ЛЕКЦИЯ 1. Интеллектуальный анализ данных: задача ассоциации

Ассоциативные правила.

Аффинитивный анализ (affinity analysis) – один из распространенных методов Data Mining. Его название происходит от английского слова affinity, которое в переводе означает «близость», «сходство». Цель данного метода исследование взаимной связи между событиями, которые происходят совместно. Разновидностью аффинитивного анализа является анализ рыночной корзины, цель которого обнаружить ассоциации между различными событиями, то есть найти правила для количественного описания взаимной связи между двумя или более событиями. Такие правила называются ассоциативными правилами.

Примерами приложения ассоциативных правил могут быть следующие задачи:

- выявление наборов товаров, которые в супермаркетах часто покупаются вместе или никогда не покупаются вместе;
- определение доли клиентов, положительно относящихся к нововведениям в их обслуживании;
- определение профиля посетителей веб-ресурса;
- определение доли случаев, в которых новое лекарство оказывает опасный побочный эффект.

Базовым понятием в теории ассоциативных правил является *транзакция* – некоторое множество событий, про исходящих совместно. Типичная транзакция приобретение клиентом товара в супермаркете. В подавляющем большинстве случаев клиент покупает не один товар, а набор товаров, который называется рыночной корзиной. При этом возникает вопрос: является ли покупка одного товара в корзине следствием или причиной покупки другого товара, то есть связаны ли данные события? Эту связь и устанавливают ассоциативные правила, Например, может быть обнаружено ассоциативное правило, утверждающее, что клиент, купивший молоко, с вероятностью 75 % купит и хлеб.

Следующее важное понятие – *предметный набор*. Это непустое множество предметов, появившихся в одной транзакции.

Анализ рыночной корзины – это анализ наборов данных для определения комбинаций товаров, связанных между собой, Иными словами, производится поиск товаров, присутствие которых в транзакции влияет на вероятность наличия других товаров или комбинаций товаров.

Современные кассовые аппараты в супермаркетах позволяют собирать информацию о покупках, которая может храниться в базе данных. Затем

накопленные данные могут использоваться для построения систем поиска ассоциативных правил.

В табл. 2.1 представлен простой пример, содержащий данные о рыночной корзине. В каждой строке указывается комбинация продуктов, приобретенных за одну покупку. Хотя на практике приходится иметь дело с миллионами транзакций, в которых участвуют десятки и сотни различных продуктов, пример ограничен 10 транзакциями, содержащими 13 видов продуктов: чтобы проиллюстрировать методику обнаружения ассоциативных правил, этого достаточно.

Таблица 2.1. Пример набора транзакций

№	Транзакция
1	Сливы, салат, помидоры
2	Сельдерей, конфеты
3	Конфеты
4	Яблоки, морковь, помидоры, картофель, конфеты
5	Яблоки, апельсины, салат, конфеты, помидоры
6	Персики, апельсины, сельдерей, помидоры
7	Фасоль, салат, помидоры
8	Апельсины, салат, морковь, помидоры, конфеты
9	Яблоки, бананы, сливы, морковь, помидоры, лук, конфеты
10	Яблоки, картофель

Визуальный анализ примера показывает, что все четыре транзакции, в которых фигурирует салат, также включают помидоры и что четыре из семи транзакций, содержащих помидоры, также содержат салат. Салат и помидоры в большинстве случаев покупаются вместе. Ассоциативные правила позволяют обнаруживать и количественно описывать такие совпадения,

Ассоциативное правило состоит из двух наборов предметов, называемых условием и следствием, записываемых в виде $X \to Y$, что читается так: «Из X следует Y», Таким образом, ассоциативное правило формулируется в виде: «Если условие, то следствие».

Условие может ограничиваться только одним предметом. Правила обычно отображаются с помощью стрелок, направленных от условия к следствию, например, помидоры → салат. Условие и следствие часто называются соответственно: левосторонним и правосторонним компонентами ассоциативного правила.

Ассоциативные правила описывают связь между наборами предметов, соответствующие условию и следствию. Эта связь характеризуется двумя показателями – *поддержкой* (support) и *достоверностью* (confidence).

Обозначим базу данных транзакций как D, а число транзакций в этой базе – как N. Каждая транзакция D_i представляет собой некоторый набор предметов. Обозначим через S поддержку, через C – достоверность.

