Федеральное государственное автономное образовательное учреждение

высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт Космических и Информационных Технологий

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №7

Методы и алгоритмы распознавания и обработки данных

Тема

# Поиск ассоциативных правил

Преподаватель

В.В. Тынченко

подпись, дата инициалы, фамилия

Студент КИ22-04-2М Д.С. Шабалин

номер группы, зачетной книжки подпись, дата инициалы, фамилия

Красноярск 2022

# ВВЕДЕНИЕ

**Цель работы:** научиться использовать ассоциативные правила для анализа данных.

# Ход работы Описание метода

Ассоциативные правила позволяют находить закономерности между связанными событиями. Примером такого правила служит утверждение, что покупатель, приобретающий 'Хлеб', приобретет и 'Молоко' с вероятностью 75%. Впервые эта задача была предложена для поиска ассоциативных правил для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis).

Транзакция — это множество событий, произошедших одновременно. Пусть имеется база данных, состоящая из покупательских транзакций. Каждая транзакция – это набор товаров, купленных покупателем за один визит. Такую транзакцию еще называют рыночной корзиной.

После определения понятия транзакция можно перейти к определению ассоциативного правила. Пусть имеется список транзакций. Необходимо найти закономерности между этими событиями. Как в условии, так и следствии правила должны находиться элементы транзакций.

Пусть I={i1, i2, ... , in} - множество элементов, входящих в транзакцию.

D - множество транзакций.

Ассоциативным правилом называется импликация XY (читается «X дает Y» или «из X следует Y»), где X  I, Y  I и X  Y=O .

Правило X  Y имеет поддержку s (support), если s% транзакций из D содержат X  Y, supp(X  Y)=supp(X  Y).

Достоверность правила показывает, какова вероятность того, что из X следует Y. Правило X  Y справедливо с достоверностью (confidence) с, если с% транзакций из D, содержащих X, также содержат Y, conf(X  Y)=supp(X  Y)/supp(X).

Покажем на конкретном примере: 75% транзакций, содержащих хлеб, также содержат молоко. 3% от общего числа всех транзакций содержат оба товара сразу. 75% - это достоверность (confidence) правила, 3% это поддержка (support) или «Если 'Хлеб', то 'Молоко'» с вероятностью 75%.

Другими словами, целью анализа является установление следующих зависимостей: если в транзакции встретился некоторый набор элементов X, то на основании этого можно сделать вывод о том, что другой набор элементов Y также должен появиться в этой транзакции. Установление таких зависимостей дает нам возможность находить очень простые и интуитивно понятные правила.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил причем поддержка и достоверность этих правил должны быть выше некоторых заранее определенных порогов, называемых, соответственно минимальной поддержкой (minsupport) и минимальной достоверностью (minconfidence). Аналогично, поддержка и достоверность ограничиваются сверху порогами максимальной поддержки (maxsupport) и максимальной достоверности (maxconfidence). В результате получаются два окна, в которые должны попасть поддержка и достоверность правила, чтобы оно было предъявлено аналитику.

Значения для параметров минимальная (максимальная) поддержка и минимальная (максимальная) достоверность выбираются таким образом, чтобы ограничить количество найденных правил. Если поддержка имеет большое значение, то алгоритмы будут находить правила, хорошо известные аналитикам или настолько очевидные, что нет никакого смысла проводить такой анализ. С другой стороны, низкое значение поддержки ведет к генерации огромного количества правил, что, конечно, требует существенных вычислительных ресурсов. Большинство интересных правил находится именно при низком значении порога поддержки, хотя слишком низкое значение поддержки ведет к генерации статистически необоснованных правил. Ассоциативные правила с высокой поддержкой могут применяться для формализации хорошо известных правил, например, в автоматизированных системах для управления процессами или персоналом. Надо отметить, что понятия «высокая» и «низкая» поддержка или достоверность очень сильно зависят от предметной области. Например, в торговле 1% вероятности совместного приобретения хлеба и молока не значит ничего, в то время как вероятность в 1% отказа двигателя самолета совершенно неприемлема, и такое правило становится чрезвычайно важным.

