування проектів з використанням інтелектуального аналізу даних в розглянутих сферах виділяють значні фінансові кошти.

1.4.4 Застосування інтелектуального аналізу даних для вирішення Web-завдань

Web Mining можна перевести як "видобуток даних в Web". Здатність визначати інтереси і переваги кожного відвідувача,

спостерігаючи за його поведінкою, є серйозною перевагою конкурентної боротьби на ринку електронної комерції.

Системи Web Mining можуть відповісти на багато питань, наприклад, хто з відвідувачів є потенційним клієнтом Web-магазина, яка група клієнтів Web-магазину приносить найбільший дохід, які інтереси певного відвідувача або групи відвідувачів.

Технологія Web Mining охоплює методи, які здатні на основі даних сайту виявити нові, раніше невідомі знання і які в подальшому можна буде використовувати на практиці. Іншими словами, технологія Web Mining застосовує технологію інтелектуального аналізу даних для аналізу неструктурованої, неоднорідної, розподіленої і значної за обсягом інформації, що міститься на Web-вузлах.

Згідно таксономії Web Mining, можна виділити два основних напрямки: Web Content Mining і Web Usage Mining.

Web Content Mining передбачає автоматичний пошук і витяг якісної інформації з різноманітних джерел Інтернету, перевантажених "інформаційним шумом". Зазвичай використовуються методи кластерізації.

Другий напрямок **Web Usage Mining** передбачає виявлення закономірностей в діях користувача Web-вузла або їх групи. Аналізується наступна інформація: які сторінки переглядав користувач; яка послідовність перегляду сторінок.

При використанні Web Mining перед розробниками виникає два типи завдань. Перший стосується збору даних, другий – використання методів персоніфікації. В результаті збору деякого об'єму персоніфікованих ретроспективних даних про конкретного клієнта, система накопичує певні знання про нього і може рекомендувати йому, наприклад, певні набори товарів або послуг. На основі інформації про всіх відвідувачів сайту Web-система може виявити певні групи відвідувачів і

також рекомендувати їм товари або ж пропонувати товари в розсилках.

Задачу Web Mining можна поділити на такі категорії:

– попередня обробка даних для Web Mining;

– виявлення шаблонів і відкриття знань з використанням асоціативних правил, тимчасових послідовностей, класифікації та кластеризації;

– аналіз отриманого знання.

1.4.5 Text Mining

Text Mining охоплює нові методи для виконання семантичного аналізу текстів, інформаційного пошуку і управління. Синонімом поняття Text Mining є KDT (Knowledge Discovering in Text – пошук або виявлення знань в тексті).

На відміну від технології інтелектуального аналізу даних, яка передбачає аналіз впорядкованої інформації, технологія Text Mining аналізує великі і надвеликі масиви неструктурованої інформації.

Програми, що реалізують цю задачу, повинні деяким чином оперувати природною людською мовою і при цьому розуміти семантику аналізованого тексту. Один з методів, на якому засновані деякі Text Mining системи, – пошук підрядка в рядку.

Питання для самоконтролю

1. Дайте визначення інтелектуальному аналізу даних.
2. Порівняйте інтелектуальний аналіз даних та статистику.
3. Порівняйте інтелектуальний аналіз даних та машинне навчання.
4. Порівняйте інтелектуальний аналіз даних та штучний інтелект.
5. Назвіть стадії інтелектуального аналізу даних.
6. Назвіть основні задачі інтелектуального аналізу даних.
7. Назвіть основні задачі інтелектуального аналізу даних.

8. Що спільного в задачах кластерізації та класифікації?
9. В чому відмінність задач кластерізації та класифікації?
10. Назвіть логічні методи.
11. Назвіть статистичні методи.
12. Які методи належать до кібернитичних?
13. Що таке масштабованість методу?
14. В яких сферах використовують методи інтелектуального аналізу даних?

Самостійна робота №1
Тема: Вступ до інтелектуального аналізу даних

Завдання для виконання

1. Ознайомиться з найбільш відомими порталами з інтелектуального аналізу даних.

- http://www.kdnuggets.com/ – найбільший портал з інтелектуального аналізу даних, підтримуваний Григорієм Пятецкий-Шапіро, одним з ідеологів Data Mining.
- http://www.iapr.org/ (International Association for Pattern Recognition) – Міжнародна асоціація розпізнавання образів.
- http://homepages.inf.ed.ac.uk/rbf/IAPR/index.php – колекція посилань на освітні ресурси з розпізнавання образів, машинного навчання, обробці сигналів, обробки зображень та комп'ютерного зору, підтримувана Міжнародною асоціацією розпізнавання образів.
- http://www.kdnet.org/ (Knowledge Discovery Network of Excellence) – міжнародний проект, який об'єднує представників науки і бізнесу, які вирішують практичні завдання інтелектуального аналізу даних.

– www.MachineLearning.ru – прогнозування, розпізнавання, класифікація.

2. Підготувати доповідь за матеріалами розділу «Most Popular Last Week» сайту http://www.kdnuggets.com.

Практична (семінарська) робота №1
Тема: Вступ до інтелектуального аналізу даних

Хід роботи

1. Бесіда з теоретичних питань теми.
2. Доповіді студентів за результатами виконаного домашнього завдання.
3. Відповіді на запитання.
4. Опрацювання інформації одногрупників.

Питання для обговорення

1. Актуальні задачі інтелектуального аналізу даних.
2. Нові технології в аналізі даних.
3. Попередня обробка даних.
4. Навички необхідні для фахівця з аналізу даних.
5. Показники гарної та поганої роботи з аналізу даних.

Тест до теми 1

1. Машинне навчання – це
   а) наука, яка вивчає комп'ютерні алгоритми, що автоматично покращуються під час роботи;
   б) наука про методи збору даних, їх обробку і аналіз для виявлення закономірностей, властивих досліджуваному явищу;
   в) науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного

моделювання видів людської діяльності, що традиційно вважаються інтелектуальними.

2. Штучний інтелект – це
   а) наука, яка вивчає комп'ютерні алгоритми, що автоматично покращуються під час роботи;
   б) наука про методи збору даних, їх обробку і аналіз для виявлення закономірностей, властивих досліджуваному явищу;
   в) науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, що традиційно вважаються інтелектуальними.
3. Статистика – це
   а) наука, яка вивчає комп'ютерні алгоритми, що автоматично покращуються під час роботи;
   б) наука про методи збору даних, їх обробку і аналіз для виявлення закономірностей, властивих досліджуваному явищу;
   в) науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, що традиційно вважаються інтелектуальними.
4. Виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних та за цими ознаками новий об'єкт можна віднести до того чи іншого класу в задачах
   а) класифікації;
   б) візуалізації;
   в) асоціації;
   г) прогнозування.
5. На основі особливостей історичних даних оцінюються пропущені або ж майбутні значення цільових чисельних показників в задачах

а) класифікації;
б) візуалізації;
в) асоціації;
г) прогнозування.

6. На основі даних створюється графічний образ аналізованих даних в задачах
   а) класифікації;
   б) візуалізації;
   в) асоціації;
   г) прогнозування.
7. Логічним методом є:
   а) баєсівські мережі;
   б) дерева рішень;
   в) регресійний аналіз.
8. Статистичним методом є:
   а) баєсівські мережі;
   б) дерева рішень;
   в) регресійний аналіз.
9. Методом крос-табуляції є:
   а) баєсівські мережі;
   б) дерева рішень;
   в) регресійний аналіз.

Рекомендована література

1. Big Data Coursera [Електронний ресурс] Режим доступу: https://www.coursera.org/specializations/big-data
2. Big Data for Data Engineers Coursera [Електронний ресурс] Режим доступу:
3. Hadoop Starter Kit [Електронний ресурс] Режим доступу: https://www.udemy.com/hadoopstarterkit/
https://www.coursera.org/specializations/big-data-engineering
4. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. – СПб.: БХВ-Петербург, 2009. – 512 с.
5. Замятин А.В. Интеллектуальный анализ данных: учеб. пособие. – Томск: Издательский Дом Томского государственного университета, 2016. – 120 с.
6. Олійник А. О. Еволюційні обчислення та програмування: Навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2010. – 324 с.
7. Олійник А. О. Інтелектуальний аналіз даних : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – 278 с.
8. Паклин Н. Бизнес-аналитика. От данных к знаниям / Н. Паклин, В. Орешков – СПб: Питер, 2013. – 704 с.
9. Руденко О. Г. Штучні нейронні мережі / О. Г. Руденко, Є. В. Бодянський. – Харків : Компанія СМІТ, 2006. – 404 с.
10. Скобцов Ю. А. Основы эволюционных вычислений /Ю. А. Скобцов. – Донецк : ДонНТУ, 2008. – 330 с.
11. Чубукова И. А. Data Mining: учебное пособие //М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний. – 2006. – 368с.

ТЕМА 2 НАБІР ДАНИХ ТА ЇХ ВЛАСТИВОСТІ

Теоретичний матеріал до лекції

План

2.1 Атрибути даних

2.2 Шкали вимірювання

2.3 Типи наборів даних

2.4 Описова статистика

Основні поняття: атрибут, змінна, значення, генеральна сукупність, вибірка, параметри, залежна змінна, незалежна змінна, вимірювання, дискретні дані, неперервні дані, номінальна шкала, порядкова шкала, інтервальна шкала, відносна шкала, дихотомічна шкала, табличні дані, графічні дані, текстові дані, бази даних, середнє значення, медіана, дисперсія, ексцес, викиди.

2.1 Атрибути даних

В широкому розумінні дані представляють собою факти, текст, графіки, картинки, звуки, аналогові або цифрові відео-сегменти. Дані можуть бути отримані в результаті вимірів, експериментів, арифметичних і логічних операцій. Дані повинні бути представлені у формі, придатній для зберігання, передачі і обробки.

Іншими словами, дані – це необроблений матеріал, що надається постачальниками даних і використовується споживачами для формування інформації на основі даних.

У таблиці 1.1 представлена двомірна таблиця, що представляє собою набір даних. По горизонталі таблиці розташовуються атрибути об'єкта або його ознаки. По вертикалі таблиці – об'єкти.

Об'єкт описується як набір атрибутів. Об'єкт також відомий як запис, випадок, приклад, рядок таблиці і т.д.

Атрибут – властивість, що характеризує об'єкт. Наприклад: колір очей людини, температура води і т.д. Атрибут також називають змінною, полем таблиці, вимірюванням, характеристикою.

Таблиця 1.1. Двомірна таблиця "об'єкт-атрибут"

Атрибути					
Об'єкти	Код клієнту	Вік	Сімейний стан	Дохід	Клас
	1	18	Не заміжня	12500	1
	2	22	Одружена	10000	1
	3	30	Не заміжня	7000	1
	4	32	Одружена	12000	1
	5	24	Розведена	9500	2
	6	18	Одружена	6000	1
	7	22	Розведена	22000	1
	8	30	Не заміжня	8500	2
	9	32	Одружена	7500	1
	10	40	Не заміжня	9000	2

В результаті операціоналізації понять, тобто переходу від загальних категорій до конкретних величин, виходить набір змінних досліджуваного поняття.

Змінна (variable) – властивість або характеристика, загальна для всіх досліджуваних об'єктів, прояв якої може змінюватися від об'єкта до об'єкта.

Значення (value) змінної є проявом ознаки.

При аналізі даних, як правило, немає можливості розглянути всю цікаву для нас сукупність об'єктів. Вивчення дуже великих обсягів даних є дорогим процесом, що вимагає великих затрат часу, а також неминуче призводить до помилок, пов'язаних з людським фактором.

Цілком достатньо розглянути деяку частину всієї сукупності, тобто вибірку, і отримати цікаву для нас інформацію на її підставі.

Однак розмір вибірки повинен залежати від різноманітності об'єктів, представлених в генеральної сукупності. У вибірці повинні бути представлені різні комбінації і елементи генеральної сукупності.

Генеральна сукупність (population) – вся сукупність досліджуваних об'єктів, яка цікавить дослідника.

Вибірка (sample) – частина генеральної сукупності, певним способом відібрана з метою дослідження і отримання висновків про властивості та характеристики генеральної сукупності.

Параметри – числові характеристики генеральної сукупності.

Статистики – числові характеристики вибірки.

Часто дослідження ґрунтуються на гіпотезах. Гіпотези перевіряються за допомогою даних.

Гіпотеза – припущення щодо параметрів сукупності об'єктів, яке повинно бути перевірено на її частині.

Гіпотеза – частково обґрунтована закономірність знань, що служить або для зв'язку між різними емпіричними фактами, або для пояснення факту, групи фактів.

Приклад гіпотези: між показниками тривалості життя і якістю харчування є зв'язок. У цьому випадку метою дослідження може бути пояснення змін конкретної змінної, в даному випадку – тривалості життя. Припустимо, існує гіпотеза, що **залежна змінна** (тривалість життя) змінюється в залежності від деяких причин (якість харчування, спосіб життя, місце проживання і т.д.), які і є **незалежними змінними**.

Однак змінна спочатку не є залежною або незалежною. Вона стає такою після формулювання конкретної гіпотези. Залежна змінна в одній гіпотезі може бути незалежною в інший.

2.2 Шкали вимірювання

Вимірювання – процес присвоєння характеристикам досліджуваних об'єктів згідно з визначеним правилом числових значень. У процесі підготовки даних вимірюється не сам об'єкт, а його характеристики.

Шкала – правило, згідно з яким об'єктам присвоюються числа.

При імпорті даних з інших джерел програмні інструменти аналізу даних пропонують вибрати тип шкали для кожної

змінної і/ або вибрати тип даних для вхідних і вихідних змінних (символьні, числові, дискретні, неперервні). Для якісного розв'язання задач аналізу даних необхідно володіти цими поняттями.

Змінні можуть бути **числовими** даними або **символьними**.

Числові дані, в свою чергу, можуть бути дискретними і неперервними.

Дискретні дані є значеннями ознак, загальне число яких скінчено або нескінченно, але може бути підраховано за допомогою натуральних чисел від одного до нескінченності.

Неперервні дані – дані, значення яких можуть брати яке завгодно значення в деякому інтервалі. Вимірювання неперервних даних передбачає велику точність.

Шкали

Існує п'ять типів шкал вимірювань: номінальна, порядкова, інтервальна, відносна і дихотомічна.

Номінальна шкала (nominal scale) – шкала, яка містить тільки категорії; дані в ній не можуть упорядковуватися, з ними не можуть бути зроблені ніякі арифметичні дії.

Номінальна шкала складається з назв, категорій, імен для класифікації та сортування об'єктів або спостережень за певною ознакою.

Приклад такої шкали: професії, місто проживання, сімейний стан.

Для цієї шкали можна застосовувати лише такі операції: рівність (=), не рівність (\ ne).

Порядкова шкала (ordinal scale) – шкала, в якій об'єктам присвоюються числа для позначення відносної позиції об'єктів, але не величини відмінностей між ними.

Шкала вимірювань дає можливість ранжувати значення змінних. Вимірювання за допомогою порядкової шкали містять інформацію тільки про порядок проходження величин, але не дозволяють сказати "наскільки одна величина більша за іншу", або "наскільки вона менше іншій".

Приклад такої шкали: місце (1, 2, 3-е), яке команда отримала на змаганнях, номер студента в рейтингу успішності (1-й, 23-й, і т.д.), при цьому невідомо, наскільки один студент успішніше іншого, відомий лише його номер в рейтингу.

Для цієї шкали можна застосовувати лише такі операції: ріне (=), не рівне (\ ne), більше (>), менше (<).

Інтервальна шкала (interval scale) – шкала, різниці між значеннями якої можуть бути обчислені, проте їхні відношення не мають сенсу.

Ця шкала дозволяє знаходити різницю між двома величинами, має властивості номінальної і порядкової шкал, а також дозволяє визначити кількісну зміну ознаки.

Приклад такої шкали: температура води в морі вранці – 19 градусів, ввечері – 24, тобто вечірня на 5 градусів вище, але не можна сказати, що вона в 1,26 разів вище.

Номінальна і порядкова шкали є дискретними, а інтервальна шкала – неперервна, вона дозволяє здійснювати точні вимірювання ознаки і здійснювати арифметичні операції додавання, віднімання, множення, ділення.

Для цієї шкали можна застосовувати лише такі операції: рівне (=), не рівне (\ ne), більше (>), менше (<), операції додавання (+) і віднімання (-).

Відносна шкала (ratio scale) – шкала, в якій є певна точка відліку (існує абсолютний нуль) і можливе відношення між значеннями шкали.

Приклад такої шкали: вага новонародженої дитини (4 кг і 3 кг). Перший в 1,33 рази важче.

Ціна на картоплю в супермаркеті вище в 1,2 рази, ніж ціна на базарі.

Відносні і інтервальні шкали є числовими.

Для цієї шкали можна застосовувати лише такі операції: рівне (=), не рівне (\ ne), більше (>), менше (<), операції додавання (+) і віднімання (–), множення (*) і ділення (/) .

Дихотомічна шкала (dichotomous scale) – шкала, яка містить тільки дві категорії.

Приклад такої шкали: стать (чоловічий і жіночий).

Приклад використання різних шкал для вимірювання властивостей різних об'єктів, наведено в таблиці даних, зображеної в таблиці 1.2.

Таблиця 1.2. Вимірювання властивостей різних об'єктів

Номер об’єкта (інтервальна)	Професія (номінальна)	Середній бал (інтервальна)	Освіта (порядкова)
1	водій	22	середня
2	доктор	55	вища
3	учитель	47	вища

Приклад використання різних шкал для вимірювання властивостей однієї системи, в даному випадку температурних умов, наведено в таблиці даних, зображеної в таблиці 1.3.

Таблиця 1.3. Вимірювання властивостей однієї системи

Дата виміру (інтервальна)	Хмарність (номінальна)	Температура в 8 годин ранку (інтервальна)	Сила вітру (порядкова)
1 вересня	ясна	23 C	Вітер дуже сильний
2 вересня	похмура	17 C	Вітер слабкий
3 вересня	хмарна	22 C	Вітер сильний

2.3 Типи наборів даних

Дані, що складаються із записів

Найбільш часто зустрічаються дані – дані, що складаються із записів. Приклади таких наборів даних: табличні дані, матричні дані, документальні дані, транзакційні або операційні.

Табличні дані – дані, що складаються із записів, кожна з яких складається з фіксованого набору атрибутів.

Приклад табличних даних наведено в таблицях 1.2, 1.3.

Транзакційні дані представляють собою особливий тип даних, де кожен запис, що є транзакцією, включає набір значень.

Приклад трансакційної бази даних, що містить перелік покупок клієнтів магазину, наведено на рис. 1.1.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Рис. 1.1. Приклад трансакційних даних

Графічні дані

Приклади графічних даних: WWW-дані; молекулярні структури; графи, карти.

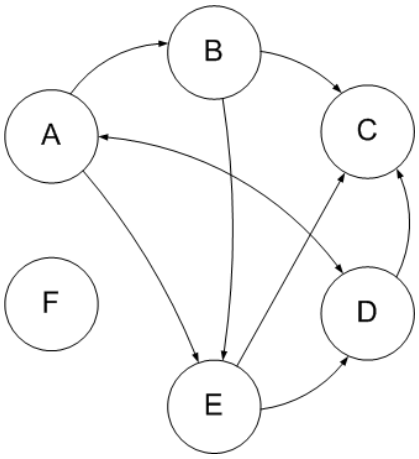


Рис. 1.2. Приклад графічних даних

За допомогою карт, наприклад, можна відстежити зміни об'єктів в часі і просторі, визначити характер їх розподілу на площині або в просторі. Перевагою графічного представлення даних є велика простота їх сприйняття, ніж, наприклад, табличних даних.

Документальні дані

Документальна база даних містить інформацію різного типу: текстову, графічну, звукову, мультимедійну.

Згідно з опитуванням на сайті Kdnuggets, www.kdnuggets.com (2017 року) **"Типи даних, що**

аналізуються", найбільше число типів даних, проаналізованими в 2017 р, були:

дані таблиць (фіксовані стовпці), 69,8%;

текст, 46,4% – перемістився на 2 місце в порівнянні з 2014;

часовий ряд, 45,6%, опустився на 3-е місце;

JSON*, 25,5%, з 7-го місця;

анонімні дані, 22,8%, з 10-го місця;

розташування / гео, 22,6%.

У порівнянні з аналогічним опитуванням KDnuggets 2014 року ми бачимо найбільше зростання частки для:

фото / відео, з 4,9% до 14,1%, зростання на 186%;

анонімні дані, з 14,0% до 22,8%, до 63%;

JSON, з 17,0% до 25,5%, підвищення на 50%;

географія: з 19,7% до 22,6%, зростання на 14,9%.

Найбільше зниження в порівнянні з опитуванням 2014 року:

набори/ транзакції, з 26,5% до 20,1%, зниження на 24%;

веб-потік / веб-журнал, з 12,5% до 10,0%, зниження на 20%;

twitter, з 17,8% до 14,7%, 17% вниз;

XML, з 14% до 12%, 14% вниз.

*JSON (англ. JavaScript Object Notation) - текстовий формат обміну даними, заснований на JavaScript. Як і багато інших текстових форматів, JSON легко читається людьми. Незважаючи на походження від JavaScript, формат вважається незалежним від мови і може використовуватися практично з будь-якою мовою програмування.

Сайт Kdnuggets, визнаний одним з найбільш авторитетних і відомих сайтів в сфері інтелектуального аналізу даних. Діаграма порівняння результатів опитування 2014 та 2017 років наведено в рисунку 1.3.

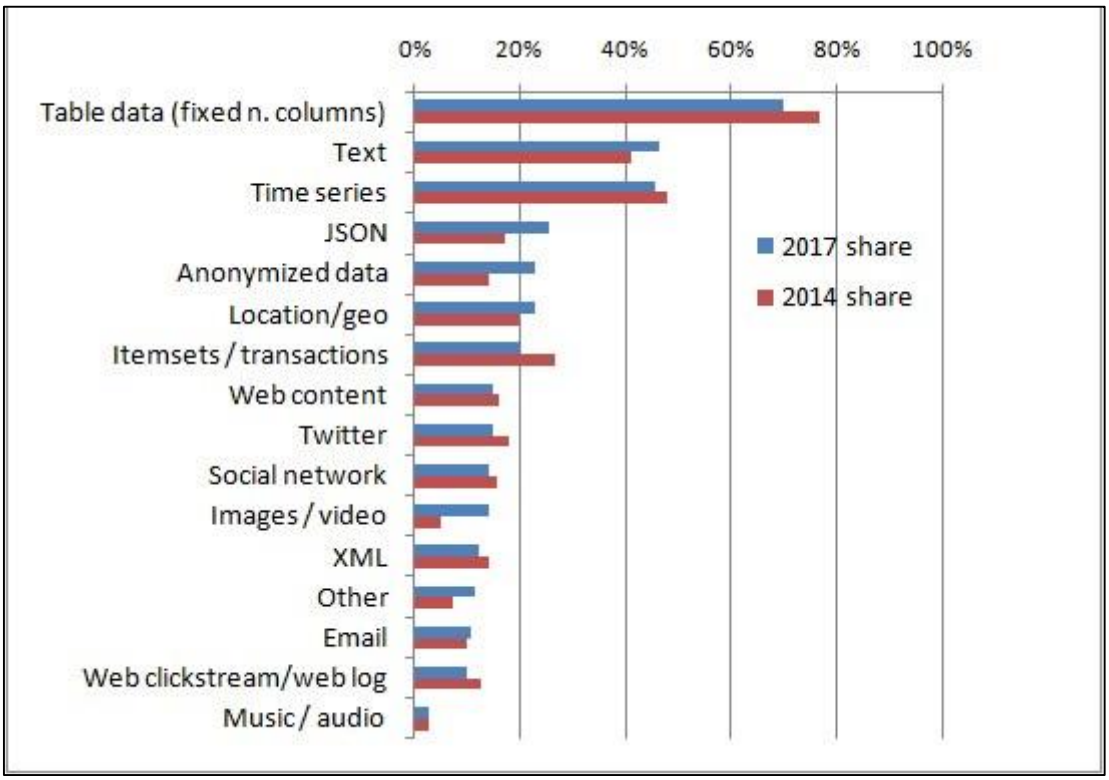


Рис. 1.3. Результати опитування "Типи даних, що аналізуються"

Формати зберігання даних

Одна з основних особливостей даних сучасного світу полягає в тому, що їх стає дуже багато. Можливі чотири аспекти роботи з даними: визначення даних, обчислення, маніпулювання і обробка (збір, передача і ін.).

При маніпулюванні даними використовується структура даних типу "файл". Файли можуть мати різні формати.

Як вже було зазначено раніше, більшість інструментів інтелектуального аналізу даних дозволяють імпортувати дані з різних джерел, а також експортувати підсумкові дані в різні формати. Дані для експериментів зручно зберігати в якомусь одному форматі.

Найбільш поширені формати, згідно з опитуванням "Формати зберігання даних":

формат тієї бази даних, яку вони використовують – 23%;

формат Text, – 18%;

формат CSV – 14%;

формат Excel – 9%;
SPSS – 8%;
R – 4%;
Weka ARFF – 6%;
інші формати – 2%.

Як бачимо з результатів опитування, найбільш поширеним форматом зберігання даних для Інтелектуального аналізу даних виступають бази даних.

2.4 Описова статистика

Описова статистика (Descriptive statistics) – техніка збору і підсумовування кількісних даних, яка використовується для перетворення маси цифрових даних в форму, зручну для сприйняття і обговорення.

Мета описової статистики – узагальнити первинні результати, отримані в результаті спостережень і експериментів.

Нехай дано набір даних А, представлений в таблиці 1.4.

Таблиця 1.4

Набір даних А

x	**y**
3	9
2	7
4	12
5	15
6	17
7	19
8	21
9	23,4
10	25,6
11	27,8

До складу описової статистики входять такі характеристики: середнє; стандартна помилка; медіана; мода;

стандартне відхилення ; дисперсія вибірки; ексцес; асиметричність; інтервал; мінімум; максимум; сума; рахунок.

Звіт "Описова статистика" для двох змінних набору даних А наведено в таблиці 1.5.

Таблиця 1.5.

Описова статистика для набору даних

	X	Y
Середнє	6,5	17,68
Стандартна помилка	0,957427108	2,210922382
Медіана	6,5	18
Стандартне відхилення	3,027650354	6,991550456
Дисперсія вибірки	9,166666667	48,88177778
Ексцес	-1,2	-1,106006058
Асиметричність	0	-0,128299221
Інтервал	9	20,8
Мінімум	2	7
Максимум	11	27,8
Сума	65	176,8
Рахунок	10	10
Найбільший (1)	11	27,8
Найменший (1)	2	7
Рівень надійності (95,0%)	2,16585224	5,001457714

Центральна тенденція

Вимірювання **центральної тенденції** полягає у виборі числа, яке найкращим способом описує всі значення ознаки набору даних. Таке число має як свої переваги, так і недоліки. Ми розглянемо дві характеристики цього виміру, а саме: середнє значення і медіану, ці поняття будуть використовуватися нами в наступних лекціях.

Головна мета **середнього** – представлення набору даних для подальшого аналізу, зіставлення і порівняння.

Значення середнього легко обчислюється і може бути використано для подальшого аналізу. Воно може бути обчислено для даних, вимірюваних по інтервального шкалою, і для деяких даних, вимірюваних по порядкової шкалою. Середнє значення розраховується як середнє арифметичне набору даних: сума всіх значень вибірки, поділена на обсяг вибірки. "Стискаючи" дані таким чином, ми втрачаємо багато інформації.

Середнє значення дуже інформативне і дозволяє робити висновок щодо всього досліджуваного набору даних. За допомогою середнього ми отримуємо можливість порівнювати кілька наборів даних або їх частин.

Середнє значення визначається за формулою:

$$x_{ср} = \frac{\sum x_i}{n}$$

де: $x_{ср}$ – середнє арифметичне;

x_i – вимірювана ознака;

n – кількість вимірювань.

При аналізі даних середнім не слід зловживати, необхідно враховувати його властивості та обмеження. Відомі характеристики "середня температура по лікарні" або "середня висота будинку", що показують некоректність використання цього заходу центральної тенденції для деяких випадків.

Властивості середнього

– при розрахунку середнього не допускаються пропущені значення даних;

– середнє може обчислюватися тільки для числових даних і для дихотомічних шкал;

– для одного набору даних може бути розраховане одне і тільки одне значення середнього.

Інформативність середнього значення змінної висока, якщо відомий її довірчий інтервал. Довірчим інтервалом для

середнього значення є інтервал значень навколо оцінки, де з даним рівнем довіри знаходиться "справжнє" середнє популяції. Обчислення довірчих інтервалів ґрунтується на припущенні нормальності спостережуваних величин.

Ширина довірчого інтервалу залежить від розміру вибірки і від розкиду даних.

Зі збільшенням розміру вибірки точність оцінки середнього зростає. Зі збільшенням розкиду значень вибірки надійність середнього падає. Якщо розмір вибірки досить великий, якість середньої збільшується незалежно від виконання припущення нормальності вибірки.

Медіана – точна середина вибірки, яка ділить її на дві рівні частини по числу спостережень.

Обов'язковою умовою знаходження медіани є упорядкованість вибірки.

Таким чином, для непарної кількості спостережень медіаною виступає спостереження з номером (n + 1) / 2, де n – кількість спостережень у вибірці.

Для парного числа спостережень медіаною є середнє значення спостережень n / 2 і (n + 2) / 2.

Властивості медіани:

– для одного набору даних може бути розраховане одне і тільки одне значення медіани;

– медіана може бути розрахована для неповного набору даних, для цього необхідно знати номери спостережень по порядку, загальна кількість спостережень і кілька значень в середині набору даних.

Характеристики варіації даних

Найбільш простими характеристиками вибірки є максимум і мінімум.

Мінімум – найменше значення вибірки.

Максимум – найбільше значення вибірки.

Розмах – різниця між найбільшим і найменшим

значеннями вибірки.

Дисперсія – це середнє арифметичне квадратів відхилень значень від їх середнього. Дана статистична величина є важливою характеристикою розсіяння варіаційного ряду. Дисперсія σ^2 визначається за формулою:

$$\sigma^2 = \frac{\sum (x_i - x_{cp})^2}{n-1}$$

Стандартне відхилення – квадратний корінь з дисперсії вибірки – міра того, наскільки широко розкидані точки даних відносно свого середнього. Середнє квадратичне відхилення розраховується за наступною формулою:

$$\sigma = \pm\sqrt{\frac{\sum (x_i - x_{cp})^2}{n-1}},$$

де x_i – вимірювана ознака;

x_{cp} – середня арифметична ознака для даної групи;

n – кількість вимірювань.

Ексцес показує "гостроту піку" розподілу, характеризує відносну загостреними або згладжена розподілу в порівнянні з нормальним розподілом. Позитивний ексцес позначає щодо гостре розподіл (пік загострений). Негативний ексцес позначає щодо згладжене розподіл (пік закруглений).

Якщо ексцес істотно відрізняється від нуля, то розподіл має або більш закруглений пік, ніж нормальне, або, навпаки, має більш гострий пік (можливо, є кілька піків). Ексцес нормального розподілу дорівнює нулю.

Асиметрія або асиметричність показує відхилення розподілу від симетричного. Якщо асиметрія істотно відрізняється від нуля, то розподіл несиметрично, нормальний розподіл абсолютно симетрично. Якщо розподіл має довгий правий хвіст, асиметрія позитивна; якщо довгий лівий хвіст – негативна.

Викиди (outliers) – дані, що різко відрізняються від основного числа даних.

При виявленні викидів перед дослідником стоїть дилема: залишити спостереження-викиди або від них відмовитися. Другий варіант вимагає серйозної аргументації і опису. Корисним буде провести аналіз даних з викидами і без і порівняти результати.

Слід пам'ятати, що при застосуванні класичних методів статистичного аналізу, які, як правило, не є робастними (стійкими), наявність викидів в наборі даних призводить до некоректних результатів. Якщо набір даних відносно малий, виняток даних, які вважаються викидами, може помітно вплинути на результати аналізу.

Наявність викидів в наборі даних може бути пов'язано з появою так званих "зсунутих" значень, пов'язаних із систематичною помилкою, помилок введення, помилок збору даних і т.д. Іноді до викидів можуть ставитися найменші і найбільші значення набору даних.

Питання для самоконтролю

1. Дайте визначення поняттям атрибут, змінна.
2. Поясніть зв'язок між вибіркою та генеральною сукупністю.
3. Що таке параметри?
4. Які існують шкали вимірювання?
5. В чом різниця між дискретними та неперервними даними?
6. Які операції можна застосовувати для змінних, які вимірюються в номінальній шкалі?
7. Які операції можна застосовувати для змінних, які вимірюються в порядковій шкалі?
8. Які операції можна застосовувати для змінних, які вимірюються в інтервальній шкалі?

9. Які операції можна застосовувати для змінних, які вимірюються в відносній шкалі?
10. Які операції можна застосовувати для змінних, які вимірюються в дихотомічній шкалі?
11. Данні якого типу найчастіше аналізуються?
12. В яких форматах зберігаються дані?
13. Для чого використовується описова статистика?
14. Які властивості середнього значння?
15. Які характеристики у варіаційних даних?
16. Як пов'язані стандартне відхилення і середнє значення?
17. Що характеризують ексцес і асиметрія?