Поддержка ассоциативною правила – это число транзакций, которые содержат kar условие, так и следствие. Например, для ассоциации $A \to B$ можно записать:

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

Достоверность ассоциативного правила $A \to B$ представляет собой меру точности правила и определяется как отношение количества транзакций, содержащих и условие, и следствие, к количеству транзакций, содержащих только условие:

$$C(A \to B) = \mathrm{P}(A \cap B) \Big/ \mathrm{P}(A) = \frac{\text{количество транзакций, содержащих A и B}}{\text{количество транзакций, содержащих только A}}$$

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

Вычислим поддержку и достоверность для ассоциаций из табл. 2.1,

Возьмем ассоциацию салат помидоры. Поскольку количество транзакций, содержащих как салат, так и помидоры, равно 4, а общее число транзакций 10, то поддержка данной ассоциации будет:

$$S(\text{салат} \to \text{помидоры}) = 4/10 = 0{,}4$$

Поскольку количество транзакций, содержащих только салат (условие), равно 4, то достоверность данной ассоциации будет:

$$C(\text{салат} \to \text{помидоры}) = 4/4 = 1$$

Иными словами, все наблюдения, содержащие салат, также содержат и помидоры, из чего делаем вывод о том, что данная ассоциация может рассматриваться как правило. С точки зрения интуитивного поведения такое правило вполне объяснимо, поскольку оба продукта широко используются для приготовления растительных блюд и часто покупаются вместе.

Теперь рассмотрим ассоциацию конфеты → помидоры, в которой содержатся, в общем то, слабо совместимые в гастрономическом плане продукты: тот, кто решил приготовить растительное блюдо, вряд ли станет покупать

конфеты, а тот, кто желает приобрести что-нибудь к чаю, скорее всего, не станет покупать помидоры.

Поддержка данной ассоциации: $S = 4/10 = 0{,}4$, а достоверность: C=4/6=0,67. Таким образом, сравнительно невысокая достоверность данной ассоциации дает повод усомниться в том, что она является правилом.

Аналитики могут отдавать предпочтение правилам, которые имеют только высокую поддержку или только высокую достоверность либо, что является наиболее частым, оба этих показателя. Правила, для которых значения поддержки или достоверности превышают определенный, заданный пользователем порог, называются *сильными правилами*. Например, аналитика может интересовать, какие товары, покупаемые вместе в супермаркете, образуют ассоциации с минимальной поддержкой 20% и минимальной достоверностью 70 %. А при анализе с целью обнаружения мошенничеств может потребоваться уменьшить поддержку до 1%, поскольку с мошенничеством связано сравнительно небольшое число транзакций.

Значимость ассоциативных правил.

Методики поиска ассоциативных правил обнаруживают все ассоциации, которые удовлетворяют ограничениям на поддержку и достоверность, наложенным пользователем. Это приводит к необходимости рассматривать десятки и сотни тысяч ассоциаций, что делает невозможным обработку такого количества данных вручную. Число правил желательно уменьшить таким образом, чтобы проанализировать только наиболее значимые из них. Значимость часто вычисляется как разность между поддержкой правила в целом и произведением поддержки только условия и поддержки только следствия.

Если условие и следствие независимы, то поддержка правила примерно соответствует произведению поддержек условия и следствия, то есть $S_{AB} \approx S_A S_B$. Это значит, что, хотя условие и следствие часто встречаются вместе, не менее часто они встречаются и по отдельности. Например, если товар A встречался в 70 транзакциях из 100, а товар B в 80 и в 50 транзакциях из 100 они встречаются вместе, то, несмотря на высокую поддержку ($S_{AB} = 0{,}5$), это не обязательно правило. Просто эти товары покупаются независимо друг от друга, но в силу их популярности часто встречаются в одной транзакции. Поскольку произведение поддержек условия и следствия$S_A S_B = 0{,}7 \times 0{,}8 = 0{,}56$, то есть отличается от $S_{AB} = 0{,}5$ всего на 0,06, предположение о независимости товаров A и B достаточно обоснованно.

