Поиск ассоциативных правил

Лекция 7 (1 час)

Емельянова М.Г.

Основные понятия

Ассоциативные правила – установление закономерностей между связанными событиями.

Впервые эта задача была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому её ещё называют анализом рыночной корзины (market basket analysis).

Пусть имеется база данных, состоящая из покупательских транзакций.

Каждая **транзакция** – это набор товаров, купленных покупателем за один визит.

Предметный набор – это непустое множество предметов, появившихся в одной транзакции.

Основные понятия

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

Ассоциативное правило состоит из двух наборов предметов:
условие (antecedent) и **следствие** (consequent).

«Если условие, то следствие»; A → B; «Из A следует B».

Условие и следствие часто называются соответственно левосторонним (left-hand side – LHS) и правосторонним (right-hand side – RHS) компонентами ассоциативного правила.

Объективные показатели значимости ассоциативных правил

Поддержка – количество или процент транзакций, содержащих как условие, так и следствие.
Правило A → B имеет поддержку S (support), если:

$$S(A \to B) = \frac{\text{количество транзакций, содержащих } A \text{ и } B}{\text{общее количество транзакций}}$$

Достоверность ассоциативного правила (confidence) C представляет собой меру точности правила и определяется, как отношение количества транзакций, содержащих условие и следствие, к количеству транзакций, содержащих только условие:

$$C(A \to B) = \frac{\text{количество транзакций, содержащих } A \text{ и } B}{\text{количество транзакций, содержащих только } A}$$

Объективные показатели значимости ассоциативных правил

Пример. Пусть 75 % транзакций, содержащих хлеб, также содержат молоко, а 3 % от общего числа всех транзакций содержат оба товара, тогда 75 % – это достоверность правила, а 3 % – это поддержка.

Если поддержка и достоверность достаточно высоки, можно с большой вероятностью утверждать, что любая будущая транзакция, которая включает условие, будет также содержать и следствие.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил вида A → B, причем поддержка и достоверность этих правил должны находиться в рамках некоторых наперед заданных границ, называемых **минимальной и максимальной поддержкой, минимальной и максимальной достоверностью**.

Субъективные показатели значимости ассоциативных правил

Лифт (lift) – это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом:

$L(A \rightarrow B) = C(A \rightarrow B) / S(B)$.

Лифт является обобщенной мерой связи двух предметных наборов: при значениях лифта > 1 связь положительная, при 1 она отсутствует, а при значениях < 1 – отрицательная. Значения лифта большие, чем единица, показывают, что условие чаще появляется в транзакциях, содержащих следствие, чем в остальных.

Субъективные показатели значимости ассоциативных правил

Левередж (leverage) – это разность между наблюдаемой частотой, с которой условие и следствие появляются совместно (то есть поддержкой ассоциации), и произведением частот появления (поддержек) условия и следствия по отдельности: $T(A \rightarrow B) = S(A \rightarrow B) - S(A)S(B)$.

Если Левередж≈0, то правило не значимо.

Алгоритм Apriori

Поиск ассоциативных правил – алгоритм Apriori (1994 г.).

Прокопенко Н.Ю. Системы поддержки принятия решений [Электронный ресурс]: учеб. пособие /Н. Ю. Прокопенко; Нижегор. гос. архитектур.-строит. ун-т. – Н. Новгород: ННГАСУ, 2017. – 188 с. ISBN 978-5-528-00202-6.

Стр. 78-82.

https://loginom.ru/blog/apriori

Пример поиска ассоциативных правил в АП Deductor Studio Academic

Supermarket.txt содержит данные о продажах товаров.

Необходимо решить задачу анализа потребительской корзины с целью последующего применения результатов для стимулирования продаж. Для этого производится поиск товаров, присутствие которых в транзакции влияет на вероятность наличия других товаров или комбинаций товаров.

Фрагмент данных

Номер чека	Товар
160698	КЕТЧУПЫ, СОУСЫ, АДЖИКА
160698	МАКАРОННЫЕ ИЗДЕЛИЯ
160698	ЧАЙ
160747	МАКАРОННЫЕ ИЗДЕЛИЯ
160747	МЕД
160747	ЧАЙ
161217	КЕТЧУПЫ, СОУСЫ, АДЖИКА
161217	МАКАРОННЫЕ ИЗДЕЛИЯ

Шаги поиска ассоциативных правил в АП Deductor Studio Academic

Шаг 1.
Импортируем данные из текстового файла, представляем в виде таблицы.

Шаг 2.
Запустим мастер обработки. В нём выберем тип обработки «Ассоциативные правила».