Самостійна робота №2
Тема: Набір даних та їх властивості

Завдання для виконання

1. На сайті Державної служби статистики України (http://www.ukrstat.gov.ua) обрати цікаву для себе інформацію у форматі файлу Microsoft Excel.
2. Завантажити дані.
3. Проаналізувати атрибути даних (тип, шкала вимірювання).
4. Провести попередню підготовку даних (фіксована кількість рядків та стовбців, відсутність пустих даних).

Практична (лабораторна) робота №2
Тема: Набір даних та їх властивості

Практичне завдання виконується або в Microsoft Excel за допомогою стандартних функцій, або в MATLAB.

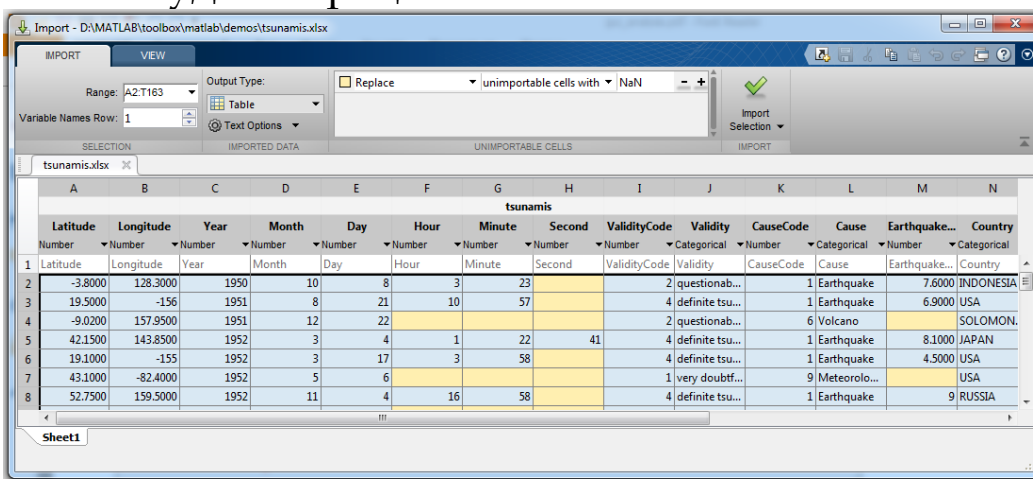
MATLAB це не тільки система математичного розрахунків і моделювання, а й потужний інструмент аналізу даних.

Імпорт даних з файлу Excel в MATLAB

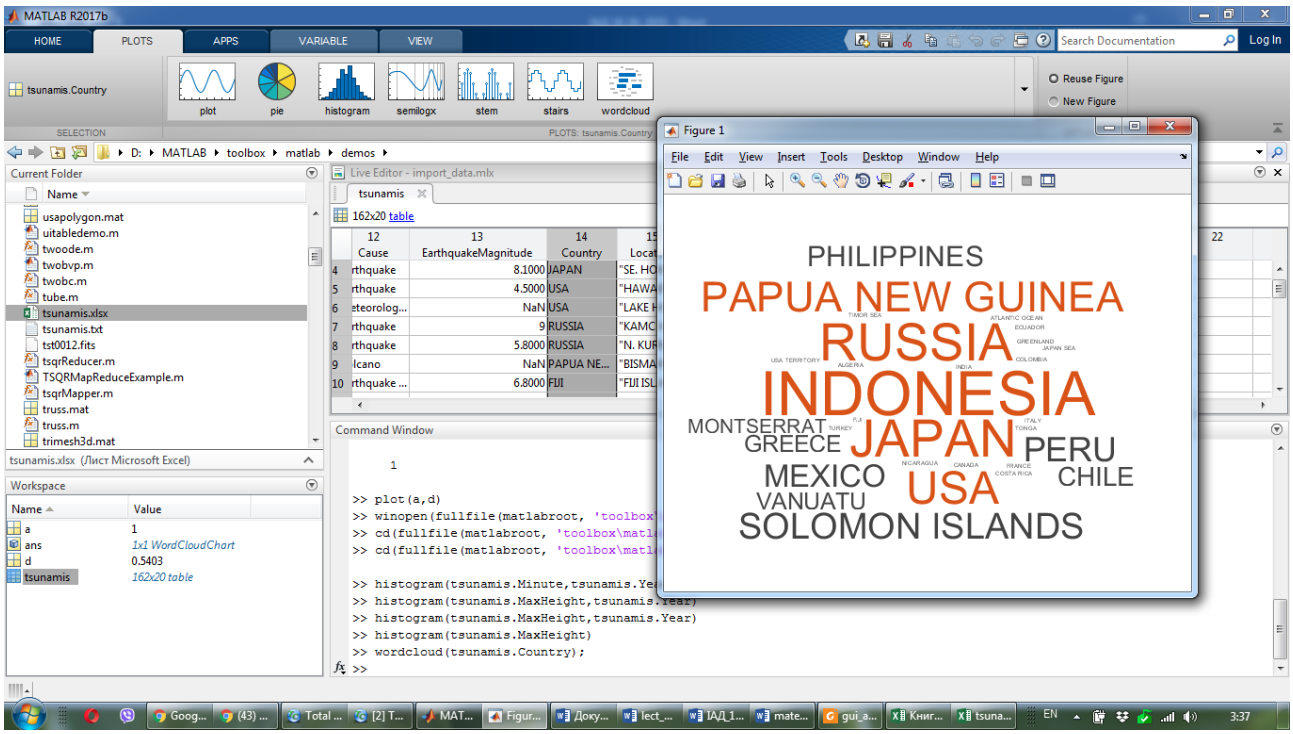
1. Пошук файлу. Копіюємо шлях розташування файлу.
2. Можна змінити шлях командою.

```
cd(fullfile(matlabroot, 'toolbox\matlab\demos'))
```

3. Знаходимо в вікні вмісту робочої папки файл tsunamis.xlsx, і куликаємо по ньому 2 рази, щоб завантажити, відкриється вікно майстра імпорту.
4. Тиснемо зелену кнопку і отримуємо змінну tsunamis, з якої будемо працювати.



5. Після імпорту в робочій області головного вікна з'явилася змінна tsunamis. відкриємо її подвійним клікомі і вивчимо таблицю, яка в ній зберігається.
6. Виділимо стовпець на кладці PLOTS доступні графіки
7. Оцінимо кількісно значення, побудувавши гістограму – графік histogram.
8. Якщо данні в таблиці мають номінальну шкалу можно побудувати графік wordcloud. Хмара слів наочно показує, які слова спостерігаються найчастіше.



Завдання для виконання

1. За попередньо підготовленими даними в самостійній роботі провести аналіз двох атрибутів даних.
2. Визначити середнє значення, розмах вибірки, стандартне відхилення.
3. Побудувати гістограму частот.

Приклад виконання роботи

Дана вибірка значень випадкової величини обсягу 20:

12, 14, 19, 15, 14, 18, 13, 16, 17, 12
18, 17, 15, 13, 17, 14, 14, 13, 14, 16

1) Ранжуємо вибірку: 12, 12, 13, 13, 13, 14, 14, 14, 14, 14, 15, 15, 16, 16, 17, 17, 17, 18, 18, 19.

2) Знаходимо частоти варіантів і будуємо дискретний варіаційний ряд.

x_i	12	13	14	15	16	17	18	19
Частоти, n_i	2	3	5	2	2	3	2	1

$$\sum_{i=1}^{8} n_i = 20$$

$p_i^* = \frac{n_i}{n}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\sum_{i=1}^{8} p_i = 1$

3) За результатами таблиці знаходимо:

$R = 19 - 12 = 7$,

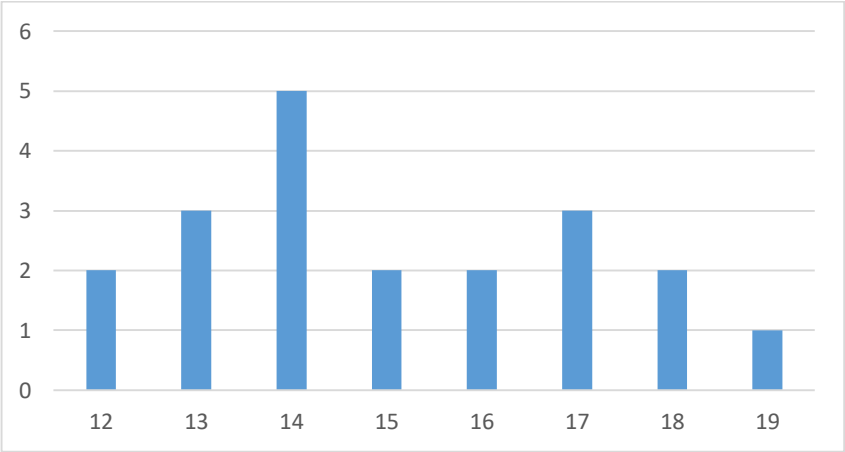
$x_{cp} = \frac{\sum x_i}{n}$,

$X_{ср}=15,05$,

$\sigma = \pm\sqrt{\frac{\sum (x_i - x_{cp})^2}{n-1}}$

$\sigma = 2,08$

4) Будуємо полігон частот.



Тест до теми 2

1. Частина генеральної сукупності, певним способом відібрана з метою дослідження і отримання висновків про властивості та характеристики генеральної сукупності– це
   а) вибірка;
   б) змінна;
   в) параметр.
2. Числові характеристики генеральної сукупності – це
   а) вибірка;
   б) змінна;
   в) параметри.

3. Статистика – це
   а) наука, яка вивчає комп'ютерні алгоритми, що автоматично покращуються під час роботи;
   б) наука про методи збору даних, їх обробку і аналіз для виявлення закономірностей, властивих досліджуваному явищу;
   в) науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, що традиційно вважаються інтелектуальними.
4. Значення, загальне число яких скінчено або нескінченно, але може бути підраховано за допомогою натуральних чисел..
   а) дискретні;
   б) неперервні;
   в) символьні.
5. Дані, значення яких можуть брати яке завгодно значення в деякому інтервалі..
   а) дискретні;
   б) неперервні;
   в) символьні.
6. Шкала, яка містить тільки категорії; дані в ній не можуть упорядковуватися, з ними не можуть бути зроблені ніякі арифметичні дії.
   а) номінальна;
   б) порядкова;
   в) інтервальна;
   г) відносна;
   д) дихотомічна.
7. Шкала, в якій об’єктам присвоюються числа для позначення відносної позиції об'єктів, але не величини відмінностей між ними.

а) номінальна;
б) порядкова;
в) інтервальна;
г) відносна;
д) дихотомічна.

8. Шкала, в якій є певна точка відліку (існує абсолютний нуль) і можливе відношення між значеннями шкали.
   а) номінальна;
   б) порядкова;
   в) інтервальна;
   г) відносна;
   д) дихотомічна.
9. Шкала, різниці між значеннями якої можуть бути обчислені, проте їхні відношення не мають сенсу.
   а) номінальна;
   б) порядкова;
   в) інтервальна;
   г) відносна;
   д) дихотомічна.
10. Шкала, яка містить тільки дві категорії.
    а) номінальна;
    б) порядкова;
    в) інтервальна;
    г) відносна;
    д) дихотомічна.
11. Різниця між найбільшим і найменшим значеннями вибірки –
    а) дисперсія;
    б) медіана;
    в) розмах;
    г) мода.

Рекомендована література

1. Жалдак М.І., Кузьміна Н.М., Берлінська С.Ю. Теорія ймовірностей і математична статистика з елементами інформаційної технології.–К.: Вища школа,1995.–352с.

2. Жлуктенко В.І., Наконечний С.І. Теорія ймовірностей і математична статистика: Навч.–метод. посібник. У 2 ч. – Ч.1. Теорія ймовірностей. – К.: КНЕУ, 2000. – 304 с.

3. Ковальчук А.М. Основи проектування та розробки інформаційних систем: Збірка навчальних матеріалів./ Ковальчук А.М., Левицький В.Г., Самолюк І.І., Янчук В.М.– Ж.: ЖДТУ, 2009. – 54с.

4. Лепский А.Е., Броневич А.Г. Математические методы распознавания образов: Курс лекций. – Таганрог: Изд-во ТТИ ЮФУ, 2009. – 155 с.

5. Олійник А. О. Еволюційні обчислення та програмування: Навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2010. – 324 с.

6. Олійник А. О. Інтелектуальний аналіз даних : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – 278 с.

7. Путятін Є.П. Методи та алгоритми комп'ютерного зору: навч. посіб. — Х.: ТОВ «Компанія СМІТ», 2006. — 236 с.

8. Руденко О. Г. Штучні нейронні мережі / О. Г. Руденко, Є. В. Бодянський. – Харків : Компанія СМІТ, 2006. – 404 с.

9. Скобцов Ю. А. Основы эволюционных вычислений /Ю. А. Скобцов. – Донецк : ДонНТУ, 2008. – 330 с.

10. Ус. С.А. Функціональний аналіз [Текст]: навч. посібник / С.А. Ус. – Д. : Національний гірничий університет, 2013. – 236 с.

11. Чубукова И. А. Data Mining: учебное пособие //М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний. – 2006. – 368с.

ТЕМА 3 ЕЛЕМЕНТИ ТЕОРІЇ МНОЖИН

Теоретичний матеріал до лекції
План

3.1 Множина та її елементи
3.2 Множина і підмножини
3.3 Операції над множинами
3.4 Основні закони алгебри множин
3.5 Добуток множин

Основні поняття: скінчені множини, нескінчені множини, рівні множини, характеристична властивість множин, порожня множина, множина натуральних чисел, множина цілих чисел, множина раціональних чисел, множина дійсних чисел, клас множин, підмножина множини, включення, невласні підмножини, власні підмножини, множина підмножин, універсальна множина, доповнення, об’єднання, перетин, різниця, диз'юнктивна сума, потужність множини, декартовий добутку множин, кортеж.

3.1 Множина та її елементи

Множина є найбільш загальною з усіх математичних структур і з'являється майже в кожній математичній моделі. Часто задачі інтелектуального аналізу даних моделюються в термінах множин, а алгоритми розв’язку формулюються в термінах основних операцій на цих множинах.

Поняття множини є базисом математики, теорія множин – її фундаментом. Дати точне визначення цьому поняттю ми не можемо, найбільше, що ми можемо, – пояснити це поняття такими словами: множина є сукупністю різних об'єктів, що мають щось загальне, що дозволяє об'єднати їх в одну сукупність.

Об'єкти, з яких складається множина, називаються її елементами. Подібно до того, як в геометрії терміни *крапка* і *лінія* є невизначеними, так і в теорії множин основні початкові

поняття теорії – *множина* і *елемент* – не можна визначити: немає більш фундаментальних і елементарних понять, за допомогою яких можна було б дати визначення елементу і множині.

Множини прийнято позначати великими латинськими літерами, елементи множин– малими латинськими літерами.

Множина, яка містить скінчене (нескінчене) число елементів, називається скінченою (нескінченою).

Будемо вважати множину заданною, якщо для любого елемента є можливість визначити, чи являється він елементом цієї множини чи ні.

Способи задання множин.

Перерахуванням її елементів. Для деяких скінчених множин використовується спосіб задання множин простим перерахуванням її елементів.

Твердження, що множина A складається з різних елементів $a_1, a_2, ..., a_n$ (і лише з цих елементів), умовно записується

$$A = \{a_1, a_2, ..., a_n\},$$

тобто множину $A = \{a_1, a_2, ..., a_n\}$ можна задати простим перерахуванням її елементів. При цьому елементи, які перераховуються, беруться у фігурні дужки.

Приклади скінчених множин

Цифри десяткової системи {0, 1, 2, 9}.

Цифри двійкової системи {0, 1}.

Букви алфавіту { а, б,…, я}.

Розв'язок рівняння $x^2 – x = 0$: {0, 1}.

Приклади нескінчених множин

Натуральні числа {1, 2, 3, …}.

Парні натуральні числа {2, 4, 6, 8 …}.

Цілі числа {0, 1, –1, 2, –2, …}.

Визначення. Дві множини рівні тоді і тільки тоді, коли вони складаються з одних і тих самих елементів.

{0, 1, 2} = {2, 1, 0} = {0, 1, 2, 1, 2 } ≠ {3, 1, 2}.

Елементи множин можуть бути записані у будь-якому порядку.

Однакові елементи в множині не розрізняються.

Дві множини *A* і *B* рівні ***(тотожні)***, *A = B*, тоді і тільки тоді, коли кожен елемент *A* є елементом *B* і навпаки.

Це означає, що множина однозначно визначається своїми елементами.

Визначальна (характеристична) властивість. Інший спосіб задання множини полягає в описі елементів *визначальною властивістю P(x) (формою від x)* загальною для всіх елементів. Зазвичай *P(x)* – це вислів, в якому щось стверджується про *x*, або деяка функція змінної *x*. Якщо при заміні змінної *x* на елемент *a* вислів *P*(*a*) стає істинним або значення функції дорівнює заданому елементу, то *a* є елемент даної множини. Множина, задана за допомогою форми *P*(*x*), позначається як *X* = {*x* | *P*(*x*)}, або *X* = {*x: P*(*x*)} і читається як «множина всіх *x*, таких, що *P*(*x*) істинно».

Наприклад:

A = {$x \mid x^2 = 2$} – множина чисел, квадрат яких рівний двом,

B = {*x* | *x* – є тварина з хоботом} – множина слонів,

C = {0, 1, 2 …, 9} = {i | i – цифра десяткової системи},

D = {1, 2, 3 …} = {n | n – натуральне число}.

Запис множини у висловлюваній формі несе більше інформації, чим запис перерахуванням елементів, тобто в {$x \mid x^2 - x = 0$} інформації більше, ніж в {*0, 1*}. Це викликано тим, що деяка множина може бути виражена більш ніж однією формою. Так, {*0, 1*} породжується як записом {*b | b – число двійкової системи*}, так і записом {*c | c = 0* або *c = 1*}.

Відповідно до визначення наступні множини

{x | x^2 – x = 0}

{b | b – число двійкової системи}

{ c | c = 0 або c = 1}

рівні між собою.

Елементи множини зображуються рядковими латинськими буквами, а самі множини – прописними.

Наприклад:

I = {i | i – цифра десяткової системи}

E = { x | x < 1 і x > 2}

L = {a, b …, z}.

Належність елемента множині ***(відношення належності)*** позначається символом ∈, тобто $a_1 \in A$, $a_2 \in A$, ..., $a_n \in A$, або коротше a_1,

a_2, ..., $a_n \in A$. Якщо b не є елементом A, то пишуть $b \notin A$.

$k \in L$, $9 \in I$, $5 \notin L$, $5 \notin E$, $-5 \notin I$.

Коли множина задається своєю характеристичною властивістю, то не завжди можна сказати заздалегідь, чи існує хоча б один елемент, що має таку властивість. Тому для зручності був введений символ порожньої множини – що не має елементів.

Множина, яка не містить жодного елементу, називається **порожньою множиною**. Позначається як {} або символом Ø.

$\{x \mid x < 1 \text{ і } x > 2\} = Ø$

$\{x \mid x \neq x\} = Ø$

$\{0\} \neq Ø$, оскільки множина {*0*} містить один елемент, а саме число нуль.

$Ø \neq 0$, оскільки Ø – множина, а 0 не є множиною.

$\{Ø\} \neq Ø$, оскільки множина {Ø} містить один елемент.

Роль порожньої множини ∅ аналогічна ролі числа нуль. Це поняття можна використовувати для визначення свідомо неіснуючої сукупності елементів (наприклад, множина зелених слонів, дійсного кореня рівняння $x^2 + 1 = 0$).

Деякі з множин чисел настільки часто використовуються, що мають стандартні назви та позначення.

Ø — пуста множина;

N = {1, 2, 3, ...} — множина *натуральних чисел;*

Z – {O, ±1, ±2, ±3, ...} — множина *цілих чисел;*

Q – *{p/q:p,q∈ Z, q ≠ 0}* — множина *раціональних чисел;*

R = {всі десяткові дроби} — множина *дійсних чисел.*

Класом або **сімейством** множин називається множина, елементи якої самі є множинами. Позначенням сімейств служать рукописні прописні букви**.**

Наприклад:

1.A = {0}, {1, 2}, {3, 4, 5}, {6, 7, 8, 9}.
2. *B* = {0, 9}, {1, 8}, {2, 7}, {3, 5}, {8, 9}.
3. *E* = {1, 2, 3}, {1, 2}, {2, 3}, {3, 1}, {1}, {2}, {3}, {∅}, {8, 9}.
4. *D* = { A | A = { x | x – буква в слові природної мови}.

3.2 Множина і підмножини

Якщо *A* і *B* – множини, то говорять, що *A* міститься у *B* (і пишуть $A \subset B$) в тому і лише тому випадку, якщо кожен елемент *A* є елементом *B*.

Множина *A*, всі елементи якої належать і множині *B*, називається ***підмножиною*** (*частиною*) множини *B*.

Це відношення між множинами називають ***включенням*** і позначають символом ⊂, тобто $A \subset B$ *(A* включено у B*)* або $B \supset A$ **(***B* включає *A*).

Наприклад, множина додатних чисел – це підмножина множини дійсних чисел.

Відношення $A \subset B$ допускає і тотожність *(A = B),* тобто будь-яку множину можна розглядати як підмножину самої себе { $A \subset A$*}.* Вважають також, що підмножиною будь-якої множини є порожня множина ∅, тобто $\varnothing \subset A$.

Одночасне виконання співвідношення $A \subset B$ та $B \subset A$ можливо тільки при A = *B.* І навпаки *A* = *B*, якщо $A \subset B$ і $B \subset A$**.** Це може служити визначенням рівності двох множин через відношення включення**.**

Разом з $A \subset B$, в літературі можна зустріти і інше позначення $A \subseteq B$. При цьому під $A \subset B$ розуміють таке відношення включення, яке не допускає рівності *A* і *B* (*строге включення*). Якщо допускається *A* = *B,* то пишуть $A \subseteq B$ (*нестроге включення*).

Відношення включення множин має такі властивості як рефлективність транзитивність та антисиметричність.

Будь-яка не порожня множина *A* має, принаймні, дві різні підмножини: сама *A* і порожня множина ∅. Ці підмножини називаються *невласними,* а всі інші підмножини *A* називають *власними* (ця термінологія пов'язана із словами «власне

підмножини», а не зі словом «власність»). Скінченні власні підмножини утворюються всілякими поєднаннями поодинці, два, три і т.д. елементів даної множини.

Елементи множини самі можуть бути деякими множинами.

Множину, елементами якої є всі підмножини множини A, називають **множиною підмножин** (*множиною-ступенем*) *A* і позначають через *P*(*A*). Так, для трьохелементної множини *A = { a, b, c}* маємо P (A)= {∅, {a}, {b}, {c}, {a, b}, {a, c}, {b, c}, {a, b, c}}.

У разі скінченої множини *A*, що складається з *n* елементів, множина підмножин P(*A*) містить 2^n елементів. Доказ грунтується на сумі всіх коефіцієнтів розкладання бінома Ньютона або на представленні підмножин *n*-розрядними двійковими числами, в яких 1 (або 0) відповідає елементам підмножин.

Слід підкреслити відмінність між відношеннями ***належності*** і ***включення***. Як вже згадувалось, множина *A* може бути своєю підмножиною ($A \subset A$), але вона не може входити до складу своїх елементів ($A \notin A$). Навіть у разі одноелементних підмножин слід розрізняти множину $A = \{a\}$ і її єдиний елемент *a*.

Відношення включення має властивість *транзитивності*: якщо $A \subset B$ *і* $B \subset C$, то $A \subset C$. Відношення належності цієї властивості не має.

Наприклад, множина $A = \{1, \{2, 3\}, 4\}$ у числі своїх елементів містить множину {2, 3}, тому можна записати: $2, 3 \in \{2, 3\}$ і $\{2, 3\} \in A$. Але з цього зовсім не слідує, що елементи 2 і 3 містяться в *A* (у приведеному прикладі ми не знаходимо 2 і 3 серед елементів множини *A*, тобто $2, 3 \notin A$).

Універсальна множина

Вже в самому визначенні конкретної множини явно або неявно обмежується сукупність допустимих об'єктів. Тому часто при застосуванні теорії множин обмежуються розглядом елементів одного типу.

Універсальною множиною або універсумом називається множина всіх об'єктів, що відносяться до окремої прикладної області. (У теорії ймовірності і статистиці вона називається множиною результатів, або вибірковим простором). Так, універсумом арифметики служать числа, зоології — світ тварин, лінгвістики — слова і т.д.

Універсальна множина позначається символом $\boldsymbol{U}$.

Наприклад:

U = {1, 2, 3, 4, 5, 6}. Вона може бути сукупністю можливих результатів в деякому експерименті.

Розглядаючи всілякі футбольні команди, ми за універсальну множину можемо прийняти множину всіх футболістів; з цієї множини ми можемо як завгодно вибирати підмножини.

В геометрії на площині (планіметрії) універсальною множиною є множина всіх точок площини.

Сукупність елементів, що належать універсальній множині U, але не належать підмножині A, називається **доповненням** множини A. Позначається як A' або $\bar{A}$.

$A' = \{x : x \in U, x \notin A\}$.

Відмітимо, що A і A' непересічні множини і обидві є підмножинами множини U.

3.3 Операції над множинами

Множини можна визначати також за допомогою операцій над деякими іншими множинами. Нехай є дві множини A і B.

Об'єднання (*сума*) $A \cup B$ є множина всіх елементів, що належать A або B. Наприклад: $\{1, 2, 3\} \cup \{2, 3, 4\} = \{1,2, 3, 4\}$.

Перетин $A \cap B$ є множина всіх елементів, що належать одночасно як A, так і B. Наприклад: $(1, 2, 3\} \cap \{2, 3, 4\} = \{2, 3\}$.

Множини, що не мають загальних елементів ($A \cap B = \varnothing$), називаються непересічними.

Різниця $A \setminus B$ (або $A - B$) є множина, що складається зі всіх елементів A, що не входять у B, наприклад $\{1, 2, 3\} \setminus \{2, 3, 4\} = \{1\}$. Її можна розглядати як *відносне доповнення B* до A. Якщо $A \subset U$, то

$\overline{A}$

множина $U \setminus A$ називається *абсолютним доповненням* $\overline{A}$ (або просто *доповненням*) множини A і позначається через . Вона містить усі $\overline{A}$ елементи універсуму U, окрім елементів множини A. Доповнення визначається запереченням властивості $P(x)$, за допомогою якої визначається A. Очевидно $A \setminus B = A \cap \overline{B}$

Диз'юнктивна сума (*симетрична різниця*) $A + B$ (або $A \oplus B$) є множина всіх елементів, що належать або A, або B (але не обом разом).

Наприклад: {1, 2, 3} + {2, 3, 4} = (1, 4}. Диз'юнктивну суму отримуємо об'єднуючи елементи множин за винятком тих, які зустрічаються двічі.

Треба відмітити, що об'єднання, перетин, доповнення є операціями, які утворюють нові множини з заданих.

Використовуючи ці операції, можна виражати одні множини через інші, при цьому спочатку виконується одномісна операція доповнення, потім перетину і тільки потім операції об'єднання (різниці). Для того, щоб змінити цей порядок, використовують дужки.

Включення множин **не є операцією.** Воно описує деяку умову. Тоді, як $A \cap B$ є певна множина, елементи якої належать як A, так і B, $A \subseteq B$ – не множина, цей запис показує, що кожен елемент множини A є також елементом і B.

3.4 Основні закони алгебри множин

Операції над множинами, як і операції над числами, мають деякі властивості (табл. 3.1.). Ці властивості виражаються сукупністю тотожностей, справедливих, незалежно від конкретного змісту вхідних в них множин, які є підмножинами деякого універсуму U.

Тотожності (1а)– (3а) виражають відповідно ***комутативний*** *(перестановний),* ***асоціативний*** *(сполучний)* і ***дистрибутивний*** *(розподіленість деякої операції над іншою)* закони для об'єднання, а тотожність (1б) – (3б) ті ж закони для перетину. Співвідношення

(4а)– (7а) визначають властивості пустої множини ∅ і універсуму *U* щодо об'єднання, а співвідношення (4б) – (7б) щодо перетину.

Вирази (8а) і (8б), названі ***законами ідемпотентності***, дозволяють записувати формули з множинами без коефіцієнтів і показників степеня. Залежності (9а) і (9б) представляють ***закони поглинання***, а (10а) і (10б) ***— теореми де Моргана.***

Таблиця 3.1. Основні властивості над множинами

1а) $A \cup B = B \cup A$	1б) $A \cap B = B \cap A$
2а) $A \cup (B \cup C) = (A \cup B) \cup C$	2б) $A \cap (B \cap C) = (A \cap B) \cap C$
3а) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	3б) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
4а) $A \cup \varnothing = A$	4б) $A \cap U = A$
5а) $A \cup \overline{A} = U$	5б) $A \cap \overline{A} = \varnothing$
6а) $A \cup U = U$	6б) $A \cap \varnothing = \varnothing$
7а) $\overline{\varnothing} = U$	7б) $\overline{U} = \varnothing$
8а) $A \cup A = A$	8б) $A \cap A = A$
9а) $A \cup (A \cap B) = A$	9б) $A \cap (A \cup B) = A$
10а) $\overline{A \cup B} = \overline{A} \cap \overline{B}$	10б) $\overline{A \cap B} = \overline{A} \cup \overline{B}$
11) якщо $A \cup B = U$ і $A \cap B = \varnothing$, то $B = \overline{A}$	
12) $\overline{A} = U \setminus A$	
13) $\overline{\overline{A}} = A$	
14) $A \setminus B = A \cap B$	
15) $A + B = (A \cap \overline{B}) \cup (\overline{A} \cap B)$	
16) $A+B = B+A$	
17) $(A+B)+C = A+(B+C)$	
18) $A+ \varnothing = \varnothing + A = A$	
19) $A \subset B$, якщо і тільки якщо $A \cap B = A$ або $A \cup B = B$ або $A \cap \overline{B} = \varnothing$	
20) $A = B$, якщо і тільки якщо $(A \cap \overline{B}) \cup (\overline{A} \cap B) = \varnothing$	

Співвідношення (11)–(20) відображають властивості доповнення, різниці, диз'юнктивної суми, включення і рівності.

Потужністю скінченої множини A називається кількість елементів цієї множини та позначається $|A|$. Множини, що

містять однакову кількість елементів, називаються **рівно потужними**.

Для того, щоб знайти кількість елементів об'єднання двох множин, існує формула, яка має вигляд:

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

Більш загальна формула для обчислення кількості елементів об'єднання декількох множин має вигляд:

$$|A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n| = (|A_1| + |A_2| + \ldots + |A_n|) -$$
$$- (|A_1 \cap A_2| + |A_1 \cap A_3| + \ldots + |A_{n-1} \cap A_n|) +$$
$$+ (|A_1 \cap A_2 \cap A_3| + |A_1 \cap A_2 \cap A_4| + \ldots + |A_{n-2} \cap A_{n-1} \cap A_n| -$$
$$+ (-1)^{n-1}(|A_1 \cap A_2 \cap \ldots \cap A_n|.$$

Задача. Кожен студент в групі або дівчина, або блондин, або любить англійську мову. В групі 20 студенток. З них 12 блондинок і одна любить англійську мову. Всього в групі 24 студента блондина, англійську мову люблять з них 12. А всього студентів (хлопців та дівчат), які люблять англійську мову 17, з них 6 дівчат. Скільки студентів в групі?

Розв'язок. Нехай A – множина студенток, B – блондинів, C – студентів, які люблять англійську мову. Тоді $|A \cup B \cup C|$ - кількість студентів в групі. $|A \cap B|$ множина блондинок, $|A \cap C|$ множина студенток, які люблять англійську мову. $|A \cap B \cap C|$ множина блондинок, які люблять англійську мову.

$$|A \cup B \cup C| = |A| + |C| + |B| - (|A \cap B| + |A \cap C| + |C \cap B|) +$$
$$+ |A \cap B \cap C| = 20 + 24 + 17 - (12 + 6 + 12) + 1 = 32.$$

Отже, в групі 32 студенти.

Перевіримо, чи дійсно це так. На основі цього результату визначимо, скільки в групі студенток. За умовою кількість студенток дорівнює 20.

$$|A| = |A \cup B \cup C| - |C| - |B| + (|A \cap B| + |A \cap C| + |C \cap B|) -$$
$$+ |A \cap B \cap C| = 32 - 24 - 17 + (12 + 6 + 12) - 1 = 20.$$

Що і треба було довести.

3.5 Добуток множин

Одним з важливих понять теорії множин є поняття *декартового добутку* множин.

Нехай є дві множини A і B (не обов'язково різних).

Добутком множин (його також називають декартовим добутком) $A \times B$ є множина всіх впорядкованих пар елементів (a, b), з яких перший елемент a належить множині A, а другий $в$ – множині B.

Нехай $A = \{a_1, a_2, a_3, a_4\}$ і $B = \{b_1, b_2\}$. Тоді $A \times B = \{(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_2, b_2), (a_3, b_1), (a_3, b_2), (a_4, b_1), (a_4, b_2)\}$. Порядок слідування пар довільний, але розташування елементів в кожній парі визначається порядком слідування множин, що перемножуються. Тому $A \times B \neq B \times A$, якщо $B \neq A$.