Поиск ассоциативных правил совсем не тривиальная задача, как может показаться на первый взгляд. Одна из проблем - алгоритмическая сложность при нахождении часто встречающих наборов элементов, т.к. с ростом числа элементов экспоненциально растет число потенциальных наборов элементов.

Обычные ассоциативные правила — это правила, в которых как в условии, так и в следствии присутствуют только элементы транзакций и при вычислении которых используется только информация о том, присутствует ли элемент в транзакции или нет. Фактически все приведенные выше примеры относятся к обычным ассоциативным правилам.

Для поиска обычных ассоциативных правил в программе служит обработчик «Ассоциативные правила».

Настройки.

Для начала необходимо указать, что является идентификатором (ID) транзакции, а что - элементом транзакции. Например, идентификатор транзакции — это номер чека или код накладной. А элемент — это наименование товара в чеке или накладной.

Затем следует настройка параметров поиска правил. Всего четыре параметра:

* Минимальная и максимальная поддержка. Ассоциативные правила ищутся только в некотором множестве всех транзакций. Для того чтобы транзакция вошла в это множество, она должна встретиться в исходной выборке количество раз, больше минимальной поддержки и меньше максимальной. Например, минимальная поддержка равна 1%, а максимальная - 20%. Количество элементов «Хлеб» и «Молоко» столбца

«Товар» с одинаковым значением столбца «Номер чека» встречаются в 5% всех транзакций (номеров чека). Тогда эти две строки войдут в искомое множество.

* Минимальная и максимальная достоверность. Это процентное отношение количества транзакций, содержащих все элементы, которые входят в правило, к количеству транзакций, содержащих элементы, которые входят в условие. Если транзакция — это заказ, а элемент - товар, то достоверность характеризует, насколько часто покупаются товары, входящие в следствие, если заказ содержит товары, вошедшие во всё правило.

Пример.

Дана таблица транзакций.

Таблица 1 – Таблица Транзакций



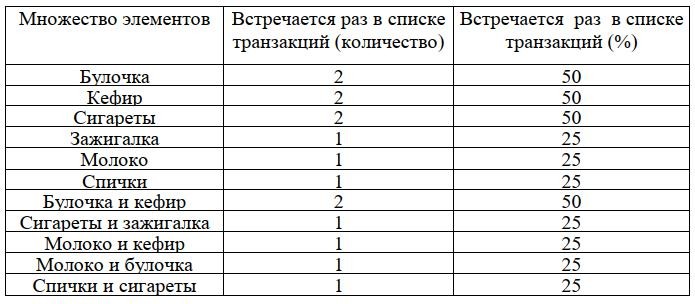
Список транзакций будет выглядеть так:

Таблица 2 – Список транзакций



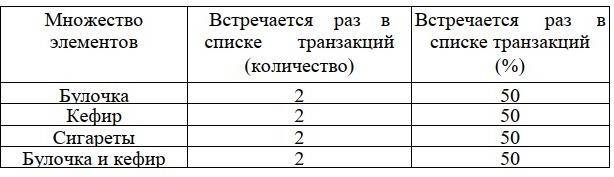
Всего исходные данные состоят из 4 транзакций. Полный список множеств для поиска правил выглядит так:

Таблица 3 – Полный список правил



Если установить минимальную поддержку 30%, а максимальную - 60%, то останется только часть списка множеств, так называемые часто встречающиеся множества:

Таблица 4 – Таблица часто встречающихся множеств



Правила будут искаться именно в этом последнем списке часто встречающихся множеств. Первые три множества в таблице одноэлементные, а последнее - двухэлементное. Ассоциативное правило можно построить только на основе 2-х и более элементного множества. Соответственно, если будут найдены только одноэлементные множества, то количество ассоциативных правил будет равно нулю. В этом случае следует уменьшить минимальную поддержку и/или увеличить максимальную. Тогда список множеств будет увеличен и, возможно, в него попадут двух и более элементные множества.

