**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра МО ЭВМ**

отчет

**по лабораторной работе №4**

**по дисциплине «Машинное обучение»**

Тема: **Ассоциативный анализ**

|  |  |  |
| --- | --- | --- |
| Студентка гр. 8304 |  | Сергеев А. Д. |
| Преподаватель |  | Жангиров Т. Р. |

Санкт-Петербург

2021

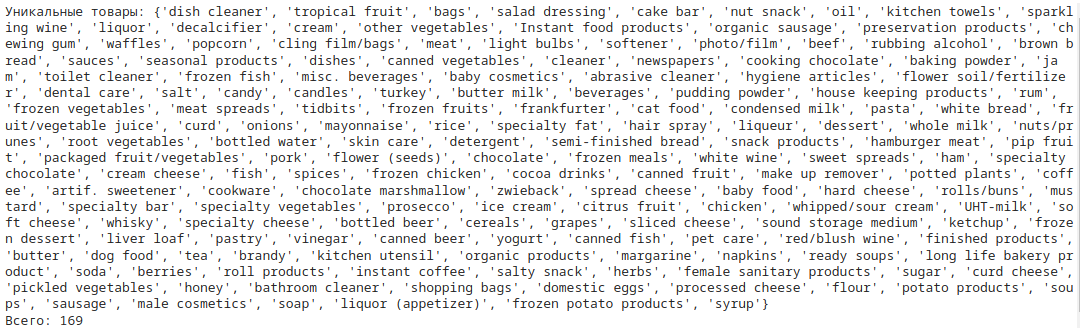
**Цель работы**

Ознакомиться с методами ассоциативного анализа из библиотеки MLxtend.

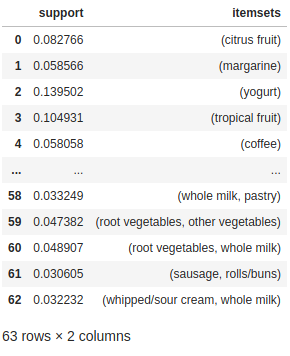
**Ход работы**

* 1. **Загрузка данных**

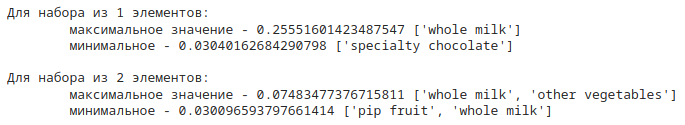
1. Был загружен датасет по ссылке, где данные представлены в виде csv таблицы.
2. Данные были переформированы и из них были удалены все значения NaN.
3. Был получен список всех уникальных товаров.

Рисунок 1 - Уникальные товары

1. **FPGrowth и FPMax**
2. Данные были преобразованы к виду, удобному для анализа.
3. Был проведен ассоциативный анализ, используя алгоритм FPGrowth при уровне поддержки 0.03.

Рисунок 2 – FPGrowth при уровне поддержки 0.03

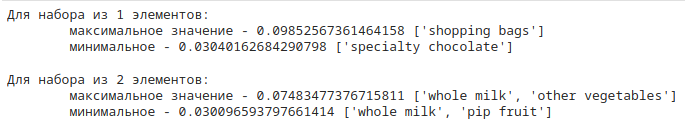
1. Были проанализированы получившиеся варианты. Было определено минимальное и максимальное значения для уровня поддержки для набора из 1,2, и.т.д. объектов.

Рисунок 3 – Минимальный и максимальный уровень поддержки для наборов различной длины в алгоритме FPGowth

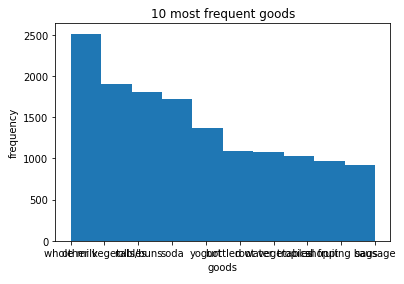
1. Был проведен аналогичный анализ, используя алгоритм FPMax. Результат представлен на рисунках 4 и 5.