Операцію добутку множин можна узагальнити на будь-яку кількість множин $A_1, A_2, ..., A_n$ і записати як

$$\prod_{i=1}^{n} A_i = A_1 \times A_2 \times A_3 \times ... \times A_n$$

В результаті отримаємо множину впорядкованих сукупностей елементів $(a_1, a_2, \ldots a_n)$ які називаються ***кортежом, послідовністю, вектором***. Для добутку множин не виконуються закони *комутативності* та *асоціативності*, але виконуються закони *дистрибутивності* відносно операцій об'єднання, перетину і різниці:

$(A_1 \cup A_2) \times B = (A_1 \times B) \cup (A_2 \times B)$;

$(A_1 \cap A_2) \times B = (A_1 \times B) \cap (A_2 \times B)$:

$(A_1 \setminus A_2) \times B = (A_1 \times B) \setminus (A_2 \times B)$.

Для добутку n однакових множин A використовується позначення через степінь $A^n = A \times A \times ... \times A$, де A повторюється n разів. В цьому випадку A^n містять елементи множини A, серед яких можуть бути однакові елементи. Так, якщо $A = \{a_1, a_2\}$, то $A^3 = \{(a_1, a_1, a_1), (a_1, a_1, a_2), (a_1, a_2. a_1), (a_1, a_2, a_2), (a_2, a_1, a_1), (a_2, a_1, a_2), (a_2, a_2, a_1), (a_2, a_2, a_2)\}$.

Слід звернути увагу на істотну різницю добутку множин від згаданих раніш операцій над множинами. В результаті таких операцій як об'єднання, перетин і т. і. завжди отримуємо

множину, елементи якої (якщо вона не порожня) належать вихідним множинам.

Елементи добутку множин істотно відрізняються від елементів множників і представляють собою об'єкти зовсім іншої категорії. Нехай N – множина натуральних чисел. Тоді $N \times N$ буде множиною пар натуральних чисел (p, q), кожна з яких визначає самі різні об'єкти, наприклад, дробові p / q, суми $p + q$, номера будинків p і квартир q (пара p, q визначає частину адреси), пари учасників шахматного турніру у відповідності з с жеребкуванням (p грає білими, а q – чорними) і т. і При цьому $(p, q) \neq (q, p)$. Якби це правило не виконувалося, то чисельники могли б стати знаменниками, номера будинків – номерами квартир і т. і.

Питання для самоконтролю

1. Поясніть поняття множина.
2. Яка множина називається скінченою (нескінченою)?
3. Що називається потужністю множини?
4. Які існують способи задання множин?
5. Що таке характеристична властивість множини?
6. Поясніть поняття підмножина?
7. Наведіть приклади множин і підмножин.
8. Які існують операції над множинами?
9. Сформулюйте комутативний закон для об'єднання.
10. Сформулюйте асоціативний закон для об'єднання.
11. Сформулюйте дистрибутивний закон для об'єднання.
12. Сформулюйте комутативний (перестановний), асоціативний і дистрибутивний закони для перетину.
13. Що називається кортежом?
14. Дайте визначення декартового добутку множин.

Самостійна робота №3
Тема: Елементи теорії множин

Завдання для виконання

1. Самостійно опрацювати тему «Нечіткі множини».
2. Підготувати доповідь «Властивості нечітких множин».

Нечіткі множини

Теорія множин є потужним інструментом математики. Але, аксіома виключеного третього, яка стверджує, що елемент або належить множині або не належить, часто робить цю теорію непридатною в реальних задачах, в яких застосовуються нечіткі оцінки, такі як: «більший прибуток», «високий тиск», «помірна температура», «надійні інструменти» і т.п. На жаль, подібні висловлювання не можуть бути адекватно формалізовані звичайними математичними методами.

Якщо ми хочемо врахувати точне значення нечіткого терма, то чіткий поділ елементів (наприклад значень тиску) на ті, які належать терму «високе», і ті, які не належать, є штучним.

Спроба розвитку формального апарату для залучення часткової власності на теорію множин була зроблена в середині 60-х років. Було введено поняття нечіткої множини як збір елементів, які можуть належати цій множині зі ступенем від 0 до 1. Причому 0 позначає абсолютну неналежність, а 1 – абсолютну приналежність множині. Це було зроблено шляхом застосування поняття функції приналежності, яка ставить у відповідність кожному елементу універсальної множини число з інтервалу [0,1], що позначає ступінь приналежності. Поняття функції приналежності є узагальненням поняттям характеристичної функції чіткої множини, яка оперує значеннями {0,1}. Тому основні властивості і операції над нечіткими множинами є узагальненнями відповідних властивостей і операцій класичної теорії множин.

Нехай $X=\{x\}$ універсальна множина, тобто повна множина, що охоплює всю проблемну область.

Нечітка множина $A \subseteq X$ являє собою набір пар $\{(x, \mu^A(x))\}$, де $x \in X$ і $\mu^A : X \to [0,1]$ – функція приналежності, яка представляє собою деяку суб'єктивну міру відповідності елемента *x* нечіткій множині *A*.

$\mu^A(x)$ може приймати значення від нуля, який позначає абсолютну не приналежність, до одиниці, яка, навпаки, говорить про абсолютну приналежності елемента нечіткій множині. Іноді зручно розглядати значення $\mu^A(x)$ як ступінь сумісності елемента *x* з розмитим поняттям, представленим нечіткою множиною *A*.

Часто нечітка множина $A \subseteq X$ і його функцію приналежності $\mu^A(x)$ розглядають як взаємозамінні поняття.

Якщо множине [0,1] замінити на {0,1}, то функція приналежності буде являти собою характеристичну функцію звичайної (НЕ нечіткої) множини.

Якщо нечітка множина $\underline{A}$ визначено на скінченій універсальній множині $X = \{x_1, x_2, ..., x_n\}$, то її зручно позначати наступним чином:

$$A = \mu^A(x_1)/x_1 + \mu^A(x_2)/x_2 + ... + \mu^A(x_n)/x_n = \sum_{i=1}^{n} \mu^A(x_i)/x_i,$$

де $<\mu^A(x_i)/x_i>$ – пара <функція приналежності / елемент>, звана Сінглтоном, а <+> – позначає сукупність пар.

Приклад. Нехай *X={1,2,…,10}*.Тоді нечітка множина <великі числа> може бути представлена в такий спосіб:

A=<Великі числа> = 0.2 / 6 + 0.5 / 7 + 0.8 / 8 + 1/9 + 1/10.

Це можна розуміти так: 9 і 10 з абсолютною впевненістю можна віднести до <великих чисел>, 8 є <велике число> зі ступенем 0.8 і т.д. 1,2, ... 5 абсолютно не є <великими числами>.

Практична робота №3
Тема: Елементи теорії множин

Завдання для виконання

1. Задано множини: A = {1, 3, 5, 7}; B = {3, 5}; E = {2}; D = {5, 7, 9}. Покажіть, чи будуть істинні наступні твердження:
а) $B \subset A$; б) $E \subset B$; в) $B \subset D$; г) $D \subset A$; д) $B \in A$; е) $\varnothing \in B$; ж) $\varnothing \subset B$; з) $\varnothing \in D$; $E \not\subset B$.
2. Серед наступних множин вкажіть множини, що не перетинаються.
A = {x, y, z}; B = {1, 2, 4, y}; C = {a, b, c, 2}; D = {1, x, a};
E = { Ø}; F = {2, 3, b, c}.
3. За допомогою кругів Ейлера покажіть, що:
а) $\varnothing \subset A \cap B \subset A \cup B$;
б) A+A = Ø;
в) якщо $A \cap B = C$, то $C \subset A$ і $C \subset B$;
г) $(M \setminus N) \cap (N \setminus M) = \varnothing$.
4. Доведіть за = допомогою тотожних перетворень співвідношення:
а) $A \setminus (A \setminus B) = B \setminus (B \setminus A)$;
б) $(A \setminus B) \setminus C = (A \setminus C) \setminus (B= \setminus C)$.
Результат перевірте за допомогою кругів Ейлера.
5. В якому співвідношенні знаходяться множини A і B, якщо $A \setminus B = B \setminus A = \varnothing$?
6. Покажіть справедливість тотожностей:
а) $\overline{A \cap \overline{B}} \cup B = \overline{A} \cup B$;
б) $(A \cap B \cap C) \cup (\overline{A} \cap B \cap C) = B \cap C$;
в) $(A \cap B \cap C \cap \overline{X}) \cup (\overline{A} \cap C) \cup (\overline{B} \cap C) \cup (C \cap X) = C$.
7. Докажіть тотожності.
а) $\overline{(A \cap \overline{X}) \cup (B \cap \overline{X})} = (\overline{A} \cup X) \cap (\overline{B} \cup X)$;
б) $((A \cap X) \cup (B \cap \overline{X})) \cup ((C \cap X) \cup (D \cap \overline{X})) = ((A \cup C) \cap X) \cup ((B \cup D) \cap \overline{X})$;
8. Нехай $S = \{p, q, r, s, t, u\}$. Покажіть серед наступних класів підмножин такі, що становлять розподіл:
а) $\{A_1 = \{p, s, t\}, A_2 = \{q, r\}, A_3 = \{t, u\}\}$;

б) $\{B_1 = \{p\}, B_2 = \{q\}, B_3 = \{r, u\}, B_4 = \{s, t\}\}$;
в) $\{C_1 = \{p, q, t\}, C_2 = C_1' \}$;
г) $\{\{r, s, t\}, \{p, q\}, \{u\}, \varnothing\}$;
д) $\{\{p, q, r, s, t, u\}\}$.

9. Нехай S – множина всіх натуральних чисел {1, 2, 3, …}. Покажіть, що клас множин $\{ C_1, C_2, C_3\}$, де $C_1 = \{3n \mid n = 1, 2, 3, \ldots\}$, $C_2 = \{3n-1 \mid n = 1, 2, 3, \ldots\}$, $C_3 = \{ 3n-2 \mid n = 1, 2, 3, \ldots\}$ становить розподіл множини S.

10. Спростити наступні вирази алгебри множин. (U – універсальна множина).
а) $((A \cup B) \cap (A \cup U)) \cup ((A \cup B) \cap (B \cup \varnothing))$;
б) $((A \cup B) \cap (B \cup U)) \cup (A \cup \varnothing)$;
в) $(A \cup B) \cap (A \cup B')$;
г) $(A \setminus B) \cup (B \setminus C) \cup (A \cap B \cap C)$;
д) $(A \cap B) \cup (A \setminus B)$;
е) $((A \setminus B) \cap (A' \cup B))'$.

11. Довести тотожності, використовуючи основні теореми і аксіоми алгебри множин.
а) $(A \setminus B) \cup (A \cap B) = A$;
б) $(A \cup B) \setminus (A \cap B) = (A \setminus B) \cup (B \setminus A)$;
в) $A \setminus (A \setminus B) = (A \cap B)$;
г) $B \cup (A \setminus B) = (A \cup B)$;
д) $(A \cap B) \setminus (A \cup B) = \varnothing$;
е) $(A \cup B) \setminus (A \setminus B) = B$.

12. Чи є наступне відношення
R ={(1,1), (2,2), (3,3), (4,4), (5,5), (1,5), (5,1), (2,5), (5,2), (1,2), (2,1)}
еквівалентністю на A ={1, 2, 3, 4, 5}? Якщо так, то покажіть відповідний йому розподіл.

13. Вкажіть всі такі бінарні відношення R на A = {1, 2, 3}, які мають властивість $R \circ R = R$.

.14. Задано множини: A = {1. 3, 5, 7}; B = (3, 5}; E = {2}; D= {5, 7, 9}. Вкажіть серед наступних тверджень які істинні, а які ні:

(а) $B \subset A$; (6) $E \subset B$; (в) $B \subset D$; (г) D ⊂ A; (д) $B \in A$; (е) $\varnothing \in B$; (ж) $\varnothing \subset B$; (з) $\varnothing \in$ D; (и) E ⊂ B; (к) множина A скінчена; (л) D не є підмножиною множини A.

15. Серед наступних множин вкажіть пари множин, які не перетинаються.

A = {x, y, z}; B = {1, 2, 4, *y*); C = (a, b, c, 2}; D = {1, x, *a*); $E = \varnothing$; F = {2, 3, b,c}.

16. Виберіть пусті множини:

X = {x : x є розв'язок рівняння $x^2 = 0$};

Y = {y : y є замужньою старою дівою};

Z = {$\varnothing$ };

W = {x : x є людиною, чий вік більше ніж 200 років}.

17. Серед наступних множин вкажіть однакові:

A = {*1,2, a,b*}; B = {*1,0, 2, a, b*}; C={*b,b,2,1,a,a*}; D={*1,2,3,a,b*}.

18. Задано множину A = {1, {2, 3}, 4}. Чи можна її описати як клас?

19. Які з наступних тверджень є вірними, а які ні:

(1) {2, 3} $\in A$; (2) {{2, 3}} $\subset A$; (3) {2, 3} $\subset A$; (4) 3 $\in A$;

(5) $\varnothing \in A$; (6) $\varnothing \subset A$; (7) 4 $\notin$ A?

19. Визначте, скінчені чи нескінчені наступні множини:

множина слів в англійській мові;

множина всіх комбінацій букв алфавіту;

множина точок на відрізку;

множина прямих, які проходять через точку (1,1);

множина розв'язків рівняння sin x = 0;

множина всіх прямокутних трикутників;

множина травинок на крикетному полі.

20. Перепишіть наступні твердження, використовуючи операції над множинами:

S – підмножина S;

x належить множині P;

множина Y не є підмножиною множини X;

множина S — підмножина множини T;

z не належить множині Z;

X є підмножиною множини D.

Тести до теми 3

Оберіть тотожні множини:
а) {1, 2, 3} та {3, 1, 2};
б) {1, 2, 3} та {2, 1, 2, 3, 1};
в) {1, 2, 3} та {{1}, 2, 3};
г) Ø та {Ø}.

2. Оберіть вірне твердження:
а) $Ø \in \{1, 2, 3\}$; $Ø \subseteq \{1, 2, 3\}$; $Ø \subset \{1, 2, 3\}$;
б) $Ø \in \{Ø, 1, 2\}$; $Ø \subseteq \{Ø, 1, 2\}$; $Ø \subset \{Ø, 1, 2\}$;
в) $Ø \in \{\{Ø\}, 1, 2\}$; $Ø \subseteq Ø$; $Ø \subset Ø$.

3. Оберіть хибне твердження:
а) відношення належності $\in$ рефлексивне;
б) відношення належності $\in$ транзитивне;
в) відношення належності $\in$ сіметричне;
г) відношення належності $\in$ антисіметричне.

4. Нехай Z –множина натуральних чисел буде універсумом, Z_1 – множина всіх парних чисел, $A = \{x \mid x < 10\}$. Який визаз описує множину {2, 4, 6, 8}:
а) $Z_1 \cap A$;
б) $Z_1 \cup A$;
в) $Z_1 \setminus A$;
г) $A \setminus Z_1$.

5. Нехай Z –множина натуральних чисел буде універсумом, Z_1 – множина всіх парних чисел, $A = \{x \mid x < 10\}$. Який визаз описує множину {1, 3, 5, 7}:
а) $Z_1 \cap A$;
б) $Z_1 \cup A$;
в) $Z_1 \setminus A$;
г) $A \setminus Z_1$.

6. В множині А 7 елементів, в В – 9. В перетині цих множин 6 елементів. Оберіть вірне твердження.
а) $|A \cap B| > 0.$;
б) $|A \cap B| = 0.$;
в) $B \subset A$;
г) $|A \cup B| > |A| + |B|$.

7. В одній множині 7 елементів, в другій – 9. В об'єднанні цих множин 16 елементів. Оберіть вірне твердження.
а) $|A \cap B| > 0.$;
б) $|A \cap B| = 0.$;
в) $|A \cup B| = 0$;
г) $|A \cup B| > |A| + |B|$.

8. В множині А 6 елементів, а в В – 10. В перетині 6 елементів. Оберіть вірне твердження.
а) $|A \cap B| > 0.$;
б) $|A \cap B| = 0.$;
в) $B \subset A$;
г) $|A \cup B| > |A| + |B|$.

9. Множина В = {{1,2,3}, {9,10), {3,4}} представляє клас множин. Кожний з його елементів є множиною. Оберіть вірне твердження.
а) $\varnothing \in B$;
б) $\varnothing \subset B$;
в) {,2,3} $\subset B$;
г) {,2,3} $\in B$.

Рекомендована література

1. Борисенко О.А. Диференціальна геометрія і топологія. – Х.: Основа, 1995. – 304 с.

2. Бояришева Т.В., Гудивок Т.В., Погоріляк О.О. Функціональний аналіз. Навчальний посібник для студентів

спеціальностей «математика», «прикладна математика», «статистика». – Ужгород, 2013. – 125 с.

3. Журавлёв Ю.И. Распознавание. Математические методы. Программная система. Практические применения. / Журавлёв Ю.И., Рязанов В.В., Сенько О.В. — М.: Изд. «Фазис», 2006. — 176 с.

4. Зиновьев А. Ю. Визуализация многомерных данных. / Зиновьев А. Ю. — Красноярск: Изд. Красноярского государственного технического университета, 2000. — 180 с.

5. Ильин В.А., Позняк Э.Г. Линейная алгебра: Учеб. для вузов. – М.: Наука, 1999. – 296 с.

6. Лепский А.Е., Броневич А.Г. Математические методы распознавания образов: Курс лекций. – Таганрог: Изд-во ТТИ ЮФУ, 2009. – 155 с.

7. Олійник А. О. Інтелектуальний аналіз даних : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – 278 с.

8. Ус. С.А. Функціональний аналіз [Текст]: навч. посібник / С.А. Ус. – Д. : Національний гірничий університет, 2013. – 236с.

ТЕМА 4 МЕТРИЧНІ ОСНОВИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Теоретичний матеріал до лекції

План

4.1 Відношення

4.2 Метричні простори

4.3 Приклади метрик

4.4 Відкриті і замкнуті множини

4.5 Зменшення вимірності. Векторна репрезентація мови

Основні поняття: бінарні відношення, область визначення, область значень, повне (універсальне) відношення, тотожне (діагональне) відношення, порожнє відношення, об'єднання відношень, перетин відношень, обернене відношення, доповнення відношення, симетричне відношення, рефлексивне відношення, транзитивне відношення, багатомісні відношення, метрика, гранична точка, ізольована точка, замкнена множина, обмежена множина, зв'язна множина, збіжність в метричному просторі, повний метричний простір, ізометрія, стискуюче відображення.

4.1 Відношення

Багато задач інтелектуального аналізу даних отримують зручну інтерпретацію на мові теорії відношень. Всі арифметичні операції – це по суті деякі відношення між числами. Елементи множини можуть знаходитися в деяких відношеннях між собою або з елементами інших множин. У найзагальнішому сенсі відношення означає *будь-який зв'язок між предметами або поняттями.* Різноманітні відношення складаються між людьми — батьки і діти, начальники і підлеглі, вчителі і учні.

Бінарні відношення. Відношення між парами об'єктів називають *бінарними (двомісними).* Вище вже були розглянуті два

таких відношення – належність ($a \in A$) і включення $A \subset B$. Перше з них визначає зв'язок між множиною і його елементами, а друге – між двома множинами. Прикладами бінарних відношень є рівність (=), нерівності (< або ≤), а також такі вирази як «бути братом», «ділитися (на якесь число)», «входити до складу (чого-небудь)» і т. і.

Бінарним відношенням R між елементами множин A та B називається підмножина добутку $A \times B$.

Для будь-якого бінарного відношення можна записати відповідне йому *співвідношення* (для відношення нерівності співвідношенням буде $x < y$, для відношення «бути братом» співвідношення запишеться як «*x* брат *y*»). У загальному вигляді співвідношення можна записати як xRy, де R– відношення, що встановлює зв'язок між елементом x з множини X $(x \in X)$ і елементом y з множини Y $(y \in Y)$. Ясно, що таке відношення може бути задане деякою сукупністю впорядкованих пар (x, y), які є елементами множини $X \times Y$. Тому будь-яке бінарне відношення R можна розглядати як **множину *впорядкованих пар (x, y).***

Наприклад, вирази $3 < 7$ і $(3, 7) \in <$ означає, що число 3 менше числа семи, але перше з них більш звичне. В той же час вираз $(7, 3) \in <$ означало б що сім менше трьох, що невірно.

Таким чином, в загальному випадку переставляти елементи в парі (x, y) не можна, що і підкреслюється назвою цієї пари – впорядкована. Елемент x називають *першою координатою,* а елемент y – *другою координатою* впорядкованої пари.

Відношенням можуть бути притамані деякі загальні властивості (наприклад, відношення включення і відношення рівності транзитивні). Визначаючи ці властивості і комбінуючи їх, можна виділити важливі типи відношень, вивчення яких в загальному вигляді замінює розгляд величезної множини частинних відношень.

Областю визначення $D_o(R)$ *та* областю значень $D_3(R)$ відношення R називаються наступні множини:

D_0 ={x | для деякого y, (x, y) ∈ *R*}

$D_з$ = {y | для деякого x, (x, y) ∈ *R*}

Тобто множина перших координат відношення *R* є областю визначення $D_o(R)$, а множина других координат – областю значень $D_3(R)$ відношення *R*. Якщо $x \in X$ і $y \in Y$, то $D_o(R) \subset X$ і $D_3(R) \subset Y$. У таких випадках говорять, що *R* є відношення від *X* до *Y*. Його називають також *відповідністю* і позначають $X \to Y$. Якщо $Y = X$, то будь-яке відношення *xRy* є підмножиною множини $X \times X$ і називається відношенням в *X* або універсальним відношенням в *X* .

Наприклад, $X = \{2, 3\}$ і $Y = \{3, 4, 5, 6\}$. Добуток цих множин $X \times Y = \{(2, 3), (2, 4), (2, 5), (2, 6), (3, 3), (3, 4), (3, 5), (3, 6)\}$.

Відношення «бути дільником» є множина *R* = {(2, 4), (2, 6), (3, 3), (3, 6)}, відношення « = » є множина *B* = {(3, 3)}, а відношення « > » є порожня множина ∅. Області визначення і значень відношення *R* – це відповідно множина $D_o(R)= \{2, 3\} = X$ і $D_3(R) = \{3, 4, 6\} \subset Y$.

Якщо область визначення відношення співпадає з деякою множиною *X*, то говорять, що відношення *визначене на X*. Подібний випадок має місце в приведеному вище прикладі відношення *R* «бути дільником».

Заслуговують уваги три окремі випадки відношень в *X:*

1) ***повне (універсальне) відношення*** $P = X \times X$, яке має місце для кожної пари (x_1, x_2) елементів з *X* (наприклад, відношення «працювати в одному відділі» на множині співробітників даного відділу);

2) ***тотожне (діагональне) відношення*** *E*, рівносильне $x = x$ (наприклад, рівність на множині дійсних чисел);

3) ***порожнє відношення*** ∅, якому не задовольняє жодна пара елементів з *X* (наприклад, відношення «бути братом» на множині жінок). Очевидно, для будь-якого відношення *R* в *X* справедливо $\varnothing \subset R \subset P$.

Розглянемо відношення $R \subset X \times Y$; якщо $x_i \in X$, то *перерізом* по x_i, відношення *R*, (позначимо $R(x_i)$), є множина $y \in Y$ таких, що $(x_i, y) \in R$. Множину всіх перерізів відношення *R* називають

фактор-множиною множини Y по відношенню R і позначають Y / R. Вона повністю визначає відношення R.

Наприклад: нехай задані множини $X = \{x_1, x_2, x_3, x_4, x_5\}$, $Y = \{y_1, y_2, y_3, y_4\}$ та відношення $R = \{(x_1, y_1), (x_1, y_3), (x_2, y_1), (x_2, y_3), (x_2, y_4), (x_3, y_1), (x_3, y_2), (x_3, y_4), (x_4, y_3), (x_5, y_2), (x_5, y_4)\}$. Очевидно, $R(x_1) = \{y_1, y_3\}$; $R(x_2) = \{y_1, y_3, y_4\}$ і та ін. Якщо записати під кожним елементом з X відповідний переріз відношення R, то елементи другого рядка утворять фактор-множину Y / R:

x_1	x_2	x_3	x_4	x_5
$\{y_1, y_3\}$	$\{y_1, y_3, y_4\}$	$\{y_1, y_2, y_4\}$	$\{y_3\}$	$\{y_2, y_4\}$

Об'єднання перерізів за елементами деякої підмножини $B \subset X$ є перерізом $R(B)$ цієї підмножини, тобто $R(B) = \bigcup R\{x\}$, $(x \in B)$. Так, із попереднього прикладу маємо $R(x_2, x_3) = R(x_2) \cup R(X_3) = \{y_1, y_2, y_3, y_4\}$.

Операції над відношенням

Над відношенням, як і над множинами можна виконувати операції, формуючи при цьому складніші відношення.

1. **Об'єднання** 2-х відношень R_1 *і* R_2 $(R_1 \cup R_2)$ утворюється об'єднанням відповідних множин впорядкованих пар.

Наприклад: нехай R_1: $a < b$, R_2: $a = b$, тоді $R_1 \cup R_2$: $a \leq b$.

2. **Перетином** 2-х відношень R_1 *і* R_2 $(R_1 \cap R_2)$ називають відношення, що визначається перетином відповідних множин впорядкованих пар.

Нехай R_1: $a \leq b$, R_2 : $a = b$, тоді $R_1 \cap R_2$: $a = b$.

3. Відношення **включення** також існує для відношень. Відношення R_1 включається *в* R_2, якщо кожна пара $(a,в) \in R_1$ входить і до відношення R_2, тобто $R_1 \subset R_2$.

Наприклад: нехай $R_1 = \{x$ молоде дерево лісу $y\}$, а $R_2 = \{x$ дерево лісу $y\}$, тоді $R_1 \subset R_2$.

4. **Обернене** відношення R^{-1} – це відношення, для якого $aR^{1}в$ справедливо тоді і тільки тоді, коли $вRa$. Таке відношення ще називають *симетричним*.

Наприклад: якщо R: $a < b$, то обернене йому відношення R^{-1}: $a > b$ або $b < a$.

5. **Доповненням** відношення R є множина $\overline{R} = (A \times B) \setminus R$, яка також є відношенням, тобто $a \overline{R} в$ виконується для всіх пар $(a,в)$ $a \in A$, $в \in B$, які не входять до R.

Наприклад: $A = \{1, 2, 3, 4\}$, $B = \{0, 3, 5\}$, $R = \{(2, 0), (1, 3), (2, 3), (2, 5), (3, 3), (4, 3), (4, 5)\}$, тоді $\overline{R} = \{(1, 0), (1, 5), (3, 0), (4, 0), (3,5)\}$

Способи представлення бінарних відношень

З вище сказаного ясно, що скінчене відношення можна представити за допомогою множин впорядкованих пар. Інший спосіб – матричний заснований на представленні відношення відповідній йому прямокутною таблицею (матрицею). Її стовпці відповідають першим координатам, а рядки — другим координатам. На перетині i-го стовпця і j-го рядка ставиться одиниця, якщо виконано співвідношення $x_i R y_j$, и нуль, якщо це співвідношення не виконується (нульові клітки можна залишити порожніми). Ця матриця містить всю інформацію про відношення R. Наприклад, матриця має вигляд:

$$C = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{matrix}$$

Ненульові елементи i-го стовпця указують на сукупність елементів $y \in Y$, що є перерізом $R(x_i)$, наприклад, $R(x_5) = \{y_2, y_4\}$.

Повному відношенню відповідає матриця, всі клітки якої заповнені одиницями, ***тотожному*** – одинична матриця, а *порожньому* – нульова матриця.

Симетричне відношення

Оскільки відношення — це множини, то над ними можна виконувати всі теоретико-множинні операції. Окрім цього, визначаються специфічні для відношень операції: *симетризация* і *композиція*.

Відношення, ***симетричне*** (*обернене*) деякому відношенню $R \subset X \times Y$, позначається через A^{-1} і є підмножиною множини $Y \times X$, утвореною тими парами $(y, x) \in У \times X$, для яких $(x, y) \in R$.

Перехід від R до R^{-1} здійснюється взаємною перестановкою координат кожної впорядкованої пари. Так, обернене відношення для «x є дільник y» буде «y ділиться на x» і для приведеного в параграфі 1.4.2. прикладу виражається множиною {(4, 2), (6, 2), (3, 3), (6, 3)}.

При переході від R до R^{-1} область визначення стає областю значень, і навпаки. Матрицю симетричного відношення отримують транспонуванням початкової матриці. Граф симетричного відношення знаходиться з початкового графа заміною напрямів всіх дуг на протилежні.

Наприклад, якщо $(a,b) \in R$, то $(b,a) \in R^{-1}$, або R^{-1}= *{(a,b) / (b,a) R}.*

Відношення «бути батьком», а симетричне – «бути сином».

Загальні властивості відношень

Нехай R – бінарне відношення на множині X. Відношення R на множині X називається

рефлексивним, якщо xRx для всіх $x \in X$,

тобто воно завжди виконується між об'єктом і ним самим (означає рівність, самообслуговування);

антирефлексивне, якщо xRx не виконується ні для одного $x \in X$, тобто із умови xRy випливає строга нерівність $x \neq y$;

симетричним, якщо з x_iRx_j випливає x_jRx_i для всіх $x_i,x_j \in X$ (відстань між двома точками, «бути братом»);

антисиметричним, якщо з x_iRx_j і x_jRx_i виходить, що $x_i = x_j$ для всіх $x_i,x_j \in X$,

транзитивним, якщо з x_iRx_j і $x_jRx_к$ випливає x_iRx_k для всіх $x_i,x_j\, x_k \in X$.

(відношення передається по «ланцюжку»: «вище» – «нижче», «тепліше» – «холодніше»).

Приклади:

1. Відношення «*x брат y*» на множині родичів **антирефлексивне** (ніхто не є собі братом) і **симетричне**. З першого погляду це відношення здається транзитивним: Але це не так. Нехай в одному сімействі є три брати: Андрій, Микола та Олег. Здавалось би, з того, що *Андрій – брат Миколи* і *Микола – брат Олега* випливає *Андрій – брат Олега*. Але з твердження *Андрій – брат Миколи* випливає, що *Микола – брат Андрія*, і якщо вважати відношення транзитивним, то виходить: *Андрій – брат Андрія,* що безглуздо.

2. Нехай *R* – відношення: курс *а* читається раніше курсу *в* на множині курсів, що читаються в університеті. Відношення *R* анти рефлексивне і транзитивне. Воно не симетричне.

3. Відношення ≤ на множині чисел рефлексивне, антисиметричне і транзитивне. Відношення < антирефлексивне, антисиметричне і транзитивне.

Відношення «бути батьком» не транзитивне. Якби *A* – батько *B*, а *B* – батько *C*, то виходить, що *A* не батько *C*, а дідусь.

Доповнення відношення *R* до транзитивного, називається **транзитивним замкненням** $R \to \hat{R}$**.** Ми добавляємо деякі пари, щоб відношення стало транзитивним**.**

Що таке транзитивне замкнення цього відношення? «Бути предком по чоловічій лінії». Таким чином, якщо *A* – предок *B*, *B* – предок *C*, то *A* – предок *C*.

Відношення «бути сусідом в аудиторії» не транзитивне, бо «сусід сусіда» вже не «сусід». Його транзитивне замкнення – «сидіти за одним столом».

І знамените феодальне право: «Вассал мого вассала не мій вассал» говорить про те, що відношення «бути вассалом» не є транзитивне. Замкнути – «бути підданим».

Для *рефлексивного* відношення всі елементи матриці на головній діагоналі – одиниці, а для антирефлексивного – *нулі.*

Якщо відношення *симетричне,* то *симетрична* і матриця, щодо *головної діагоналі.*

Матриця *антисиметричного* відношення характеризується тим, що у неї *нульові елементи по діагоналі* і *немає жодної пари*

одиниць, які знаходяться *на симетричних місцях по відношенню до головної діагоналі*.

Матриця т*ранзитивного* відношення характеризується тим, що якщо $a_{ij} = 1$ *і* $a_{mi} = 1$, *то* $a_{mj} = 1$.

Багатомісні відношення

Відношення може бути визначене не тільки для пар об'єктів, але і для трійок, четвірок і т.д. Відношення *n* об'єктів (*n-місне відношення*) визначається як множина *n-мірних* векторів (x_1, x_2,…, x_n), що є підмножиною добутку $X_1 \times X_2 \times .. \times X_n$, причому $x_1 \in X_1, x_2 \in X_2$ і т.д.

Багатовимірні вектори можна визначити в термінах впорядкованих пар, наприклад трійка (x_1, x_2, x_3) розглядається як впорядкована пара ((x_1 x_2), x_3), де перша координата (x_1, x_2) сама є впорядкованою парою, причому (x_1, x_2) $\in X_1 \times X_2$. Взагалі, *n-мірний* вектор виражається як впорядкована пара через ((x_1, x_2., x_{n-1}), x_n), якщо визначено (x_1, x_2, x_{n-1}).

Прикладом тримісних (тернарних) відносин є: арифметичні операції над числами, відношення між батьками і дітьми (батько, мати, дитина) і т.і.

Відношенням **еквівалентності** називається рефлексивне, симетричне і транзитивне відношення. Позначається знаком ~, або записується $x \equiv y$. Читається як «x еквівалентне y».

