Центральноукраїнського державного університету
імені Володимира Винниченка

Інтелектуальний аналіз даних Data Mining

Навчально-методичний посібник
Укладач І.В.Лупан

Кропивницький
2022

УДК 004.021:004.6:519 (075.8)
Л 85

Лупан І.В.
Інтелектуальний аналіз даних Data Mining: навчально-методичний посібник. – Кропивницький, ФОП Піскова М. А., 2022. – 112 с.

У посібнику наведено завдання до лабораторних робіт з курсу «Інтелектуальний аналіз даних» для бакалаврів спеціальності 122.Комп'ютерні науки. Представлено теоретичні відомості та завдання до тем «Вступ. Задачі аналізу даних», «Пошук асоціативних правил», «Кластерний аналіз», «Дискримінантний аналіз», «Аналіз зв'язків», тощо. До лабораторних робіт додаються вправи і контрольні запитання та тестові завдання, розглянуто базовий інструментарій аналізу даних у пакетах Rapid Miner та KNIME.

Матеріали посібника будуть корисними також для бакалаврів та магістрів спеціальності «Статистика».

Рецензенти:

Підгорна Т.В. *– доктор педагогічних наук, доцент, професор кафедри комп'ютерних наук та інформаційних систем Державного торговельно-економічного університету;*

Акбаш К.С. *– кандидат фізико-математичних наук, доцент кафедри математики та методики її викладання Центральноукраїнського державного університету імені Володимира Винниченка.*

Гаєвський М.В. *– кандидат фізико-математичних наук, консультант департаменту розширеної аналітики (АА) компанії RBC Group.*

Рекомендовано методичною радою Центральноукраїнського державного університету імені Володимира Винниченка (протокол №2 від 7.12.2022)

Зміст

Присвячується моїм колегам, вчителям і просто гарним друзям ***Ользі Валентинівні Авраменко*** *та* ***Степану Дмитровичу Паращуку****.*

Дякую Вам щиро за те, що вели за собою і продовжуєте надихати та підтримувати.

Бажаю Вам міцного здоров'я, нових творчих планів та звершень, простого людського щастя

ТЕМА 1. ВСТУП. ЗАДАЧІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

1.1. Інформація та дані

Основним притиріччям сучасності є помітний надлишок даних при дефіциті інформації та знань.

Під інформацією *(лат. Informatio* – роз'яснення, виклад*)* розуміють «повідомлення» у будь-якій формі; відомості, які є об'єктом зберігання, переробки та передавання.

Іншими поясненнями цього поняття є такі:

- Абстрактне значення виразів, висловлень, графічних зображень тощо (М.Брой[1]).
- Кількісна міра усунення невизначеності (ентропія), міра організації системи (К.Шеннон).
- Будь-які раніше невідомі відомості (знання) про подію (сутність або процес), які є об'єктами операцій, для яких існує змістова інтерпретація.
- «Під інформацією розуміють документовані або публічно оголошені відомості про події та явища, що відбуваються у суспільстві, державі та навколишньому природному середовищі»[2].
- «Інформація – це відомості, що передаються усно, письмово чи якимось іншим способом, *за допомогою умовних сигналів* з використанням технічних засобів також, та сам процес передавання чи здобування цих відомостей[3]».
- У кібернетиці: «*Інформація* – це будь-яка сукупність сигналів, впливів або відомостей, які деяка система сприймає від оточуючого середовища (вхідна інформація), видає в оточуюче середовище (вихідна інформація) або зберігає в собі (внутрішня інформація)» (А.О.Дородніцин[4]).

Під *даними* розуміють представлення деяких фактів у формалізованому вигляді, придатному для зберігання, обробки та передавання, тобто дані – це *зареєстровані* сигнали будь-якої природи, використані для передавання змісту повідомлень. Це факти, тексти, графіки, картинки, звуки, аналогові та цифрові відео-матеріали, представлені у формі, придатній для зберігання, передавання та обробки.

Підсумовуючи, слід зауважити, що *інформація є продуктом взаємодії даних та адекватних методів інтерпретації*.

Інтерпретування, в свою чергу, – це перехід від подання (зовнішньої форми) до значення (абстрактного змісту) інформації. Близьким за значенням, але більш глибоким за сутністю є *розуміння* – встановлення зв'язку між інформацією, поданою у деякій зовнішній формі, та реальним світом.

---

[1] Брой М. Информатика, 1996.
[2] Стаття 1 Закону України «Про інформацію».
[3] Большой Энциклопедический Словарь. – М., 1997, с. 455.
[4] Кибернетика. Становление информатики, 1986.

На відміну від даних інформація має деякий контекст. Інформація – *це спосіб та система обробки даних.* За Б.Лангефорсом[5], інформація – це особлива форма знання, придатна для транспортування у суспільстві:

$$I = i(D, S, t), \qquad (1.1)$$

де D – сукупність даних, які передають повну інформацію;
I – інформація, виражена в даних (D);
S – сприймаюча структура, що включає систему інтерпретації даних;
t – час, потрібний реципієнту (приймачу) для інтерпретації та розуміння;
i – інформаційна функція.

Тобто інформація «*не просто знання, а змінення знань, розширення уявлень спостерігача відносно цілей та шляхів їх досягнення»* (Ю.М.Канигін[6]).

За Бертрамом Бруксом[7] також інформація – це розширення знань:

$$K(S) + \Delta I = K(S + \Delta S), \qquad (1.2)$$

де S – структура знань;
K(S) – наявний обсяг знань;
ΔI – інформація;
ΔS – деформація структури знань, спричинена інформацією.

Отже, під *знаннями* розуміють певним чином структуровану сукупність інформації про світ, властивості об'єктів, закономірності процесів і явищ, а також правила їх використання для прийняття рішень. Тобто знання представляють собою 1) сукупність понять, фактів, закономірностей та евристичних правил, організовану у деяку структуру за допомогою підходящих відношень; 2) продукт інформаційної діяльності, реалізований як система суджень або тверджень стосовно об'єктів, процесів або явищ.

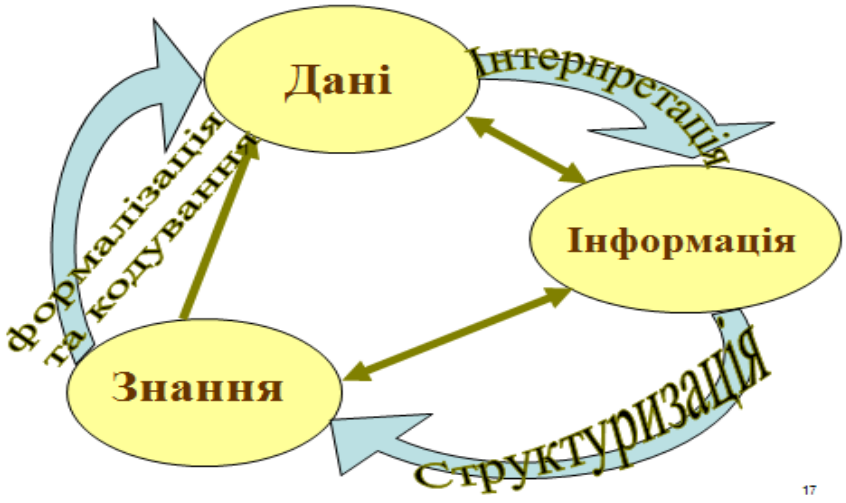


Рис. 1.1

Тобто поняття дані, інформація та знання тісно пов'язані між собою, але не тотожні (Рис. 1.1).

[5] Börje Langefors – швецький інженер та науковець, професор Стокгольмського університету.
[6] Кибернетика. Становление информатики, 1986.
[7] Brookes, B. C. (1980). The foundations of information science. Part I. Philosophical aspects. Journal of Information Science, 2(3-4), 125-133 pp.

Дані перетворюються в *інформацію* за допомогою категоризації, калькуляції, контекстуалізації, коректування та конденсації.

Інформація перетворюється в *знання* за допомогою порівняння, наслідку, зв'язку, судження та усвідомлення, що включає збір інформації, аналіз, синтез, обмін та використання.

Таблиця 1.1

Інформація	***Знання***
Генерується людьми, комп'ютерами, тощо	Генеруються тільки людьми
Завжди пов'язана з даними	Мають відношення до даних та інформації, але не завжди з ними пов'язані
Знаходиться скрізь	Дефіцитні
Може залежати від контексту, а може не залежати	Завжди пов'язані з контекстом
Легко сприймається і передається	Важкі
Частіше статична	Динамічні
Може бути легко взаємопов'язана	Вимагають меж розуміння
Має вартість створення і підтримки	Дуже дорогі і ціна не фіксована
Може використовуватися ким завгодно і коли завгодно	Мають термін використання і цільове призначення

Основними властивостями інформації називають:

- *Повноту* – достатність для прийняття рішення;
- *Достовірність* – відповідність реальному стану справ, яка знижується наявністю інформаційного шуму, недосконалими методами передавання та інтерпретації, інформаційними втратами та спотвореннями при зберіганні та обробці, тощо;
- *Цінність* та *корисність* – для конкретної категорії користувачів, що оцінюється за тим, наприклад, наскільки інформація підвищує імовірність досягнення поставлених цілей, або за матеріальним ефектом від використання інформації;
- *Адекватність* = повнота+достовірність;
- *Актуальність* – відповідність теперішньому моменту часу;
- *Об'єктивність* – незалежність від джерела та приймача;
- *Доступність* – складається з доступності даних та доступності методів;
- *Надлишковість* – фактично надлишковість даних – допускає зменшення обсягу повідомлень при збереженні їхнього абстрактного значення.

Знання можна поділити на такі типи:

1) причини, цілі (бачення), що дають відповідь на питання «чому?», а також дають підстави для структурування проблем та прагнення до досягнення успіху;

2) предмет знання (факти, концепції, теорії, конструкції), що дають відповідь на питання «що?»;

3) алгоритми (процедури, методи, ноу-хау, технології, уміння виконати на практиці), що дають відповідь на питання «як зробити?»;

4) альтернативи (варіанти, нюанси), що дають відповідь на питання «хто?», «де?», «коли?», «в яких умовах?».[8]

Для знань важливими властивостями є:

- Структурованість;
- Зручність для доступу та засвоєння;
- Лаконічність;
- Несуперечливість;
- Наявність процедур обробки знань.

Явні знання (кодифіковані та формалізовані) виражаються у словах, цифрах, знаках, формулах, схемах, образах і т.д. Їх легко передавати та поширювати, вони належать усьому людству та впливають на продуктивну діяльність.

Люди у процесі мислення та практичної діяльності в основному оперують *неявними* знаннями, що знаходяться у їхній свідомості.

Таблиця 1.2. Трансформація знань[9]

	В неявні знання	***В явні знання***
З неявних знань	Соціалізація (обмін знаннями)	Екстерналізація (кодифікація знань)
З явних знань	Інтерналізація (навчання)	Комбінація (обробка інформації)

У теорії інформації – науці, заснованій Клодом Шенноном у 1948 році, що вивчає кількісні закономірності, пов’язані з отриманням, передаванням, обробкою та зберіганням інформації, – однією з основних задач є стиснення даних, тобто пошук найбільш ефективних способів кодування інформації (фактично, даних). Тому необхідно вміти кількісно вимірювати обсяг даних, що зберігаються або передаються.

Для цього введено поняття *ентропії*, як *універсальної міри різноманітності системи та невизначеності наших знань про її можливі стани* (К.Шеннон, У.Уівер).

Ентропія – це *міра кількості інформації, що виробляється джерелом, пропускається каналом або потрапляє до користувача.*

За формулою К.Шеннона інформація *зменшує ентропію* (тобто невизначеність) *системи*:

8 Орехов В.Д. Прогнозирование развития человечества с учетом фактора знания. Монография. – Жуковский: МИМ ЛИНК, 2015. – http://world-evolution.ru/monograph/monograph_chapter_1.pdf

9 Там само.

$$H = K\sum_{i=1}^{n} p_i \log p_i \,, \qquad (1.3)$$

де H – ентропія джерела інформації (набору повідомлень);
n – кількість можливих повідомлень;
p_i – імовірність появи повідомлення з номером і;
log – функція логарифма;
K – коефіцієнт пропорційності, який залежить лише від вибору одиниць вимірювання та основи логарифма.

Інформація та ентропія системи пов'язані кількісно та еквівалентні: отримання інформації супроводжується рівносильним зменшенням ентропії і, навпаки, зростання ентропії пов'язане з втратою інформації.

Величина ентропії не залежить від особливостей кодування. Кількості та імовірності повідомлень залежать лише від природи джерела інформації (а не від способу кодування).

Кількість інформації у будь-якому наборі даних – це мінімальна границя, до якої прямує кількість даних при найбільш ефективному способі кодування. Отже фактично тут під *кількістю інформації* розуміють *кількість (обсяг) даних*.

В аналізі даних мають справу з даними, що характеризують властивості об'єктів. У різних галузях для позначення властивостей застосовують терміни «*атрибут*», «*вимір*», «*характеристика*», «*змінна*».

Атрибутам прийнято приписувати значення відповідно до визначеного способу – шкали вимірювання.

За типом розрізняють шкали номінативні, порядкові, інтервальні та абсолютні. Перші дві застосовують до якісних ознак, а решту до кількісних.

Номінативну шкалу (шкалу назв) застосовують до атрибутів, що позначають приналежність об'єкта до деякого класа. Класи зручно нумерувати, але відношення між номерами ніяк не пов'язані з властивостями об'єктів.

Прикладами таких атрибутів можуть бути колір очей, розмір одягу та взуття, стать, клінічні діагнози, автомобільні номери, номери телефонів, типи темперамента, тощо.

Частковим випадком номінативної шкали є дихотомічна (бінарна) шкала. Атрибути, що вимірюються за такою шкалою, можуть приймати одне з двох значень (наприклад, так/ні, істинне/хибне, 0/1, +/– і т.п.). Бінарні дані можуть бути симетричними та асиметричними. Симетричні – у випадку, коли ваги значень однакові (наприклад, для атрибута стать значення чоловіча/жіноча мають однакову вагу). У випадку різної ваги (значимості) найбільш важливому (і водночас найбільш рідкісному) значенню атрибута приписують одиницю, а іншому – нуль. Такий атрибут є асиметричним. Наприклад, реакція на щеплення є асиметричним атрибутом.

У ***порядковій шкалі*** (*ранговій*) номери об’єктів відображують кількісні характеристики властивостей, а саме відношення “більше-менше”. Але однакові різниці між числами не означають однакових різниць у кількостях властивостей.

Прикладами рангових атрибутів є нагороди за заслуги, військові ранги, твердість мінералів, рейтинги, академічні ранги.

Оцінки (*задовільно/добре/відмінно*) або рівні (*низький/середній/високий*), можуть відповідати порядковій шкалі, якщо ці значення мають чіткі границі і не перетинаються. У разі нечітких границь таку шкалу називають ***лінгвістичною*** і тут ми її не розглядатимемо.

Якщо існує одиниця вимірювання, за допомогою якої значення властивості можна не лише впорядкувати, але й приписати їм числа так, щоб однакові різниці між числами відображували однакові відмінності у кількостях вимірюваної ознаки, то така шкала буде ***інтервальною***. Однак нульова точка інтервальної шкали довільна і не вказує на відсутність властивості, тому різницю між значеннями таких атрибутів для різних об’єктів визначити можна, а відношення між ними – ні.

Прикладами інтервальних величин є календарний час, шкали температур за Фаренгейтом та Цельсієм, номери шкільних класів (років навчання) тощо.

У ***шкалі відношень*** (*абсолютній*) числа, присвоєні об’єктам, мають усі властивості об’єктів інтервальної шкали, та крім того на шкалі визначено абсолютний нуль. Значення нуля свідчить про повну відсутність оцінюваної ознаки, тому відношення між числами відображують кількісні відношення між відповідними значеннями атрибута.

Прикладами таких атрибутів є зріст, вага, час, температура за Кельвіном, кількість допущених у тесті помилок, тощо.

Інтервальна та абсолютна шкали є числовими, але значення, залежно від природи атрибута, можуть бути неперервними або дискретними.

Тип шкали накладає обмеження на допустимі перетворення значень атрибутів та статистичні методи, які можна до них застосувати (Таблиця 1.3):

Таблиця 1.3

Тип шкали	**Допустимі статистики**	**Допустимі перетворення**
номінативна	Мода, Хі-квадрат	One to One (equality (=))
порядкова	Медіана, Процентилі	Монотонне зростання (порядок (<))
інтервальна	Середнє, Стандартне відхилення, Кореляція, Регресія, Дисперсійний аналіз	Лінійні перетворення
відношень	Усі статистики інтервальної шкали, а також Середнє геометричне, Середнє гармонійне, Коефіцієнт варіації, логарифмування	Перетворення подібності

1.2. Задачі аналізу даних

Інтелектуальним аналізом даних (*Data Mining*) називають процес визначення нових, коректних та потенційно корисних знань на основі великих масивів даних. «Інтелектуальний аналіз даних» деякі дослідники вважають синонімом ще одного популярного терміна – виявлення знань у даних – "*Knowledge Discovery in Databases"* (KDD)[10], на думку інших – інтелектуальний аналіз даних є лише важливим кроком у процесі виявлення знань[11].

Виявлені в результаті інтелектуального аналізу знання називають *патерном* (зразком). Тобто задача інтелектуального аналізу полягає в ефективному виявленні *осмислених* патернів з наявного масиву даних великого розміру. Отримані знання мають бути *цікавими*.

Ознаками цікавих знань є

- *несподіваність* – отримані знання мають дивувати (бути нетривіальними) та нести нову інформацію;
- *застосовність* – нові знання мають бути придатними до застосування для досягнення поставлених цілей, або до формулювання на їхній основі нових корисних цілей.

Аналіз даних є природним результатом еволюції інформаційних технологій. Розвиток баз даних та технологій керування (менеджменту) потребує також розвитку технологій збору та зберігання даних, керування та аналізу даних, оскільки серйозною проблемою інформаційного суспільства залишається «*суперечність між збільшенням обсягів інформації та зменшенням темпів зростання обсягів істинних знань»*[12].

Етапами інтелектуального аналізу є

1. Вивчення предметної області.
2. Збір даних.
3. Попередня обробка даних.
   а) очищення даних – виключення «шумів» та суперечностей;
   б) інтеграція даних – об'єднання даних з різних джерел в одному сховищі;
   в) перетворення даних до підходящої форми (агрегація, стиснення, скорочення розмірності, дискретизація атрибутів, тощо).
4. Аналіз даних з метою виявлення патернів.
5. Інтерпретація знайдених патернів (візуалізація, відбір корисних патернів відповідно до функції корисності).
6. Використання нових знань.

---

[10] Засновником напряму є Григорій Пятецький-Шапіро (Gregory I. Piatetsky-Shapiro) – американський вчений, президент та головний редактор сайта KDnuggets.com

[11] Кислова О. Н. Интеллектуальный анализ данных: история становления термина / Ольга Кислова // Український соціологічний журнал. – 2011. – №1-2. – С. 83-94. – URL: http://www.sau.kiev.ua/docs/magazine/USM2011-1-2.pdf

[12] Шендрик А. И. Информационное общество и его культура: противоречия становления и развития – Информационный гуманитарный портал «Знание. Понимание. Умение» – Культурология, №4, 2010. – URL: http://www.zpu-journal.ru/e-zpu/2010/4/Shendrik

Компонентами системи інтелектуального аналізу є (Рис. 1.2):

1. База даних та OLTP – On-Line Transaction Processing – підсистема операційної (транзакційної) обробки даних, тобто СУБД, якими можуть бути реляційна база даних, сховище даних, транзакційна, об'ктно-орієнтована, об'єктно-реляційна, просторова (Spatial databases), часова (Temporal databases), текстова або мультимедійна бази даних; всесвітня павутина, тощо.
2. Сервер бази даних або сховища даних, який відповідає за добування істотних даних на основі користувацького запиту.
3. База знань – знання про предметну область, за якими визначають шляхи пошуку та оцінюють корисність отриманих патернів.
4. Служба добування знань – містить набір функціональних модулів для здійснення власне процедур інтелектуального аналізу даних.
5. Модуль оцінки патернів – визначає міру корисності патернів.
6. Графічний користувацький інтерфейс.

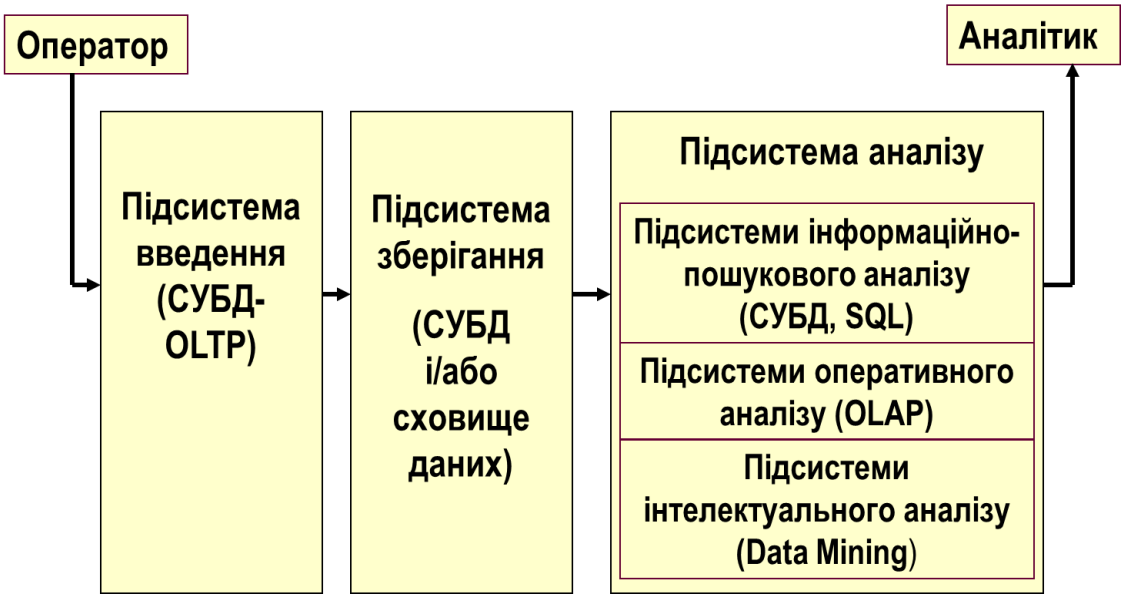


Рис. 1.2

Приклади сфер застосування інтелектуального аналізу великих даних:

- електронні бібліотеки;
- архіви зображень;
- бази даних геномних досліджень;
- медичні зображення;
- фінансові дані;
- бази даних підприємств;
- телекомунікаційні системи;
- всесвітня павутина;
- біометричні дані людей…

Основними напрямками (аспектами) методології аналізу даних є:

- добування різних видів нових знань;

- добування знань у багатовимірному просторі (аналіз у просторі куба даних може перевищувати потужність та гнучкість інтелектуального аналізу, особливо з урахуванням міждисциплінарного характеру досліджень);
- розширення можливостей пошуку у мережевому середовищі;
- обробка невизначеностей та шумів або неповних даних;
- оцінювання отриманих в результаті аналізу шаблонів.

Відповідно, основними типами задач аналізу даних є задачі опису та задачі прогнозування, тобто:

- *Класифікація* – процес знаходження моделей чи функцій, які описують та розрізняють класи для прогнозування класу довільно заданого об'єкта з відомими атрибутами на основі навчаючої вибірки;
- *Кластеризація* – виявлення ознак, за якими можна буде здійснювати класифікацію, шляхом групування "схожих" між собою об'єктів, генерування міток класів на основі відстаней між об'єктами;
- *Регресія* – встановлення залежностей неперервних результуючих змінних від вихідних;
- *Асоціація* – пошук закономірностей між декількома подіями, що відбуваються одночасно;
- *Послідовність* – пошук часових закономірностей між транзакціями;
- *Прогнозування* – оцінювання пропущених або майбутніх значень цільових чисельних показників;
- *Виявлення відхилень або викидів*;
- *Оцінювання* – передбачення неперервних значень ознаки;
- *Аналіз зв'язків* – пошук залежностей у наборі даних;
- *Візуалізація* – створення графічного образу аналізованих даних;
- *Підведення підсумків* – опис конкретних груп об'єктів з аналізованого набору;
- *Еволюційний аналіз* – опис та моделювання регулярностей та трендів для об'єктів, чия поведінка змінюється у часі.

До глибокого аналізу даних відносять групи задач, розв'язування яких передбачає

- відтворення «портрета» об'єкта у середовищі, тобто виведення моделі, яка «прозоро» інтерпретується і пояснює функціонування об'єкта;
- виявлення структури в даних, наприклад, ідентифікація системи зв'язків та впливів між характеристиками об'єкта у середовищі;
- знаходження закономірностей поведінки системи (об'єкта) – регулярності, періодичності, інваріантів, пошук аномалій [1].

Основними методами аналізу даних є:

- Множинний регресійний аналіз;
- Дерева рішень;
- Дискримінантний аналіз;
- Кластерний аналіз;
- Факторний аналіз;
- Виведення правил асоціації;

- Нейронні мережі;
- Аналіз часових рядів;
- Генетичні алгоритми;
- Візуалізація даних;
- Нечітка логіка...

Рис. 1.3

На Рис. 1.3 наведено неповний перелік дисциплін, що вплинули і продовжують впливати на формування змісту та методів аналізу даних, а на Рис. 1.4 наведено основні галузі аналізу даних:

- *OLAP* – On-Line Analitical Processing – технологія оперативної аналітичної обробки багатовимірних даних.
- *Data mining* – дослідження та виявлення “машиною” нових нетривіальних, необхідних для прийняття рішень, практично корисних, доступних для інтерпретації людиною знань, прихованих у “сирих” даних.
- *Візуалізація* – представлення багатовимірного розподілу даних на двовимірній площині, при якому відображено, принаймні якісно, основні закономірності вхідного розподілу – його кластерна структура, топологічні особливості, внутрішні зв’язки між ознаками, інформація про розташування даних у вхідному просторі тощо.

Рис. 1.4

1.3. Візуалізація статистичних даних та описова статистика

Візуалізація даних становить важливу складову аналізу, оскільки 80% інформації люди сприймають саме зором.

У статистиці візуалізація первинних даних здійснюється за допомогою гістограми, полігона частот, блочної діаграми, квартильних графіків, діаграми розсіювання тощо.

Нехай маємо деякий набір статистичних даних – результати статистичного дослідження, – причому для кожного об'єкта визначено два атрибути X та Y (Таблиця 1.4):

Таблиця 1.4

X	***Y***	***X***	***Y***	***X***	***Y***	***X***	***Y***
73	291	57	219	61	241	68	264
69	270	71	281	62	243	62	240
72	279	66	262	63	245	70	277
72	282	76	302	71	282	70	279
65	254	70	275	65	252	65	253
67	264	68	267	70	276	70	275
56	216	74	290	70	276	63	248
70	276	68	266	63	246	63	243
63	248	69	270	73	284	67	264
64	253	71	283	68	271	68	267
70	276	60	237	59	227	55	213
67	262	56	222	64	256	56	218
60	234	71	281	79	309	58	223
63	243	68	269	77	300	70	278
80	313	66	257	78	310	59	236

Побудуємо *варіаційний ряд* розподілу однієї з ознак, наприклад, X, підрахувавши частоту появи кожного значення X для діапазону значень від min(X) до max(X):

Значення	Частота	Значення	Частота	Значення	Частота	Значення	Частота	Значення	Частота
55	1	60	2	65	3	70	9	75	0
56	3	61	1	66	2	71	4	76	1
57	1	62	2	67	3	72	2	77	1
58	1	63	6	68	6	73	2	78	1
59	2	64	2	69	2	74	1	79	1
								80	1

Для побудови *інтервального ряду* слід визначити необхідну кількість та ширину класових інтервалів[13]. У даному випадку візьмемо 6 класів шириною 5, тобто першим буде клас 51-55, а останнім – 76-80:

Діапазон значень	51-55	56-60	61-65	66-70	71-75	76-80
Частота X	1	9	14	22	9	5

[13] Детальніше про визначення класових інтервалів див., наприклад, [11, Тема 1].

На Рис. 1.5 представлено *гістограму* та *полігон* для варіаційного ряду, а на Рис. 1.6 – для інтервального ряду.

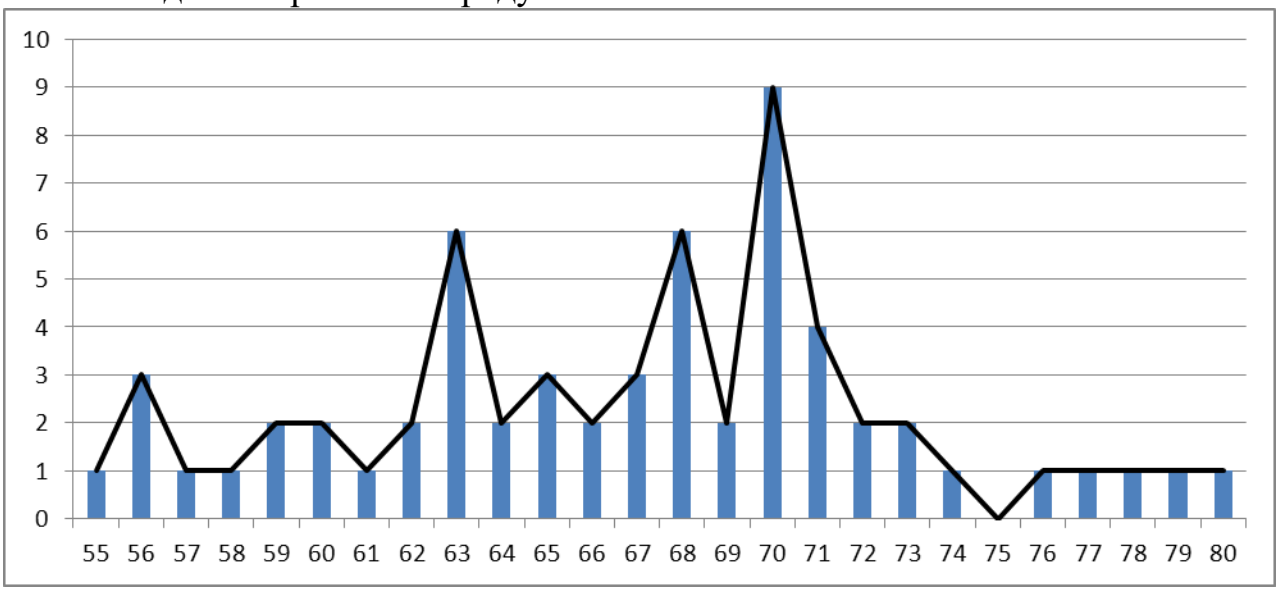


Рис. 1.5

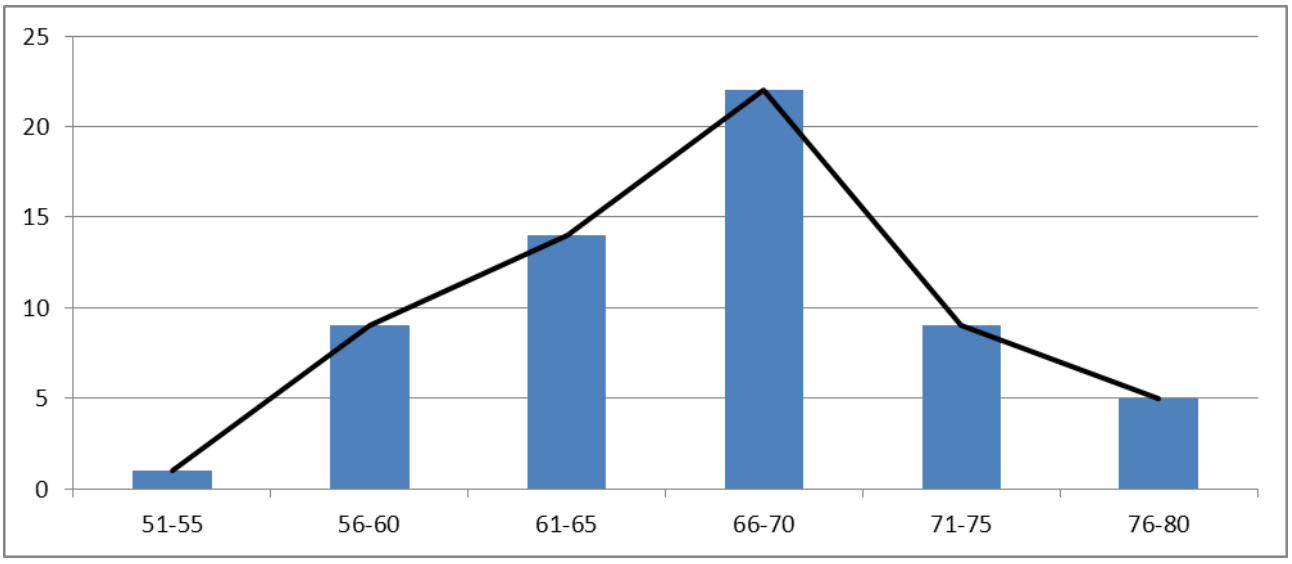


Рис. 1.6

Блочна діаграма (boxplot), розроблена у 70-ті роки Джоном Тьюки, – це графік, що компактно зображує одновимірний розподіл імовірностей. За блочною діаграмою зручно оцінювати медіани, квантилі, дисперсію та асиметрію розподілу, а також виявляти викиди. Для побудови блочної діаграми для змінної необхідно визначити 5 чисел: мінімальне та максимальне значення, перший, другий (медіану або середнє) та третій квартиль і стандартне квадратичне відхилення. В електронних таблицях для цього виконують функції МИН(діапазон), МАКС(діапазон), КВАРТИЛЬ(діапазон;1), МЕДИАНА(діапазон), КВАРТИЛЬ(діапазон,3), СТАНДОТКЛОН(діапазон)[14].

Довжину «вусиків» визначають декількома способами:

1) як мінімальне та максимальне значення по вибірці (у такому разі викиди відсутні);

[14] Порядок побудови блочної діаграми в Excel див.: https://exceltip.ru/создание-диаграммы-ящик-с-усами-в-excel/ або https://edwvb.blogspot.com/2014/12/kak-stroit-boxplot-v-excel.html.

2) довжину першого з «вусиків» діаграми визначають як різницю першого квартиля та півтори міжквартильних відстані, а другого – як суму третього квартиля та півтори міжквартильних відстані:

$$X_1=Q_1-k(Q_3-Q_1),\ X_2=Q_3+k(Q_3-Q_1),$$

де X_1 – нижня границя «вусика», X_2 – верхня границя «вусика», Q_1 – перший квартить, Q_3 – третій квартиль, k – коефіцієнт (зазвичай обирають k=1,5).

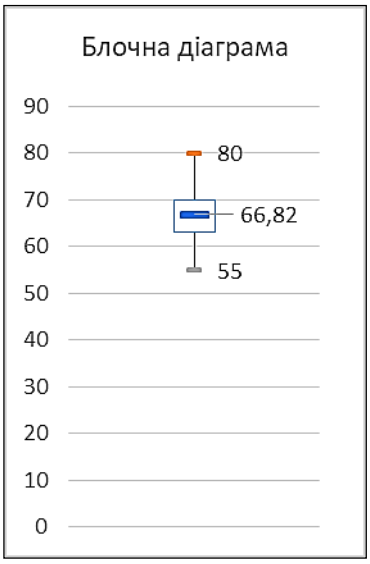
3) як середнє арифметичне (або медіану) по вибірці плюс/мінус одне стандартне квадратичне відхилення:

$$X_1 = \bar{X} - \sigma, X_2 = \bar{X} + \sigma \quad (\text{або } X_1 = Q_2 - \sigma, X_2 = Q_2 + \sigma).$$

Щоб побудувати таку блочну діаграму в Excel можна скористатись інструментом «Біржева діаграма».

Дані для графіка слід подати у такому порядку:

Перший квартиль (Q1)	Максимальне	Мінімальне	Третій квартиль (Q3)	Середнє
63	80	55	70	66,82



Значення відповідних параметрів розраховано для змінної X з таблиці 1.4. Рядки з даними слід декілька разів продублювати: для біржевої діаграми кількість рядків має перевищувати кількіть параметрів побудови, – а після побудови зайві діаграми можна не показувати.

Стовбець із значеннями середнього додають окремо до вже побудованої діаграми та у переліку рядів даних переміщують на передостаннє місце, щоб не зменшити висоту блока, оскільки положення його верхньої планки визначається останнім параметром. Залишається для ряду «Середнє» встановити маркер та додати підписи даних.

Для побудови *квантиль-квантильного графіка* (QQ-plot) необхідно поставити у відповідність квантилям нормального розподілу квантилі емпіричного розподілу.

Для цього 1) у першому стовпці розташуємо у порядку зростання (неспадання) усі значення масиву X – від 55 до 80, – усього 60 значень;

2) у другому стовпці запишемо проміжки розбиття одиничного інтервала, наприклад, на 20 рівних частин (назвемо цю змінну k)[15];

3) у третьому стовпці обчислимо значення k-того квантиля для розподілу X, тобто визначимо, яке значення досліджуваної змінної поділяє усю вибірку у співвідношенні k/(1-k) за формулою ПЕРСЕНТИЛЬ(масив X;k)[16];

4) в останньому стовпці слід обчислити відповідні квантилі нормального розподілу (формула НОРМСТОБР(k) або NORM.S.INV(k));

5) за даними двох останніх стовпців (Рис. 1.7) будуємо точкову діаграму (діаграму розсіювання) (див. Рис. 1.8).

[15] Одиничний інтервал можна розбити на довільну кількість частин, але бажано кратну 5.

[16] в інших версіях MS Office - PERCENTILE.INC(масив X;k)

X	K	PERCENTILE.EXC($A$2:$A$61;K)	NORM.S.INV(k)
55	0,05	56	-1,644853627
56	0,1	58,1	-1,281551566
56	0,15	60	-1,036433389
56	0,2	62	-0,841621234
57	0,25	63	-0,67448975
58	0,3	63	-0,524400513
59	0,35	64,35	-0,385320466
59	0,4	65,4	-0,253347103
60	0,45	67	-0,125661347
60	0,5	68	0
61	0,55	68	0,125661347
62	0,6	69	0,253347103
62	0,65	70	0,385320466
63	0,7	70	0,524400513
63	0,75	70	0,67448975
63	0,8	71	0,841621234
63	0,85	72	1,036433389
63	0,9	73,9	1,281551566
63	0,95	77,95	1,644853627
79			
80			

Рис. 1.7

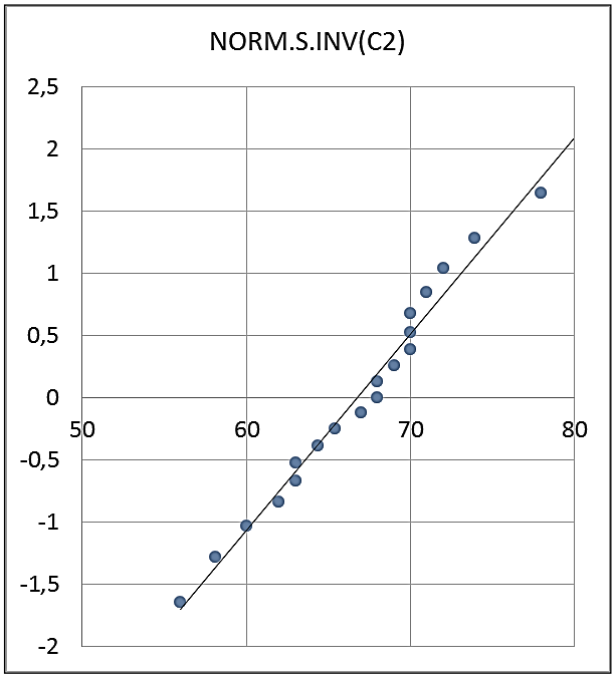


Рис. 1.8

QQ-plot використовують для первинного візуального порівняння емпіричного розподілу з нормальним: якщо точки побудованого графіка добре вкладаються на пряму, то можна припустити, що емпіричний розподіл близький до нормального.

Діаграму розсіювання (Scatter plot) також застосовують для візуалізації статистичних зв'язків між змінними. Значення змінних X та Y задають координати точок на площині. Отже для набору даних (Таблиця 1.4) матимемо таку діаграму (Рис. 1.9):

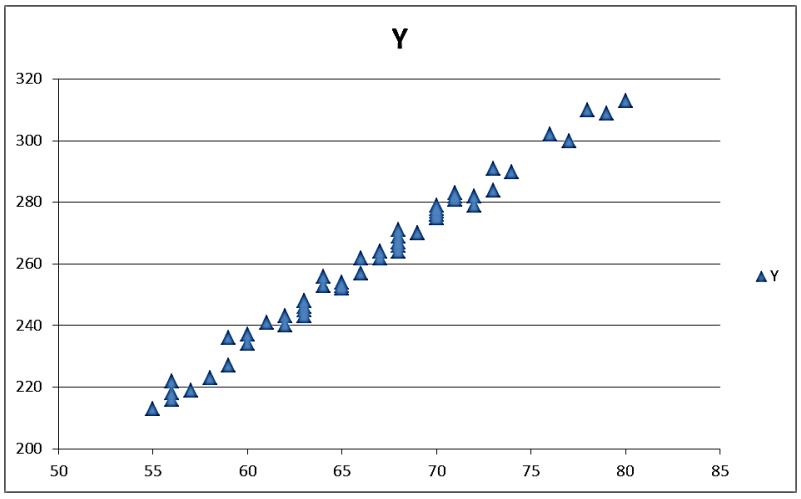


Рис. 1.9

Іншими видами статистичних графіків є кругова та пелюсткова діаграми тощо.

1.4. Аналіз даних з пакетом RapidMiner

Початкові відомості.

Пакет RapidMiner, раніше відомий як YALE (*у перекладі з англ.: "Ще одне навчальне середовище"*), почали розробляти у 2001 році Ральф Клінкенберг, Інго Мірсва та Симон Фішер у відділі штучного інтелекту Дортмундського технічного університету. У 2006 році Ральф Клінкенберг та Інго Мірсва заснували компанію Rapid-I. Відповідно, у 2007 році назву програмного продукту було змінено з YALE на RapidMiner. У 2013 році компанію також було перейменовано на RapidMiner.

