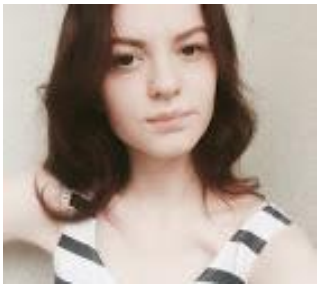
УДК 336.74:004.738.5

АФФИНИТИВНЫЙ АНАЛИЗ ДАННЫХ. ПОИСК АССОЦИАТИВНЫХ ПРАВИЛ



А.Ю. Залога
Магистрант кафедры математического и информационного обеспечения экономических систем УО ГрГУ.им. Я.Купалы

Н.В. Марковская
Доцент кафедры математического и информационного обеспечения экономических систем УО ГрГУ.им. Я.Купалы

УО Гродненский государственный университет имени Янки Купалы
E-mail: Zaloga147@gmail.com , n.markovskaya@grsu.by

А.Ю. Залога
Окончила Гродненский государственный университет имени Янки Купалы. Магистрант кафедры математического и информационного обеспечения экономических систем УО ГрГУ им. Я. Купалы.

Н.В. Марковская
Доцент кафедры математического и информационного обеспечения экономических систем УО ГрГУ им.Я.Купалы, кандидат физико-математических наук, доцент.

Аннотация. В данной работе рассмотрены ассоциативные правила. **Associations rules learning** — ARL представляет из себя, простой, но довольно часто применимый в реальной жизни метод поиска взаимосвязей (ассоциаций) в датасетах, или, если точнее, айтемсетах (itemsests). История развития: впервые подробно об этом заговорил Piatesky-Shapiro G в работе “Discovery, Analysis, and Presentation of Strong Rules.” (1991) Более подробно тему развивали Agrawal R, Imielinski T, Swami A в работах “Mining Association Rules between Sets of Items in Large Databases” (1993) и “Fast Algorithms for Mining Association Rules.” (1994). Одним из ограничений стандартного подхода к обнаружению ассоциаций является то, что при поиске в большом числе возможных ассоциаций набора объектов, которые могут быть ассоциированными, есть большой риск нахождения большого числа случайных ассоциаций.

Ключевые слова: ассоциативных правила, алгоритм apriori, полное доверие, оллективная мощь, убеждённость, рычаг, лифт.

Обучение на ассоциативных правилах (далее Associations rules learning – ARL) представляет из себя, с одной стороны, простой, с другой – довольно часто применимый в реальной жизни метод поиска взаимосвязей (ассоциаций) в датасетах, или, если точнее, айтемсетах (itemsests). Впервые подробно об этом заговорил Piatesky-Shapiro G в работе “Discovery, Analysis, and Presentation of Strong Rules.” (1991) Более подробно тему развивали Agrawal R, Imielinski T, Swami A в работах “Mining Association Rules between Sets of Items in Large Databases” (1993) и «Fast Algorithms for Mining Association Rules.» (1994).

В общем виде ARL можно описать как «Кто купил x, также купил y». В основе лежит анализ транзакций, внутри каждой из которых лежит свой уникальный itemset из набора items. При помощи ARL алогритмов находятся те самые «правила» совпадения items внутри одной транзакции, которые потом сортируются по их силе.

За этой простотой, однако, могут скрываться поразительные вещи, о которых common

sense даже не подозревал [1].

Классический случай такого когнитивного диссонанса описан в статье D.J. Power «Ask Dan!», опубликованной в DSSResources.com .

История

Концепция ассоциативного правила стала популярной благодаря статье 1993 года Агравала, Имелинского, Свами, на которую, согласно Google Scholar, к августу 2015 насчитывалось более 18.000 ссылок, и она является одной из наиболее цитируемых статей в области Data Mining (поиска закономерностей в базах данных). Однако то, что ныне называется «ассоциативными правилами» было введено ещё в статье 1966 года о системе GUHA, общем методе анализа данных, разработанном Петром Гайеком с сотрудниками.