При цьому $x \sim y$ означає, що впорядкована пара (x, y) належить множині $R \subset M \times M$, яка є відношенням еквівалентності на множині M.

Властивості еквівалентності записуються таким чином:

$x \sim x$ рефлективність

якщо $x \sim y$, то $y \sim x$ симетричність

з $x \sim y$ і $y \sim z$ слідує $x \sim y$ транзитивність

Нехай A – деяка множина. Сукупність підмножин

$R = \{A_1, A_2, \ldots\}$

множини A називається *розбиттям* множини A на класи, якщо

1) будь-які дві різні множини не перетинаються

2) об'єднання всіх підмножин співпадає з A.

Приклад. A={1, 2, 3}.

а) R_1 = {{1}, {2}, {3}} – розбиття A.

б) R_2 = {{1},{2},{1,3}} – не є розбиття.

в) R_3 = {{1}, {2}} – не є розбиття

г) R_4 = {{1, 2},{3}} – є розбиття.

д) R_5 = {1, {2, 3}} не є розбиття.

е) R_6 = {{1, 2, 3}} – є розбиття.

Класи еквівалентності

Найважливіше значення еквівалентності полягає в тому, що це відношення визначає ознаку, яка допускає *розбиття* множини M на **непересічні підмножини, звані класами еквівалентності**. Навпаки, всяке розбиття множини M на непересічні підмножини визначає між елементами цієї множини деяке відношення еквівалентності.

Нехай A – деяка множина, а R – деяке відношення еквівалентності на множині A. Для кожного елемента $a \in A$ визначимо підмножину $\bar{a}$ множини A наступним чином

$\bar{a} = \{x \mid x \in A \text{ і } (a, x) \in R\}$

Підмножина $\bar{a}$ називається **класом еквівалентності** R, *я*кий визначається елементом a. Елемент a називається представником цього класу.

Прикладами відношення еквівалентності можуть бути паралельність прямих, твердження «бути таким же».

Розглянемо відношення «порівняння по модулю m на множині натуральних чисел». Це можна записати як $x = y(mod\ m)$ і означає: x порівняно з y по модулю m (m– ціле додатне число, не рівне нулю), якщо $x - y$ ділиться на m. Цілі числа, порівняні по модулю m, пов'язані відношенням $x = y + km$ (k– ціле число) і утворюють підмножину цілих чисел, які мають однакову остачу j, якщо поділити на m. Ці множини не перетинаються, вони є класами еквівалентності, і представником кожного з них природно вибрати остачу $j = 0, 1, 2, \ldots, m - 1$.

Таким чином *відношення порівняння по модулю m* визначає розподіл множини натуральних чисел на m класів M_0, M_1, M_2, …, M_{m-1}, де $M_j = (j, j+m, j+2m, \ldots)$.

Наприклад, при $m=4$ маємо $M_0 = \{0,4,8,\ldots\}$; $M_1 = \{1,5,9,13,\ldots\}$; $M_2 = \{2,6,10,\ldots\}$; $M_3 = \{3,7,11,\ldots\}$. Представниками класів еквівалентності є числа 0,1,2,3, Таким чином, множина цілих чисел розбивається *відношенням порівняння по модулю 4* на чотири класи еквівалентності. Всередині кожного класу ці числа не розрізняються одне від одного (4 ~ 0, 5~1, 6 ~ 2, 7 ~ 3 і т.д.).

При $m=1$ розподіл складається з одного класу, який співпадає з вихідною множиною.

Відношення $x=y$ $(mod2)$ розбиває множину цілих чисел на два класи: парних і непарних чисел.

4.2 Метричні простори

Метричні методи широко застосовуються при вирішенні задач інтелектуального аналізу даних. Пошук оптимальної в деякому сенсі функції відстані у просторі аналізованих об'єктів є однією з фундаментальних задач аналізу даних.

На основі адекватно підібраних метрик будується модель простору, що дозволяє ефективно вирішувати завдання класифікації, кластеризації, ранжування та інші. Методи, засновані на пошуку найближчих сусідів, і алгоритм *k* середніх – класичні приклади метричних підходів, успішно використовуваних у великому числі різних додатків.

Ефективний інформаційний пошук на основі подібності стає все більш важливим інструментом у зв'язку з швидким зростанням обсягів інформації, що зберігається і обробляється.

Основні підходи до визначення подібності – це функціональний підхід (двомісні функції, що задовольняють аксіомам), геометричний підхід (визначення в просторі множин точок), табличний підхід (матриці попарної подібності над кінцевими множинами).

Як правило, пошук за точним збігом у багатьох випадках не може задовольнити запити користувача. Функції відстані повинні також враховувати семантичну схожість запиту і кандидатів. Для завдань інформаційного пошуку в масивах неструктурованої інформації різних видів (тексти, зображення, аудіо, відео) розроблено і розробляється велика кількість різних функцій подібності між об'єктами, що дозволяють врахувати специфіку вирішуваних задач і складну структуру сутностей.

Називаючи деяку множину простором зазвичай наділяють її одним або декількома властивостями звичайних просторів, що вивчаються в елементарній геометрії. Основні властивості простору – це:

1) в просторі визначено відстань між будь-якими двома точками;

2) з будь-якої точки простору можна безперервно (не виходячи з цього простору) перейти в будь-яку іншу точку; при цьому кожну точку можна розмістити в деякому як завгодно малому «околі» цієї точки, що є підмножиною безлічі всіх точок цього простору;

3) в просторі визначено поняття вектора (елемента) простору і операції додавання елементів (векторів) простору і множення вектора на число.

Наділяючи абстрактний простір якимось одним, або будь-якими двома, або трьома з цих властивостей, ми отримуємо різні типи просторів. Спираючись на відомі властивості відстані між точками в тривимірному просторі аналогічно введемо поняття відстані між двома точками в будь-якому просторі.

На множині X визначено структуру метричного простору, якщо задана функція пари аргументів $\rho: X \times X \to R^1$, яка задовольняє таким властивостям:

1) $\rho(x, y) \geq 0; \rho(x, y) = 0 \Leftrightarrow x = y$;

2) $\rho(x, y) = \rho(y, x)$;

3) $\rho(x,z) \le \rho(x,y) + \rho(y,z)$;

Функція $\rho(x,y)$ називається **метрикою** або функцією відстані між точками x і y. Тоді пара (X,ρ) утворює метричний простір.

4.3 Приклади метрик

Існує два основних класи метрик: евклидові та неевклидові. Евклідова відстань визначається на основі положення точок в просторі. Неевклидова відстань визначається на основі властивостей точок, але не на основі їх положення в просторі.

Евклідові метрики

Евклідова метрика (L_2), де відстань обчислюється наступним чином:

$$\rho(x_i, x_j) = \sqrt{\sum_{t=1}^{m} (x_{it} - x_{jt})^2}$$

Метрика Хеммінга (L_1) – це відстань є просто середнім різниці за координатами. У більшості випадків дана міра відстані приводить до таких же результатів, як і для звичайного відстані Евкліда, проте для неї вплив окремих великих різниць (викидів) зменшується (так як вони не зводяться в квадрат). Відстань за Хеммінгом обчислюється за формулою:

$$\rho(x_i, x_j) = \sum_{t=1}^{m} |x_{it} - x_{jt}|$$

Метрика Чебишева (L_∞). Дана метрика може виявитися корисною, коли бажають визначити наскільки два об'єкти є "різними", якщо вони розрізняються за якоюсь однією координатою (будь-яким одним виміром). Відстань Чебишева обчислюється за формулою:

$$\rho_\infty(x_i, x_j) = \max_{l<i<m} |x_{it} - x_{jt}|$$

Неевклідові метрики

Косинусна відстань. Косинус подібності — коефіцієнт подібності двох не нульових векторів, який обчислюється як косинус кута між ними. Косинус 0° дорівнює 1, а для всіх інших значень кута в інтервалі (0,п] буде менше за 1.

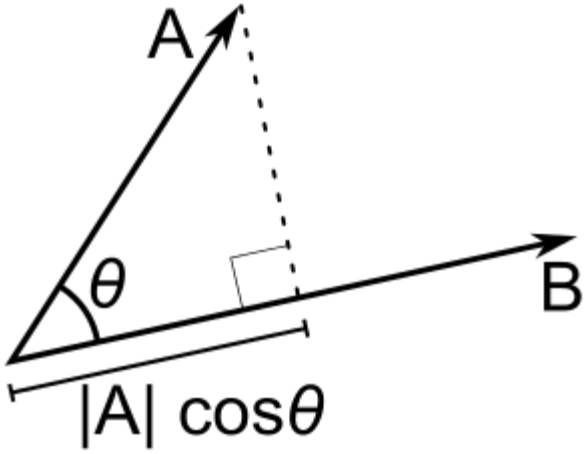


Рис. 4.1 Кут між векторами

Скалярний добуток двох векторів обчислюється за формулою $A \cdot B = |A| \cdot |B| \cdot cos\theta$.

$$\rho(A, B) = \theta = \arccos(\frac{A \cdot B}{|A| \cdot |B|})$$

Косинус подібності часто використовують в позитивному просторі, для якого результат обмежений проміжком . Одиничні вектори максимально «подібні», якщо вони паралельні і максимально «різні», якщо вони ортогональні (перпендикулярні). Найчастіше косинус подібності використовується у багатовимірних додатних просторах. Наприклад, при інформаційному пошуку та аналізі тексту.

Однією з переваг косинуса подібності є низька складність обчислення, особливо для розріджених векторів: достатньо брати лише координати з ненульовим значенням.

Edit Distance (редакційна відстань або дистанція редагування). Edit distance – це число вставок, видалень, яке потрібно зробити, щоб перетворити один рядок в інший. LCS (longest common subsequence) – це найбільша спільна підпослідовність.

$$\rho(A,B) = |A| + |B| - 2|LCS(A,B)|$$

Приклад. x= abcde; y = bcduve

Перетворення x в y: видалити a, вставити u і v після d. Edit distance = 3. Або LCS(x,y) = bcde. Edit distance =5 + 6 –2*4 = 3.

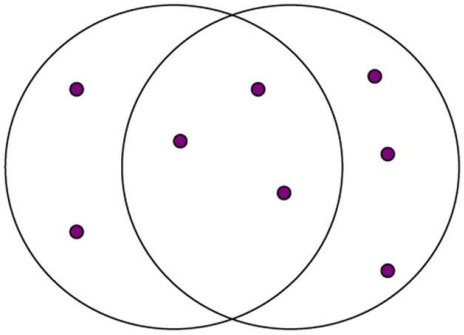
Jaccard Distance.

Jaccard Similarity – бінарна міра подібності, запропонована Полем Жаккаром в 1901 році. Запропонований метод здобув поширення і нині використовується для оцінки подібності скінченних множин, для пошуку подібних документів, плагіату, тощо

$$Sim(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Jaccard Distance між двома наборами – це 1 мінус Jaccard Similarity

$$\rho(A,B) = 1 - Sim(A,B).$$



Приклад.

Перетин = 3

Об'єднання = 8

Jaccard Similarity = 3/8

Jaccard Distance = 5/8

4.4 Відкриті і замкнуті множини

Відкритою кулею радіуса r з центром в точці x називається $B_r(x) = B(r,x) = \{y : \rho(x,y) < r\}$

Замкнена куля: $\overline{B}(r,x)=\{y:\rho(x,y)\le r\}$.

Околом точки $x\in X$ називається будь-яка відкрита множина, що містить цю точку. Позначка: $N(x)$ – окіл точки x; $N_\varepsilon(x)$ – окіл точки x радіусу ε.

Нехай $Y\subset X$, тоді точка $x\in X$ називається **граничною точкою** множини Y, якщо кожний окіл точки x містить принаймні одну точку y: $y\in Y, y\neq x$.

Точка $y\in Y$ називається **ізольованою точкою** множини Y, якщо існує окіл точки y, який не містить жодної точки з Y окрім самої точки y.

Точка $y\in Y\subset X$ називається **внутрішньої**, якщо вона міститься в Y разом з деяким своїм околом.

Множина в метричному просторі називається **замкненою**, якщо воно містить всі свої граничні точки.

Множина в метричному просторі називається **відкритою**, якщо всі її точки внутрішні.

Замикання множини Y (позначення: $\overline{Y}$) – є перетин всіх замкнених множин, що містять Y, тобто замикання – це найменша з усіх замкнених множин, яка містить Y.

Нехай (X,ρ) – метричний простір, на множині Y, що належить метричному простору X, визначена та сама метрика ρ, (тому Y теж буде метричним простором) тоді пара (Y,ρ) називається **підпростором** (X,ρ).

Множина $A\subset X$ називається **обмеженою**, якщо існує такий елемент $x_0\in X$ і постійна $c>0$, що $\rho(x,x_0)<c\forall x\in X$.

Множина X називається **зв'язною**, якщо її не можна представити у вигляді суми двох непустих замкнених (або двох непустих відкритих) підмножин, що не мають спільного перетину.

Нехай A і B – дві множини в метричному просторі X. A називається щільною в множині B, якщо $B \subset \overline{A}$ і **всюди щільним в просторі** X, якщо $\overline{A} = X$.

Простір, в якому існують скінчені, усюди щільні множини, називається **сепарабельним**.

Множина A називається **ніде не щільною** в метричному просторі X, якщо будь-яка відкрита множина цього простору містить іншу відкриту множину, цілком вільну від точок множини A.

Множина A, розташована в метричному просторі називається **досконалою**, якщо вона замкнена і якщо кожна точка цієї множини є його граничною точкою.

Об'єднання відкритих множин $\bigcup_\alpha G_\alpha$ таке, що $A \subset \bigcup_\alpha G_\alpha$ називається **відкритим покриттям** множини A.

Множинна A називається **компактною**, якщо будь-яке її відкрите покриття містить скінчене покриття (підпокриття).

Зауваження: Будь-яка компактна множина обмежена.

Зауваження: Будь-яка компактна множина замкнена.

Зауваження: Протилежне твердження, в загальному випадку невірно.

Збіжність і неперервність відображень в метричному просторі

Послідовність $\{x_n\}$ точок метричного простору X називається **збіжною** до точки $x \in X$, якщо будь-який окіл точки x: $N_\varepsilon(x)$ містить всі точки цієї послідовності, починаючи з деякого номера, (за винятком кінцевого їх числа), тобто $\rho(x_n, x) \to 0,\ n \to \infty$.

Послідовність $\{x_n\}$ елементів метричного простору називається **фундаментальною**, якщо для неї виконується умова Коші: $\rho(x_n, x_m) \to 0, при\ \ n, m \to \infty$.

Зауваження. Будь-яка збіжна послідовність в метричному просторі є фундаментальною, але не всяка фундаментальна послідовність елементів метричного простору буде збіжною в цьому просторі.

Нехай відображення $g: R \to R_0$ – відображення метричного простору $R = (X, \rho)$ в метричний простір $R_0 = (Y, \rho_0)$. Неперервність еквівалентна наступній властивості: якщо $x_0, x_n \in X, n = 1, 2, \ldots$ і $x_n \to x_0$, то і $g(x_n) \to g(x_0)$.

Поповнення метричних просторів

Метричний простір (X, ρ) називається **повним**, якщо в ньому будь-яка фундаментальна послідовність збігається до елементу даного простору.

Зауваження. Не кожен метричний простір є повним.

Зауваження. Дві метрики на множині елементів простору *X* називаються еквівалентними, якщо збіжність елементів по одній означає збіжність і по інший.

Взаємнооднозначне відображення одного метричного простору (X, ρ) в інше метричний простір (Y, ρ_0) – $g: (X, \rho) \to (Y, \rho_0)$ називається **ізометрією**, якщо $\forall x_1, x_2 \in X$ виконується $\rho(x_1, x_2) = \rho_0(g(x_1), g(x_2))$. В цьому випадку (X, ρ) і (Y, ρ_0) називаються ізометричними один до одного.

У багатьох прикладних задачах доцільно мати справу з повними просторами. Тому природно виникає питання про можливість розширення неповного метричного простору (поповнення) до повного.

Розглянутий приклад, вказує шлях, який у багатьох випадках призводить до мети, – включити в даний простір додаткові елементи, що представляють собою межі всіх фундаментальних послідовностей, що належать деякому простору, який містить даний простір, з тією ж метрикою, що і даний простір.

Повний метричний простір (Y,ρ_0) називається поповненням метричного простору (X,ρ) (з тією ж метрикою), якщо (X,ρ) є підпростір простору (Y,ρ_0) і замикання (X,ρ) збігається з простором (Y,ρ_0). Значить, будь-який метричний простір можна поповнити, тобто він може бути вкладений в інший повний метричний простір Y такий, що в метричному просторі Y існує всюди щільний підпростір X_0 ізометричний вихідного простору X.

Теорема. Для будь-якого метричного простору існує його поповнення, причому це поповнення єдино з точністю до ізометрії.

Відображення називається **стискуючим** відображенням, якщо існує таке число $\alpha \in(0,1)$, що $\rho(g(x),g(y))<\alpha\rho(x,y),$ де $x,y\in X$.

Теорема (принцип стискуючих відображень): Будь-яке стискуюче відображення повного метричного простору (X,ρ) самого в себе має одну і тільки одну нерухому точку, тобто таку точку $x\in X$: $g(x)=x$.

4.5 Зменшення вимірності. Векторна репрезентація мови

Припустимо, ми маємо сто тисяч точок у стовимірному просторі, і кожній точці відповідає стовимірний вектор, і нам потрібно знати, як ці точки розподілені – які знаходяться поряд чи далеко, чи є закономірності. Один із методів вирішення даної задачі – взяти попарні відстані між усіма точками, отримаємо матрицю відстаней. В цій матриці ми ігноруємо конкретну інформацію про ознакове описання значення кожного конкретного вектора, оскільки для нас цінність несе тільки позиція кожної точки відносно всіх інших. Тут є нюанс: людина досить погано сприймає інформацію в більш ніж тривимірних

проекціях і може нормально розібратися із структурою даних у дво- чи тривимірній репрезентації.

Часто для того, щоб зрозуміти наскільки ефективно певний алгоритм перетворює дані в багатьох вимірах, нам потрібно зменшити вимірність.

Ми можемо дістати з наших даних матрицю відстаней і на її основі сконструювати меншвимірну репрезентацію точок, розподіл яких буде максимально відображати розподіл цих точок в оригінальній багатовимірній розмірності векторів.

Як правило, нас цікавить інформація про те, які точки знаходяться поряд, а які – далеко. Не завжди потрібно використовувати всю матрицю відстаней і не завжди необхідно враховувати відстань до дуже віддалених точок, особливо у випадку багатовимірних репрезентацій.

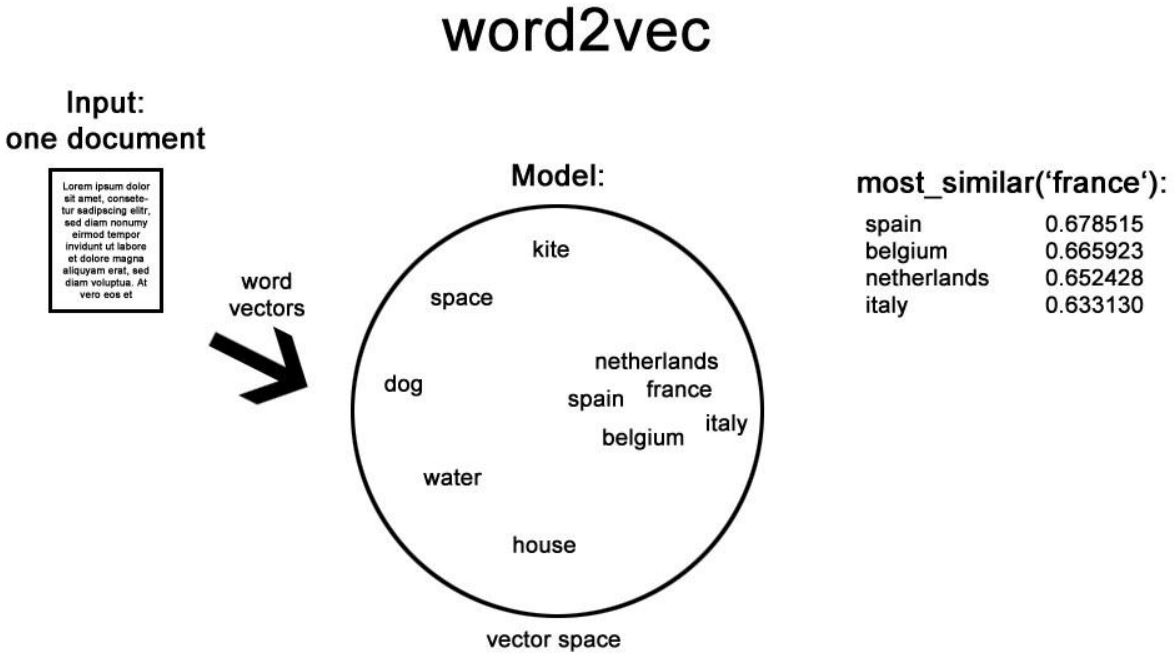
Часто використовуються відстані до найближчих сусідів: ми беремо набір точок і до кожної точки дивимось, припустимо, 100 найближчих сусідів, зберігаємо їхні індекси і відстані до них. Інші точки ми ігноруємо.

Як показує практика, якщо ми маємо справу із реально великою кількістю точок, 100 найближчих сусідів достатньо для того, щоб викрити локальну і нелокальну структуру в розподілі даних.

Ідея в тому, що не потрібно знати, на скільки Ви знайомі із кожним із 7 мільярдів людей на Планеті, достатньо знати список друзів однієї конкретної людини. І, якщо ми візьмемо весь граф списку друзів кожної людини, то ми можемо в сукупності відтворити структуру загального графу соціальних зв’язків людей на всій Землі.

Аналогічна проблема існує з обробкою тексту. Символи в слові мало говорять про сенс цього слова. Відповідно потрібно перекодувати слова в таку репрезентацію, в якій вектор, що відповідає слову, міститиме більше інформації про сенс цього слова і його зв'язок з іншими словами.

Відомий алгоритм для цього – **word2vec**, який дозволяє навчити комп'ютер працювати зі словами на більш високоабстрактному рівні.



Припустимо, ми маємо весь текст української Вікіпедії. Після фільтрації від усіх непотрібних символів та частин її розмір складе порядку 8–9 Гб тексту. Відповідно ми маємо основу для навчання без вчителя – велика кількість даних із купою прихованих закономірностей. Ми хочемо навчити комп'ютер читати слова на більш високоабстрактному рівні, ніж просто набір символів.

Що ми робимо? Ми робимо словник, де записуємо не тільки всі слова, що вживалися у Вікіпедії, але й скільки разів вживалося кожне слово. Візьмемо, наприклад, всі слова, які зустрічались принаймні 10 разів. В результаті отримуємо словник, припустимо, на 100 тисяч слів. Відповідно весь наш гігантський дата сет на 9 Гб ми можемо перекодувати і замінити кожне слово на номер в словнику. Відповідно вся Вікіпедія виглядатиме, як величезна послідовність індексів слів.

List

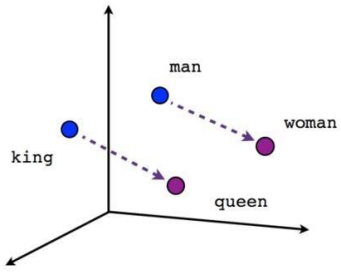
index	value
0	"Eggs"
1	"Milk"
2	"Cheese"
3	"Yogurt"
4	"Butter"
5	"More Cheese"

Dictionary

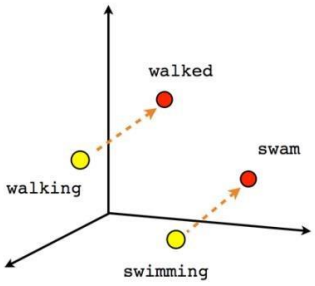
key	value
'Eggs'	2.59
'Milk'	3.19
'Cheese'	4.80
'Yogurt'	1.35
'Butter'	2.59
'More Cheese'	6.19

Сформулюємо для неї задачу наступним чином: маючи індекс певного слова, спрогнозувати, які слова його будуть оточувати. Ці дані у нас наявні, відповідно ми автоматично отримуємо мітки із сусідніх слів або центрального слова.

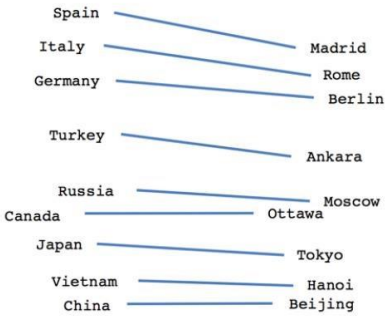
Стовимірні вектори, що відповідають словам «автомобіль» і «машина», будуть досить подібні і набагато ближчі, ніж, наприклад, слово «гуляти». В результаті навчання системи, ми можемо пройтись по всьому нашому словнику, дістати стовимірну векторну репрезентацію кожного слова і отримати словник стовимірної карти: сто тисяч точок (слів), яким відповідатимуть асоціативно пов'язані слова, які зустрічаються в подібних контекстах. Більше того, в даних векторних просторах ми отримуємо багато переваг у порівнянні із звичайною роботою із текстами. Окрім того, що сусідні за значенням слова проектуватимуться у сусідні точки, досить часто паралельним векторам відповідатимуть аналогії. Наприклад, якщо взяти вектор від точки «чоловік» і точки «король», і додати його до точки «жінка», то ми помітимо, що він вкаже на точку «королева». Відповідно паралельним перенесенням відповідатимуть аналогії.



Male-Female



Verb tense



Country-Capital

Якщо конструювати таку модель тексту з декількох мов, то паралельне перенесення відповідатиме перекладам з однієї мови на іншу. Звісно, точність методу залежить від різноманітності даних, кількості тренувань тощо.

Але сама ідея векторної репрезентації слів для обробки і аналізу тексту принципово змінила підхід до всіх задач natural language processing – обробки природньої мови. Тепер ми можемо думати про слова, як певні точки в гіперпросторах, додавати, віднімати, брати середнє значення від декількох слів і екстрагувати точки, які відповідають певним неявним сенсам. Ми можемо спостерігати закономірності в формулюваннях речень, яким відповідають специфічні траєкторії в гіперпросторі (кожне слово – точка в багатовимірному просторі, речення – траєкторія руху по цим точкам в цьому просторі). Аналізуючи траєкторії, закономірності, відхилення, ми можемо класифікувати сенс, який міститься у реченнях, а не просто аналізувати його на рівні правил заміни, символів, префіксів тощо.

Питання для самоконтролю

1. Дайте визначення поняття відношення.
2. Що таке бінарне відношення?
3. Поясніть поняття область визначення відношення.
4. Поясніть поняття область значень відношення.
5. Що таке фактор множина?

6. Яке відношення називається повним?
7. Яке відношення називається тотожним?
8. Яке відношення називається порожнім?
9. Які існують операції над відношеннями?
10. Які існують способи представлення відношень?
11. Що таке метричній простір?
12. Наведіть приклади метрик.
13. Чим відрізняються евклідові метрики від неевклідових?
14. Яка множина називається відкритою в метричному просторі?
15. Яка множина називається замкненою в метричному просторі?
16. Яка множина називається обмеженою?
17. Яка множина називається зв'язною?
18. Який простір називається повним?
19. Що таке ізометрія?
20. Дайте визначення стискаючого відображення.
21. Поясніть принцип стискуючих відображень.

Самостійна робота №4
Тема: Метричні основи інтелектуального аналізу даних

Завдання для виконання

1. Написати програму для обчислення Евклідової метрики (L2).
2. Написати програму для обчислення метрики Хеммінга (L1).
3. Порівняти швидкість обчислення на даних різної розмірності.

Приклад програми обчислення Евклідової метрики на мові Python.

```
import matplotlib
import numpy
import perfplot
```

```
    from scipy.spatial import distance
    def linalg_norm(data):
        a, b = data
        return numpy.linalg.norm(a-b, axis=1)
    def sqrt_sum(data):
        a, b = data
        return      numpy.sqrt(numpy.sum((a-b)**2,
axis=1))
    def scipy_distance(data):
        a, b = data
        return list(map(distance.euclidean, a, b))
    def mpl_dist(data):
        a, b = data
        return  list(map(matplotlib.mlab.dist,  a,
b))
    def sqrt_einsum(data):
        a, b = data
        a_min_b = a - b
        return  numpy.sqrt(numpy.einsum('ij,ij->i',
a_min_b, a_min_b))
    perfplot.show(
        setup=lambda n: numpy.random.rand(2, n, 3),
        n_range=[2**k for k in range(15)],
        kernels=[linalg_norm,      scipy_distance,
mpl_dist, sqrt_sum, sqrt_einsum],
        logx=True,
        logy=True,
        xlabel='len(x), len(y)'
    )
```

Практична робота №4
Тема: Метричні основи інтелектуального аналізу даних

Завдання для виконання

1. Для бінарних відношень визначити їх властивості. Для відношень еквівалентності знайти класи еквівалентності та фактор – множини**.**

Відношення визначено на множині $N \times N$: $< a,b > R < c, d > <=>$ [(ad = bc і b ≠ 0 і d ≠ 0) або (a = c, b = 0, d = 0)];

Відношення визначено на множині *Z*: $xRy <=> x \le y + 1$;

Відношення визначено на множині *N*: $xRy <=>$ НЗД(x,y) ≠ 1;

Відношення визначено на множині *R*: $xRy <=> y = |x|$;

Відношення визначено на множині *Z*: $xRy <=> (x^2 - y^2)/5$;

Відношення визначено на множині *N*: $xRy <=> x/y$;

Відношення визначено на множині *R*: $xRy <=> |x - 2y| \in N$;

Відношення визначено на множині *Z*: $xRy <=> 2x = 3y$;

Відношення визначено на множині *N*: $xRy <=> |x + 5| \ge |3 - y|$;

Відношення визначено на множині *R*: $xRy <=> xy > 1$;

Відношення визначено на множині *Z*: $xRy <=> 3/(x - y)$;

Відношення визначено на множині *N*: $xRy <=> (x - y)/m, m > 0$;

Відношення визначено на множині *Z*: $xRy <=> 3/(x + y)$;

Відношення визначено на множині *N*: $xRy <=>$ *НЗД(x,y) = x*;

Відношення визначено на множині $N \times N$: $< a,b > R < c, d > <=> a + d = d + c$;

Відношення визначено на множині { 5, 7, 9, 10, 13, 15, 18, 19,20}; $xRy <=> |x - 2y|/5$;

Відношення визначено на множині { 5, 8, 9, 12, 13, 16, 18, 19,20}; $xRy <=>> |x - y|/4$;

Відношення визначено на множині {7, 9, 10, 14, 15, 18, 19,21}; $xRy <=> |x - y|/7$;

Відношення визначено на множині {6, 9, 10, 12, 15, 18, 19,20}; $xRy <=> |x - y|/6$;

Відношення визначено на множині {2, 2,5, 3, 3,5, 5, 5,5, 8, 8,5}; $xRy <=> |x - y| = k \in N$;

Тест до теми 4

1. В R^1 метрика може бути задана формулою

1) $\rho(x_1, x_2) = x_1 - x_2$

2) $\rho(x_1, x_2) = \sqrt{{x_1}^2 + {x_2}^2}$

3) $\rho(x_1, x_2) = |x_1 - x_2|$
4) $\rho(x_1, x_2) = 1$
5) $\rho(x_1, x_2) = x_1 + x_2$

2. Відкритими в R^1 є множини
1) R^1 и $\{1\}$
2) $\emptyset$ и $(0; 3)$
3) R^1 и $[0; +\infty)$
4) $(0; 3) \cup [5; 6]$ и $\emptyset$
5) $(0; +\infty)$ и $\{0\}$

3. Замкнутими в R^1 є множини
1) $\emptyset$ и $(0; 1]$
2) R^1 и $[0; 1]$
3) $[0; 1] \cup (2; 3)$ и $[0; 5]$
4) $\emptyset$ и $(0; +\infty)$
5) R^1 и $[0; 5]$

4. Щільними в R^1 є множини
1) N
2) Z
3) $[0; 1]$
4) Q
5) $(0; 3)$

5. Відображення f метричного простору M в себе називається стискаєючим якщо
1) $\forall x_1, x_2 \in M: \rho(f(x_1), f(x_2)) < \rho(x_1, x_2)$
2) $\exists 0 < \alpha < 1: \forall x_1, x_2 \in M: \rho(f(x_1), f(x_2)) \leq \alpha\rho(x_1, x_2)$
3) $\forall x_1, x_2 \in M: \rho(f(x_1), f(x_2)) \leq \rho(x_1, x_2)$
4) $\forall x_1 \in M\ \exists x_2 \in M: \rho(f(x_1), f(x_2)) < \rho(x_1, x_2)$
5) $\exists 0 < \alpha < 1: \forall x_1 \in M\ \exists x_2 \in M: \rho(f(x_1), f(x_2)) \leq \alpha\rho(x_1, x_2)$

6. Відображення $f(x) = 4x + 4x^2$ відображає в себе простір
1) $[0; 1]$

2) $[-1; 1]$
3) $[-1; 0]$
4) $[-1/2; 0]$
5) $[0; 4]$

Рекомендована література

1. Борисенко О.А. Диференціальна геометрія і топологія. – Х.: Основа, 1995. – 304 с.
2. Бояришева Т.В., Гудивок Т.В., Погоріляк О.О. Функціональний аналіз. Навчальний посібник для студентів спеціальностей «математика», «прикладна математика», «статистика». – Ужгород, 2013. – 125 с.
3. Журавлёв Ю.И. Распознавание. Математические методы. Программная система. Практические применения. / Журавлёв Ю.И., Рязанов В.В., Сенько О.В. — М.: Изд. «Фазис», 2006. — 176 с.
4. Зиновьев А. Ю. Визуализация многомерных данных. / Зиновьев А. Ю. — Красноярск: Изд. Красноярского государственного технического университета, 2000. — 180 с.
5. Ильин В.А., Позняк Э.Г. Линейная алгебра: Учеб. для вузов. – М.: Наука, 1999. – 296 с.
6. Лепский А.Е., Броневич А.Г. Математические методы распознавания образов: Курс лекций. – Таганрог: Изд-во ТТИ ЮФУ, 2009. – 155 с.
7. Олійник А. О. Інтелектуальний аналіз даних : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – 278 с.
8. Ус. С.А. Функціональний аналіз [Текст]: навч. посібник / С.А. Ус. – Д. : Національний гірничий університет, 2013. – 236с.