Однако если условие и следствие независимы, то правило вряд ли

представляет интерес независимо от того, насколько высоки его поддержка и достоверность. Например, если статистика дорожно-транспортных происшествий показывает, что из 100 аварий в 80 участвуют автомобили марки ВАЗ, то на первый взгляд это выглядит как правило *«если авария, то ВАЗ»*. Но если учесть, что парк автомобилей ВАЗ составляет, скажем, 80% от общего числа легковых автомобилей, то такое правило вряд ли можно назвать значимым.

По этой причине при поиске ассоциативных правил используются дополнительные показатели, позволяющие оценить значимость правила. Можно выделить объективные и субъективные меры значимости правил, Объективными являются такие меры, как поддержка и достоверность, которые могут применяться независимо от конкретного приложения. Субъективные меры связаны со специальной информацией, определяемой пользователем в контексте решаемой задачи, Такими субъективными мерами являются лифт (lift) и левередж (от англ., leverage «плечо», «рычаг»).

Лифт (оригинальное название — *интерес*) вычисляется следующим образом:

$$L(A \to B) = C(A \to B)/S(B)$$

Лифт – отношение частоты появления условия в транзакциях, которые также содержат и следствие к частоте появления следствия в целом. Значения лифта большие 1 показывают, что условие чаще появляется в транзакциях, содержащих следствие, чем в остальных. Можно утверждать, что лифт является обобщенной мерой связи двух предметных наборов: при значениях лифта больше 1 связь положительная, при 1 она отсутствует, а при значениях меньше 1 – отрицательная.

Рассмотрим ассоциацию помидоры → салат из табл. 2.1.

$S(\text{салат}) = 4/10 = 0{,}4;\ \text{C}(\text{помидоры} \to \text{салат}) = 4/7 = 0{,}57.$

Следовательно, $L(\text{помидоры} \to \text{салат}) == 0{,}57/0{,}4 == 1{,}425.$

Теперь рассмотрим ассоциацию помидоры → конфеты.

$S(\text{конфеты}) = 0{,}6;\ \text{C}(\text{помидоры} \to \text{конфеты}) = 4/7 = 0{,}57.$

Тогда $L(\text{помидоры} \to \text{конфеты}) = 0{,}57/0{,}6 = 0{,}95.$

Большее значение лифта для первой ассоциации показывает, что помидоры больше влияют на частоту покупок салата, чем конфет.

Хотя лифт используется широко, он не всегда оказывается удачной мерой значимости правила. Правило с меньшей поддержкой и большим лифтом может быть менее значимым, чем альтернативное правило с большей поддержкой и меньшим лифтом, потому что последнее применяется для большего числа

покупателей. Значит, увеличение числа покупателей приводит к возрастанию связи между условием и следствием.

Другой мерой значимости правила, предложенной Г. Пятецким-Шапиро, является левередж:

$$\mathrm{T}(\mathrm{A} \to \mathrm{B}) = S(A \to \mathrm{B}) - S(A)S(B).$$

Левередж – 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.

Рассмотрим ассоциации $\text{морковь} \to \text{помидоры}$ и $\text{салат} \to \text{помидоры}$, которые имеют одинаковую поддержку $C = 1$, поскольку салат и морковь всегда продаются вместе с помидорами (см, табл. 2.1). Лифты для данных ассоциаций так же будут одинаковыми, поскольку в обеих ассоциациях поддержка следствия $S(\text{помидоры}) = 7/10 = 0{,}7$.

Тогда лифт равен:

$L(\text{морковь} \to \text{помидоры}) = L(\text{салат} \to \text{помидоры}) = 1/0{,}7 = 1{,}43.$

Последняя ассоциация представляет больший интерес, так как она встречается чаще, то есть применяется для большего числа покупателей.

$$S(\text{морковь} \to \text{помидоры}) = 3/10 = 0{,}3;$$

$$S(\text{морковъ}) = 0{,}3; S(\text{помидоры}) = 0{,}7.$$

Таким образом, левередж равен:

$\mathrm{T}(\text{морковь} \to \text{помиидоры}) = 0{,}3 - 0{,}3 \times 0{,}7 = 0{,}09.$

$S(\text{салат} \to \text{помидоры}) = 0{,}4;\ S(\text{салат}) = 0{,}4;\ S(\text{помидоры}) = 0{,}7.$

Следовательно, $\mathrm{T}(\text{салат} \to \text{помидоры}) = 0{,}4 - 0{,}4 \times 0{,}7 = 0{,}12.$

Итак, значимость второй ассоциации больше, чем первой.