В начале (примерно) 1989 года для поиска минимальной поддержки и доверия для поиска всех ассоциативных правил использовалась система «Характеристическое моделирование» (англ. Feature Based Modeling), которая находит все правила со значениями supp(x) и conf(X=>Y), которые больше заданных пользователем границ.

Альтернативные меры интересности

Кроме доверия, были предложены и другие меры интересности для правил.

Некоторые популярные меры:

1. Полное доверие (англ. All-confidence)
2. Коллективная мощь (англ. Collective strength)
3. Убеждённость (англ. Conviction)
4. Рычаг (англ. Leverage)
5. Лифт (первоначально назывался интересом)

Несколько других мер представили и сравнили Тан, Кумар и Сривастана, а также Хаслер. Поиск техник, которые могут моделировать, что пользователю известно (и использовать это в качестве меры интересности) в настоящее время является активным трендом исследований под названием «Субъективная интересность».

Статистически обоснованные ассоциации.

Одним из ограничений стандартного подхода к обнаружению ассоциаций является то, что при поиске в большом числе возможных ассоциаций набора объектов, которые могут быть ассоциированными, есть большой риск нахождения большого числа случайных ассоциаций. Это наборы объектов, которые оказываются вместе с неожиданной частотой в данных, но чисто случайно. Например, предположим, что мы рассматриваем набор из 10.000 объектов и ищем правило, содержащее два объекта в левой части и один объект в правой части. Имеется примерно 1.000.000.000.000 таких правил. Если мы применим статистический тест независимости с уровнем 0,05 это означает, что имеется только 5 % шанса принять правило при отсутствии ассоциации. Если мы предполагаем, что нет никаких ассоциаций, мы должны, тем не менее, ожидать нахождения 50.000.000.000 правил. Статистически обоснованное обнаружение ассоциаций контролирует этот риск, в большинстве случаев сокращая риск нахождения любой случайной ассоциации для заданного пользователем уровня значимости [2].

Алгоритмы

Было предложено много алгоритмов для генерации ассоциативных правил.

Несколько алгоритмов хорошо известны, это Apriori, Eclat и FP-Growth, но они делают только половину работы, поскольку они предназначены для отыскания часто встречающихся наборов объектов. Нужно сделать ещё один шаг после того, как часто встречающиеся наборы найдены в базе данных.

Алгоритм Apriori

Алгоритм Apriori – алгоритм поиска ассоциативных правил. Алгоритм Apriori ищет ассоциативные правила и применяется по отношению к базам данных, содержащим огромное количество транзакций.

Пример: Скажем, у нас есть база данных транзакций супермаркета. Вы можете представить себе базу данных как огромную таблицу, в которой каждая строка – это номер транзакции, а каждый столбик представляет собой отдельные покупки.

Transaction ID	Chips	Dip	Soda	Apples	Milk
1	X	X	X		
2	X	X			X
3	X		X		

Рисунок 1. База данных транзакций супермаркета

Применяя алгоритм Apriori, мы можем определить товары, купленные вместе – то есть установить ассоциативные правила [3].

Что это дает:

Вы можете определить товары, которые часто покупают вместе. Основная задача маркетинга – заставить клиентов покупать больше. Связанные товары называются наборами.

Например:

Вы можете заметить, что чипсы, чипсы с соусом и газировка часто стоят на прилавках рядом. Это называется двухэлементным набором. Когда база данных достаточно большая, будет гораздо сложнее «увидеть» взаимосвязи, в особенности, когда вы имеете дело с трёхэлементными или более крупными наборами. Как раз для этого и создан алгоритм Apriori.

Как же работает алгоритм Apriori? Перед тем, как перейти к сути алгоритма, вам нужно определить 3 параметра:

Во-первых, нужно установить размер набора. Вы хотите определить двухэлементный, трёхэлементный набор или какой-нибудь еще?

Во-вторых, определить поддержку – это число транзакций, входящих в набор, разделенное на общее количество транзакций. Набор, который равен поддержке, является самым часто встречаемым набором.

