Ассоциативные правила [M.101]

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* (market basket analysis), 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* (association rules).

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 показывает опасный побочный эффект.

Базовым понятием в теории ассоциативных правил является *транзакция* — некоторое множество событий, происходящих 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 таблице 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 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

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 *условие* (antecedent) и *следствие* (consequent), записываемых в виде $X \to Y$, что читается следующим образом: «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 (left-hand side — LHS) и правосторонним (right-hand side — RHS) компонентами ассоциативного правила.

Ассоциативные правила описывают связь между наборами предметов, соответствующими 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. Зададим, что S — поддержка, C — достоверность.

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

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

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:

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

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 таблицы 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:

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

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

$$C(салат \rightarrow помидоры) = 4 / 4 = 1.$$

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* (strong rules). 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 $S_A S_B = 0{,}7 \cdot 0{,}8 = 0{,}56$, 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.

Пример

Рассмотрим проект Data Mining в области продаж, который призван объяснить связь между покупками женского белья и других товаров. Простой ассоциативный анализ (рассматривающий отдельные корзины закупок всех клиентов за год) обнаруживает с очень высокой степенью поддержки и достоверности, что женское белье ассоциировано с конфетами. Сначала это вызывает некоторое недоумение, но последующий анализ показывает, что все основные категории товаров ассоциированы с конфетами с высоким уровнем достоверности. Большинство клиентов покупают конфеты в любом случае. Ассоциации с конфетами как следствием с поддержкой $S = 0{,}87$ оказываются малоинтересны, поскольку $87\ \%$ покупателей всегда приобретают конфеты. Чтобы быть значимой, ассоциация с конфетами в качестве следствия должна иметь поддержку большую, чем $0{,}87$.

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, также встречается термин «улучшение») вычисляется следующим образом:

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

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

Рассмотрим ассоциацию помидоры → салат из таблицы 1.

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

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

Теперь рассмотрим ассоциацию помидоры → конфеты.

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

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

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: $T(A \rightarrow B) = S(A \rightarrow B) - S(A)S(B)$.

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 морковь → помидоры и салат → помидоры, которые имеют одинаковую поддержку $C = 1$, поскольку салат и морковь всегда продаются вместе с помидорами (см. таблицу 1). Лифты для данных ассоциаций также будут одинаковыми, поскольку в обеих ассоциациях поддержка следствия $S(помидоры) = 7/10 = 0{,}7$.

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

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

$S(морковь \rightarrow помидоры) = 3/10 = 0{,}3$; $S(морковь) = 0{,}3$; $S(помидоры) = 0{,}7$.

Таким образом, $T(морковь \rightarrow помидоры) = 0{,}3 - 0{,}3 \cdot 0{,}7 = 0{,}09$.

S(салат → помидоры) = 0,4; S(салат) = 0,4; S(помидоры) = 0,7.

Следовательно, $T(салат \rightarrow помидоры) = 0{,}4 - 0{,}4 \cdot 0{,}7 = 0{,}12$.

Итак, значимость второй ассоциации больше, чем первой.

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 $k \cdot 2^{k-1}$ 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 $100 \cdot 299 \approx 6{,}4 \cdot 1031$. 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*.