Національний технічний університет України «КПІ»

Кафедра автоматизованих систем обробки інформації і управління

Лабораторна робота №4

з дисципліни «[Аналіз даних в інформаційних управляючих системах](https://do.ipo.kpi.ua/course/view.php?id=5470)»

на тему: «МЕТОДИ ВИДОБУВАННЯ АСОЦІАТИВНИХ ПРАВИЛ

З ВЕЛИКИХ МАСИВІВ ДАНИХ»

Виконав:

студент групи ІС-23

Шимків М.В.

Викладач:

Гавриленко О.В

Київ 2024

**Мета роботи:** вивчити основні методи видобування асоціативних правил з великих масивів даних, навчитися використовувати спеціалізовані програмні засоби для видобування знань з масивів даних.

**Завдання до роботи**

1. Ознайомитися з конспектом лекцій та рекомендованою літературою за темою

роботи, а також з додатком до лабораторної роботи, що містить опис програмного

забезпечення для видобування асоціативних правил з великих масивів даних.

2. Сформувати набір даних для обробки та аналізу.

3. Розробити за допомогою середовища Matlab програмне забезпечення для

видобування асоціативних правил з великих масивів даних або вивчити рекомендоване

програмне забезпечення (пакет Armada модулю Matlab) та здійснити обробку набору

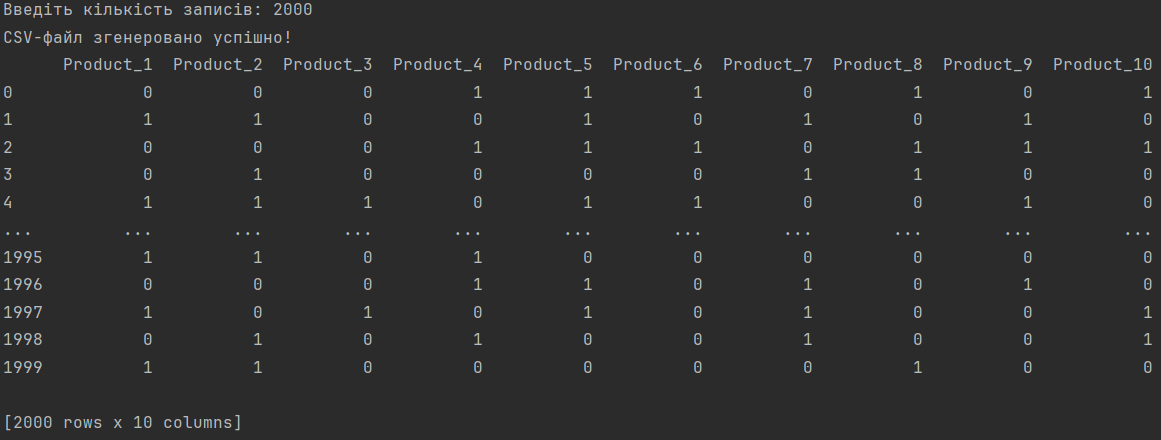
даних з метою виділення асоціативних правил.

4. Оформити звіт з роботи.

5. Відповісти на контрольні питання.