Зараз RapidMiner (RM) – це програмна платформа, яка забезпечує інтегроване середовище для підготовки даних, машинного навчання, глибокого навчання, виведення тексту та прогнозної аналітики (Рис. 1.10). Вона використовується для бізнесу і комерційних застосувань, а також для досліджень, освіти, підготовки кадрів, швидкого створення прототипів та розробки додатків, а також підтримує всі етапи процесу машинного навчання, включаючи підготовку даних, візуалізацію результатів, перевірку моделей та оптимізацію моделей. RapidMiner Studio Free Edition, яка обмежується одним логічним процесором та 10 000 рядків даних, доступна під ліцензією AGPL. Детальну інформацію з установки продукту розміщено на сайті компанії: https://rapidminer.com/products/studio/

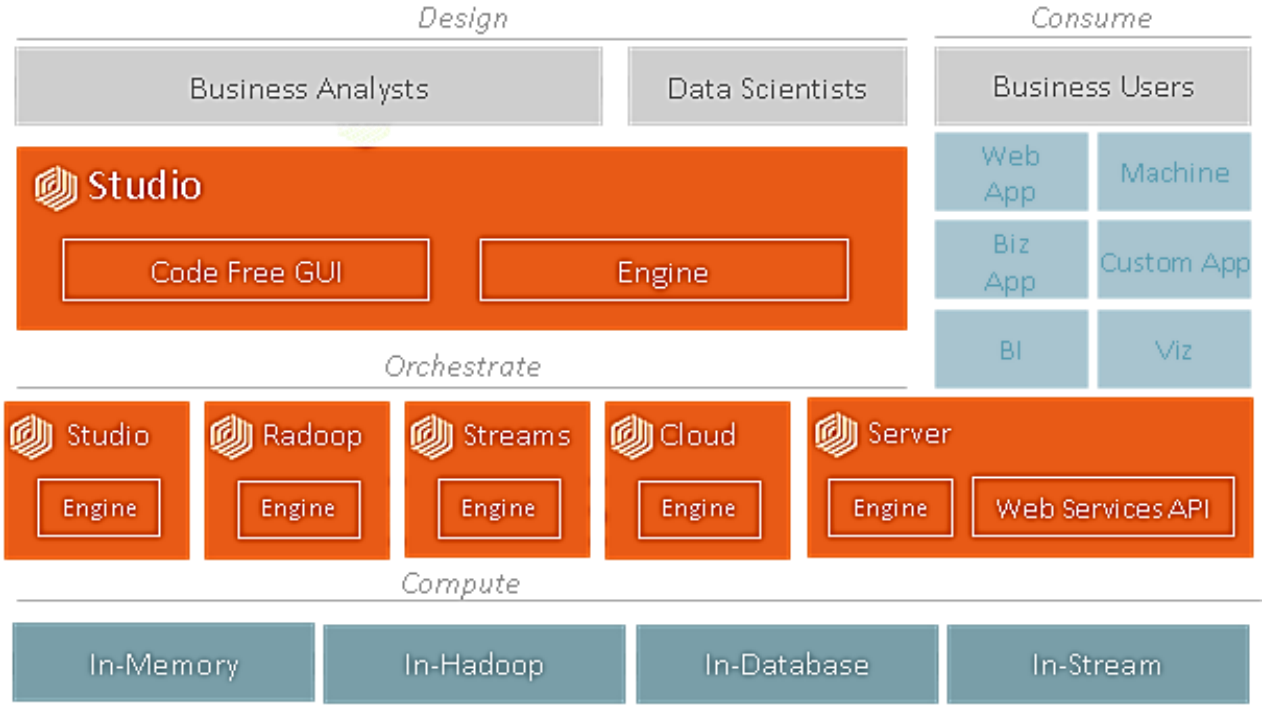


Рис. 1.10 Архітектура пакета RapidMiner

Основні поняття:

Основними об'єктами в RM є процес, оператор та репозиторій.

Процес – це сукупність операторів, з'єднаних між собою у заданому порядку для виконання потрібної задачі аналізу або обробки даних (Рис. 1.11).

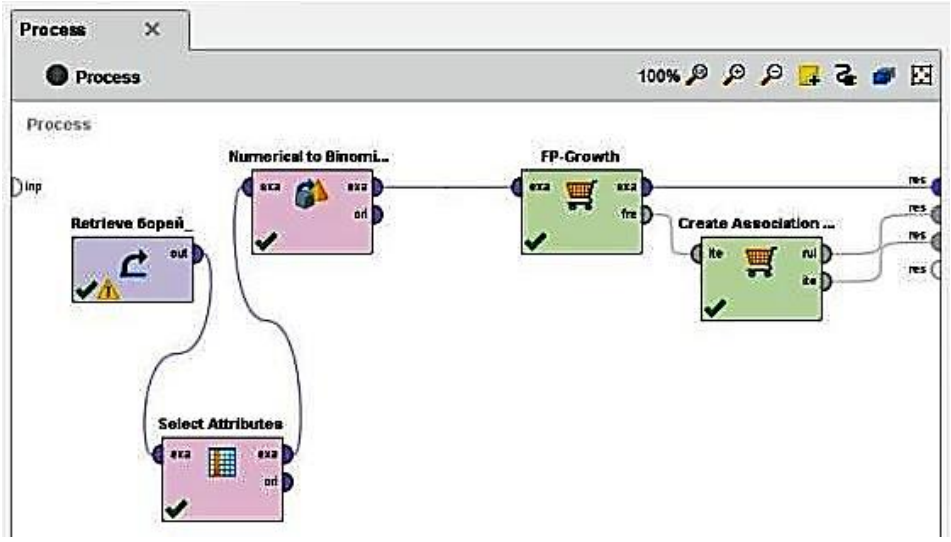


Рис. 1.11

Оператор – логічна одиниця процесу. Оператор виконує дії над даними, у нього є вхід-вихід (так звані «порти»). На вхід дані приходять з попереднього оператора або первинного джерела даних, а на вихід передаються результати виконання. Таким чином можна створювати ланцюжки опрацювання даних, в тому числі і розпаралелювані.

В інтерфейсі програми операторам відповідає вкладка ***Operators***, де вони згруповані за функціональністю. Для використання оператора його необхідно перенести в робочу область процеса.

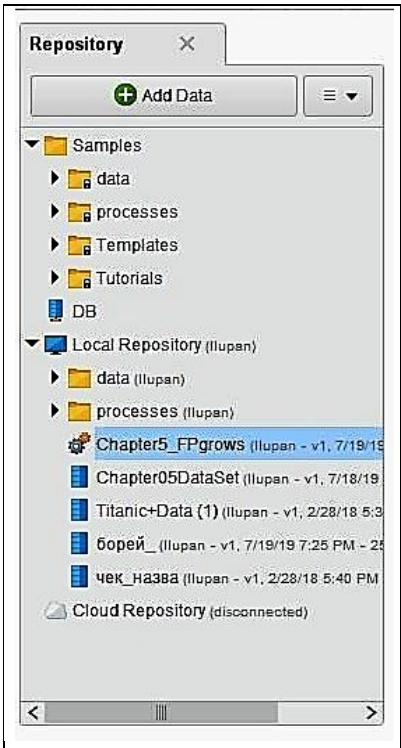


Рис. 1.12

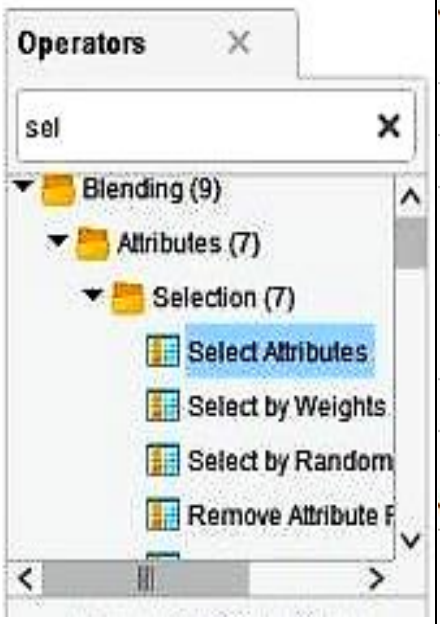


Рис. 1.13

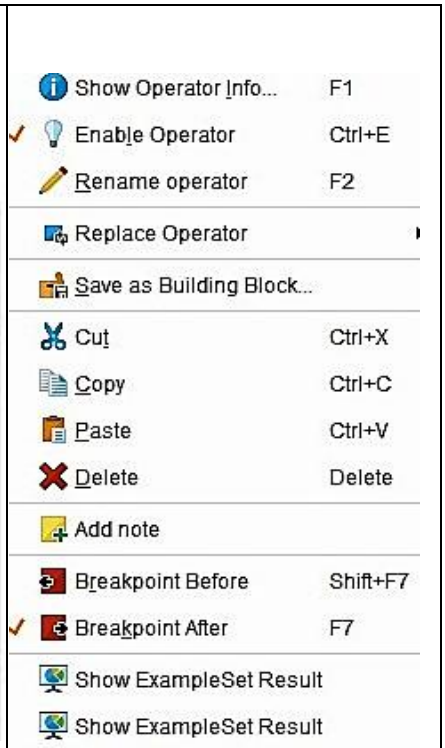


Рис. 1.14

За початковими літерами назви оператора його швидко можна знайти (Рис. 1.13).

Репозиторій – місце для зберігання процесів RM та даних, – може бути локальним, а також віддаленим (RapidMiner Server), для якого можливе виконання процесів на боці сервера і т.п. (Рис. 1.12).

У вкладці **Repositories** в RM можна побачити набір прикладів (*Samples*), поточні з'єднання з базами даних, визначені через Tools → Manage Database Connections (*DB*) та місце для зберігання власних процесів на комп'ютері (*Local Repository*).

Процес може бути перенесений у вікно **Process** як один оператор (перетягуванням піктограми) або розгорнутий у вигляді послідовності операторів (подвійним «кліком»).

На Рис. 1.15 бачимо виділений оператор **Execute**, який виконуватиме процес агломеративного кластерного аналізу (04_AgglomerativeHierarchicalClustering), а вище – той самий процес кластерного аналізу у розгорнутому вигляді як послідовність операторів Retrieve (отримати дані) та **AgglomerativeClustering.**

Для виділеного у вікні **Process** процеса або оператора можна за допомогою розташованого справа вікна **Parameters** переглянути та налаштувати параметри.

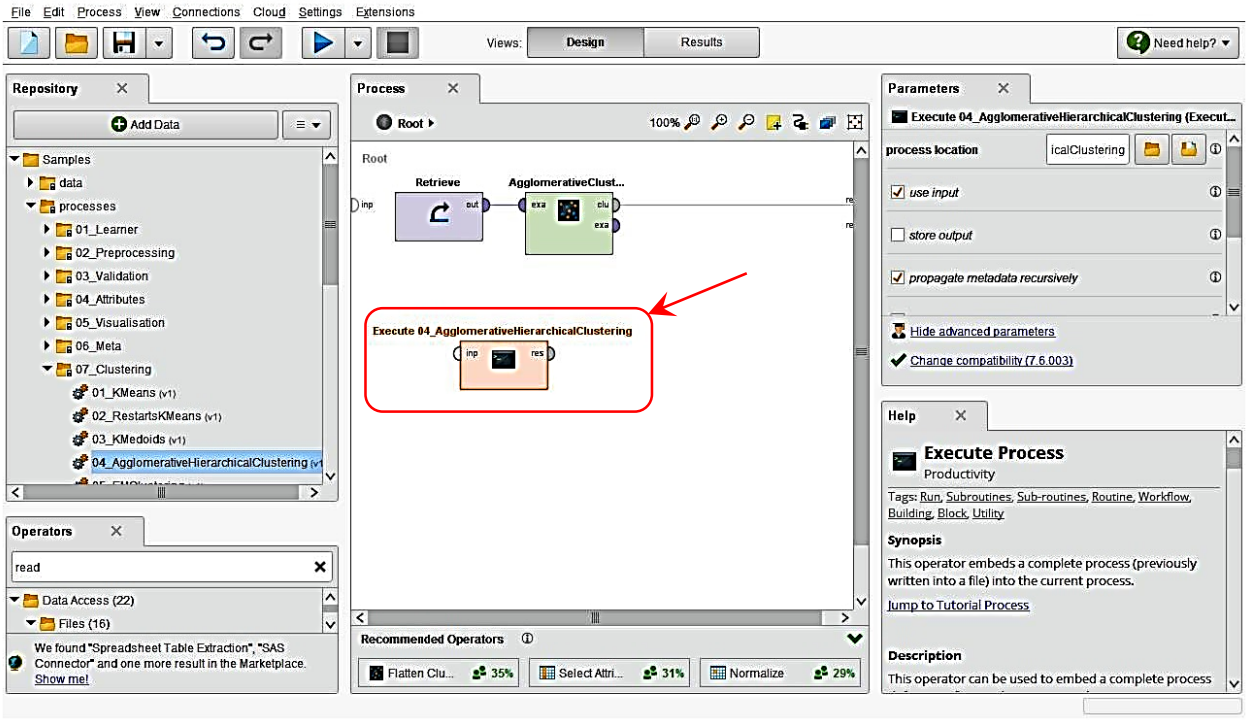


Рис. 1.15

Запуск процеса на виконання здійснюється кнопкою  або клавішею F11. У режимі **Design** здійснюється налаштування процеса в цілому та окремих операторів. Для перегляду результатів слід перейти у режим **Result**.

Початок роботи. Перегляд вхідних даних.

Першою дією в аналізі є перегляд статистичних даних та здійснення процедур описової статистики. Дані можна завантажити до репозиторія, застосувавши команду **Add data** у вікні **Repository** та вказавши у відповідному діалоговому вікні місце знаходження даних (відповідну папку). В результаті дані будуть переміщені до відповідного репозиторія і зберігатимуться у спеціальному форматі. При цьому їх можна буде редагувати, зокрема, вилучати стовпці, змінювати типи та ролі атрибутів.

При перетягуванні піктограми, що відповідає набору даних, з вікна репозиторія до вікна процесу автоматично створюється оператор **Retrieve**.

В RM також можна використовувати дані у різних форматах без переміщення до репозиторія. У такому разі використовують оператор **Read**. Якщо форма представлення цих даних не відповідає цільовому оператору обробки даних, то доведеться застосовувати додатково оператори попередньої обробки.

Далі слід з'єднати кружечок ***out*** оператора **Retrieve Data** (або іншого, за допомогою якого отримано доступ до даних) з кружечком ***res*** на правому боці вікна **Process**. На цьому дизайн процесу завершено, процес можна запустити на виконання (**Run**), а результати переглянути у режимі **Results**.

На Рис. 1.16 для даних з репозиторія за допомогою контекстного меню відкрито **Data Editor**, у якому для будь-якого із стовпців дозволено виконати вилучення, змінення типу та заповнення випадковими даними чи константами. Між тим для того, щоб вилучити стовпець із даних, отриманих у форматі SPSS, довелося застосовувати оператор **Remove Attribute Range** з параметром, що задає діапазон стовпців, які слід вилучити.

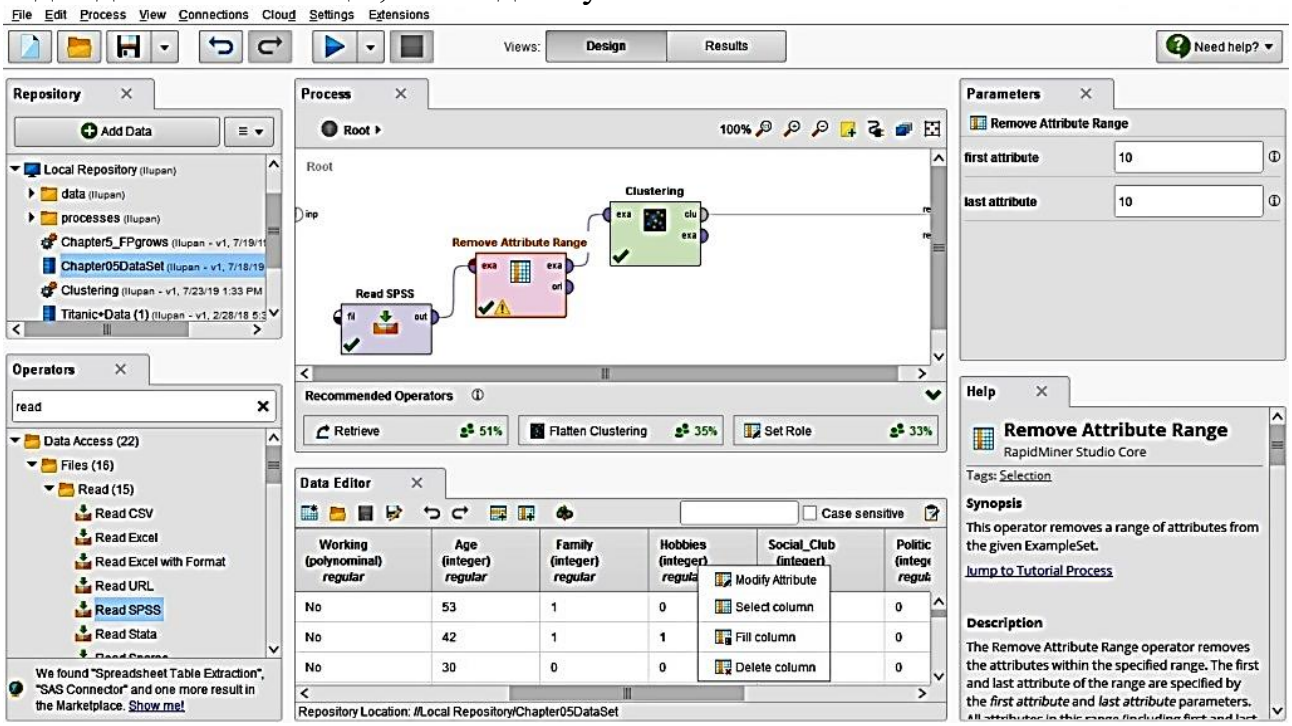
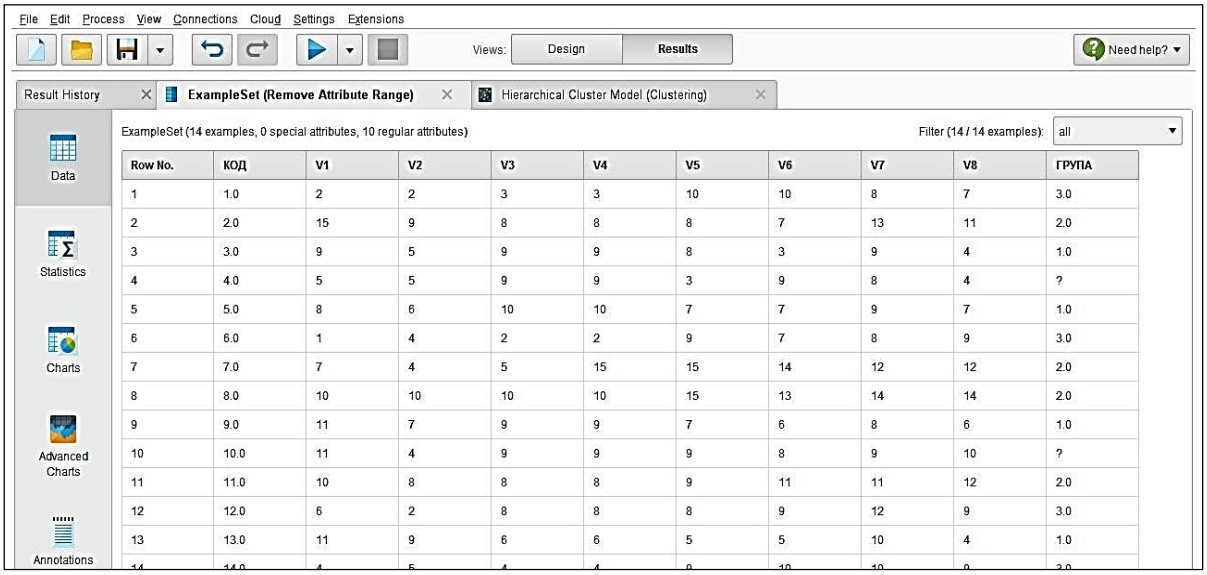


Рис. 1.16

Результатами для процеса перегляду даних є Data – власне таблиця даних (Рис. 1.17); Statistics – короткі статистичні дані по кожній змінній – тип даних, мінімальне та максимальне значення, середнє та стандартне квадратичне відхилення для числових змінних, гістограму розподілу (Рис. 1.18). У закладці Charts – надається можливість побудувати необхідні графіки (Рис. 1.19).



Row No.	КОД	V1	V2	V3	V4	V5	V6	V7	V8	ГРУПА
1	1.0	2	2	3	3	10	10	8	7	3.0
2	2.0	15	9	8	8	8	7	13	11	2.0
3	3.0	9	5	9	9	8	3	9	4	1.0
4	4.0	5	5	9	9	3	9	8	4	?
5	5.0	8	6	10	10	7	7	9	7	1.0
6	6.0	1	4	2	2	9	7	8	9	3.0
7	7.0	7	4	5	15	15	14	12	12	2.0
8	8.0	10	10	10	10	15	13	14	14	2.0
9	9.0	11	7	9	9	7	6	8	6	1.0
10	10.0	11	4	9	9	9	8	9	10	?
11	11.0	10	8	8	8	9	11	11	12	2.0
12	12.0	6	2	8	8	8	9	12	9	3.0
13	13.0	11	9	6	6	5	5	10	4	1.0

Рис. 1.17

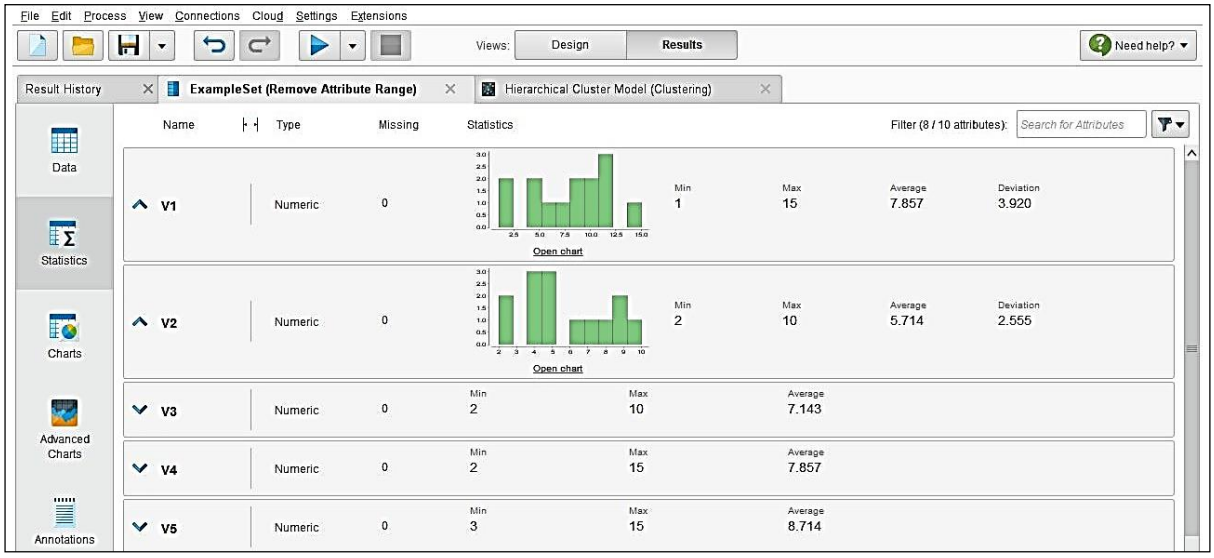


Рис. 1.18

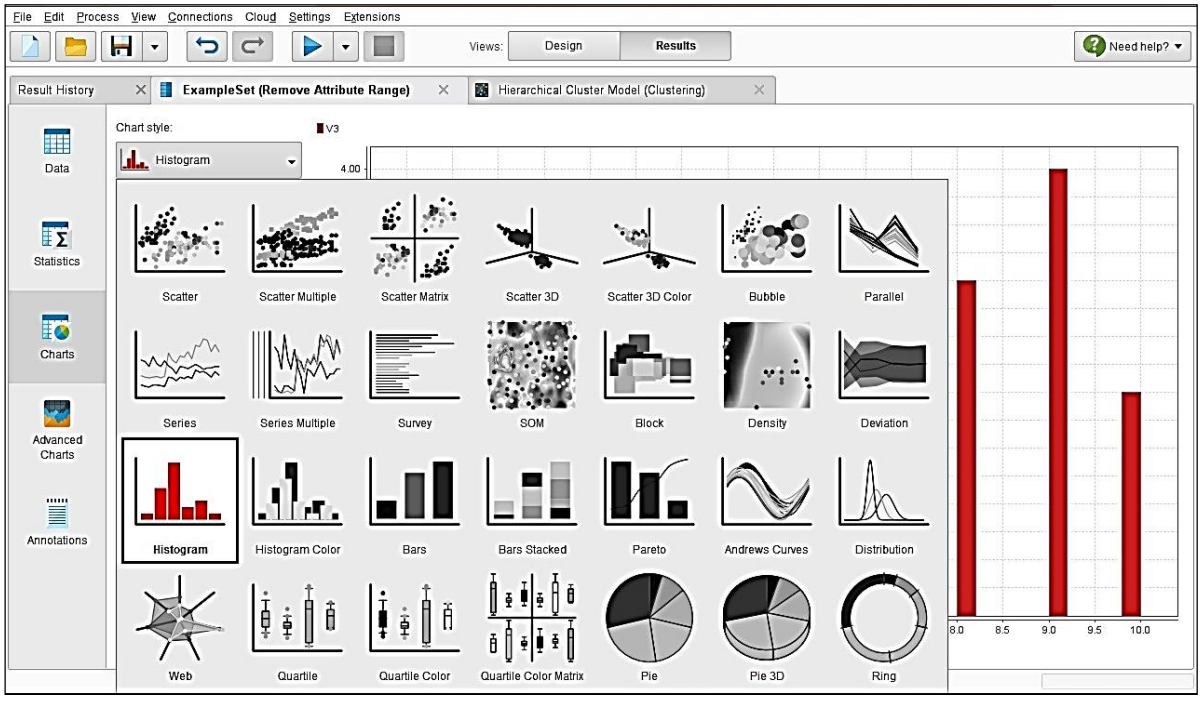


Рис. 1.19

Таким чином можна переглянути набори даних, що надаються разом з пакетом RapidMiner у навчальних прикладах.

Щоб переглянути проміжі результати виконання окремих операторів складного процесу, після його запуску можна за допомогою контекстного меню до вибраного оператора застосувати команду **Show ExamplesSet Result** (Рис. 1.14). Причому в одному випадку отримаємо вигляд даних до, а в іншому – після виконання оператора.

Також є можливість призупинити виконання процеса для перегляду проміжних результатів, встановивши конрольні точки до або після виконання вибраного оператора (**Breackpoint**).

1.5. Основи роботи з пакетом KNIME

KNIME[17] – вільно поширювана аналітична платформа з широким набором засобів інтелектуального аналізу даних (доступу до даних різноманітних форматів, трансформації даних, аналітичних функцій, засобів візуалізації та підготовки звітів).

Процес аналізу, тобто перетворення вхідних даних у результати, в KNIME представлено потоком робіт (workflow), який графічно зображено в основному вікні програми (Рис. 1.20):

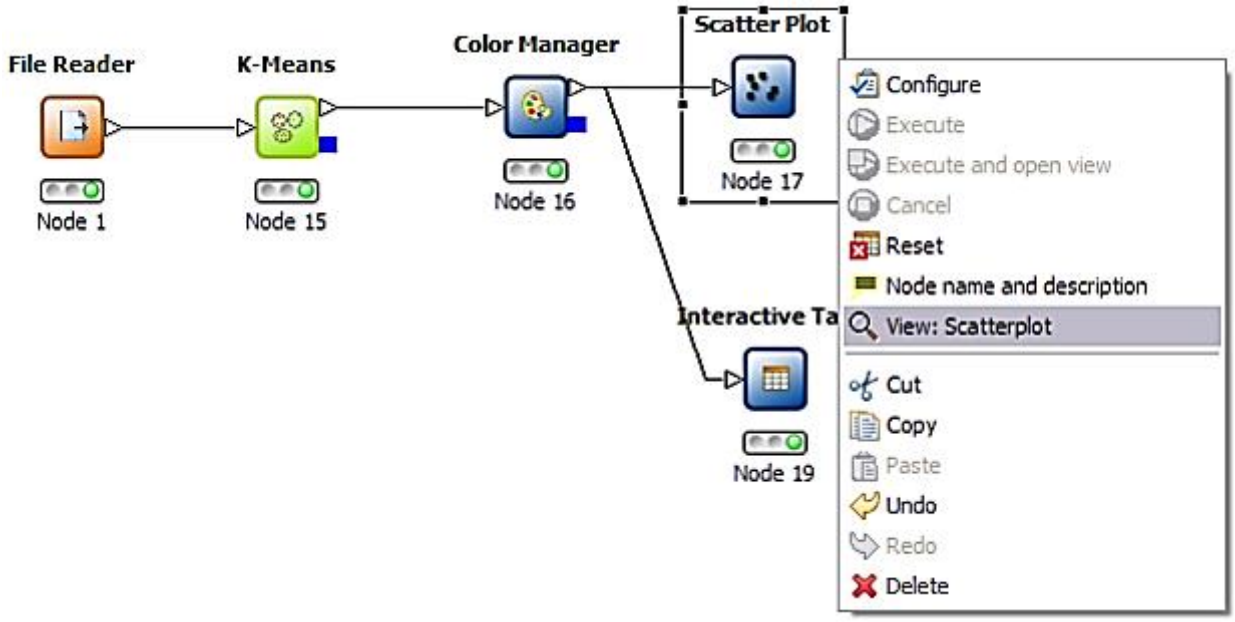


Рис. 1.20. Простий потік робіт

Зображення складається з вузлів (nodes), які інкапсулюють операції над даними, та стрілок, що з'єднують «порти» вузлів у напрямку обміну даними між операціями. Кожен вузол має бути налаштований (відконфігурований – Configure) відповідним чином. Вузол також можна знищити, скопіювати, переіменувати, використовуючи контекстне меню, або переглянути результати його виконання.

Стан виконання тої чи іншої операції або її готовність до виконання (node status) візуалізується індикатором (Рис. 1.21):

[17] http://www.knime.org/

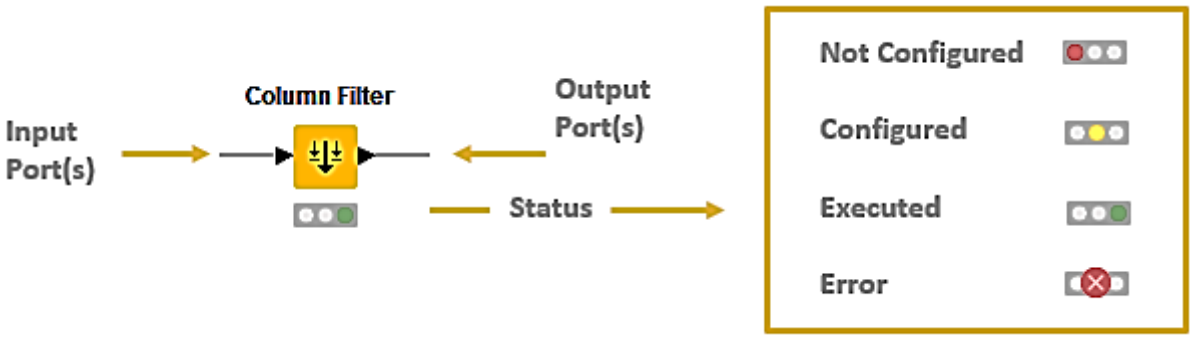


Рис. 1.21

Рекомендації до початку роботи можна отримати на сайті розробника: https://www.knime.com/getting-started.

1.6. Завдання для виконання

Оцінювання

Вид роботи	**Вид звітності**	**Оцінка**
Індивідуальне завдання	Реферативна доповідь	до 5 балів
Вправи	Письмовий звіт	2 бали за кожну
Лабораторна робота	Звіт з висновками	до 10 балів
Тест		до 5 балів
Максимальна оцінка за тему	**20 балів**	***формується із завдань на вибір студента***

Домашнє завдання

Для виконання наступних лабораторних робіт знайдіть дані для аналізу з різних галузей, наприклад, на одному з безкоштовних джерел. Перелік деяких з таких джерел наведено у Додатку А.

Завдання до лабораторної роботи №1. Попередній статистичний аналіз.

1. Ознайомтеся з цікавими для Вас даними на сайті https://www.google.com/publicdata/. Розгляньте різні способи візуалізації цих даних. Додайте до звіту ваші міркування (інтерпретацію) стосовно розглянутих даних.
2. Знайдіть цікаві для Вас статистичні дані. Виконайте статистичний аналіз цих даних (див. пункт 3).
3. Виконайте процедури описової статистики у пакеті MS Excel:
   а) обчисліть середнє, дисперсію, стандартне квадратичне відхилення, мінімальне та максимальне значення, моду та медіану за допомогою вбудованих функцій;
   б) застосуйте процедури описової статистики з Пакета Аналізу;
   в) побудуйте варіаційний ряд;
   г) побудуйте гістограму та полігон розподілу частот;
   д)* побудуйте блочну діаграму;

е)* побудуйте квантильний графік;
ж) побудуйте діаграму розсіювання;
з) побудуйте куб даних, використовуючи засіб «Зведена таблиця».

Завдання до лабораторної роботи №2. Ознайомлення з інтерфейсом пакетів RapidMiner та KNIME або SPSS та Statistica.

1. Розгляньте декілька наборів даних з прикладів, представлених у репозиторії пакета RapidMiner або KNIME.
2. Для обраних наборів створіть процеси для перегляду та отримайте результати (описову статистику та графіки).
3. З'ясуйте, у які формати можна експортувати результати та з яких форматів дані можна отримати (імпортувати).

4. Виконайте процедури описової статистики для цих даних у пакеті SPSS або Statistica:
   а) отримайте значення середнього, дисперсії, стандартного квадратичного відхилення;
   б) отримайте гістограму розподілу;
   в) отримайте блочну діаграму (box plot);
   г) отримайте квантильний графік.
5. Оформіть звіт, у якому
   а) опишіть розглянуті набори даних (кількість та типи змінних, кількість кортежів у наборі даних, отримані характеристики);
   б) зробіть висновки стосовно інструментарію для здійснення процедур описової статистики розглянутих пакетів.

Індивідуальне завдання

Ознайомитися з розширеним інструментарієм для побудови графіків у пакеті Rapid Miner: https://docs.rapidminer.com/downloads/RapidMiner-5.2-Advanced-Charts.pdf

Питання для обговорення

1. Опишіть етапи інтелектуального аналізу даних.
2. Наведіть приклад, де успіх бізнесу залежить від застосування технології інтелектуального аналізу даних. Чи можна було б обійтись простим статистичним аналізом чи запитом до бази даних?
3. Уявіть, що Ви розробляєте програмне забезпечення для деякого університету, і вашим завданням є створення системи для інтелектуального аналізу з відомостями про студентів (ім'я, адреса, рік вступу, вивчені курси з оцінками). Опишіть архітектуру, яку б ви вибрали. Яке призначення кожного компонента?
4. У чому полягає функціональність перелічених задач аналізу даних: характеризація, дискримінація, аналіз асоціацій, класифікація, прогнозування, кластеризація, еволюційний аналіз. Наведіть приклад застосування такої функціональності, використовуючи деяку реальну базу даних, яку ви знаєте.

5. У чому різниця та схожість між дискримінацією та класифікацією? Між узагальненням та кластеризацією? Між класифікацією та прогнозуванням?
6. Викиди часто вважають шумом. Однак те, що в одній ситуації є шумом, може виявитися цінним в іншій ситуації. Наприклад, винятки в операціях з кредитними картами можуть допомогти виявити випадки шахрайського використання. Розглядаючи виявлення шахрайства як приклад, запропонуйте два методи, які можуть бути використані для виявлення викидів та обговоріть, який з них більш надійний.
7. Опишіть три виклики для інтелектуального аналізу даних щодо питань взаємодії з користувачами.
8. Які основні проблеми у видобутку величезної кількості даних (наприклад, мільярди кортежів) у порівнянні з видобуванням невеликої кількості даних (наприклад, набору даних у кілька сотень кортежів)?
9. Охарактеризуйте основні виклики досліджень інтелектуального аналізу даних в одній конкретній області застосування, такі як аналіз потокових / сенсорних даних, просторово-часовий аналіз даних або біоінформатика.
10. Чим сховище даних відрізняється від бази даних?

Вправи

1.1. Наведіть три додаткові загальноприйняті статистичні показника, які використовують для характеристики розмаху (варіативності) даних. Обговоріть, як їх можна ефективно обчислити у великих базах даних [22, 2.1].

1.2. Припустимо, що дані для аналізу включають атрибут віку із значеннями: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. [22, 2.2]

а) Яке середнє значення цих даних? Яка медіана?

б) Яка мода даних? Прокоментуйте модальність розподілу даних (тобто, унімодальний, бімодальний, тримодальний тощо).

в) Який розмах даних?

г) Чи можна знайти (приблизно) перший квартиль (Q1) і третій квартиль (Q3) цих даних?

д) Обчисліть та назвіть п’ять значень, що описують розподіл цих даних (мінімальне, Q1, медіану, Q3 та максимальне).

е) Побудуйте блочну діаграму (box plot).

1.3. Припустимо, що значення для деякого набору даних згруповані в інтервали. Інтервали та відповідні частоти такі:

Вік	Частота
1-5	200
6-15	450
16-20	300
21-50	1500
51-80	700
81-110	44

Обчисліть медіану даного набору значень [22, 2.3].

1.4. Нехай у лікарні зібрано дані про вік та ступінь ожиріння для 18 випадково відібраних хворих: [22, 2.4]

Вік	23	23	27	27	39	41	47	49	50
% жиру	9,5	26,5	7,8	17,8	31,4	25,9	27,4	27,2	31,2
Вік	52	54	54	56	57	58	58	60	61
% жиру	34,6	42,5	28,8	33,4	30,2	34,1	32,9	41,2	35,7

а) обчисліть середнє, медіану та стандартне квадратичне відхилення для двох змінних;

б) зобразіть блочні діаграми для двох змінних;

в) побудуйте діаграму розсіювання та квантиль-квантильний графік на основі двох змінних.

1.5. Медіана є одним з найважливіших цілісних показників статистичного аналізу даних. Запропонуйте кілька методів для медіанної апроксимації. Проаналізуйте їхню складність при різних значеннях параметрів і визначте, якою мірою реальне значення можна апроксимувати. Крім того, запропонуйте евристичну стратегію, щоб збалансувати точність і складність, а потім застосувати її до всіх методів, які ви назвали. [22, 2.7]

1.6. Розмір взуття (довжину ступні) вказано у сантиметрах: 23, 24, 24, 25, 25, 26, 26, 22, 23, 28, 23, 23, 27, 27.

Як ці дані можуть бути використані для аналізу? До якого типу шкали слід віднести такі дані? Відповідь поясніть.

1.7. Якість даних може бути оцінена зокрема за точністю, повнотою та узгодженістю. Для кожної з трьох вищезазначених властивостей обговоріть, як оцінка якості даних може залежати від передбачуваного використання даних, наведіть приклади. Запропонуйте два інших аспекта оцінювання якості даних. [22, 3.1]

1.8. У реальних даних кортежі з пропущеними значеннями для деяких атрибутів є звичайним явищем. Опишіть різні методи для вирішення цієї проблеми. [22, 3.2]

1.9. Для атрибута ВІК задано такі значення: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

а) Виконайте згладжування даних, використовуючи метод ковзного середнього з глибиною 3.

б) Як виявити викиди в даних?

в) Які ще існують методи згладжування? [22, 3.3]

1.10. Обговоріть питання, які необхідно враховувати при інтеграції даних. [22, 3.4]

1.11. Надійне завантаження даних є проблемою в системах баз даних, оскільки вхідні дані часто забруднені. У багатьох випадках вхідний запис може пропускати кілька значень; деякі записи можуть бути забруднені, деякі значення даних виходять за межі допустимого діапазону значень або мають інший типу, ніж очікувалося. Розробіть автоматизований алгоритм очищення та завантаження даних, щоб помилкові дані були відмічені, а забруднені дані не були помилково вставлені в базу даних під час завантаження. [22, 3.14]

Література до теми 1:

При підготовці даного розділу були використані збірка наукових статей [9], навчальні посібники і підручники [2, 3, 4, 5, 8, 11, 15, 18, 19, 22, 23], стаття [1] та монографія [13] (згідно нумерації загального списку – див. с. 93).

- ✓ Балабанов О.С. Аналітика великих даних: принципи, напрямки і задачі (огляд). //Проблеми програмуванння. – №2. – 2019. – С.47-68. – URL: https://doi.org/10.15407/pp2019.02.047
- ✓ Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. пособие / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2007. – 384 с.
- ✓ Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.
- ✓ Бауэр Ф., Гооз Г. Информатика, 1990.
- ✓ Брой М. Информатика, 1996.
- ✓ Інтелектуальний аналіз даних: підручник /О.І.Черняк, П.В.Захарченко. – К.: Знання, 2014.
- ✓ Кибернетика. Становление информатики, 1986.
- ✓ Кислова О. Н. Интеллектуальный анализ данных: история становления термина / Ольга Кислова // Український соціологічний журнал. – 2011. – №1-2. – С. 83-94. – URL: http://www.sau.kiev.ua/docs/magazine/USM2011-1-2.pdf
- ✓ Лупан І.В., Авраменко О.В., Акбаш К.С. Комп'ютерні статистичні пакети: навчально-методичний посібник. – 2-е вид. – Кіровоград, «КОД», 2015. – 236 с.
- ✓ Орехов В.Д. Прогнозирование развития человечества с учетом фактора знания. Монография. – Жуковский: МИМ ЛИНК, 2015. – 210 с. – URL: http://world-evolution.ru/monografiya/
- ✓ Представление и использование знаний/ Под ред. Х.Уэно, 1989.
- ✓ Чубукова И.А. Data Mining (Курс лекций) – URL: http://kist.ntu.edu.ua/textPhD/Chubukova-Data_Mining.pdf
- ✓ Han J., Kamber M., Pei J. Data mining: Concepts and Techniques. – Morgan Kaufmann Publishers. – 2012. – 740 p.
- ✓ Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. – URL: http://www.mmds.org
- ✓ Matthew North. Data Mining for the Masses. – 2012. – 264 p. – URL: https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf

ТЕМА 2. КЛАСТЕРНИЙ АНАЛІЗ

2.1. Теоретичні відомості.

Призначення кластерного аналізу

Кластерним аналізом (від англ. *cluster* – гроно, кущ, рій, скупчення, жмуток) називають сукупність методів групування об'єктів на основі схожості між об'єктами. Таким чином кластерний аналіз дозволяє виявляти структуру даних.

Задачами кластерного аналізу є:

- розробка типології або класифікації;
- дослідження корисних концептуальних схем групування об'єктів;
- представлення гіпотез на основі дослідження даних;
- перевірка гіпотез про існування виділених деяким способом типів досліджуваних даних.

Кластеризація є логічним продовженням класифікації, але її відносять до методів "навчання без вчителя", оскільки при формуванні кластерів не використовують "навчаючу множину" (множину об'єктів, кластерна приналежність яких заздалегідь відома).

У наукових дослідженнях кластеризація при правильному застосуванні дозволяє навіть відкривати нові перспективні напрямки. Яскравим прикладом є періодична таблиця елементів: безперечною заслугою Д.Менделєєва є те, що у 1869 р. він поділив 60 відомих на той час елементів на кластери або періоди за схожими характеристиками. Вивчення причин об'єднання елементів у явно виражені кластери визначило пріоритети наукових досліджень на роки вперед. І лише через 50 років засобами квантової фізики вдалося науково обгрунтувати такий поділ [3].

Однак кластеризація є і поки залишається описовою, розвідувальною процедурою: на її основі не можна робити статистичних висновків. Правильність кластеризації оцінюють лише непрямими методами:

- шляхом встановлення контрольних точок;
- перевіркою стабільності кластеризації шляхом введення в модель нових змінних;
- порівнянням результатів застосування різних методів кластеризації (створення схожих кластерів при використанні різних методів вказує на правильність кластеризації).

Ієрархічні методи кластерного аналізу

На сьогоднішній день розроблено більше ста різноманітних алгоритмів кластеризації, які поділяють, зокрема, на ієрархічні та неієрархічні (ітераційні) алгоритми.

При застосуванні ієрархічного агломеративного (об'єднувального) кластерного аналізу (AGglomerative NESting, AGNES) на першому кроці кожен об'єкт розглядають як окремий кластер. На наступних кроках найбільш схожі

між собою об’єкти об’єднують у кластери. Процес продовжується доти, поки всі об’єкти не об’єднаються в один кластер.

При застосуванні ієрархічного дивізимного (подільного) кластерного аналізу (Divisive ANAlysis, DIANA) на першому кроці уся сукупність досліджуваних об’єктів розглядається як один кластер, який на наступних кроках поділяється на менші.

Ієрархічні методи є найбільш наочними, однак застосовуються як правило на невеликих наборах даних. Результати ієрархічного кластерного аналізу зручно представляти у вигляді дендрограм (Рис. 2.4, Рис. 2.5). Кожен рівень дендрограми відповідає одному кроку кластеризації.