Шаг 3.
В мастере обработки укажем, какой столбец является идентификатором транзакции (номер чека), а какой элементом транзакции (товар).

Шаг 4.
В мастере обработки настроим параметры построения ассоциативных правил: минимальную и максимальную поддержку, минимальную и максимальную достоверность, а также максимальную мощность множества.

Шаги поиска ассоциативных правил в АП Deductor Studio Academic

Шаг 5.

Запустим процесс поиска ассоциативных правил. На экране будет отображаться информация о количестве множеств, количестве найденных правил, а также гистограмма распределения найденных часто встречающихся множеств по мощности.

Шаг 6.

После завершения процесса поиска, проанализируем полученные результаты, используя появившиеся специальные визуализаторы **«Популярные наборы», «Правила», «Дерево правил», «Что если».**

Визуализаторы

Популярные наборы – это множества, состоящие из одного и более элементов, которые наиболее часто встречаются в транзакциях одновременно. Насколько часто встречается множество в исходном наборе транзакций, можно судить по поддержке. Данный визуализатор отображает множества в виде списка.

Номер множества	Элемент	Поддержка		Мощность
		Кол-во	%	
1	ВАФЛИ	14	31,82	1
3	ВАФЛИ СУХАРИ	10	22,73	2
2	СУХАРИ	14	31,82	1

Визуализаторы

Визуализатор **«Правила»** отображает ассоциативные правила в виде списка правил. Этот список представлен таблицей со столбцами: «номер правила», «условие», «следствие», «поддержка, кол-во», «поддержка, %», «достоверность», «лифт».

Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
			Кол-во	%		
1	ВАФЛИ	СУХАРИ	10	22,73	71,43	2,245
2	СУХАРИ	ВАФЛИ	10	22,73	71,43	2,245
3	КЕТЧУПЫ, СОУСЫ	МАКАРОННЫЕ ИЗДЕЛИЯ	20	45,45	86,96	1,594

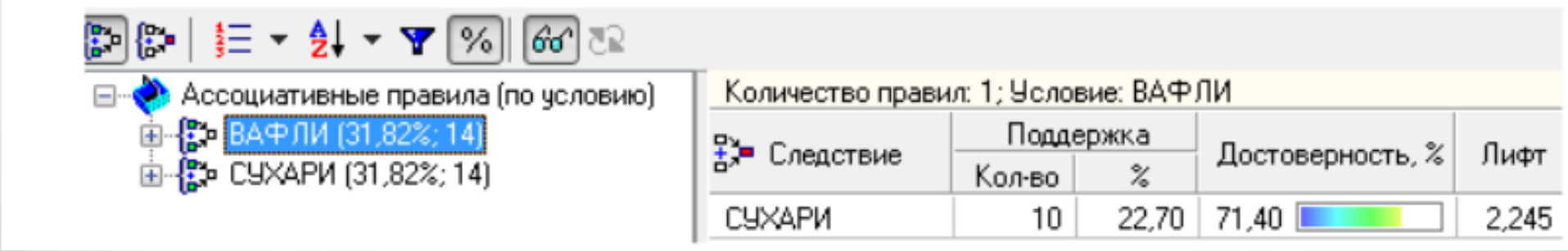
Таким образом, эксперту предоставляется набор правил, которые описывают поведение покупателей. Например, если покупатель купил вафли, то он с вероятностью 71,4 % также купит и сухари.

Визуализаторы

Визуализатор **«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.
Второй вариант дерева правил – дерево, построенное **по следствию**. Здесь на первом уровне располагаются узлы со следствием. Справа от дерева находится список правил, построенный по выбранному узлу дерева.

Для каждого правила отображаются поддержка и достоверность.