Рисунок 4 - FPMax при уровне поддержки 0.03

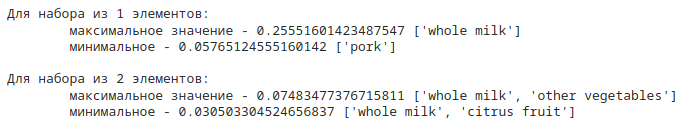
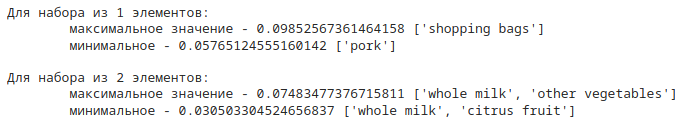
1. Были проанализированы получившиеся варианты. Было определено минимальное и максимальное значения для уровня поддержки для набора из 1,2, и.т.д. объектов.

Рисунок 5 - Минимальный и максимальный уровень поддержки для наборов различной длины в алгоритме FPGowth

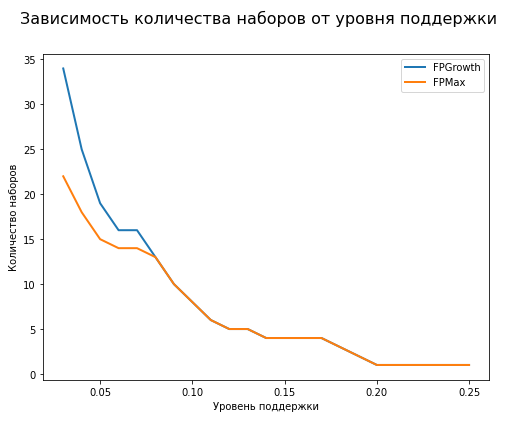
1. Из рисунков 3 и 5 видно, что отличается только максимальный уровень поддержки для длины набора 1. Так произошло из-за того, что в алгоритме FPMax один набор не может быть частью другого. Наиболее часто встречающиеся наборы длины 1 вошли в наборы длины 2.
2. Была построена гистограмма для каждого товара. Столбцы на гистограмме были упорядочены по уменьшению частоты. Сравнивая результаты гистограммы и рисунка 2, можно прийти к выводу, что уровень поддержки прямо пропорционален количеству товара.

Рисунок 6 - Гистограмма 10 самых часто встречающихся товаров

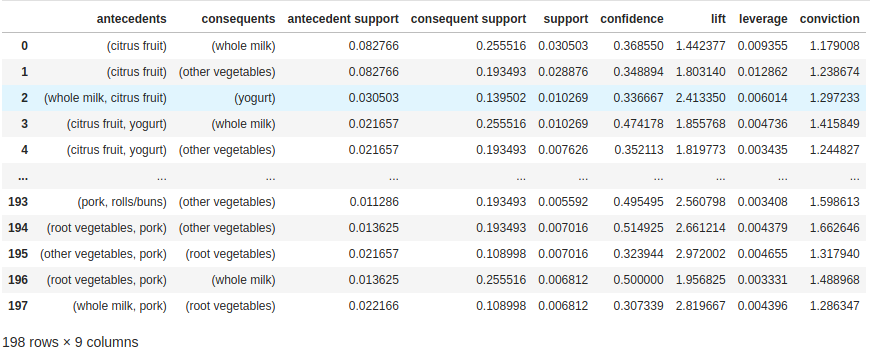
1. Был преобразован набор данных так, чтобы он содержал ограниченный набор товаров.
2. Был проведен анализ FPGrowth и FPMax для нового набора данных. Т.к. были удалены товары с минимальным уровнем поддержки, то минимальные значения для FPGrowth и FPMax увеличились. Максимальные остались без изменений.

Рисунок 7 и 8 - Анализ для ограниченного списка товаров

1. Был построен график изменения количества получаемых правил от уровня поддержки. Результат представлен на рисунке 8. На графике отдельно отображены кривые для набора товаров 1, 2, и т.д. Количество наборов уменьшается с увеличением уровня минимальной поддержки.

Рисунок 9 - Зависимость количества наборов от уровня поддержки

1. **Ассоциативные правила**
2. Были получены частоты наборов используя алгоритм FPGrowth.
3. Был проведен ассоциативный анализ.

Рисунок 10 – Ассоциативный анализ

1. *Поддержка* — это показатель, показывающий то, насколько часто набор объектов встречается в базе данных. Поддержка набора X определяется как отношение числа транзакций, содержащих набор X, к общему числу транзакций.