Усі методи кластерного аналізу базуються на двох основних припущеннях:

1) досліджувані ознаки в принципі допускають поділ досліджуваної сукупності об’єктів на кластери;

2) шкали вимірювання усіх досліджуваних ознак співрозмірні, тобто представлені у нормалізованому (стандартизованому) масштабі, але з урахуванням важливості (ваги) тої чи іншої ознаки.

Найчастіше для стандартизації усі дані ділять на стандартне квадратичне відхилення досліджуваної ознаки, а для врахування ваги – множать на ваговий коефіцієнт або експертну оцінку її важливості:

$$\textit{стандартизоване значення: } x'_i = \frac{x_i}{\sigma}, \text{ де } \sigma = \sqrt{\frac{\sum_n (x_i - \bar{x})^2}{n-1}}.$$

Формальна постановка задачі кластеризації полягає у тому, що є набір даних з такими властивостями:

- кожен екземпляр виражений числовим значенням;
- клас для кожного екземпляра невідомий.

Необхідно знайти:

- спосіб порівняння даних між собою;
- спосіб кластеризації (об’єднання/поділ на кластери);
- розподіл даних по кластерах.

Міри схожості об’єктів

Критерієм схожості об’єктів кластеризації є “відстань” між ними у просторі досліджуваних змінних.

Найбільш популярною мірою схожості є коефіцієнт кореляції Пірсона: $r_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}}$. Основним недоліком такої міри є чутливість до форми за рахунок зниження чутливості до величини різниці між змінними.

Іншими мірами схожості є, наприклад:

1) евклідова відстань $d_{2ij} = \sqrt{\sum_{t=1}^{m}(x_{it} - x_{jt})^2}$, де t – розмірність простору. У просторі двох змінних $d_{2ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Тут d_{ij} – відстань між і-тим та j-тим об'єктами, X та Y – досліджувані ознаки.
2) відстань за Хеммінгом (або Манхеттенська або відстань міських кварталів): $d_{Hij} = \sum_{t=1}^{m}\left|x_{it} - x_{jt}\right|$, – на неї менше впливають окремі "викиди" (великі різниці), оскільки їх не підносять до степеня;
3) відстань Чебишова: $d_{\infty ij} = \max_{1\le t\le m}\left|x_{it} - x_{jt}\right|$, – дозволяє розрізнити об'єкти, які відрізняються лише однією координатою;
4) відстань Махаланобіса: $d_M(x_i, x_j) = (x_i - x_j)\, S^{-1}\, (x_i - x_j)^t$, – дає хороший результат при застосуванні на конкретному класі і погано працює на усій множині вхідних даних;
5) пікова відстань: $d_{Lij} = \frac{1}{m}\sum_{t=1}^{m}\frac{\left|x_{it} - x_{jt}\right|}{x_{it} + x_{jt}}$, – відстань в ортогональному просторі, тобто необхідною умовою її застосування є незалежність змінних
6) відстань Мінковського: $\rho(x, y) = \left(\sum_{i=1}^{n}\left|x_i - y_i\right|^q\right)^{1/q}$.

Коректно обрати міру відстані можна лише при врахуванні характеристик вхідних даних.

Для бінарних змінних відстані обчислюють за коефіцієнтами асоціативності [17]. Позначивши літерами *a*, *b*, *c*, *d* комірки у 2x2-таблиці асоціативності, обчислюють простий коефіцієнт $S = \frac{a+d}{a+b+c+d}$, або коефіцієнт Жаккара $J = \frac{a}{a+b+c}$.

1 – наявність ознаки, 0 – відсутність ознаки	1	0
1	*a*	*b*
0	*c*	*d*

Коефіцієнт Жаккара виявився частковим випадком (коли усі дані двійкові) для коефіцієнта Гауера: $s_{ij} = \sum_{k=1}^{p} S_{ijk} / \sum_{k=1}^{p} W_{ijk}$, де W_{ijk} – вагова змінна, яка дорівнює 1, якщо порівняння об'єктів за ознакою *k* слід враховувати, та 0 – у протилежному випадку; S_{ijk} – "внесок" у схожість об'єктів *i* та *j* ознаки *k*.

У сферах, де доводиться мати справу з бінарними даними, користуються також імовірнісними коефіцієнтами схожості – при створенні кластерів обчислюється "інформаційний виграш" об'єднання.

Міри схожості кластерів

Об'єднання у кластери (визначення міри схожості між двома кластерами, а не об'єктами) здійснюється за правилами об'єднання кластерів. Такими є, наприклад,

1) *метод найближчого сусіда* (*поодинокого зв'язку*) – відстань між кластерами визначається відстанню між двома найбільш близькими об'єктами у різних кластерах;
2) *метод найвіддаленішого сусіда* (*повного зв'язку*) – відстань між кластерами визначається найбільшою відстанню між двома довільними об'єктами у різних класах;
3) *метод незваженого попарного середнього* – відстань між кластерами визначається за середнім між усіма парами об'єктів в них;
4) *метод зваженого попарного середнього* – до середнього між усіма парами об'єктів в кластерах застосовується ваговий коефіцієнт (кількість об'єктів у кластері);
5) *незважений центроїдний метод* – за відстань між кластерами береться відстань між їхніми центрами ваги;
6) *метод Варда* – для оцінки відстаней використовує дисперсійний аналіз: за відстань між кластерами беруть приріст суми квадратів відстаней об'єктів до центрів кластерів, що отримується в результаті об'єднання.

Визначення відстані між кластерами: $d_{rs}=\alpha_p d_{ps}+\alpha_q d_{qs}+\beta d_{pq}+\gamma|d_{ps}-d_{qs}|$

Назва методу	α_p	α_q	β	γ
метод найближчого сусіда (Nearest neighbor)	1/2	1/2	0	–1/2
метод найвіддаленішого сусіда (Furthest neighbor)	1/2	1/2	0	1/2
Метод медіан – центром кластера є середнє усіх об`єктів (Median clustering)	1/2	1/2	–1/4	0
Середня відстань між кластерами (Between-groups linkage)	1/2	1/2	0	0
Середня відстань між усіма об'єктами пари кластерів з урахуванням відстаней всередині кластерів (Within-groups linkage)	$\frac{k_p}{k_p+k_q}$	$\frac{k_p}{k_p+k_q}$	0	0
Відстань між центрами кластерів (Centroid clustering)	$\frac{k_p}{k_p+k_q}$	$\frac{k_p}{k_p+k_q}$	$-\frac{k_p k_q}{k_p+k_q}$	0
Метод Варда (Ward`s method)	$\frac{k_r+k_p}{k_r+k_p+k_q}$	$\frac{k_r+k_p}{k_r+k_p+k_q}$	$-\frac{k_r}{k_r+k_p+k_q}$	0

Позначення методів кластеризації у статистичних пакетах

SPSS	***Statistica***	
Ієрархічний кластерний аналіз		
Cluster Method	***Amalgamation (linkage) rule***	***Методи об'єднання***
Between-group linkage	Weighted pair-group average	Міжгрупове зв'язування
Within-group linkage	Unweighted pair-group average	Внутрігрупове зв'язування

Nearest neightbor	Single linkage	Поодиноке зв'язування (найближчого сусіда)
Furthest neighbor	Complete linkage	Повне зв'язування
Centroid clustering	Unweighted pair-group centroid	Центроїдний метод
Median clustering	Weighted pair-group centroid (median)	Медіанний метод
Ward's method	Ward's method	Метод Варда
Measure	***Distance measure***	***Міри відстані***
Interval	***Для числових даних***	
Squared Euclidian distance	Squared Euclidian distance	Нормалізована Евклідова відстань
Euclidian distance	Euclidian distance	Евклідова відстань
Block	City-block (Manhattan) distances	Відстань міських кварталів
Chebyshev	Chebyshev distance metric	Відстань Чебишова
Minkowski	Power: SUM(ABS(x-y)**p)**1/r	Метрика Мінковського
	Percent disagreement	Відсоток відмінностей
Pearson correlation	1-Pearson r	1 – коефіцієнт кореляції за Пірсоном
Cosine		Косинусна відстань (1 – косинус кута між об'єктами)
Count	***Для номінативних даних***	
Chi-square measure		Міра Хі-квадрат
Phi- square measure		Міра фі-квадрат
Binary[18]	***Для дихотомічних (бінарних) даних***	
Jaccard		Міра Жаккара

Ітеративні методи кластерного аналізу

При великій кількості спостережень застосовують ітеративні методи кластерного аналізу: формування нових кластерів припиняють при досягненні певної їх кількості.

Найбільш популярним серед ітераційних методів є алгоритм k-середніх. В результаті його застосування буде побудовано k кластерів (потрібно попередньо мати гіпотезу про імовірну кількість кластерів) максимально віддалених один від одного.

На першому кроці алгоритму випадковим чином обирають k точок – центрів майбутніх кластерів. Усі об'єкти розподіляють по кластерах.

На наступних кроках ітераційного процесу центри кластерів переобчислюють з урахуванням віднесених до кластеру об'єктів, і об'єкти знову

[18] Вказано лише одну міру для дихотомічних даних з переліку, який має пакет SPSS.

перерозподіляють між кластерами. Процес припиняється, коли кластери стабілізуються, тобто після перерозподілу об'єкти залишаються у тих кластерах, де і були, або виконано максимальну кількість ітерацій.

Показником якості кластеризації є достовірна різниця середніх, обчислених для різних кластерів.

Метод k-середніх є простим, зрозумілим, прозорим, але надзвичайно чутливим до викидів (екстремальних даних) та повільним для великих баз даних.

Узагальненням ітеративних методів є методи нечітких середніх, у яких кожен кластер є нечіткою множиною, і кожен елемент належить різним кластерам з різним степенем приналежності. Серед нечітких методів можна назвати метод Fuzzy C-Means та метод кластеризації за Гюстафсоном-Кесселем [3]. Огляд деяких нових алгоритмів кластеризації та посилання на літературу про них можна знайти в [19].

Перспективним розвитком методів кластеризації є застосування теорії графів для удосконалення обчислювальних процедур та створення нуль-гіпотези для перевірки кількості кластерів у матриці схожості [17].

2.2. Завдання.

Примітка: Завдання виконати в одному з пакетів на вибір: KNIME, SPSS, RapidMiner або іншому.

Завдання 1: Ієрархічний кластерний аналіз

1. Підібрати дані для виконання кластерного аналізу (відомості про 20-30 об'єктів, результати вимірювань за 5-10 змінними).
2. Виконати ієрархічний кластерний аналіз: застосувати різні способи обчислення відстаней або різні способи об'єднання кластерів (не менше двох варіантів).
3. До звіту включити опис змінних, посилання на джерело даних, дендрограми.
4. Зробити висновки, проінтерпретувати отримані результати, оцінити (візуально) достатню кількість кластерів та достовірність отриманої кластеризації.

Завдання 2: Ітеративний кластерний аналіз k-середніх

1. Для даних з попереднього завдання визначити бажану кількість кластерів (наприклад, 2-4).
2. Виконати аналіз k середніх (k-means).
3. За даними таблиці кінцевих центрів кластерів побудувати лінійний графік середніх (профіль) для кожного кластера.
4. Проінтерпретувати отримані результати.
5. Порівняти процедури проведення та результати кластерного аналізу, виконаного двома методами. Зробити висновки.

Контрольні запитання

1. Опишіть призначення кластерного аналізу. Чому кластерний аналіз називають методом "класифікації без навчання"?
2. Назвіть основні групи методів кластерного аналізу та етапи кластеризації.

3. Як графічно представляють результати кластерного аналізу?
4. Які основні кроки алгоритмів агломеративних методів кластеризаццїї?
5. Які основні кроки алгоритмів дивізимних методів кластеризації?
6. Які основні кроки алгоритмів методу k-середніх?
7. Які є способи обчислення відстані між об'єктами?
8. Які є критерії об'єднання у кластери?
9. Як визначається якість кластеризації?

Вправи

1. Два об'єкта представлено кортежами (22, 1, 42, 10) та (20, 0, 36, 8). Обчисліть для них
   а) евклідову відстань;
   б) манхеттенську відстань;
   в) відстань Мінковського, використовуючи q=3. [22, 2.6]
2. Дано координати на площині восьми точок, які необхідно розбити на три кластери методом k-середніх: (2, 10), (2, 5), (8, 4), (5, 8), (7, 5), (6, 4), (1, 2), (4, 9).
   Нехай початковими центрами кластерів визначено точки (2, 10), (5, 8), (1, 2). Використовуючи Евклідову відстань на площині знайдіть: [16], [22, 10.2]
   а) центри трьох кластерів після виконання першої ітерації в методі k-середніх;
   б) результуючі три кластери.

2.3. Приклади виконання завдання у різних пакетах

Приклад 1: Виконаємо ієрархічний (агломеративний) кластерний аналіз для поданих нижче даних, які є результатами психологічного тестування (Таблиця 2.1).

Таблиця 2.1

код	лідерство	впевненість	вимогливість	скептицизм	поступливість	довірливість	добросердя	чуйність
1	2	2	3	3	10	10	8	7
2	15	9	8	8	8	7	13	11
3	9	5	9	9	8	3	9	4
4	5	5	9	9	3	9	8	4
5	8	6	10	10	7	7	9	7
6	1	4	2	2	9	7	8	9
7	7	4	5	15	15	14	12	12
8	10	10	10	10	15	13	14	14
9	11	7	9	9	7	6	8	6
10	11	4	9	9	9	8	9	10
11	10	8	8	8	9	11	11	12
12	6	2	8	8	8	9	12	9
13	11	9	6	6	5	5	10	4
14	4	5	4	4	9	10	10	9

Тут значення від 0 до 4 балів означають низький, а 5-8 балів – помірний рівень вираженості якості, тобто в цілому адаптивну поведінку. Бали 9-12 – високий рівень, тобто екстремальну поведінку, а бали 13-16 – поведінку екстремальну до патології. Наприклад, значення за шкалою лідерство можна інтерпретувати як «*тенденцію до лідерства-владність-деспотизм*», за шкалою впевненість як «*впевненість у собі-самовпевненість-самозакоханість*», за шкалою вимогливість – «*вимогливість-непримиримість-жорстокість*», за шкалою скептицизм – «*скептицизм-упертість-негативізм*», за шкалою поступливість – «*поступливість-покірливість-пасивне підкорення*», за шкалою довірливість – «*довірливість-слухняність-залежність*», за шкалою добросердя – «*добросердя-несамостійність-надмірний конформізм*», за шкалою чуйність (альтруїзм) – «*співчутливий, турботливий-безкорисливий-жертвенний*»[19].

Виконання у пакеті SPSS

У пакеті SPSS в основному діалоговому вікні кластерного аналізу потрібно вибрати змінні, за якими буде здійснюватися кластеризація (**Рис. 2.1**), та у допоміжних вікнах – методи обчислення відстаней і об'єднання кластерів (Рис. 2.2):

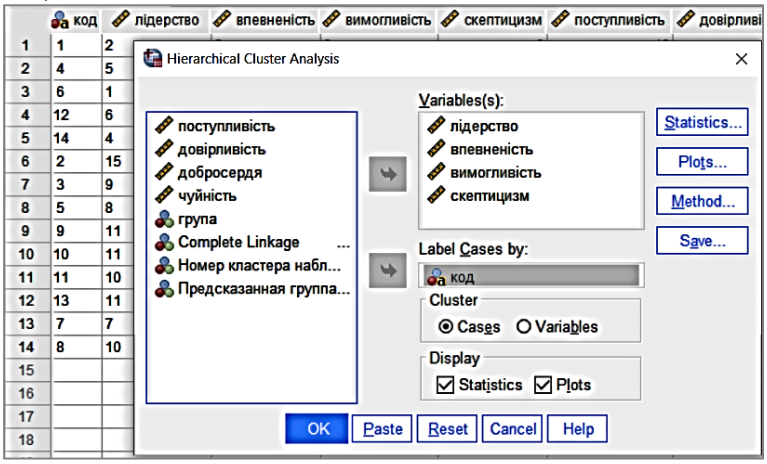


Рис. 2.1

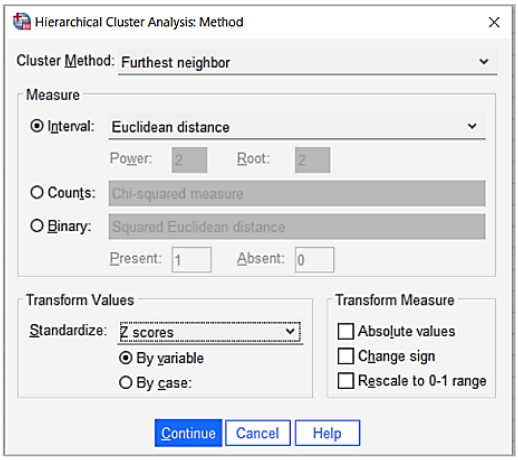


Рис. 2.2

Результатами кластерного аналізу будуть таблиця «Порядок агломерації» та бурулькова діаграма, що показує кількість кластерів, до складу яких входить об'єкт (Рис. 2.3а, Рис. 1.3б):

Agglomeration Schedule

Stage	Cluster Combined Cluster 1	Cluster 2	Coefficients	Stage Cluster First Appears Cluster 1	Cluster 2	Next Stage
1	9	11	2,000	0	0	5
2	7	8	2,000	0	0	3
3	7	10	2,236	2	0	5
4	1	3	2,646	0	0	10
5	7	9	2,828	3	1	6
6	7	12	3,162	5	0	7
7	7	14	3,464	6	0	8
8	2	7	3,464	0	7	9
9	2	4	3,464	8	0	11
10	1	5	3,873	4	0	12
11	2	6	4,690	9	0	12
12	1	2	6,708	10	11	13
13	1	13	7,416	12	0	0

Рис. 2.3а

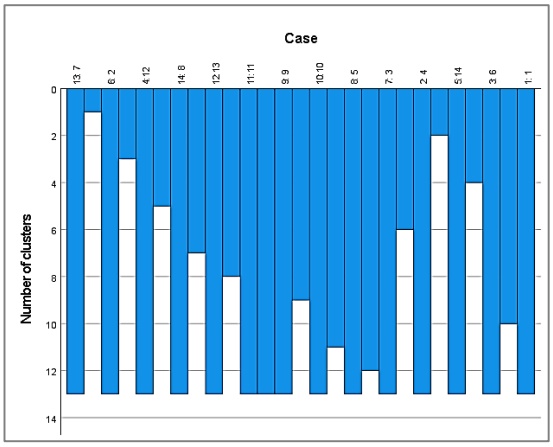


Рис. 2.3б

[19] https://ru.wikipedia.org/wiki/Тест_Лири

Найбільш наочним результатом кластерного аналізу є дендрограма. На **Рис. 2.4** видно, що метод найближчого сусіда не дозволяє виділити окремі кластери – проявляється основний недолік даного методу: “ланцюговий ефект”, тобто утворення великого довгастого кластера, до якого один за одним додаються нові об’єкти. У такому разі кластеризацію слід повторити з іншими методами об’єднання. Наприклад, при застосуванні методу найдальшого сусіда кластерна структура більш приваблива (Рис. 2.5):

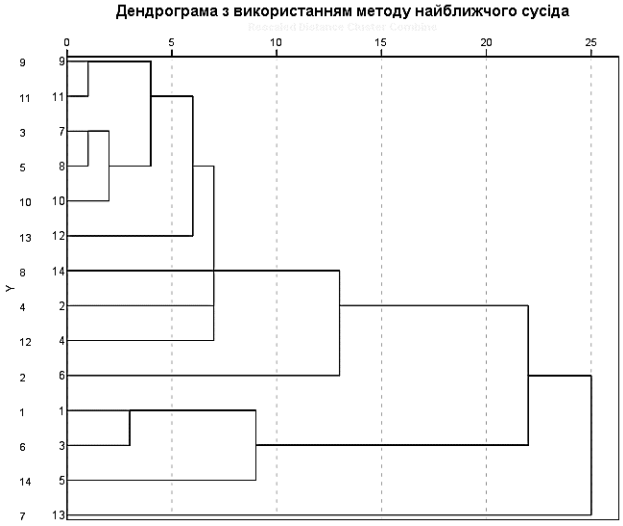


Рис. 2.4

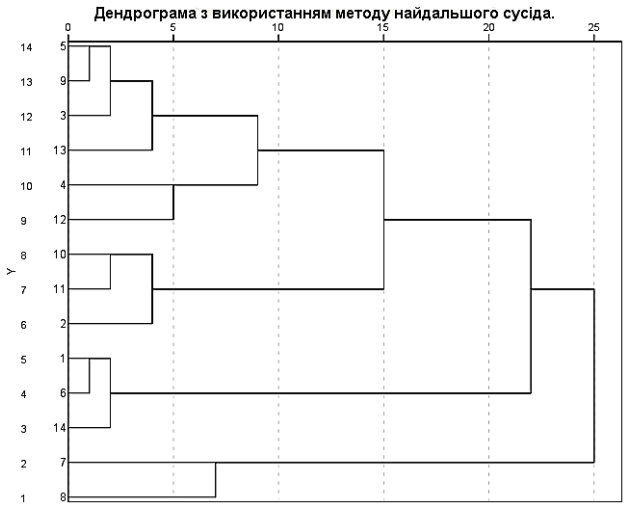


Рис. 2.5

Якщо застосувати до тих самих даних аналіз k-means, то цікавим результатом, крім розподілу об’єктів за кластерами, буде графічна інтерпретація центрів кластерів (Рис. 2.6):

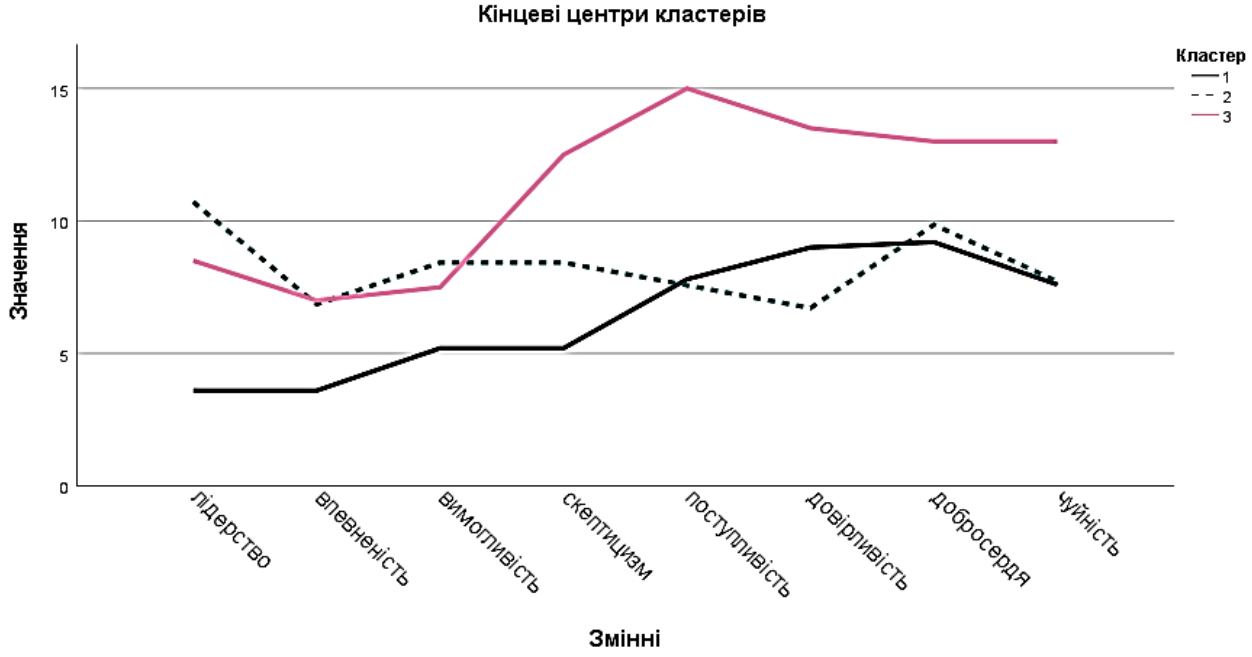


Рис. 2.6

За цим графіком видно, що у представників першого кластеру перші чотири якості, умовно назвемо їх «твердість», виражені слабо, а другі чотири, назвемо їх «м’якість», – виражені середньо. У представників другого кластеру досить яскраво виражені як якості першої групи, так і другої. А третій кластер відзначається екстремальними значеннями «м’яких» якостей.

Виконання кластерного аналізу у пакеті Rapid Miner

Для тих самих даних, збережених у форматі SPSS, виконаємо агломеративний кластерний аналіз у пакеті RapidMiner. Для цього необхідно створити процес, як показано на Рис. 1.16, з операторів 1) **Read SPSS** – імпорту даних з SPSS за допомогою оператора; 2) **Remote Attribute Range** – для вилучення стовпчика номер 10 зі змінною, в якій було збережено приналежність об'єктів досліджуваної множини до відповідного кластера; 3) **Agglomerative Clustering** – оператора кластерного аналіза. Налаштування для оператора кластерного аналізу показано на Рис. 2.7.

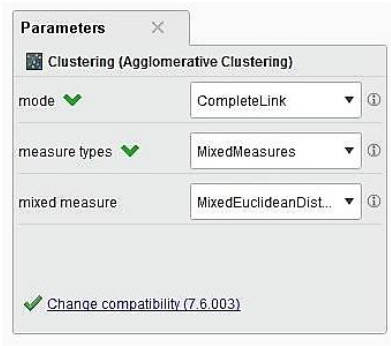


Рис. 2.7

В результаті буде отримано 27 кластерів, які можна представити у вигляді ієрархії папок, дендрограми (Рис. 2.8) та графа (Рис. 2.9):

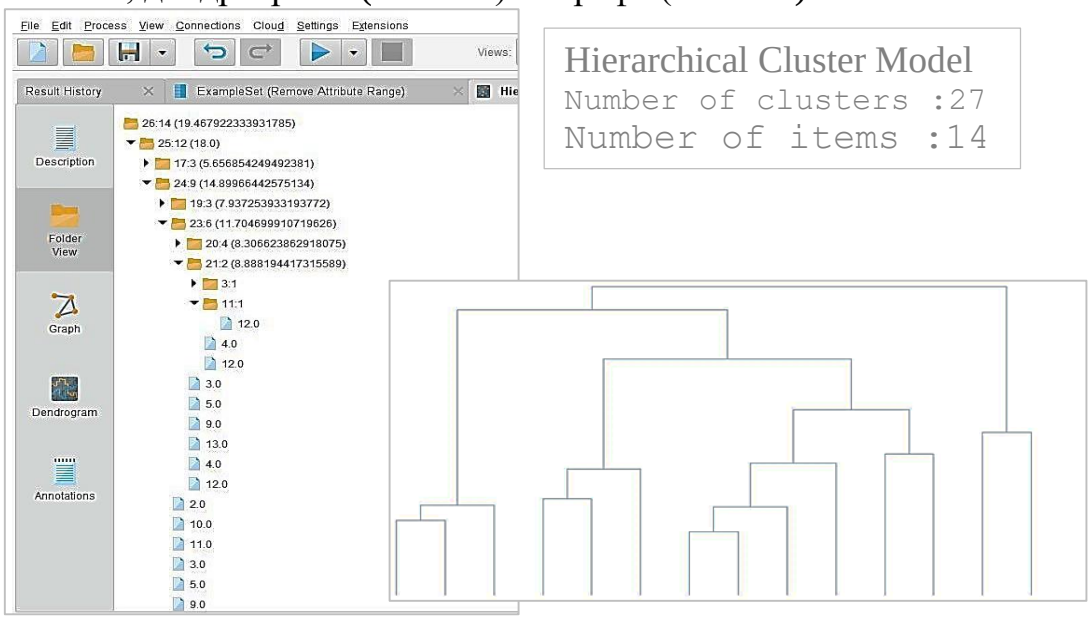


Рис. 2.8

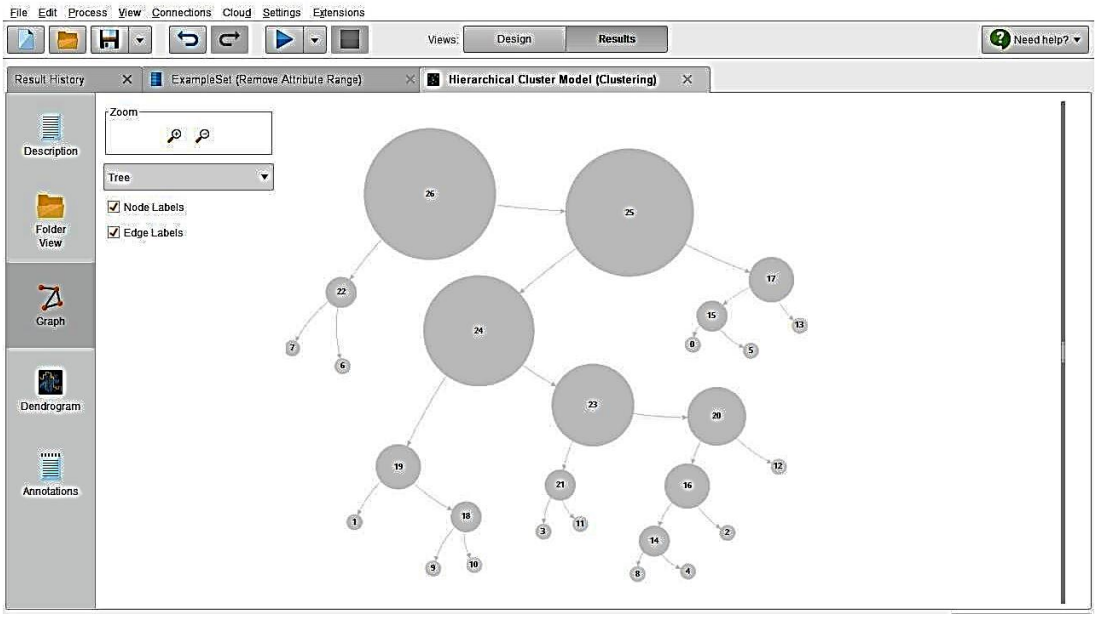


Рис. 2.9

Для виконання кластерного аналізу *k-середніх* у процесі слід замінити один оператор. Добавимо після нього оператор Write Excel, щоб зберегти результати кластеризації (Рис. 2.10). На рисунку видно, які параметри встановлено для оператора кластеризації.

Після виконання процесу буде створено файл формату xls або xlsx такого виду:

КОД	V1	V2	V3	V4	V5	V6	V7	V8	id	cluster
1.0	2,0	2,0	3,0	3,0	10,0	10,0	8,0	7,0	1,0	cluster_2
2.0	15,0	9,0	8,0	8,0	8,0	7,0	13,0	11,0	2,0	cluster_1
3.0	9,0	5,0	9,0	9,0	8,0	3,0	9,0	4,0	3,0	cluster_1
…	…	…	…	…	…	…	…	…	…	…

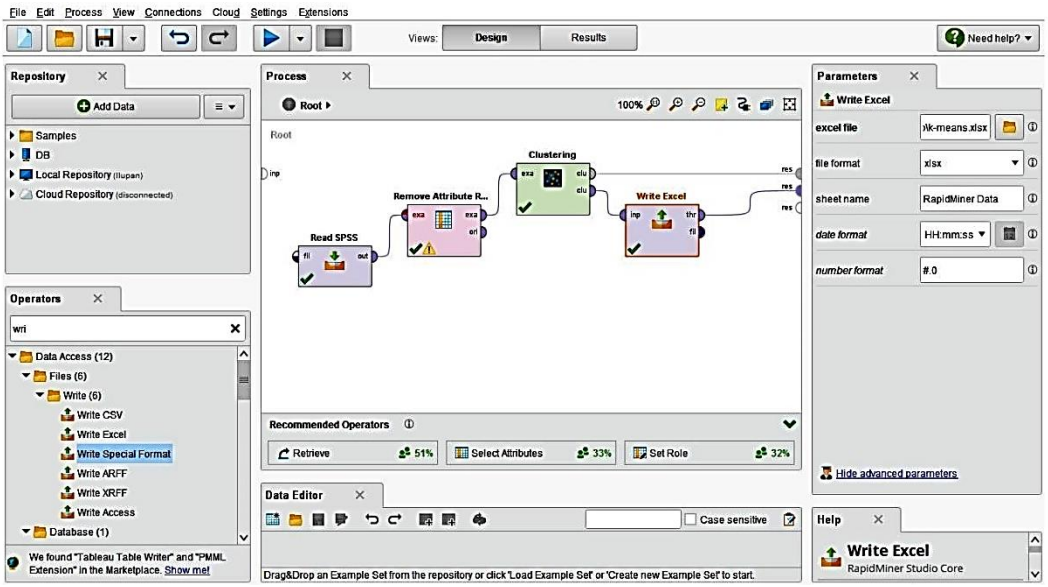


Рис. 2.10

На графіках об’єкти, що потрапили до різних кластерів, будуть зображені точками різного кольору. На тривимірній точковій діаграмі (Scatter 3D) множину об’єктів можна візуалізувати у тривимірному просторі обраних змінних (Рис. 2.11).

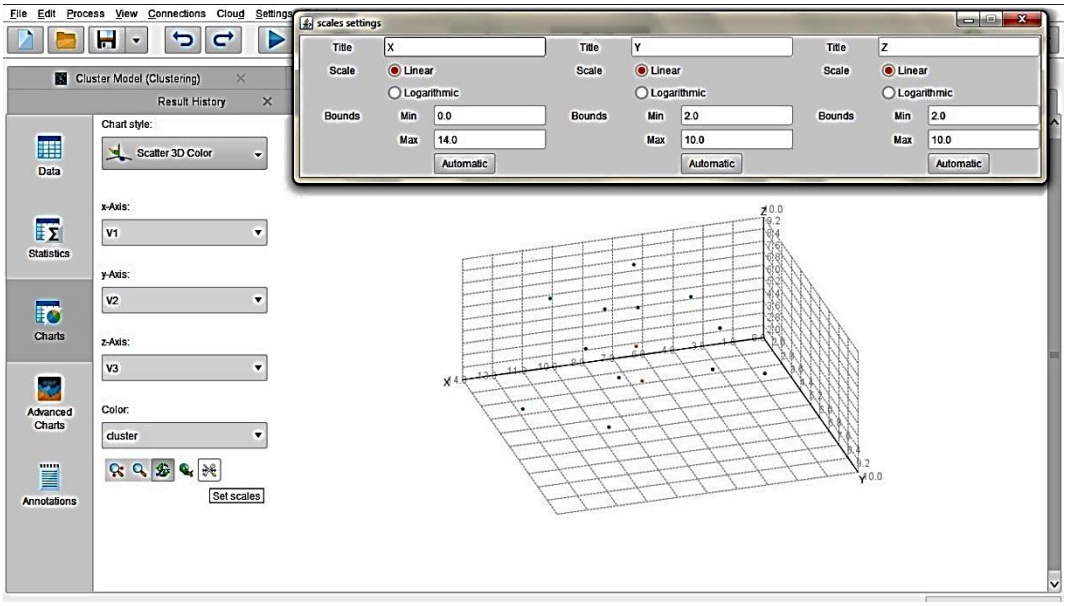


Рис. 2.11

Виконання кластерного аналізу у пакеті KNIME

Виконаємо кластерний аналіз для тих самих даних у пакеті KNIME. Для цього створимо новий проект та сформуємо потік робіт як показано на Рис. 2.12:

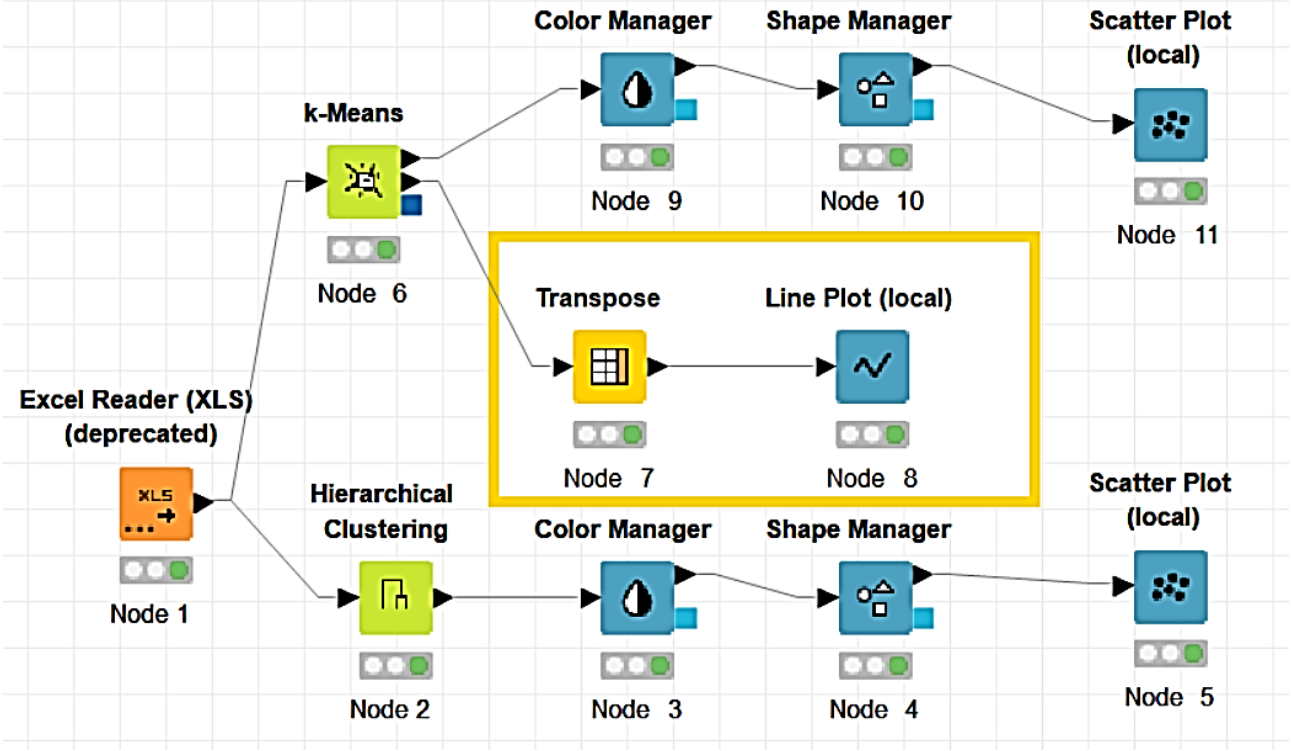


Рис. 2.12

Спочатку потрібно отримати дані, наприклад, з електронної таблиці (вузол Excel Reader, який слід конфігурувати на конкретний файл). Отримані дані передаються процедурам кластерного аналізу, відповідно вузлам Hierarchical Clustering (вузол 2) та k-Means (вузол 6).

Конфігурування вузла Hierarchical Clustering полягає у тому, що слід обрати змінні, які будуть використані для обрахунку відстаней між об'єктами (Include), кількість кластерів, приналежність до яких буде зафіксовано у результатах аналізу, та способи визначення відстані між об'єктами (Euclidean, Manhattan) та об'єднання у кластери (Single Linkage, Complete Linkage, Average Linkage) (Рис. 2.13).

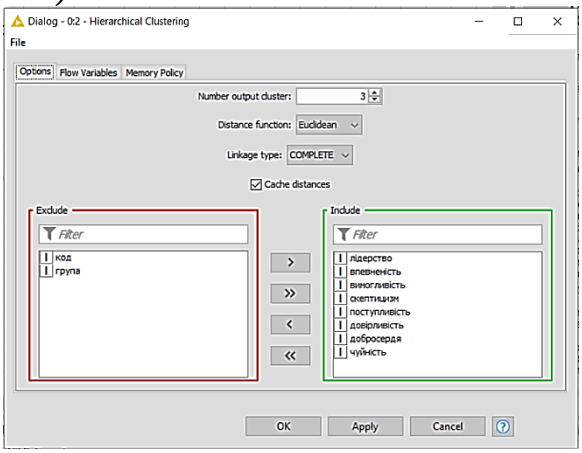


Рис. 2.13

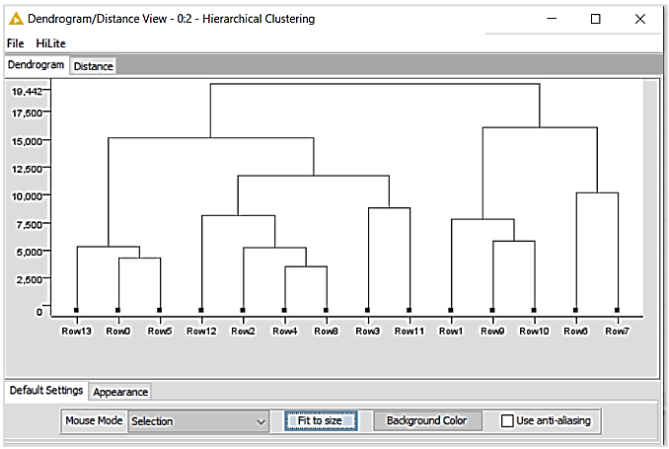


Рис. 2.14

Результатом ієрархічного аналізу буде дендрограма (Рис. 2.14).

Крім того на вкладці Distance виводиться графік, за допомогою якого можна оцінити оптимальну кількість кластерів: він відображає відстані між кластерами для кожної їх кількості. У місцях злиття різних груп на графіку будуть раптові стрибки рівня подібності. Координата *y* – це відстань злиття, координата *x* – номер злиття, тобто рівень ієрархії.

На жаль, у випадках, коли графік спадає плавно, визначити кількість кластерів візуально буде складно або і неможливо. Так для даних, аналізованих у прикладі, графік досить плавний, можна помітити лише невеликий злам на другому рівні, на якому в дендрограмі спостерігається два кластера (Рис. 2.15):

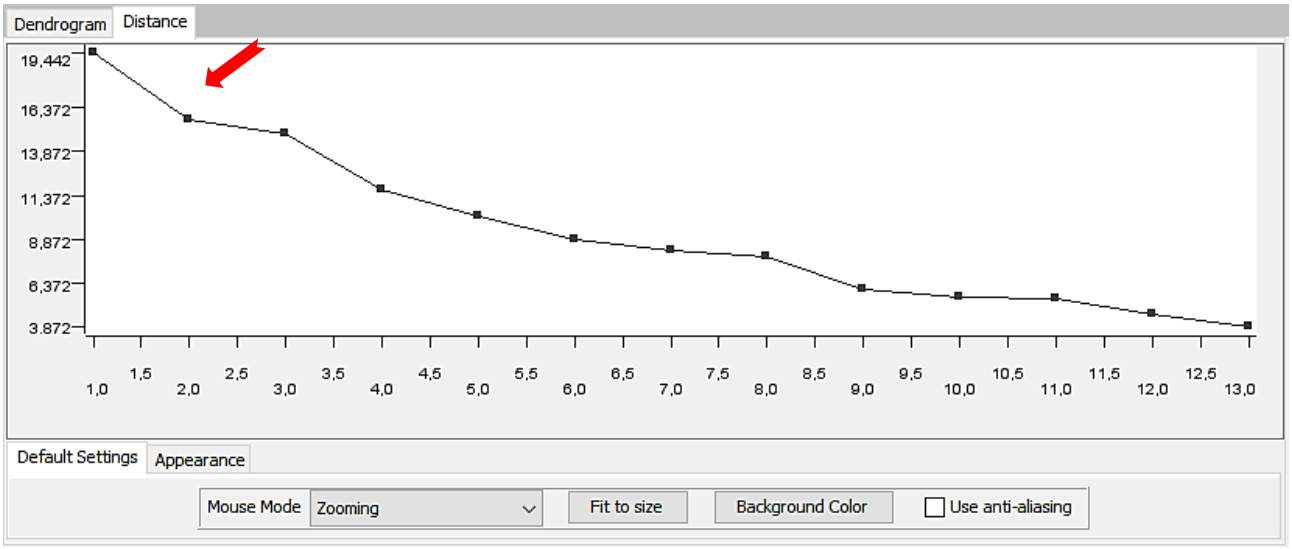


Рис. 2.15

У налаштуванні аналізу k-means слід вказати змінні, використовувані при обрахунках, та кількість кластерів. Результати кластеризації тут можуть бути представлені лінійним графіком профілів кластерів (Linea Plot) (Рис. 2.16), який, однак, потребує транспонування таблиці координат центрів кластерів.

