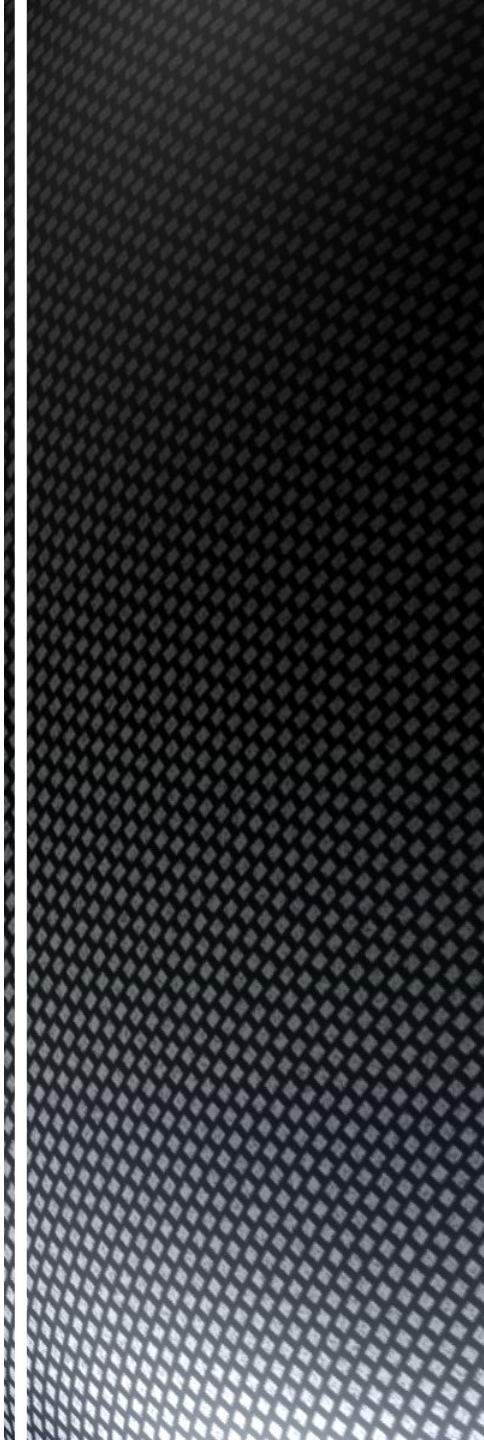
Аффинитивный анализ. Алгоритм Apriori



Определение

Аффинитивный анализ (affinity analysis) — методы исследования взаимной связи (ассоциаций) между событиями происходящими совместно и их количественная (т.е. в виде числа) оценка.

Результальт выполенния аффинитивного анализа – набор ассоциативных правил.

«affinity», в переводе означает «близость», «сходство».

Сфера применения

- Торговая сфера. Для выявление наборов товаров, которые в супермаркетах часто покупаются вместе или никогда не покупаются вместе.
- Медицина. Выявление причинно-следственных связей по возникновению побочных эффектов лекарств.
- Производственная сфера. Выявление связи между параметрами оборудования и получаемыми качественными характеристиками продукта.
- и многие другие сферы

Понятие транзакции

Ключевое понятие – транзакция – множество событий происходящих одновременно (совместно)

Например: если мы анализируем деятельность торговой площадки, то в качестве транзакции можно рассматривать отдельный чек отдельного покупателя – **совместная** покупка отдельных товаров

Тогда проанализировав множество транзакций можно определить - является ли покупка одного товара следствием или причиной покупки другого товара. (клиент, купивший молоко, с

Исходные данные – множество 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	Лук

Ассоциативные правила

Импликация (бинарная логическая связка)
$X \rightarrow Y$, где $X \subset I$, $Y \subset I$ и $X \cap Y = \varnothing$,
I – множество всех событий
T - транзакция
X – множества событий транзакции, называемых
условием (antecedent)
Y – множества событий транзакции, называемых
Следствием (consequent)

Читается правило: «Из X следует Y»

Связь между наборами предметов

$X \rightarrow Y$

Ассоциативные правила описывают связь между
наборами событий X и Y.

Связь оценивается численно с помощью набора
показателей:

Основных:

- Поддержка (support), обозначение S
- Достоверность (confidence), обозначение C

и вспомогательных:

- Лифт (lift), обозначение L
- Левередж (leverage), обозначение T

Основные показатели:

Поддержка S (support) правила $A \rightarrow B$, рассчитывается так:

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

Достоверность C (confidence) правила $A \rightarrow B$, рассчитывается так:

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

№	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	Лук

Пример расчета **поддержки** S для правила *салат→помидоры*

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

№	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	Лук

Пример расчета **достоверности** C для правила

сал

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

Проверка зависимости A от B в правиле $A \rightarrow B$

$$S(A,B) \approx S(A) \cdot S(B)$$

- Если выполняется, то A и B независимы друг от друга и правило $A \rightarrow B$ непригодно.

Пример:
Всего транзакций 100 штук.
A и B встречаются совместно в 50 транзакциях:
$S(A,B)=50/100$
A встречается в 70 транзакциях: $S(A)=70/100$
B встречается в 80 транзакциях: $S(B)=80/100$

Проверим по правилу выше:
$S(A,B) \approx S(A) \cdot S(B)$
$0,5 \approx 0,7 \cdot 0,8$
$0,5 \approx 0,56$. Наше правило выполняется, это значит, что условие A и следствие B часто встречаются вместе, не менее часто они встречаются и по отдельности. Правило

Лифт, L для правила $A \rightarrow B$ – это отношение

$$\frac{C(A \rightarrow B)}{S(B)}$$

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 пример использования лифта для меры связи в двух правилах:

1. *Помидоры →салат*
2. *Помидоры → конфеты*

№	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	Лук

S(салат) = 4/10 = 0,4; C(помидоры → салат) = 4/7 = 0,57.
Следовательно, L(помидоры → салат) = 0,57/0,4 = 1,425. **>1, хорошо**

№	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	Лук

S(конфеты) = 6/10; C(помидоры → конфеты) = 4/7 =0,57.
Тогда L(помидоры → конфеты) = 0,57/0,6 = 0,95. **<1**
плохо

Противоречие использование меры лифт

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.

Мера левередж, T для правила A→B – это разность

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

- Левередж применяется для сравнения значимости двух и более правил, у которых поддержка и достоверность одинаковые.
- Чем левередж больше, тем значимее правило.

Сравним значимость двух правил:

1. морковь → помидоры
2. салат → помидоры

И определим, какое из правил значимее (“сильней”)

№	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	Лук

C(морковь → помидоры) = 3 / 3 = 1

L(…)=1/S(помидоры)=1/(6/10)

C(салат → помидоры) =3 / 3 = 1 L(…)=1/S(помидоры)=1/(6/10)

№	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	Лук

T(морковь → помидоры) = S(м…)- S(м…)· S(п…) =0,3 – 0,3 · 0,6 =0,12

T(салат → помидоры) = S(с…) –S(с…) · S(п…) =0,4 – 0,4 · 0,6 =0,16

0,16>0,12