Министерство образования и науки 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

ОТЧЕТ

По лабораторной работе №2

|  |  |  |  |
| --- | --- | --- | --- |
| Поиск ассоциативных правил | | | |
|  | | | |
|  |  | Руководитель: | |
|  |  | | \_\_\_\_\_\_\_\_\_\_\_\_\_М.Л. Цымблер  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2020 г.  Автор работы  студент группы КЭ-220  \_\_\_\_\_\_\_\_\_\_\_\_А.В.Витомсков  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2020 г. |

Челябинск 2020

1. *Выполните поиск ассоциативных правил для наборов данных из задания по поиску частых наборов. 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).*

*Отфильтруйте правила, в которых антецедент и консеквент суммарно включают в себя не более семи объектов.*

*Проанализируйте и изложите содержательный смысл полученного результата.*

Для исследования мной был использован набор данных о транзакциях UCI. Репозиторий машинного обучения UCI создал этот набор данных, содержащий фактические транзакции за 2010 и 2011 годы. Набор данных поддерживается на их сайте, где его можно найти под названием «Интернет-магазин». Ссылка на набор: <https://www.kaggle.com/carrie1/ecommerce-data> Набор данных содержит 25900 записей. Так он выглядит:

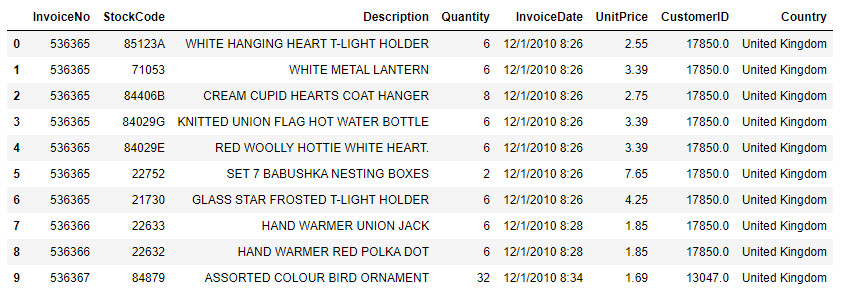


Рисунок 1 – Набор данных UCI

Для поиска частых наборов мной был использован алгоритм apriori из библиотеки mlxtend. Порог минимальной поддержки был взят 0,75 %, так как транзакции данного набора данных содержат много разнородных объектов, и для получения необходимого количества объектов нужен небольшой порог поддержки.

Для поиска ассоциативных правил использована реализация association\_rules из библиотеки mlxtend. Значения достоверности выбраны из следующего списка: [0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70].

Для вычисления среднего времени выполнения, алгоритм поиска ассоциативных правил запускается 1000 раз с одним и тем же значением min\_confidence. Время работы алгоритмов и прочая статистика сохраняется в списке tables.

|  |
| --- |
| Набор данных UCI: сводная таблица времени выполнения  name confidence time rows\_count max\_len count\_different\_len  0 ass\_rules\_execute 0.00 5.247144 128 3 {2: 122, 3: 6}  1 ass\_rules\_execute 0.05 5.402812 128 3 {2: 122, 3: 6}  2 ass\_rules\_execute 0.10 5.880640 128 3 {2: 122, 3: 6}  3 ass\_rules\_execute 0.15 6.256241 127 3 {2: 121, 3: 6}  4 ass\_rules\_execute 0.20 5.359500 118 3 {2: 112, 3: 6}  5 ass\_rules\_execute 0.25 5.236558 108 3 {2: 102, 3: 6}  6 ass\_rules\_execute 0.30 5.509993 97 3 {2: 91, 3: 6}  7 ass\_rules\_execute 0.35 5.102470 76 3 {2: 70, 3: 6}  8 ass\_rules\_execute 0.40 5.046048 54 3 {2: 48, 3: 6}  9 ass\_rules\_execute 0.45 4.916446 40 3 {2: 35, 3: 5}  10 ass\_rules\_execute 0.50 5.570649 32 3 {2: 28, 3: 4}  11 ass\_rules\_execute 0.55 5.260444 23 3 {2: 19, 3: 4}  12 ass\_rules\_execute 0.60 4.810339 20 3 {2: 16, 3: 4}  13 ass\_rules\_execute 0.65 4.600535 9 3 {2: 6, 3: 3}  14 ass\_rules\_execute 0.70 4.521380 7 3 {2: 5, 3: 2} |