**Набір даних:**

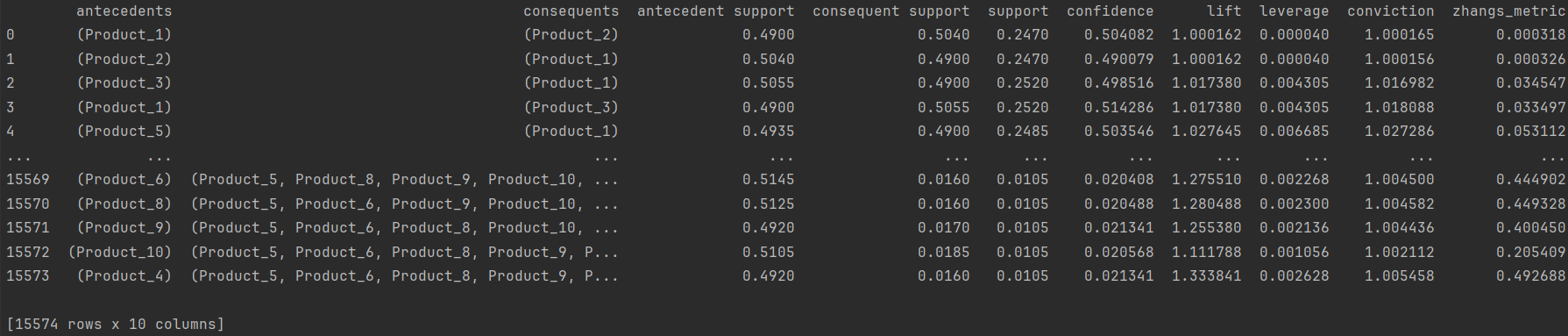
Набір даних генерується в програмі(при першому запуску потрібно вказати кількісить записів, якщо запуск програми відбувається не вперше, то програма просто прочитає вже згенеровані дані)



Запуск програми відбувається так:



Отримані асоціативні правила з великих масивів даних та здійснення обробку набору даних з метою виділення асоціативних правил:



**Код програми:**

import pandas as pd  
from mlxtend.frequent\_patterns import apriori  
from mlxtend.frequent\_patterns import association\_rules  
import csv  
import random  
import os  
  
# Шлях до файлу  
file\_path = 'Sales.csv'  
  
# Перевіряємо, чи файл вже існує  
if not os.path.exists(file\_path):  
 # Кількість записів  
 num\_records = int(input("Введіть кількість записів: "))  
  
 # Кількість продуктів  
 num\_products = 10  
  
 # Назви продуктів  
 products = [f"Product\_{i}" for i in range(1, num\_products + 1)]  
  
 # Відкриваємо CSV-файл для запису  
 with open(file\_path, 'w', newline='') as csvfile:  
 # Створюємо об'єкт для запису в CSV-файл  
 writer = csv.writer(csvfile)  
  
 # Записуємо назви колонок  
 writer.writerow(products)  
  
 # Генеруємо випадкові записи  
 for \_ in range(num\_records):  
 # Генеруємо рандомну послідовність чисел 0 або 1 для кожного продукту  
 record = [random.randint(0, 1) for \_ in range(num\_products)]  
  
 # Записуємо запис у CSV-файл  
 writer.writerow(record)  
  
 print("CSV-файл згенеровано успішно!")  
else:  
 print("Файл уже існує, генерація не потрібна.")  
  
data = pd.read\_csv("Sales.csv")  
print(data)  
  