Выявление действительно интересных правил — это одна из главных подзадач при вычислении ассоциативных зависимостей. Для того чтобы

получить действительно интересные зависимости, нужно разобраться с несколькими эмпирическими правилами:

1. Уменьшение минимальной поддержки приводит к тому, что увеличивается количество потенциально интересных правил, однако это требует существенных вычислительных ресурсов. Одним из ограничений уменьшения порога минимальной поддержки является то, что слишком маленькая поддержка правила делает его статистически необоснованным.
2. Уменьшение порога достоверности также приводит к увеличению количества правил. Значение минимальной достоверности также не должно быть слишком маленьким, так как ценность правила с достоверностью 5% чаще всего настолько мала, что это и правилом считать нельзя.
3. Правило со слишком большой поддержкой с точки зрения статистики представляет собой большую ценность, но, с практической точки зрения, это, скорее всего, означает то, что-либо правило всем известно либо товары, присутствующие в нем, являются лидерами продаж, откуда следует их низкая практическая ценность.
4. Правило со слишком большой достоверностью практической ценности в контексте решаемой задачи не имеет, т.к. товары, входящие в следствие, покупатель, скорее всего, уже купил.

Если значение верхнего предела поддержки имеет слишком большое значение, то в обнаруженных правилах основную часть будут составлять товары - лидеры продаж. При таком раскладе не представляется возможным уменьшить минимальный порог поддержки до того значения, при котором могут появляться интересные правила. Причиной тому является просто огромное число правил и, как следствие, нехватка системных ресурсов. Причем получаемые правила процентов на 95 содержат товары - лидеры продаж.

Варьируя верхним и нижним пределами поддержки, можно избавиться от очевидных и неинтересных закономерностей. Как следствие, правила, генерируемые алгоритмом, принимают приближенный к реальности вид.

При большом ассортименте товара важно отобразить построенные правила в удобном виде. Для этого в Deductor Studio служат два визуализатора: «Дерево правил» и «Что-если».

Дерево правил — это всегда двухуровневое дерево. Оно может быть построено либо по условию, либо по следствию. При построении дерева

правил по условию на первом (верхнем) уровне находятся узлы с условиями, а на втором уровне - узлы со следствием.

Пример.



Рисунок 1 – Пример дерева правил

В этом примере узлы «Шпаклевки», «Эмали», «Грунтовки» находятся на верхнем уровне дерева и представляют собой условия. А «Герметики»,

«Клей - ж.гвозди» и т.д. - следствия. Это означает, что человек, купивший эмаль, купит еще и герметик с достоверностью 54.3%, клей с достоверностью 52.2% и т.д. В окне слева расположен список со следствиями для конкретного узла с условием. Для каждого следствия указана поддержка и достоверность. Например, в исходной выборке данных и эмаль и пена монтажная встретились одновременно в 25 транзакциях (чеках).

Второй вариант дерева правил - дерево, построенное по следствию.

Здесь на первом уровне располагаются узлы со следствием.



Рисунок 2 – Пример дерева правил

Например, для того чтобы человек приобрел герметик, он должен купить хотя бы один предмет из следующего списка: пигменты, пистолеты для герметиков, решетки радиаторные, уплотнители, шпаклевки, эмали и т.д. И для каждого из этих правил отображены поддержка и достоверность.

Чтобы перестраивать дерево по условию или по следствию служат две кнопки на панели инструментов: «Группировать по условию» и

«Группировать по следствию».

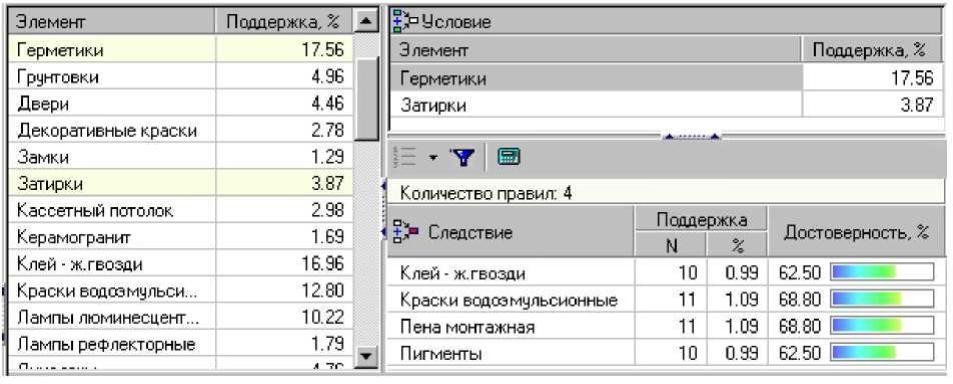


Рисунок 3 – Перестроенное дерево правил

Наиболее удобным и оперативным инструментом использования ассоциативных правил является анализ «что-если». Внешний вид формы для проведения такого анализа представлен на рисунке.