За отриманим графіком видно, що до кластера_2 віднесено елементи з помірними значеннями за усіма змінними (назвемо цей кластер «помірні»). У кластері_1 знаходяться елементи з помірними значеннями за першими чотирма змінними та високими значеннями за останніми чотирма змінними (назвемо їх «добросерді»). А до кластера_0 – елементи з низькими значеннями перших чотирьох змінних та помірними – для останніх чотирьох змінних.

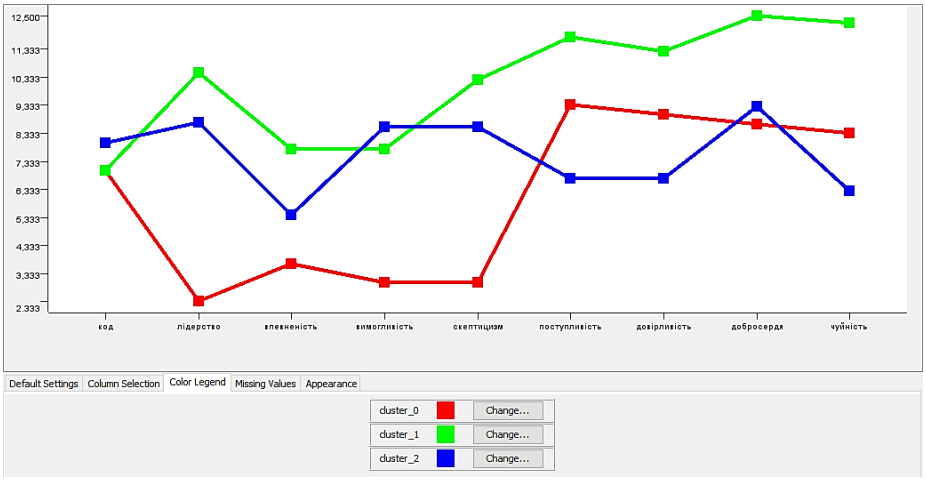


Рис. 2.16

Хотілося б отримати графічне зображення кластерів. Однак, спочатку слід зменшити розмірність: зобразити точки у 8-вимірному просторі неможливо. Зниження розмірності за допомогою методів дискримінантного та факторного аналізу буде здійснено у наступних темах.

Але на даному етапі аналізу можна побачити розподіл елементів кластерів у двовимірному просторі, визначеному обраними змінними. Для цього спочатку слід скористатися вузлами Color Manager та Shape Manager, а потім побудувати діаграму розсіювання (Scatter Plot) як за результатами ієрархічного кластерного аналізу (вузли 3, 4, 5), так і за результатами аналізу k-середніх (вузли 9, 10, 11). Відповідні діаграми розсіювання у площині змінних *лідерство-чуйність* представлено на Рис. 2.17 та Рис. 2.18 відповідно:

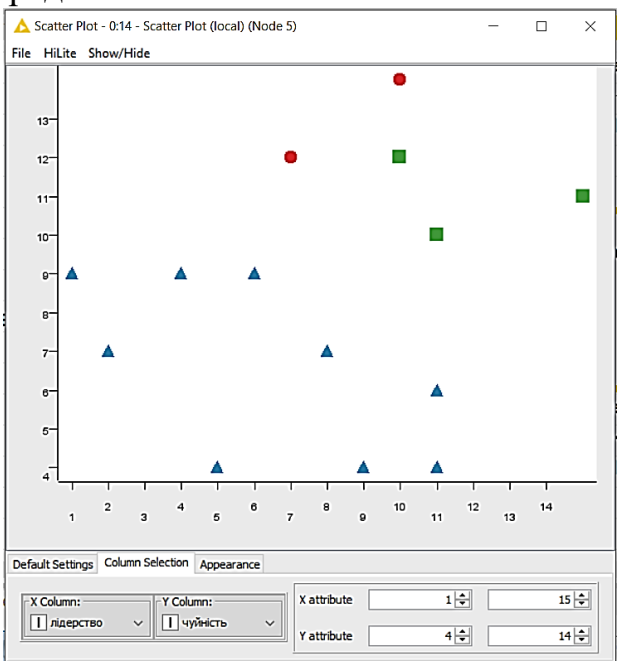


Рис. 2.17

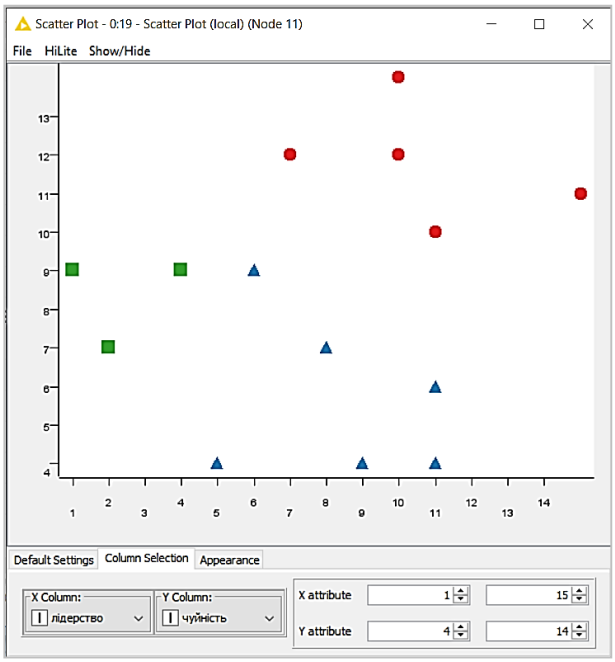


Рис. 2.18

2.4. Література до теми:

При підготовці даного розділу були використані навчальні посібники і підручники [3, 6, 8, 11, 16, 17, 19, 21, 22] (згідно нумерації загального списку – див. с. 93).

- ✓ Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP: пособие / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2007. – 384 с.
- ✓ Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: 2003.
- ✓ Інтелектуальний аналіз даних: підручник /О.І.Черняк, П.В.Захарченко. – К.: Знання, 2014.
- ✓ Лупан І.В., Авраменко О.В., Акбаш К.С. Комп'ютерні статистичні пакети: навчально-методичний посібник. – 2-е вид. – Кіровоград, «КОД», 2015. – 236 с.

- ✓ Степанов Р.Г. Технология Data Mining: Интеллектуальный анализ данных. – Казань, 2008. – 58 с. (pdf)
- ✓ Факторный, дискриминантный и кластерный анализ/ Дж.-О.Ким, Ч.У.Мьюллер, У.Р.Клекка и др. – М.,1989. – 215 с.
- ✓ Чубукова И.А. Data Mining (Курс лекций) – URL: http://kist.ntu.edu.ua/textPhD/Chubukova-Data_Mining.pdf
- ✓ Шитиков В. К., Мастицкий С. Э. (2017) Классификация, регрессия, алгоритмы Data Mining с использованием R. – URL: https://github.com/ranalytics/data-mining
- ✓ Han J., Kamber M., Pei J. Data mining: Concepts and Techniques. – Morgan Kaufmann Publishers. – 2012. – 740 p.

ТЕМА 3. ДИСКРИМІНАНТНИЙ АНАЛІЗ

3.1. Теоретичні відомомсті.

Призначення та основні припущення дискримінантного аналізу

Задачі дискримінантного аналізу полягають у тому, щоб визначити правила, які б дозволили за значенням дискримінантних (метричних) змінних віднести кожен з досліджуваних об'єктів (в тому числі об'єктів, класова приналежність яких невідома) до одного з відомих класів, та визначити "вагу" кожної дискримінантної змінної у такому поділі на класи. Тобто дискримінантний аналіз – це група статистичних процедур, які відносять до методів "класифікації з навчанням" або "розпізнавання образів". В результаті виконання дискримінантного аналізу можна отримати корисні відомості про окремі об'єкти, про відмінності між класами та про здатність змінних як таких розрізняти класи.

Початкові припущення дискримінантного аналізу

Позначимо через g кількість класів, p – кількість дискримінантних змінних, n_i – кількість об'єктів у класі i, N – загальну кількість об'єктів. Тоді у моделі дискримінантного аналізу має бути [17]:

- кількість класів не менше двох: $g \geq 2$;
- принаймні два об'єкти у кожному класі: $n_i \geq 2$;
- довільна кількість дискримінантних змінних, за умови, що їх не більше ніж кількість об'єктів мінус 2: $0 < p < (N–2)$;
- вимірювання дискримінантних змінних за шкалою не нижче інтервальної;
- лінійна незалежність дискримінантних змінних;
- приблизна рівність між коваріаційними матрицями для кожного класу;
- багатовимірна нормальність закону розподілу дискримінантних змінних для кожного класу.

Геометрична інтерпретація методу

Геометрична інтерпретація методу полягає у тому, що кожен об'єкт дослідження можна представити у *p*-вимірному просторі дискримінантних змінних точкою. Точки, координатами яких є середні значення дискримінантних змінних для класу, називають "*центроїдами*" класів. Об'єкт приписують до того класу, до центроїда якого він найближчий.

Для того, щоб розрізняти відносне положення центроїдів достатньо розмірності, на одиницю меншої за кількість класів. Отже одним із завдань дискримінантного аналізу є перехід від *h*-вимірного простору дискримінантних змінних до (*g–1*)-вимірного простору канонічних дискримінантних функцій. Початком координат нового простору є головний центроїд – точка, координатами якої є середні значення усіх дискримінантних змінних. Першу канонічну вісь орієнтують у напрямку, в якому усі центроїди класів розрізняються максимально, другу – перпендикулярно до першої також у напрямку максимальної відмінності класів (центроїдів) і т.д.

Кожна канонічна функція є лінійною комбінацією дискримінантних змінних і має вигляд:

$$f_{km}=u_0+u_1X_{1km}+u_2X_{2km}+\ldots+u_pX_{pkm},$$

де f_{km} – значення канонічної дискримінантної функції для m-ного об'єкта у групі k; X_{ikm} – значення дискримінантної змінної X_i для m-ного об'єкта у групі k; u_i – канонічні коефіцієнти.

Відстані між об'єктами та центроїдами класів визначають за мірою Махаланобіса, для якої необхідне виконання умови рівності для усіх класів коваріаційних матриць:

$$D^2(X|G_k)=(N-g)\sum_{i=1}^{p}\sum_{j=1}^{p}a_{ij}(X_i-X_{ik})(X_j-X_{jk}),$$

де $D^2(X|G_k)$ – квадрат відстані від даного об'єкта X до центроїда класу k.

Про критерії оцінки якості класифікації та особливості виконання дискримінантного аналізу у пакеті SPSS див. [11].

3.2. Завдання.

Завдання виконати в одному з пакетів на вибір: KNIME, SPSS, RapidMiner або іншому.

1. Ознайомитися з прикладами застосування дискримінантного аналізу у [17] або в інших джерелах[20].
2. Використати дані, для яких було виконано кластерний аналіз, наприклад, дані, наведені у додатку, або у рейтинговій таблиці університетів України[21].
3. Виконати дискримінантний аналіз. Побудувати за результатами аналізу діаграму розсіювання елементів кластерів у просторі визначених канонічних функцій.
4. Проінтерпретувати отримані результати.
5. Проаналізувати результати дискримінантного аналізу. Зробити висновки про якість прогнозування значень групуючої змінної від незалежних: яка із незалежних змінних має найбільший вплив на класифікацію? Чи всі обрані змінні мають вплив на групуючу змінну? Наскільки точний отриманий прогноз? Чи можна в подальшому класифікувати об'єкти лише за допомогою обраних змінних? Спробувати проінтерпретувати отримані канонічні функції.

Контрольні запитання

1. Опишіть призначення дискримінантного аналізу. Чому дискримінантний аналіз називають методом "класифікації з навчанням"?
2. Що таке "центроїд"? В чому полягає геометрична інтерпретація правила інтерпретації?

---

[20] Наприклад: Пустовойт В.И., Самойлов А.С., Ключников М.С. Скрининг-диагностика функционального состояния спортсменов-дайверов с преобладанием автономного типа регуляции. //Медицина экстремальных ситуаций. 2019: 21 (2), С. 320-328. – URL: https://cyberleninka.ru/article/n/skrining-diagnostika-funktsionalnogo-sostoyaniya-sportsmenov-dayverov-s-preobladaniem-avtonomnogo-tipa-regulyatsii/viewer

[21] Рейтингова таблиця вищих навчальних закладів «Топ-200 Україна» — 2021 рік. – URL: http://www.euroosvita.net/index.php/?category=49&id=6867

3. Як будуються канонічні дискримінантні функції? Скільки їх можна побудувати?
4. Критеріями чого є “власне значення”, λ-Уілкса та χ^2? Якими повинні бути їхні значення?
5. Що показують структурні коефіцієнти канонічних функцій? Що показує кореляція визначених канонічних функцій з незалежними змінними?
6. Що таке “толерантність”? Які значення вона може мати? При яких її значеннях змінну слід вилучити з аналізу? Чому (як проінтерпретувати вилучення)?
7. Показником чого є статистика F-вилучення?
8. Які основні результати дискримінантного аналізу?
9. Як можна оцінити точність прогнозу, отриманого в результаті дискримінантного аналізу?
10. Як можна графічно представити результати дискримінантного аналізу?

3.3. Приклади виконання завдання

Виконання дискримінантного аналізу у пакеті KNIME

Виконаємо дискримінантний аналіз для тих даних, які були використані у кластерному аналізі. У прикладі нам довелося взяти файл з більшою кількістю даних, оскільки у пакеті KNIME є обмеження: розмір найменшої групи об’єктів має бути більшим за кількість дискримінантних змінних, тобто $n_i \geq p$. З цією ж метою були змінені налаштування ієрархічного кластерного аналізу: замість евклідової відстані (Euclidean) використано манхеттенську (Manhattan) та встановлено 3 кластери для виведення (у вихідній таблиці дані буде розподілено на три кластери-групи). Використані дані наведено у Таблиця 3.1 (стор. 53).

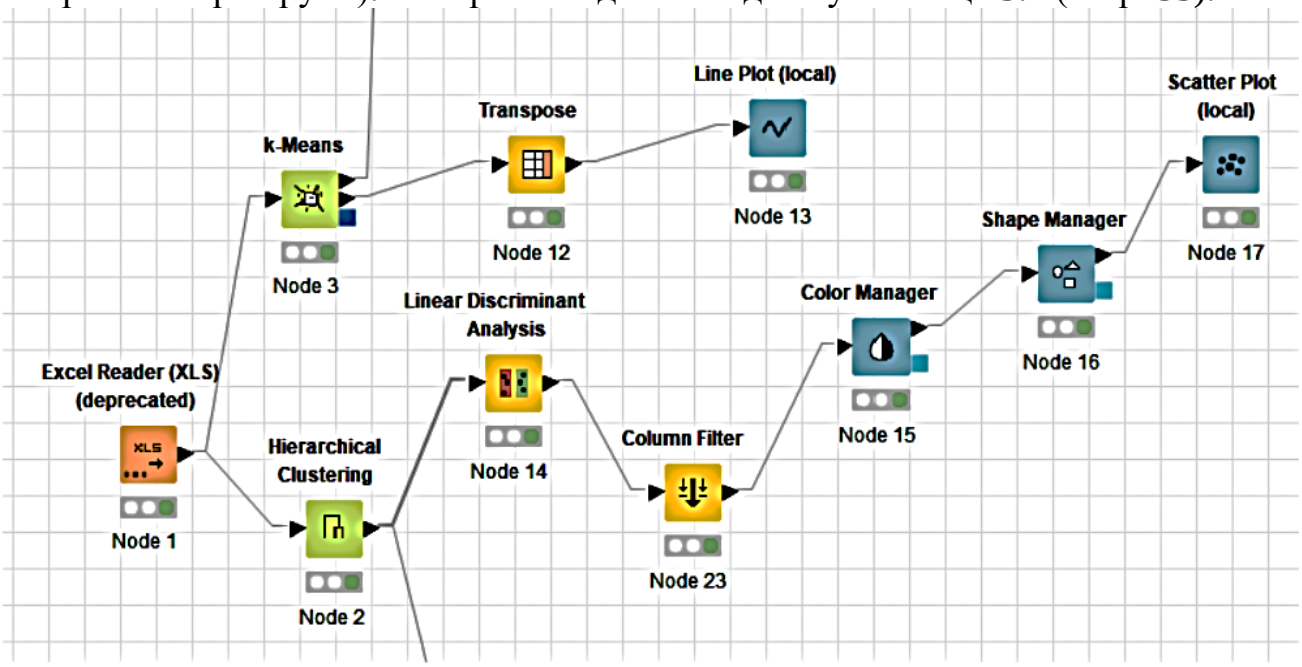


Рис. 3.1

Для виконання дискримінантного аналізу доповнимо створений раніше проект новими вузлами, як показано на Рис. 3.1 та встановимо налаштування: вкажемо цільову розмірність – 2 (кількість результуючих канонічних дискримінантних функцій має бути на одиницю меншою за кількість визначених

груп об’єктів), та оберемо групуючу змінну (номінативну; у даному випадку вона одна – це приналежність до кластеру).

В результаті виконання кластерного аналізу отримаємо таку дендрограму (Рис. 3.2):

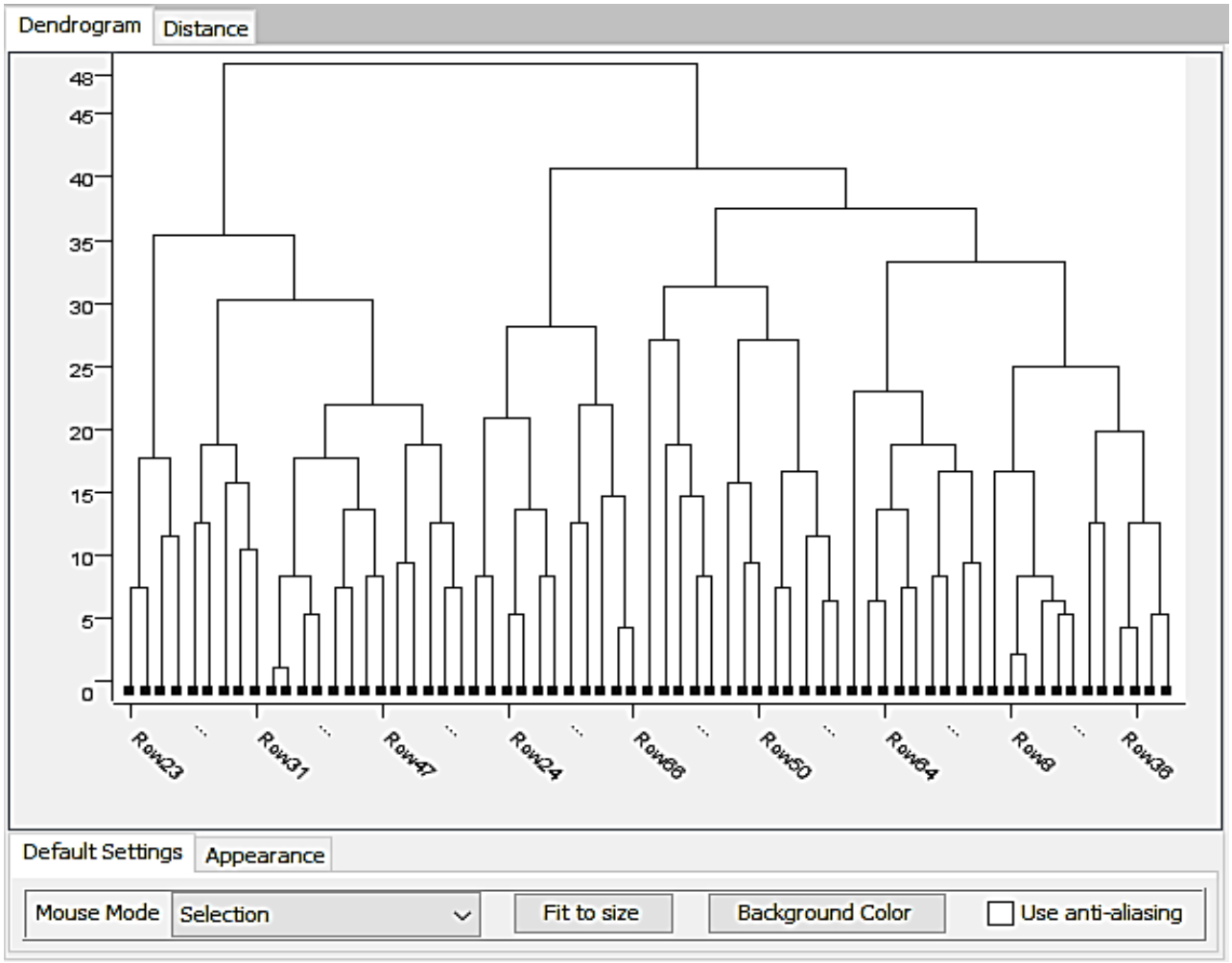


Рис. 3.2

Допоміжні вузли Color Manager та Shape Manager потрібні для налаштування графічного подання об’єктів різних груп на діаграмі розсіювання, а вузол Column Filter – для того, щоб відсіяти зайві змінні (слід залишити тільки отримані канонічні дискримінантні функції). Сама діаграма буде мати такий вигляд як показано на Рис. 3.3.

Як бачимо, об’єкти, що належать різним кластерам досить непогано розрізняються, хоча границі кластерів не надто чіткі і мають перетини.

У пакеті SPSS для тих самих даних діаграма розсіювання матиме такий самий вигляд (Рис. 3.4).

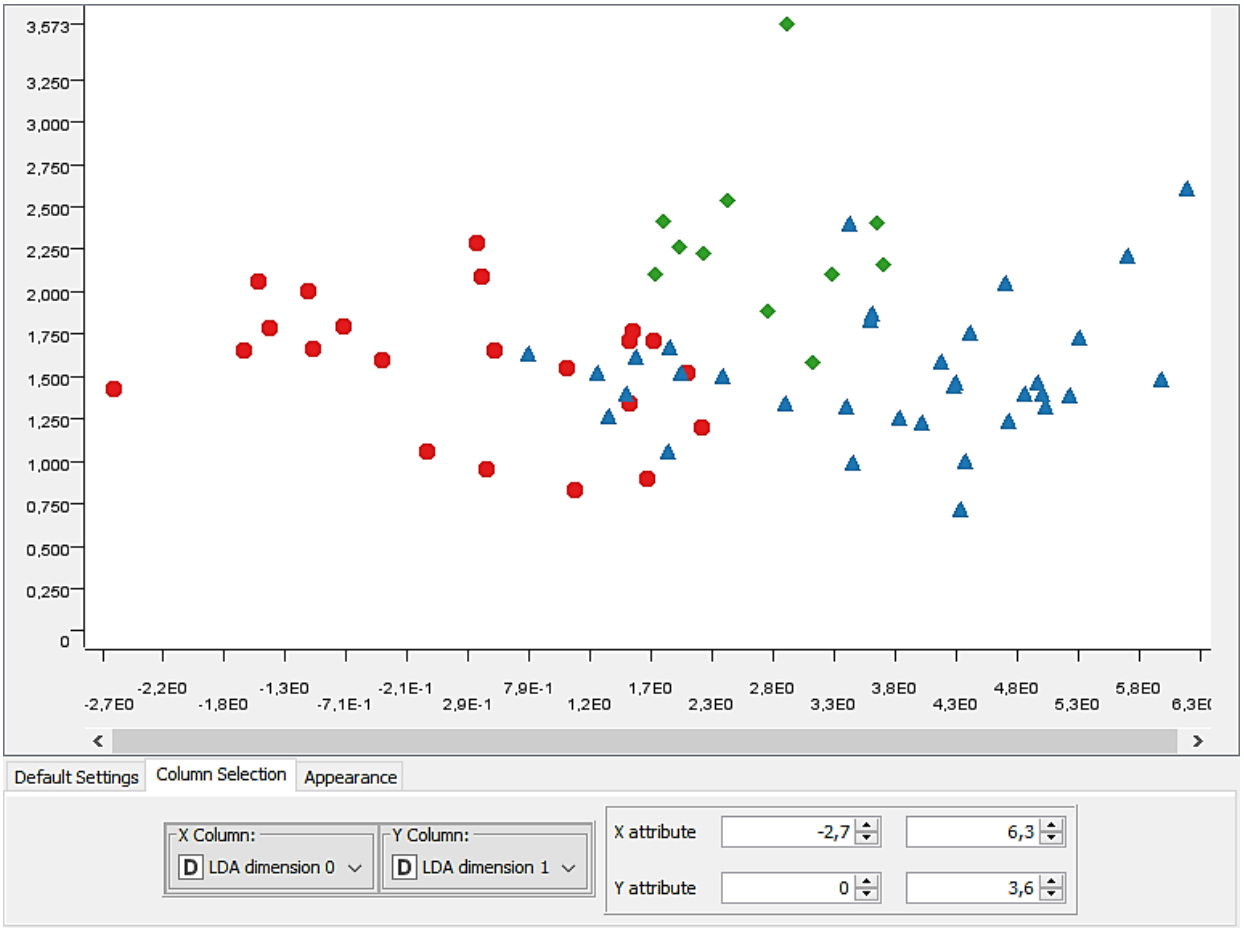

3,573
3,250
3,000
2,750
2,500
2,250
2,000
1,750
1,500
1,250
1,000
0,750
0,500
0,250
0
-2,7E0
-2,2E0
-1,8E0
-1,3E0
-7,1E-1
-2,1E-1
2,9E-1
7,9E-1
1,2E0
1,7E0
2,3E0
2,8E0
3,3E0
3,8E0
4,3E0
4,8E0
5,3E0
5,8E0
6,3E
Default Settings
Column Selection
Appearance
X Column:
LDA dimension 0
Y Column:
LDA dimension 1
X attribute
-2,7
6,3
Y attribute
0
3,6

Рис. 3.3

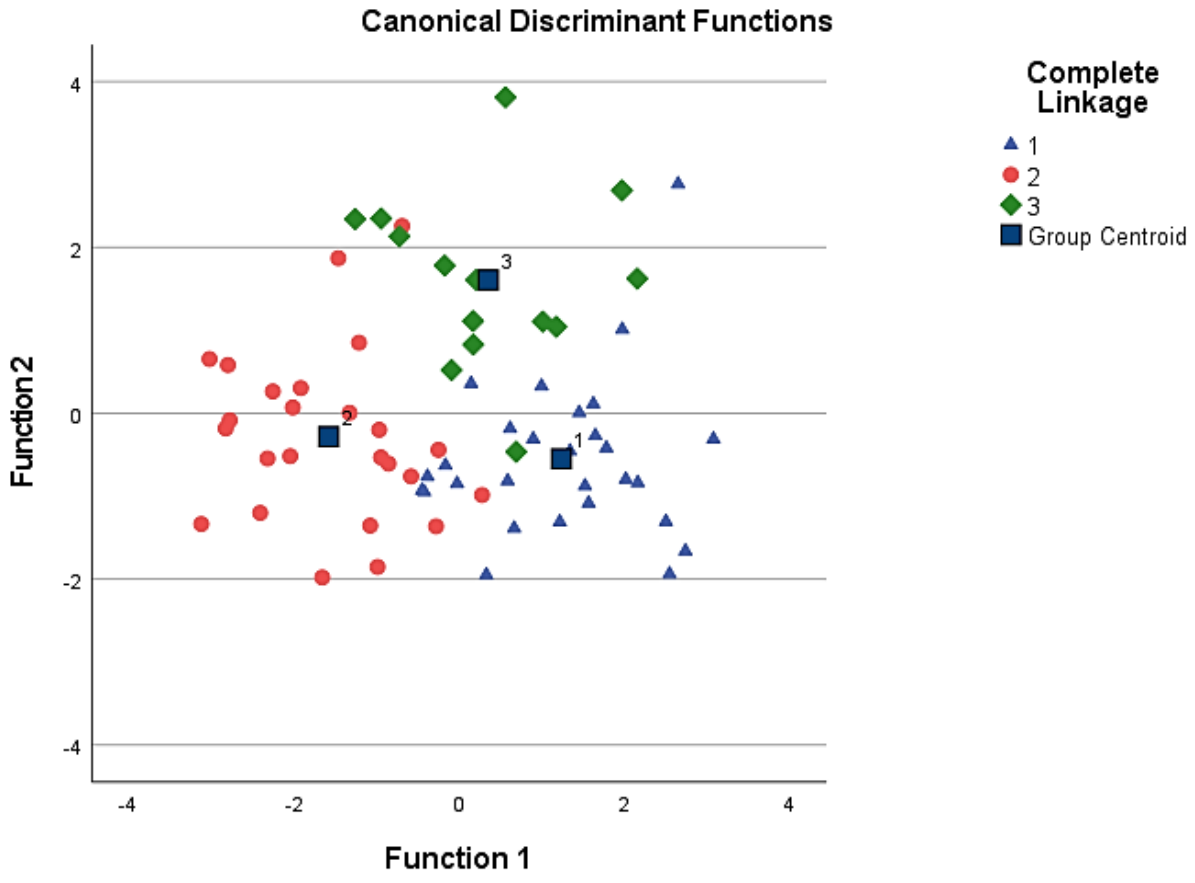

Canonical Discriminant Functions
Complete Linkage
1
2
3
Group Centroid
Function 2
Function 1
4
2
0
-2
-4
-4
-2
0
2
4

Рис. 3.4

Однак, крім діаграми SPSS дає таблиці коефіцієнтів канонічних дискримінантних функцій, структурну матрицю, яка показує кореляцію кожної з вхідних змінних з утвореними канонічними, координати центроїдів:

Standardized Canonical Discriminant Function Coefficients		
	Function	
	1	2
лідерство	,230	-,046
впевненість	-,164	,795
вимогливість	-,731	-1,304
скептицизм	,466	,962
поступливість	,742	-,412
довірливість	,104	,859
добросердя	,298	,035
чуйність	,264	,118

Structure Matrix		
	Function	
	1	2
поступливість	,777*	-,086
добросердя	,503*	,106
вимогливість	-,368*	-,083
скептицизм	-,365*	-,048
чуйність	,286*	-,087
довірливість	,407	,562*
впевненість	-,260	,548*
лідерство	,061	,208*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

Functions at Group Centroids		
	Function	
Complete Linkage	1	2
1	1,236	-,551
2	-1,581	-,283
3	,351	1,606

Unstandardized canonical discriminant functions evaluated at group means

Цікавим результатом дискримінантного аналізу в SPSS є таблиця результатів класифікації, у якій підраховано відсоток елементів первинних груп, які залишаться в них згідно із обчисленими канонічними функціями. По ній видно, що тільки 86,6% об'єктів залишилося у кластерах, визначених кластерним аналізом.

Classification Results[a]						
			Predicted Group Membership			
		Complete Linkage	1	2	3	Total
Original	Count	1	23	3	2	28
		2	1	22	2	25
		3	1	0	13	14
	%	1	82,1	10,7	7,1	100,0
		2	4,0	88,0	8,0	100,0
		3	7,1	,0	92,9	100,0
a. **86,6%** of original grouped cases correctly classified.						

3.4. Література до теми:

- ✓ А.М.Дубров, В.С.Мхитарян, Л.И.Трошин Многомерные статистические методы. – М.: 2003.
- ✓ Інтелектуальний аналіз даних: підручник /О.І.Черняк, П.В.Захарченко. – К.:Знання, 2014.
- ✓ Лупан І.В., Авраменко О.В., Акбаш К.С. Комп'ютерні статистичні пакети: навчально-методичний посібник. – 2-е вид. – Кіровоград, «КОД», 2015. – 236 с.
- ✓ Факторный, дискриминантный и кластерный анализ/ Дж.-О.Ким, Ч.У.Мьюллер, У.Р.Клекка и др. – М.,1989. – 215 с.
- ✓ Чубукова И.А. Data Mining (Курс лекций) – URL: http://kist.ntu.edu.ua/textPhD/Chubukova-Data_Mining.pdf
- ✓ Пустовойт В.И., Самойлов А.С., Ключников М.С. Скрининг-диагностика функционального состояния спортсменов-дайверов с преобладанием автономного типа регуляции. //Медицина экстремальных ситуаций. 2019: 21 (2), С. 320-328. – URL: https://cyberleninka.ru/article/n/skrining-diagnostika-funktsionalnogo-sostoyaniya-sportsmenov-dayverov-s-preobladaniem-avtonomnogo-tipa-regulyatsii/viewer
- ✓ Рейтингова таблиця вищих навчальних закладів «Топ-200 Україна» – 2021 рік. – URL: http://www.euroosvita.net/index.php/ ?category=49&id=6867

3.5. Дані, використані у прикладі

Таблиця 3.1

Код	лідерство	впевне-ність	вимогли-вість	скепти-цизм	поступли-вість	довірли-вість	добро-сердя	чуйність
10001	9	5	10	10	7	7	7	8
10003	8	5	10	10	12	8	12	11
10004	10	6	7	7	4	5	8	9
10005	10	9	6	6	4	7	9	10
10006	2	7	8	8	8	7	9	12
10007	12	10	4	4	11	10	10	11
10009	5	8	11	11	4	9	9	12
10010	10	4	5	5	10	5	12	13
10011	9	5	9	9	6	6	10	10
10013	12	5	8	8	4	5	9	10
10015	6	5	4	4	8	5	7	10
10016	7	5	4	4	7	7	10	11
10017	8	9	11	11	3	4	9	3
10018	9	7	5	5	6	7	8	11
10020	7	3	6	6	8	7	13	9
10022	9	8	6	6	10	8	11	11
10023	8	4	6	6	2	7	5	6
10024	5	4	9	9	8	7	7	7
10025	6	7	7	7	12	12	10	8
10027	4	10	5	5	12	4	7	6
10028	9	6	8	8	6	6	7	5

10030	10	7	7	7	11	7	10	7
10032	13	7	8	8	6	8	12	13
10033	13	12	12	12	6	4	6	5
10034	10	7	6	6	7	4	9	13
10035	7	10	6	6	7	4	9	13
10037	9	7	8	8	7	8	11	3
10038	10	9	8	8	5	9	7	9
10039	5	4	9	9	6	5	9	5
10040	11	5	10	10	6	7	8	11
10041	4	6	7	7	4	3	6	4
10042	7	9	10	10	5	6	9	10
10043	7	6	8	8	9	6	7	8
10046	5	5	6	6	5	2	5	7
10047	6	5	7	7	6	6	8	11
10051	3	3	8	8	13	8	10	11
10052	7	3	6	6	9	8	12	11
10053	9	4	5	5	5	6	10	14
10054	12	10	12	12	3	5	6	6
10055	9	4	7	7	9	10	11	8
10056	11	10	6	8	7	10	12	0
10057	8	8	10	10	5	11	9	5
10059	5	10	9	9	5	3	7	8
10060	8	6	8	8	7	9	12	14
10061	7	2	7	7	10	10	7	10
10062	10	7	8	8	9	7	11	11
10065	11	4	8	8	3	4	8	11
10067	5	6	8	8	5	6	5	12
10068	9	5	9	9	6	6	8	9
10069	6	8	12	12	3	2	6	8
10073	8	6	6	6	2	3	8	7
10074	11	8	13	13	2	5	9	5
10075	10	4	10	10	8	6	8	9
10078	4	6	9	9	2	7	6	6
10079	8	4	7	7	8	6	12	9
10082	6	5	5	5	5	4	9	10
10083	5	4	7	7	8	6	12	6
10084	9	4	4	4	8	7	10	8
10085	10	6	7	7	4	4	9	9
10086	9	4	8	8	11	10	15	10
10087	11	7	6	6	3	9	11	7
10088	6	5	5	5	10	6	12	11
10089	5	6	6	6	8	11	9	4
10090	1	9	10	10	4	2	9	10
10091	8	5	6	6	10	8	8	11
10093	8	7	6	6	10	13	10	10
10094	10	9	9	9	7	9	8	9

ТЕМА 4. ПОШУК АСОЦІАТИВНИХ ПРАВИЛ

4.1. Задача про ринковий кошик

Асоціативні правила – це твердження типу «*якщо→то»*, які допомагають виявити взаємозв'язки між даними, тобто логічними атрибутами об'єктів. Асоціативне правило має дві частини, *антецедент* або *умову* (якщо) і *наслідок* (то). Антецедент – це елемент (або набір елементів), знайдений у даних. Наслідок – це елемент (або набір елементів), який знаходять у поєднанні з антецедентом.

Поштовхом для створення процедур пошуку асоціативних правил (ПАП) стала задача аналізу ринкового кошика *(Market basket analysis)*. З цим пов'язана і використовувана термінологія: об'єкти, стосовно яких здійснюється аналіз, називають товарами (*item*); задача аналізу полягає у тому, щоб з'ясувати, які товари, з тих, що продаються на ринку, найчастіше потрапляють в один кошик (*basket*), тобто які з товарів купують одночасно. Оскільки відомості про покупки фіксують у відповідних базах даних, то кошик є синонімом транзакції. Отже транзакція – це повний набір товарів в індивідуальному кошику на момент покупки. Звичайно деякі покупки можуть бути випадковими, не пов'язаними між собою, тому завдання аналізу виявити саме зв'язані підмножини таких наборів (*itemsets*)[22].

З погляду маркетологів цікавими з є завдання

- ефективно розмістити товари на території супермаркета;
- розробити систему знижок на одні товари, щоб стимулювати продаж інших;
- сформувати рекламу на товар, врахувавши товари, з якими його зазвичай купують разом.

Іншими прикладами застосування методів ПАП є виявлення шахрайства у різноманітних сферах, зокрема в операціях з кредитними картками та страховими випадками; визначення причин збоїв у телекомунікаційних та інших складних системах, аналіз ДНК живих організмів, обробка даних соціологічних досліджень і т.п.

Такі задачі поєднує те, що серед усіх можливих зв'язків між об'єктами бази даних транзакцій слід відібрати лише ті, що становлять інтерес для подальших досліджень або прийняття рішень. Щоб відсіяти велику кількість банальних правил, в процедурах ПАП застосовують такі показники [22]:

Підтримка *(Support)* – це частота появи елемента (або набору елементів) у базі даних. Підтримку набору елементів X, тобто supp(X), визначають як частку у базі даних транзакцій, що містять itemset X.

Наприклад, для даних, наведених у прикладі (стор. 60), одноелементні набори мають підтримку, відповідно, *supp(I1)=6/9, supp(I2)=7/9, supp(I3)=6/9, supp(I4)=2/9, supp(I5)=2/9*, а двоелементні – *supp(I1→I2)=4/9, supp(I1→I3)=4/9, supp(I2→I3)=4/9, supp(I1→I5)=2/9*.

[22] Зазначимо, що у загальному випадку шукана підмножина може складатися як з одного елемента, так і з двох, трьох і т.д. елементів.

Достовірність (*Confidence*) – є оцінкою P(Y | X) – імовірності появи Y, за умови, що дана транзакція вже містить X, тобто умовної імовірності.

$$conf(X \rightarrow Y) = supp(X \cup Y)/supp(X),$$

де $supp(X \cup Y)$ означає, що транзакція одночасно мість і X, і Y. Значення достовірності може коливатись у межах від 0 до 1.

Отже, для даних з табліці 4.1 маємо: $conf(I1 \rightarrow I2)=4/6=2/3$, $conf(I1 \rightarrow I3)=4/6=2/3$, $conf(I2 \rightarrow I3)=4/6=2/3$, $conf(I1 \rightarrow I5)=2/6=1/3$.

На жаль вказаних характеристик недостатньо для оцінювання корисності виявлених асоціацій. Остаточну «несподіваність» чи «застосовність» правила оцінює експерт. Тим не менше додатково визначають ще й такі показники[23]:

Ліфт або *покращення* (*lift, improvement*) – є мірою кореляції між елементами, яку визначають наступним чином: вибір набору елементів X ***не залежить*** від вибору набору елементів Y, якщо $P(X \cup Y) = P(X)P(Y)$; в іншому випадку, набори X та Y залежні і корелюють як події. Це визначення можна легко розширити до більш ніж двох наборів. Ліфт між виборами X та Y може бути обчислений так:

$$lift(X \rightarrow Y) = supp(X \cup Y)/(supp(X)\ supp(Y)) = conf(X \rightarrow Y)\ /supp(Y).$$

Тобто ліфт визначають як відношення спостережуваної підтримки до тої, яку очікують у разі незалежності X та Y. Значення показника може знаходитися у діапазоні $[0, +\infty)$. Значення, близькі до 1, означають, що X та Y незалежні, тобто правило – не цікаве; значення більше за 1 означає пряму кореляцію, а менше за 1 – обернену кореляцію між наборами елементів, тобто елементи замінюють один одного . Іншими словами ліфт показує, наскільки підвищується імовірність виявити Y у наборі, де вже є X.

Для даних з Таблиці 4.1 маємо:

$lift(I1 \rightarrow I2) = conf(I1 \rightarrow I2)/supp(I2) = 2/3 \cdot 9/7=6/7 < 1;$

$lift(I1 \rightarrow I3) = conf(I1 \rightarrow I3)/supp(I3) = 2/3 \cdot 9/6=6/6 = 1;$

$lift(I2 \rightarrow I3) = conf(I2 \rightarrow I3)/supp(I3) = 2/3 \cdot 9/6=6/7 = 1;$

$lift(I1 \rightarrow I5) = conf(I1 \rightarrow I5)/supp(I5) = 1/3 \cdot 9/2=9/6 = 3/2 > 1.$

Отже, серед наведених правил найбільш переспективним є правило I1→I5, воно на 50% потужніше правила, яке містить тільки I1, хоча достовірність його невелика. Тобто I1 та I5 зустрічаються в одній транзакції у 1,5 рази частіше, ніж цього можна було б очікувати, якби вони були статистично незалежними.

Інтерес (*interest*) – це абсолютне значення різниці між достовірністю правила $conf(X \rightarrow Y)$ та підтримкою Y (яку очікують при незалежному виборі товарів):

$$Int(X \rightarrow Y) = |conf(X \rightarrow Y) - supp(Y)|.$$

Якщо X не впливає на Y, то частка кошиків, які містять X та Y, така сама, як і частка кошиків, які містять Y. Інтерес такого правила дорівнює нулю.

[23] Association_rules: Association rules generation from frequent itemsets. – URL: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

Наприклад, для правила *{dog}→{cat}* (див. Вправу 4.0) матимемо *conf({dog}→{cat})*=5/7, *supp({cat})*=6/8, відповідно *Interes({dog}→{cat})*=|5/7-3/4|=|0,714-0,75|≈0,036.

Для даних з Таблиці 4.1 маємо:

Int(I1→I2) = |conf(I1→I2) – supp(I2)|=|2/3-7/9|=1/9≈0,11;

Int(I1→I3) = |conf(I1→I3) – supp(I3)|=|2/3-6/9|=0;

Int(I2→I3) = |conf(I2→I3) – supp(I3)|=|2/3-6/9|=0;

Int(I1→I5) = |conf(I1→I5) – supp(I5)|=|1/3-2/9|=1/9≈0,11.

Тут також бачимо, що правило I1→I5 має значення інтересу відмінне від нуля, в той час, як для правил I1→I3 та I2→I3 цей показник дорівнює нулю.

Впевненість *(conviction)* – степінь імплікації асоціативного правила:

$$conv(X \rightarrow Y) = (1 - supp(Y))/(1 - conf(X \rightarrow Y)).$$

Показник може приймати значення у діапазоні [0, +∞]. Високе значення показника означає, що наслідок сильно залежить від антецедента, тобто умови. Як і для ліфта, значення впевненості рівне 1 означає незалежність антецедента та наслідка.

