ЛАБОРАТОРНА РОБОТА № 8
ПОШУК АСОЦІАТИВНИХ ПРАВИЛ

Мета роботи

На практиці вивчити роботу алгоритмів пошуку асоціативних правил і навчитися інтерпретувати результати їх роботи.

Основні теоретичні відомості

У лабораторній роботі розглядаються два методи пошуку асоціативних правил:

- алгоритм Apriori;
- алгоритм FPGrowth.

Параметри налаштування алгоритмів

Розглянемо параметри налаштування використовуваних алгоритмів пошуку асоціативних правил в WEKA (табл. 8.1).

Таблиця 8.1 – Параметри налаштування алгоритмів

Метод	**Параметри**
Apriori	*car* – пошук класових (зі значенням цільового атрибута в правій частині) або звичайних асоціативних правил. *classIndex* – індекс цільового атрибута. Якщо встановлено значення -1, буде обраний останній атрибут. *delta* – ітеративно зменшувати значення порогу підтримки на дане значення. Зменшення буде відбуватися до тих пір, поки не буде досягнуто мінімальне значення підтримки чи не буде згенеровано задану кількість правил. *lowerBoundMinSupport* – нижня межа порогу підтримки. *metricType* – встановлює тип метрики, за якою будуть ранжуватися правила (Confidence, Lift, Leverage, Conviction). *minMetric* – мінімальне граничне значення для обраної метрики. *numRules* – кількість правил, які необхідно знайти. *outputItemSets* – чи виводити часті набори. *removeAllMissingCols* – прибирати чи колонки (атрибути) в яких всі значення відсутні.

Продовження табл.8.1.

Метод	**Параметри**
	significanceLevel – рівень значущості (тільки для достовірності). *upperBoundMinSupport* – верхня межа мінімальної підтримки. Ітеративне зменшення підтримки починається з цього значення.
FPGrowth	*delta* – ітеративно зменшувати значення порогу підтримки на дане значення. Зменшення буде відбуватися до тих пір, поки не буде досягнуто мінімальне значення підтримки чи не буде згенеровано задану кількість правил. *findAllRulesForSupportLevel* – знайти всі правила, які задовольняють нижній межі мінімального значення підтримки та мінімальному значенню метрики. Включення цього режиму скасує виконання ітеративного зменшення підтримки для знаходження заданого кількості правил. *lowerBoundMinSupport* - нижня межа порогу підтримки як частка кількості примірників. *maxNumberOfItems* – максимальна кількість примірників у частому наборі; значення -1 означає без обмежень. *metricType* – встановлює тип метрики, за якою будуть ранжуватися правила. *minMetric* – мінімальне граничне значення для метрики. *numRulesToFind* – кількість правил, які необхідно знайти. *positiveIndex* – встановлює індекс бінарного атрибуту, який буде розглядатися як позитивний. *rulesMustContain* – виводити правила, які містять задані об'єкти (список об'єктів, розділених комою). *transactionsMustContain* – для роботи алгоритму використовувати транзакції (примірники), які містять задані об'єкти. *upperBoundMinSupport* – верхня межа мінімальної підтримки. Ітеративне зменшення підтримки починається з цього значення. *useORForMustContainList* – - використовувати логічний зв'язку «або» замість «і» для списків обов'язкових елементів у транзакціях і правилах.

Інтерпретація результатів пошуку асоціативних правил в

WEKA

Секція «*Associator model*» містить інформацію про побудовану модель.

Для алгоритму Apriori секція містить значення мінімальної підтримки, мінімальне значення обраної метрики (зазвичай достовірність), кількість циклів алгоритму. Далі розташовані знайдені часті набори даних та знайдені правила.

Для алгоритму FPGrowth буде вказана загальна кількість знайдених правил та задана кількість найліпших з них.

Представимо асоціативне правило у вигляді X => Y.

Для кожного з правил будуть відображені деякі числові параметри: число перед стрілкою вказує кількість екземплярів вибірки, для яких ліва умовна частина правила вірна (N_X), а число після стрілки вказує кількість екземплярів вибірки, для яких вірна ліва і права частини правила (N_{XUY}).