В-третьих, определить достоверность, то есть условную вероятность определенного товара оказаться в корзине с другими товарами. Пример: чипсы в вашем наборе имеют 67%-ную вероятность оказаться в одной корзине с газировкой.

Простой алгоритм Apriori состоит из трех шагов:

Объединение. Просмотр базы данных и определение частоты вхождения отдельных товаров.

Отсечение. Те наборы, которые удовлетворяют поддержке и достоверности, переходят на следующую итерацию с двухкомпонентными наборами,

Повторение. Предыдущие два шага повторяются для каждой величины набора, пока не будет повторно получен ранее определенный размер.

$$
\begin{aligned}
&\textit{Apriori}\,(T, \varepsilon)\\
&\quad L_1 \leftarrow \{\ \textit{large 1-itemsets that appear in more than } \varepsilon \textit{ transactions}\ \}\\
&\quad k \leftarrow 2\\
&\quad\quad \textit{while } L_{k-1} \neq \varnothing\\
&\quad\quad\quad C_k \leftarrow \textit{Generate}(L_{k-1})\\
&\quad\quad\quad \textit{for transactions } t \in T\\
&\quad\quad\quad\quad C_t \leftarrow \text{Subset}(C_k, t)\\
&\quad\quad\quad\quad \textit{for candidates } c \in C_t\\
&\quad\quad\quad\quad\quad \text{count}[c] \leftarrow \text{count}[c] + 1\\
&\quad\quad\quad L_k \leftarrow \{c \in C_k | \ \text{count}[c] \geq \varepsilon\}\\
&\quad\quad\quad k \leftarrow k + 1\\
&\textit{return} \bigcup_k L_k
\end{aligned}
$$

Рисунок 2. Повторения алгоритма apriori

Пакет «arules» системы R представляет основу для создания и преобразования входных данных: обеспечивает фундамент для представления, преобразования и анализа транзакционных данных и моделей - частых наборов и ассоциативных правил, - так же обеспечивает интерфейс для реализации в C основанных на идее ассоциативных правил алгоритмов Apriori и Eclat. Эти алгоритмы могут быть использованы для формирования частых наборов, максимальных частых наборов (maximal frequent itemsets), объемлющих частых наборов (closed frequent itemsets) и ассоциативных правил. Ниже представлена установка пакета «arules»

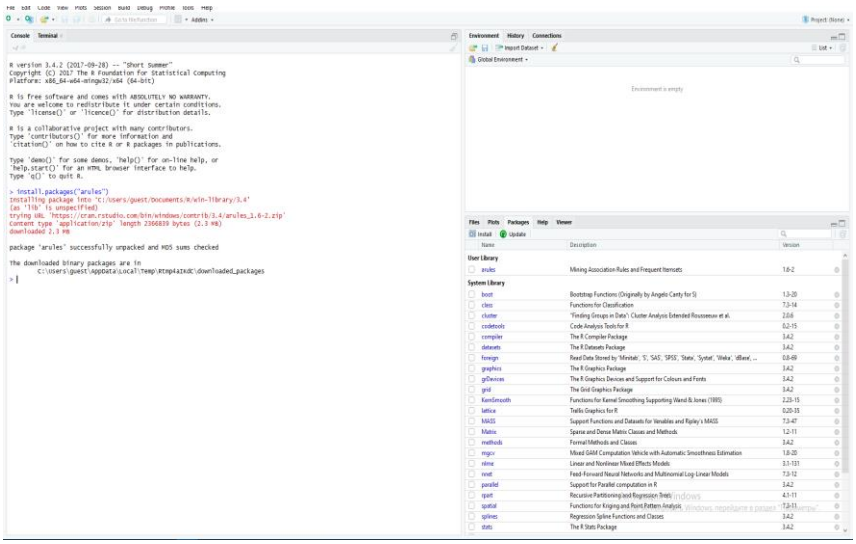


Рисунок 3. Установка пакета Arules в RStudio

Ниже представлена установка пакета ArulesViz в RStudio.