Характеристика є чутливою до напрямку правила, тобто $conv(X \rightarrow Y) \neq conv(Y \rightarrow X)$. Чим отримане значення вище за 1, тим краще.

Для розглядуваного прикладу маємо:

conv(I1→I2)=(1-supp(I2)/(1-conf(I1→I2))=(1-7/9)/(1-2/3)=2/9·3/1=2/3;

conv(I1→I3)=(1-supp(I3)/(1-conf(I1→I3))=(1-6/9)/(1-2/3)=3/9·3/1=1;

conv(I2→I3)=(1-supp(I3)/(1-conf(I2→I3))=(1-6/9)/(1-2/3)=2/9·3/1=1;

conv(I1→I5)=(1-supp(I5)/(1-conf(I1→I5))=(1-2/9)/(1-1/3)=7/9·3/2=7/6≈1,17.

Тут маємо, що правило I1→I5 буде на 17% (у 1,17 рази) більш правильним, ніж проста випадковість потрапляння елементів I1 та I5 в одну транзакцію.

Для правила *I5→I1* впевненість прямує до нескінченості, що виключає будь-яку випадковість потрапляння I1 у транзакцію, що містить I5:

conv(I5→I1)=(1-supp(I1)/(1-conf(I5→I1))=(1-6/9)/(1-1)=3/9/0→∞.

Посилення (*leverage*) – визначає різницю між спостережуваною частотою появи X і Y разом із частотою, яка очікувалася б, якби X та Y були незалежними. Значення показника може знаходитися у діапазоні [-1, +1]. Значення, рівне 0, вказує на незалежність.

$$leverage(X \rightarrow Y) = supp(X \rightarrow Y) - supp(X)\ supp(Y).$$

Для розглядуваного прикладу маємо:

leverage(I1→I2) = supp(I1→I2)–supp(I1) ·supp(I2)=4/9-(6/9)·(7/9)=(36-42)/81 ≈-0,074;

leverage(I1→I3) = supp(I1→I3)–supp(I1) ·supp(I3)=4/9-(6/9)·(6/9)=(36-36)/81 ≈0;

leverage(I2→I3) = supp(I2→I3)–supp(I2) ·supp(I3)=4/9-(7/9)·(6/9)=(36-42)/81 ≈-0,074;

leverage(I1→I5) = supp(I1→I5)–supp(I1) ·supp(I5)=2/9-(6/9)·(2/9)=(18-12)/81 ≈0,074.

Як бачимо, найменш цікавим виявилося правило I1→I3: незалежність елементів, що його формують, підтверджується значеннями посилення, ліфта, впевненості та інтереса, хоча підтримка цього правила досить висока.

4.2. Алгоритми пошуку асоціативних правил

Базовим алгоритмом пошуку асоціативних правил є алгоритм Apriori, розроблений у 1993 р.: *R. Agrawal, T. Imielinski, A. Swami* (1993). Mining Associations between Sets of Items in Massive Databases. In Proc. of the 1993 ACM-SIGMOD Int'l Conf. on Management of Data, 207-216.; *R. Agrawal, R. Srikant*. "Fast Discovery of Association Rules", In Proc. of the 20th International Conference on VLDB, Santiago, Chile, September 1994.

Основними кроками алгоритма Apriori є формування k-елементних наборів, які зустрічаються у базі даних, та відсікання наборів з підтримкою, нижчою за встановлений мінімальний поріг. Пошук здійснюється для наборів з k елементів, починаючи від k=1.

На першому проході алгоритма підраховують елементи (items), а потім визначають, які з них є частими. На другому проході підраховують лише пари елементів, обидва з яких часто зустрічаються на першому проході. Інші пари ігнорують на підставі монотонності[24] частих наборів. Далі, беручи до уваги монотонність, пошук наборів розміром k елементів здійснюють лише серед наборів розміром k-1, які були визнані частими.

У псевдокоді алгоритм Apriori можна предствити так (Алгоритм 4-1) [3]:

Алгоритм 4-1

L_1={одноелементні набори, що часто зустрічаються}

для (k=2; L_{k-1}<>∅; k++)

C_k=AprioriGen(F_{k-1}) // генерація кандидатів

для всіх транзакцій t∈D виконати

C_t=subset(C_k, t) //вилучення надлишкових правил

для всіх кандидатів c∈C_t виконати c.count ++

кінець ***для всіх***

кінець ***для всіх***

L_k={c∈C_k |c.count>=$Supp_{min}$} // відбір кандидатів

Кінець ***для***

Результат = $\bigcup_k L_k$

[24] Тобто елементи, які рідко зустрічаються самі по собі, не можуть утворювати у пари з більшою частотою.

Або так (Алгоритм 4-2):

Алгоритм Apriori: Пошук частих наборів з використанням ітеративного рівневого підходу, заснованого на генеруванні кандидатів [22, Chapter 6.2].

Алгоритм 4-2

Вхідні дані:
D – база даних транзакцій;
min sup – мінімальний поріг підтримки.
Результати: *L* – часті набори у базі *D*.
Метод:

```
(1) L1 = find.frequent.1-itemsets(D);
(2) for (k = 2;Lk-1 ≠ 0;k++) {
(3)     Ck = apriori_gen(Lk-1);
(4)     for each transaction t ∈ D { // перегляд D для підрахунку
(5)         Ct = subset(Ck, t); // формування множини, де t є кандидатом
(6)         for each candidate c ∈ Ct
(7)             c.count++;
(8) }
(9) Lk = {c ∈ Ck|c.count ≥ min_sup}
(10) }
(11) return L = ∪kLk;

procedure apriori_gen(Lk-1:frequent (k -1)-itemsets)
(1) for each itemset l1 = Lk-1
(2)     for each itemset l2 ∈ Lk-1
(3)         if (l1[1] = l2[1])^(l1[2] = l2[2])
                ^...^(l1[k -2] = l2[k -2])^(l1[k -1] < l2[k -1]/ then {
(4)             c = l1 ⊗ l2; // об'єднання з метою формування кандидатів
(5)             if has_infrequent_subset(c, Lk-1) then
(6)                 delete c; // скорочення (вилучення) некорисного
    кандидата
(7)             else add c to Ck;
(8)         }
(9) return Ck;

procedure has_infrequent_subset(c: candidate k-itemset;
        Lk-1: frequent (k -1)-itemsets); // використання попередніх знань
(1) for each (k -1)-subset s of c
(2)     if s ∉ Lk-1 then
(3)         return TRUE;
(4) return FALSE;
```

Розглянемо приклад.

Нехай маємо базу транзакцій з елементами I1, I2, I3, I4, I5:

Таблиця 4.1

TID	набір елементів
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Нехай мінімальний поріг підтримки визначено рівним 2, тобто min_supp=2.

При першому проході алгоритма будуть визначені елементи бази даних та підраховані їхні частоти. Ці елементи складуть множину одноелементних наборів C_1. Оскільки підтримка елементів цієї множини не менша за порогове значення, то усі вони становитимуть множину частих одноелементних наборів L_1:

Одноелементні набори:

C_1	Частоти
I1	6
I2	7
I3	6
I4	2
I5	2

Часті одноелементні набори:

L_1	Частоти
I1	6
I2	7
I3	6
I4	2
I5	2

Наступними кроками є формування множини C_2 – двоелементних наборів з елементів множини L_1, підрахунок їхніх підтримок та формування множини частих двоелементних наборів L_2:

Множина C_2:

C_2	Частоти
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

Множина L_2:

L_2	Частоти
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

Нарешті з елементів множини L_2 буде сформовано множину 3-елементних наборів C_3, підраховані підтримки її елементів та сформовано множину L_3:

Множина C_3:	Частоти (підтримки)	*Множина L_3:*
{I1, I2, I3}	2	{I1, I2, I3}
{I1, I2, I5}	2	{I1, I2, I5}

Продовжувати далі немає сенсу, оскільки не всі підмножини кандидита до набору C4={I1, I2, I3, I5} є частими, зокрема підмножина {I2, I3, I5}. Отже знайдено усі часті набори з одного, двох та трьох елементів вхідної бази даних, підтримка яких не менша за встановлене порогове значення.

Не зважаючи на те, що в кожній наступній ітерації алгоритма Apriori кількість кандидатів зменшується, що призводить до хорошого підвищення продуктивності, тим не менше у класичному підході доводиться 1) багаторазово сканувати всю базу даних і перевіряти великий набір кандидатів на відповідність шаблону; 2) створювати величезну кількість наборів кандидатів (наприклад, якщо є 10^4 частих одноелементних наборів, алгоритму Apriori потрібно буде згенерувати більше 10^7 двоелементних наборів для подальшого аналізу). – В реальних умовах, враховуючи великий розмір та, можливо, розподіленість баз даних, здійснення таких дій є проблематичним та надзвичайно витратним по часу та обчислювальних ресурсах.

Для подолання цих проблем та покращення часової і просторової ефективності алгоритмів пошуку асоціативних правил застосовують, зокрема, такі підходи [22, 23]:

1. ***Хешування***. Техніка на основі хешування може бути використана для зменшення кількості кандидатів до k-елементного набору, C_k, для $k > 1$.

На першому етапі під час сканування кожної транзакції в базі даних для генерації частих одноелементів наборів L_1, можна генерувати також і всі двоелементні набори для кожної транзакції, розміщуватити їх у різні сегменти хеш-таблиці та збільшувати лічильники відповідних сегментів. Нехай для даних прикладу (Таблиця 4.1) хеш функцію визначено так:

$$h(X, Y)=((\text{номер } X*10+\text{номер } Y) \bmod 7.$$

Тоді при скануванні бази даних крім множини C_1 (див. с.60) буде створено хеш-таблицю:

Таблиця 4.2

Номер сегмента	0	1	2	3	4	5	6
Лічильник	2	2	4	2	2	4	4
Вміст сегмента	{I1, I4} {I3, I5}	{I1, I5} {I1, I5}	{I2, I3} {I2, I3} {I2, I3} {I2, I3}	{I2, I4} {I2, I4}	{I2, I5} {I2, I5}	{I1, I2} {I1, I2} {I1, I2} {I1, I2}	{I1, I3} {I1, I3} {I1, I3} {I1, I3}

Двоелементний набір, якому відповідає лічильник із значенням нижчим за поріг підтримки, не може бути частим, і тому його слід вилучити з набору кандидатів. Така методика на основі хешу може істотно зменшити кількість досліджуваних k-елементних наборів (особливо, коли $k = 2$).

Саме такий підхід застосовано, наприклад, в алгоритмі PCY (J.Park, M.Chen, P.Yu – 1995) – алгоритмі прямого хешування та скороченння даних (Direct

Hashing and Pruning) – відкидання k-елементних наборів, підтримки яких не досягли мінімального значення.

2. ***Скорочення транзакцій*** (зменшення кількості сканованих транзакцій у майбутніх ітераціях): транзакція, яка не містить частих k-елементних наборів, не може містити будь-якого (k-1)-елементного частого набору. Отже, таку транзакцію можна якимось чином позначити або видалити з подальшого розгляду, оскільки наступні сканування бази даних для j-елементних наборів, де j> k, не потребуватимуть розгляду такої транзакції.

Скорочення транзакцій за принципом «*поділяй і володарюй*» реалізовано в алгоритмі **FP-grows** (frequent pattern grows – зростання частих шаблонів). Алгоритм, по-перше, стискає базу даних частих елементів у дерево частих шаблонів або FP-дерево, яке зберігає інформацію про асоціації наборів елементів. Потім він ділить стиснуту базу даних на набір умовних баз даних (особливий вид проектованої бази даних), кожна з яких пов’язана з одним частим елементом або «фрагментом шаблону», і міняє кожну базу даних окремо. Для кожного «фрагмента шаблону» потрібно досліджувати лише набори даних, пов’язані з ним. Отже такий підхід спрямований на істотне зменшення розміру множини наборів даних, що підлягають пошуку, із «зростанням» в той же час кількості досліджуваних шаблонів.

Якщо застосувати алгоритм FP-grows до даних прикладу, наведеного на стор. 60 (Таблиця 4.1), то результати першого проходу будуть такими ж, як і для алгоритма Apriori: буде створено множину одноелементних частих наборів та визначено підтримки елементів. Нехай встановлено мінімальний поріг підтримки рівний 2. Тоді буде отримано набір частих елементів, відсортований у порядку зменшення величини підтримки: L_1={{I2, 7}, {I1, 6}, {I3, 6}, {I4, 2}, {I5, 2}}.

FP-дерево будують так: по-перше, створюють корінь дерева, позначений «null», і сканують базу даних D вдруге, обробляючи елементи кожної транзакції в порядку L (тобто відповідно до спадання значення підтримки), та для кожної транзакції створюють гілку. Наприклад, сканування першої транзакції «T100: I1, I2, I5», яка містить три елементи (I2, I1, I5 у порядку L), призведе до побудови першої гілки дерева з трьома вузлами, (I2: 1), (I1: 1) та (I5: 1), де I2 є нащадком кореневого вузла, I1 є нащадком вузла I2, а I5 є нащадком вузла I1. Друга транзакція, T200, містить елементи I2 та I4 у порядку L, що призведе до створення гілки, де I2 пов’язаний з коренем, а I4 – з вузлом I2. Однак ця гілка має загальний префікс I2 з існуючим шляхом для T100. Тому замість створення нового вузла слід збільшити лічильник вузла I2 на 1 та створити новий вузол, (I4: 1), як дочірній до вузла (I2: 2). Загалом, розглядаючи гілку, яку слід додати для транзакції, лічильники кожного вузла вздовж загального префікса слід збільшити на 1, а для елементів, що слідують за префіксом, створюють пов’язані з префіксом вузли (Рис. 4.1).

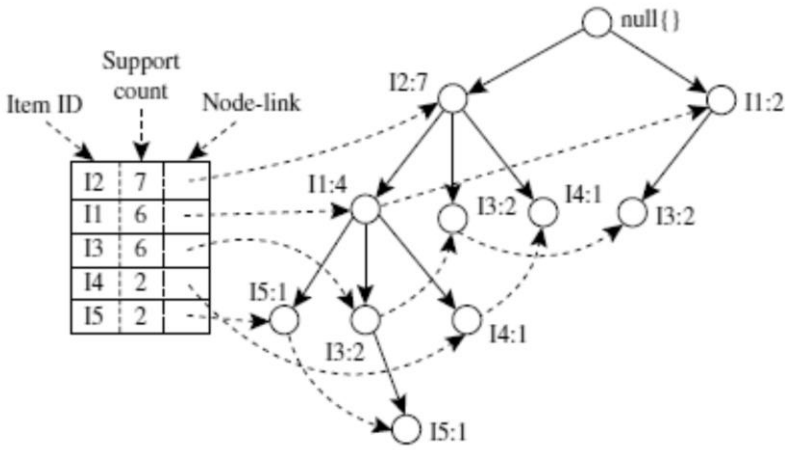


Рис. 4.1[25]

Детальніше опис алгоритму FP-grows наведено у підручнику [22, chapter 6.2]. В цілому метод перетворює проблему пошуку довгих частотних моделей у рекурсивний пошук більш коротких моделей у набагато менших за розміром умовних базах даних з подальшим об'єднанням суфіксу. Метод істотно знижує витрати на пошук.

Однак, коли база даних велика, для побудови основного FP-дерева може не вистачити пам'яті. У такому разі пропонують розділити базу даних на набір баз даних, а потім сконструювати та проаналізувати FP-дерево для кожної з них. Дослідження показують, що метод FP-grows ефективний і масштабований для добування довгих і коротких частих шаблонів і приблизно на порядок швидший, ніж алгоритм Апріорі.

3. ***Розбиття*** (поділ даних для пошуку наборів елементів) – алгоритм, побудований на основі такого підходу, складається з двох фаз (Рис. 4.2). На фазі I алгоритм ділить усі транзакції бази D на n розділів (частин), що не перетинаються. Якщо мінімальний відносний поріг підтримки для транзакцій у D є *min_sup*, то мінімальним значенням підтримки для розділу є *min_sup* × *число_число_транзакцій_у_цьому_розділі*. Для кожного розділу знаходять усі локальні часті набори елементів (тобто набори елементів, які часто зустрічаються в розділі). Локальний набір елементів може бути, а може і не бути частим відносно всієї бази даних D. Однак, будь-який набір елементів, який є потенційно частим відносно D, має зустрічатися як частий набір елементів принаймні в одному з розділів. Тому всі локальні часті набори елементів – це набори кандидатів по D. Колекція частих наборів елементів з усіх розділів утворює глобальні набори елементів кандидатів по D. У II фазі проводиться друге сканування D, в якому оцінюється фактична підтримка кожного кандидата та визначаються глобальні часті набори елементів. Розмір розділу та кількість розділів встановлюють таким чином, щоб кожен розділ вміщувався в основну пам'ять і читався лише один раз на кожному етапі [22].

[25] Див. [22] – Fig. 6.7.

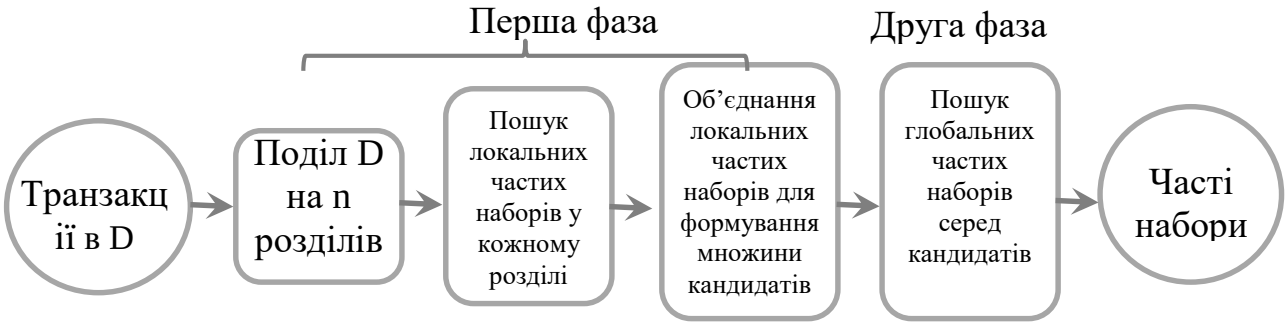


Рис. 4.2

Саме такий підхід реалізовано в алгоритмі SON (A.Savasere, E.Omiecinski, S.Navathe – 1995). Цей алгоритм особливо підходить для технології MapReduce.

4. ***Рандомізація*** – у рандомізованих алгоритмах перегляд даних обмежують випадковою вибіркою кошиків, достатньо малою, щоб можна було зберігати як вибірку, так і необхідну кількість наборів елементів в основній пам'яті. Поріг підтримки у такому разі зменшують пропорційно до розміру вибірки. Оскільки репрезентативність вибірки залежить від випадкових факторів, при такому підході можливі як помилково позитивні, так і помилково негативні результати, коли набір елементів є частим у вибірці, але не у повному наборі, або, навпаки, є частим, але у вибірці майже не зустрічається.

Наприклад, алгоритм Тойвонена (H.Toivonen, 1996) полягає у пошуку частих наборів елементів у вибірці, але з пониженим порогом, тому імовірність пропустити набір елементів, який є частим в цілому, мала. Далі досліджують весь файл транзакцій, підраховуючи не тільки набори елементів, які є частими у вибірці, але і *негативну границю* – набори елементів, чиї підмножини виявилися частими. Якщо жоден член негативної границі не зустрічається часто в цілому, то задовольняються вже знайденою множиною частих наборів. Але якщо член негативної границі виявляється частим, то весь процес повторюють з іншою вибіркою.

Індивідуальних завдання №1

Ознайомитися з одним з алгоритмів пошуку асоціативних правил:

- ✓ **AIS** (R.Agrawal, T.Imielinsky, A.Swami – 1993, IBM) – кандидатів множини наборів генерують і підраховують «на льоту» в процесі сканування БД.
- ✓ **SETM** – для формування кандидатів використано операцію об'єднання мови SQL.
- ✓ **Apriori** (R.Agrawal, R.Shrikant – 1994) – формування та підрахунок кандидатів (декілька етапів).
- ✓ **AprioriTid** – генерування наборів з великих наборів, знайдених на попередньому кроці (БД сканують лише при першому проході).
- ✓ **AprioriHybrid** – на перших етапах Apriori, потім – AprioryTid.
- ✓ **PCY** (J.Park, M.Chen, P.Yu – 1995) – алгоритм прямого хешування та обрізання даних (Direct Hashing and Pruning) – відкидання к-наборів, які не досягли мінімального значення підтримки.
- ✓ **SON** (A.Savasere, E.Omiecinski, S.Navathe – 1995).
- ✓ **FP-grows**.

- ✓ **DIC** (S.Brin, R.Motwani, J.Ullman, S.Tsur – 1997).

Підготувати доповідь, в якій вказати 1) історію створення алгоритма (хто і коли його розробив); 2) базові підходи та припущення методу; 3) власне алгоритм; 4) переваги та недоліки алгоритму (ефективність, вимоги до обчслювальних ресурсів, тощо); 5) шляхи подальшого удосконалення.

4.3. Пошук асоціативних правил в пакеті RapidMiner

В пакеті RapidMiner пошук асоціативних правил здійснюють за допомогою операторів **FP-Growth** – пошуку частих елементів та наборів елементів у базі даних транзакцій, – та **Create Association Rules**, який генерує набір асоціативних правил із сформованої оператором **FP-Growt** множини частих наборів.

Проблема полягає у тому, що дані для аналізу зазвичай формуються як результат виконання запиту до бази даних і мають такий вигляд:

ID_транзакції	ID_товару
100	хліб
100	молоко
200	печиво
200	…

А оператор **FP-Growt** в RM потребує такого вигляду вхідних даних:

ID_транзакції	хліб	молоко	печиво	…
100	true	true	false	…
200	false	false	true	…
…	…	…	…	…

Рис. 4.3

Отже дані, передані для аналізу, незалежно від формату – текстового чи у вигляді електронної таблиці, – потребують попередньої підготовки. У шаблоні *Market basket analysis* пакета RM пропонується послідовність операторів, показана на Рис. 4.4, яка при запуску на виконання формує правила за тестовим набором даних. Однак підготовчі дії для іншого набору даних (з іншими атрибутами) можуть відрізнятися.

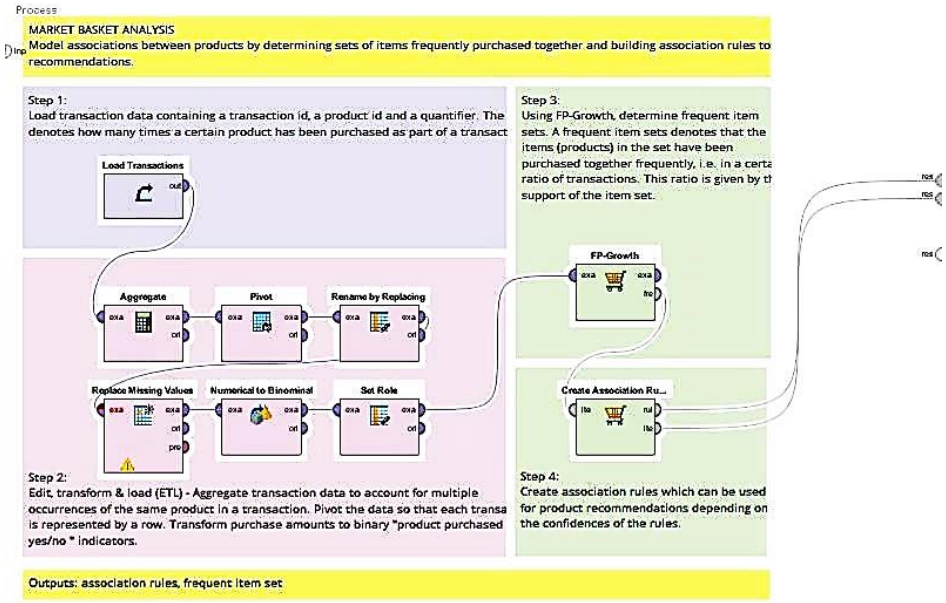


Рис. 4.4

Взявши за основу шаблон, виконаємо пошук асоціативних правил для даних, отриманих з деякої бази даних, як збережений у текстовому форматі

результат виконання запиту, що містить фіскальний номер товарного чека (ідентифікатор транзакції) та назву товару (Рис. 4.5). Результат запиту можна зберегти і у форматі електронної таблиці.

Файл з даними можна завантажити до локального репозиторія (**Add data**) і потім перетягнути в область виконання проекта (**Retrieve**), або відкрити з Excel оператором **Read Excel**. В результаті дані однаково будуть перетворені до вигляду як на Рис. 4.6:

фіс_ном	назва
601	яйце
601	банан
580	ківі
580	грейпфрут рожевий
580	банан
51403	яйце
51403	ківі
51403	банан
51398	грейпфрут рожевий
51398	банан
478	ківі
478	банан
387	ківі
387	Кардамон
387	банан

Рис. 4.5

Row No.	фіс_ном (integer) regular	назва (polynomial) regular
	601	яйце
	601	банан
	580	ківі
	580	грейпфрут рожевий
	580	банан
	51403	яйце
	51403	ківі
	51403	банан
	51398	грейпфрут рожевий
	51398	банан
	478	ківі
	478	банан
	387	ківі
	387	Кардамон
	387	банан

Рис. 4.6

Якщо слідувати шаблону, то послідовність операторів для наведеного набору даних буде такою, як на Рис. 4.7:

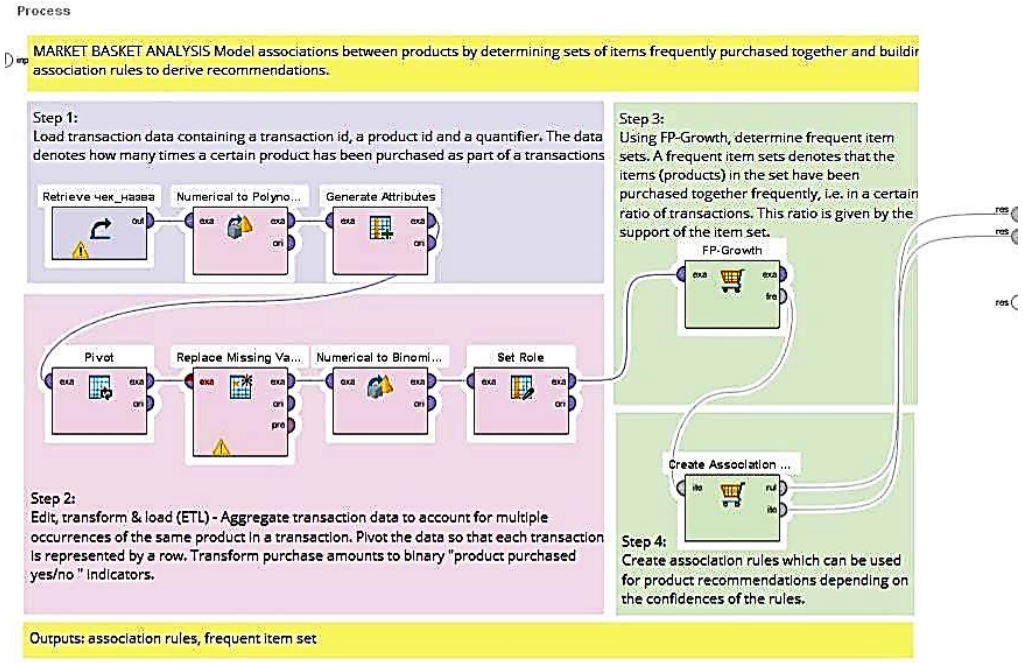


Рис. 4.7

Тут

1) **Retrieve** – завантаження первинних даних з репозиторію;

2) **Numerical to Polynomial** – перетворення атрибута фіс_ном з числового типу до номінативного;

3) **Generate Attribute** – добавлення атрибута item із значенням 1 для кожного рядка вхідних даних, заповнивши відповідні значення function description у параметрах оператора (це потрібно для створення зведення у наступному операторі – у шаблоні перед цим ще здійнюють агрегацію даних для випадку, якщо товар в одному чеку буде вказаний більше одного разу);

4) **Pivot** – оператор перетворення таблиці до вигляду, де кожній транзакції відповідає рядок, а атрибутами є назви продуктів:

Row No.	фіс_ном	item_банан	item_яйце	…
1	601	1	1	?
2	580	1	?	?
…	…	…	…	…

Про входження продукту до транзакції свідчить значення 1, яке утворюється додаванням відповідних значень атрибуту item (згенерованого попереднім оператором), оскільки фактично формується зведена таблиця, в якій атрибут, за яким буде здійснюватися групування (у даному випадку необхідно вказати це «фіс_ном»), та атрибут для індексування (у даному випадку це «назва»). Атрибут для об'єднання та узагальнюючу операцію можна вказати явно, але у нашому випадку, якщо налаштувати параметри оператора як на Рис. 4.8, його буде визначено автоматично і об'єднання здійснюватиметься за тим атрибутом, що залишився, тобто «item».

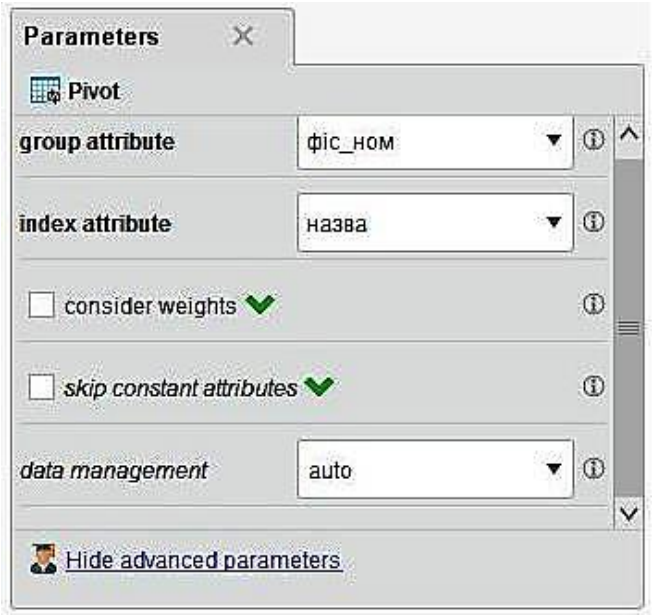


Рис. 4.8

У разі відсутності товару в транзакції і неможливості виконати об'єднання значень, відповідна комірка таблиці не заповнюється (на екрані виводиться знак «?»). – Це пропущене значення, яке слід заповнити, для чого і призначено наступний оператор;

5) **Replace Missing Value** – за замовченням усі пропущені значення заповнюються нулем;

6) **Numerical to Binomial** – нулі та одиниці будуть замінені відповідно значеннями *false* та *true* для перетворення таблиці до вигляду, представленого на Рис. 4.3.

Попередня підготовка завершена.

Далі один за одним виконуватимуться оператори **FP-Growth** та **Create Association Rules**.

Для формування множини частих наборів товарів у параметрах оператора **FP-Growth** необхідно встановити порогове значення мінімальної підтримки. За замовченням встановлено значення 0,95, якому у нашому прикладі не відповідає

жоден елемент, тому встановлено значення 0,09. В результаті отрималося 62 набори з одного, двох та трьох елементів (Рис. 4.9):

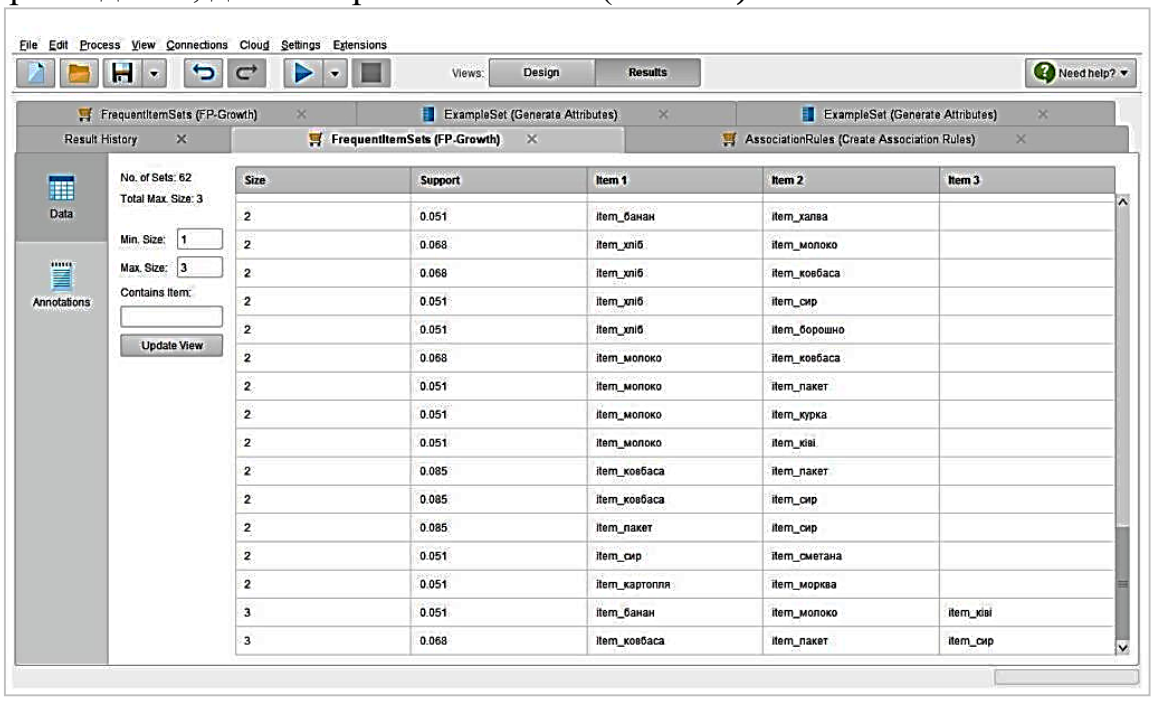


Рис. 4.9

Далі потрібно встановити порогове значення для одного з параметрів оператора **Create Association Rules**. Наприклад, у наведеному випадку встановлено значення достовірності=0,2, але його можна регулювати після виконання процеса, зменшуючи відповідно кількість правил, побудованих на сформованій множині частих шаблонів (Frequent Patterns). Так, за вказаними параметрами буде створено 57 правил (Рис. 4.10):

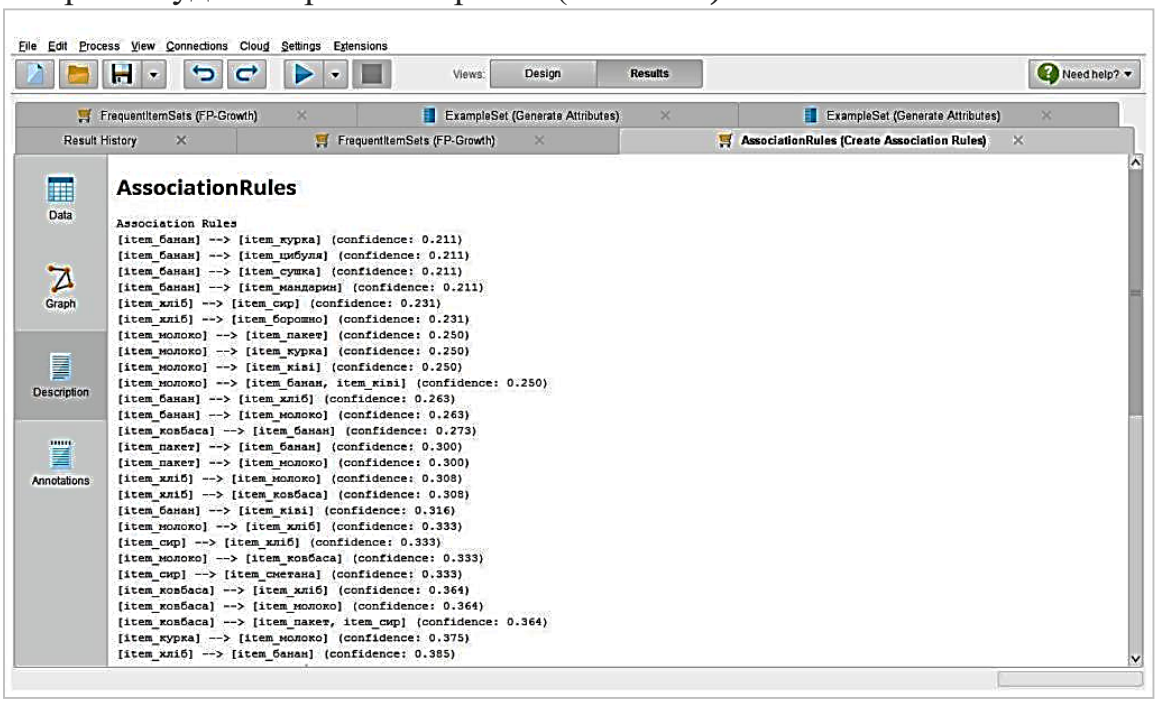


Рис. 4.10

Але, для аналізу, регулюючи обраний параметр повзунком, можна залишити лише найбільш значущі правила, які в результатах процесу будуть представлені

у вигляді таблиці (Рис. 4.11) або у вигляді графа (Рис. 4.12). Для кожного правила наведено антецедент (тут *Premise*) та наслідок (тут *Conclusion*), підтримку та достовірність, а у таблиці, крім того ще й впевненість, ліфт, а також gain, laplace та ps:

- gain (виграш, підсилення) – коефіцієнт підсилення, який обчислюється за допомогою тета-параметра коефіцієнта підсилення;
- laplace – обчислюється за допомогою параметра k laplace;
- ps – ще один критерій для вибору правил.

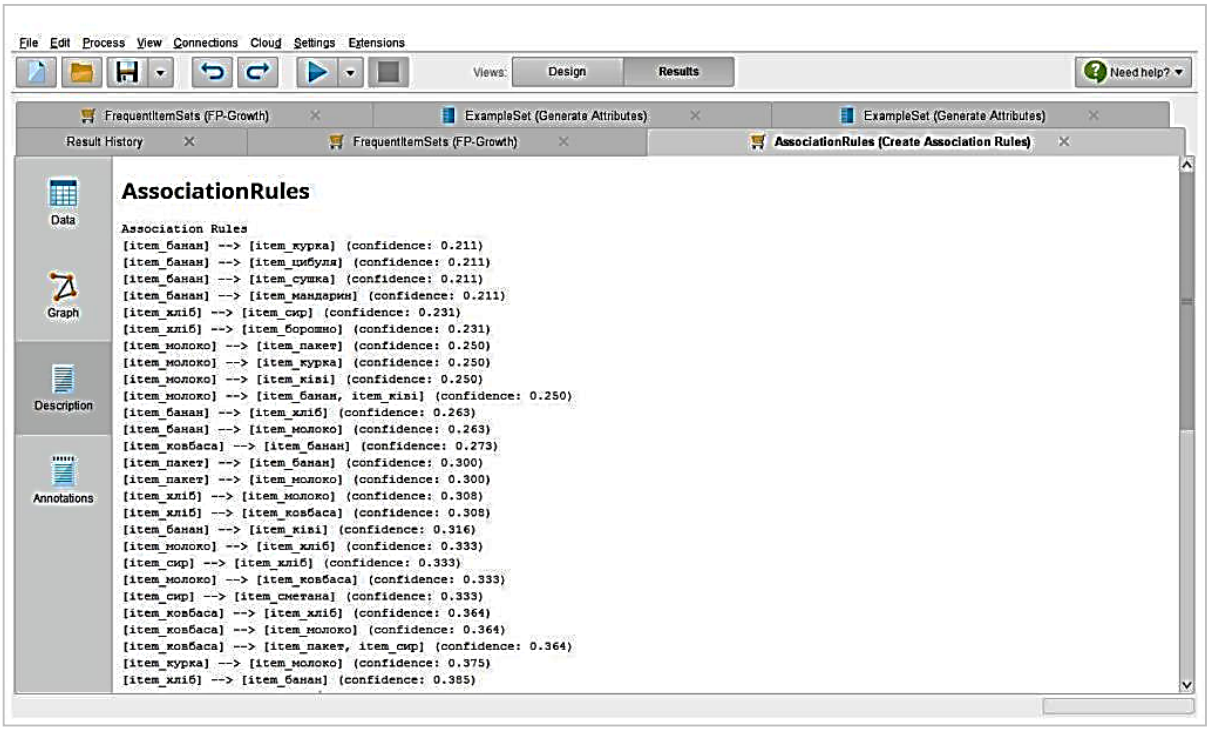


Рис. 4.11

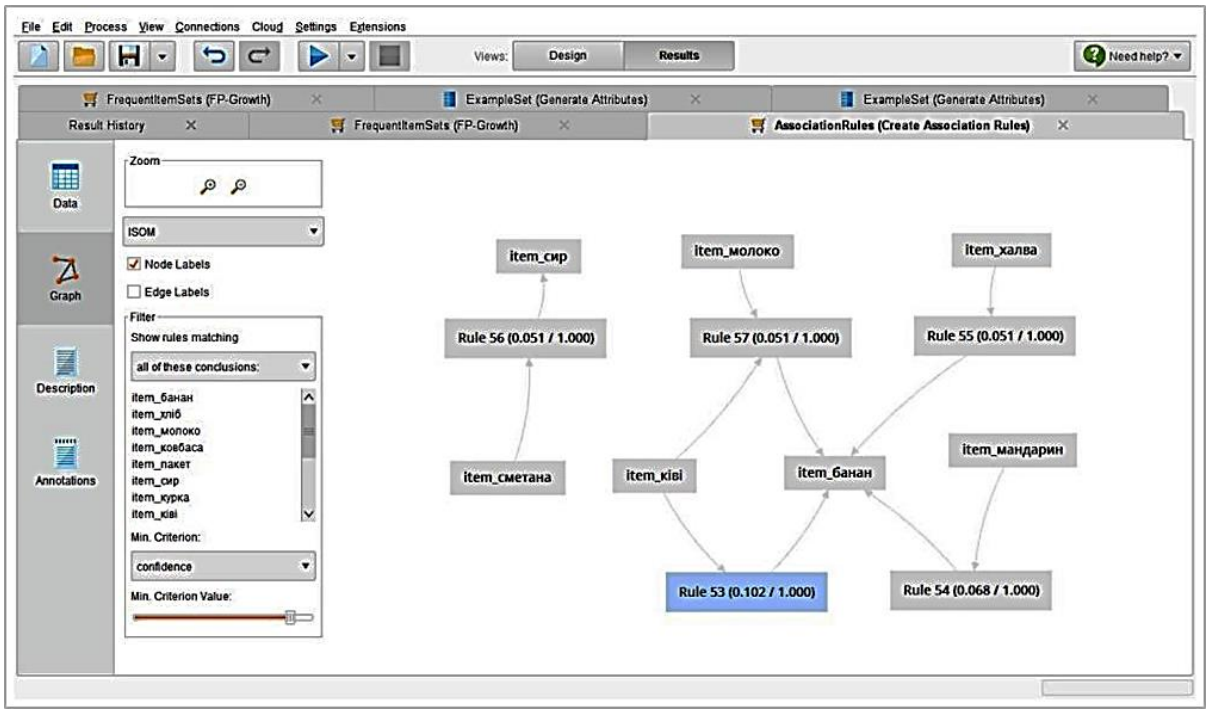


Рис. 4.12

Слід зауважити, що справді цікавих правил у наведеному наборі не знайдено – усі вони доволі банальні, але і проаналізована база транзакцій містить лише близько трьосот записів.

4.4. Пошук асоціативних правил з пакетом KNIME

Ознайомитися з особливостями застосування процедур пошуку асоціативних правил в пакеті KNIME можна на блозі розробника[26] (Рис. 4.13) та проаналізувавши приклади з Explorer'а.