Иногда лифт, представляющий собой меру полезности ассоциативного правила, называют улучшением и вычисляют подобно левереджу, только берется не разность, а отношение наблюдаемой частоты и частот появления по отдельности:

$$I(A \to B) = C(A \to B)/S(A) \times S(B)$$

Улучшение (лифт) показывает, полезнее ли правило случайного угадывания,

Если $I(A \to B) > 1$, это значит, что вероятнее предсказать наличие набора B с помощью правила, чем угадать случайно.

Такие меры, как лифт и левередж, могут использоваться для последующего ограничения набора рассматриваемых ассоциаций путем установки порога значимости, ниже которого ассоциации отбрасываются.

Поиск ассоциативных правил.

В процессе поиска ассоциативных правил может производиться обнаружение всех ассоциаций, поддержка и достоверность для которых превышают заданный минимум. Простейший алгоритм поиска ассоциативных правил рассматривает все возможные комбинации условий и следствий, оценивает для них поддержку и достоверность, а затем исключает все ассоциации, которые не удовлетворяют заданным ограничениям, Число возможных ассоциаций с увеличением числа предметов растет экспоненциально. Если в базе данных транзакций присутствует k предметов и все ассоциации являются бинарными (то есть содержат по одному предмету в условии и следствии), то потребуется проанализировать $k \times 2^{k-1}$ ассоциаций. Поскольку реальные базы данных транзакций, рассматриваемые при анализе рыночной корзины, обычно содержат тысячи предметов, вычислительные затраты при поиске ассоциативных правил огромны. Например, если рассматривать выборку, содержащую всего 100 предметов, то количество ассоциаций, образуемых этими предметами, составит $100 \times 2^{99} \approx 6{,}4 \times 10^{31}$. Поиск ассоциативных правил путем вычисления поддержки и достоверности для всех возможных ассоциаций и сравнения их с заданным пороговым значением малоэффективен из-за больших вычислительных затрат.

Поэтому в процессе генерации ассоциативных правил широко используются методики, позволяющие уменьшить количество ассоциаций, которое требуется проанализировать. Одной из наиболее распространенных является методика, основанная на обнаружении так называемых частых наборов, когда анализируются только те ассоциации, которые встречаются достаточно часто. На этой концепции основан известный алгоритм поиска ассоциативных правил Apriori.

Алгоритм Apriori.

При практической реализации систем поиска ассоциативных правил используют различные методы, которые позволяют снизить пространство поиска до размеров, обеспечивающих приемлемые вычислительные и временные затраты, например, алгоритм Apriori.

В основе алгоритма Apriori лежит понятие частого набора, который также можно назвать частым предметным набором, часто встречающимся множеством (соответственно, он связан с понятием частоты). Под частотой понимается простое количество транзакций, в которых содержится данный предметный набор. Тогда частыми наборами будут те из них, которые встречаются чаще, чем в заданном числе транзакций.

Частый предметный набор – предметный набор с поддержкой больше заданного порога либо равной ему. Этот порог называется минимальной поддержкой.

Методика поиска ассоциативных правил с использованием частых наборов состоит из двух шагов:

1. Следует найти частые наборы.

2. На их основе необходимо сгенерировать ассоциативные правила, удовлетворяющие условиям минимальной поддержки и достоверности. Чтобы сократить пространство поиска ассоциативных правил, алгоритм Apriori использует свойство антимонотонности. Свойство утверждает, что если предметный набор Z не является частым, то добавление некоторого нового предмета A к набору Z не делает его более частым. Другими словами, если Z не является частым набором, то и набор $Z \cup A$ также не будет являться таковым, Данное полезное свойство позволяет значительно уменьшить пространство поиска ассоциативных правил.

Пусть имеется множество транзакций, представленное в табл. 2.2.

