

Лекция 9. Ассоциативные правила и анализ корзины покупок

Тема: Алгоритм Apriori, FP-Growth, поддержка и доверие правил

1. Введение

В современном мире данные о покупках, транзакциях и пользовательском поведении играют ключевую роль в развитии бизнеса и аналитике.

Одним из важных направлений интеллектуального анализа данных (Data Mining) является **поиск ассоциативных правил** — закономерностей, описывающих совместное появление элементов в наборах данных.

Самый распространённый пример — **анализ корзины покупок (Market Basket Analysis)**, который помогает определить, какие товары покупаются вместе.

Эта техника активно используется в розничной торговле, электронной коммерции и маркетинге для:

- оптимизации выкладки товаров,
 - построения систем рекомендаций,
 - повышения объёма продаж.
-

2. Основные понятия ассоциативных правил

2.1. Транзакции и элементы

Пусть имеется множество товаров $I = \{i_1, i_2, \dots, i_m\}$

и база транзакций $D = \{T_1, T_2, \dots, T_n\}$

где каждая транзакция $T_j \subseteq I$ — это набор товаров, купленных вместе.

2.2. Ассоциативное правило

Ассоциативное правило имеет вид:

$$A \Rightarrow B$$

где A и B — подмножества множества товаров, причём $A \cap B = \emptyset$.

Интерпретация: если покупатель приобрёл товары из множества AAA, то с высокой вероятностью он купит и товары из множества BBB.

Пример:

$(\text{Молоко, Хлеб}) \Rightarrow (\text{Масло}) \setminus \text{text}\{(\text{Молоко, Хлеб})\} \rightarrow \setminus \text{text}\{(\text{Масло})\}(\text{Молоко, Хлеб}) \Rightarrow (\text{Масло})$

Это значит, что покупатели, купившие молоко и хлеб, часто покупают и масло.

3. Метрики качества правил

Чтобы оценить силу и надёжность ассоциативных правил, используются ключевые показатели: **поддержка, доверие и лифт**.

3.1. Поддержка (Support)

Показывает, насколько часто товары встречаются вместе в базе данных.

$\text{support}(A \Rightarrow B) = |\{T \in D : A \cup B \subseteq T\}| / |D|$
 $\text{support}(A \Rightarrow B) = |D| \cdot |\{T \in D : A \cup B \subseteq T\}| / |D|^2$

Пример:

если 2 из 10 покупателей купили и хлеб, и масло, то поддержка = 0.2 (20%).

3.2. Доверие (Confidence)

Оценивает вероятность покупки BBB, если уже куплены товары из AAA:

$\text{confidence}(A \Rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)}$
 $\text{confidence}(A \Rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)}$

Пример:

Если 3 человека купили хлеб, и из них 2 также купили масло, то доверие = $2/3 = 66.7\%$.

3.3. Лифт (Lift)

Показывает, насколько сильно правило отличается от случайной покупки.

$$\text{lift}(A \Rightarrow B) = \frac{\text{confidence}(A \Rightarrow B) \cdot \text{support}(B)}{\text{support}(A)} \quad \text{lift}(A \Rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A) \cdot \text{support}(B)}$$

- Если **лифт** > 1 $\rightarrow$ между товарами есть положительная зависимость.
- Если **лифт** $= 1$ $\rightarrow$ товары независимы.
- Если **лифт** < 1 $\rightarrow$ товары взаимно исключают друг друга.

4. Алгоритм Apriori

4.1. Основная идея

Apriori — один из первых и наиболее известных алгоритмов для поиска ассоциативных правил.

Он основан на принципе:

Если множество является частым, то все его подмножества также часты.

Это свойство позволяет эффективно сокращать пространство поиска.

4.2. Основные шаги алгоритма

1. **Вычислить поддержку** для всех товаров и выбрать частые (частота $\geq \text{min_support}$).
2. **Сгенерировать кандидатов** на 2-элементные наборы, состоящие из частых элементов.
3. **Отфильтровать** наборы, чья поддержка ниже порога.
4. Повторять шаги, увеличивая размер наборов, пока не останется частых.
5. **Формировать правила** $A \Rightarrow B$ и вычислять их доверие.
6. **Отбирать лучшие правила**, удовлетворяющие порогам по поддержке и доверию.