ТЕМА 5 СТАТИСТИЧНІ МЕТОДИ АНАЛІЗУ ДАНИХ

Теоретичний матеріал до лекції
План
5.1 Кореляційний аналіз
5.2 Регресійний аналіз

Основні поняття: кореляційна залежність, коефіцієнт кореляції, парна кореляція, множинна кореляція, додатна кореляція, від'ємна кореляція, регресія, лінійна регресія, нелінійна регресія, інтерполяція, екстраполяція, залишок.

Статистичний аналіз включає велику різноманітність методів. Існує велика різноманітність прикладних пакетів, що реалізують широкий спектр статистичних методів, їх також називають універсальними пакетами або інструментальними наборами.

5.1 Кореляційний аналіз

Кореляційний аналіз застосовується для кількісної оцінки взаємозв'язку двох наборів даних, представлених в безрозмірному вигляді. Кореляційний аналіз дає можливість встановити, чи асоційовані набори даних по величині.

Кореляційною залежністю Y від X називають функцію $Y=F(X)$. Рівняння $Y=F(X)$ називають рівнянням регресії Y на X, а її графік – лінією регресії Y на X.

Кореляційний аналіз розглядає два завдання.

Перше завдання теорії кореляції – встановити форму кореляційного зв'язку, тобто вид функції регресії (лінійна, квадратична і так далі).

Друге завдання теорії кореляції – оцінити силу (щільність) кореляційного зв'язку. Щільність кореляційного зв'язку (залежності) оцінюється за величиною розсіювання значень

навколо умовного середнього.

Щільність зв'язку визначають за величиною коефіцієнта кореляції, який може приймати значення від –1 до +1 включно.

Коефіцієнт кореляції, завжди позначається латинською буквою r, використовується для визначення наявності взаємозв'язку між двома властивостями.

Відповідно до «Таблиці Чеддока» кореляція вважається:

дуже високою	при r = 0,9 – 0,99,
високою	при r = 0,7 – 0,89,
значною	при r = 0,5 – 0,69,
помірною	при r = 0,3 – 0,49,
слабкою	при r = 0,1 – 0,29.

Парна кореляція – це зв'язок між двома ознаками: результативною і факторною або двома факторними.

Коефіцієнт кореляції Пірсона r, який є безрозмірним індексом в інтервалі від –1,0 до 1,0 включно, відображає ступінь лінійної залежності між двома множинами даних.

Показник щільності зв'язку між двома ознаками визначається за формулою лінійного коефіцієнта кореляції:

$$r_{y/x} = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

де x – значення факторної ознаки;

y – значення результативної ознаки;

n – число пар даних.

Варіанти зв'язку, що характеризують наявність або відсутність лінійного зв'язку між ознаками:

– великі значення з одного набору даних пов'язані з великими значеннями іншого набору (додатна кореляція) – наявність прямого лінійного зв'язку;

– малі значення одного набору пов'язані з великими значеннями іншого (від'ємна кореляція) – наявність від'ємного лінійного зв'язку;

– дані двох діапазонів ніяк не пов'язані (нульова кореляція) – відсутність лінійного зв'язку.

Наприклад, візьмемо набір даних А. Необхідно визначити наявність лінійного зв'язку між ознаками x і y.

x	**y**
3	9
2	7
4	12
5	15
6	17
7	19
8	21
9	23,4
10	25,6
11	27,8

Для графічного представлення зв'язку двох змінних використана система координат з осями x і y. Побудований графік, який називається діаграмою розсіювання, показаний на рис. 5.1.

Дана діаграма показує, що низькі значення змінної x відповідають низьким значенням змінної y, високі значення змінної x відповідають високим значенням змінної y. Цей приклад демонструє наявність явного зв'язку.

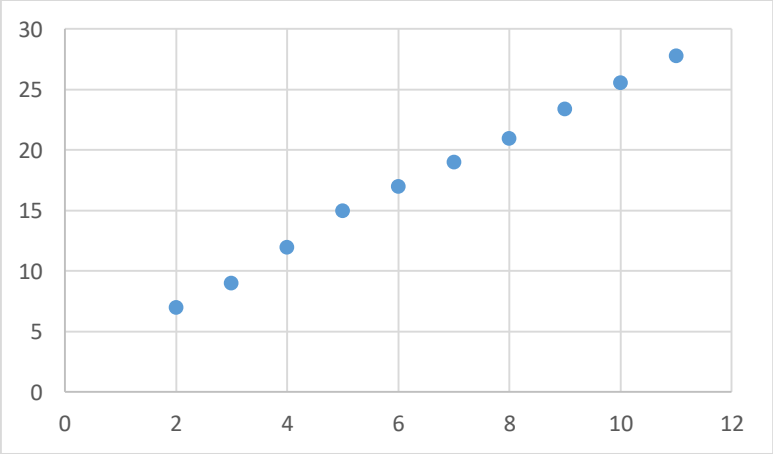


Рис. 5.1. Діаграма розсіювання

Таким чином, ми можемо встановити залежність між змінними x і y. Розрахуємо коефіцієнт кореляції Пірсона між двома масивами (x і y). В результаті отримуємо значення коефіцієнту кореляції 0,99.

Множинна кореляція. Множинною кореляцією називається кореляційний зв'язок між одним результативним і декількома факторними ознаками. Щільність зв'язку між результативною ознакою і факторними ознаками оцінюється коефіцієнтом множинної кореляції. Рівняння множинної кореляції буде мати вигляд: $\overline{y}_{x_1,x_2} = a_0 + a_1 x_1 + a_2 x_2$.

Коефіцієнт множинної кореляції буде мати вигляд:

$$R_{y/x_1x_2} = \sqrt{\frac{r^2_{y/x_1} + r^2_{y/x_2} - 2r_{y/x_1} \cdot r_{y/x_2} \cdot r_{x_1/x_2}}{1 - r^2_{x_1/x_2}}},$$

де r_{y/x_1}, r_{y/x_2}, r_{x_1/x_2} – лінійні коефіцієнти кореляції.

Будь-яка залежність між змінними володіє двома важливими властивостями: величиною і надійністю. Чим сильніше залежність між двома змінними, тим більше величина залежності і тим легше передбачити значення однієї змінної за значенням іншої змінної. Величину залежності легше виміряти, ніж надійність.

Надійність залежності не менш важлива, ніж її величина. Надійність залежності характеризує, наскільки ймовірно, що ця залежність буде знову знайдена на інших даних. З ростом величини залежності змінних її надійність зазвичай зростає.

5.2 Регресійний аналіз

Основна особливість регресійного аналізу, в тому, що можна отримати конкретні відомості про те, яку форму і характер має залежність між досліджуваними змінними.

Послідовність етапів регресійного аналізу

1. Формулювання завдання. На цьому етапі формуються

попередні гіпотези про залежність досліджуваних явищ.

2. Визначення залежних і незалежних змінних.

3. Формулювання гіпотези про форму зв'язку (проста або множинна, лінійна або нелінійна).

4. Визначення функції регресії (полягає в розрахунку чисельних значень параметрів рівняння регресії)

5. Оцінка точності регресійного аналізу.

6. Інтерпретація отриманих результатів. Отримані результати регресійного аналізу порівнюються з попередніми гіпотезами. Оцінюється коректність і правдоподібність отриманих результатів.

7. Передбачення невідомих значень залежної змінної.

За допомогою регресійного аналізу можливе вирішення задач прогнозування і класифікації.

Прогнозні значення обчислюються шляхом підстановки в рівняння регресії незалежних змінних.

Рішення задачі класифікації здійснюється таким чином: лінія регресії ділить всю множину об'єктів на два класи, і та частина множини, де значення функції більше нуля, належить до одного класу, а та, де воно менше нуля, – до іншого класу.

Завдання регресійного аналізу

Розглянемо основні завдання регресійного аналізу: встановлення форми залежності, визначення функції регресії, оцінка невідомих значень залежної змінної.

Встановлення форми залежності

Характер і форма залежності між змінними можуть утворювати такі різновиди регресії:

- додатна лінійна регресія (виражається в рівномірному зростанні функції);
- додатна рівноприскорена зростаюча регресія;
- додатна рівноуповільнена зростаюча регресія;
- від'ємна лінійна регресія (виражається в рівномірному

падінні функції);

– від'ємна рівноприскорена спадна регресія;
– від'ємна рівноуповільнена спадна регресія.

Однак згадані різновиди зазвичай зустрічаються не в чистому вигляді, а в поєднанні один з одним. У такому випадку говорять про комбіновані форми регресії.

Визначення функції регресії

Друге завдання зводиться до з'ясування впливу на залежну змінну головних чинників або причин, при незмінних інших умов, і за умови виключення впливу на залежну змінну випадкових елементів. Функція регресії визначається у вигляді математичного рівняння того або іншого типу.

Лінійна регресія:

$$Y=a+bx$$

Нелінійна регресія. Нелінійною регресією називається зв'язок, що виражається нелінійною залежністю.

Рівняння експоненційної регресії:

$$Y=a*exp(bx).$$

Рівняння гіперболічної регресії:

$$Y=a+b/x.$$

Рівняння показникової регресії:

$$Y=a*b^x.$$

Рівняння логарифмічної регресії:

$$Y=a+b*log(x).$$

Рівняння параболічної регресії:

$$Y=a+b_1x+b_2x^2.$$

Рівняння степеневої регресії

$$Y=ax^n.$$

Таке різноманіття форм зв'язку необхідно для того, щоб отримати найкраще наближення того чи іншого рівняння зв'язку до емпіричної лінії регресії.

Оцінка невідомих значень залежної змінної

Вирішення цього завдання зводиться до вирішення задачі

одного з типів:

– оцінка значень залежної змінної всередині розглянутого інтервалу вхідних даних, при цьому вирішується **задача інтерполяції**;

– оцінка майбутніх значень залежної змінної, тобто знаходження значень поза заданого інтервалу вихідних даних; при цьому вирішується **задача екстраполяції**.

Розглянемо деякі припущення, на які спирається регресійний аналіз.

Припущення лінійності, тобто передбачається, що зв'язок між розглянутими змінними є лінійним. Так, в розглянутому прикладі ми побудували діаграму розсіювання і змогли побачити явний лінійний зв'язок. Якщо ж на діаграмі розсіювання змінних ми бачимо явну відсутність лінійного зв'язку, тобто присутній нелінійний зв'язок, слід використовувати нелінійні методи аналізу.

Припущення про нормальність залишків. Припускаємо, що розподіл різниці передбачених і спостережуваних значень є нормальним. Для візуального визначення характеру розподілу можна скористатися гістограмами залишків.

При використанні регресійного аналізу слід враховувати його основне обмеження. Воно полягає в тому, що регресійний аналіз дозволяє виявити лише залежності, а не зв'язки, що лежать в основі цих залежностей.

Регресійний аналіз дає можливість оцінити ступінь зв'язку між змінними шляхом обчислення передбачуваного значення змінної на підставі кількох відомих значень.

Рівняння лінійної регресії.

Рівняння регресії має такий вигляд: $Y = a + b * X$

За допомогою цього рівняння змінна Y виражається через константу a і кут нахилу прямої (або кутовий коефіцієнт) b, помножений на значення змінної X. Константу a також

називають вільним членом, а кутовий коефіцієнт – коефіцієнтом регресії або *B*-коефіцієнтом.

Параметри *a* і *b* знаходимо методом найменших квадратів. Метод найменших квадратів: лінія повинна бути проведена так, щоб сума квадратів відхилень фактичних даних від вирівняних була найменшою. Параметри *a* і *b* знаходимо за формулами:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \frac{\sum_{i=1}^{n} x_i^2 \cdot \left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

У більшості випадків (якщо не завжди) спостерігається певний розкид спостережень щодо регресійної прямої.

Залишок – це відхилення окремої точки (спостереження) від лінії регресії (передбаченого значення).

Таблиця 5.1

Регресійна статистика	
Множинний R	0,998364
R–квадрат	0,99673
Нормований R–квадрат	0,996321
Стандартна помилка	0,42405
Спостереження	10

Спочатку розглянемо верхню частину розрахунків, представлену в таблиці 5.1, – регресійну статистику.

Величина R-квадрат, звана також мірою визначеності, характеризує якість отриманої регресійної прямої. Це якість

виражається ступенем відповідності між вхідними даними і регресійною моделлю (розрахунковими даними). Міра визначеності завжди знаходиться в межах інтервалу [0; 1].

У більшості випадків значення R-квадрат знаходиться між цими значеннями, званими екстремальними, тобто між нулем і одиницею.

Якщо значення R-квадрата близько до одиниці, це означає, що побудована модель пояснює майже всю мінливість відповідних змінних. І навпаки, значення R-квадрата, близьке до нуля, означає погану якість побудованої моделі.

У нашому прикладі міра визначеності дорівнює 0,99673, що говорить про дуже хорошій підгонці регресійної прямої до вихідних даних.

Множинний R – коефіцієнт множинної кореляції R – виражає ступінь залежності незалежних змінних (X) і залежною змінною (Y).

Множинний R дорівнює квадратному кореню з коефіцієнта детермінації, ця величина приймає значення в інтервалі від нуля до одиниці.

У простому лінійному регресійному аналізі множинний R рівний коефіцієнту кореляції Пірсона. Дійсно, множинний R в нашому випадку рівний коефіцієнту кореляції Пірсона з попереднього прикладу (0,998364).

Тепер розглянемо частину розрахунків, представлену в таблиці 5.2. Тут дані коефіцієнт регресії b (2,305454545) і зміщення по осі ординат, тобто константа a (2,694545455).

Виходячи з розрахунків, можемо записати рівняння регресії таким чином: Y = x * 2,305454545 + 2,694545455.

Таблиця 5.2

Коефіцієнти регресії

	Коефіцієнти	*Стандартна*	*t-статистика*

		помилка	
Y-перетин	2,694545455	0,33176878	8,121757129
Змінна X 1	2,305454545	0,04668634	49,38177965

Напрямок зв'язку між змінними визначається на підставі знаків (від'ємний або додатний) коефіцієнтів регресії (коефіцієнта b).

У нашому випадку знак коефіцієнта регресії додатний, отже, зв'язок також є додатним.

Якщо знак при коефіцієнті регресії - від'ємний, зв'язок залежної змінної з незалежної є від'ємним (обернений).

Слід враховувати, що даний приклад є досить простим і далеко не завжди можлива якісна побудова регресійної прямої лінійного виду.

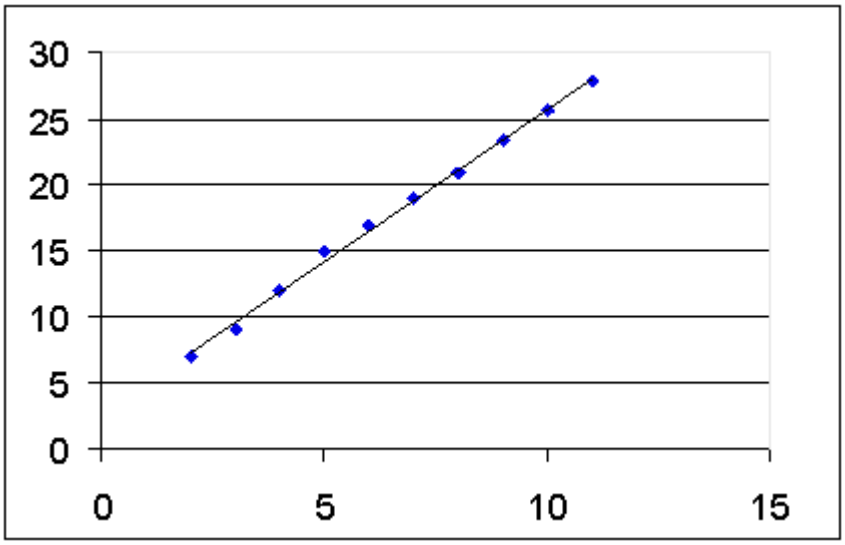


Рис. 5.2. Вихідні дані і лінія регресії

Маючи рівняння регресії, завдання прогнозування зводиться до вирішення рівняння Y = x * 2,305454545 + 2,694545455 з відомими значеннями x.

Питання для самоконтролю

1. Дайте визначення кореляційної залежності.
2. Назвіть основні завдання кореляційного аналізу.
3. Які значення може приймати коефіцієнт кореляції?

4. Які існують рівні щільності кореляційного зв'язку?
5. В чому відмінність парної та множинної кореляції?
6. Для якого види зв'язку використовують коефіцієнт кореляції Пірсона?
7. Назвіть основні етапи регресійного аналізу.
8. Назвіть завдання регресійного аналізу.
9. В чому полягає особливість інтерполяції.
10. В чому полягає особливість екстраполяції.
11. Що таке залишок?
12. Які величини характеризують якість регресійного аналізу?

Самостійна робота №5
Тема: Статистичні методи аналізу даних

Завдання для виконання

1. Для даних з самостійної роботи №2 , які завантажено з мережі Інтернет. Здійснити конвертацію **Excel в ARFF.**

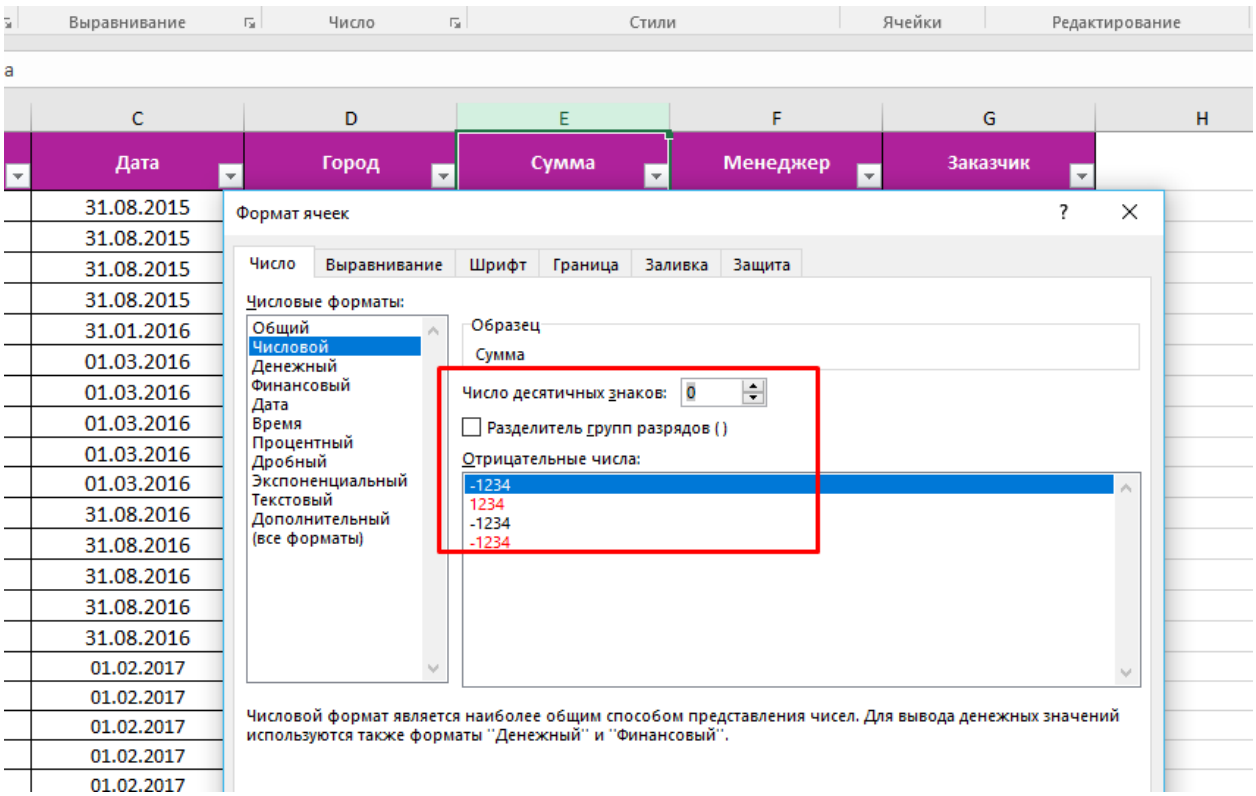
Конвертація Excel в ARFF

Для того що б конвертація пройшла успішно нам потрібно перевести базу на англійську мову. В іншому випадку програмне забезпечення Weka покаже таблицю з незрозумілими символами, так виходить через неправильного встановленого кодування. Нижче наведені приклад перекладу:

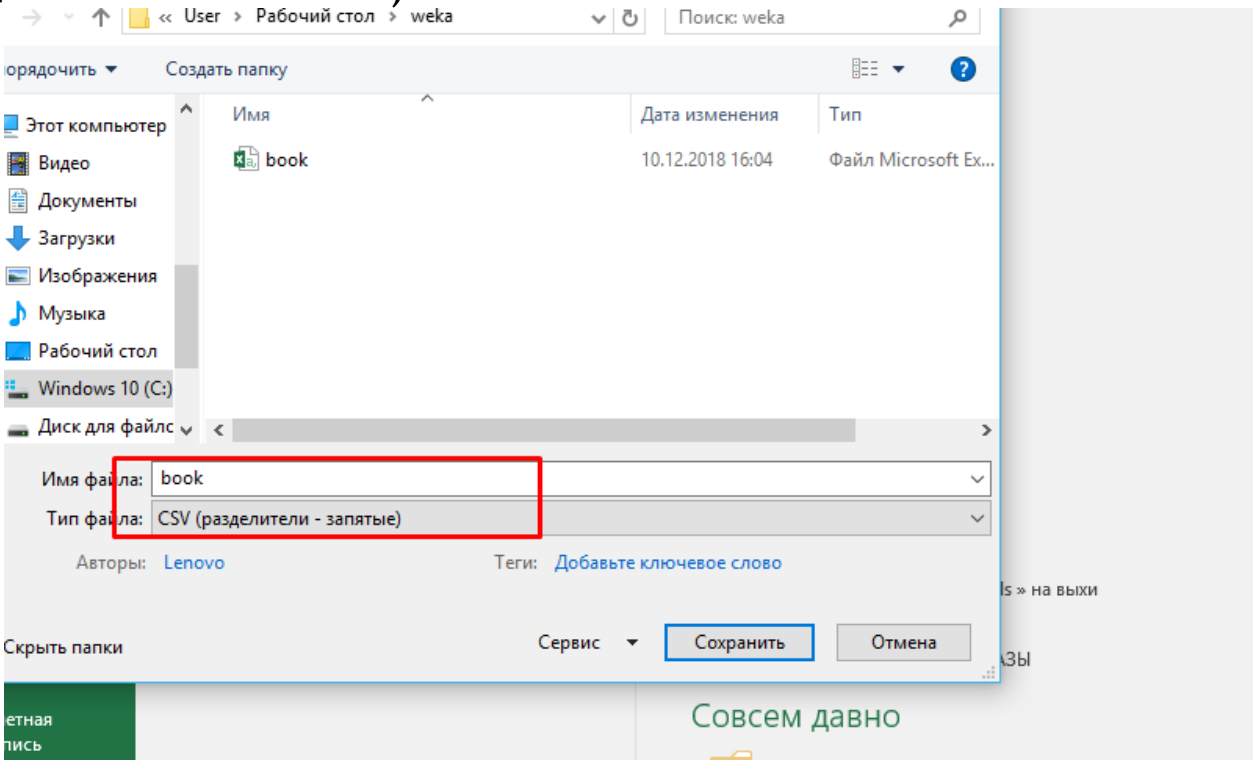
Товар	Категория	Дата	Город	Сумма	Менеджер
JiaYu S2	телефон	31.08.2015	Луганск	4 950 UAH	Дмитрий
Lenovo A588t	телефон	31.08.2015	Суммы	3 175 UAH	Дмитрий
Jeep Z6	телефон	31.08.2015	Бердичев	2 675 UAH	Дмитрий
Leagoo Lead 5	телефон	31.08.2015	Днепр	2 600 UAH	Дмитрий
Elephone M1	телефон	31.01.2016	Артемовск	3 873 UAH	Артур
Lenovo s856	телефон	01.03.2016	Березенка	2 436 UAH	Артур
NO.1 X-men X1	наушники	01.03.2016	Мариуполь	378 UAH	Артур
NO.1 X-men X1	наушники	01.03.2016	Днепр	378 UAH	Артур
iPhone 6S	телефон	01.03.2016	Днепр	3 724 UAH	Артур
Meizu M1 Note	телефон	01.03.2016	Днепр	3 920 UAH	Артур
Sony Xperia Z1	телефон	31.08.2016	Суммы	4 185 UAH	Екатерина
HTC One (E8)	телефон	31.08.2016	Луганск	3 375 UAH	Екатерина
NO.1 X-men X1	наушники	31.08.2016	Запорожье	359 UAH	Екатерина
UMI Hammer	телефон	31.08.2016	Чернигов	3 350 UAH	Екатерина
Sony Xperia Z1	телефон	31.08.2016	Донецк	4 185 UAH	Екатерина
Oukitel K6000 Pro	телефон	01.02.2017	Киев	4 176 UAH	Роман
Huawei Honor 6X	телефон	01.02.2017	Днепр	6 500 UAH	Роман
Lenovo Vibe P1	телефон	01.02.2017	Кривой рог	4 060 UAH	Роман
Blackview BV6000	телефон	01.02.2017	Тернополь	5 568 UAH	Роман

	A	B	C	D	E	F	
	Product	Category	Data	City	Total	Manager	C
	JiaYu S2	phone	31.08.2015	Lugansk	4950	Dmitrii	I
	Lenovo A588t	phone	31.08.2015	Summy	3175	Dmitrii	I
	Jeep Z6	phone	31.08.2015	Berdichev	2675	Dmitrii	I
	Leagoo Lead 5	phone	31.08.2015	Dnepr	2600	Dmitrii	I
	Elephone M1	phone	31.01.2016	Artemovsk	3873	Artur	I
	Lenovo s856	phone	01.03.2016	Berezenka	2436	Artur	I
	X-men X1	headphone	01.03.2016	Mariupol	378	Artur	I
	X-men X1	headphone	01.03.2016	Dnepr	378	Artur	I
	iPhone 6S	phone	01.03.2016	Dnepr	3724	Artur	I
	Meizu M1 Note	phone	01.03.2016	Dnepr	3920	Artur	I
	Sony Xperia Z1	phone	31.08.2016	Summy	4185	Katerina	W
	HTC One (E8)	phone	31.08.2016	Lugansk	3375	Katerina	W
	X-men X1	headphone	31.08.2016	Zaporozhe	359	Katerina	W
	UMI Hammer	phone	31.08.2016	Chernigov	3350	Katerina	W
	Sony Xperia Z1	phone	31.08.2016	Doneck	4185	Katerina	W
	Oukitel K6000 Pro	phone	01.02.2017	Kiev	4176	Roman	I
	Huawei Honor 6X	phone	01.02.2017	Dnepr	6500	Roman	I
	Lenovo Vibe P1	phone	01.02.2017	Krivoj rog	4060	Roman	I
	Blackview BV6000	phone	01.02.2017	Ternopol	5568	Roman	I

В даному випадку ми використовуємо цифри (сума) і дати. У таблиці Excel нам потрібно прибрати символ валюти і вказати число десяткових знаків 0. Для цього вибираємо стовпець **«ПКМ – Формат ячеек – Числовий»:**

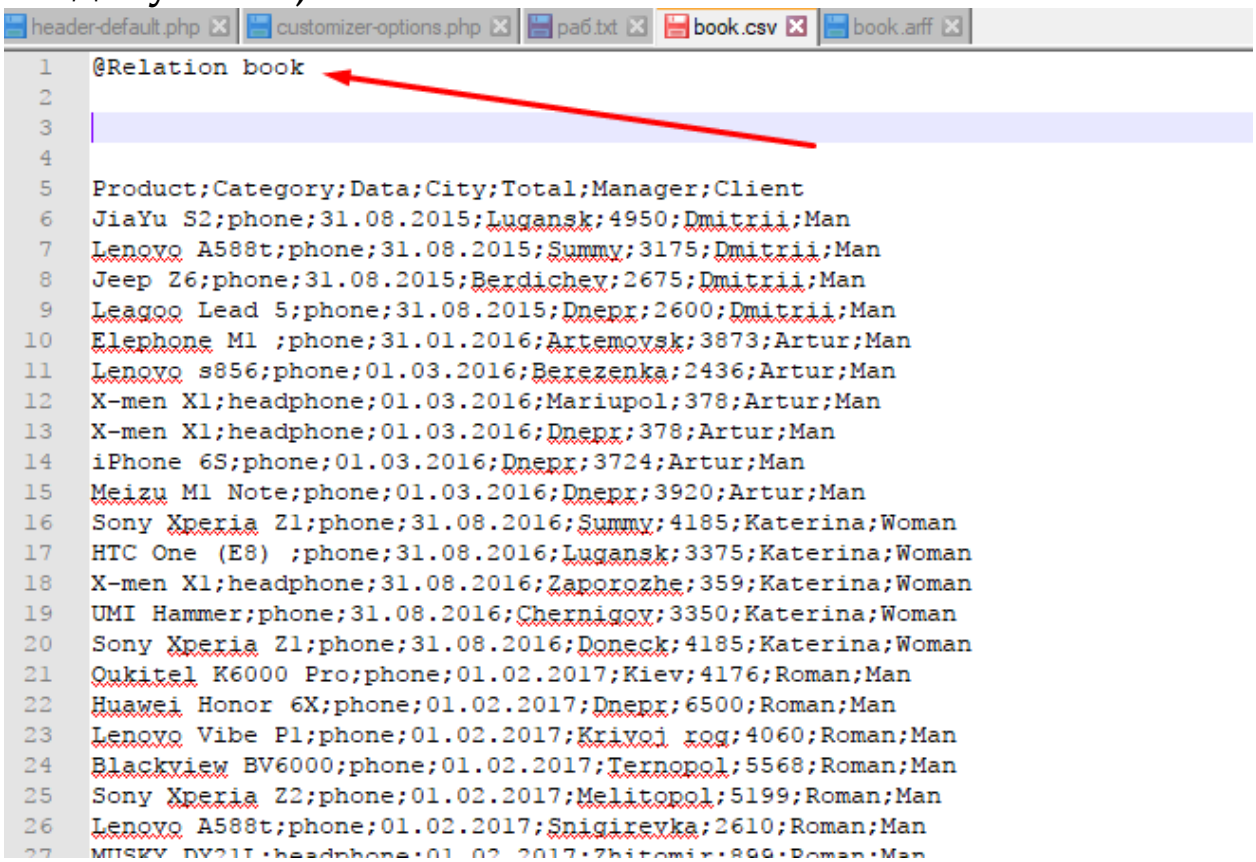


Після перекладу, зберігаємо таблицю в CSV, для цього переходимо **«Главная – Сохранить как – Тип файла (CSV разделители – запятые)»:**

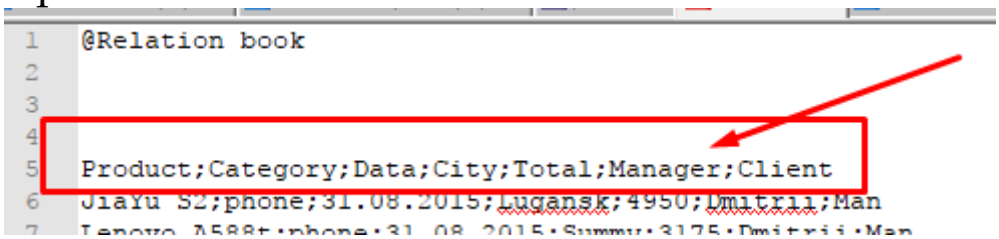


Переходимо до збереженого і відкриваємо через **«Блокнот»,** або будь–який інший текстовий редактор, наприклад **«Notepad ++».** Переходимо до першого рядка,

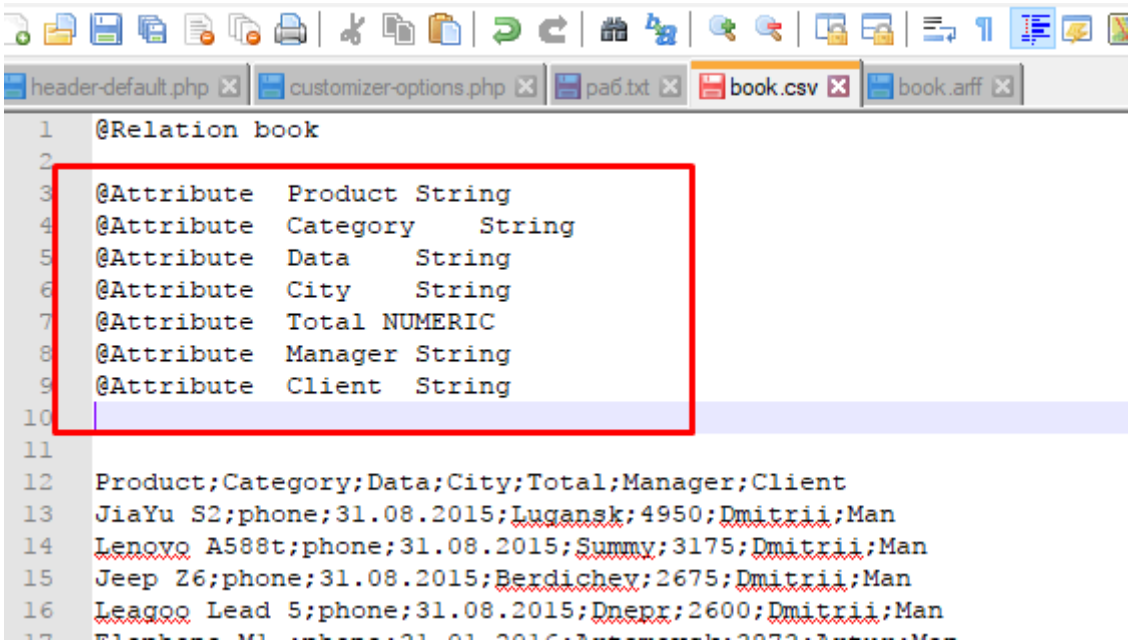
натискаємо **«Enter»** і вставляємо команду **«@Relation (назва вашого документа)»**.