Таблица 2.2. Множество транзакций

№ транзакции	***Предметные наборы***
1	Капуста, перец, кукуруза
2	Спаржа, кабачки, кукуруза
3	Кукуруза, помидоры, фасоль, кабачки
4	Перец, кукуруза, помидоры, фасоль
5	Фасоль, спаржа, капуста
6	Кабачки, спаржа, фасоль, помидоры
7	Помидоры, кукуруза
8	Капуста, помидоры, перец
9	Кабачки, спаржа, фасоль
10	Фасоль, кукуруза
11	Перец, капуста, фасоль, кабачки
12	Спаржа, фасоль, кабачки
13	Кабачки, кукуруза, спаржа, фасоль
14	Кукуруза, перец, помидоры, фасоль, капуста

Будем считать частыми наборы, которые встречаются в D более чем $f = 4$ раза. Сначала найдем частые однопредметные наборы. Для этого представим базу данных транзакций из табл. 2.2 в нормализованном виде, который демонстрируется в табл. 2.3.

Таблица 2.3. Нормализованный вид множества транзакций

№	Спаржа	Фасоль	Капуста	Кукуруза	Перец	Кабачки	Помидоры
1	0	0	1	1	1	0	0
2	1	0	0	1	0	1	0
3	0	1	0	1	0	1	1
4	0	1	0	1	1	0	1
5	1	1	0	0	0	0	1
6	1	1	0	0	0	1	1
7	0	0	0	1	0	0	1
8	0	0	1	0	1	0	1
9	1	1	0	0	0	1	0
10	0	1	0	1	0	0	0
11	0	1	1	0	1	1	0
12	1	1	0	0	0	1	0
13	1	1	0	1	0	1	0
14	0	1	1	1	1	0	1

На пересечении строки транзакции и столбца предмета ставится 1, если данный предмет присутствует в транзакции, и 0 – в противном случае. Тогда, просуммировав значения в каждом столбце, мы получим частоту появления каждого предмета.

Поскольку все суммы равны или превышают 4, все предметы можно рассматривать как частые однопредметные наборы. Обозначим их в виде множества

$F_1 = \{\text{спаржа, фасоль, капуста, кукуруза, перец, кабачки, помидоры}\}.$

Теперь переходим к поиску частых 2-предметных наборов, Вообще, для поиска F_k то есть k-предметных наборов, алгоритм Apriori сначала создает множество F_k кандидатов в k-предметные наборы путем связывания множества F_{k-1} с самим собой. Затем F_k сокращается с использованием свойства антимонотонности. Предметные наборы множества F_k, которые остались после сокращения, формируют F_k.

Множество F_2 содержит все комбинации предметов, представленные в табл. 2.4.

Таблица 2.4. Предметные наборы

Набор	**Кол-во**	**Набор**	**Кол-во**
Спаржа, фасоль	5	Капуста, кукуруза	2
Спаржа, капуста	1	Капуста, перец	4
Спаржа, кукуруза	2	Капуста, кабачки	1
Спаржа, перец	0	Капуста, помидоры	2
Спаржа, кабачки	5	Кукуруза, перец	3
Спаржа, помидоры	1	Кукуруза, кабачки	3
Фасоль, капуста	3	Кукуруза, помидоры	4
Фасоль, кукуруза	5	Перец, кабачки	1
Фасоль, перец	3	Перец, помидоры	3
Фасоль, кабачки	6	Кабачки, помидоры	2
Фасоль, помидоры	4		

Поскольку $f = 4$, из табл. 6.4 в множество F_2 (то есть множество 2-предметных наборов) войдут только те наборы, которые встречаются в исходной выборке 4 раза или более, Таким образом:

F_2 = *{спаржа, фасоль};*
{спаржа, кабачки};
{фасоль, кукуруза};
...{фасоль, кабачки};
{фасоль, помидоры};
{капуста, перец};
{кукуруза, помидоры}.

Далее мы используем частые 2-предметные наборы из множества F_2 для генерации множества F_3 предметных наборов. Для этого нужно связать множество F_2 с самим собой, где предметные наборы являются связываемыми, если у них первые $k-1$ предметов общие (предметы должны следовать в алфавитном порядке), Например, наборы *{спаржа, фасоль}* и *{спаржа, кабачки}*, для которых $k = 2$, чтобы быть связываемыми, должны иметь $k - 1 = 1$ общий первый элемент, которым и является спаржа. В результате связывания пары 2-предметных наборов мы получим:

{спаржа, фасоль} + *{спаржа, кабачки}*=*{спаржа, фасоль, кабачки}*.