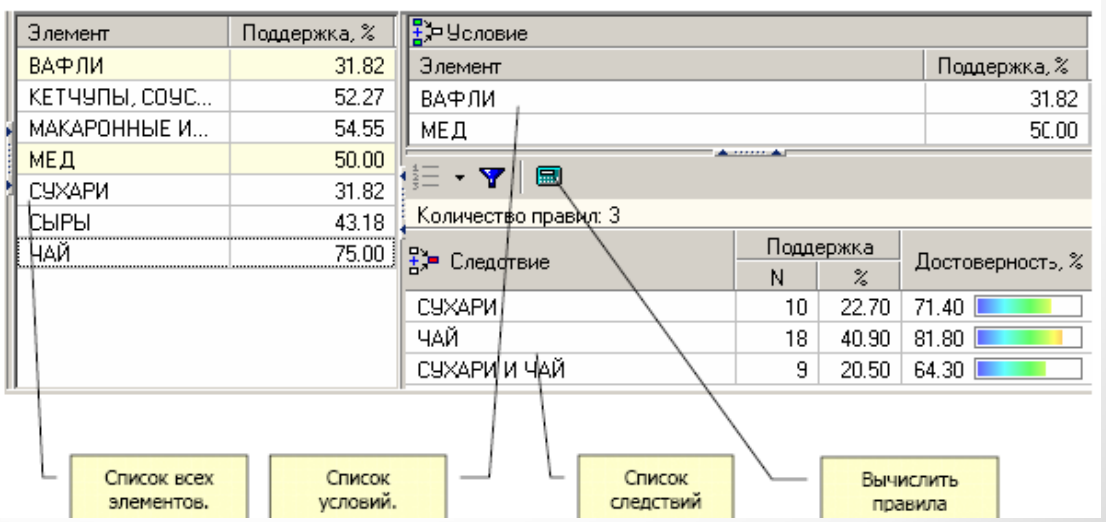


Если покупатель приобрел вафли, то он с вероятностью 71 % также приобретет сухари.

Визуализаторы

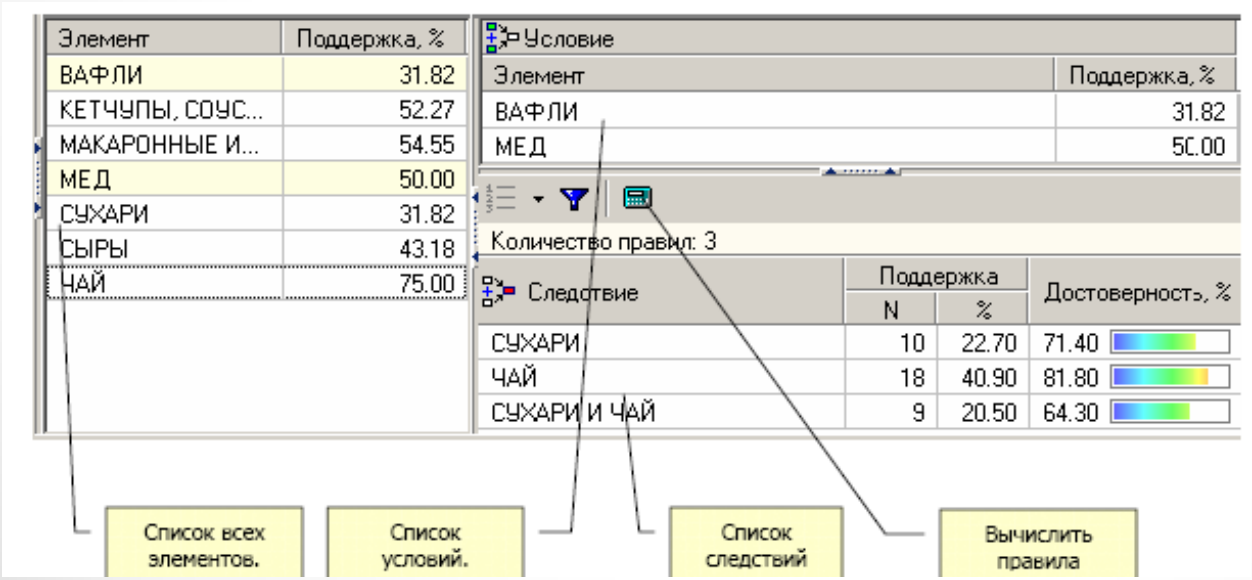
Анализ **«Что если»** в ассоциативных правилах позволяет ответить на вопрос: «Что получим в качестве следствия, если выберем данные условия? Например. Какие товары приобретаются совместно с выбранными товарами?

В окне слева расположен список всех элементов транзакций. Справа от каждого элемента указана поддержка. В правом верхнем углу расположен список элементов, входящих в условие. В правом нижнем углу расположен список следствий. Справа от элементов списка отображается поддержка и достоверность.



Визуализаторы

Пусть необходимо проанализировать, что, возможно, забыл покупатель приобрести, если он уже взял вафли и мёд. Для этого необходимо добавить в список условий эти товары и затем нажать на кнопку «Вычислить правила». При этом в списке следствий появятся товары, совместно приобретаемые с данными. Появятся «сухари», «чай», «сухари и чай».



Визуализаторы

Существующий в АП Deductor набор визуализаторов позволяет эксперту (аналитику) найти интересные, необычные закономерности, понять, почему так происходит и применить их на практике.

Результаты анализа можно применить, например, для анализа предпочтений клиентов, для планирования расположения товаров в супермаркетах.

Правила выявления интересных зависимостей

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. 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.