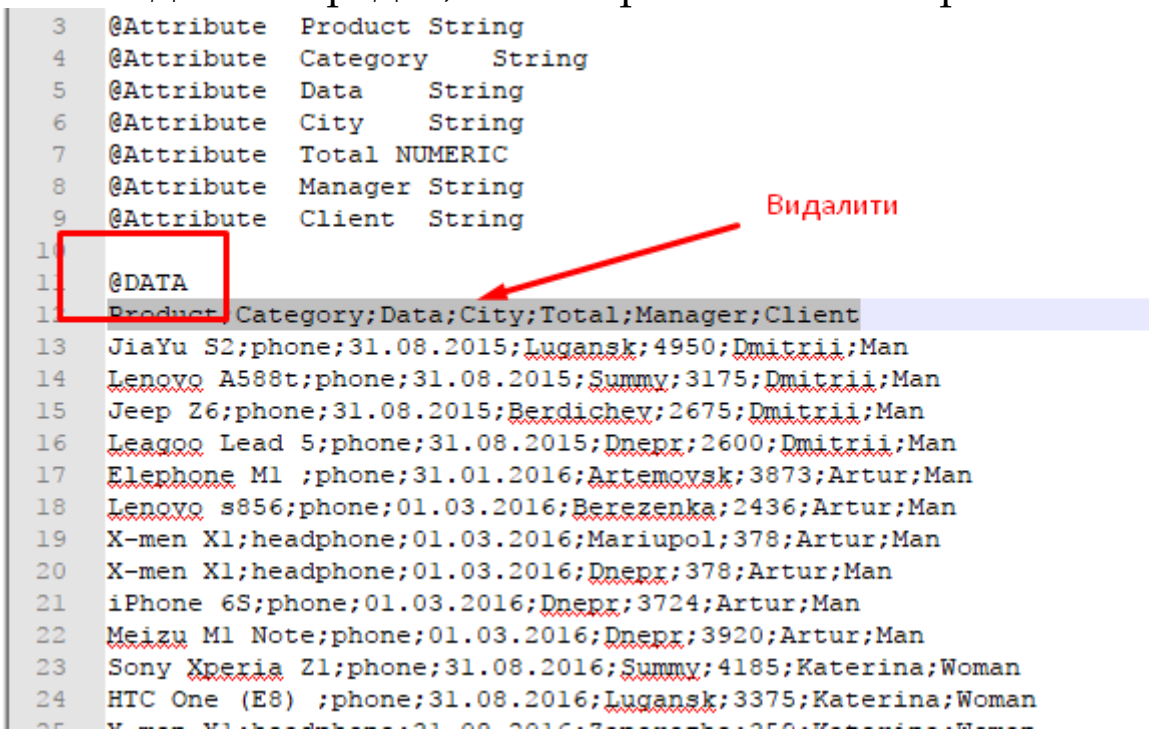
Далі ми спостерігаємо атрибути, які повинні відобразитися в програмі Weka:



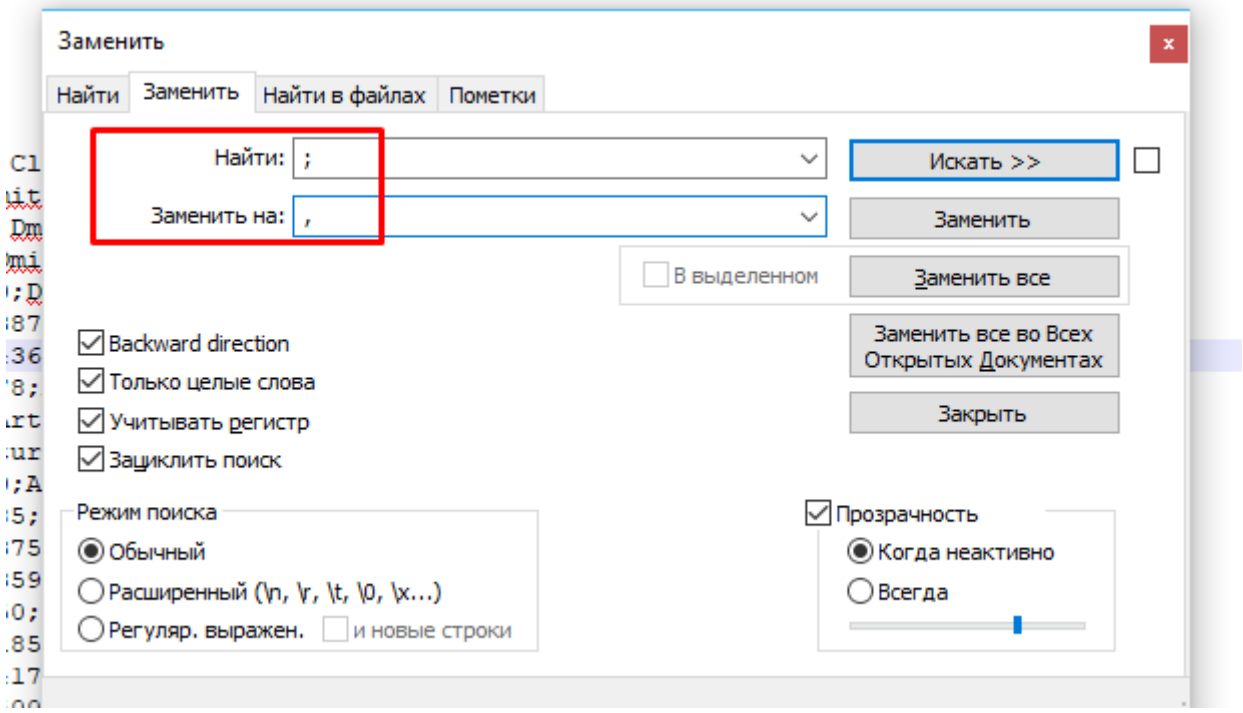
На жаль програма не зможе зрозуміти, що ми від неї хочемо і тут так само потрібно прописати вкладки, які будемо використовувати. Так само нам потрібно зрозуміти, що буде перебувати в цій категорії. Якщо це строковий тип (назва товару, міста і тд.) То команда буде виглядати так – **@Attribute (назва вкладки) String**. Якщо ж це числові значення (сума, кількість і тд.) То команда буде виглядати так – **@Attribute (назва вкладки) NUMERIC**. Виходячи з даної інформації ми прописуємо категорії, наприклад:



Далі нам потрібно до створених категоріям підключити базу, яку програма буде аналізувати. Дописуємо команду **@DATA** і видаляємо рядок, в якій прописані категорії:



Тепер нам потрібно замінити крапку з комою на звичайну кому. Для цього натискаємо **Ctrl + H** або **«Правка – Заменить»**. І робимо заміну с **;** на **,**



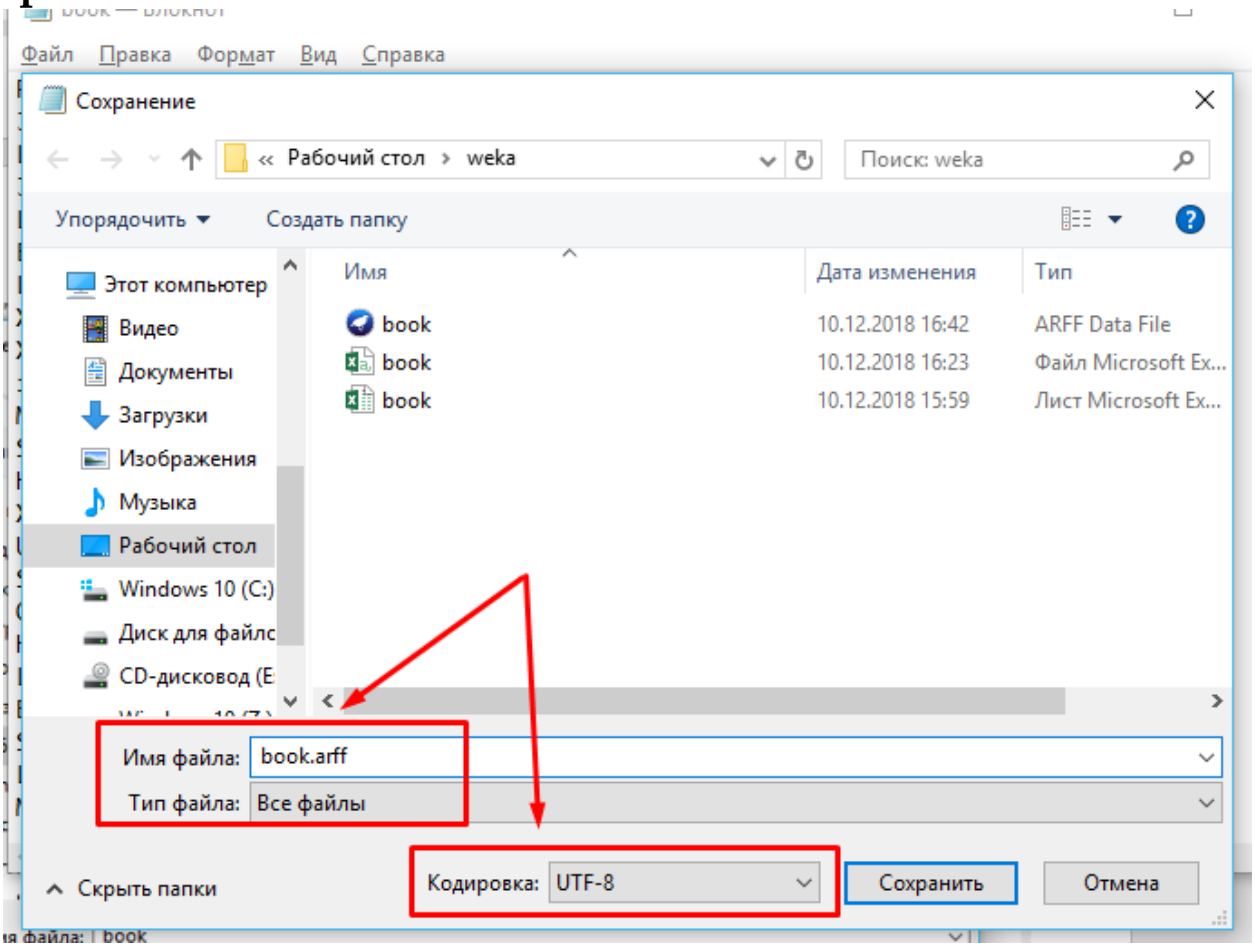
І так само нам потрібно замінити Пропуск на будь-який інший символ, наприклад тіре –

```
@DATA
LG,phone,31.08.2015,Lugansk,4950,Dmitrii,Man
Lenovo-A588t,phone,31.08.2015,Summy,3175,Dmitrii,Man
Jeep-Z6,phone,31.08.2015,Berdichev,2675,Dmitrii,Man
Leagoo-Lead-5,phone,31.08.2015,Dnepr,2600,Dmitrii,Man
Elephone-M1-,phone,31.01.2016,Artemovsk,3873,Artur,Man
Lenovo-s856,phone,01.03.2016,Berezenka,2436,Artur,Man
X-men-X1,headphone,01.03.2016,Mariupol,378,Artur,Man
X-men-X1,headphone,01.03.2016,Dnepr,378,Artur,Man
iPhone-6S,phone,01.03.2016,Dnepr,3724,Artur,Man
Meizu-M1-Note,phone,01.03.2016,Dnepr,3920,Artur,Man
Sony-Xperia-Z1,phone,31.08.2016,Summy,4185,Katerina,Woman
HTC-One-E8,phone,31.08.2016,Lugansk,3375,Katerina,Woman
X-men-X1,headphone,31.08.2016,Zaporozhe,359,Katerina,Woman
UMI-Hammer,phone,31.08.2016,Chernigov,3350,Katerina,Woman
Sony-Xperia-Z1,phone,31.08.2016,Doneck,4185,Katerina,Woman
Oukitel-K6000-Pro,phone,01.02.2017,Kiev,4176,Roman,Man
Huawei-Honor-6X,phone,01.02.2017,Dnepr,6500,Roman,Man
Lenovo-Vibe-P1,phone,01.02.2017,Krivoj-rog,4060,Roman,Man
Blackview-BV6000,phone,01.02.2017,Ternopol,5568,Roman,Man
Sony-Xperia-Z2,phone,01.02.2017,Melitopol,5199,Roman,Man
Lenovo-A588t,phone,01.02.2017,Snigirevka,2610,Roman,Man
MUSKY-DY21L,headphone,01.02.2017,Zhitomir,899,Roman,Man
PPTV-King-7,phone,01.02.2017,Pavlograd,3538,Roman,Man
IUNI-i1,phone,01.02.2017,Pavlograd,2697,Roman,Man
Bluedio-Legend-Version,phone,31.05.2017,Herson,2080,Denis,Man
Samsung-Galaxy-S6-Active,phone,31.05.2017,Mariupol,6344,Roman,Man
DIGOOR-DG1,phone,31.05.2017,Harkov,4134,Roman,Man
Cubot-X15,phone,31.05.2017,Kiev,2860,Roman,Man
```

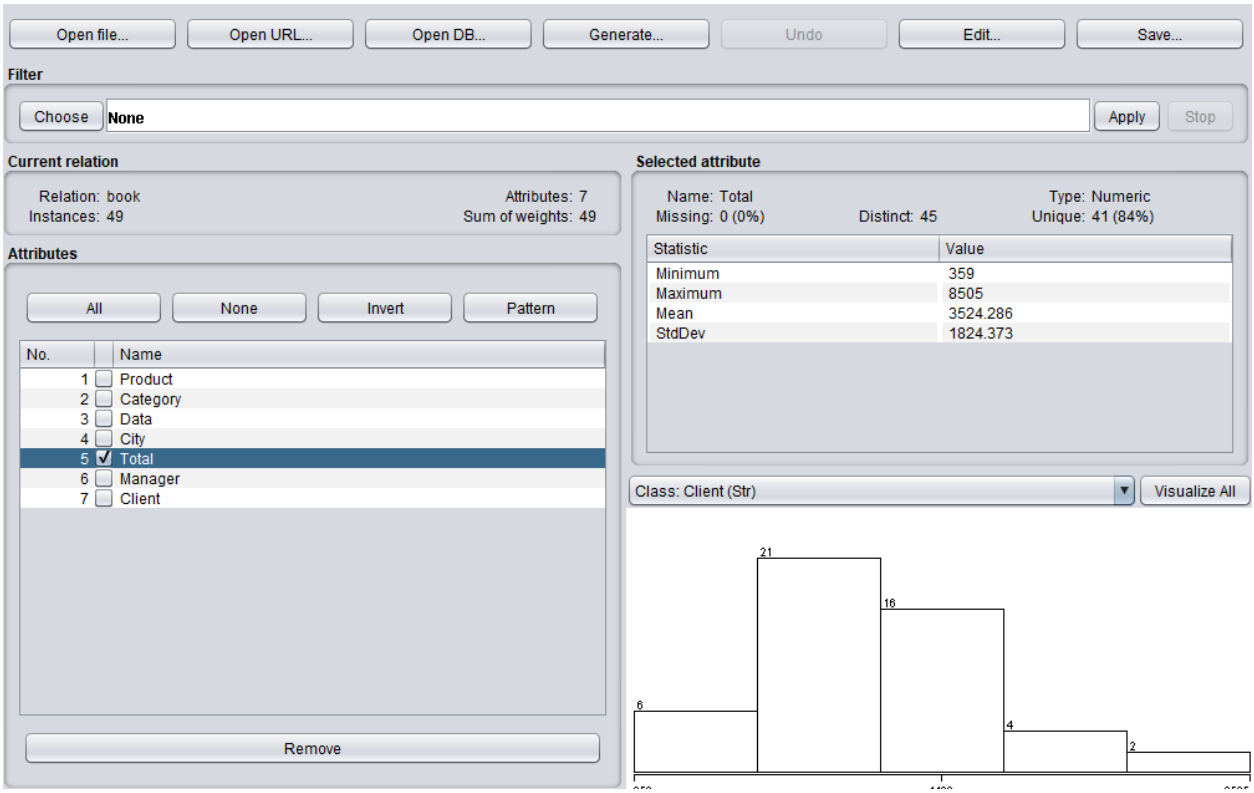
Зверніть увагу, що рядки де прописані команди, не повинні бути з прочерком або будь-яким іншим символами. Якщо напочатку у Вас з'явилися прочерк, то видаліть вручну.

```
1  @Relation book
2
3  @Attribute  Product String
4  @Attribute  Category    String
5  @Attribute  Data    String
6  @Attribute  City    String
7  @Attribute  Total NUMERIC
8  @Attribute  Manager String
9  @Attribute  Client  String
10
11 @DATA
```

Після заміни, зберігаємо файл в кодуванні **UTF-8** і даємо назву документу **.arff**. Переходимо **«Файл - Сохранить как – Кодировка UTF-8»**:



Запускаємо програмне забезпечення Weka, переходимо в **«Explorer – Open File – (назва вашого файла.arff) – OK»**. Якщо Ви зробили все вірно, то відобразитися готова таблиця адаптована під ПО Weka.



Помилки, які зустрічаються часто:

1. Невірно вибрано кодування, при збереженні.
2. Некоректно перекладена база.
3. Допущено пробіл \ символ в базі.

Практична (лабораторна) робота №5
Тема: Статистичні методи аналізу даних

Waikato Environment for Knowledge Analysis (WEKA), є вільно поширюваним програмним пакетом з відкритим вихідним кодом для аналізу даних. Завнтажити **WEKA** можливо на сторінці

http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

WEKA забезпечує графічний користувальницький інтерфейс для роботи з файлами даних і генерації візуальних результатів (у вигляді таблиць і графіків). Крім того, можливо інтегрувати WEKA, як і будь-яку іншу бібліотеку, у свої власні додатки, наприклад, для автоматизації аналізу даних на стороні сервера, використовуючи стандартний API. WEKA поширюється по ліцензії GNU General Public License (GPL).

Ця програма дає можливість виконувати завдання аналізу даних таких як:

- підготовка даних – попередня обробка;
- відбір ознак;
- кластеризація;
- класифікація, зокрема, дерева рішень;
- пошук асоціативних правил;
- регресійний аналіз;
- візуалізація результатів.

Рис 5.3 Стартове вікно WEKA

При запуску WEKA, пакет пропонує вам на вибір 4 графічних інтерфейси для роботи з WEKA і даними. Будемо використовувати опцію **Explorer**.

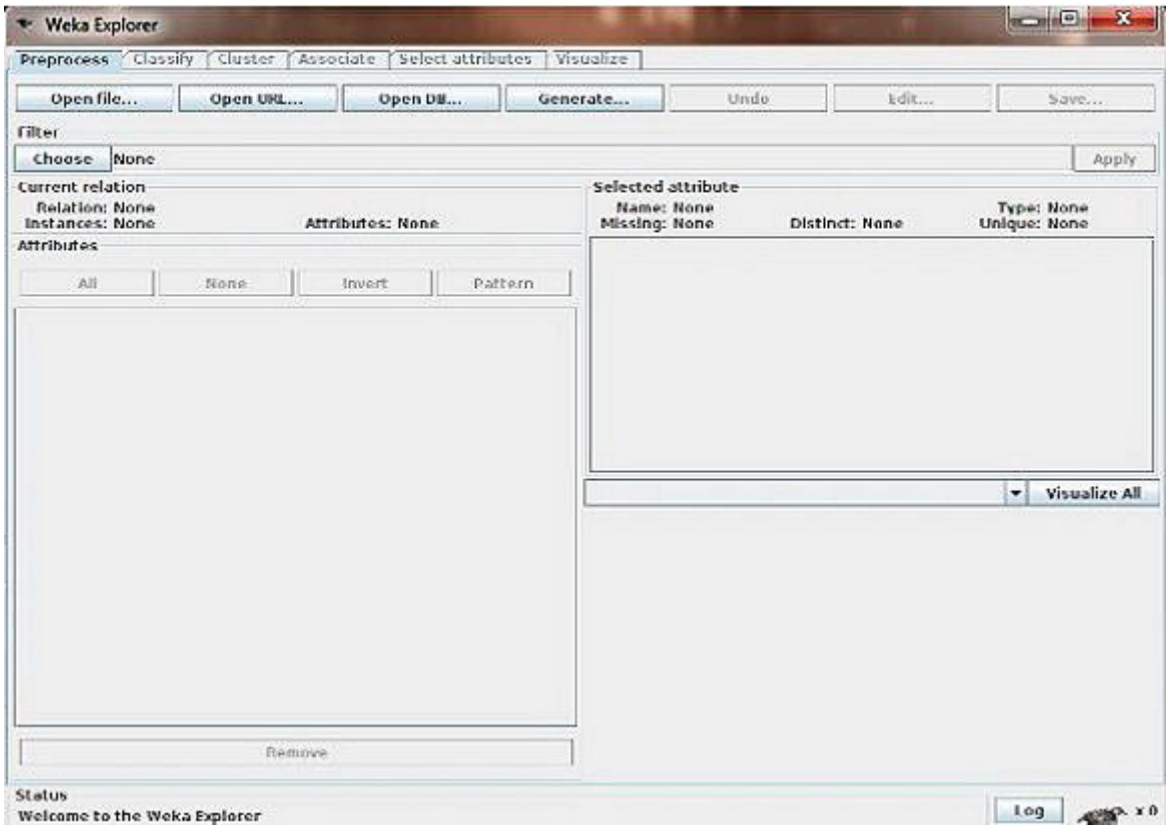


Рис 5.3 Вікно WEKA Explorer

Завдання для виконання

Для того щоб завантажити дані в WEKA, їх слід перетворити у формат, зрозумілий для цього програмного пакету. Найбільш підходящим форматом для завантаження даних в WEKA є формат Attribute-Relation File Format (ARFF), який спочатку визначає тип завантажуваних даних, а потім вказує власне дані. У файлі формату ARFF ви вказуєте назву і тип даних для кожного стовпця таблиці, а потім дані по рядках. У моделях регресійного аналізу використовуються всього два типи даних: **NUMERIC і DATE**. Після того, як ви описали всі стовпці таблиці, ви додаєте дані по рядках, використовуючи як роздільник кому. Нижче наведено файл ARFF з даними про ціни на будинки, які ми будемо використовувати для побудови нашої тестової моделі. Зверніть увагу, що в списку відсутній рядок з даними будинку, ціну для якого необхідно встановити. Зараз ми створюємо регресійну модель на базі відомих параметрів і, отже, не можемо включити в неї параметри нашого

будинку, оскільки ціна його невідома.

Файл даних для завантаження в WEKA

```
@RELATION house

@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE lotSize NUMERIC
@ATTRIBUTE bedrooms NUMERIC
@ATTRIBUTE granite NUMERIC
@ATTRIBUTE bathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC

@DATA
3529,9191,6,0,0,205000
3247,10061,5,1,1,224900
4032,10150,5,0,1,197900
2397,14156,4,1,0,189900
2200,9600,4,0,1,195000
3536,19994,6,1,1,325000
2983,9365,5,0,1,230000
```

Завантаження даних в WEKA

Тепер, коли файл з даними готовий, його потрібно завантажити в WEKA. Запустіть WEKA і виберіть опцію **Explorer**. В результаті відкриється закладка **Preprocess** вікна Explorer. Клацніть на кнопці **Open File** і виберіть створений вами ARFF-файл. Вікно WEKA Explorer із завантаженими даними про будинках показано на рис. 5.4.

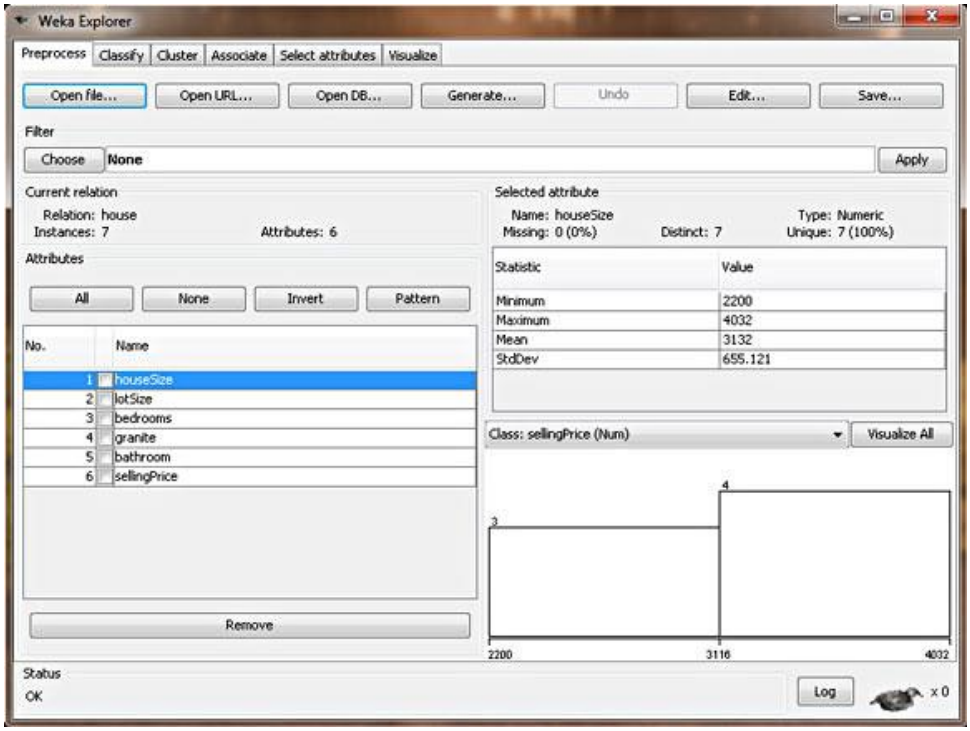


Рис. 5.4. Вікно WEKA Explorer із завантаженими даними про будинки

У цьому вікні ви можете перевірити дані, на підставі яких ви збираєтеся будувати модель. У лівій частині вікна **Explorer** показані параметри об'єктів (**Attributes**), які відповідають заголовкам стовпців нашої вихідної таблиці, а також вказано кількість об'єктів (**Instances**), тобто рядків таблиці. Якщо ви клацнете мишкою на одному із заголовків стовпців, то в правій панелі буде виведена повна інформація про набір даних в даному стовпці. Наприклад, якщо ми виберемо стовпець **houseSize** в лівій панелі (він обраний за замовчуванням), то в правій панелі відобразиться додаткова статистична інформація з цього стовпця. Буде показано максимальне значення в стовпці (4032 кв.футів) і мінімальне значення (2200 кв.футів). Крім того, буде підраховано середнє значення (3131 кв.фут) і стандартне відхилення (655 кв.футів) (стандартне відхилення – статистичний показник розсіювання значень випадкової величини). Нарешті, тут же вам пропонується можливість

візуального аналізу даних (кнопка **Visualize All**). Оскільки в нашій таблиці даних не так багато, то їх візуальне відображення не дає такої наочної аналітичної картини, як у випадку використання сотень або тисяч показників.

Давайте перейдемо від розгляду даних до створення моделі і визначимо, нарешті, вартість мого будинку.

Створення регресійної моделі в WEKA

Для того щоб створити модель, відкрийте закладку **Classify**. В якості першого кроку, нам треба вибрати тип моделі для аналізу, щоб вказати WEKA, яким чином ми хочемо аналізувати наші дані, і яку модель побудувати:

1. Клацніть на копанні **Choose** і розгорніть меню **functions**.
2. Виберіть опцію **LinearRegression** .

Таким чином, ми вказали WEKA, що ми хочемо створити модель регресійного аналізу. Меню включає безліч моделей. Якщо ви вибрали правильну модель, то вікно WEKA Explorer має виглядати так, як показано на рис. 5.5.

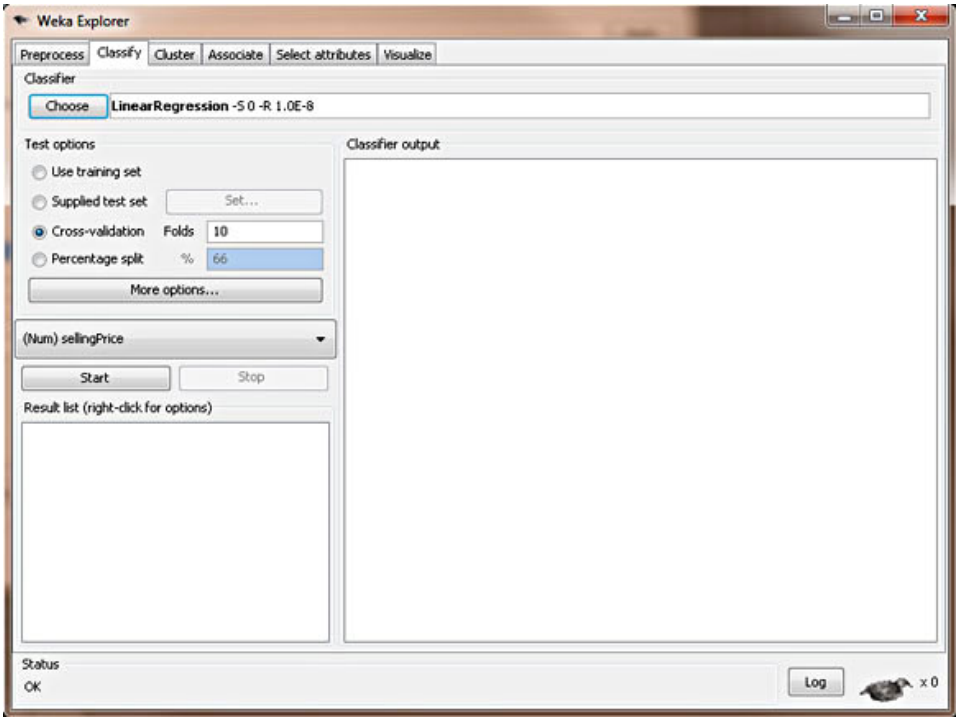


Рис. 5.5 Модель лінійного регресійного аналізу WEKA

Після того, як ми вибрали тип моделі, потрібно вказати

WEKA, які дані повинні використовуватися для її створення. Незважаючи на те, що відповідь на це питання для нас цілком очевидна – потрібно взяти дані зі створеного нами ARFF-файлу – існує кілька інших, більш складних, можливостей надання даних для аналізу. Опція **Supplied test set** дозволяє вказати додатковий набір тестових даних для моделі, опція **Cross-validation** використовує кілька наборів даних, усереднює їх і будує модель на основі середніх значень, а опція **Percentage split** використовує в якості бази для моделі процентилі набору даних. Ці способи застосовуються для створення аналітичних моделей. У разі регресійного аналізу нам потрібна опція **Use training set** . У цьому випадку WEKA створить модель на базі даних із завантаженого ARFF-файлу.

Завершальний етап створення моделі – вибір залежної змінної (колонка, в якій знаходиться невідоме нам значення, яке потрібно розрахувати). У нашому прикладі – це ціна будинку, оскільки, саме це значення ми і хочемо дізнатися. Відразу після секції **Test options** знаходиться список, що розкривається, в якому вам потрібно вибрати залежний параметр. Типово повинен бути вибраний атрибут **sellingPrice** . Якщо це не так, виберіть самі цей параметр.

Ми визначили всі параметри і можемо приступити до створення моделі. Натисніть кнопку **Start** . У результаті вікно WEKA має виглядати так, як показано на рис. 5.6.