Аналогично *{фасоль, кукуруза}* и *{фасоль, кабачки}* могут быть объединены в 3-предметный набор *{фасоль, кукуруза, кабачки}*, и наконец, так

же формируются остальные 3-предметные наборы *{фасоль, кабачки, помидоры}* и *{фасоль, кукуруза, помидоры}*. Таким образом:

$F_3 =$ {спаржа, фасоль, кабачки};
{фасоль, кукуруза, кабачки};
{фасоль, кабачки, помидоры};
{фасоль, кукуруза, помидоры}.

Затем F_3 также сокращается с помощью свойства антимонотонности. Для каждого предметного набора s из множества F_3 создаются и проверяются поднаборы размером $k-1$. Если любой из этих поднаборов не является частым и, следовательно, наборы s также не могут быть частыми (в соответствии со свойством антимонотонности), то он должен быть исключен из рассмотрения. Например, пусть s = *{спаржа, фасоль, кабачки}*. Тогда поднаборы размера $k-1=2$, сгенерированные на основе набора s – *{спаржа, фасоль}* кол-во = 5, *{спаржа, кабачки}* кол-во = 5 *и {фасоль, кабачки}* кол-во = 6 (табл. 2.4).

Видно, что все эти поднаборы являются частыми, т.к. кол-во $f > 4$. Значит, и набор s = *{спаржа, фасоль, кабачки}* будет частым и сокращению не подлежит. Таким же образом можно убедиться, что и набор s = *{фасоль, кукуруза, помидоры}* является частым.

Рассмотрим набор s = *{фасоль, кукуруза, кабачки}*. Поднабор {кукуруза, кабачки} появляется всего три раза (см. табл. 2.4), поэтому не является частым. Тогда в соответствии со свойством антимонотонности и набор s = *{фасоль, кукуруза, кабачки}* не будет частым, и мы должны его отбросить.

Теперь рассмотрим набор *{фасоль, кабачки, помидоры}*. Поскольку под набор *{кабачки, помидоры}* не является частым (частота его появления – всего 2), набор *{фасоль, кабачки, помидоры}* также не является частым и вследствие этого будет исключен из рассмотрения,

Таким образом, в множество F_3 3-предметных частых наборов попадают два набора *{спаржа, фасоль, кабачки}* и *{фасоль, кукуруза, помидоры}*. Их уже нельзя связать, поэтому задача поиска частых предметных наборов на исходном множестве транзакций решена.

Генерация ассоциативных правил

После того как все частые предметные наборы, найдены, можно переходить к генерации на их основе ассоциативных правил. Для этого к каждому частому предметному набору s, нужно применить процедуру, состоящую из двух шагов.

1, Генерируются все возможные поднаборы s.

2, Если поднабор ss является непустым поднабором s, то рассматривается

ассоциация $R: ss \rightarrow (s - ss)$, где $s - ss$ представляет собой набор s без поднабора ss. R считается ассоциативным правилом, если удовлетворяет условию заданного минимума поддержки и достоверности. Данная процедура повторяется для каждого подмножества ss из s.

Рассмотрим предметные наборы – кандидаты в ассоциативные правила, содержащие два предмета в условии, например набор s = *{спаржа, фасоль, кабачки}* из множества 3-компонентных предметных наборов F_3, полученных на этапе поиска частых наборов. Соответствующими поднаборами s являются: {спаржа}, {фасоль}, {кабачки}, {спаржа, фасоль}, {спаржа, кабачки}, {фасоль, кабачки} (табл. 2.5).

Таблица 2.5. Ассоциативные правила с двумя предметами в условии

Если условие, то следствие	**Поддержка**	**Достоверность**
Если {спаржа и фасоль}, то {кабачки}	4/14 == 28,6 %	4/5==80%
Если {спаржа и кабачки}, то {фасоль}	4/14 == 28,6 %	4/5==80%
Если {фасоль и кабачки}, то {спаржа}	4/14 == 28,6 %	4/6 == 66,7 %

Для первого ассоциативного правила в табл. 2.5 предположим, что ss = *{спаржа, фасоль}*, и тогда $(s - ss)$ = *{кабачки}*.