# Застосовуємо алгоритм Apriori  
frequent\_itemsets = apriori(data, min\_support=0.01, use\_colnames=True)  
  
# Знаходимо асоціативні правила  
rules = association\_rules(frequent\_itemsets, metric="lift", min\_threshold=1)  
  
# Виводимо асоціативні правила  
print(rules)

**Контрольні питання:**

1. **Що таке асоціативне правило?** Асоціативне правило - це правило у вигляді "якщо/тоді", яке встановлює зв'язок між певними подіями чи об'єктами в базі даних. Воно вказує на те, що якщо відбувається одна подія, то ймовірно, що відбудеться інша.
2. **Для чого призначені асоціативні правила?** Асоціативні правила використовуються для виявлення кореляційних зв'язків між різними елементами в базі даних і можуть бути використані для рекомендаційних систем, виявлення спільних покупок тощо.
3. **Дати визначення понять підтримки та достовірності правила.**
   * Підтримка (support) - це ймовірність, що обидва події, які описує асоціативне правило, відбудуться разом.
   * Достовірність (confidence) - це умовна ймовірність того, що друга подія відбудеться, за умови, що відбулася перша.
4. **Яке призначення алгоритмів пошуку асоціативних правил?** Алгоритми пошуку асоціативних правил призначені для виявлення цих правил в базах даних.
5. **На які підзадачі розбивається задача знаходження асоціативних правил?** Задача знаходження асоціативних правил розбивається на підзадачі пошуку часто зустрічаються наборів (фріквентних наборів), виявлення асоціативних правил та їх оцінка.
6. **Які методи використовуються для знаходження асоціативних правил?** До методів пошуку асоціативних правил відносяться Apriori, FP-Growth, Eclat та інші.
7. **Яким чином обираються значення параметрів minsupport та minconfidence?** Значення параметрів minsupport та minconfidence обираються експериментальним шляхом з урахуванням конкретних вимог до якості асоціативних правил та обсягу даних.
8. **Що таке числові асоціативні правила?** Числові асоціативні правила встановлюють зв'язок між числовими даними, наприклад, діапазонами числових значень.
9. **Поясніть поняття “узагальнене асоціативне правило”.** Узагальнене асоціативне правило - це правило, яке включає в себе інші правила і включає в себе більш широкий набір даних.
10. **Що називається ієрархією елементів?** Ієрархія елементів - це структура даних, в якій кожен елемент має батьківські елементи та може мати декілька дочірніх елементів.
11. **Які переваги дає введення додаткової інформації про групування елементів?** Введення додаткової інформації про групування елементів дозволяє враховувати ієрархічну структуру даних та здійснювати пошук асоціативних правил з урахуванням цієї структури.
12. **Поясніть, які проблеми можуть виникнути при безпосередньому застосуванні алгоритмів знаходження асоціативних правил.** При безпосередньому застосуванні алгоритмів знаходження асоціативних правил можуть виникнути проблеми з обробкою великого обсягу даних та збільшенням обчислювальної складності.
13. **В чому полягає сутність виявлення узагальнених асоціативних правил?** Сутність виявлення узагальнених асоціативних правил полягає в пошуку правил, які описують більш широкий набір даних та мають загальні шаблони.
14. **Яким чином визначають “цікаві” правила? В чому полягає актуальність такого процесу?** "Цікаві" правила визначаються на основі їх підтримки та достовірності. Актуальність цього процесу полягає в тому, щоб відфільтрувати найбільш інформативні правила з великої кількості знайдених.
15. **Дати визначення понять батьківського правила (пращура) та найближчого батьківського правила.**
    * Батьківське правило (пращур) - це вихідне асоціативне правило, з якого походять інші правила в результаті узагальнення.
    * Найближче батьківське правило - це найменший асоціативний шаблон, який включає в себе даний набір даних.
16. **Порівняйте поняття цікавого та частково цікавого правила.**
    * Цікаве правило - це правило, яке задовольняє задані критерії цікавості (наприклад, велика підтримка та достовірність).
    * Частково цікаве правило - це правило, яке задовольняє лише деякі з критеріїв цікавості.
17. **Які проблеми усуває алгоритм обчислення узагальнених асоціативних правил?** Алгоритм обчислення узагальнених асоціативних правил допомагає врахувати ієрархічну структуру даних та зменшити кількість правил шляхом їх узагальнення.
18. **З яких етапів складається процес обчислення узагальнених асоціативних правил?** Процес обчислення узагальнених асоціативних правил складається з етапів виявлення правил, узагальнення правил та оцінки їх цікавості.