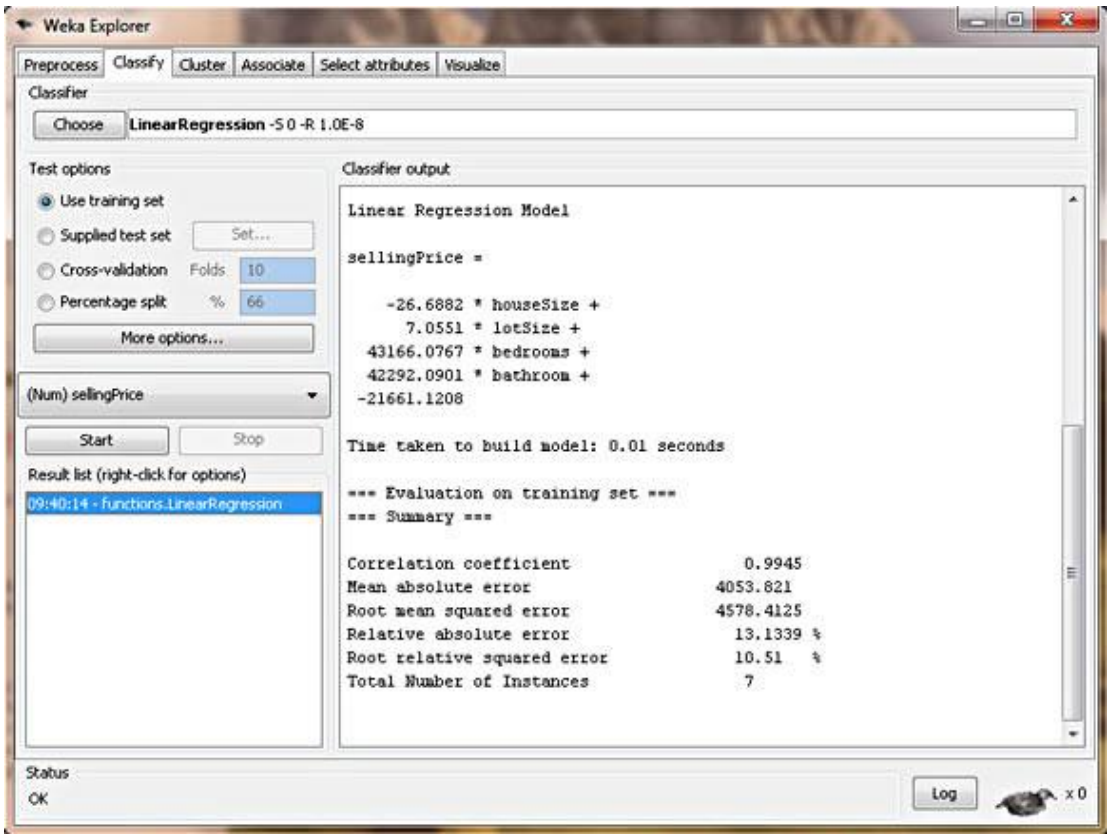


Рис. 5.6 Регресійна модель WEKA для розрахунку вартості будинку

Розберемо, які дані включені в результуючий висновок **Готова модель регресійного аналізу**

```
sellingPrice = (-26.6882 * houseSize) +
               (7.0551 * lotSize) +
               (43166.0767 * bedrooms) +
               (42292.0901 * bathroom)
              - 21661.1208
```

Далі в отриману модель для визначення вартості ми підставляємо параметри нашого будинку.

Тест до теми 5

1. Статистика – це
   а) наука, яка вивчає комп'ютерні алгоритми, що автоматично покращуються під час роботи;
   б) наука про методи збору даних, їх обробку і аналіз для виявлення закономірностей, властивих досліджуваному явищу;
   в) науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, що традиційно вважаються інтелектуальними.
2. Частина генеральної сукупності, певним способом відібрана з метою дослідження і отримання висновків про властивості та характеристики генеральної сукупності– це
   а) вибірка;
   б) змінна;
   в) параметр.
3. Числові характеристики генеральної сукупності – це
   а) вибірка;
   б) змінна;
   в) параметри.
4. Якщо коефіцієнт кореляції дорівнює 0,2 , то така кореляція вважається…
   а) високою;
   б) значною;
   в) слабкою.
5. Якщо коефіцієнт кореляції дорівнює 0,8 , то така кореляція вважається…
   а) високою;
   б) значною;
   в) слабкою.
6. Якщо коефіцієнт кореляції дорівнює 0,4 , то така кореляція вважається…

а) високою;
б) помірною;
в) слабкою.

7. Якщо коефіцієнт кореляції дорівнює 0,99 , то така кореляція вважається…
   а) дуже високою;
   б) значною;
   в) слабкою.
8. Кореляційний зв'язок між одним результативним і декількома факторними ознаками називається
   а) парною кореляцією;
   б) від′ємною кореляцією;
   в) множинною кореляцією.
9. Зв'язок між двома ознаками: результативною і факторною або двома факторними називається
   а) парною кореляцією;
   б) від′ємною кореляцією;
   в) множинною кореляцією.
10. Задача оцінки значень залежної змінної всередині розглянутого інтервалу вхідних даних
    а) екстраполяція;
    б) інтерполяція;
    в) кореляція.

Рекомендована література

1. Жалдак М.І., Кузьміна Н.М., Берлінська С.Ю. Теорія ймовірностей і математична статистика з елементами інформаційної технології.-К.: Вища школа,1995.–352с.

2. Жлуктенко В.І., Наконечний С.І. Теорія ймовірностей і математична статистика: Навч.-метод. посібник. У 2 ч. – Ч.1. Теорія ймовірностей. – К.: КНЕУ, 2000. – 304 с.

3. Лепский А.Е., Броневич А.Г. Математические методы распознавания образов: Курс лекций. – Таганрог: Изд–во ТТИ ЮФУ, 2009. – 155 с.

4. Олійник А. О. Інтелектуальний аналіз даних : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – 278 с.

5. Руденко О. Г. Штучні нейронні мережі / О. Г. Руденко, Є. В. Бодянський. – Харків : Компанія СМІТ, 2006. – 404 с.

6. Скобцов Ю. А. Основы эволюционных вычислений /Ю. А. Скобцов. – Донецк : ДонНТУ, 2008. – 330 с.

ТЕМА 6 ЗАДАЧІ КЛАСИФІКАЦІЇ Й КЛАСТЕРИЗАЦІЇ

Теоретичний матеріал до лекції
План
6.1 Задача класифікації
6.2 Постановка задачі кластеризації
6.3 Методи кластерного аналізу
6.3 Алгоритм k –середніх
6.4 Алгоритм с-середніх
6.5 Метод опорних векторів

Основні поняття: класифікація, метод опорних векторів, кластерізація, ієрархічні алгоритми, неієрархічні алгоритми, чіткі алгоритми, нечіткі алгоритми, метод ближнього сусіда, метод найбільш віддалених сусідів, центроїдний метод, дендограмма, алгоритм k-середніх, алгоритм с-середніх.

6.1 Задача класифікації

Класифікація є найбільш простою і одночасно найбільш часто розв'язуваною задачею аналізу даних. Задача класифікації відноситься до стратегії "навчання з учителем".

Класифікація – впорядкована за деяким принципом множина об'єктів, які мають подібні класифікаційні ознаки (одна або кілька властивостей), обраних для визначення подібності або відмінності між цими об'єктами.

Завданням класифікації часто називають передбачення категоріальної залежної змінної (тобто залежною змінною, яка є категорією) на основі вибірки неперервних і/або категоріальних змінних.

Наприклад, можна передбачити, хто з клієнтів фірми є потенційним покупцем певного товару, а хто – ні, хто скористається послугою фірми, а хто – ні, і т.д. Цей тип завдань відноситься до завдань бінарної класифікації, в них залежна

змінна може приймати тільки два значення (наприклад, так чи ні, 0 або 1).

Інший варіант класифікації виникає, якщо залежна змінна може приймати значення з деякої множини визначених класів. Наприклад, коли необхідно передбачити, яку марку автомобіля захоче купити клієнт. У цих випадках розглядається множина класів для залежної змінної.

Класифікація може бути одновимірною (за однією ознакою) і багатовимірною (за двома і більше ознаками).

Розглянемо задачу класифікації на простому прикладі. Припустимо, є база даних про клієнтів туристичного агентства з інформацією про вік і дохід за місяць. Є рекламний матеріал двох видів: дорожчий, комфортний відпочинок і дешевший, молодіжний відпочинок. Відповідно, визначені два класа клієнтів: клас 1 і клас 2. База даних приведена в таблиці 6.1.

Таблиця 6.1.

База даних клієнтів туристичного агентства

Код клієнта	**Вік**	**Дохід**	**Клас**
1	18	25	1
2	22	100	1
3	30	70	1
4	32	120	1
5	24	15	2
6	25	22	1
7	32	50	2
8	19	45	2
9	22	75	1
10	40	90	2

Завдання. Визначити, до якого класу належить новий клієнт і який з двох видів рекламних матеріалів йому варто надсилати.

Для наочності представимо нашу базу даних в двомірному вимірі (вік і дохід), у вигляді множини об'єктів, що належать класам 1 (помаранчева мітка) і 2 (сіра мітка). На рис. 3.1 наведені об'єкти з двох класів.

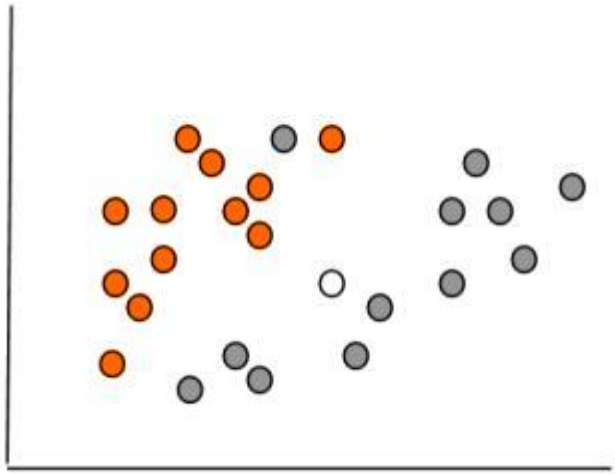


Рис. 6.1 Множина в двомірному просторі

Рішення нашої задачі буде полягати в тому, щоб визначити, до якого класу належить новий клієнт, на малюнку позначений білою міткою.

Набір вихідних даних (або вибірку даних) розбивають на дві множини: навчальну і тестову.

Навчальна множина (training set) – множина, яке включає дані, що використовуються для навчання (конструювання) моделі. Така множина містить вхідні та вихідні значення прикладів. Вихідні значення призначені для навчання моделі.

Тестова (test set) множина також містить вхідні та вихідні значення прикладів. Тут вихідні значення використовуються для перевірки працездатності моделі.

Процес класифікації складається з двох етапів: конструювання моделі та її використання.

1. Конструювання моделі:

– опис множини визначених класів;
– кожен приклад набору даних відноситься до одного класу;
– використовується навчальна множина, на якій

відбувається конструювання моделі;
- отримана модель представлена класифікаційними правилами, деревом рішень або математичною формулою.

2. Використання моделі:
- класифікація нових або невідомих значень;
- оцінка правильності (точності) моделі;
- відомі значення з тестового прикладу порівнюються з результатами використання отриманої моделі;
- рівень точності – відсоток правильно класифікованих прикладів в тестовій множині;
- тестова множина не повинна залежати від навчальної множини;
- якщо точність моделі допустима, можливе використання моделі для класифікації нових прикладів, клас яких невідомий.

Найбільш поширені методи, що застосовуються для вирішення задач класифікації: статистичні методи, зокрема, лінійна регресія, метод опорних векторів, класифікація за допомогою дерев рішень, класифікація за допомогою методу найближчого сусіда, класифікація за допомогою генетичних алгоритмів.

Схематичне рішення задачі класифікації методом лінійної регресії наведено на рис.6.1.

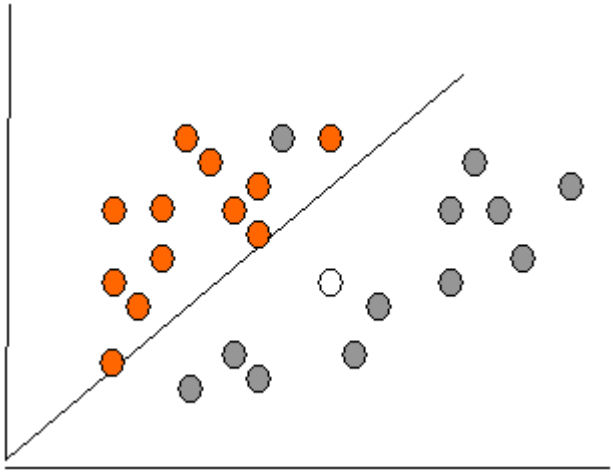


Рис. 6.2 Рішення задачі класифікації

методом лінійної регресії

Схематичне рішення задачі класифікації методом дерев рвшень:

```
if X > 5 then grey
    else if Y > 3 then orange
            else if X > 2 then grey
                    else orange
```

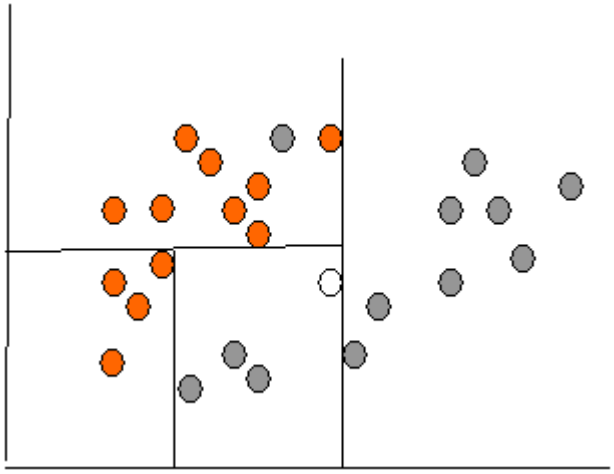


Рис. 6.3 Рішення задачі класифікації методом дерев рішень

6.2 Метод опорних векторів

Якщо у нас є певна множина точок, де кожна точка належить до одного з двох класів, ми відповідно маємо певну множину точок, яку можна розділити на дві частини певною прямою або гіперплощиною (якщо ми маємо справу з багатовимірними даними), яка буде знаходитись посередині між класами і відстань до найближчих точок буде максимальною.

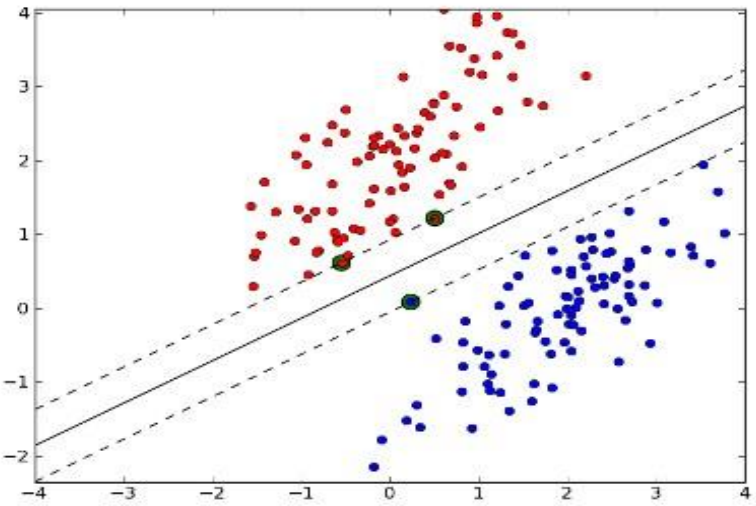


Рис. 6.4 Поділ двомірних даних на два класи

Метод опорних векторів (англ. SVM, support vector machine) вирішує задачу знаходження такої гіперплощини, відстань від якої до найближчих точок з кожного боку площини буде максимальною.

Методом опорних векторів ми шукаємо таку функцію, яка для кожного із прикладів першого класу буде більше 0, а із другого класу – менше 0. Мітки класів ми можемо розглядати відповідно як +1 та –1. В результаті нам потрібно знайти таку функцію, щоб відстань до найближчих векторів з кожного боку була максимальною. Саме такі вектори і називаються **опорними**, бо мають максимальний вплив на те, якою буде пряма, що розділяє два класи відповідає на питання, до якого класу належить точка. Опорні вектори зображені на рис. 6.5.

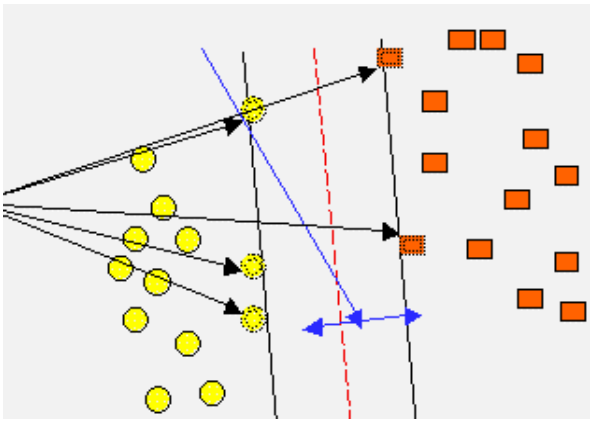


Рис. 6.5 Опорні вектори

Процес вирішення задачі методом опорних векторів зводиться до переформулювання задачі в систему рівнянь, в якій ми шукаємо такий набір вагових коефіцієнтів.

Звідки беруться вагові коефіцієнти? Нагадаємо, що пряма на площині задається рівнянням $y=kx+b$, де k – відповідає за нахил прямої, а b – за зміщення прямої вздовж осі. У випадку багатовимірної площини замість k ми маємо певну матрицю коефіцієнтів. Якщо замість x ми підставимо конкретний приклад і помножимо на транспоновану матрицю вагових коефіцієнтів, на виході отримаємо додатне чи від'ємне число. Таким чином ми можемо стверджувати, до якого з класів належить точка.

В реальності досить часто виникає ситуація, коли точки не можуть бути розділені лінійно. Буває, що класи мають невеликий перетин або не є ідеально лінійно роздільними. Таким чином ми намагаємось знайти таку пряму чи гіперплощину, в яких окрім того, що найближчі вектори з кожного боку будуть максимально віддаленими, сума різниць векторів з протилежного боку мають бути мінімальними. Це так зване «епсілон», яке ми додаємо в нашу систему рівнянь.

В такому випадку наша пряма чи гіперплощина не буде ідеально розділяти два класи, але це буде оптимальна пряма чи гіперплощина, яка на більшій кількості прикладів видасть нам необхідний розподіл.

Метод опорних векторів використовується досить широко через його неймовірну простоту та доступність реалізації. У випадку даних, які можна розділити лінійно і ми точно впевнені, що можна провести одну пряму чи гіперплощину і отримати точний результат належності точки до одного з класів, в процесі знаходження такої прямої чи гіперплощини ми

шукаємо певний набір вагових коефіцієнтів. Підставивши їх в наше рівняння, ми вважаємо, що наша пряма чи гіперплощина повинна проходити чітко посередині між найближчими прикладами з кожного класу.

Такі найближчі приклади будуть лежати на паралельних прямих чи гіперплощинах і називаються **опорними векторами**, оскільки вони впливають на те, під яким кутом знаходиться пряма чи гіперплощина.

Попри те, що після етапу формулювання системи рівнянь методи вирішення цієї системи не такі очевидні і описання математичного апарату такого рішення потребує значної кількості сил та часу, для того, щоб використовувати цей метод для вирішення прикладних задач із використанням доступних бібліотек.

Точність класифікації

Точність класифікації тестової множини порівнюється з точністю класифікації навчальної множини. Якщо класифікація тестової множини дає приблизно такі ж результати по точності, як і класифікація навчальної множини, вважається, що дана модель пройшла крос-перевірку.

6.3 Постановка задачі кластеризації

Кластеризація призначена для розбиття сукупності об'єктів на однорідні групи (кластери або класи). Якщо дані вибірки уявити як точки в просторі ознак, то задача кластеризації зводиться до визначення "згущень точок".

У таблиці 6.2 наведено порівняння задач класифікації та кластеризації.

Таблиця 6.2

Порівняння класифікації та кластеризації

Характеристика	**Класифікація**	**Кластеризация**
Контрольованість навчання	Контрольоване навчання	Неконтрольоване навчання
Стратегія	Навчання з вчителем	Навчання без вчителя

Наявність позначки класу	Навчальна множина супроводжується міткою, яка вказує клас, до якого належить спостереження	Мітки класу навчальної множини невідомі
Підстава для класифікації	Нові дані класифікуються на підставі навчальної множини	Дано множина даних з метою встановлення існування класів або кластерів даних

Дано набір даних з наступними властивостями:

– кожний екземпляр даних виражається однаковою кількістю параметрів із значенням;
– клас для кожного конкретного екземпляру даних до початку дослідження невідомий.

Потрібно знайти:

– спосіб порівняння даних між собою(міру схожості);
– спосіб кластеризації;
– розбиття даних на кластери.

Формально задача кластеризації описується наступним чином. Дано множину об'єктів даних *I*, кожен з яких представлений набором атрибутів. Потрібно побудувати множину кластерів *C* і відображення *F* множини *I* на множину *C, F: I –> C*. Відображення *F* задає модель даних, що являється власне вирішенням задачі.

Множина I визначається наступним чином

$$I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\},$$

де i_j – об'єкт, що досліджується.

Кожен із об'єктів характеризується набором параметрів:

$$i_j = \{x_1, x_2, \ldots, x_h, \ldots, x_m\}$$

Задача кластеризації полягає в побудові множини:

$$C = \{c_1, c_2, \ldots, c_k, \ldots, x_g\},$$

де c_k – кластер, який має в собі похожі один на одного об'єкти з множини *I*:

$$c_k = \{i_l, i_p \mid i_l \in I, i_p \in I, d(i_l, i_p) < \sigma\},$$

де σ – величина, яка визначає міру близькості для

включення об’єктів в один кластер, $d(i_l, i_p)$ – відповідно міра близькості.

6.4 Методи кластерного аналізу

В загальному можна виділити дві основні класифікації алгоритмів кластерного аналізу за ієрархією та чіткістю.

Ієрархічні і неієрархічні. *Ієрархічні* алгоритми (так звані алгоритмами таксономії) будують не одне розбиття вибірки на непересічні кластери, а систему вкладеного розбиття. На виході ми отримуємо дерево кластерів, коренем якого є вся вибірка, а листям – найбільш дрібні кластера. *Неієрархічні* алгоритми будують одне розбиття об'єктів на кластери;

Чіткі і нечіткі. Чіткі (або непересічні) алгоритми кожному об'єкту вибірки ставлять у відповідність номер кластера, тобто кожен об'єкт належить тільки одному кластеру. *Нечіткі* (або пересічні) алгоритми кожному об'єкту ставлять у відповідність набір значень, що показує ступінь відношення об'єкта до кластерів. Тобто кожен об'єкт відноситься до кожного кластера з певною ймовірністю.

Методи ієрархічних алгоритмів.

Метод ближнього сусіда або одиночний зв'язок. Тут відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) в різних кластерах. Цей метод дозволяє виділяти кластери складної форми за умови, що різні частини таких кластерів з'єднані ланцюжками близьких один до одного елементів. В результаті роботи цього методу кластери представляються довгими ланцюжками.

Метод найбільш віддалених сусідів або повний зв'язок. Тут відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (найбільш віддаленими сусідами). Метод добре використовувати, коли об'єкти дійсно відбуваються з різних груп. Якщо ж кластери мають подовжену форму то цей метод не слід використовувати.

Метод Варда. За відстань між кластерами береться приріст суми квадратів відстаней об'єктів до центрів кластерів, що

отримується в результаті їх об'єднання. На відміну від інших методів кластерного аналізу для оцінки відстаней між кластерами, тут використовуються методи дисперсійного аналізу. На кожному кроці алгоритму об'єднуються такі два кластери, які призводять до мінімального збільшення цільової функції, суми квадратів елементів кожної групи. Цей метод направлений на об'єднання близько розташованих кластерів і використовується для створення кластерів невеликого розміру.

Метод невиваженого попарного середнього. Як відстань між двома кластерами береться середня відстань між усіма парами об'єктів в них. Цей метод слід використовувати, якщо об'єкти дійсно належать до різних груп, коли кластери мають подовжену форму, та якщо є припущення щодо істотних відмінностей в розмірах кластерів.

Метод зваженого попарного середнього. Цей метод схожий на метод невиваженого попарного середнього, різниця полягає лише в тому, що тут в якості вагового коефіцієнта використовується розмір кластера (число об'єктів, що містяться в кластері).

Незважений центроїдний метод. Як відстань між двома кластерами в цьому методі береться відстань між їх центрами тяжкості. Цей метод переважно використовувати у випадках, якщо є припущення щодо істотних відмінностей в розмірах кластерів.

Зважений центроїдний метод. Цей метод схожий на попередній, різниця полягає в тому, що для обліку різниці між розмірами кластерів (числа об'єктів в них), використовуються ваги. Цей метод переважно використовувати у випадках, якщо є припущення щодо істотних відмінностей в розмірах кластерів.

Методи ієрархічних алгоритмів

Серед методів ієрархічної кластеризації виділяються два основні типи: висхідні і низхідні алгоритми.

Низхідні алгоритми працюють за принципом «зверху-вниз»: на початку всі об'єкти поміщаються в один кластер, який

потім розбивається на всі більш дрібні кластери.

Більш поширені висхідні алгоритми, які на початку роботи поміщають кожен об'єкт в окремий кластер, а потім об'єднують кластери в усе більші, поки всі об'єкти вибірки не будуть міститися в одному кластері. Результати таких алгоритмів зазвичай представляють у вигляді дерева – дендрограми.

Для обчислення відстаней між кластерами частіше за все користуються двома відстанями: одиночній зв'язок або повний зв'язок. До недоліку ієрархічних алгоритмів можна віднести систему повного розбиття, яка може бути зайвою в контексті розв'язуваної задачі.

Велику популярність при вирішенні задач кластеризації набули алгоритми, засновані на пошуку оптимального розбиття множини даних на кластери (групи). Дані алгоритми намагаються групувати дані в кластери таким чином, щоб цільова функція алгоритму розбиття досягала екстремуму (мінімуму). В даних алгоритмах використовуються наступні базові поняття:

- вхідна множина даних;
- метрика відстані;
- вектор центрів кластерів;
- матриця розбиття кластерів U;
- цільова функція;
- набір обмежень.

Однією з основних проблем кластерного аналізу є визначення кількості кластерів, що для багатьох алгоритмів є вхідним параметром. Розбиття на кластери можуть суттєво відрізнятися в залежності від обраної кількості кластерів.

Існує ряд методів для обчислення оптимальної кількості кластерів заснованих на :

- індексах, які порівнюють ступені "розкиду" даних усередині кластерів і між кластерами;
- розрахунку значень характеристик (функцій стійкості), що показують відповідність призначених кластерів для окремих елементів;

– статистиках, що визначають найбільш ймовірне рішення;
– оцінюванні цільності розподілів.

Розглянемо декілька відомих алгоритмів на основі індексів.

Перший підхід (Calinski-Harabasz) обирає кількість кластерів як значення аргументу, максимізує функцію *CH(K) :*

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)}$$

де *B(K)* і *W(K)*, відповідно, зовнішня і внутрішня суми квадратів елементів даних з *K* кластерами. Це один з найперших запропонованих методів. Він виявляється ефективним при даних невеликої розмірності.

Підхід Krzanowski-Lai максимізує функцію *KL(K):*

$$KL(K) = \left|\frac{Diff(K)}{Diff(K+1)}\right|$$

де

$$Diff(K) = (K-1)^{\frac{2}{p}}W(K-1) - (K)^{\frac{2}{p}}W(K)$$

де, p – розмірність, *W(K)* – внутрішня сума квадратів елементів даних з *K* кластерами. Основна ідея полягає в вимірі порядку мінливості внутрішніх дисперсій.

Суть одного з можливих удосконалень полягає в знаходженні більш "хороших" початкових значень центроїдів кластерів.

Точний алгоритм виглядає наступним чином:

– вибрати перший центр ваги випадковим чином (серед усіх точок);
– для кожної точки знайти значення квадрата відстані до найближчого центроїда (з тих, які вже обрані) – dx^2;
– вибрати з цих точок наступний центроїд так, щоб ймовірність вибору точки була пропорційна обчисленому для неї квадрату відстані. Це можна зробити наступним чином. На кроці 2 потрібно паралельно з розрахунком dx^2 підраховувати суму $Sum(dx^2)$. Після накопичення суми знайти значення Rnd = random (0.0,1.0) * $Sum(dx^2)$. Rnd

випадковим чином вкаже на число з інтервалу [0; Sum), і нам залишається тільки визначити, якій точці це відповідає. Для цього потрібно знову почати підраховувати суму S (dx^2) до тих пір, поки сума не перевищить Rnd. Як тільки це станеться, підсумовування зупиняється, і ми можемо взяти поточну точку в якості центроїда. При виборі кожного наступного центроїда спеціально стежити за тим, щоб він не збігся з однією з вже обраних в якості центроїда точок, не потрібно, тому що ймовірність повторного вибору деякої точки дорівнює 0;

– повторювати кроки 2 і 3 до тих пір, поки не будуть знайдені всі необхідні центроїди.

6.4.1 Алгоритм k –середніх

Алгоритм k-середніх (англ. k-means) був запропонований ще в 1979 році, проте він не втрачає своєї актуальності і сьогодні. Ідея, що лежить в його основі, досить проста, і алгоритм працює досить ефективно. Однак, оптимальність рішень, отриманих за допомогою алгоритму k-середніх, не гарантована. Крім того, недоліком алгоритму є необхідність знати число кластерів заздалегідь.

Формально рішення задачі кластеризації полягає в тому, щоб «помітити» кожен з наявних об'єктів – приписати йому номер певного класу. Алгоритм k-середніх передбачає таке розбиття об'єктів на класи, при якому мінімізуються відмінності «відстані» між об'єктами одного і того ж класу та максимізуються вдстань між об'єктами різних класів.

Даний алгоритм являється прообразом практично всіх алгоритмів нечіткої кластеризації, і його розгляд допоможе нам краще зрозуміти принципи, що закладені в більш складні алгоритми.

В загальному алгоритм представляє собою ітераційну процедуру:

Крок 1. Проініціалізувати початкову матрицю розбиття U випадковим чином і обрати точність δ, що буде

використовуватися для завершення алгоритму, встановити номер ітерації $l = 0$;

Крок 2. Визначити центри кластерів за наступною формулою:

$$c_l^{(i)} = \frac{\sum_{j=1}^{d} u_{ij} m_j}{\sum_{j=1}^{d} u_{ij}}, 1 \le i \le c \qquad (6.1)$$

де $c_l^{(i)}$ – центри кластерів, d – вимір об'єкта, що досліджується, c – кількість кластерів, l – номер поточної ітерації, u – матриця розбиття, m – об'єкт, що досліджується;

Крок 3. Оновити матрицю розбиття:

$$u_{ij}^{(l)} = \begin{cases} 1, & \text{при } d(m_j, c_i) = \min\limits_{l \le k \le c} d(m_j, c_k) \\ 0, & \text{в інших виппадках} \end{cases} \qquad (6.2)$$

де $u_{ij}^{(l)}$ – елемент матриці розбиття, $d(x,y)$ – обрана метрика, c – кластер, m – об'єкт, що досліджується;

Крок 4. Перевірити умову $\left\| U^{(l)} - U^{(l-1)} \right\| < \delta$, де U –матриця розбиття, і δ – обрана точність. Якщо умова виконується – завершити процес, якщо ні – перейти до кроку 2 з номером ітерації $l = l + 1$.

Також існує альтернативний варіант даного алгоритму:

Крок 1. Випадковим чином обрати центри кластерів із елементів вхідних даних і обрати точність δ, що буде використовуватися для завершення алгоритму, встановити номер ітерації $l = 0$;

Крок 2. Оновити матрицю розбиття(формула 6.2);

Крок 3. Визначити центри кластерів(формула 6.1);

Крок 4. Перевірити наскільки змістилися центри кластерів. Якщо вони змістилися менше ніж на точність δ – завершити процес, якщо ні – перейти до кроку 2 з номером ітерації $l = l + 1$.

Також існує набір обмежень: $u_{ij}^{(l)} \{0,1\}$; $\sum_{j=1}^{c} u_{ij} = 1; 0 < \sum_{j=1}^{d} u_{ij} < d$, який визначає, що кожен вектор даних може належати тільки одному кластеру і не належати іншим. В кожному кластері повинно бути не менше одного елемента даних і не більше загальної кількості елементів.

Основний недолік даного алгоритму через дискретність елементів матриці розбиття – великий розмір просторового розбиття.

Одним із способів усунення даного недоліку є представлення елементів матриці розбиття числами із інтервалу від 0 до 1. Тобто належність елемента даних до кластера визначається функцією належності – елемент даних може належати декільком кластерам з різною степеню належності. Даний підхід реалізований в алгоритмі нечіткої кластеризації с-середніх(Fuzzy C-Means).

6.4.2 Алгоритм с-середніх

Алгоритм с-середніх(англ. fuzzy c-means)відрізняється від попереднього тим, що кластери тепер являються нечіткими множинами і кожний елемент даних належить до різних кластерів з різною степеню приналежності.