4.3. Пример

Транзакция	Содержимое
T1	Молоко, Хлеб, Масло

Транзакция	Содержимое
T2	Хлеб, Масло
T3	Молоко, Хлеб
T4	Молоко, Масло
T5	Хлеб, Масло

Если задать минимальную поддержку 0.4, частыми будут пары: (Хлеб, Масло), (Молоко, Хлеб), (Молоко, Масло).

4.4. Достоинства и недостатки

Преимущества:

- Простота и прозрачность.
- Хорошо интерпретируемые результаты.

Недостатки:

- Высокая вычислительная сложность при большом количестве товаров.
 - Много проходов по базе данных.
 - Возможна генерация огромного числа кандидатов.
-

5. Алгоритм FP-Growth (Frequent Pattern Growth)

5.1. Идея метода

FP-Growth — более эффективная альтернатива Apriori. Он устраняет необходимость в явной генерации кандидатов, используя структуру FP-дерева (Frequent Pattern Tree).

5.2. Этапы алгоритма

1. **Первичный проход по данным:** вычисляется частота всех элементов и отбрасываются редкие.
2. **Построение FP-дерева:**
 - Каждая транзакция добавляется в дерево в порядке убывания частоты элементов.
 - Одинаковые пути объединяются.
3. **Извлечение частых наборов:**
 - Для каждого элемента строится условное дерево.

- Рекурсивно выделяются частые комбинации без генерации всех кандидатов.

5.3. Преимущества FP-Growth

- Быстрее Apriori на больших наборах данных.
- Требуется меньше проходов по базе (обычно два).
- Не создаёт огромное число промежуточных наборов.

Недостаток:

Сложность реализации и необходимость хранения FP-дерева в памяти.

6. Интерпретация ассоциативных правил

Ассоциативные правила ценны не только для маркетинга, но и для стратегических решений.

Их можно использовать для:

- кросс-продаж (например, «Купивший смартфон часто покупает чехол»),
- планирования складских запасов,
- персонализированных рекомендаций,
- анализа поведения клиентов.

Важно учитывать не только силу правил, но и их бизнес-смысл.

Иногда высокая поддержка не означает полезность — например, правило, что «все покупатели покупают хлеб», не помогает в принятии решений.

7. Применение в разных областях

Область	Пример применения
Розничная торговля	Определение сопутствующих товаров для акций
Интернет-маркетинг	Персональные рекомендации товаров
Финансы	Анализ совокупных транзакций клиентов
Медицина	Поиск взаимосвязей между симптомами и диагнозами
Кибербезопасность	Выявление типичных комбинаций событий атак

8. Визуализация правил

Для анализа и интерпретации ассоциативных правил часто используются:

- **Диаграммы поддержки-доверия,**
 - **Графы правил** (где узлы — товары, а рёбра — ассоциации),
 - **Таблицы лифта,**
 - **Интерактивные дашборды** (например, в Tableau, Power BI или Orange Data Mining).
-

9. Заключение

Ассоциативные правила — мощный инструмент анализа данных, позволяющий находить скрытые взаимосвязи между объектами и событиями. Алгоритмы **Apriori** и **FP-Growth** являются фундаментальными методами в этой области и лежат в основе многих современных рекомендательных систем.

- **Apriori** эффективен для небольших и средних наборов данных, прост в реализации.
- **FP-Growth** подходит для анализа больших объёмов транзакций, обеспечивая более высокую производительность.

Комбинированное использование метрик — поддержки, доверия и лифта — позволяет не только обнаружить закономерности, но и оценить их значимость и ценность для бизнеса.

Список литературы

1. Han, J., Kamber, M., Pei, J. *Data Mining: Concepts and Techniques*. — Morgan Kaufmann, 2012.
2. Agrawal, R., Imieliński, T., Swami, A. *Mining Association Rules Between Sets of Items in Large Databases*. — ACM SIGMOD, 1993.
3. Borgelt, C. *Efficient Implementations of Apriori and Eclat*. — Workshop on Open Source Data Mining, 2003.
4. Tan, P.-N., Steinbach, M., Kumar, V. *Introduction to Data Mining*. — Pearson, 2019.
5. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. — O'Reilly, 2022.