Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(национальный исследовательский университет)
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

ОТЧЁТ ПО ЗАДАНИЮ №2

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Поиск ассоциативных правил

Выполнил
студент группы КЭ-120
Глизница Максим Николаевич
E-mail: letadllo@mail.ru

Челябинск 2021

1. Задание

Задание 2. Поиск ассоциативных правил

Выполните поиск ассоциативных правил для наборов данных из задания 1. Зафиксируйте значение пороговое значение поддержки (например, 10%), варьируйте пороговое значение достоверности (например, от 70% до 95% с шагом 5%). Получите список результирующих правил в удобочитаемом виде (антецедент➡консеквент).

1. Подготовьте список правил, в которых антецедент и консеквент суммарно включают в себя не более семи объектов (разумное количество). Проанализируйте и изложите содержательный смысл полученного результата.

2. Выполните визуализацию полученных результатов в виде следующих диаграмм:

- сравнение быстродействия поиска правил на фиксированном наборе данных при изменяемом пороге достоверности;
- общее количество найденных правил на фиксированном наборе данных при изменяемом пороге достоверности;
- максимальное количество объектов в правиле на фиксированном наборе данных при изменяемом пороге достоверности;
- количество правил, в которых антецедент и консеквент суммарно включают в себя не более семи объектов, на фиксированном наборе данных при изменяемом пороге достоверности. поддержки.

2. Краткие сведения о наборах данных

Использованные наборы данных:

Groceries dataset (https://www.kaggle.com/heeraldedhia/groceries-dataset). Содержит данные о покупках продуктов в формате «покупатель-дата-продукт». Каждая запись содержит один продукт, поэтому для получения списка транзакций требуется предобработка данных. Всего содержит 14963 транзакции, средняя длина транзакции: 2.54.

Dataset for Apriori Algorithm - Frequent Itemsets (https://www.kaggle.com/akalyasubramanian/dataset-for-apriori-algorithm-frequent-itemsets). Также содержит данные о покупках продуктов, но уже сгруппированные в транзакции. Всего содержит 7501 транзакций, средняя длина – 3.91.

MyAnimeList Dataset (https://www.kaggle.com/azathoth42/myanimelist?select=AnimeList.csv) – содержит информацию о различных аниме, взятую с сайта

myanimelist.net, из которой был использован список жанров. Таким образом, каждое аниме было рассмотрено как транзакция, а ассортимент товаров составили жанры, использующиеся на сайте (такие как Action, Adventure и т.д.). Всего содержит 14414 транзакции, средняя длина – 2.91.

3. Краткие сведения о средствах реализации

Для реализации методов была использована библиотека PyFIM, автор Christian Borgelt (https://borgelt.net/pyfim.html). Библиотека содержит используемые в задании алгоритмы Apriori, ECLAT и FP-growth, а также некоторые другие.

Репозиторий по дисциплине: https://github.com/Airpllane/DAAlgorithms.

Каталог для задания: 2. Association.

4. Частые правила

В ходе анализа первого набора данных о продуктах были обнаружены следующие правила с поддержкой выше 0.03% и достоверностью выше 80%:

- (Тропические фрукты + Йогурт + Хлеб) → Колбаса;
- (Выпечка + Газ. вода + Овощи) → Молоко;
- (Плавленый сыр + Выпечка) → Молоко;
- (Хлебный полуфабрикат + Фрукт./Овощ. Сок) → Хлеб.

Из этих правил можно сделать вывод, что клиенты, покупающие множество различных продуктов, имеют высокую вероятность купить такую популярную еду, как колбаса, молоко и хлеб.

В ходе анализа второго набора данных о продуктах были обнаружены следующие интересные правила с поддержкой выше 5% и достоверностью выше 30%:

- Фарш → Спагетти;
- Оливковое масло → Спагетти;
- Пищевое масло → Спагетти;
- Тёртый сыр → Спагетти.