Слева находится список всех элементов транзакций, то есть весь ассортимент товара, который фигурирует в часто встречающихся множествах. Для каждого элемента указана поддержка - количество транзакций (чеков), в которых встречался данный элемент. Предположим, что человек купил герметики и затирки. Двойным щелчком мыши по элементу он переносится в список условий. Используя ассоциативные правила, можно предложить этому человеку сопутствующий товар, который приобретался совместно с тем, что он заказал. Этот товар отображается в списке следствий в правом нижнем окне, т.е. этому человеку можно предложить купить еще и клей, либо краски водоэмульсионные, либо пену монтажную, либо пигменты. Список следствий можно отсортировать по следствию, поддержке или достоверности либо отфильтровать, оставив в нем часть следствий. Для вычисления следствий по условиям служит кнопка «Обновить правила» на панели инструментов списка следствий. В связи с тем, что список элементов может быть очень большим, для быстрого поиска нужного элемента можно отсортировать список или воспользоваться поиском. Для этого нужно воспользоваться кнопками «Сортировка элементов» и «Найти...» на панели инструментов.

Пусть имеется база данных, состоящая из покупательских транзакций. Каждая транзакция – это набор товаров, купленных покупателем за один визит. Такую транзакцию еще называют рыночной корзиной. Целью анализа является установление следующих зависимостей: если в транзакции встретился некоторый набор элементов X, то на основании этого можно сделать вывод о том, что другой набор элементов Y также должен появиться в этой транзакции. Установление таких зависимостей дает нам возможность находить очень простые и интуитивно понятные правила.

Основными характеристиками таких правил являются поддержка и достоверность. Правило "Из Х следует Y" имеет поддержку s, если s% транзакций из всего набора содержат наборы элементов Х и Y. Достоверность правила показывает, какова вероятность того, что из X следует Y. Правило "Из X следует Y" справедливо с достоверностью с, если c% транзакций из всего множества, содержащих набор элементов X, также содержат набор элементов Y. Покажем на конкретном примере: пусть 75% транзакций, содержащих хлеб, также содержат молоко, а 3% от общего числа всех транзакций содержат оба товара. 75% – это достоверность правила, а 3%

— это поддержка.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил вида "из X следует Y", причем поддержка и достоверность этих правил должны находиться в рамках некоторых наперед заданных границ, называемых соответственно минимальной и максимальной поддержкой и минимальной и максимальной достоверностью.

Границы значений параметров поддержки и достоверности выбираются таким образом, чтобы ограничить количество найденных правил. Если поддержка имеет большое значение, то алгоритмы будут находить правила, хорошо известные аналитикам или настолько очевидные, что нет никакого смысла проводить такой анализ. С другой стороны, низкое значение поддержки ведет к генерации огромного количества правил, что, конечно, требует существенных вычислительных ресурсов. Тем не менее, большинство интересных правил находится именно при низком значении порога поддержки, хотя слишком низкое значение поддержки ведет к генерации статистически необоснованных правил. Таким образом, необходимо найти компромисс, обеспечивающий, во-первых, интересность правил и, во-вторых, их статистическую обоснованность. Поэтому значения этих границ напрямую зависят от характера анализируемых данных и подбираются индивидуально. Еще одним параметром, ограничивающим

Количество найденных правил является максимальная мощность часто встречающихся множеств. Если этот параметр указан, то при поиске правил будут рассматриваться только множества, количество элементов которых будет не больше данного параметра.