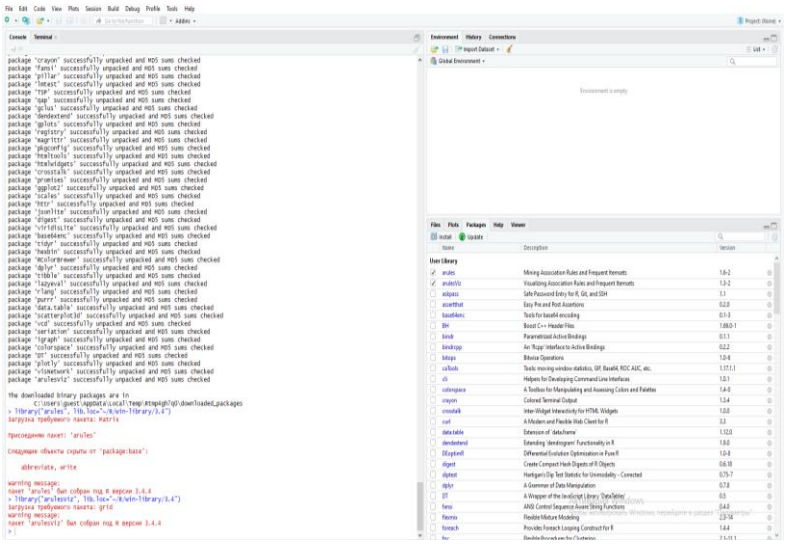


Рисунок 4. Установка пакета ArulesViz в RStudio

Возьмем базу Groceries, содержащуюся в пакете «arules»

```
> data("Groceries")
> dim(Groceries)
[1] 9835  169
> Groceries
transactions in sparse format with
 9835 transactions (rows) and
 169 items (columns)
```

Рисунок 5. База данных Groceries

Для анализа данных используем алгоритм Apriori с минимальной поддержкой 0.001, и значимостью 0.8

```
> rules <- apriori(Groceries, parameter = list(supp = 0.001, conf = 0.8))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen
        0.8    0.1    1 none FALSE            TRUE       5   0.001      1     10
 target   ext
  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [410 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Рисунок 6. Анализ данных алгоритмом apriori

В результате работы алгоритм вывел 410 правил

```
> summary(rules)
set of 410 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6
 29 229 140  12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   4.000   4.329   5.000   6.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001017   Min.   :0.8000   Min.   : 3.131   Min.   :10.00
 1st Qu.:0.001017   1st Qu.:0.8333   1st Qu.: 3.312   1st Qu.:10.00
 Median :0.001220   Median :0.8462   Median : 3.588   Median :12.00
 Mean   :0.001247   Mean   :0.8663   Mean   : 3.951   Mean   :12.27
 3rd Qu.:0.001322   3rd Qu.:0.9091   3rd Qu.: 4.341   3rd Qu.:13.00
 Max.   :0.003152   Max.   :1.0000   Max.   :11.235   Max.   :31.00