19. **Проаналізуйте базовий алгоритм пошуку множин, що зустрічаються часто.** Базовий алгоритм пошуку множин, що зустрічаються часто, використовує перебір всіх можливих комбінацій елементів для знаходження фріквентних наборів.
20. **Опишіть алгоритм генерації кандидатів.** Алгоритм генерації кандидатів полягає у створенні нових наборів шляхом об'єднання чи злиття фріквентних наборів.
21. **Яким чином використовується хеш-дерево для підрахунку підтримки кандидатів? Як відбувається процес побудови такого дерева?** Хеш-дерево використовується для швидкого підрахунку підтримки кандидатів шляхом зберігання підсумків підтримки в кожному вузлі. Процес побудови такого дерева полягає у вставці елементів згідно з їх хеш-кодами.
22. **Виконайте порівняльний аналіз базового та покращеного алгоритмів пошуку множин, що зустрічаються часто.** Покращені алгоритми, такі як FP-Growth або Eclat, зазвичай ефективніші за базові алгоритми, такі як Apriori, оскільки вони використовують структури даних, що дозволяють уникнути повного перебору.
23. **За рахунок яких оптимізацій відбувається покращення базового алгоритму пошуку множин, що зустрічаються часто?** Покращення базового алгоритму пошуку множин, що зустрічаються часто, відбувається за рахунок оптимізацій у вигляді використання структур даних (наприклад, FP-дерева, хеш-дерева) та інших методів для скорочення обсягу перебору.
24. **В чому полягає сутність маштабованого алгоритму пошуку асоціативних правил Apriori?** Сутність маштабованого алгоритму Apriori полягає в його здатності обробляти великі обсяги даних шляхом ефективного виявлення фріквентних наборів та генерації асоціативних правил.
25. **Яким чином перетворюються дані для можливості використання алгоритму Apriori?** Дані перетворюються в бінарний формат, де кожен елемент має значення "1" або "0" в залежності від того, чи належить він до певного набору.
26. **Яка властивість використовується в алгоритмі Apriori? Для чого вона використовується?** В алгоритмі Apriori використовується властивість антимонотонності, за допомогою якої можна визначити, що якщо підмножина не часто зустрічається, то й більші множини, що її містять, також не часто зустрічаються.
27. **Наведіть послідовність виконання алгоритму Apriori.** Послідовність виконання алгоритму Apriori:
    * Знаходження фріквентних одиничних елементів.
    * Генерація кандидатів для фріквентних наборів.
    * Підрахунок підтримки для кожного кандидата.
    * Вибір фріквентних наборів.
    * Генерація асоціативних правил.
28. **Опишіть функцію генерації кандидатів в алгоритмі Apriori.** Функція генерації кандидатів в алгоритмі Apriori включає в себе об'єднання фріквентних наборів попереднього рівня для створення нових наборів.
29. **Як відбувається підрахунок підтримки для кожного кандидата в алгоритмі Apriori? Для чого в цій процедурі використовують хеш-дерево?** Підрахунок підтримки для кожного кандидата в алгоритмі Apriori полягає у переборі бази даних та визначенні кількості його збігів. Для цієї процедури використовується хеш-дерево для ефективного пошуку кандидатів.
30. **Як здійснити добування правил з набору, що часто зустрічається?** Добування правил з набору, що часто зустрічається, здійснюється шляхом генерації всіх можливих підмножин та оцінки їх цікавості.
31. **Проаналізуйте внутрішню структуру модулю Armada пакету Matlab: основні змінні, параметри, методи та функції, їх призначення та використання.** Модуль Armada для пакету Matlab - це набір інструментів та функцій, що призначений для аналізу даних, статистичних обробок, машинного навчання та візуалізації результатів в середовищі Matlab. Цей модуль надає різноманітні функції для завантаження, очищення та підготовки даних, виконання статистичних обчислень, розв'язання задач машинного навчання, таких як класифікація та кластеризація, а також побудову графіків та візуалізацію результатів. Модуль Armada може бути корисним для дослідників, аналітиків даних та інженерів, які працюють з обробкою даних та аналізом в середовищі Matlab.

**Висновок:**

Отже, під час роботи ми зрозуміли важливість попередньої підготовки даних, вибору параметрів алгоритму та правильного інтерпретування отриманих результатів. Застосування методів видобування асоціативних правил дозволило нам виявити цікаві залежності та асоціації між різними елементами наших даних.Отримані результати підтвердили ефективність застосування методів видобування асоціативних правил для аналізу великих масивів даних. Це дозволить нам у подальшому здійснювати більш глибокий аналіз даних, виявляти цікаві залежності та використовувати отримані знання для прийняття рішень в різних сферах діяльності.