Можно увидеть, что в комплекте со спагетти часто покупаются продукты, которые можно использовать вместе со спагетти в кулинарии. Это же можно увидеть в следующих более длинных наборах:

- (Заморож. овощи + Фарш) → Спагетти;
- (Креветки + Фарш) → Спагетти;
- (Оливковое масло + Заморож. овощи) → Спагетти.

Помимо спагетти, очень популярным товаром в правилах является минеральная вода, однако найти смысл в её сочетаниях с другими продуктами сложно.

В ходе анализа набора данных о жанрах аниме при пороге поддержки 3% и пороге достоверности 70% были обнаружены такие правила, как:

- Mecha → Sci-Fi;
- Super Power → Action;
- Space → Sci-Fi;
- Parody → Comedy.

В этих зависимостях можно легко увидеть смысл; так, например, аниме жанров Mecha и Space вполне ожидаемо сочетается с более широкой категорией Sci-Fi.

В более длинных наборах были обнаружены похожие сочетания, например:

- (Police + Shounen + Adventure) → Mystery;
- (Demons + Magic) → Fantasy;
- (Samurai + Drama) → Historical.

Во всех этих наборах можно проследить связь: к примеру, вполне ожидаемо, что сочетание жанров Demons и Magic влечёт за собой жанр Fantasy.

5. Визуализация

Для визуализации были использованы пороговые значения достоверности от 1% до 96% с шагом 5%. Порог поддержки был установлен на значение 0.2%. Результаты визуализации для первого набора данных приведены на рис. 1.

На рисунке можно увидеть, что алгоритмы практически не отличаются по временным затратам, хотя Apriori показывает несколько худшие результаты, чем два остальных. Общее количество правил при низких значениях поддержки очень высоко, но при увеличении порога резко падает и достигает нуля около 40% (это можно увидеть из графика максимального размера набора).

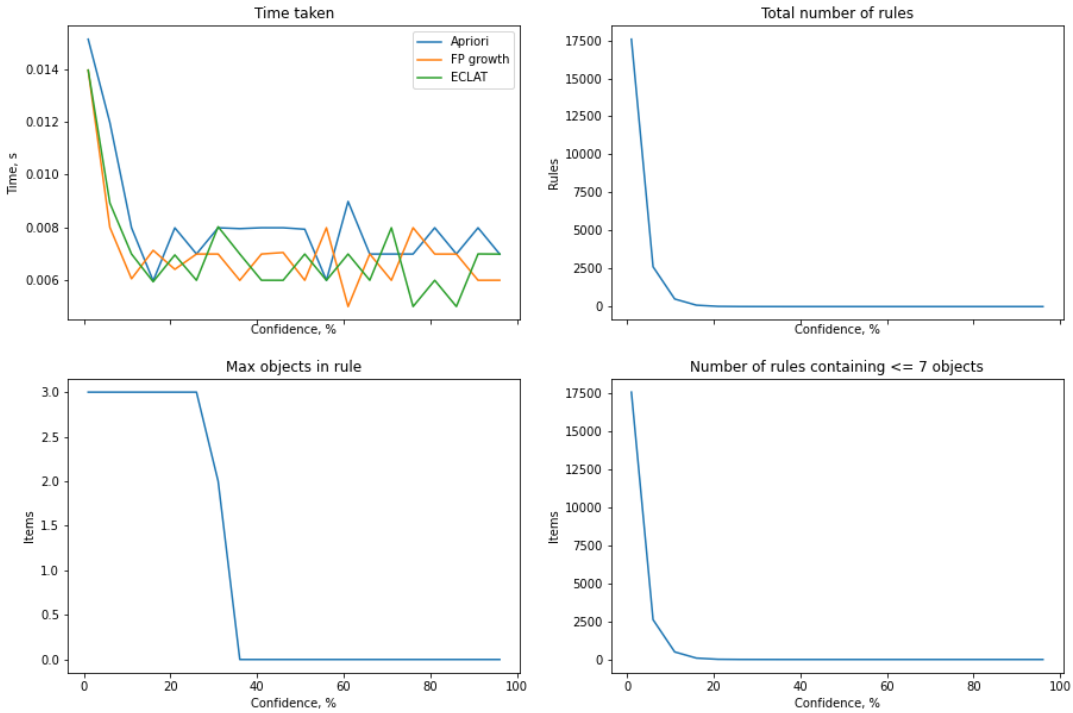

Рис. 1. Результаты визуализации для первого набора данных

Результаты визуализации для второго набора данных приведены на рис. 2.