mining info:
      data ntransactions support confidence
 Groceries          9835   0.001        0.8
```

Рисунок 7. Список правил(rules)

Изобразим список правил графически:

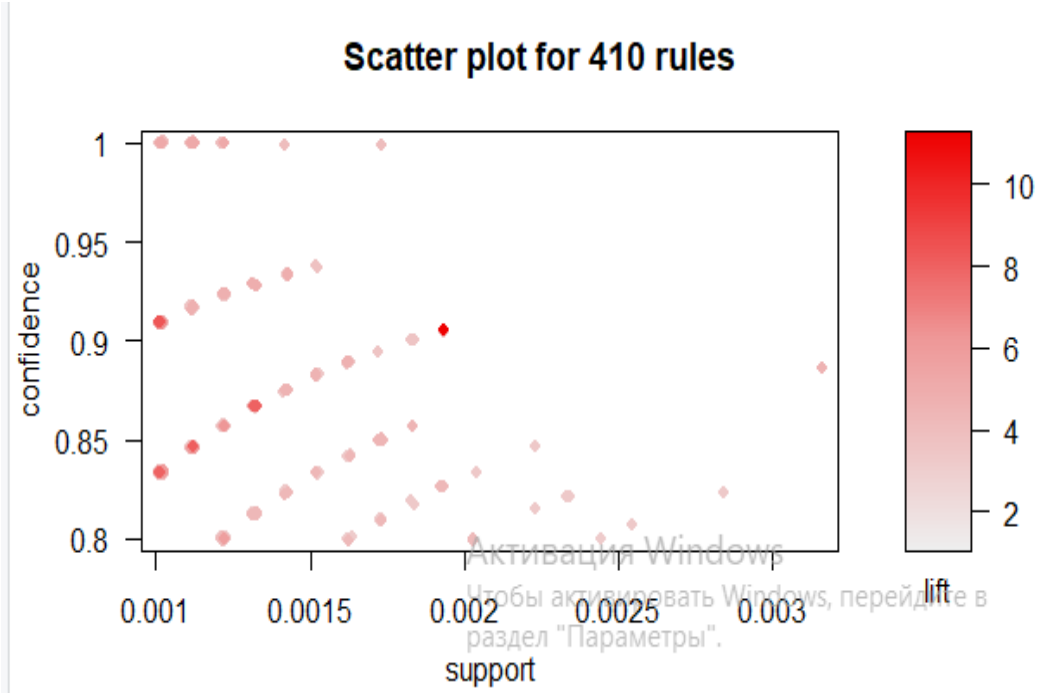


Рисунок 8. График правил

Заключение

Выполнив сортировку и выбрав пять лучших правил, можно сделать вывод, что чаще всего люди, покупающие ликер, красное или белое вино, обязательно купят бутылочку пива. Те, кто покупают творог и хлопья, приобретут молоко. С меньшей вероятностью молоко приобретут вместе с йогуртом и хлопьями.

```
> options(digits=2)
> inspect(rules[1:5])
    lhs                          rhs           support confidence lift count
[1] {liquor,red/blush wine} => {bottled beer} 0.0019  0.90       11.2 19
[2] {curd,cereals}          => {whole milk}   0.0010  0.91        3.6 10
[3] {yogurt,cereals}        => {whole milk}   0.0017  0.81        3.2 17
[4] {butter,jam}            => {whole milk}   0.0010  0.83        3.3 10
[5] {soups,bottled beer}    => {whole milk}   0.0011  0.92        3.6 11
```

Рисунок 9. Сортировка

Литература

[1] Интернет ресурс - https://docplayer.ru/36774167-Associativnye-pravila-algoritm- apriori-m-102.html. Режим доступа – 11.01.2019.

[2] Интернет ресурс - http://economics.studio/osnovyi-marketinga/assotsiativnyie-metodyi- 13000.html. Режим доступа – 15.01.2019.

[3] Интернет ресурс - https://ru.wikipedia.org/wiki/Apriori. Режим доступа – 15.01.2019.

AFFINITY DATA ANALYSIS. DISCOVERING ASSOCIATION RULES

A.J. ZALOGA
Master student of the Department of Mathematical and Information Support of Economic Systems, YKSUG

N.V. MARKOVSKAYA
Associate professor of the Department of Mathematical and Information Support of Economic Systems, YKSUG

Yanka Kupala State University of Grodno, Republic of Belarus
E-mail: Zaloga147@gmail.com , n.markovskaya@grsu.by

Abstract. In this paper, we consider associative rules. Associations rules learning - ARL is a simple, but quite often applicable in real life method of finding relationships (associations) in datasets, or, more precisely, data sets (itemsests). Development history: Piatesky-Shapiro G talked about this for the first time in detail in “Discovery, Analysis, and Presentation of Strong Rules.” (1991) Agrawal R, Imielinski T, Swami A developed in more detail in “Mining Association Rules between Sets of Items in Large Databases” (1993) and “Fast Algorithms for Mining Association Rules.” (1994). One of the limitations of the standard approach to detecting associations is that when searching for a large number of possible associations of a set of objects that may be associated, there is a big risk of finding a large number of random associations.

Keywords: association rules, apriori algorithm, full trust, collective power, conviction, lever, lift.