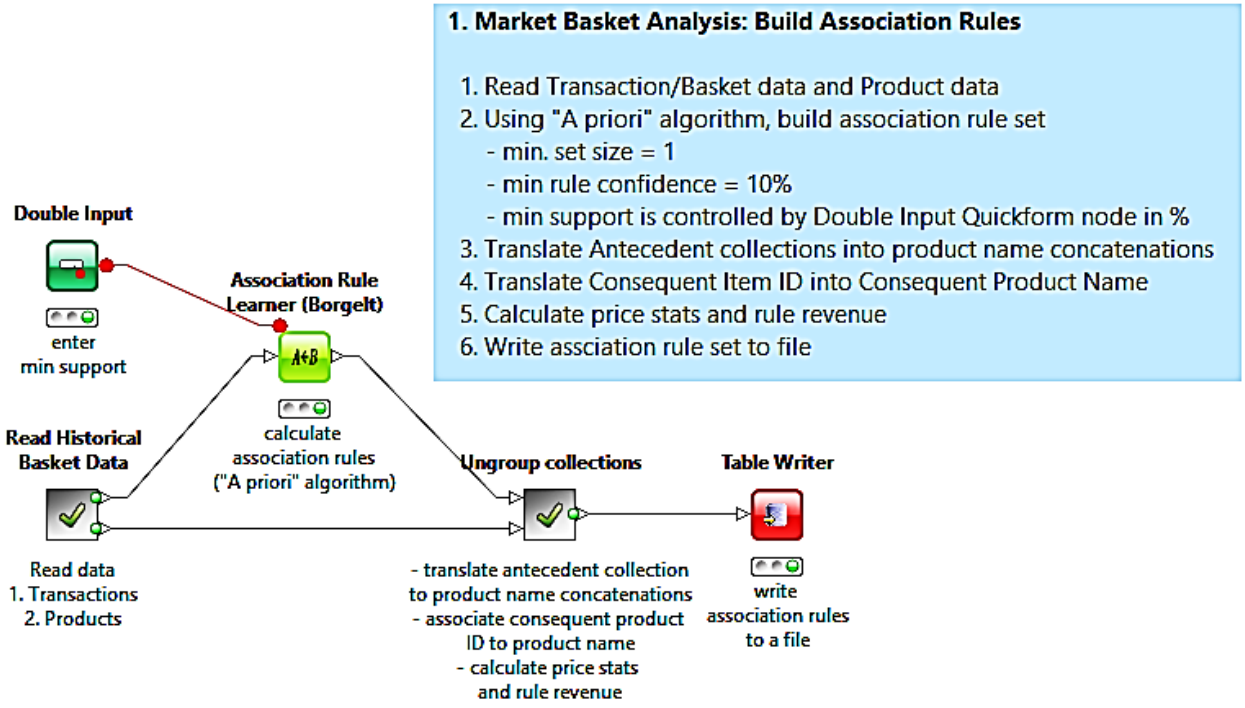


Рис. 4.13

Пропонується використати дані транзакцій. Наприклад, для вправи 4.1. дані матимуть вигляд (Рис. 4.14):

	A
1	Транзакції
2	1
3	1,2
4	1,3
5	1,2,4,
6	1,5
7	1,2,3,6
8	1,7
9	1,2,4,8
10	1,3,9,
11	1,2,5,10
12	1,11
13	1,2,3,4,6,12
14	1,13
15	1,2,7,14
16	1,3,5,15
17	1,2,4,8,16,
18	1,17
19	1,2,3,6,9,18

Рис. 4.14

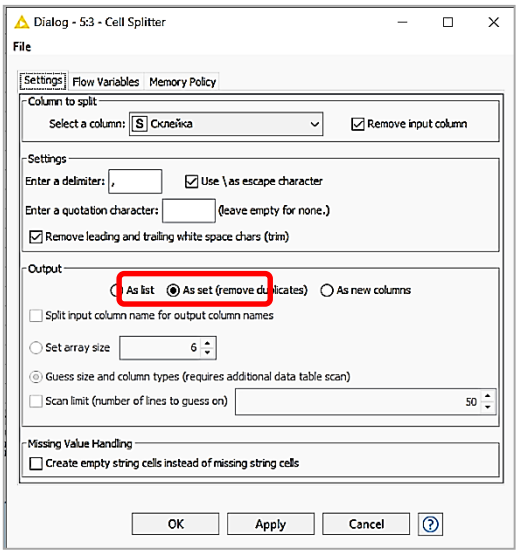


Рис. 4.15

[26] Market Basket Analysis and Recommendation Engines - March 30, 2015 – by Rosaria Silipo. – URL: https://www.knime.com/blog/market-basket-analysis-and-recommendation-engines

З файла Excel такі дані можна прочитати, використовуючи вузол **Excel Reader**. Дані мають бути записані як рядки чисел, розділених комами. Кожен рядок відповідає транзакції.

Наступний крок – перетворити рядки чисел у множини, застосовуючи **Cell Splitter**. Важливо правильно відконфігурувати цей вузол (Рис. 4.15): слід вказати, що послідовності елементів потрібно перетворити *у множину без повторень*.

Після цього можна приступити до побудови асоціативних правил або за допомогою вузла **Association Rule Learner**, який реалізує алгоритм Apriori в традиційній реалізації, або в реалізації **Association Rule Learner (Borgelt)**, яка пропонує кілька покращень продуктивності порівняно з традиційною реалізацією алгоритму. Однак набір правил асоціації вихідних даних залишається незмінним. При конфігуруванні цих вузлів слід вказати граничні значення для підтримки та достовірності для пошуку асоціативних правил (Рис. 4.16, Рис. 4.17).

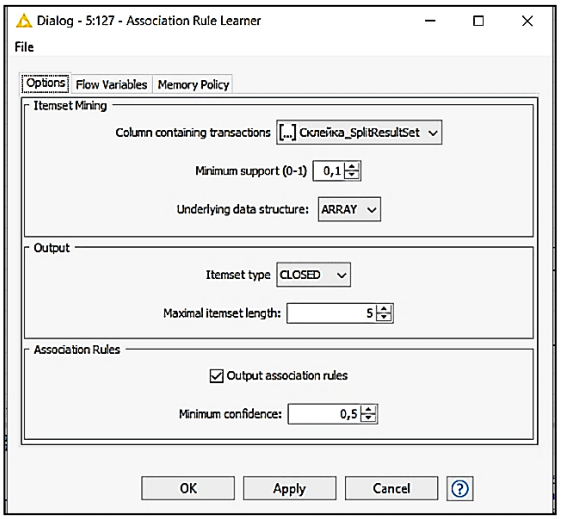


Рис. 4.16

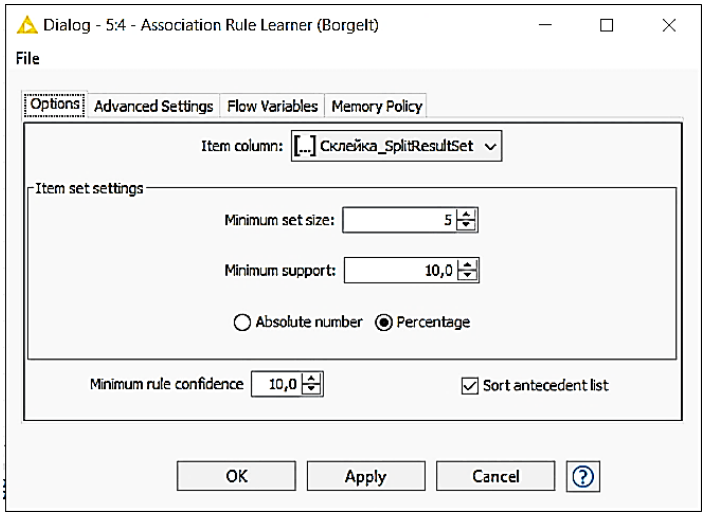


Рис. 4.17

Якщо все налаштовано коректно (Рис. 4.19), то вузли виконаються без зауважень у консольному вікні (Рис. 4.18).

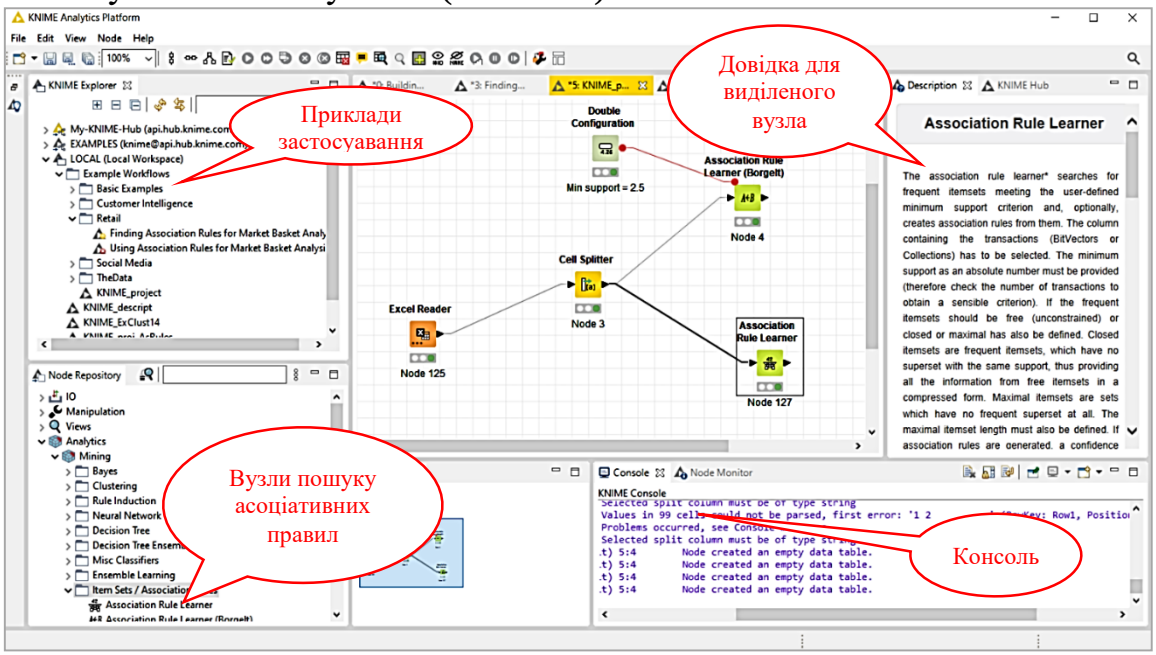


Рис. 4.18

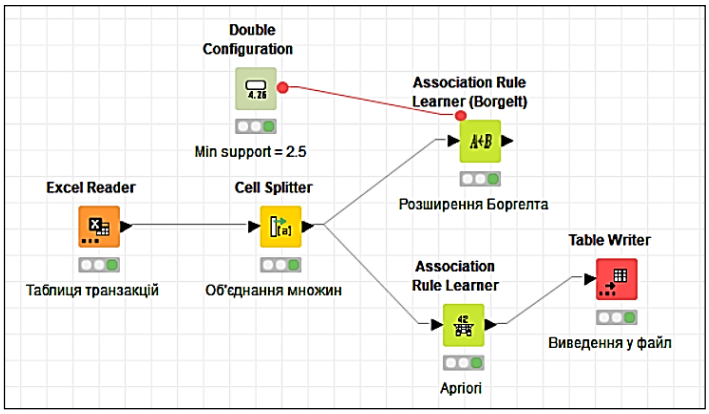


Рис. 4.19

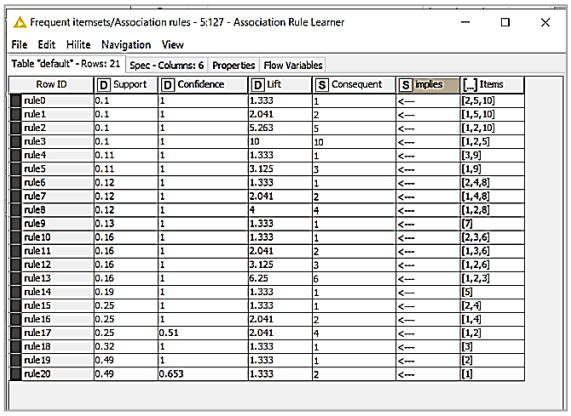


Рис. 4.20

Результатом застосування вузла **Association Rule Learner** буде таблиця (Рис. 4.20).

Розглянемо приклад, наведений у п. 4.3 для пакета Rapid Miner: завантажимо файл **чек_Запит**.xlsx формату *.xls або *.xlsx (вузол **Excel Reader**). Потік команд для виконання аналізу асоціативних правил наведено на Рис. 4.21:

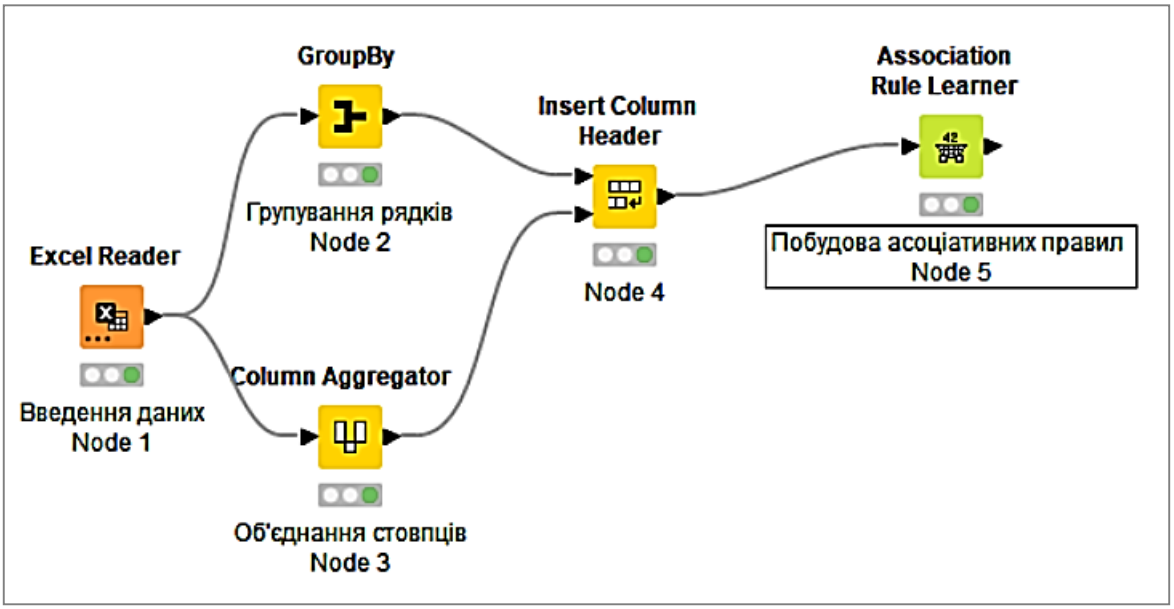


Рис. 4.21

Результатом виконання Вузла 1 (Введення даних з файла Excel) буде таблиця (Рис. 4.22).

Для групування рядків (Вузол 2) за номерами чеків застосовують групову операцію «Set» – утворення множини (Рис. 4.23).

На Рис. 4.24 та Рис. 4.25 показано налаштування цього вузла та результати відповідно.

Вузол 3 - Агрегація стовпців (товари) – призначений для об’єднання значень, розміщених у стовпцях. Тут застосовуємо операцію склеювання (Concatenate) (Рис. 4.26, Рис. 4.27).

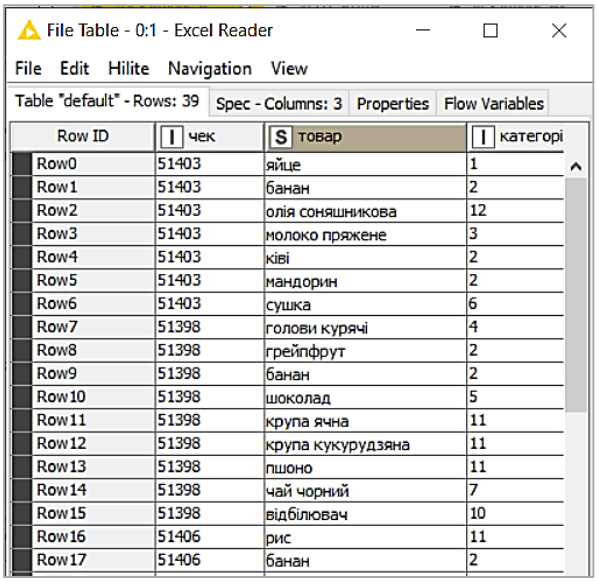


Рис. 4.22

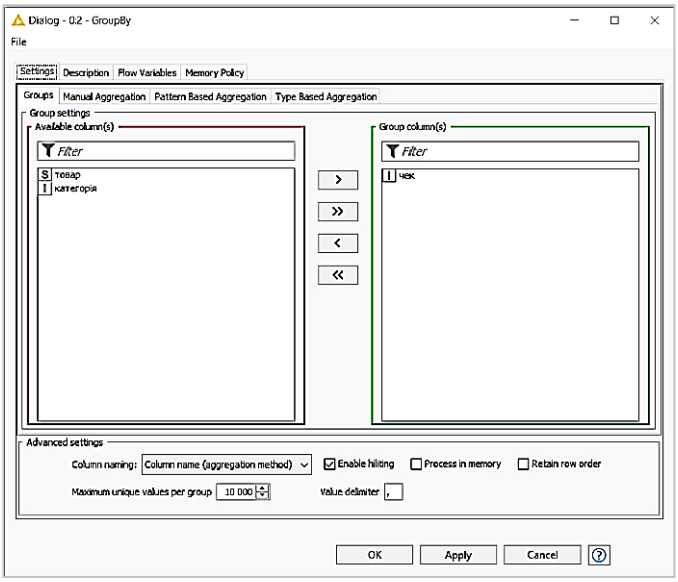


Рис. 4.23

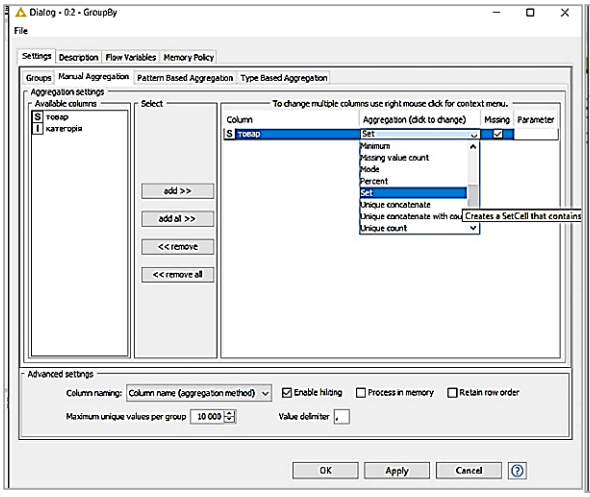


Рис. 4.24

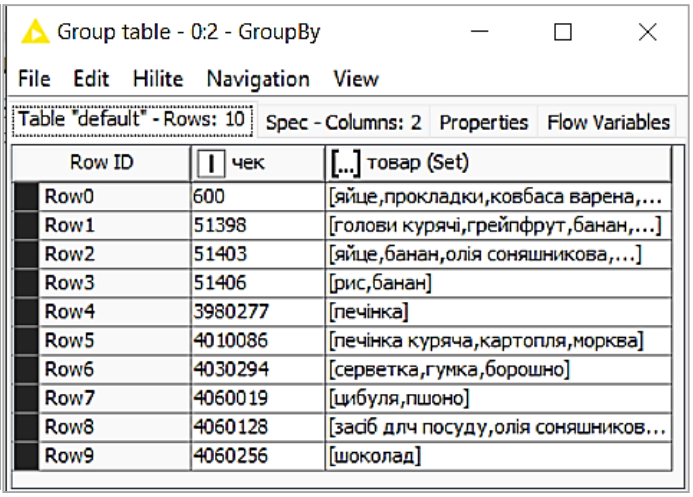


Рис. 4.25

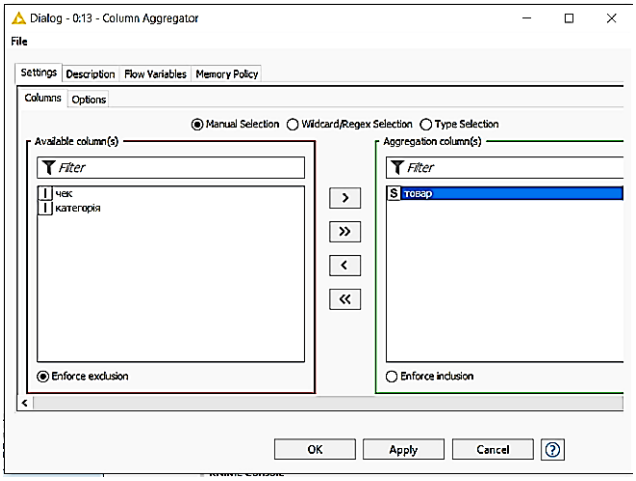


Рис. 4.26

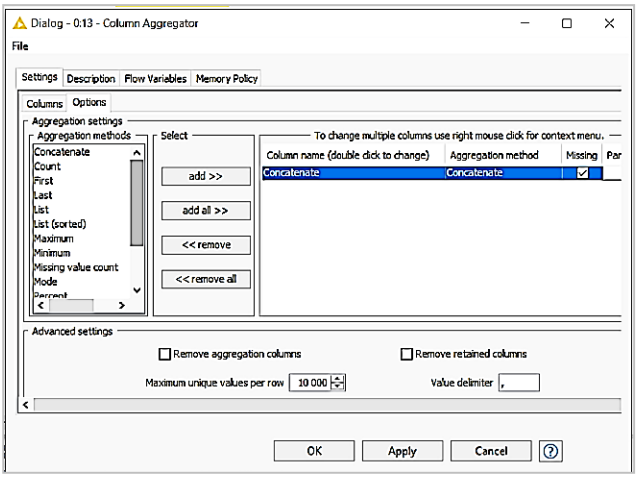


Рис. 4.27

Вузол 4 призначений для того, щоб визначити назви стовпців. У даному прикладі назвами будуть товари з утворених на попередніх кроках об’єднань (Рис. 4.28).

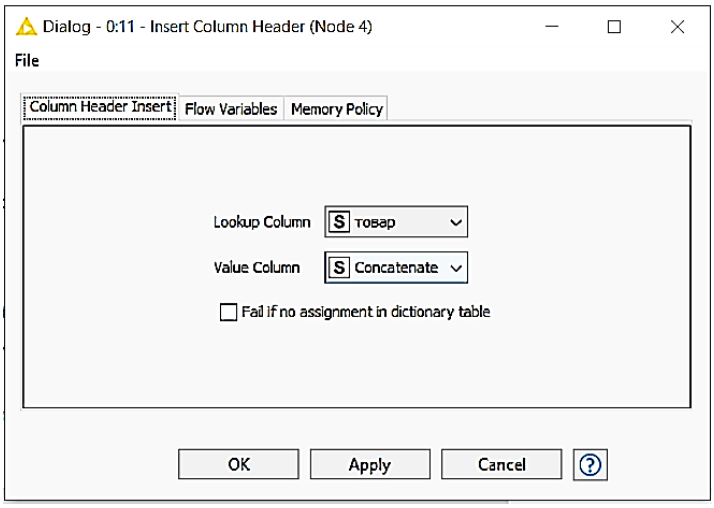


Рис. 4.28

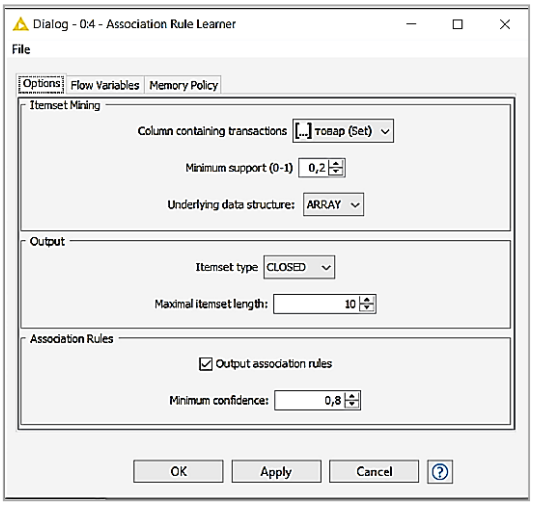
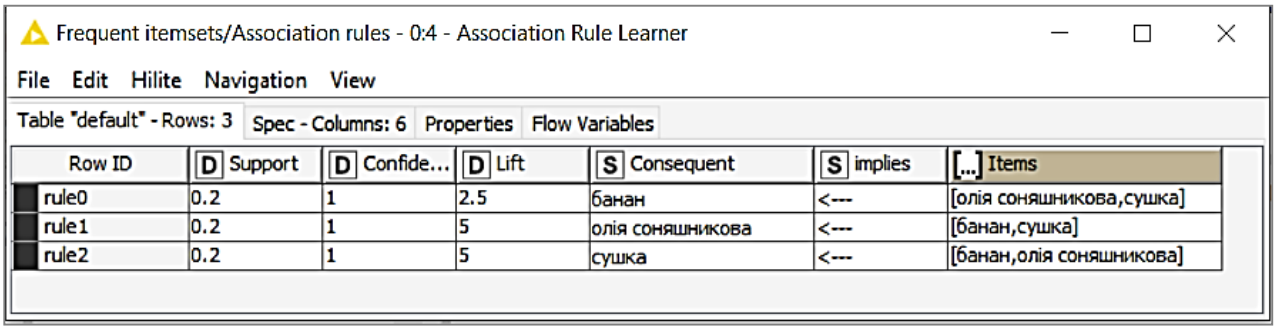


Рис. 4.29

Після виконання усіх підготовчих дій можна приступати до побудови асоціативних правил (Вузол 5): необхідно спочатку налаштувати параметри пошуку – визначити мінімальні значення підтримки та достовірності правил (Рис. 4.29).

Для наведених параметрів (мінімальна підтримка=0,2 та мінімальна достовірність=0,8) отримаємо результати як на Рис. 4.30. При зменшенні мінімальних значень кількість правил збільшиться. Можливо серед них будуть і цікавіші.



Row ID	Support	Confide...	Lift	Consequent	implies	Items
rule0	0.2	1	2.5	банан	<---	[олія соняшникова,сушка]
rule1	0.2	1	5	олія соняшникова	<---	[банан,сушка]
rule2	0.2	1	5	сушка	<---	[банан,олія соняшникова]

Рис. 4.30

Якщо дані для аналізу подано у форматі *.csv, то потік команд буде таким, як показано на Рис. 4.31:

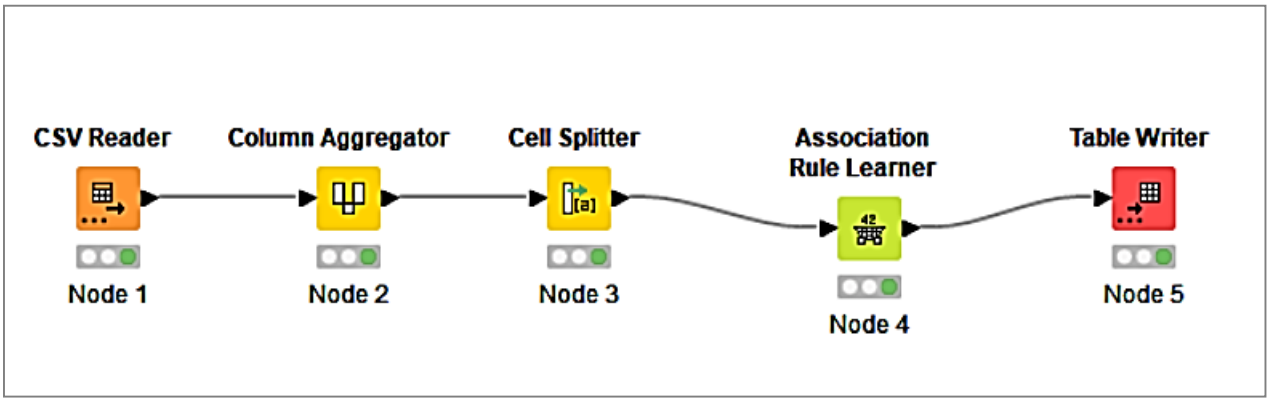


Рис. 4.31

Тут у Вузлі 1 необхідно вказати назву файла, розділовий знак, який використано для відокремлення значень, та наявність або відсутність заголовків

стовпців (Рис. 4.32). В результаті вміст файлу (розглянемо для прикладу файл groceries.csv[27]) буде прочитано у таблицю (Рис. 4.33):

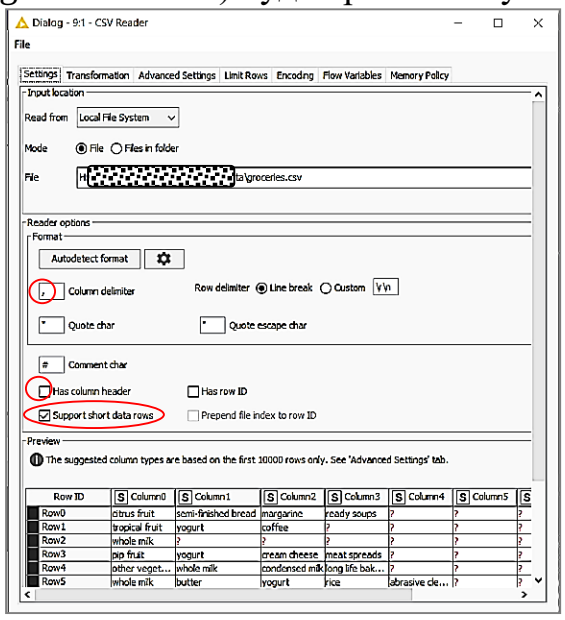


Рис. 4.32

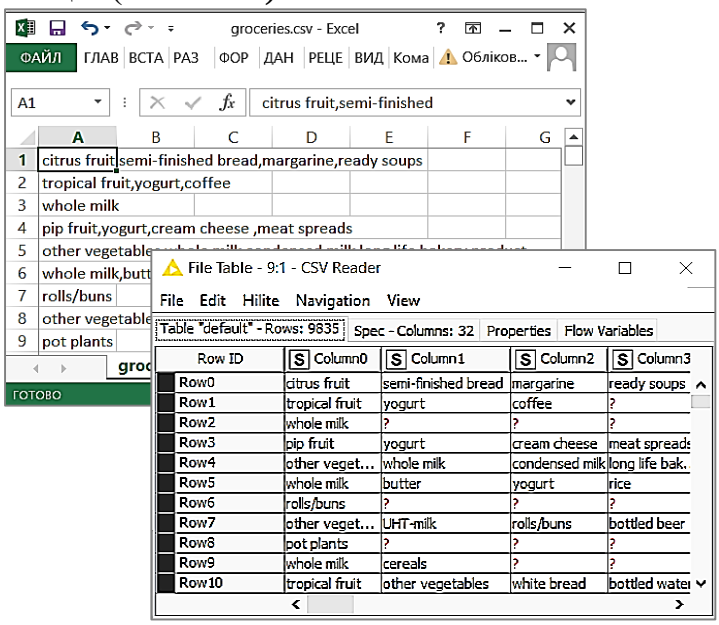


Рис. 4.33

У Вузлі 2 (Column Aggregator) до значень стовпців буде застосовано склеювання («Concatenate»), а у Вузлі 3 (Cell Splitter) зі склеєних значень буде утворено множину (Рис. 4.34):

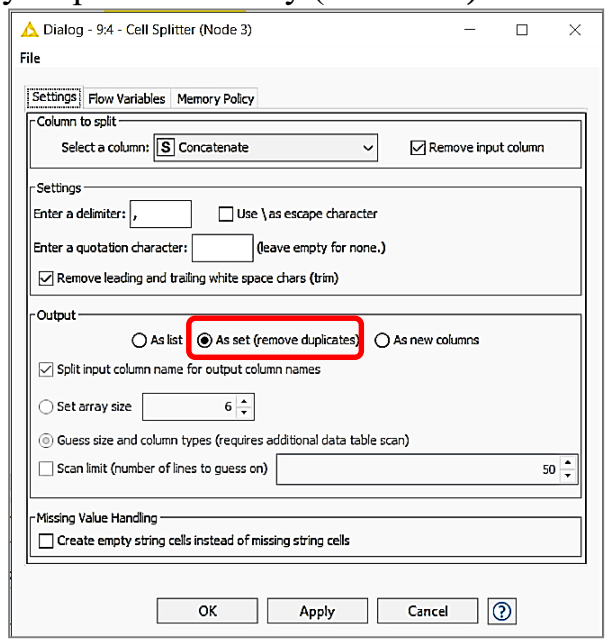


Рис. 4.34

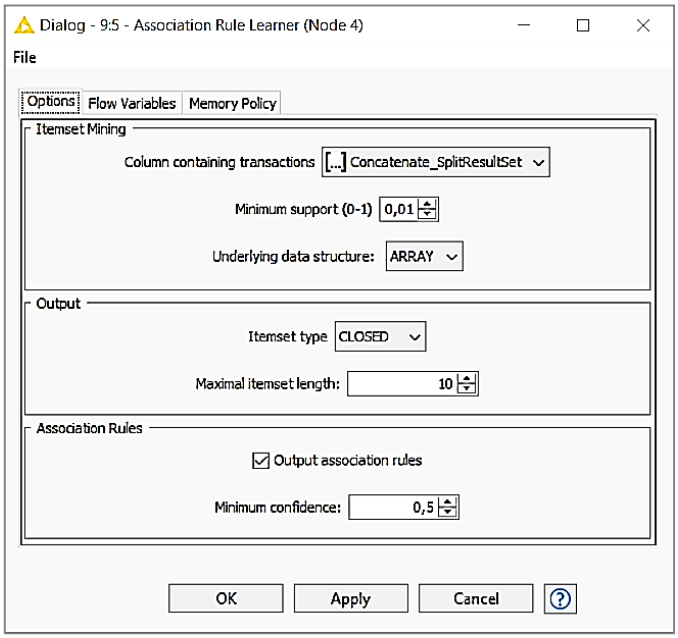
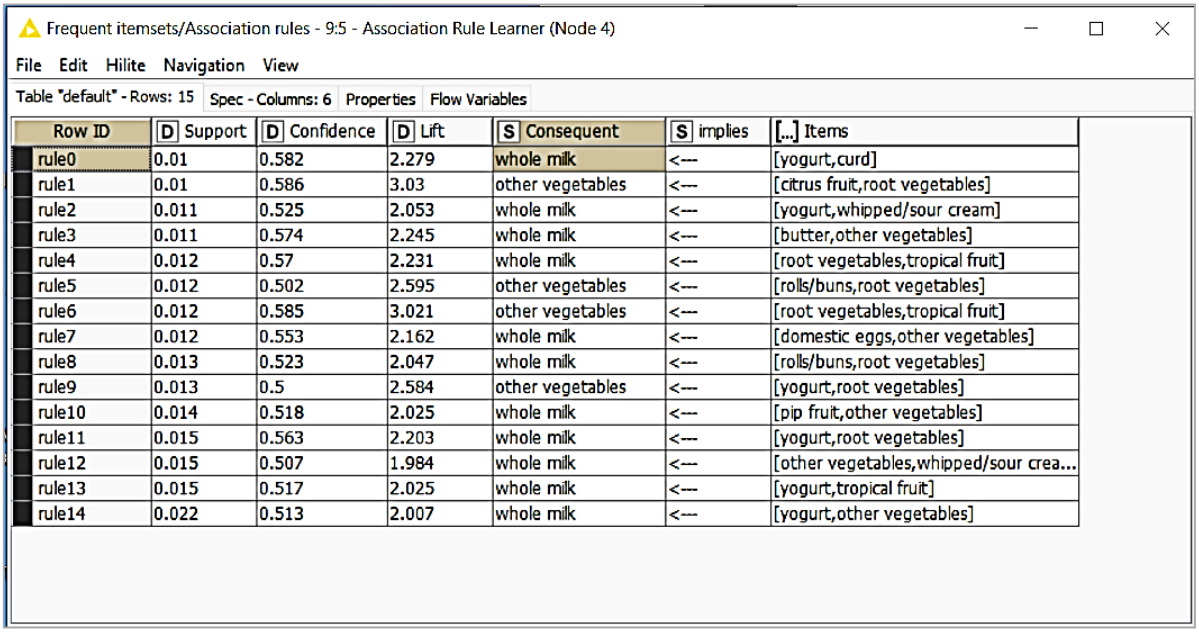


Рис. 4.35

У Вузлі 4 (Association Rules Learner) потрібно налаштувати параметри для пошуку асоціативних правил (Рис. 4.35). Для наведених параметрів (*мінімальна підтримка = 0,01* та *мінімальна достовірність = 0,5*) отримаємо 14 правил (Рис. 4.36). Всі вони мають значення ліфта більші за одиницю, тобто потенційно цікаві. Остаточне рішення про корисність отриманих правил має приймати експерт.

[27] https://onlinefreetutorial.com/association-rule/groceries.csv



Row ID	Support	Confidence	Lift	Consequent	implies	Items
rule0	0.01	0.582	2.279	whole milk	<---	[yogurt,curd]
rule1	0.01	0.586	3.03	other vegetables	<---	[citrus fruit,root vegetables]
rule2	0.011	0.525	2.053	whole milk	<---	[yogurt,whipped/sour cream]
rule3	0.011	0.574	2.245	whole milk	<---	[butter,other vegetables]
rule4	0.012	0.57	2.231	whole milk	<---	[root vegetables,tropical fruit]
rule5	0.012	0.502	2.595	other vegetables	<---	[rolls/buns,root vegetables]
rule6	0.012	0.585	3.021	other vegetables	<---	[root vegetables,tropical fruit]
rule7	0.012	0.553	2.162	whole milk	<---	[domestic eggs,other vegetables]
rule8	0.013	0.523	2.047	whole milk	<---	[rolls/buns,root vegetables]
rule9	0.013	0.5	2.584	other vegetables	<---	[yogurt,root vegetables]
rule10	0.014	0.518	2.025	whole milk	<---	[pip fruit,other vegetables]
rule11	0.015	0.563	2.203	whole milk	<---	[yogurt,root vegetables]
rule12	0.015	0.507	1.984	whole milk	<---	[other vegetables,whipped/sour crea...
rule13	0.015	0.517	2.025	whole milk	<---	[yogurt,tropical fruit]
rule14	0.022	0.513	2.007	whole milk	<---	[yogurt,other vegetables]

Рис. 4.36

4.5. Завдання до виконання

Оцінювання:

Вид роботи	**Вид звітності**	**Оцінка**
Індивідуальне завдання	Реферативна доповідь	до 5 балів
Вправи	Письмовий звіт	5 балів за кожну
Проект на створення програми	Програма	до 10 балів
Лабораторна робота	Звіт з висновками	до 10 балів
Тест		до 5 балів
Максимальна оцінка за тему	**20 балів**	формується із завдань на вибір студента

Контрольні питання

1. Наведіть приклади асоціацій.
2. Дайте формальне визначення поняття «асоціативне правило».
3. Які асоціативні правила називають корисними?
4. Що таке транзакція?
5. Назвіть характеристики асоціативних правил.
6. Що таке «підтримка» асоціативного правила?
7. Що таке «достовірність» асоціативного правила?
8. Для чого потрібна характеристика «покращення»?
9. Які є типи асоціативних правил?
10. Які правила називають узагальненими? Що можна сказати про характеристики узагальнених правил?
11. У чому полягає сутність алгоритму Apriori?
12. Чим інші відомі алгоритми пошуку асоціативних правил відрізняються від Apriori?

13. Перелічіть алгоритми пошуку асоціативних правил та назвіть їхні особливості.
14. Яким чином можуть бути представлені результати пошуку асоціативних правил?
15. У чому полягає алгоритм FP-grows? Яким чином мають бути представлені дані для нього?
16. З яких кроків складається процес пошуку асоціативних правил в пакеті RapidMiner?
17. Які результати дають процедури пошуку асоціативних правил? Як їх інтерпретувати?

Завдання до лабораторної роботи :

1. Ознайомтеся з прикладами виконання аналізу асоціативних правил у пакеті KNIME[28].
2. Підберіть в інтернеті дані для аналізу або скористайтеся даними, згенерованими до вправ 4.1 та 4.2 (див. Вправи), чи наборами, запропонованими у літературі до даної роботи (див. далі).
3. Підготуйте дані для аналізу у вигляді таблиць. За необхідності, виконайте дії щодо перетворення даних у своєму наборі даних. Переконайтеся, що всі змінні мають узгоджені для подальших процедур аналізу дані та типи.
4. Виконайте пошук асоціативних правил засобами KNIME або RapidMine та ще одного з пакетів (наприклад, RulesWizard, Deductor Studio, WizWhy, See5 або інш.).
5. Налаштуйте параметри пошуку правил (підтримку та достовірність) так, щоб визначити найбільш цікаві правила або отримати невелику кількість (10-20) правил для інтерпретації. Розгляньте інші показники сили правил, такі як, наприклад, Lift або Conviction.
6. *Візуалізуйте результати пошуку усіма доступними в обраному пакеті способами. З’ясуйте переваги та недоліки різних способів візуалізації правил.
7. Проінтерпретуйте отримані результати, виявлення цікавих правил.
8. Зробіть та задокументуйте свої висновки: Які правила ви знайшли? Які атрибути найбільш сильно пов’язані один з одним. Які правила виявилися несподіваними для Вас? Чому, на вашу думку, таке може бути? Скільки різних значень підтримки та достовірності довелося перевірити, перш ніж були знайдені деякі асоціативні правила? Чи були якісь із знайдених правил достатньо корисними, щоб їх можна було використати для прийняття рішень? Чому так, або чому ні? Які способи представлення асоціативних правил були найбільш переконливими або зрозумілими для вас? Чим сподобався/не сподобався використаний інструментарій, тощо.
9. Оформіть звіт до лабораторної роботи за таким планом:
    1. Опис даних (посилання на джерело даних, перелік та опис використовуваних змінних).

[28] https://hub.knime.com/knime/spaces/Examples/latest/50_Applications/16_MarketBasketAnalysis~4Gfp_9KDLU0Hijq7/

2. Використовуваний інструмент (вказати пакет та використані у ньому оператори чи процедури).
3. Встановлені параметри пошуку асоціативних правил:

Підтримка	Достовірність	Кількість правил

4. Отримані візуальні результати аналізу (скріншоти або інше).
5. Висновки.

10. Підготуйтеся до тестування з теми, відповівши на контрольні запитання.
11. Виконайте 2-3 вправи із наведених нижче.

Вправи

4.0. Розглянемо 8 кошиків, до складу яких входять товари-слова [23]:

1. {cat, and, dog, bites}.
2. {Yahoo, news, claims, a, cat, mated, with, a, dog, and, produced, viable, offspring}.
3. {cat, killer, likely, is, a, big, dog}.
4. {Professional, free, advice, on, dog, training, puppy, training}.
5. {cat, and, kitten, training, and, behavior}.
6. {dog, &, cat, provides, dog, training, in, Eugene, Oregon}.
7. {dog, and, cat, is, a, slang, term, used, by, police, officers, for, a, male–female, relationship}.
8. {Shop, for, your, show, dog, grooming, and, pet, supplies}.

а) визначити одноелементні набори з підтримкою не нижче 3;
б) визначити двоелементні набори з підтримкою не нижче 3;
в) визначити триелементні набори з підтримкою не нижче 3;
г) визначити достовірність правила {cat, dog} ⇒ {a};
д) визначити достовірність правила {cat } ⇒ {kitten};
е) визначити достовірність правила {dog} ⇒ {cat};
ж) визначити інтерес правила {dog} ⇒ {cat}.

4.1. Нехай маємо 100 товарів, пронумерованих від 1 до 100, та 100 кошиків, також пронумерованих від 1 до 100. Вважатимемо, що товар ***i*** входить до кошика ***b***, тоді і тільки тоді, коли ***i*** ділить ***b*** без остачі. Таким чином товар №1 входить до всіх кошиків, товар №2 входить до складу 50-ти кошиків з парними номерами і так далі. Кошик №12 складається з товарів {1, 2, 3, 4, 6, 12}, оскільки усі ці числа є дільниками числа 12 [23, 6.1.1].