- Крок 1. Вибрати кількість кластерів $2 \le c \le d$, і обрати точність δ, що буде використовуватися для завершення алгоритму, встановити номер ітерації $l = 0$;
- Крок 2. Вибрати коефіцієнт нечіткості $w = 2$;
- Крок 3. Проініціалізувати початкову матрицю розбиття U випадковим чином;
- Крок 4. Визначити центри кластерів за наступною формулою:

$$c_l^{(i)} = \frac{\sum_{j=1}^{d}(u_{ij})^w m_j}{\sum_{j=1}^{d}(u_{ij})^w}, 1 \le i \le c \qquad (1.15)$$

де $c_l^{(i)}$ – центри кластерів, d – вимір об'єкта, що досліджується, c – кількість кластерів, l – номер поточної ітерації, u – матриця розбиття, m – об'єкт, що досліджується, w – коефіцієнт нечіткості;

- Крок 5. Для всіх елементів даних обчислити квадрати відстаней до всіх центрів кластерів:

$$d_A^2\left(m_j, c_l^{(i)}\right) = \left(c_l^{(i)} - m_j\right)^i A\left(c_l^{(i)} - m_j\right); \qquad (1.16)$$

- Крок 6. Оновити матрицю розбиття:

$$u_{ij}^{(l)} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_A^2\left(m_j, c^{(l)}\right)}{d_A^2\left(m_j, c^{(k)}\right)}\right)^{\frac{1}{w-1}}}, \text{для } 1 \le i \le c, 1 \le j \le d \quad (1.17)$$

де $u_{ij}^{(l)}$ – елемент матриці розбиття, *d(x,y)* – обрана метрика, *c* – кластер, *m* – об'єкт, що досліджується, *w* – коефіцієнт нечіткості;

– Крок 7. Перевірити умову $\left\|U^{(l)} - U^{(l-1)}\right\| < \delta$. Якщо умова виконується – завершити процес, якщо ні – перейти до кроку 4 з номером ітерації $l = l + 1$.

Також існує набір обмежень: $u_{ij}^{(l)}$ $[0,1]$; $\sum_{j=1}^{c} u_{ij} = 1; 0 < \sum_{j=1}^{d} u_{ij} < d$, який визначає, що кожен вектор даних може належати різним кластерам з різною степеню приналежності, сума степенів приналежності елементів даних до різних кластерів повинна бути рівною одиниці. В кожному кластері повинно бути не менше одного елемента даних і не більше загальної кількості елементів.

Основі недоліки даних алгоритмів:

- припущення, що кластери мають певну сталу форму, не завжди правдиве, тому результат кластеризації для даних які мають складне взаємне розташування можу бути взагалі не точним і його не можливо буде інтерпретувати;
- припущення, що кластери мають центр – певну точку, яка має рівень приналежності до певного кластера рівним одиниці, а інші точки не можуть належати до кластера з таким же самим рівнем приналежності, веде до того, що результат кластеризації для даних які мають складне взаємне розташування може бути не прийнятним;
- дані алгоритми засновані не на основі взаємного розміщення точок один від одного, а тільки на розміщенні точок відносно центра кластерів.

Питання для самоконтролю

1. Яка мета задачі класифікації?
2. Для чого потрібна навчальна множина?
3. Для чого потрібна тестова множина?
4. Які найбільш поширені методи розв'язання задачі класифікації?
5. Поясніть метод класифікації з використанням лінійної регресії.
6. Поясніть метод дерев рішень.
7. Поясніть метод опорних векторів.
8. В чому різниці між задачею класифікації та кластерізації?
9. Які методи кластерізації відносяться до ієрархічних?
10. Які методи кластерізації відносяться до неієрархічних?
11. Яка різниця між чіткими та нечіткими методами кластерізації?
12. Поясніть метод k-середніх.
13. Поясніть метод c-середніх.

Самостійна робота №6
Тема: Задачі класифікації й кластеризації

Завдання для виконання

1. Для даних з самостійної роботи №2, 5 розв'язати задачу регресії або класифікації.
2. Зробити власні висновки.
3. Результати оформити у вигляді тез на наукову студентську конференцію.

Практична (лабораторна) робота №6
Тема: Задачі класифікації й кластеризації

Модель класифікаційного дерева використовує підхід, аналогічний тому, який ми застосовували при створенні моделі

регресійного аналізу, а саме створення моделі на основі навчальної послідовності (training set). Цей підхід використовує відомі дані для аналізу впливу вхідних значень на результат і будує модель на основі отриманих залежностей. Таким чином, коли у нас є новий набір даних з невідомим результатом, ми підставляємо наші дані в модель і отримуємо очікуваний висновок.

Метод класифікаційного аналізу використовує підхід, сутність якого в тому, що набір даних поділяється на 2-і частини. **Перша частина** містить 60-80% всіх даних і використовується для навчального набору і формування моделі. Після цього етапу дані, що залишилися «проганяються» через сформовану модель, і результати порівнюються із реальними даними. Даний підхід дозволяє зробити оцінку точності побудованої аналітичної моделі.

Додатковий крок перевірки моделі дозволяє уникнути *зайвої «підгонки»* аналітичної моделі під дані, що зібрані. У випадку використання *великих масивів* даних для розробки моделі, побудована модель буде ідеально відповідати зібраними даними. Завданням, що ставиться при цьому є створення моделі для прогнозування невідомих даних, а не для їх прогнозу. Для перевірки цього критерію використовується тестовий набір даних. За його допомогою визначається точність моделі. Отже, можемо гарантувати, що модель здатна з високою ймовірністю визначити невідомі значення. За допомогою інструментарію **WEKA** розглянемо дану технологію на конкретному прикладі.

Питання надмірної кількості даних приводить нас до обговорення ще одного важливого принципу побудови класифікаційних дерев – принципу відсікання гілок. «*Відсікання гілок*», як випливає з назви, має на увазі видалення гілок з класифікаційного дерева. Але навіщо видаляти інформацію з дерева рішень? Для того, щоб уникнути зайвої підгонки дерева під відомі дані. Із зростанням даних зростає кількість атрибутів, таким чином, що дерево стає надмірно складним і

розгалуженим. Отже, теоретично, можна створити дерево з кількістю листків leaves = (rows * attributes) , але наскільки корисне таке дерево? Воно навряд чи зможе допомогти визначити невідомий результат, так як воно всього лише визначає вже відомі дані. Необхіднім є досягнення балансу між точністю аналізу та простотою моделі. Стає потрібною аналітична модель з мінімальною кількістю вузлів та листків, яка, при цьому, відрізняється високою точністю. Тобто необхідним є дотримання певного компромісу.

Ще одне питання, на якому ми хотіли б зупинитися, перш ніж приступити до створення конкретних моделей в WEKA, це **проблема помилкового розпізнавання** (false positive і false negative). Головним сенсом проблеми помилкового розпізнавання є те, що модель розглядає будь-який атрибут, що має вплив на кінцевий результат, тоді, коли цей атрибут ніякого впливу на результат не має. І навпаки, істотний атрибут може інтерпретуватися моделлю як несуттєвий.

Таким чином, хибне розпізнавання свідчить про помилковість моделі і невірну класифікацію даних. Зауважимо, що певна похибка в класифікації є припустимою, і завданням розробника моделі є – визначення допустимого відсотку помилок. Так, наприклад, при розробці моделі оцінки медичного обладнання для лікарень будуть потрібні виключно низькі показники помилковості результатів. І, навпаки, у випадку розробки навчальної моделі для ілюстрації статті про інтелектуальний аналіз даних, достатньо високий рівень похибки моделі є припустимим. Розвиваючи цю тему, корисно визначити прийнятне процентне відношення хибно-негативного розпізнавання до хибно-позитивного. Найбільш очевидним прикладом у цьому випадку є приклад моделі визначення спаму. Хибно-позитивне розпізнавання, коли потрібний лист позначається як спам, скоріше за все, буде мати більше негативних наслідків, ніж хибно-негативні, коли лист-спам потрапить в розряд потрібних листів. У цьому випадку

може вважатися прийнятним співвідношення хибно-негативних розпізнавань до хибно-позитивних як 100 до 1.

Розглянемо реальний набір даних. Так, набір даних, який будемо використовувати для прикладу класифікаційного аналізу, містить інформацію, яка зібрана дилерським центром BMW. Центр починає рекламну кампанію, і пропонує розширену дворічну гарантію своїм постійним клієнтам. Такі кампанії вже проводилися, так що дилерський центр має 4500 показників щодо попередніх продажів із розширеною гарантією. Набір даних має наступні атрибути:

- Розподіл за доходами
  [0=$0-$30k,
  1=$31k-$40k,
  2=$41k-$60k,
  3=$61k-$75k,
  4=$76k-$100k,
  5=$101k-$150k,
  6=$151k-$500k,
  7=$501k+]
- Рік / місяць покупки першого автомобіля BMW
- Рік / місяць покупки останнього автомобіля BMW
- Чи скористався клієнт розширеною гарантією

Файл даних у форматі Attribute-Relation File Format (ARFF) буде виглядати наступним чином:

```
@attribute IncomeBracket {0,1,2,3,4,5,6,7}
@attribute FirstPurchase numeric
@attribute LastPurchase numeric
@attribute responded {1,0}

@data

4,200210,200601,0
5,200301,200601,1
...
```

Класифікація в WEKA

Завантажте файл **bmw-training.arff** в програмний пакет **WEKA**.

Зауваження: *в пропонованому файлі містяться 3000 з наявних 4500 записів. Ми розділили набір даних так, щоб частина їх використовувалася для створення моделі, а частина – для перевірки її точності, щоб переконатися, що модель не є підігнаною під конкретний набір даних.*

Після завантаження даних вікно WEKA має виглядати так, як показано на рис. 6.6.

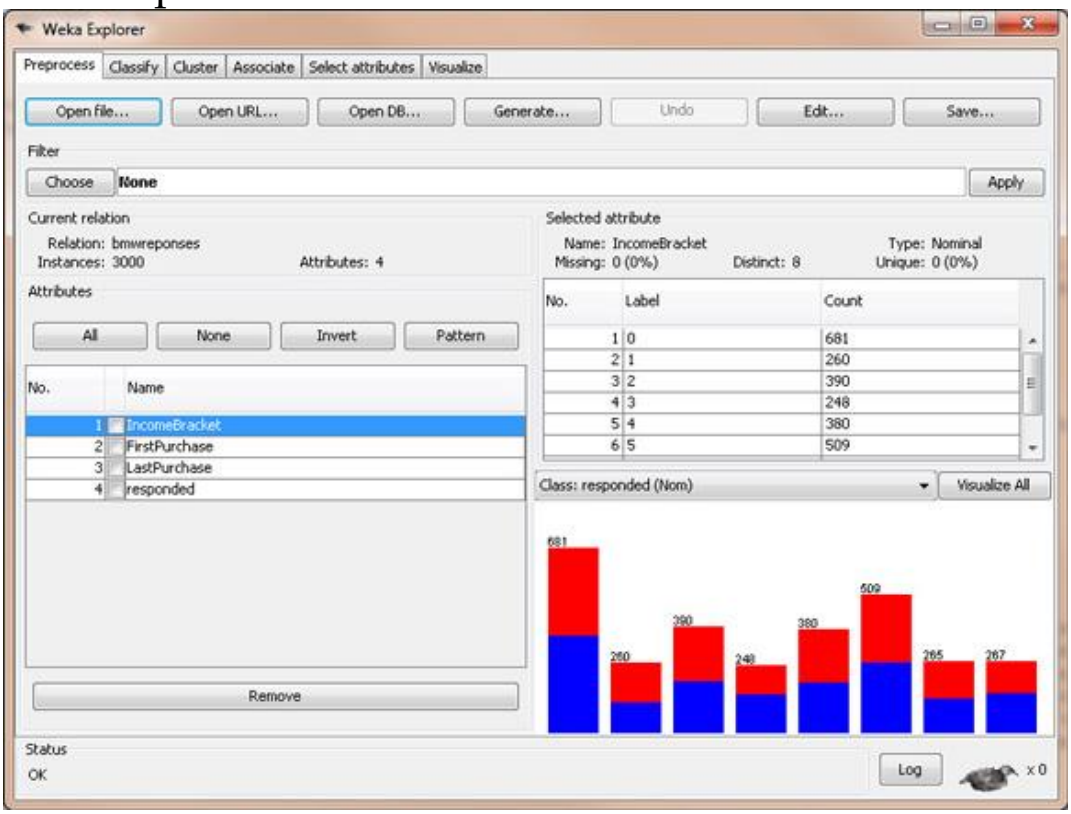


Рис. 6.6 Дані дилерського центру BMW

Аналогічно тому, як ми вибирали тип моделі для регресійного аналізу, тепер нам слід вибрати модель для класифікації: відкрийте закладку **Classify** , виберіть опцію **trees**, а потім опцію **J48**.

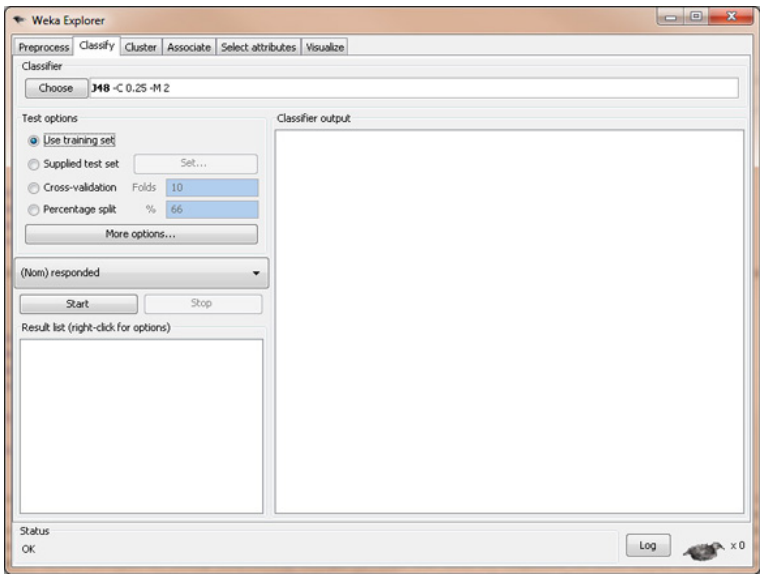


Рис. 6.7 Алгоритм класифікації даних BMW

Тепер ми готові приступити до створення конкретної моделі аналізу засобами пакету WEKA. Переконайтеся, що обрана опція **Use training set** , щоб пакет WEKA при створенні моделі використовував саме ті дані, які ми тільки що завантажили у вигляді файлу. Натисніть кнопку **Start** і надайте WEKA можливість попрацювати з нашими даними. Результуюча модель повинна виглядати так, як показано нижче:

Результат роботи класифікаційної моделі WEKA

Number of Leaves : 28
Size of the tree : 43
Time taken to build model: 0.18 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances	1774	59.1333 %
Incorrectly Classified Instances	1226	40.8667 %
Kappa statistic	0.1807	
Mean absolute error	0.4773	
Root mean squared error	0.4885	
Relative absolute error	95.4768 %	
Root relative squared error	97.7122 %	

Total Number of Instances 3000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.662	0.481	0.587	0.662	0.622	0.616	1
	0.519	0.338	0.597	0.519	0.555	0.616	0
Weighted Avg.	0.591	0.411	0.592	0.591	0.589	0.616	

=== Confusion Matrix ===

a	b	<-- classified as
1009	516	a = 1
710	765	b = 0

Що означають всі ці числа? Найбільш суттєві дані - це показники класифікації "**Correctly Classified Instances**" (59.1%) і "**Incorrectly Classified Instances**" (40.9%). Крім того, слід звернути увагу на число в першому рядку стовпця **ROC Area** (0.616). Трохи пізніше ми докладніше обговоримо ці значення, поки ж просто запам'ятайте їх. Нарешті, таблиця **Confusion Matrix** показує кількість **хибно-позитивних** (516) і **хибно-негативних** (710) розпізнавань.

Як зрозуміти, наскільки хороша отримана модель? Оскільки показник точності нашої моделі - 59,1%, то в первісному розгляді її не можна назвати досить хорошою.

Де це так зване дерево? Ви зможете побачити дерево, якщо клацнете правою кнопкою миші в панелі результуючої моделі. У контекстному меню виберіть опцію **Visualize tree** . На екрані відобразиться візуальне уявлення класифікаційного дерева нашої моделі (рис. 3), проте в даному випадку картинка мало чим нам допоможе. Ще один спосіб побачити дерево моделі – прокрутити вгору висновок у вікні **Classifier Output**, там ви знайдете текстовий опис дерева з вузлами і листками.

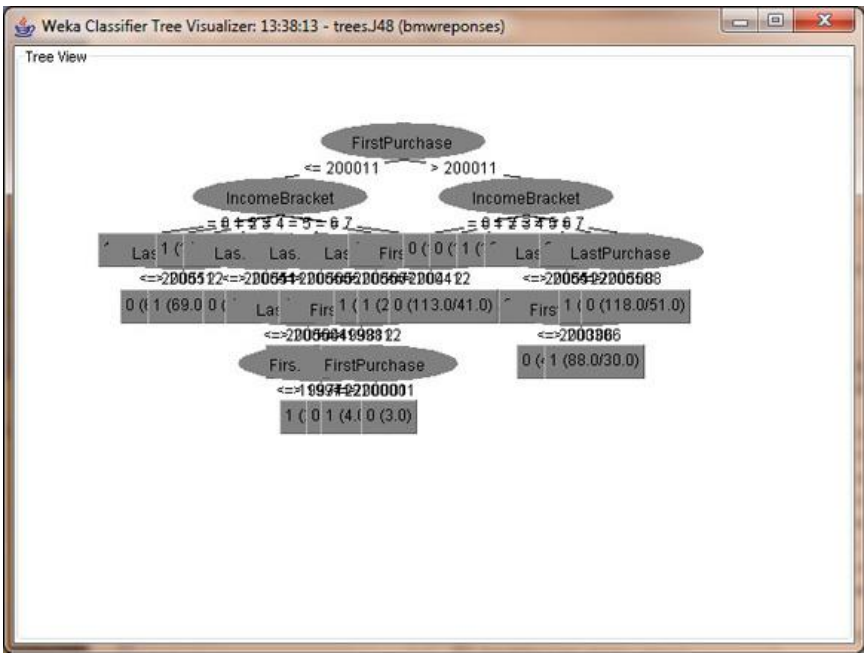


Рис. 6.8 Візуальне подання дерева класифікації

Залишився останній етап перевірки класифікаційного дерева: нам треба пропустити набір даних, що залишився через отриману модель і перевірити, наскільки результати класифікації будуть відрізнятися від реальних даних. Для цього в секції **Test options** виберіть опцію **Supplied test set** і натисніть на кнопку **Set**. Вкажіть файл **bmw-test.arff**, що містить решту 1500 даних, які не були включені в навчальний набір. При натисканні на кнопку **Start** WEKA пропустить тестові дані через модель і покаже результат роботи моделі. Давайте натиснемо на **Start** і перевіримо, що у нас вийшло.

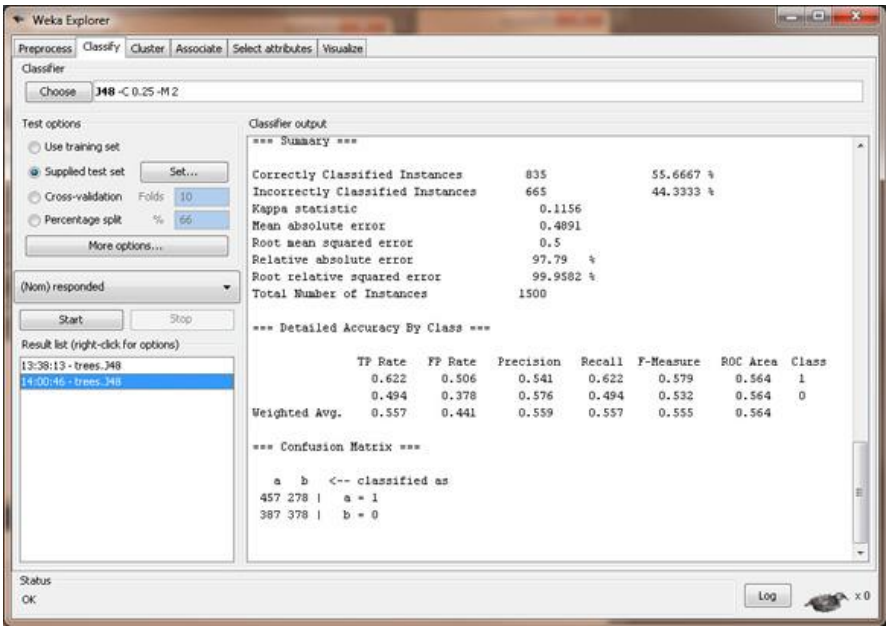


Рис 6.9 Перевірка класифікаційного дерева

Порівнюючи показник **Correctly Classified Instances** для тестового набору (55,7%) з цим же показником для навчального набору (59,1%), ми бачимо, що точність моделі для двох різних наборів даних приблизно однакова. Це означає, що нові дані, які будуть використовуватися в цій моделі в майбутньому, не знизять точність її роботи.

Однак, оскільки власне точність моделі досить низька (всього лише 60% даних класифіковано вірно), ми маємо повне право зупинитися і сказати: «На жаль, ця модель взагалі нікуди не годиться.

Отже, завжди необхідно пам'ятати, що **для того, щоб витягти корисну інформацію з великого набору даних, вам слід вибрати відповідну модель.**

Тест до теми 6

1. Впорядкована за деяким принципом множина об'єктів, які мають подібні класифікаційні ознаки (одна або кілька властивостей), обраних для визначення подібності або відмінності між цими об'єктами – це

   а) кластеризація;

б) класифікація;
в) асоціація.

2. Розбиття сукупності об'єктів на однорідні групи
   а) кластеризація;
   б) класифікація;
   в) асоціація.
3. Алгоритми, які кожному об'єкту вибірки ставлять у відповідність номер кластера, тобто кожен об'єкт належить тільки одному кластеру
   а) чіткі;
   б) ієрархічні;
   в) неієрархічні.
4. Алгоритми, які будують не одне розбиття вибірки на непересічні кластери, а систему вкладеного розбиття
   а) чіткі;
   б) ієрархічні;
   в) неієрархічні.
5. Алгоритми, які кожному об'єкту ставлять у відповідність набір значень, що показує ступінь відношення об'єкта до кластерів.
   а) чіткі;
   б) ієрархічні;
   в) неієрархічні.
6. Алгоритм с-середніх належить до..
   а) чітких;
   б) ієрархічних;
   в) неієрархічних.
7. Метод дерев рішень належить до..
   а) чітких;
   б) ієрархічних;
   в) неієрархічних.

8. Множина, яке включає дані, що використовуються для навчання (конструювання) моделі..
   г) чітка;
   д) навчальна;
   е) тестова.
9. Множина, що містить вхідні та вихідні значення прикладів. та використовуються для перевірки працездатності моделі..
   а) чітка;
   б) навчальна;
   в) тестова.

Рекомендована література

1. Борисенко О.А. Диференціальна геометрія і топологія. – Х.: Основа, 1995. – 304 с.

2. Бояришева Т.В., Гудивок Т.В., Погоріляк О.О. Функціональний аналіз. Навчальний посібник для студентів спеціальностей «математика», «прикладна математика», «статистика». – Ужгород, 2013. – 125 с.

3. Журавлёв Ю.И. Распознавание. Математические методы. Программная система. Практические применения. / Журавлёв Ю.И., Рязанов В.В., Сенько О.В. — М.: Изд. «Фазис», 2006. — 176 с.

4. Зиновьев А. Ю. Визуализация многомерных данных. / Зиновьев А. Ю. — Красноярск: Изд. Красноярского государственного технического университета, 2000. — 180 с.

5. Ильин В.А., Позняк Э.Г. Линейная алгебра: Учеб. для вузов. – М.: Наука, 1999. – 296 с.

6. Олійник А. О. Еволюційні обчислення та програмування: Навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2010. – 324 с.

7. Олійник А. О. Інтелектуальний аналіз даних : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – 278 с.

8. Руденко О. Г. Штучні нейронні мережі / О. Г. Руденко, Є. В. Бодянський. – Харків : Компанія СМІТ, 2006. – 404 с.

9. Ус. С.А. Функціональний аналіз [Текст]: навч. посібник / С.А. Ус. – Д. : Національний гірничий університет, 2013. – 236с.

10. Чубукова И. А. Data Mining: учебное пособие //М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний. – 2006. – 368с.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. – СПб.: БХВ–Петербург, 2009. – 512 с.
2. Бояришева Т.В., Гудивок Т.В., Погоріляк О.О. Функціональний аналіз. Навчальний посібник для студентів спеціальностей «математика», «прикладна математика», «статистика». – Ужгород, 2013. – 125 с.
3. Елманова Н. Введение в OLAP-технологии Microsoft / Наталия Елманова, Алексей Федоров – М.: Диалог–МИФИ, 2002. – 272 с.
4. Журавлёв Ю.И. РАСПОЗНАВАНИЕ. Математические методы. Программная система. Практические применения. / Журавлёв Ю.И., Рязанов В.В., Сенько О.В. — М.: Изд. «Фазис», 2006. — 176 с.
5. Замятин А.В. Интеллектуальный анализ данных: учеб. пособие. – Томск: Издательский Дом Томского государственного университета, 2016. – 120 с.
6. Зиновьев А. Ю. Визуализация многомерных данных. / Зиновьев А. Ю. — Красноярск: Изд. Красноярского государственного технического университета, 2000. — 180 с.
7. Ковальчук А.М. Основи проектування та розробки інформаційних систем: Збірка навчальних матеріалів./ Ковальчук А.М., Левицький В.Г., Самолюк І.І., Янчук В.М.– Ж.: ЖДТУ, 2009. – 54с.
8. Лепский А.Е., Броневич А.Г. Математические методы распознавания образов: Курс лекций. – Таганрог: Изд-во ТТИ ЮФУ, 2009. – 155 с.
9. Луценко Е. В. Интеллектуальные информационные системы. – Краснодар: КубГАУ, 2004. – 633 с
10. Олійник А. О. Еволюційні обчислення та програмування: Навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2010. – 324 с.

11. Олійник А. О. Інтелектуальний аналіз даних : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – 278 с.
12. Паклин Н. Бизнес-аналитика. От данных к знаниям / Н. Паклин, В. Орешков – СПб: Питер, 2013. – 704 с.
13. Путятін Є.П. Методи та алгоритми комп'ютерного зору: навч. посіб. — Х.: ТОВ «Компанія СМІТ», 2006. — 236 с.
14. Руденко О. Г. Штучні нейронні мережі / О. Г. Руденко, Є. В. Бодянський. – Харків : Компанія СМІТ, 2006. – 404 с.
15. Скобцов Ю. А. Основы эволюционных вычислений /Ю. А. Скобцов. – Донецк : ДонНТУ, 2008. – 330 с.
16. Ус. С.А. Функціональний аналіз [Текст]: навч. посібник / С.А. Ус. – Д. : Національний гірничий університет, 2013. – 236 с.
17. Чубукова И. А. Data Mining: учебное пособие //М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний. – 2006. – 368с.
18. Жалдак М.І., Кузьміна Н.М., Берлінська С.Ю. Теорія ймовірностей і математична статистика з елементами інформаційної технології.–К.: Вища школа,1995.–352с.
19. Жлуктенко В.І., Наконечний С.І. Теорія ймовірностей і математична статистика: Навч.-метод. посібник. У 2 ч. – Ч.1. Теорія ймовірностей. – К.: КНЕУ, 2000. – 304 с.
20. Борисенко О.А. Диференціальна геометрія і топологія. – Х.: Основа, 1995. – 304 с.
21. Ильин В.А., Позняк Э.Г. Линейная алгебра: Учеб. для вузов. – М.: Наука, 1999. – 296 с.
22. Advanced Machine Coursera [Електронний ресурс] Режим доступу: Learninghttps://www.coursera.org/specializations/aml
23. Big Data Coursera [Електронний ресурс] Режим доступу: https://www.coursera.org/specializations/big-data
24. Big Data Fundamentals [Електронний ресурс] Режим доступу: https://www.edx.org/course/big-data-fundamentals-adelaidex-bigdatax

25. Hadoop Starter Kit [Електронний ресурс] Режим доступу: https://www.udemy.com/hadoopstarterkit/
26. Ian H. Witten. Data Mining: Practical Machine Learning Tools and Techniques / Ian H. Witten, Eibe Frank and Mark A. Hall — [3rd Edition] — Morgan Kaufmann, 2011. — P. 664. — ISBN9780123748560
27. Introduction to Big Data [Електронний ресурс] Режим доступу: https://www.edx.org/course/introduction-to-big-data-2
28. Machine Learning Coursera [Електронний ресурс] Режим доступу: https://www.coursera.org/learn/machine-learning
29. Machine Learning Fundamentals [Електронний ресурс] Режим доступу: https://www.edx.org/course/machine-learning-fundamentals

ГЛОСАРІЙ

Атрибут – властивість, що характеризує об'єкт.

Біоінформатика – напрямок, метою якого є розробка алгоритмів для аналізу і систематизації генетичної інформації. Отримані алгоритми використовуються для визначення структур макромолекул, а також їх функцій, з метою пояснення різних біологічних явищ.

Відносна шкала (ratio scale) – шкала, в якій є певна точка відліку (існує абсолютний нуль) і можливе відношення між значеннями шкали.

Вибірка (sample) – частина генеральної сукупності, певним способом відібрана з метою дослідження і отримання висновків про властивості та характеристики генеральної сукупності.

Вимірювання – процес присвоєння характеристикам досліджуваних об'єктів згідно з визначеним правилом числових значень. У процесі підготовки даних вимірюється не сам об'єкт, а його характеристики.

Генеральна сукупність (population) – вся сукупність досліджуваних об'єктів, яка цікавить дослідника.

Гіпотеза – припущення щодо параметрів сукупності об'єктів, яке повинно бути перевірено на її частині.

Дискретні дані є значеннями ознак, загальне число яких скінчено або нескінченно, але може бути підраховано за допомогою натуральних чисел від одного до нескінченності.

Дисперсія – це середнє арифметичне квадратів відхилень значень від їх середнього. Дана статистична величина є важливою характеристикою розсіяння варіаційного ряду. Дисперсія $\boldsymbol{\sigma^2}$ визначається за формулою:

$$\sigma^2 = \frac{\sum (x_i - x_{cp})^2}{n-1}$$

Дихотомічна шкала (dichotomous scale) – шкала, яка містить тільки дві категорії.

Змінна (variable) – властивість або характеристика, загальна для всіх досліджуваних об'єктів, прояв якої може змінюватися від об'єкта до об'єкта.

Інтелектуальний аналіз даних (Data Mining) – це мультидисциплінарна область, що виникла і розвивається на базі таких наук як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних.

Інтервальна шкала (interval scale) – шкала, різниці між значеннями якої можуть бути обчислені, проте їхні відношення не мають сенсу.

Максимум – найбільше значення вибірки.

Масштабованість – властивість обчислювальної системи, що забезпечує передбачуваний ріст системних характеристик, наприклад, швидкості реакції, загальної продуктивності та ін. при додаванні до неї обчислювальних ресурсів.

Медіана – точна середина вибірки, яка ділить її на дві рівні частини по числу спостережень.

Мінімум – найменше значення вибірки.

Множинна кореляція. Множинною кореляцією називається кореляційний зв'язок між одним результативним і декількома факторними ознаками.

Неперервні дані – дані, значення яких можуть брати яке завгодно значення в деякому інтервалі.

Номінальна шкала (nominal scale) – шкала, яка містить тільки категорії; дані в ній не можуть упорядковуватися, з ними не можуть бути зроблені ніякі арифметичні дії.

Об'єкт описується як набір атрибутів. Об'єкт також відомий як запис, випадок, приклад, рядок таблиці і т.д.

Описова статистика (Descriptive statistics) – техніка збору і підсумовування кількісних даних, яка використовується для перетворення маси цифрових даних в форму, зручну для сприйняття і обговорення.

Параметри – числові характеристики генеральної сукупності.

Парна кореляція – це зв'язок між двома ознаками: результативною і факторною або двома факторними.

Порядкова шкала (ordinal scale) – шкала, в якій об’єктам присвоюються числа для позначення відносної позиції об'єктів, але не величини відмінностей між ними.

Потужністю скінченої множини A називається кількість елементів цієї множини та позначається |A|.

Розмах – різниця між найбільшим і найменшим значеннями вибірки.

Стандартне відхилення – квадратний корінь з дисперсії вибірки – міра того, наскільки широко розкидані точки даних відносно свого середнього. Середнє квадратичне відхилення розраховується за наступною формулою:

$$\sigma = \pm\sqrt{\frac{\sum (x_i - x_{cp})^2}{n-1}},$$

де x_i – вимірювана ознака;

x_{cp} – середня арифметична ознака для даної групи;

n – кількість вимірювань.

Статистика – це наука про методи збору даних, їх обробку і аналіз для виявлення закономірностей, властивих досліджуваному явищу.

Шкала – правило, згідно з яким об'єктам присвоюються числа.

Штучний інтелект – науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, що традиційно вважаються інтелектуальними.