Для выполнения работы использую данные из файла «Supermarket.txt»

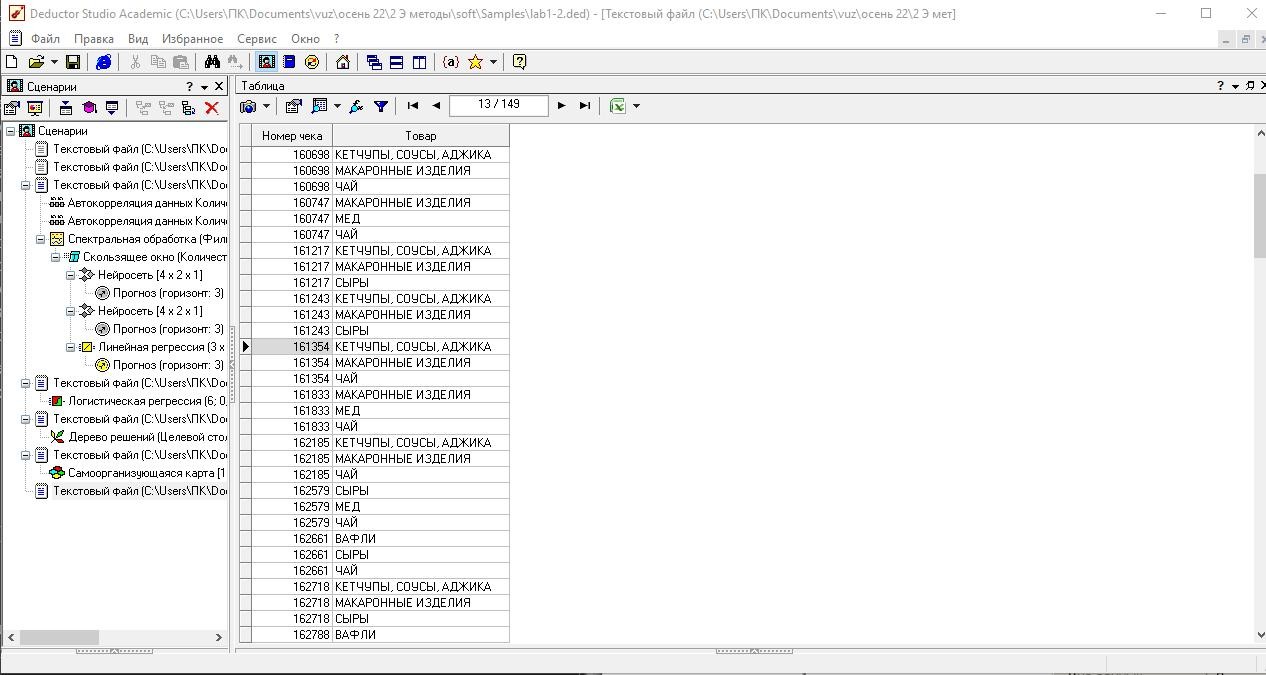


Рисунок 4 – Импортированные данные Настраиваю мастер обработки.