Дайте відповіді на запитання:

а) Якщо поріг підтримки дорівнює 5, які товари є частими?
б) Якщо поріг підтримки дорівнює 5, які пари товарів є частими?
в) Який сумарний розмір усіх кошиків?
г) який кошик найбільший?
д) яка достовірність правила {5, 7} → 2?
е) яка достовірність правила {2, 3, 4}→ 5?
ж) опишіть усі правила, що мають 100% достовірність.

4.2. Нехай маємо 100 товарів, пронумерованих від 1 до 100, та 100 кошиків, також пронумерованих від 1 до 100. Вважатимемо, що товар ***i*** входить до кошика ***b*** тоді і тільки тоді, коли ***b*** ділить ***i*** без остачі. Наприклад, кошик №12 містить товари {12, 24, 36, 48, 60, 72, 84, 96} [23, 6.1.3]. Повторіть вправу 4.1 для цих даних.

Дайте відповіді на запитання:

а) Якщо поріг підтримки дорівнює 5, які товари є частими?

б) Якщо поріг підтримки дорівнює 5, які пари товарів є частими?

в) Який сумарний розмір усіх кошиків?

г) який кошик найбільший?

д) яка достовірність правила $\{24, 60\} \rightarrow 8$?

е) яка достовірність правила $\{2, 3, 4\} \rightarrow 5$?

ж) опишіть усі правила, що мають 100% достовірність.

4.3. Нехай маємо товари, пронумеровані від 1 до 10. Нехай кожен кошик включає товар ***i*** з імовірністю 1/***i*** незалежно від інших (набори товарів жодним чином не пов'язані один з одним). Тобто усі кошики містять товар №1, половина – товар №2, третина – товар №3 і так далі. Припустимо, що кількість кошиків досить велика для того, щоб сукупність кошиків можна було розглядати як статистичну. Нехай поріг підтримки становить 1% від кількості кошиків [23, 6.1.4].

а) знайдіть часті множини;

б) перевірте, що серед даних немає цікавих правил, тобто всі правила мають *інтерес*=0.

4.4. Нехай база даних містить дані про чотири покупки [16, с.39]:

ID транзакція	Дата	Придбано товари
100	15.10.2019	{K,A,D,B}
200	15.10.2019	{D,A,C,E,B}
300	19.10.2019	{C,A,B,E}
400	22.10.2019	{B,A,D}

Якщо мінімальна частота дорівнює 60%, а мінімальна достовірність – 80%:
а) знайдіть усі набори, що часто зустрічаються, за допомогою алгоритма Apriori.
б) визначте усі асоціативні правила.

4.5. Наведіть приклад, коли елементи в асоціативному правилі корелюють негативно [22, 6.13].

4.6. У базі даних п'ять транзакцій. Нехай мінімальна підтримка = 60%, а мінімальна достовірність = 80% [22, 6.6]:

Транзакція_ID	набір елементів
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, R, I, E}

а) знайдіть усі часті набори, використовуючи алгоритми Apriori та FP-growth. Порівняйте ефективності двох процесів;

б) створіть список усіх строгих асоціативних правил[29] , що відповідають метаправилу

$$\forall x \in transaction,\ buys(X,item_1) \wedge buys(X,item_2) \Rightarrow buys(X,item_3)\ [s,c],$$

де X – змінна, що представляє покупця, а $item_i$ – представляють покупки, наприклад «A», «B» і т.п.

4.7. База даних складається з чотирьох транзакцій. Нехай мінімальна підтримка = 60%, а мінімальна достовірність = 80% [22, 6.8]:

покупець	Транзакція_ID	набір елементів
01	T100	{краби, молоко, сир, хліб}
02	T200	{сир, молоко, яблука, тістечко, хліб}
03	T300	{яблука, молоко, хліб, тістечко}
04	T400	{хліб, молоко, сир}

а) для елемента, наприклад, *молоко* за шаблоном правила

$$\forall X \in transaction,\ buys(X,item_1) \wedge buys(X,item_2) \Rightarrow buys(X,item_3)\ [s,c]$$

перелічіть усі часті множини з k елементів для максимального k та усі строгі асоціативні правила (з підтримкою s та достовірністю c), що стосуються частих k-елементних множин для максимального k;

б) для елемента, наприклад, *молоко* за шаблоном правила

$$\forall X \in transaction,\ buys(X,item_1) \wedge buys(X,item_2) \Rightarrow buys(X,item_3)\ [s,c]$$

перелічіть усі часті множини з k елементів для максимального (але не формуйте жодних правил).

4.8. Таблиця спряженості містить узагальнені дані про транзакції супермаркетів, які *містять* та *не містять* хот-доги, і транзакції, які *містять* та *не містять* гамбургери [22, 6.14]:

	з хот-догами	без хот-догів	Разом по рядку
з гамбургерами	2000	500	2500
без гамбургерів	1000	1500	2500
Разом по стовпцю	3000	2000	5000

а) припустимо, знайдено асоціативне правило «хот-дог→гамбургер». Якщо мінімальна підтримка 25%, а мінімальна достовірність 50%, чи є це правило строгим[30]?

б) чи можна вважати на основі цих даних, що покупка хот-дога не залежить від покупки гамбургера? Якщо ні, то який зв'язок (кореляція) між ними існує?

4.9. Нехай знайдені у великій базі часті набори елементів для заданого мінімального рівня частоти збережені у великій базі даних транзакцій, DB. Запропонуйте ефективні способи пошуку асоціативних правил з тою самою мінімальною підтримкою, якщо до початкової бази DB було добавлено невелику множину транзакцій, наприклад, ΔDB [22, 6.10].

4.10. Алгоритм Apriori використовує попередні знання про властивості підтримки піднабору [22, 6.3].

---

[29] Строгим вважають правило з підтримкою не меншою ***S*** і достовірністю не меншою ***C***.

[30] Див. попередню зноску.

а) доведіть, що всі непусті підмножини частих наборів повинні також бути частими;

б) доведіть, що підтримка непорожньої підмножини ***s′*** множини s має бути принаймні такою ж великою, як і підторимка ***s***;

в) дано часту множину *l* та її підмножину *s*. Доведіть, що достовірність правила *s′→(l-s′)* не може бути більшою за достовірність правила *s→(l-s)*, де *s′⊂s*;

г) один з варіантів алгоритму *Apriori* поділяє транзакції бази даних *D* на *n* підрозділів, що не перекриваються. Доведіть, що будь-який набір елементів, який часто зустрічається в D, повинен бути частим принаймні в одному підрозділі D.

4.11. Нехай набір *c* є кандидатом до множини C_k, що генерується алгоритмом Apriori. Скільки наборів довжини (*k-1*) необхідно перевірити на етапі скорочення? [22, 6.4]

4.12. Розгляньте метод генерації асоціативних правил з множини частих наборів (Apriori). Запропонуйте більш ефективний метод. Обгрунтуйте його ефективність. [22, 6.5]

Альтернативні завдання: проекти на створення програмних продуктів:

Порада: для тестування розроблених програм скористайтеся числовими множинами, наведеними на сайті http://fimi.uantwerpen.be/data/, зокрема T10I4D100K (.gz) або T40I10D100K (.gz), чи самостійно згенеруйте набір даних так, як пропонується у вправах 4.1 чи 4.2. Зверніть увагу на файл ВЕЛИКОГО розміру webdocs.dat.gz (488 MB!)

4.п1. Використовуючи мову програмування, яку ви добре знаєте, наприклад, C++ або Java, реалізуйте один з алгоритмів пошуку частих наборів: 1) Apiori, 2) FP-growth, 3) PCY або інш. Порівняйте продуктивності цих алгоритмів на різних наборах великих даних. Напишіть звіт з аналізом ситуації (розмір даних, розподіл даних, мінімальний поріг підтримки, кількість отриманих правил (*pattern density*), де один з алгоритмів працює краще за інші і чому [22, 6.7]. *(три учасника)*

4.п2. Було запропоновано багато методів для подальшого поліпшення ефективності алгоритмів добування частих наборів елементів. Використовуючи, наприклад, алгоритми, що базуються на застосуванні FP-дерева (наприклад, FP-зростання), реалізуйте один з наведених нижче методів оптимізації. Порівняйте продуктивність вашої нової реалізації з неоптимізованою версією [22, 6.12]

а) Метод добування частих наборів, описаний у Розділі 6.2.4 [22], застосовує FP-дерево для формування бази шаблонів з використанням методу проектування знизу-вгору. Проте, можна розробити техніку проектування згори-вниз. Розробити та реалызувати такий метод побудови FP-дерев зверху вниз. Порівняйте його продуктивність з методом проектування знизу вгору.

б) Вузли і покажчики в FP-дереві в алгоритмі FP-зростання використовуються рівномірно. Однак така структура може потребувати невиправдано багато місця, коли дані є розрідженими. Однією з можливих альтернатив є

застосування гібридної реалізації на основі масиву та покажчика, де вузол може зберігати кілька елементів, якщо він не містить точки розщеплення для декількох гілок. Розробити таку реалізацію та порівняти її з оригінальною.

4.п3. База даних DBLP містить більше мільйона записів про дослідження, опубліковані на конференціях та у журналах з комп'ютерних наук. Серед них велика кількість авторів, що мають співавторські відношення. [22, 6.15]

а) Запропонуйте метод ефективного пошуку множини авторів, які тісно пов'язані, наприклад, часто публікуються разом.

б) На основі методів пошуку та методів оцінювання результатів пошуку патернів, обговорюваних у даному розділі, обговоріть, які методи можуть допомогти краще виявити моделі (патерни) співробітництва авторів краще за інші.

в) На підставі дослідження (а) розробити метод, який може приблизно зпрогнозувати відношення консультанта та консультованого та приблизний період такого консультативного нагляду.

4.п4. Набір X називають генератором на множині наборів D, якщо не існує підмножини $Y \subset X$, такої, що *support(X)=support(Y)*. Генератор X є частим генератором, якщо *support(X)* більша за мінімальний поріг підтримки. Нехай G – множина всіх частих генераторів на множині наборів даних D [22, 6.2].

а) чи можна визначити, чи є набір елементів A частим, та підтримку A, якщо він частий, використовуючи тільки G, і підраховуючи підтримки усіх частих генераторів? Якщо так, представте свій алгоритм. Якщо ні, яка ще інформація потрібна? Чи можна представити алгоритм, якщо потрібну інформацію представлено?

б) який зв'язок між закритими наборами та генераторами?

4.п5. Нехай велика мережа магазинів має базу транзакцій, розподілену між чотирма відділеннями. Транзакції кожного компонента бази даних мають однаковий формат T_j:{i_1, …, i_m}, де T_j – ідентифікатор транзакції, а i_k ($1 \le k \le m$) – ідентифікатор товару, придбаного у транзакції. Запронуйте та представте схематично ефективний алгоритм добування глобальних асоціативних правил (тобто правил, що стосуються мережі в цілому, а не окремого відділення). Алгоримт не повинен вимагати передавання усіх даних на один сайт та не повинен створювати надмірне навантаження на комунікаційну мережу. [22, 6.9]

4.п6. Зазвичай в алгоритмах добування шаблонів розглядають елементи у транзакції лише як множину, тобто без врахування кількості входжень окремих елементів. Тим не менше, випадки, коли елемент в одному кошику для покупок зустрічається декілька разів, наприклад, чотири торти і три пакети молока, може бути важливим ву аналізі транзакційних даних. Як в процедурі добування частих наборів можна ефективно врахувати такі випадки? Запропонуйте модифікації відомих алгоритмів, таких як Apriori і FP-зростання, з урахуванням такої ситуації [22, 6.11].

Література до розділу:

- ✓ Han J., Kamber M., Pei J. Data mining: Concepts and Techniques. – Morgan Kaufmann Publishers. – 2012. – 740 p.
- ✓ Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP.
- ✓ Інтелектуальний аналіз даних: підручник /О.І.Черняк, П.В.Захарченко. – К.:Знання, 2014.
- ✓ Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets http://www.mmds.org
- ✓ Прокопенко Н.Ю. Системы поддержки принятия решений //Bstudy - статьи для высших учебных заведений (info{ат}bstudy.net) © 2017 – 2023 – URL: https://bstudy.net/926773/tehnika/primenenie_assotsiativnyh_pravil_otsenki_vybora_upravlencheskih_resheniy
- ✓ Шитиков В. К., Мастицкий С. Э. (2017) Классификация, регрессия, алгоритмы Data Mining с использованием R. - https://github.com/ranalytics/data-mining
- ✓ Ассоциативные правила, или пиво с подгузниками https://habr.com/ru/company/ods/blog/353502/
- ✓ Association_rules: Association rules generation from frequent itemsets. – URL: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

Набори даних до лабораторної роботи:

1) https://onlinefreetutorial.com/association-rule/groceries.csv

2) https://gist.github.com/Harsh-Git-Hub/2979ec48043928ad9033d8469928e751#file-retail_dataset-csv

3) https://drive.google.com/file/d/1y5DYn0dGoSbC22xowBq2d4po6h1JxcTQ/view?usp=sharing

ТЕМА 5. АНАЛІЗ ЗВ'ЯЗКІВ.

5.1. Інформаційний пошук

Інформаційний пошук (ІП) (англ. *Information retrieval*) – наука про пошук *неструктурованої* документальної інформації, що включає:

- пошук інформації в документах;
- пошук самих документів;
- добування метаданих з документів;
- пошук тексту, зображень, відео та звуку у локальних реляційних базах даних, у гіпертекстових базах даних таких, як Інтернет та локальні інтранет;
- сортування документів за частотою шуканої фрази.

Етапами інформаційного пошуку є:

1) кроулінг (crawling) – збір документів;
2) синтаксичний розбір;
3) аналіз;
4) *індексування**;***
5) пошук.

Якість пошуку визначається такими показниками (Рис. 5.1):

1. Точність (*precision*) – відношення кількості добутих релевантних документів до загальної кількості добутих документів = ***RR/Rd,*** тобто *відповідність знайденого шуканому.*
2. Вибірка (*recall*) – відношення кількості добутих релевантних документів до загальної кількості релевантних документів = ***RR/Rt.***

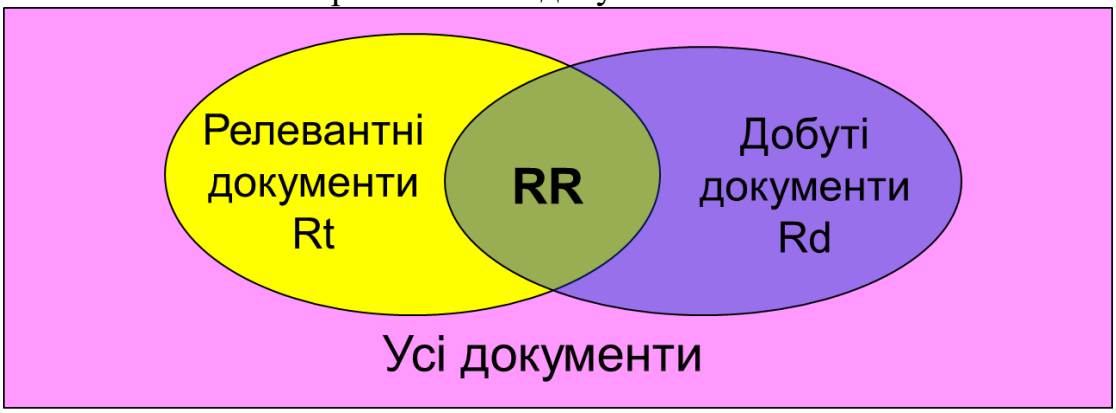


Рис. 5.1

5.2. Обчислення показників PageRank.

Степінь важливості веб-документа оцінюють кількістю підтверджень його зв'язків з іншими сторінками. Відповідно до показника PageRank[31] сторінка є

[31] *Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd.* The PageRank Citation Ranking: Bringing Order to the Web. – 1998; *Sergey Brin, Lawrence Page.* The Anatomy of a Large-Scale Hypertextual Web Search Engine. – 1998.

важливою, якщо до неї прив'язані важливі сторінки (Рис. 5.2). Іншими словами, якщо всі дороги ведуть до Рима, то Рим – важливе місто.

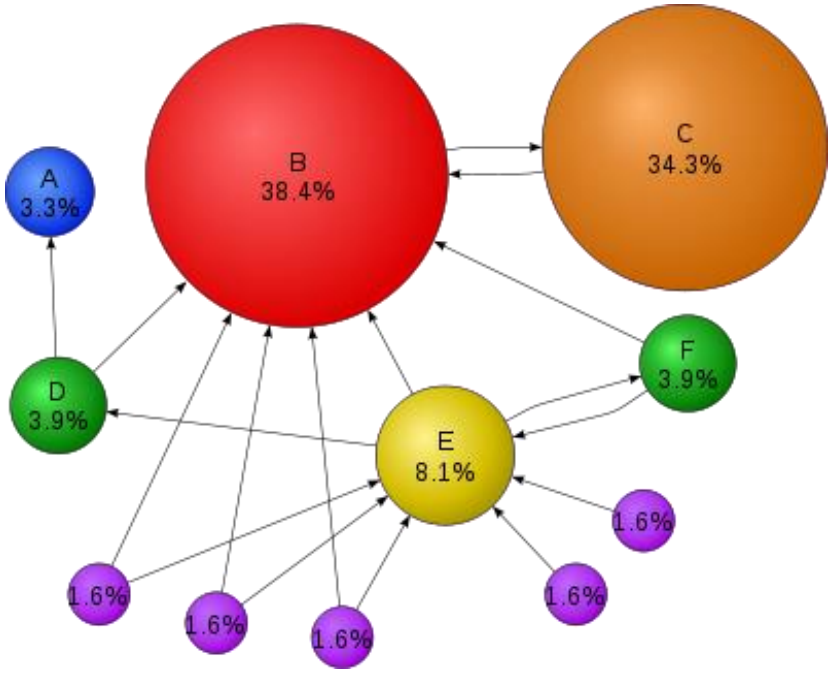


Рис. 5.2

До використання алгоритма PageRank пошукові системи здійснювали кроулінг веба та створювали списки термів, вишукуючи їх на кожній сторінці та створюючи зворотний індекс (*inverted index*). Коли пошуковий список сформований, сторінки з цим термом видобувають із зворотного індекса та ранжують відповідно до кількості входжень терма у сторінці (його релевантності). При такому підході релевантність сторінки можна штучно підвищити багатократним використанням терма, зокрема, завуальовано, наприклад, з використанням прихованого шрифту. Таке явище називають термінологічним спамом або термспамом.

Для подолання термспаму розробники алгоритму PageRank в компанії Google запропонували наступне:

1. Моделювати поведінку «випадкового серфера», який подорожує вебом, обираючи сторінки випадковим чином. Часто відвідувані сторінки вважають більш «важливими».

2. Про контент сторінки судять не тільки за термами самої сторінки, а і за термами пов'язаних сторінок, які значно важче сфальсифікувати.

Власне PageRank – це функція, яка кожній веб-сторінці ставить у відповідність дійсне число – «важливість» сторінки. Сторінка тим важливіша, чим вищий її PageRank.

Для обчислення PageRank веб слід представити як орієнтований граф. Для графа, зображеного на Рис. 5.3, матриця гіперпосилань M (*transition matrix*) показує, що з вузла A випадковий серфер може з однаковою імовірністю (p=1/3) перейти в кожен з решти вузлів B, C або D[32]. Вузли B та D пов'язані кожен з

[32] У матриці M елемент m_{ij} дорівнює імовірності, з якою випадковий серфер може потрапити з *i*-того вузла веба в *j*-тий вузол.

двома вузлами, тому імовірність переходів p=1/2. З вузла C можна перейти тільки у вузол A з імовірністю p=1. Нехай випадковий серфер у початковий момент знаходиться у будь-якому з вузлів (веб-сторінок). Тоді вектор V_0 складається з елементів 1/n, де n – загальна кількість веб-сторінок. Наступне значення вектора обчислюють множенням на матрицю гіперпосилань:

$$\boldsymbol{v}_i=\boldsymbol{M}\cdot\boldsymbol{v}_{i-1} \qquad (5.1)$$

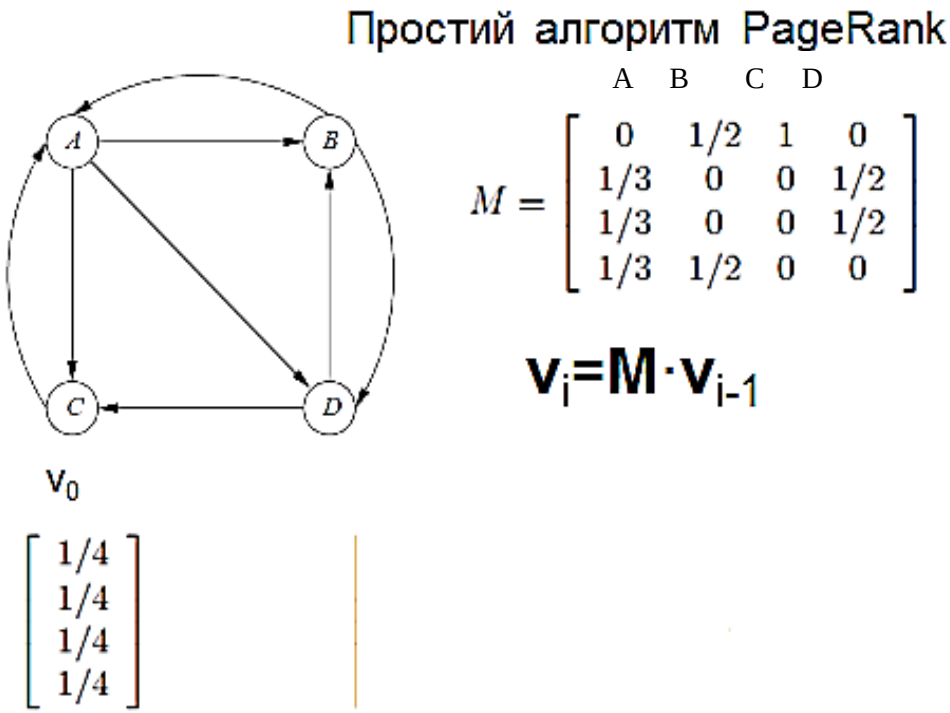


Рис. 5.3

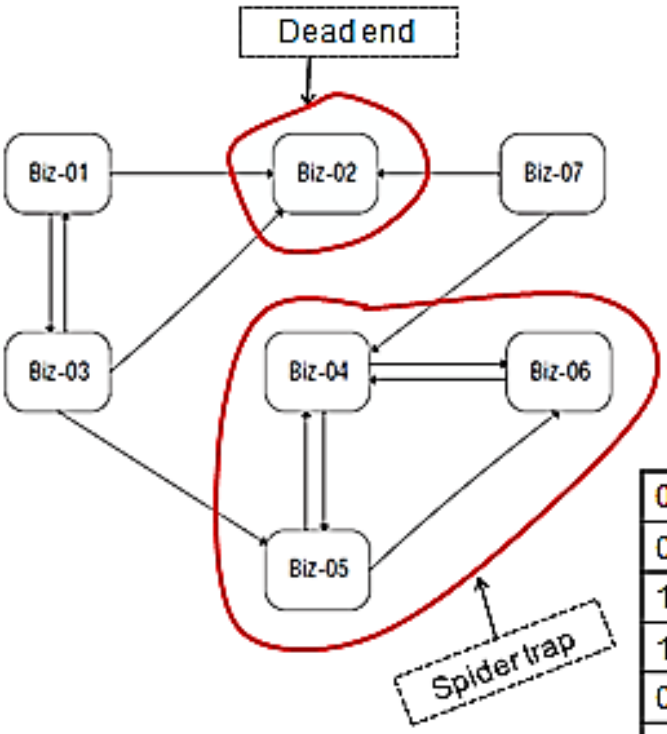
На *k*-тій ітерації отримують розподіл серфера у вебі після *k* кроків (Рис. 5.4). В решті решт результат сходиться до значення, яке і показує PageRank кожної сторінки. У даному випадку найбільший PageRank має перша сторінка, тобто сторінка A:

V_0 V_1 V_2 V_3

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}, \begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix}, \ldots, \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

Рис. 5.4

Простий PageRank коректно працює для «коректного» веба. Реальна ситуація складніша: крім строго пов'язаних компонентів (SCC – *strong connected components*) веб може містити компоненти, *з яких можна* потрапити в SCC в один бік, *в які можна* потрапити з SCC в один бік, ізольваних компонентів і т.д. Отже в реальному вебі часто виникають глухі кути або тупики (*dead ends*) та пастки (*spider traps*) (Рис. 5.5).



Приклад мережі веб-сторінок

Файл	Посилання на
Biz-01	Biz-02, Biz-03
Biz-02	Немає посилань
Biz-03	Biz-01, Biz-02, Biz-05
Biz-04	Biz-05, Biz-06
Biz-05	Biz-04, Biz-06
Biz-06	Biz-04
Biz-07	Biz-02, Biz-04

Матриця гіперпосилань =

01	02	03	04	05	06	07
0	0	1/3	0	0	0	0
1/2	0	1/3	0	0	0	1/2
1/2	0	0	0	0	0	0
0	0	0	0	1/2	1	1/2
0	0	1/3	1/2	0	0	0
0	0	0	1/2	1/2	0	0
0	0	0	0	0	0	0

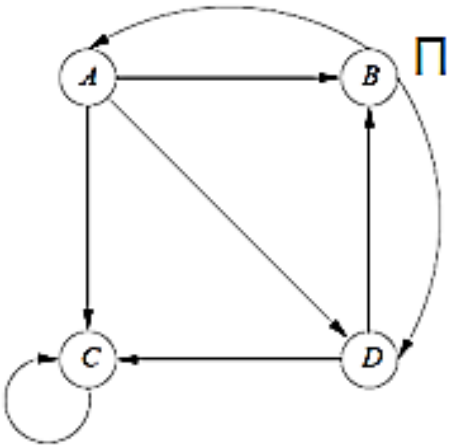
Рис. 5.5

Для обчислення PageRank з урахуванням пасток і тупиків вводять механізм «таксації» – скінчену імовірність для серфера вийти з веба на будь-якому етапі та запустити нового серфера на будь-яку сторінку. Таку імовірність позначають β.

$$\mathbf{v}' = \beta M \mathbf{v} + (1 - \beta)\mathbf{e}/n \qquad (5.2)$$

На Рис. 5.6 маємо структуру, в якій вершина C – пастка. Потрапивши на цю сторінку, серфер не зможе перейти на інші сторінки. За алгоритмом простого PageRank ця сторінка отримає найвищий коефіцієнт, а решта сторінок отримають PageRank=0. Модифікований PageRank дозволяє серферу «телепортацію» з пастки.

Покращити результати дозволяє і так званий тематично-чутливий PageRank: якщо дослідити зміст (контент) сторінок, то можна дати перевагу відвідувати лише тематично пов'язані сторінки, тобто сторінки з високою релевантністю до пошукового запиту. Так само можна обмежити відвідування сторінок, позначених як спам. Такий коефіцієнт ще називають TrustRank (Рис. 5.7). Для обчислення TrustRank формують множину S вершин, тематично пов'язаних або вільних від спаму, і в одиничному векторі e_S обнуляють компоненти не пов'язані з елементами множини S.



Подолання пасток (taxation)

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

$$\mathbf{v}' = \beta M \mathbf{v} + (1-\beta)\mathbf{e}/n$$

β=0,8
1-β=0,2
n=4

$$\mathbf{v}' = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \mathbf{v} + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/60 \\ 13/60 \\ 25/60 \\ 13/60 \end{bmatrix}, \begin{bmatrix} 41/300 \\ 53/300 \\ 153/300 \\ 53/300 \end{bmatrix}, \begin{bmatrix} 543/4500 \\ 707/4500 \\ 2543/4500 \\ 707/4500 \end{bmatrix}, \ldots, \begin{bmatrix} 15/148 \\ 19/148 \\ 95/148 \\ 19/148 \end{bmatrix}$$

Імовірність залишитися у пастці обмежена 14

Рис. 5.6

Тематично-чутливий PageRank

$$\mathbf{v}' = \beta M \mathbf{v} + (1-\beta)\mathbf{e}/n$$

$$\mathbf{v}' = \beta M \mathbf{v} + (1-\beta)\mathbf{e}_S/|S|$$

set $S = \{B, D\}$

$$\mathbf{v}' = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \mathbf{v} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$\begin{bmatrix} 0/2 \\ 1/2 \\ 0/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 2/10 \\ 3/10 \\ 2/10 \\ 3/10 \end{bmatrix}, \begin{bmatrix} 42/150 \\ 41/150 \\ 26/150 \\ 41/150 \end{bmatrix}, \begin{bmatrix} 62/250 \\ 71/250 \\ 46/250 \\ 71/250 \end{bmatrix}, \ldots, \begin{bmatrix} 54/210 \\ 59/210 \\ 38/210 \\ 59/210 \end{bmatrix}$$

Рис. 5.7

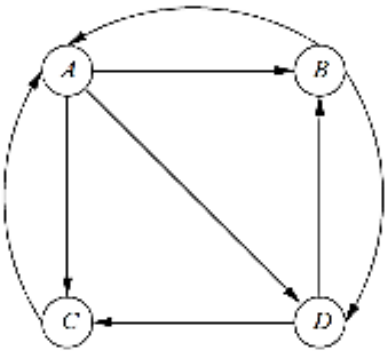
Узагальненим показником важливості сторінки є SpamMass (Рис. 5.8):

Обчислення SpamMass

SpamMass=(r-t)/r,

де r – PageRank,

t – TrustRank



Node	PageRank	TrustRank	Spam Mass
A	3/9	54/210	0.229
B	2/9	59/210	-0.264
C	2/9	38/210	0.186
D	2/9	59/210	-0.264

Тут немає спама

Якщо SM≈1, то сторінка імовірно є спамом.
Якщо SM<0, або SM≈0, то сторінка – не спам

21

Рис. 5.8

В оригінальній роботі авторів формула PageRank мала такий вигляд: припустимо, що на сторінку A вказують (тобто цитують її) сторінки $T_1 \ldots T_n$. Встановимо параметр d – коефіцієнт затухання, який може бути визначений в діапазоні від 0 до 1, – рівним 0,85. Позначимо C(A) – кількість посилань, що виходять зі сторінки A. Тоді PageRank сторінки A буде визначений так:

$$PR(A) = (1-d) + d\,(PR(T_1)/C(T_1) + \ldots + PR(T_n)/C(T_n))$$

5.3. HITS

Інший підхід до визначення важливості веб-сторінки запропоновано в алгоритмі HITS. На відміну PageRank HITS ранжує тільки відповіді на конкретний запит. Важливість сторінок визначається або її авторитетом (*authorities*) – сторінка важлива, оскільки забезпечує інформацію по темі, – або її посередництвом (*hub*) – сторінка є хабом, концентратором або посередником, якщо містить посилання на авторитетні сторінки.

Тобто, якщо для PageRank *«сторінка важлива, якщо до неї прив'язані важливі сторінки»*, то для HITS *«сторінка є гарним хабом, коли вона пов'язана з гарним авторитетом, і навпаки»*.

Для обчислення показників авторитетності та хабності спочатку будують матрицю зв'язків L (L_{ij}=1, коли сторінки i та j пов'язані) та транспонують її. Далі обчислюють $\bar{h}$ та $\bar{a}$ (Рис. 5.9). Тут λ та μ – масштабні коефіцієнти.

Граничні значення векторів $\bar{h}$ та $\bar{a}$ і будуть містити показники хабності та авторитетності для сторінок веба. Як видно з Рис. 5.9, сторінка A є найбільшим концентратором, оскільки вона пов'язана з найбільшими авторитетами. Відповідно, сторінки B та C – найбільші авторитети.

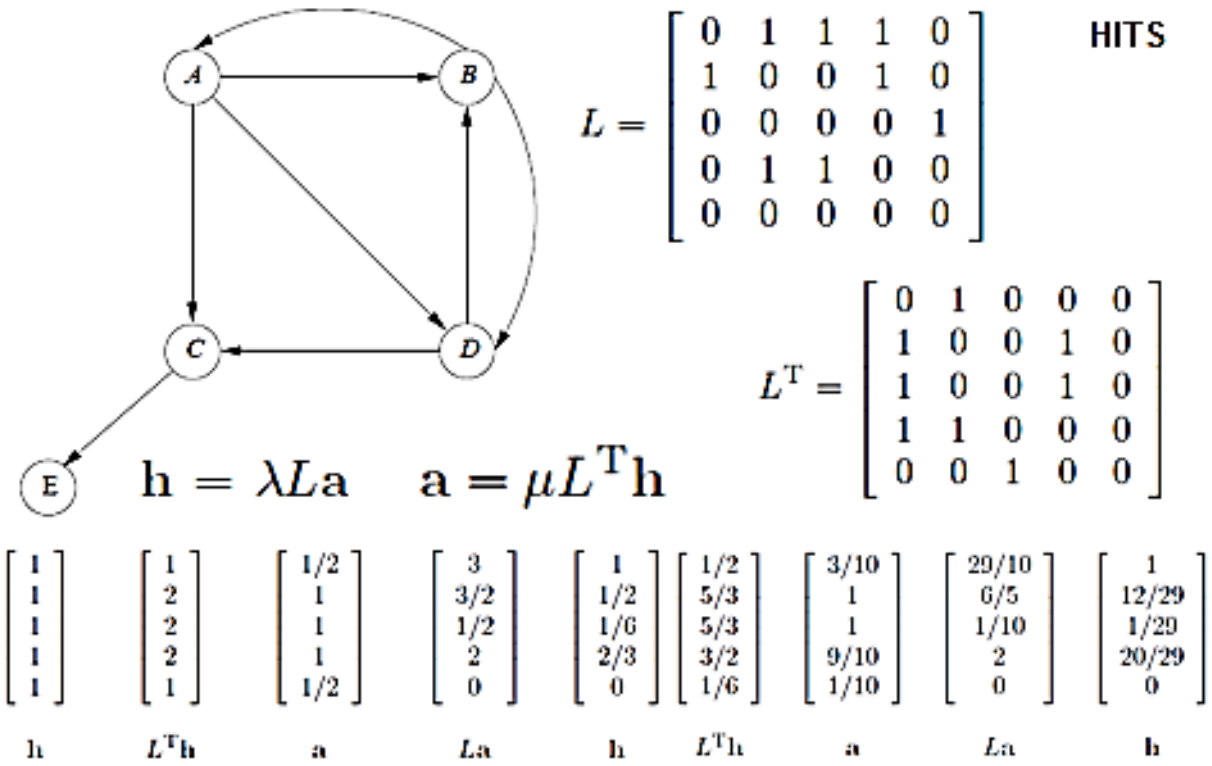


Рис. 5.9. Алгоритм HITS

5.4. Контрольні питання:

1. Що таке терм-спам. Які є стратегії подолання терм-спаму?
2. Як побудувати матрицю гіперпосилань? Що це за матриця?
3. У чому полягає алгоритм PageRank? Які базові припущення цього алгоритму?
4. Для чого при обчисленні PageRank вводять таксацію (які проблеми виникають при застосуванні PageRank до реальної мережі)?
5. Які вузли називають тупиками? Як обчислюють PageRank для мереж з тупиками?
6. Які вузли називають пастками? Як обчислюють PageRank для мереж з пастками?
7. Яку множину називають множиною телепортації?
8. У чому полягає алгоритм TrustRank? Чим він відрізняється від тематично-чутливого PageRank?
9. Що таке посилальний спам (link spam)?
10. Що таке ферма спама? Яка її структура? Для чого її створюють?
11. Як обчислюють SpamMass? Для чого це потрібно? Як інтерпретувати його значення?
12. Які основні припущення алгоритма HITS? Як проінтерпретувати його результати?
13. Що таке хабність та авторитетність? Як їх обчислити?

5.5. Домашнє завдання:

Виконати або завдання 1 або завдання 2:

Завдання1. Розробити програму для реалізації одного із алгоритмів аналізу зв'язків (однієї з модифікацій PageRank або HITS)

a. використовуючи масиви;
b. використовуючи динамічні структури;
c. використовуючи засоби розпаралелювання.

Завдання 2. Розробити граф мережі з n вузлів (n=6, 8 або 10) та дослідити її на спам. При використанні симулятора[33] дослідити не менше двох графів (з пасткою/ без пасток або з мертвими вершинами/ без них) та прокоментувати отримані результати.

5.6. Індивідуальне завдання:

Підготувати реферат на тему: «Сучасні засоби та методи боротьби із посилальним спамом» або «Практичне використання коефіцієнтів PageRank та HITS».

5.7. Вправи[34].

Вправа 1: Обчислити PageRank для кожної сторінки на Рис. 5.10 без урахування таксації.

Вправа 2: Обчислити PageRank для кожної сторінки на Рис. 5.10, враховуючи $\beta = 0.8$.

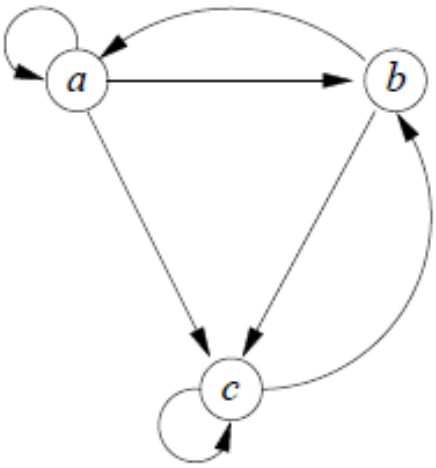


Рис. 5.10

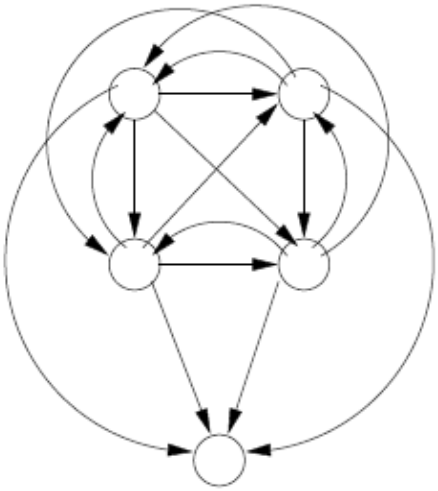


Рис. 5.11

Вправа 3*: Нехай Web складається з кліки (множини вузлів, з'єднаних кожен з кожним максимальною кількістю ребер) з n вузлів та одним вузлом, що є нащадком кожного з n вузлів кліки. На Рис. 5.11 показано граф для n = 4. Визначити PageRank кожної сторінки, як функцію від n та β.

Вправа 4*: Побудуйте для довільного цілого n, такий Web, щоб залежно від β, кожен з n вузлів міг мати найвищий PageRank серед тих n. Допускається наявність інших вузлів у Web окрім цих n.

---

[33] https://computerscience.chemeketa.edu/cs160Reader/_static/pageRankApp/index.html

[34] **Mining of Massive Datasets** / by Jure Leskovec , Anand Rajaraman, and Jeff Ullman. - http://www.mmds.org

Вправа 5: Обчисліть тематично-чутливий PageRank для графа на Рис. 5.12, вважаючи, що множина телепортації складається з:

а) тільки з вузла A;

б) тільки з вузлів A та C.

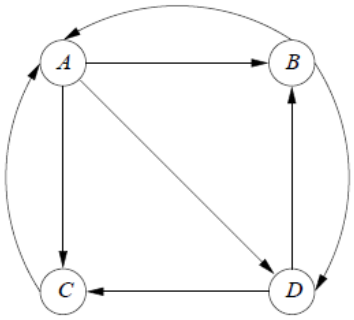


Рис. 5.12

Вправа 5: У розділі 5.4.2[35] проаналізовано ферму спаму, представлену на Рис. 5.13, де кожна опорна сторінка мала зворотний зв'язок з цільовою сторінкою. Повторіть аналіз для ферми спаму, у якій:

а) Кожна опорна сторінка посилається на себе, а не на цільову сторінку.

б) Кожна опорна сторінка не посилає ні на кого.

в) Кожна опорна сторінка посилається як на себе, так і на цільову сторінку.

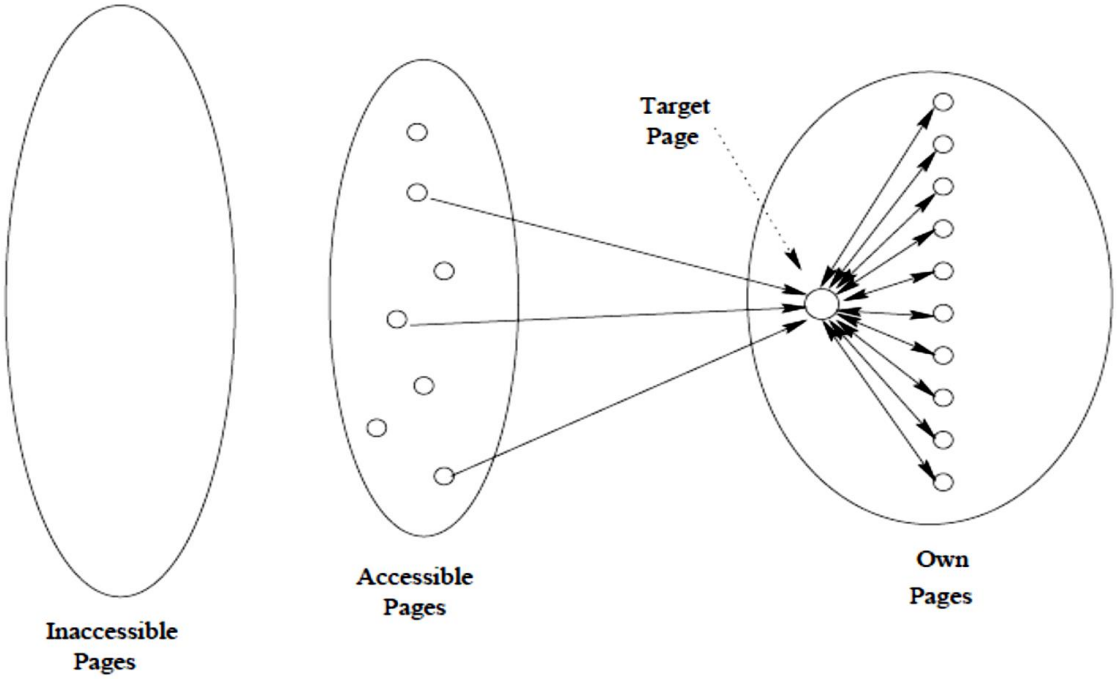


Рис. 5.13

Література до розділу:

- ✓ Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets http://www.mmds.org
- ✓ https://computerscience.chemeketa.edu/cs160Reader/_static/pageRankApp/index.html

[35] **Mining of Massive Datasets** / by Jure Leskovec , Anand Rajaraman, and Jeff Ullman. - http://www.mmds.org/

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Балабанов О.С. Аналітика великих даних: принципи, напрямки і задачі (огляд). //Проблеми програмуваннння. – №2. – 2019. – С.47-68. – URL: https://doi.org/10.15407/pp2019.02.047
2. Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.
3. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. / А. А. Барсегян, М. С. Куприянов, В.В. Степаненко, И. И. Холод. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2007. – 384 с.
4. Бауэр Ф., Гооз Г. Информатика, 1990.
5. Брой М. Информатика, 1996.
6. Дубров А.М., Мхитарян В.С., Трошин Л.И.Многомерные статистические методы. – М.: 2003.
7. Зиновьев А.Ю. Визуализация многомерных данных. – Красноярск: Изд-во КГТУ, 2000. – 180 с.
8. Інтелектуальний аналіз даних: підручник /О.І.Черняк, П.В.Захарченко. – К.:Знання, 2014. – 599 с.
9. Кибернетика. Становление информатики, 1986.
10. Кислова О. Н. Интеллектуальный анализ данных: история становления термина / Ольга Кислова // Український соціологічний журнал. – 2011. – №1-2. – С. 83-94. – URL: http://www.sau.kiev.ua/docs/magazine/USM2011-1-2.pdf
11. Лупан І.В., Авраменко О.В., Акбаш К.С. Комп'ютерні статистичні пакети: навчально-методичний посібник. – 2-е вид. – Кіровоград, «КОД», 2015. – 236 с.
12. Марманис Х., Бабенко Д. Алгоритмы интеллектуального Интернета. Передовые методики сбора, анализа и обработки данных. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 480 с.
13. Орехов В.Д. Прогнозирование развития человечества с учетом фактора знания. Монография. – Жуковский: МИМ ЛИНК, 2015. – 210 с. – URL: http://world-evolution.ru/monografiya/
14. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям (+CD): Учебное пособие. 2-е изд., испр. – СПб.: Питер, 2013. – 704 с.
15. Представление и использование знаний/ Под ред. Х.Уэно, 1989.
16. Степанов Р.Г. Технология Data Mining: Интеллектуальный анализ данных. – Казань, 2008. – 58 с.
17. Факторный, дискриминантный и кластерный анализ/ Дж.-О.Ким, Ч.У.Мьюллер, У.Р.Клекка и др. – М.,1989. – 215 с.
18. Цаленко М.Ш. Семантические и математические модели баз данных, 1985.
19. Чубукова И.А. Data Mining (Курс лекций) – URL: http://kist.ntu.edu.ua/textPhD/Chubukova-Data_Mining.pdf
20. Шамша Б.В., Гуржій А.М., Дудар З.В., Левикін В.М. Математичне забезпечення інформаційно-керуючих систем: Підручник для студентів внз, 2006.