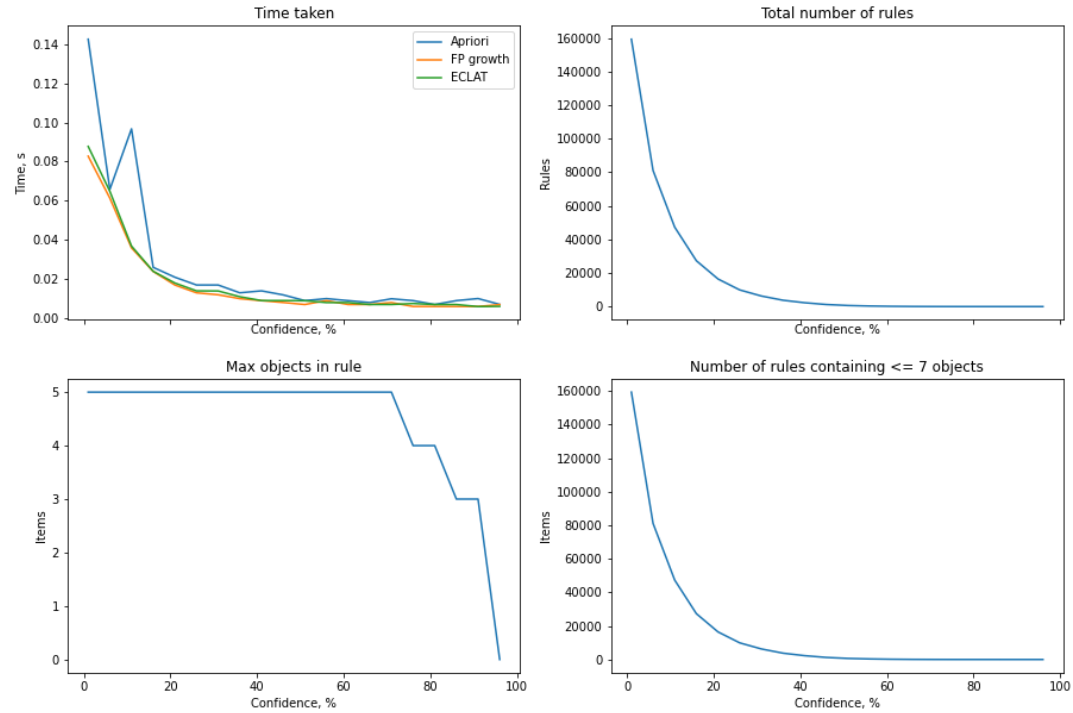

Рис. 2. Результаты визуализации для второго набора данных

В отличие от первого набора данных, во втором наборе значительно выше максимальное количество объектов в правиле, которое достигает нуля только когда порог достоверности переходит 95%. В остальном характер зависимостей не меняется.

Результаты визуализации для третьего набора данных приведены на рис. 3.