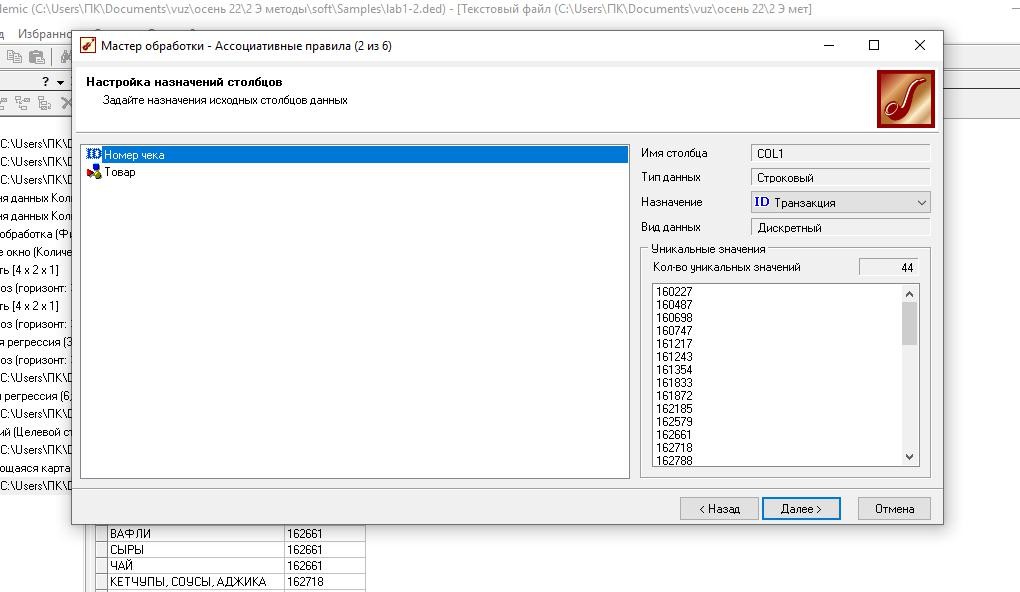


Рисунок 5 – Настройка параметров ассоциативных правил

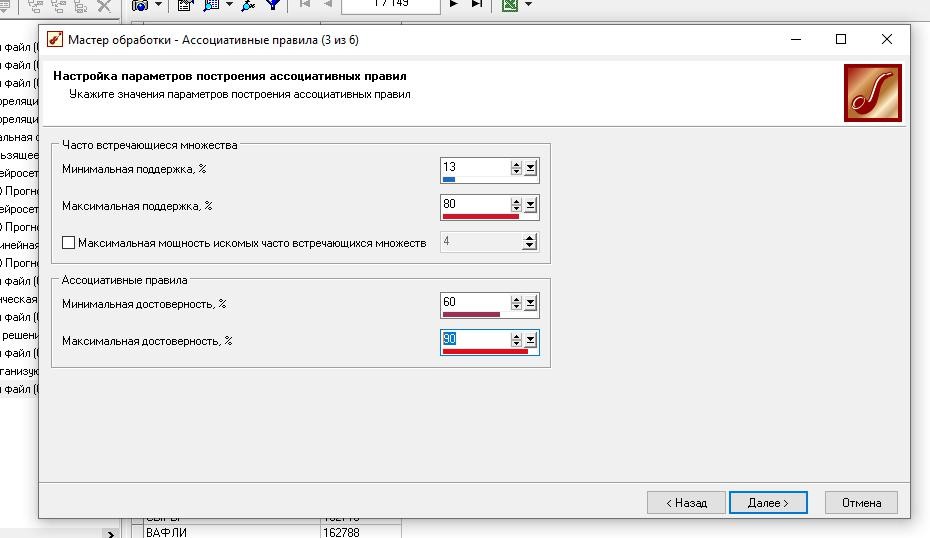


Рисунок 6 – Настройка параметров ассоциативных правил

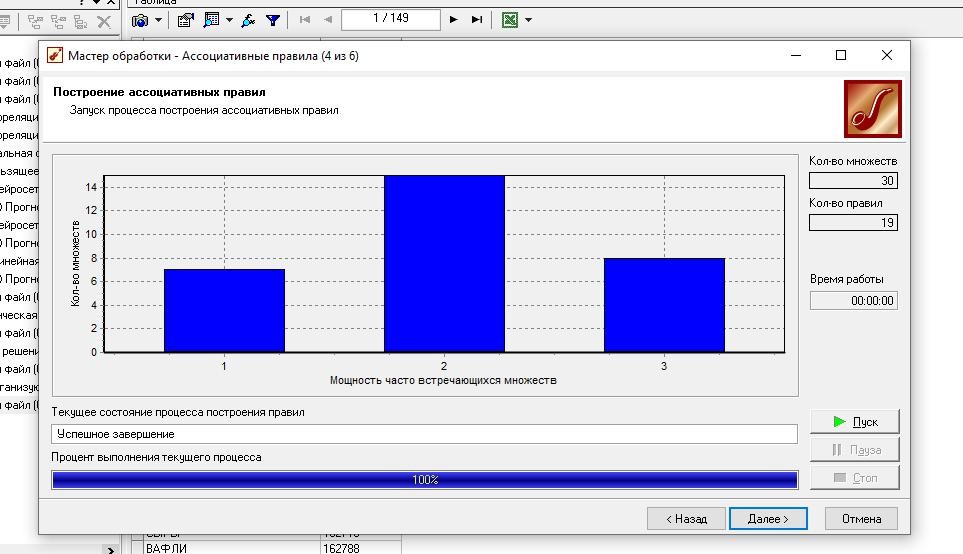


Рисунок 7 – Успешное завершение поиска

Получившиеся результаты представлены далее.