Рассмотрим правило R: *{спаржа, фасоль}* → *{кабачки}*.

Поддержка, показывающая долю транзакций, которые содержат как условие *{спаржа, фасоль}*, так и следствие {кабачки}, в общем наборе транзакций, имеющихся в базе данных, составляет 28,6 % (4 из 14 транзакций). Чтобы найти достоверность, мы должны учесть, что набор {спаржа, фасоль} появляется в 5 из 14 транзакций, 4 из которых также содержат {кабачки}, Тогда достоверность будет 4/5 == 80 %, Аналогично определяются поддержка и достоверность для остальных правил в табл. 2.5.

Если предположить, что минимальная достоверность для правила составляет 60 %, то все ассоциации, представленные в табл. 2.5, будут правилами. Если порог установить равным 80 %, то правилам и будут считаться только первые две ассоциации.

Наконец, рассмотрим кандидатов в правила, содержащих одно условие и одно следствие, Для этого применим описанную выше методику генерации ассоциативных правил к множеству F_2 2-компонентных предметных наборов, и результаты представим в табл. 2.6.

Таблица 2.6. Ассоциативные правила с одним предметом в условии

Если условие, то следствие	**Поддержка**	**Достоверность**
Если {спаржа}, то {фасоль}	5/14 = 35,7 %	5/6 = 83,3%

Если {фасоль}, то { спаржа }	5/14 = 35,7 %	5/10 = 50%
Если { спаржа }, то {кабачки}	5/14 = 35,7 %	5/6 = 83,3 %
Если {кабачки}, то {спаржа}	5/14 = 35,7 %	5/7==71,4%
Если {фасоль}, то {кукуруза}	5/14 = 35,7 %	5/10 = 50%
Если {кукуруза}, то {фасоль}	5/14 = 35,7 %	5/8 == 62,5 %
Если {фасоль}, то {кабачки}	6/14 = 42,9 %	6/10 == 60 %
Если {кабачки}, то {фасоль}	6/14 = 42,9 %	6/7 == 85,7 %
Если {фасоль}, то {помидоры}	4/14 = 28,6 %	4/10==40%
Если {помидоры}, то {фасоль}	4/14 = 28,6 %	4/6 == 66,7 %
Если {капуста}, то {перец}	4/14 = 28,6 %	4/5 == 80 %
Если {перец}, то {капуста}	4/14 = 28,6 %	4/5 == 80 %
Если {кукуруза}, то {помидоры}	4/14 = 28,6 %	4/8 == 50 %
Если {помидоры}, то {кукуруза}	4/14 = 28,6 %	4/6 == 66,7%

Чтобы проверить значимость сгенерированных правил, обычно перемножают их значения поддержки и достоверности, что позволяет аналитику ранжировать правила в соответствии с их значимостью и достоверностью, В табл. 2.7 представлен список правил, сгенерированных на основе исходного множества транзакций (см. табл. 2.2) при заданном уровне минимальной достоверности 80 %,

Таблица 2.7. Ассоциативные правила

Если условие, то следствие	**Поддержка, S**	**Достоверность, C**	**SC**
Если {кабачки}, то {фасоль}	6/14 = 42,9 %	6/7 == 85,7 %	0,3677
Если {спаржа}, то {фасоль}	5/14 = 35,7 %	5/6 = 83,3%	0,2974
Если { спаржа }, то {кабачки}	5/14 = 35,7 %	5/6 = 83,3%	0,2974
Если {капуста}, то {перец}	4/14 = 28,6 %	4/5 == 80 %	0,2288
Если {перец}, то {капуста}	4/14 = 28,6 %	4/5 == 80 %	0,2288
Если {спаржа и фасоль}, то {кабачки}	4/14 = 28,6 %	4/5 == 80 %	0,2288
Если {спаржа и кабачки}, то {фасоль}	4/14 = 28,6 %	4/5 == 80 %	0,2288

Таким образом, в результате применения алгоритма Apriori нам удалось обнаружить 7 ассоциативных правил, с достоверностью не менее 80 % показывающих, какие продукты из исходного набора чаще всего продаются вместе. Это знание позволит разработать более совершенную маркетинговую стратегию, оптимизировать закупки и размещение товара на прилавках и витринах.