Таблица 1 – Результаты тестирования для UCI\_dataset

Здесь:

name – наименование алгоритма,

confidence – минимальное значение достоверности,

time – время выполнения 10 итераций,

rows\_count – количество ассоциативных правил,

max\_len – наибольшее число элементов (сумма антецедентов и консеквентов*)* в правилах,

count\_different\_len – расписано количество ассоциативных правил при разном числе объектов в правиле. В сумме даёт rows\_count.

Видно, что все правила включают в себя не более 7 объектов (максимум – 3). Правил с меньшим количеством элементов существенно больше, чем с большим, и увеличение порога достоверности приводит к резкому уменьшению числа правил с малым количеством элементов. Таким образом, при анализе частых наборов следует больше внимания уделять наборам с большим числом элементов, так как вероятность того, что они являются ассоциативным правилом, существенно больше.

2. *Выполните визуализацию полученных результатов в виде следующих диаграмм:*

*- сравнение быстродействия поиска правил на фиксированном наборе данных при изменяемом пороге достоверности.*

Было выполнено 1000 запусков функции ass\_rules\_execute при каждом значении confidence, результат представлен на рисунке 2.

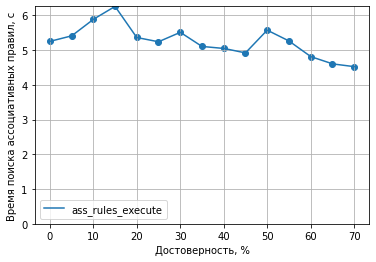
**

Рисунок 2 - Сравнение быстродействия поиска правил на фиксированном наборе данных при изменяемом пороге достоверности

Здесь имеются отличия от поиска частых наборов, и время работы алгоритма при уменьшении достоверности увеличивается лишь незначительно. Это связано с тем, что в исследуемой библиотеке операции вычисления confidence проводятся еще на этапе поиска частых наборов, а на этапе вычисления ассоциативных правил производится лишь сортировка и фильтрация частых наборов по готовым вычисленным значениям мер.

*- общее количество найденных правил на фиксированном наборе данных при изменяемом пороге достоверности;*

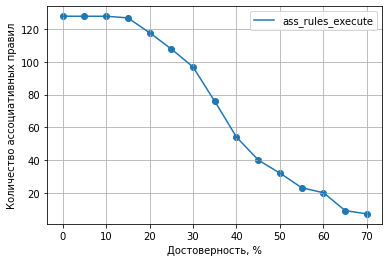
**

Рисунок 3 - 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;*

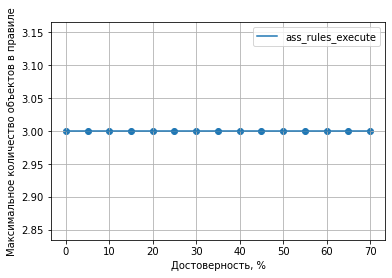
**

Рисунок 4 - 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.*

В моем случае в каждое ассоциативное правило входит не более 3-х объектов, поэтому данный график совпадает с графиком, изображенным на рисунке 3. График, показывающий количество ассоциативных правил различной длины в зависимости от порога минимальной достоверности представлен на рисунке 5.

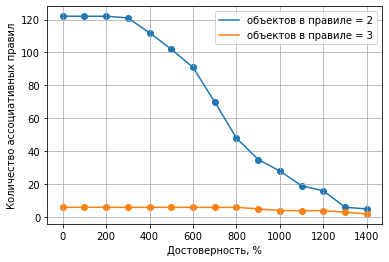
**

Рисунок 5 - Количество ассоциативных правил различной длины ass\_rules\_execute на датасете UCI при изменяемом пороге достоверности