21. Шитиков В. К., Мастицкий С. Э. (2017) Классификация, регрессия, алгоритмы Data Mining с использованием R. – https://github.com/ranalytics/data-mining
22. Han J., Kamber M., Pei J. Data mining: Concepts and Techniques. – Morgan Kaufmann Publishers. – 2012. – 740 p.
23. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets http://www.mmds.org
24. Matthew North. Data Mining for the Masses. – 2012. – 264 p. https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf

ДОДАТКИ

Додаток А. Рекомендовані джерела даних[36]

Data.gov надає загальний доступ до наборів даних з високою вартістю, машиночитаних, створених виконавчою гілкою федерального уряду. Зібрання пропонує безліч безкоштовних наборів даних у пошуковій базі даних.

UNdata – Статистичні дані Організації Об'єднаних Націй з багатьох питань, зокрема сільське господарство, злочинність, освіта, зайнятість, енергетика, довкілля, здоров'я, ВІЛ / СНІД, людський розвиток, промисловість, інформаційні та комунікаційні технології, національні рахунки, населення, біженці, Туризм, торгівля, а також База даних цілей розвитку тисячоліття.

Google Public Data Explorer – Сайт надає доступ до даних міжнародних організацій, національних статистичних органів, неурядових організацій та дослідницьких установ та дозволяє створювати візуалізації загальнодоступних даних за допомогою цього інструменту від Google.

DataHub – 8000 + безкоштовних наборів даних від Фонду відкритих знань. Різні теми. Включає багато великих наборів даних з національних урядів та численні набори даних, пов'язані з економічним розвитком.

Quandl – Пропонує безкоштовну платформу з сотнями безкоштовних наборів даних від "центральних банків, бірж, брокерських компаній, урядів, статистичних установ, аналітичних центрів, науковців, дослідницьких компаній і багато іншого".

Щоб завантажити дані, потрібно створити безкоштовний обліковий запис на сайті.

StatSci.org – "Портал статистичної науки" пропонує довгий список посилань на набори даних для навчання, а також інші ресурси зі статистики.

re3data.org - Registry of Research Data Repositories – це глобальний реєстр репозиторіїв дослідних даних, що охоплює репозиторії дослідних даних з різних навчальних дисциплін.

The Data and Story Library – це онлайн-бібліотека файлів даних та історій, які ілюструють використання методів базової статистики для широкого кола тем, які будуть цікавими для студентів. Набори даних можна переглядати за темою або шукати за ключовим словом.

Колекції даних, організованих за статистичною методикою

The Data and Story Library – Історії даних з наборами, які можна шукати за допомогою певних статистичних методів.

Understandable Statistics Data Sets – видавець цього підручника надає деякі набори даних, організовані за типом / типами даних, наприклад:

[36] Eastern Michigan University https://guides.emich.edu/data/free-data

* дані для множинної лінійної регресії
* одна змінна для великих або вибірок
* парні дані для t-тестів
* дані для односторонньої або двосторонньої ANOVA
* дані часових рядів тощо

University of Florida Statistics Professor's Miscellaneous Datasets – проф. Ларрі Віннер, департамент статистики Університету Флориди, надає посилання на довгий список наборів даних, організованих статистичною методикою.

Порада: Оскільки список досить довгий, можна допомогти використовувати Ctrl + F для пошуку сторінки за ключовим словом.

Безкоштовні джерела наборів даних в Інтернеті

American National Election Studies (ANES)

Щоб задовольнити потреби дослідників-соціологів, викладачів, студентів, політиків та журналістів, ANEC виробляє високоякісні дані з власних опитувань щодо голосування, громадської думки та політичної участі.

Data.World – це соціальна мережа даних. Люди, які реєструються, можуть шукати, копіювати, аналізувати та завантажувати набори даних.

Datasets, Instruments and Tools for Analysis - Childcare & Early Education Research Connections – Пошук наборів даних або інструментів, що використовуються в ранніх дослідженнях. Ви також можете використовувати інструмент на сайті для аналізу даних.

DataWeb - USITC – Інтерактивний тариф та торгівля USITC DataWeb надає статистику міжнародної торгівлі США та дані тарифів США. Вона включає статистику імпорту США, статистику експорту в США, тарифи США, майбутні тарифи США та інформацію про тарифи в США, а також дані про міжнародну торгівлю за 1989 рік.

Education Data Analysis Tool (EDAT) - National Center for Education Statistics – Інструмент аналізу даних освіти (EDAT) дозволяє завантажувати набори даних дослідження NMES. Включає дані декількох тривалих досліджень з питань освіти.

Energy Information Administration – Багато типів деталізованої енергетичної статистики (США та міжнародні) про постачання, ціни, споживання, торгівлю, навколишнє середовище, прогнози та аналізи. Включає статистики для багатьох видів енергії, включаючи альтернативні джерела. Часто історичні статистичні дані включаються і часто статистика може бути завантажена у файли Excel.

EUROSTAT –це статистичний офіс Європейського Союзу, розташований у Люксембурзі. У базі представлена економічна та демографічна статистика для Європи.

Federal Contract Solicitation & Award Notices

Новий набір (2013) файлів .csv, отриманих через Закон про свободу інформації від Адміністрації загальних служб. Містить заявки на відшкодування та присудження федеральних контрактів на період 2000-2013 років.

Federal Reserve Economic Data (FRED) – пропонує дані США та міжнародні часові ряди з 86 джерел.

Fiscally Standardized Cities database - Lincoln Institute of Land Studies – База даних «Стандартизовані дані міст» (FiSC) дає змогу порівняти фінанси місцевого самоврядування для 112 найбільших міст США у більш ніж 120 категоріях доходів, видатків, боргів і активів.

FIVE Project - Dartmouth

До вільних наборів даних відносяться: історичні продажі робочих місць, фотолактографія, пивоварні та суднобудування.

Global Entrepreneurship Monitor (GEM) project – Проект Глобального моніторингу підприємництва (GEM). Щорічна оцінка підприємницької діяльності, прагнень і ставлення окремих осіб у широкому діапазоні країн. Можна завантажити в SPSS.

Google's Dataset Search – Пошук набору даних дозволяє знаходити набори даних, де б вони не знаходилися, незалежно від того, чи є це сайтом видавця, цифровою бібліотекою або особистою веб-сторінкою автора. Ще в бета-тесті.

Health and Retirement Study – це лонгітюдне дослідження обстежує велику вибірку американців старше 50 років кожні 2 роки. "З моменту його запуску в 1992 році, дослідження зібрало інформацію про доходи, роботу, активи, пенсійні плани, медичне страхування, інвалідність, фізичне здоров'я та функціонування, когнітивне функціонування та витрати на охорону здоров'я".

Inforum EconData – тут можна знайти кілька тисяч економічних часових рядів, створених рядом урядових агентств США і поширених у різних форматах і засобах масової інформації. Ці серії включають національні доходи та рахунки продуктів (NIPA), статистику праці, індекси цін , поточні показники бізнесу та промислового виробництва тощо.

Innovative Data Sources for Economic Analysis – метою цього веб-сайту є інформування економічних дослідників та політиків про нові та інноваційні джерела даних та аналітичні інструменти, які мають потенціал для поліпшення розуміння динаміки економіки США, зокрема, оскільки це стосується інновацій та підприємництва. Прокрутіть вниз для посилання на категорії даних. Деякі описані тут джерела не безкоштовні.

International Macroeconomic Data Set - U.S. Dept of Agriculture Economic Research Service – міжнародний макроекономічний набір даних Міністерства сільського господарства США надає дані з 1969 по 2030 рр. для реального (з урахуванням інфляції) валового внутрішнього продукту (ВВП), населення, реальних обмінних курсів та інших змінних для 190 країн і 34 регіонів, які є найбільш важливі для торгівлі сільськогосподарською продукцією в США.

International Monetary Fund Data – Дані часових рядів Міжнародного валютного фонду (МВФ) для багатьох міжнародних економічних показників. Крім того, дані про борги, прямі інвестиції, товари, державні фінанси, експорт, курси валют і т.д.

Kilts Center Marketing Databases – для студентів доступні лише публічні бази даних. До них відносяться дані про продаж продуктового магазину, дані про купівлю домогосподарств, дані панелі сканера тощо.

MeasuringWorth.com – розроблений двома професорами економіки, цей сайт пропонує калькулятори та набори даних, пов'язані з мірками вартості протягом тривалих періодів часу. Заходи включають річні темпи зростання ІСЦ, ВВП і ціни на золото; відносна вартість долара США (або британського фунта) у порівнянні з індексом роздрібних цін, дефлятором ВВП, середнім заробітком, ВВП на душу населення або ВВП; та порівняння купівельної спроможності, рівня інфляції та Dow Jones Industrial Average.

Medical Expenditure Panel Survey (MEPS) – дослідження групи медичних витрат (MEPS) – це комплекс широкомасштабних опитувань сімей та приватних осіб, їх медичних працівників і роботодавців у США. MEPS є найбільш повним джерелом даних про вартість та використання медичних послуг та медичне страхування.

National Climatic Data Center - NOAA – сучасні та історичні дані про погоду та клімат.

National Longitudinal Surveys (U.S. Bureau of Labor Statistics) –Національні довгострокові дослідження (NLS) – це сукупність опитувань, призначених для збору інформації в декілька моментів часу про діяльність на ринку праці та інших значущих подій життя кількох груп чоловіків і жінок, як важливий інструмент для економістів, соціологів та інших дослідників.

NBER Data – Національне бюро економічних досліджень пропонує деякі дані, пов'язані з дослідженнями NBER. Включає макродані, галузеві дані, дані міжнародної торгівлі, індивідуальні дані, демографічну та життєву статистику, патентні дані тощо.

Our World in Data (OWID) – охоплює широкий спектр питань у різних галузях: тенденції в галузі охорони здоров'я, харчування, зростання та розподілу доходів, насильство, права, війни, культура, використання енергії, освіта та зміни навколишнього середовища емпірично аналізуються та візуалізуються в цій веб-публікації. Для кожної теми обговорюється якість даних і, вказуючи відвідувача на джерела, цей веб-сайт також є базою даних баз даних.

Panel Study of Income Dynamics (PSID) – PSID є національно представницьким лонгітюдним дослідженням майже 8 000 американських сімей. Для тих же сімей та окремих осіб з 1968 року, PSID збирає дані про економічну, медичну та соціальну поведінку.

Penn World Table – PWT версія 9.0 – це база даних з інформацією про відносні рівні доходів, обсягів виробництва, вхідних даних і продуктивності, що охоплюють 182 країни в період з 1950 по 2014 рік. Надається через Центр міжнародних порівнянь в Університеті Пенсільванії.

Pew Research Center For The People & The Press Data Archive – сирі дані з опитувань Pew розміщуються тут через шість місяців після публікації результатів опитування. Включає архівовані дані до 1997 року.

Surveys of Consumers (University of Michigan) – цей сайт Інституту соціальних досліджень Мічиганського університету надає звіти, пов'язані з кількома проектами дослідження, включаючи: Індекс поточних умов, Споживчі очікування, Індекс споживчих настроїв.

Tax Statistics (IRS) – включає статистику доходів, статистику податків та бізнесу, індивідуальну та добровільну статистику, статистику за формою ДІВ та багато іншого. Таблиці можна завантажити в Excel.

UK Dataservice – пропонується велика кількість рядів даних – Великобританія, Європа та міжнародний фокус.

United Nations Statistical Databases – Статистичні бази даних Організації Об'єднаних Націй. На цьому сайті можна безкоштовно отримати ряд статистичних баз даних США. Часто дані можна завантажувати. До вільних джерел входять дані з системи демографічного щорічника, Спільної ініціативи щодо даних нафти, бази даних показників тисячоліття, бази даних основних агрегатів національних рахунків, соціальні показники, бази даних населення та багато іншого. Зверніть увагу на додаткові посилання на статистичну інформацію в лівому полі.

Wolfram Data Repository – сховище даних Wolfram – це публічний ресурс, який містить розширену колекцію обчислюваних наборів даних, курируються і структуровані для негайного використання в обчисленні, візуалізації та аналізі.

World Bank Data Catalog – Каталог даних Світового банку. Дані щодо розвитку, дані щодо зміни клімату, дані ВВП, дані Світового банку про фінанси тощо.

World Resources Institute – Інститут світових ресурсів (WRI) є глобальною дослідницькою організацією, яка охоплює понад 50 країн, з офісами в Бразилії, Китаї, Європі, Індії, Індонезії та США. Пропонує широкий спектр статистичної, графічної та аналітичної інформації, пов'язаної з екологічними, соціальними та економічними тенденціями.

Додаток Б. Тестові завдання

Тема1. Інформація та дані

1.1. Основи статистики

1) Якщо використовується шкала найменувань, то класам об’єктів для їхнього позначення можна приписувати:
   а) тільки букви;
   б) тільки числа;
   в) тільки словесні позначки;
   г) будь-які символи: словесні позначки, числа, букви чи інші знаки.

2) Для результатів, одержаних за допомогою шкали порядку, як міру центральної тенденції використовують:
   а) моду і медіану;
   б) середнє арифметичне значення;
   в) дисперсію;
   г) розмах.

3) Шкала інтервалів класифікує об’єкти за принципом: «більше на певну кількість одиниць – менше на певну кількість одиниць», але не дає можливості порівнювати, у скільки разів одна величина більша чи менша за іншу, бо вона:
   а) враховує лише якісні характеристики об’єктів;
   б) не має точки абсолютного нуля;
   в) є не числовою шкалою;
   г) має масштаб.

4) При інтерпретації даних, виміряних за допомогою інтервальної шкали, як характеристики центральної тенденції можна використовувати:
   а) кореляційне відношення, коефіцієнт кореляції Пірсона;
   б) дисперсію, стандартне квадратичне відхилення;
   в) моду, медіану, середнє арифметичне значення;
   г) асиметрію, ексцес.

5) Варіаційний ряд – це ...
   а) послідовність варіант, записаних за зростанням їхніх значень;
   б) послідовність інтервалів, які записані у порядку зростання;
   в) послідовність інтервалів, які містять усю сукупність варіант;
   г) відповідність між варіантами та їхніми частотами або відносними частотами.

6) Інтервальний варіаційний ряд – це ...
   а) послідовність варіант, записаних за зростанням їхніх значень;
   б) послідовність інтервалів, які записані у порядку зростання;
   в) послідовність інтервалів, які містять усю сукупність варіант;
   г) відповідність між інтервалами та їхніми частотами або відносними частотами.

7) Щоб побудувати полігон відносних частот, потрібно:

а) з’єднати відрізками точки, ординатами яких є варіанти, а абсцисами – відповідні їм частоти;
б) побудувати ламану, відрізки якої з’єднують точки, абсцисами яких є значення середин частинних інтервалів, а ординатами – відповідні їм значення частот;
в) на осі абсцис відкласти частинні проміжки, а на осі ординат – частоти;
г) побудувати ламану, відрізки якої з’єднують точки, абсцисами яких є варіанти, а ординатами – відповідні їм значення відносних частот;

8) Щоб побудувати гістограму частот, потрібно:
а) побудувати прямокутники, основами яких є частинні інтервали, а їх висотами – відповідні частоти;
б) побудувати ламану, що з’єднує точки, абсциси яких є середини інтервалів, а їх ординати – відповідні частоти;
в) на осі ординат відкласти частинні проміжки, а на осі абсцис – частоти;
г) немає правильної відповіді.

9) Числовими характеристиками центральної тенденції є:
а) розмах, дисперсія вибірки, середнє квадратичне відхилення вибірки;
б) вибіркове середнє, вибіркове середнє геометричне, мода, медіана;
в) ексцес, асиметрія.

10) Числовими характеристиками варіації є:
а) розмах, дисперсія вибірки, середнє квадратичне відхилення вибірки;
б) вибіркове середнє вибірки, вибіркове середнє геометричнек, мода, медіана;
в) ексцес, асиметрія.

11) Числовими характеристиками форми кривої розподілу є:
а) розмах, дисперсія вибірки, середнє квадратичне відхилення вибірки;
б) вибіркове середнє вибірки, вибіркове середнє геометричнек, мода, медіана;
в) ексцес, асиметрія.

12) Дисперсія – це
а) алгебраїчна сума відхилень значень ознаки від середньої величини;
б) сума модулів усіх відхилень ознаки від середньої величини;
в) середній квадрат відхилень значень ознаки від середньої величини;
г) це відношення середнього квадратичного відхилення до середньої величини.

13) Мода – це
а) варіанта, яка ділить варіаційний ряд на дві рівні за обсягом частини;
б) варіанта, що в статистичному ряді зустрічається найчастіше;
в) відношення суми всіх результатів вимірів до обсягу вибірки;
г) відношення абсолютних характеристик варіації до середньої величини

14) Медіана – це ...
а) варіанта, яка є серединою невпорядкованого ряду;
б) варіанта, що в статистичному ряді зустрічається найчастіше;

в) варіанта, яка ділить варіаційний ряд на частини і найчастіше зустрічається;
г) варіанта, яка ділить варіаційний ряд на дві рівні за обсягом частини.

15) Якщо $A < 0$, то розподіл вважають ...
а) лівостороннім;
б) симетричним;
в) правостороннім;
г) немає правильної відповіді.

16) Якщо $E > 0$, то розподіл вважають ...
а) туповершинним;
б) гостровершинним;
в) плосковершинним;
г) нормальним розподілом.

17) Який інтервал називають довірчим (надійним) для параметра розподілу випадкової величини?
а) інтервал, який із близькою до одиниці ймовірністю „накриває" точне значення параметра;
б) інтервал, який визначає положення статистичної оцінки;
в) інтервал, за допомогою якого знаходять наближене значення параметра розподілу випадкової величини;
г) немає правильної відповіді.

18) Статистична гіпотеза це ...
а) припущення, яке пояснює спостережуване явище;
б) гіпотеза про властивості генеральної сукупності, що перевіряється на основі вибірки;
в) статистичні висновки;
г) спосіб мислення в цілому, який включає висування припущення, його розвиток і доведення.

19) Параметрична статистична гіпотеза це ...
а) статистична гіпотеза про значення параметрів ознак генеральної сукупності;
б) статистична гіпотеза, що висувається на підставі обробки вибірки про закон розподілу ознаки генеральної сукупності;
в) статистична гіпотеза, що підлягає перевірці;
г) статистична гіпотеза, яка повністю або частково логічно заперечує нульову гіпотезу.

20) Непараметрична статистична гіпотеза це ...
а) статистична гіпотеза про значення параметрів ознак генеральної сукупності;
б) статистична гіпотеза, що висувається на підставі відомостей про закон розподілу ознаки генеральної сукупності;
в) статистична гіпотеза, що підлягає перевірці;
г) статистична гіпотеза, яка повністю або частково логічно заперечує нульову гіпотезу;

д) немає правильної відповіді.

21) Яку параметричну гіпотезу називають складеною?
а) гіпотезу, яка складається з одного припущення;
б) гіпотезу, яка є однозначною;
в) гіпотезу, яка складається з скінченої чи нескінченної кількості припущень;
г) немає правильної відповіді.

22) Що таке рівень значущості α?
а) ймовірність відкинути неправильну гіпотезу;
б) ймовірність похибки першого роду;
в) ймовірність похибки другого роду;
г) немає правильної відповіді.

23) У чому полягає помилка другого роду?
а) гіпотеза H_0 відхиляється, але вона істинна;
б) гіпотеза H_0 приймається, але вона хибна;
в) гіпотеза в H_1 відхиляється, але вона неістинна;
г) гіпотеза H_1 приймається, але вона хибна.

24) Статистичним критерієм гіпотези називається …
а) випадкова величина *K*, за допомогою якої проводиться перевірка нульової гіпотези;
б) випадкова величина *K* така, щоб закон розподілу її ймовірностей був невідомий;
в) випадкова величина *K*, яка використовується для перевірки альтернативної гіпотези;
г) немає правильної відповіді.

25) Критичною областю критерію гіпотези називається …
а) сукупність значень критерію, за яких нульову гіпотезу приймають;
б) сукупність значень критерію, за яких нульова гіпотеза відхиляється;
в) сукупність значень критерію, які відокремлюють область, де приймають гіпотезу, від області, де відхиляють;
г) немає правильної відповіді.

26) Якщо критична область на числовій прямій визначається нерівністю $K < k_{kp}$, то вона називається …
правосторонньою;
лівосторонньою;
двосторонньою;
немає правильної відповіді.

27) Яку критичну область будують, якщо $H_0 : M(X) = 14$ за конкуруючої гіпотези $H_\alpha : M(X) > 14$?
правосторонню;
лівосторонню;
двосторонню;
несиметричну.

28) Яку критичну область будують, якщо $H_0 : M(X) = 14$ за конкуруючої гіпотези $H_\alpha : M(X) \neq 14$?
   а) правосторонню;
   б) лівосторонню;
   в) двосторонню;
   г) несиметричну.

29) Що таке потужність критерію?
   а) імовірність відхилення нульової гіпотези, коли правильною є конкуруюча гіпотеза;
   б) імовірність того, що нульова гіпотеза буде прийнята, коли справджується конкуруюча гіпотеза;
   в) імовірність того, що конкуруюча гіпотеза буде відхилена, коли справджується нульова гіпотеза;
   г) немає правильної відповіді.

30) Непараметричні критерії використовують тоді, коли дослідник має справу з:
   дуже малими вибірками чи з якісними даними;
   дуже великими вибірками;
   нормально розподіленими вибірками;
   вибірками, для яких вид розподілу невідомий або неможливо з'ясувати.

1.2. Параметри статистичного розподілу

Заданий статистичний розподіл зображено у вигляді таблиці

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	1	2	6	4	7	7	8	6	3	4	2

1) Знайти відносну частоту варіанти 6:
   а) 0,12;
   б) 0,16;
   в) 0,75;
   г) 0,162.
2) Визначити частоту частинного інтервалу $(2;5]$:
   а) 18;
   б) 17;
   в) 24;
   г) 26.
3) Визначити ординату полігона відносних частот для варіанти 9:
   а) 0,4;
   б) 0,9;
   в) 0,08;
   г) 0,5.
4) Знайти вибіркове середнє:
   а) 5,18;
   б) 1,1;
   в) 4,7;
   г) 6,11.
5) Обчислити вибіркову дисперсію:
   а) 6,12;
   б) 32,4;
   в) 32,94.
   г) 6,11.
6) Знайти вибіркове середнє квадратичне відхилення:
   а) 5,74;
   б) 2,47;
   в) 5,69;
   г) 5,6.
7) Знайти моду:
   а) 10;
   б) 6;
   в) 7;
   г) 4.
8) Знайти медіану:
   а) 4;
   б) 6;
   в) 7;
   г) 5.
9) Знайти асиметрію:
   а) 23,81;
   б) 9,64;
   в) 0,4762;
   г) 0,032.
10) Обчислити ексцес:
   а) 0,72;
   б) 84,9156;
   в) 11,98;
   г) 2,28

1.3. Описова статистика

1.1. Наведіть три додаткові загальноприйняті статистичні показника, які використовують для характеристики розмаху (варіативності) даних.

*Як їх можна ефективно обчислити у великих базах даних?

1.2. Припустимо, що дані для аналізу включають атрибут віку із значеннями: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

а) Яке середнє значення цих даних? Яка медіана?

б) Яка мода даних? Прокоментуйте модальність розподілу даних (тобто, унімодальний, бімодальний, тримодальний тощо).

в) Який розмах даних?

г) Чи можна знайти (приблизно) перший квартиль (Q1) і третій квартиль (Q3) даних?

д) Назвіть п’ять значень, що описують розподіл даних (мінімальне, Q1, медіану, Q3 та максимальне).

е) За отриманими даними побудуйте блочну діаграму даних (box plot).

1.3. Припустимо, що значення для даного набору даних згруповані в інтервали. Інтервали та відповідні частоти такі:

Вік	1-5	6-15	16-20	21-50	51-80	81-110
Частота	200	450	300	1500	700	44

Обчисліть медіану даного набору значень.

1.4. Нехай у лікарні зібрано дані про вік та ступінь ожиріння для 18 випадково відібраних хворих:

Вік	23	23	27	27	39	41	47	49	50
% жиру	9,5	26,5	7,8	17,8	31,4	25,9	27,4	27,2	31,2
Вік	52	54	54	56	57	58	58	60	61
% жиру	34,6	42,5	28,8	33,4	30,2	34,1	32,9	41,2	35,7

а) обчисліть середнє, медіану та стандартне квадратичне відхилення для двох змінних;

б) зобразіть блочні діаграми для двох змінних;

в) побудуйте діаграму розсіювання та квантиль-квантильний графік на основі двох змінних.

1.6. Розмір взуття (довжину ступні) вказано у сантиметрах: 23, 24, 24, 25, 25, 26, 26, 22, 23, 28, 23, 23, 27, 27.

Як ці дані можуть бути використані для аналізу? До якого типу шкали слід віднести такі дані? Відповідь поясніть.

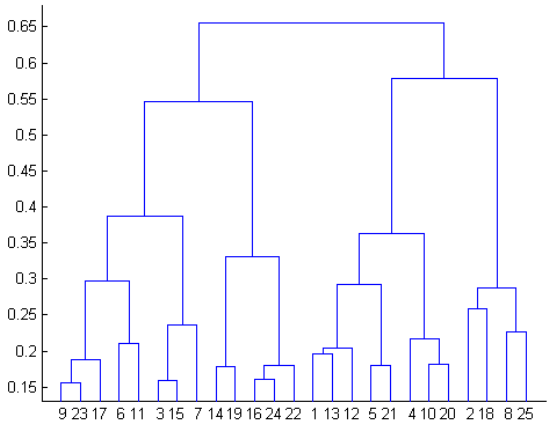
1.7. Якість даних може бути оцінена зокрема за точністю, повнотою та узгодженістю. Для кожної з трьох вищезазначених властивостей обговоріть, як оцінка якості даних може залежати від передбачуваного використання даних, наведіть приклади. Запропонуйте два інших аспекта оцінювання якості даних.

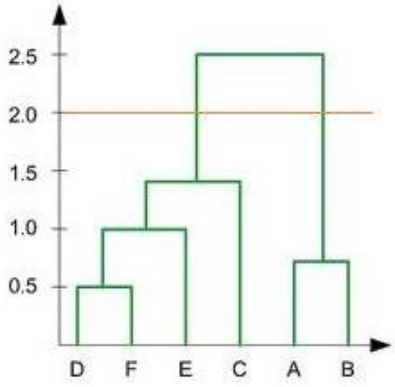
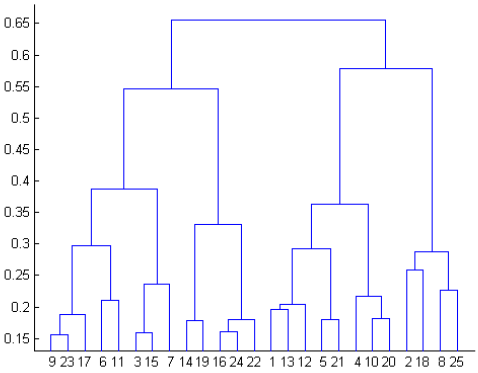
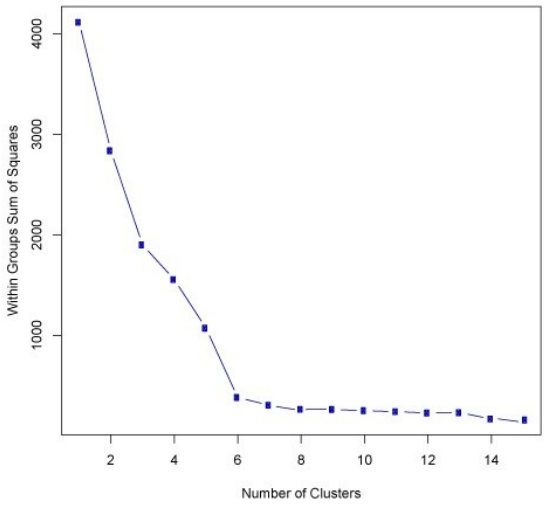
Тема «Кластерний аналіз»

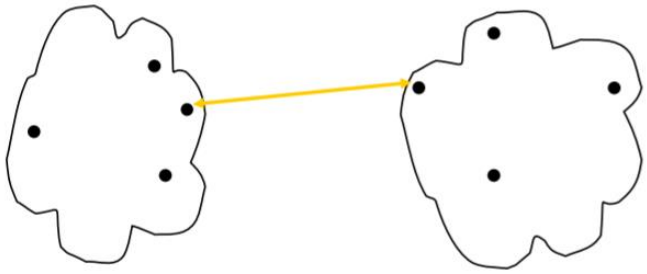
Тема 2. Кластерний аналіз

1	Під поняттям "класифікація" розуміють: а) системний розподіл предметів за будь-якими істотними ознаками для зручності їх дослідження; розподіл вихідних понять і об'єднання у певному порядку; б) системне об'єднання предметів за будь-якими істотними ознаками для зручності їх дослідження, розподілу вихідних понять і розташування у певному порядку; **в) системний розподіл предметів за будь-якими істотними ознаками для зручності їх дослідження; групування вихідних понять і розташування у певному порядку.**
2	Задача класифікації полягає у: **а) передбаченні категоріальної залежної змінної на основі вибірки безперервних і категоріальних змінних;** б) розрахунку категоріальної залежної змінної на основі вибірки неперервних категоріальних змінних; в) виявленні категоріальної залежної змінної на основі вибірки неперервних категоріальних змінних.
3	Вважають, що метою процесу класифікації є: а) побудова графічного зображення, яке використовує прогнозуючі атрибути як вхідні параметри і набуває значення залежного атрибуту; **б) побудова моделі, яка використовує прогнозуючі атрибути як вхідні параметри і набуває значення залежного атрибуту;** **в) побудова алгоритму, який використовує прогнозуючі атрибути як вхідні параметри і набуває значення залежного атрибуту.**
4	У чому полягає процес класифікації: а) у розбитті множини об'єктів на класи за різними критеріями; **б) у розбитті множини об'єктів на класи за певним критерієм;** в) в автоматичному розбитті множини об'єктів на класи.
5	Часто в процесі класифікації застосовують класифікатори, під якими розуміють: а) певну сутність, що визначає, якому із зумовлених класів належить об'єкт за зовнішнім виглядом; б) певну сутність, що визначає, якому із зумовлених класів належить об'єкт за вектором кількості об'єктів; **в) певну сутність, що визначає, якому із зумовлених класів належить об'єкт за вектором ознак.**
6	На які множини розбивають набір даних під час класифікації: а) навчальну і корисну; б) функціональну і тестову; **в) навчальну і тестову;** г) функціональну і виконавчу.
7	При конструюванні моделі класифікації розуміють, що це опис: **а) множини зумовлених класів;** б) цільової функції; в) множини обмежень класів.
8	Вибір методу класифікації можна зробити, користуючись такими характеристиками: а) швидкість, точність , інтерпретованість, надійність; б) швидкість, робастність, алгоритмічність, надійність; **в) швидкість, функціональність, інтерпретованість, надійність;** г) швидкість, робастність, інтерпретованість, стійкість; **д) швидкість, робастність, інтерпретованість, надійність.**

9	Процедуру крос-перевірки застосовують для оцінки: а) надійності класифікації на даних із тестової множини; **б) точності класифікації на даних із тестової множини;** в) швидкості класифікації на даних із тестової множини
10	Відмінність задач кластеризації від задач класифікації полягає в тому, що класи набору даних, що вивчаються: **а) заздалегідь не обумовлені;** б) заздалегідь обумовлені; **в) невідомі**
11	Технологія кластеризації призначена для розбиття сукупності об'єктів: а) на різнорідні групи; **б) на однорідні групи;** в) на однакові групи.
12	Метою кластеризації є пошук: а) однакових множин; б) функціональних залежностей; **в) наявних структур.**
13	Кластер можна охарактеризувати як групу об'єктів: а) що не мають загальних властивостей; б) що мають спеціальні властивості; **в) що мають загальні властивості.**
14	Характеристиками кластера вважають такі ознаки: а) внутрішня різнорідність і зовнішня ізольованість; **б) внутрішня функціональність і зовнішня ізольованість;** **в) внутрішня однорідність і зовнішня ізольованість;** г) внутрішня однорідність і зовнішня неізольованість.
15	Під поняттям «дендрограма» зазвичай розуміють візуальне представлення результатів таксономії у вигляді: **а) деревоподібної ієрархічної структури;** б) графічної ієрархічної структури; в) багаторівневої ієрархічної структури.
16	Важливість вибору відстані між об'єктами полягає в тому, що від неї залежить остаточний варіант: а) об'єднання об'єктів за цього алгоритму розбиття; **б) розбиття об'єктів на класи за цього алгоритму розбиття;** в) формального опису об'єктів у класах за будь-якого алгоритму розбиття.
17	Застосовуючи поняття «об'єднання або зв'язки для двох кластерів», вважають, що це є: а) правила побудови кластеру; б) правила розподілу об'єктів у кластерах; **в) правила визначення відстані між кластерами.**
18	Під оцінкою якості кластеризації розуміють: а) можливість так розподілити кластери, щоб значення вибраного функціоналу якості набуло найкращого значення; **б) можливість так приписати номери кластерів об'єктам, щоб значення вибраного функціоналу якості набуло найкращого значення;** в) можливість так приписати номери кластерів об'єктам, щоб значення вибраного функціоналу якості набуло значення нескінченності.
19	При виборі методу кластеризації керуютьс такими математичними характеристиками: **а) центр, радіус, середньоквадратичне відхилення, розмір кластера;** б) швидкість, радіус, середньоквадратичне відхилення, розмір кластера; в) центр, радіус, стійкість, розмір кластера.

20	Яка мінімальна кількість змінних/ознак необхідна для виконання кластеризації? A. 0 **B. 1** C. 2 D. 3
21	Що з наведеного нижче може виступати можливими умовами припинення в K-Means? 1. **Досягнення фіксованої кількості ітерацій.** **2. Віднесення спостережень до кластерів не змінюється між ітераціями. За винятком випадків з поганим локальним мінімумом.** **3. Центроїди не змінюються між послідовними ітераціями.** **4. Припинення, коли RSS (residual sum of squares - залишкова сума квадратів) опускається нижче порога.**
22	Після виконання аналізу K-Means на наборі даних побудовано наступну дендрограму. Який із наведених висновків можна зробити з дендрограми?  A. У кластерному аналізі було 28 точок даних B. Найкраща кількість кластерів для аналізованого набору даних – 4 C. Використовуваною функцією близькості є кластеризація середніх **D. Наведена вище інтерпретація дендрограми неможлива для кластерного аналізу K-середніх**
23	Яка може бути причина (причини) для створення двох різних дендрограм за допомогою алгоритму агломеративної кластеризації для одного набору даних? A. Використана функція близькості B. використаних точок даних C. використаних змінних D. B і C Лише **E. Все перераховане**
24	Які з тверджень справедливі по відношенню до кластеризації K-Mean? **1. K-means надзвичайно чутливий до ініціалізації центрів кластерів.** **2. Погана ініціалізація може призвести до поганої швидкості зближення.** **3. Погана ініціалізація може призвести до поганої загальної кластеризації.**
25	Що з наведеного нижче можна застосувати, щоб отримати хороші результати для алгоритму K-середніх, що відповідає глобальним мінімумам? **1. Спробуйте запустити алгоритм для іншої ініціалізації центроїда** **2. Налаштуйте кількість ітерацій** **3. З'ясувати оптимальну кількість кластерів**

26	Яка кількість кластерів відповідає значенню *y*=2?  A. 1 **B. 2** C. 3 D. 4
27	Яка кількість кластерів є найбільш підходящою для набору даних, представленого наступною дендрограмою:  A. 2 **B. 4** C. 6 D. 8
28	Який має бути найкращий вибір для кількості кластерів на основі наступних результатів**:**  A. 5 **B. 6** C. 14

29	На малюнку нижче показано два кластери. Вони з'єднані лінією, яка представляє відстань, що використовується для визначення міжкластерної подібності:  Який показник міжкластерної подібності представляє ця лінія? a. **найближчого сусіда** b. найдальшого сусіда c. групових середніх d. відстань між центроїдами
30	**.** Припустимо, потрібно об'єднати 7 спостережень у 3 кластери за допомогою алгоритму кластеризації K-Means. Після кластерів першої ітерації C1, C2, C3 має такі спостереження: C1: {(2,2), (4,4), (6,6)} C2: {(0,4), (4,0)} C3: {(5,5), (9,9)} Якими будуть центроїди кластера, якщо ви хочете перейти до другої ітерації? **A. C1: (4,4), C2: (2,2), C3: (7,7)** B. C1: (6,6), C2: (4,4), C3: (9,9) C. C1: (2,2), C2: (0,0), C3: (5,5) D. Нічого з наведеного

Тема 4. «Пошук асоціативних правил»

4.1. У якому випадку виникають асоціації:

а) якщо виникає потреба аналізу великої кількості даних;
б) у випадку, коли треба пов'язати дані з різних баз даних;
в) у разі необхідності підвищення ефективності процесу продажу товарів;
г) у випадку, коли декілька подій пов'язані одна з одною?

4.2. Асоціативні правила – це:

а) кількісний аналіз подій;
б) опис зв'язків між декількома подіями;
в) якісний аналіз подій.

4.3. Цікавими вважаються асоціативні правила, які:

а) перевищують певний поріг;
б) мають кількісні та якісні характеристики подій;
в) пов'язують найбільшу кількість подій.

4.4. Дайте формальне визначення поняттю "асоціативне правило":

а) це імплікація $X \Longrightarrow Y$, де $X \subset I, Y \subset I$ і $X \cap Y = 0$;
б) це імплікація $X \Longrightarrow Y$, де $X \not\subset I, Y \not\subset I$ і $X \cap Y = 0$;
в) це імплікація $X \Longrightarrow Y$, де $X \subset I, Y \subset I$ і $X \in Y$.

4.5. Поняття "підтримка асоціативного правила" – це:

а) правило $X \Longrightarrow Y$ має підтримку s (support), якщо s% транзакцій з D, містять $X \cap Y$, $\sup p(X \Longrightarrow Y) = \sup p(X \cap Y)$.

б) правило X $\Longrightarrow$ *Y* має підтримку s (support), якщо s% транзакцій з D, містять $X \cup Y$, sup p(X $\Longrightarrow$ *Y*)=sup p($X \cup Y$).

в) правило X $\Longrightarrow$ Y має підтримку s (support), якщо s% транзакцій з D, містять X $\in$ Y, sup p(X $\Longrightarrow$ Y)=sup p(X $\in$ Y).

4.6. Що означає поняття "достовірність"?

а) достовірність правила показує, яка частина X міститься в Y;

б) достовірність правила показує, яка ймовірність того, що з X виходить Y;

в) достовірність правила показує співвідношення X і Y;

4.7. Які з наведених правил відносяться до асоціативних:

а) арифметичні асоціативні правила;

б) булеві асоціативні правила;

в) узагальнені асоціативні правила;

г) статистичні асоціативні правила;

д) кількісні асоціативні правила;

ж) формальні асоціативні правила;

з) непрямі асоціативні правила?

4.8. В чому полягає сутність алгоритму Apriori?

а) робота цього алгоритму складається з таких основних кроків, як пошук асоціацій та формування правил;

б) робота цього алгоритму складається з таких основних кроків, як формування кандидатів та підрахунку асоціацій;

в) робота цього алгоритму складається з таких основних кроків, як формування кандидатів та підрахунок кандидатів.

4.9. В чому полягає відмінність між алгоритмом AprioriTid та алгоритмом Apriori:

а) AprioriTid покращує Apriori за рахунок того, що використотвує базу даних при багатьох проходах;

б) алгоритм AprioriTid покращує Apriori за рахунок того, що використовує базу даних лише при першому проході;

в) AprioriTid покращує Apriori за рахунок того, що використовує базу даних лише при непарних проходах;

г) AprioriTid покращує Apriori за рахунок того, що використовує базу даних лише при парних проходах?

4.10. Які нові механізми були запропоновані в алгоритмі DHP?

а) механізми прямого хешування і обрізання даних;

б) механізми оптимізації запитів і обрізання даних;

в) механізми прямого хешування і розрахунку даних?

4.11. Розглянемо 8 кошиків, до складу яких входять товари-слова:

1. {cat, and, dog, bites}.
2. {Yahoo, news, claims, a, cat, mated, with, a, dog, and, produced, viable, offspring}.
3. {cat, killer, likely, is, a, big, dog}.
4. {Professional, free, advice, on, dog, training, puppy, training}.
5. {cat, and, kitten, training, and, behavior}.
6. {dog, &, cat, provides, dog, training, in, Eugene, Oregon}.
7. {dog, and, cat, is, a, slang, term, used, by, police, officers, for, a, male–female, relationship}.
8. {Shop, for, your, show, dog, grooming, and, pet, supplies}.

а) визначити одноелементні набори з підтримкою не нижче 3;

б) визначити двоелементні набори з підтримкою не нижче 3;
в) визначити триелементні набори з підтримкою не нижче 3;
г) визначити достовірність правила {cat, dog} ⇒ {a};
д) визначити достовірність правила {cat } ⇒ {kitten};
е) визначити достовірність правила {dog} ⇒ {cat};
ж) визначити інтерес правила {dog} ⇒ {cat}.

4.12. Нехай база даних містить дані про чотири покупки:

ID транзакція	Дата	Придбано товари
100	15.10.2019	{K, A, D, B}
200	15.10.2019	{D, A, C, E, B}
300	19.10.2019	{C, A, B, E}
400	22.10.2019	{B, A, D}

Нехай мінімальна підтримка дорівнює 60%, а мінімальна достовірність – 80%.

а) знайдіть усі набори, що часто зустрічаються, за допомогою алгоритма Apriori:

б) визначте усі асоціативні правила.

Підписано до друку *14.12.2022.* Формат 60x84 $^{1}/_{16}$
Папір офсет. Цифровий друк. Ум. др. арк. 7. Обл. вид. арк. 3,74.
Віддруковано: ФОП Піскова М. А.
Свідоцтво про державну реєстрацію
№2444000000027816 від 18.08.2016