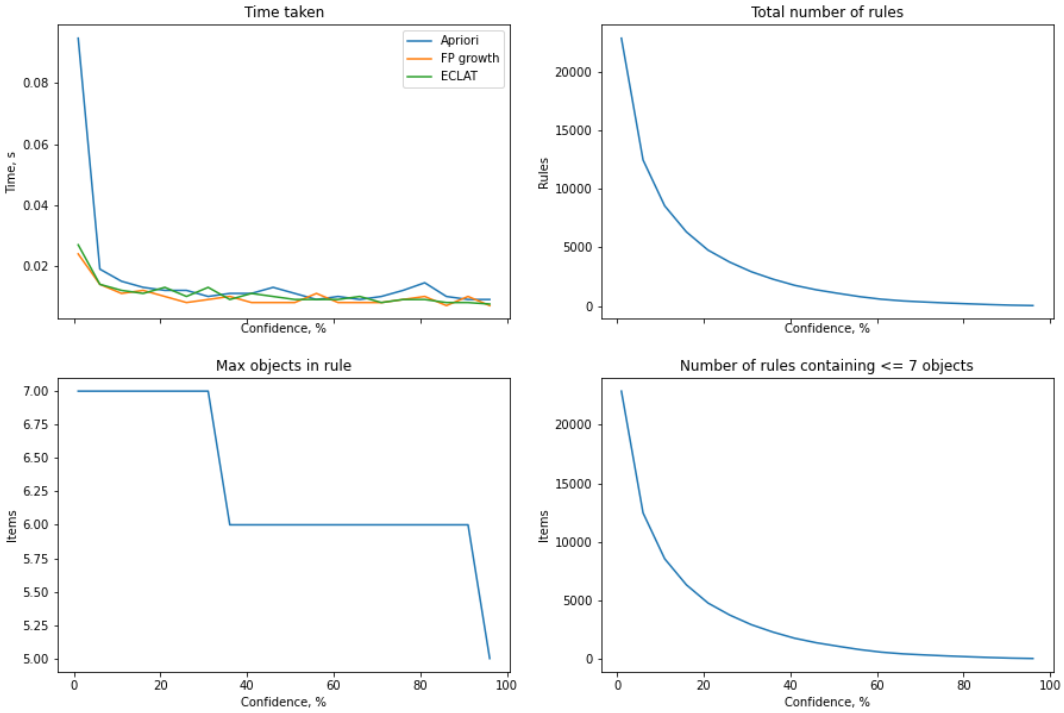

Рис. 3. Результаты визуализации для третьего набора данных

Можно увидеть, что в третьем наборе данных меньше общее количество найденных правил, но выше их средняя длина. Характер зависимостей не меняется.

Поскольку визуализация при пороге поддержки 0.2% не дала интересных результатов на графике, отражающем количество правил, содержащих менее 7 объектов, был построен другой график, содержащий количество правил, содержащих **более** 7 объектов. Этот график был построен при сниженном до 0.02% пороге поддержки. Для построения был использован третий набор как содержащий наибольшее максимальное количество объектов в правиле. График приведён на рис. 4.

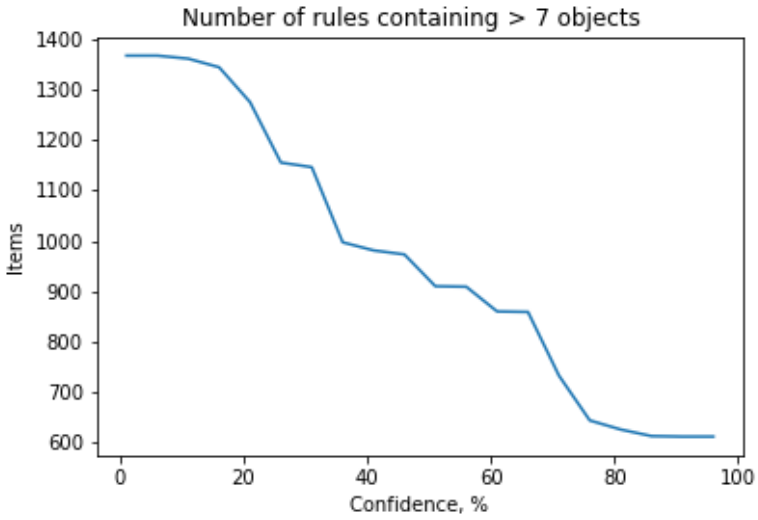

Рис. 4. Количество правил, содержащих более 7 объектов

На этом графике можно увидеть, что при увеличении порога уверенности количество настолько длинных правил падает, что совпадает с тем, что демонстрировал график максимального количества объектов в правиле.