Підтримка (support) правила дорівнює:

$$\sup(X \Rightarrow Y) = N_{X \cup Y} / N$$

Достовірність (confidence) правила – це відношення підтримки всього правила до підтримка лівої частини правила:

$$conf(X \Rightarrow Y) = \frac{\sup(X \Rightarrow Y)}{\sup(X)} = \frac{N_{X \cup Y}}{N_X}$$

Параметр Lift визначається діленням достовірності на підтримку правої частини правила:

$$lift(X \Rightarrow Y) = \frac{\sup(X \Rightarrow Y)}{\sup(X) * \sup(Y)} = \frac{conf(X \Rightarrow Y)}{\sup(Y)},$$

а параметр Conviction дорівнює:

$$conv(X \Rightarrow Y) = \frac{1 - \sup(Y)}{1 - conf(X \Rightarrow Y)}.$$

Параметр Leverage – це відношення додаткових екземплярів, що покриваються правилом, до очікуваних в тому випадку, коли ліва і права частини правила були би статистично незалежними.

Наприклад, нехай ми маємо 1000 екземплярів в вибірці (N), при

цьому ліва частина правила покриває 200 з них (NX), права окремо покриває 100 (NY) і все правило покриває 50 (NXUY). Частка екземплярів, що покриваються правилом (підтримка) дорівнює 50/1000=0,05. Кількість екземплярів, які були б покриті правилом у випадку, коли ліва і права частина правила незалежні, дорівнює 200 * 100 / 1000 = 20. Параметр leverage для цього прикладу дорівнює 50 - 20=30, що в пропорції від загальної вибірки дорівнює 30/1000=0,03.

Завдання на лабораторну роботу

Частина А

1. Виконайте наступні завдання для набору даних ‘vote.arff’.
2. Запустіть алгоритм пошуку асоціативних правил Apriori.
3. Яке значення для порогу підтримки було використано в побудованій моделі? Яке значення для порогу достовірності було використано?
4. Запишіть 10 найкращих знайдених правил, вкажіть для них значення підтримки та достовірності.
5. Що позначають числа ліворуч і праворуч від стрілки в знайдених асоціативних правилах?
6. Запустіть алгоритм пошуку асоціативних правил FPGrowth.
7. Порівняйте списки десяти найкращих правил, отриманих двома алгоритмами. Поясніть відмінність в роботі двох алгоритмів.
8. Запустіть алгоритм Apriori задавши значення car = true. Які асоціативні правила були отримані? Що знаходиться в правій частині знайдених асоціативних правил?
9. Вирішіть задачу пошуку асоціативних правил для одного з наборів даних ‘bank-data.csv’, ‘marketbasket.arff’, ‘FoodMart.arff’. Проаналізуйте та поясніть знайдені правила.
10. Яку попередню обробку необхідно провести з набором даних ‘anduin_data.arff’, чтобы иметь возможность применить к нему алгоритм Apriori? щоб мати можливість застосувати до нього алгоритм Apriori? Застосуйте необхідні фільтри і вирішіть задачу пошуку асоціативних правил. Проаналізуйте знайдені асоціативні правила.

Частина Б

11. Спробуйте відшукати у власному індивідуальному завданні нові шаблони (асоціативні правила) за допомогою двох розглянутих алгоритмів. Згенеруйте також правила, в правій частині яких буде знаходитися ваш цільовий атрибут.

Контрольні питання

1. У чому полягає задача пошуку асоціативних правил? Наведіть практичний приклад?
2. Що таке частий набір?
3. Що таке сильне асоціативне правило?
4. З яких двох кроків складається пошук асоціативних правил?
6. У чому полягає принцип Apriori?
7. Як формуються правила зі знайдених частих наборів?
8. Опишіть алгоритм Apriori
9. Що означають параметри support, confidence, lift, conviction, що застосовуються в алгоритмі Apriori?
10. Опишіть алгоритм ECLAT.
11. Опишіть алгоритм FPGrowth.

Зміст звіту

1. Тема і мета роботи.
2. Завдання до роботи.
3. Результати виконання завдань.
4. Відповіді на контрольні запитання.
5. Висновки, що відображують результати виконання роботи та їх критичний аналіз.

ПЕРЕЛІК ПОСИЛАНЬ

1. 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. – СПб.: БХВ-Петербург, 2007. – 384 с.

2. Чубукова И.А. Data Mining: учебное пособие / И.А. Чубукова. – М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.

3. Witten, I.H. Data mining: practical machine learning tools and techniques.—3rd ed. / Ian H. Witten, Frank Eibe, Mark A. Hall. – Morgan Kaufmann Publishers, 2011. – 629 p.

4. Han J. Data Mining: Concepts and Techniques (Second Edition) / J. Han, M. Kamber – Morgan Kaufmann Publishers, 2006. – 800 p.

5. Weka 3: Data Mining Software in Java [Електронний ресурс] – Режим доступу: http://www.cs.waikato.ac.nz/ml/weka.

6. Weka 3 Wiki documentation [Електронний ресурс] – Режим доступу: http://weka.wikispaces.com/