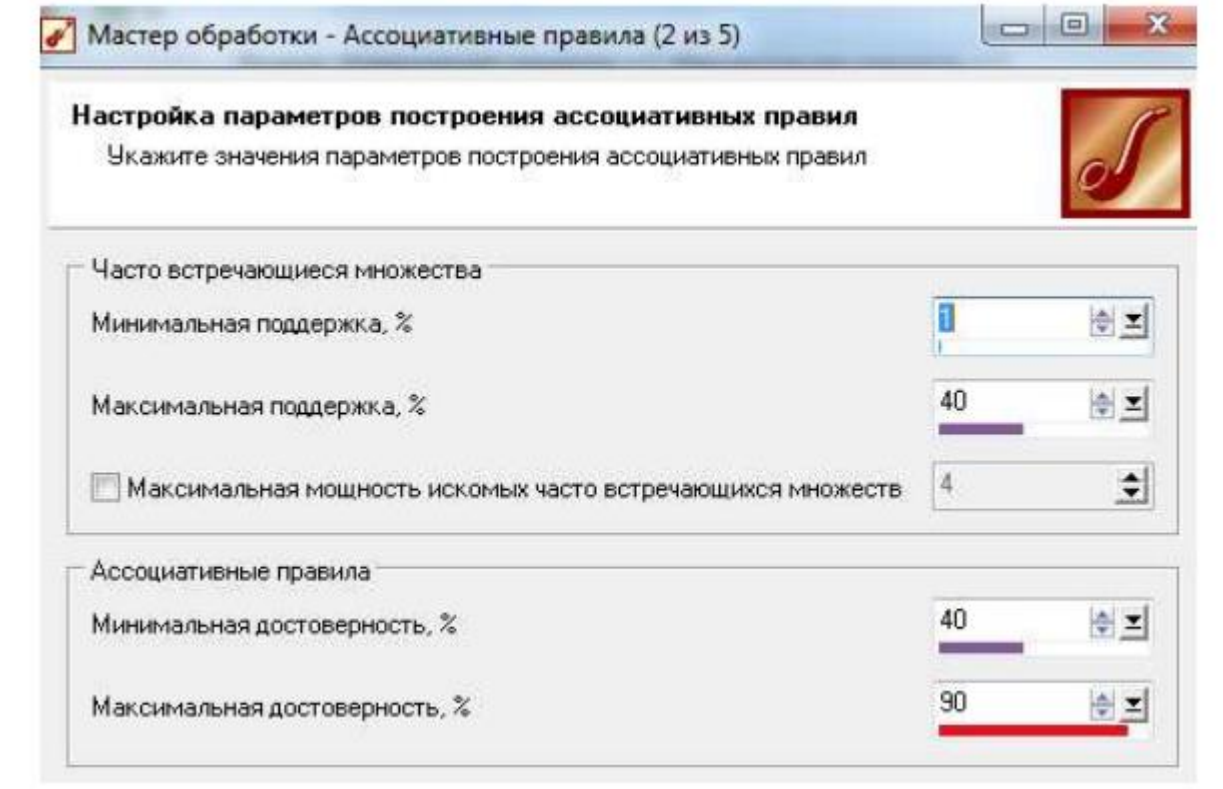
Правила выявления интересных зависимостей

3. 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.

4. Интерпретация ассоциативных правил.

Параметры построения ассоциативных правил

По умолчанию в обработчике установлены следующие границы поддержки – 1 % и 20 % и достоверности 40 % и 90 %.
Можно изменить максимальную поддержку до 40%.



Три вида ассоциативных правил

Полезные правила содержат действительную информацию, которая ранее была неизвестна, но имеет логичное объяснение. Такие правила могут быть использованы для принятия решений, приносящих выгоду.

Тривиальные правила содержат действительную и легко объяснимую информацию, которая уже известна. Такие правила, хотя и объяснимы, но не могут принести какой-либо пользы, т.к. отражают известные законы в исследуемой области или результаты прошлой деятельности. При анализе рыночных корзин в правилах с самой высокой поддержкой и достоверностью окажутся товары-лидеры продаж. Практическая ценность таких правил крайне низка.

Непонятные правила содержат информацию, которая не может быть объяснена. Такие правила могут быть получены или на основе аномальных значений, или глубоко скрытых знаний. Напрямую такие правила нельзя использовать для принятия решений, т.к. их необъяснимость может привести к непредсказуемым результатам. Для лучшего понимания требуется дополнительный анализ.

Вопросы для проверки

1. Что такое ассоциативные правила?
2. Что такое поддержка?
3. Что такое достоверность?
4. Какой алгоритм используется для поиска ассоциативных правил?
5. Какие визуализаторы используются для просмотра результатов поиска?