*Достоверность правила (доверие)* — это показатель, насколько часто правило оказывается верным, это условная вероятность того, что транзакция содержит консеквент 𝑌, при условии, что он содержит антецедент 𝑋: .

*Лифт* определяется как отношение наблюдаемой совместной вероятности X и 𝑌 к ожидаемой совместной вероятности, если бы они были статистически независимыми, то есть . Если правило имеет лифт 1, это означает, что событие в левой части независимо от события в правой части. Если лифт > 1, это позволяет нам знать степень, насколько события связаны друг с другом.

*Усиление* измеряет разницу между наблюдаемой и ожидаемой совместной вероятностью 𝑋Y при условии, что 𝑋 и 𝑌 независимы: .

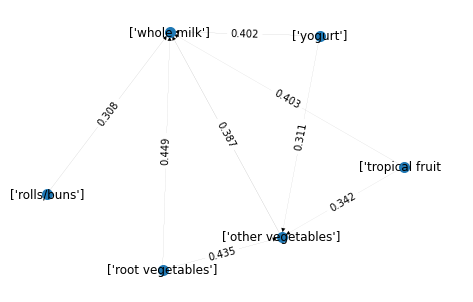
*Уверенность* измеряет ожидаемую ошибку правила, то есть, как часто 𝑋 встречается в транзакции, а 𝑌 - нет. Таким образом, это мера силы правила по отношению к дополнению консеквента. 

1. Было проведено построение ассоциативных правил для различных метрик (значение min\_threshold такое, чтобы выводилось не менее 10 правил).
2. Было рассчитано среднее значение, медиана и СКО для каждой из метрик.

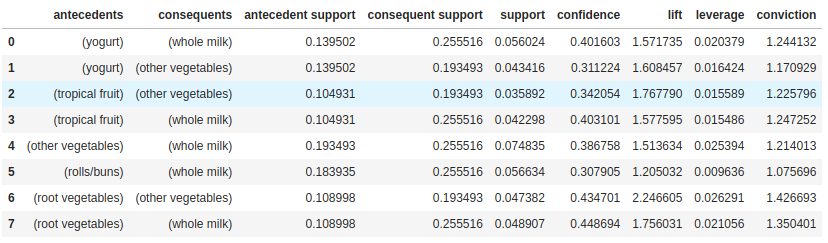
Таблица 1 – Расчет СКО, медианы и среднего значения.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Antecedent support** | **Consequent support** | **Support** | **Confidence (доверие)** | **Lift**  **(лифт)** | **Leverage**  **(усиление)** | **Conviction**  **(уверенность)** |
| Count | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| Среднее значение | 0.0125 | 0.2183 | 0.00748 | 0.60367 | 2.80204 | 0.0048 | 1.9903 |
| СКО | 0.00418 | 0.03203 | 0.00233 | 0.04207 | 0.3158 | 0.00163 | 0.20291 |
| Медиана | 0.01215 | 0.19349 | 0.00732 | 0.59625 | 2.88012 | 0.0045 | 1.94172 |

1. Был построен граф. Каждая вершина графа отображает набор товаров. Ширина ребра должна отображать уровень support, а подпись на ребре отображать confidence.

Рисунок 11 - Граф

1. Из графа можно сделать вывод, что при покупке root\_vegetables, othher\_vegetables tropical\_fruit, yogurt с вероятностью примерно 40%, а при покупке rolls/buns — с вероятностью 30%, будет куплено также whole milk. При покупке yogurt и tropical\_fruit также присутствует вероятность покупки other\_vegetables, а при покупке other\_vegetables - root\_vegetables.
2. Альтернативный способ представления правил - графический.

Рисунок 12 - Правила

**Выводы**

В ходе лабораторной работы изучены методы ассоциативного анализа из библиотеки MLxtend: алгоритмы FPGrowth и FPMax позволяют выделить часто встречающиеся наборы элементов для заданного минимального уровня поддержки. Различие данных алгоритмов заключается в том, что наборы в FPMax не могут быть частью других наборов большей длины. Ассоциативные правила можно генерировать с помощью алгоритма association\_rules, который принимает на вход метрику и ее минимальное значение для расчета.