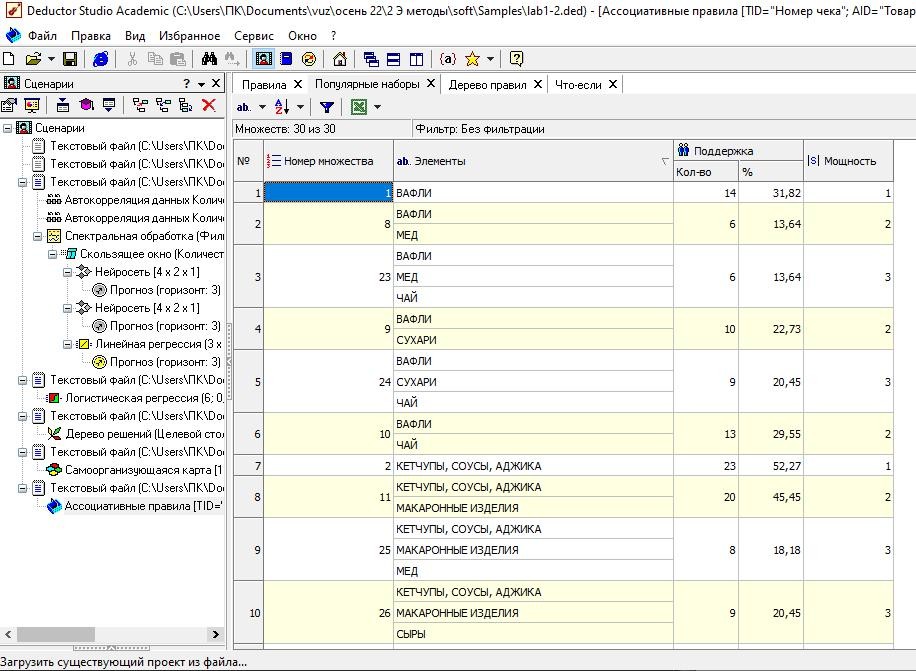


Рисунок 8 – Популярные наборы

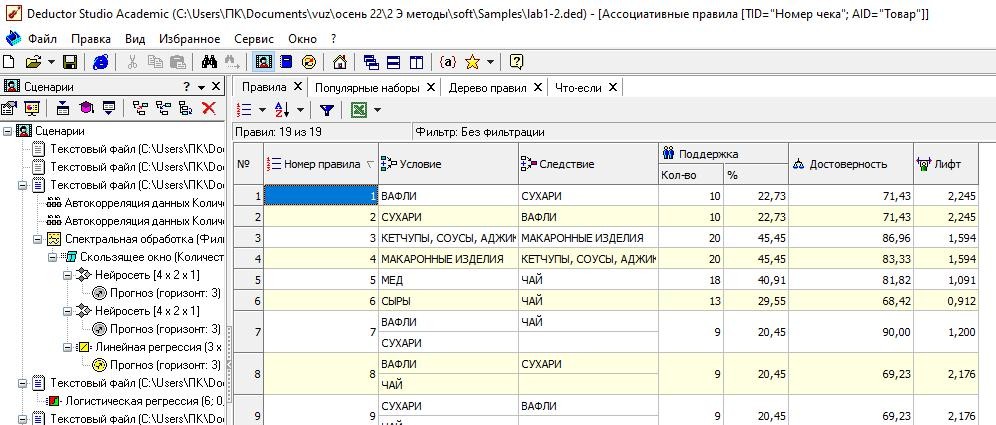


Рисунок 9 – Правила

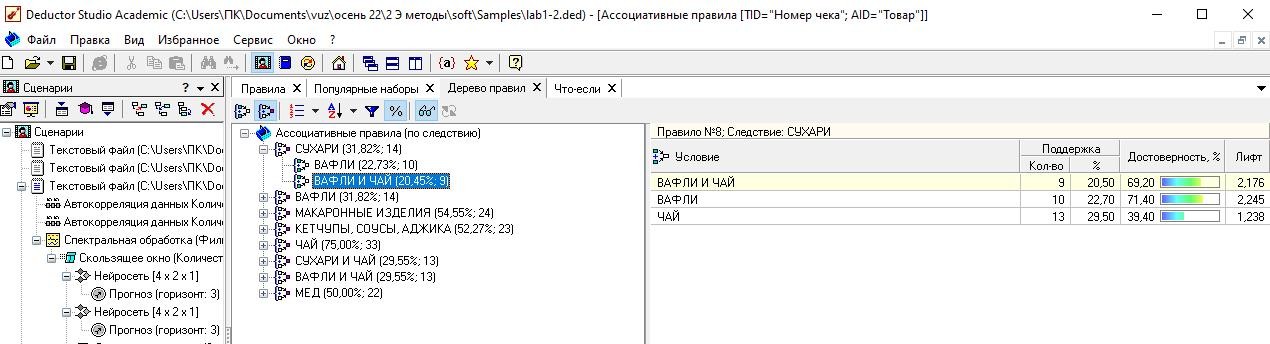


Рисунок 10 – Дерево правил

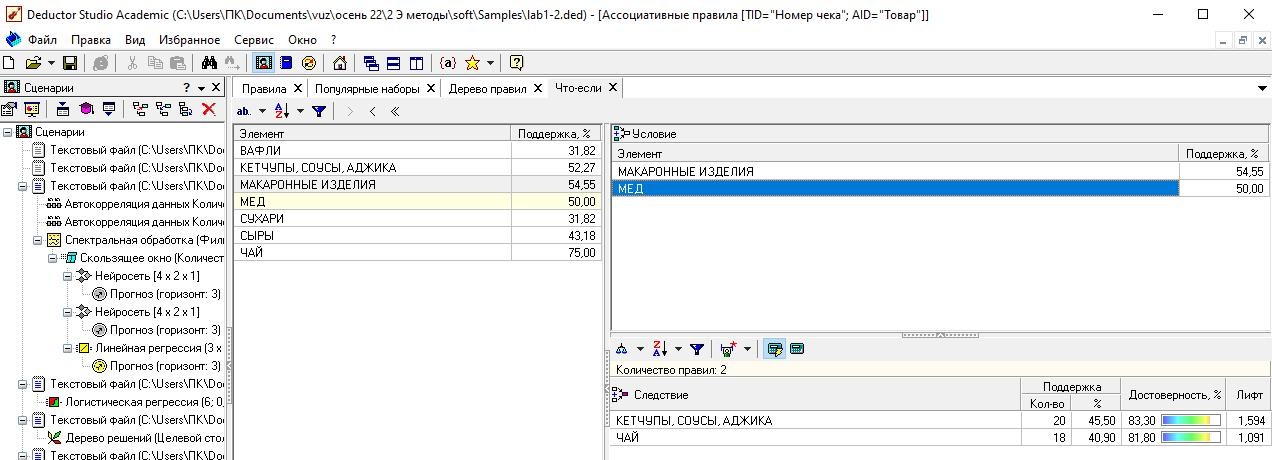


Рисунок 11 – Отображение «что-если»

# ВЫВОД

В ходе данной работы были получены знания о том, как использовать ассоциативные правила для анализа данных. Как показал данный пример, результаты анализа можно применить и для сегментации покупателей по поведению при покупках, и для анализа предпочтений клиентов, и для планирования расположения товаров в супермаркетах, кросс-маркетинге. Предлагаемый набор визуализаторов позволяет эксперту найти интересные, необычные закономерности, понять, почему так происходит, и применить их на практике.

В данной работе найденные правила можно использовать для сегментации клиентов на два сегмента: клиенты, покупающие макаронные изделия и соусы к ним, и клиенты, покупающие все к чаю. В разрезе анализа предпочтений можно узнать, что наибольшей популярностью в данном магазине пользуются чай, мед, макаронные изделия, кетчупы, соусы и аджика. В разрезе размещения товаров в супермаркете можно применить результаты предыдущих двух анализов, т. е. располагать чай рядом с медом, а кетчупы, соусы и аджику рядом с макаронными изделиями и т.д.