

АЛГОРИТМ «АПРИОРИ» КАК ИНСТРУМЕНТ ДЛЯ ПОИСКА АССОЦИАТИВНЫХ ПРАВИЛ

М. Е. Артюх

*Гродненский государственный университет имени Янки Купалы, Гродно
Научный руководитель: Н. В. Марковская, кандидат физ.-мат. наук, доцент*

История создания алгоритма. Алгоритм «Apriori» был предложен Реймондом Агривале, Томасом Импеллицерри и Аруном Свами в статье, опубликованной в 1994 году. Эта статья, которая носила название "Извлечение ассоциативных правил между наборами элементов в больших базах данных" ("Mining Association Rules Between Sets of Items in Large Databases"), описывает основные идеи и методы алгоритма «Apriori». Идея алгоритма «Apriori» основана на использовании свойства антимонотонности функции поддержки. Это свойство гласит, что если набор элементов встречается в базе данных с частотой, превышающей заданный порог, то все его подмножества также встречаются с этой частотой. Это позволяет сократить пространство поиска ассоциативных правил, начиная с более частых элементов и постепенно увеличивая длину наборов. Алгоритм «Apriori» стал одним из первых и наиболее широко использованных методов для поиска ассоциативных правил в больших базах данных. Его простота и эффективность сделали его популярным инструментом в областях анализа данных, добычи знаний и машинного обучения [1-3].

Использование алгоритма. Ассоциативные правила описывают связи между различными элементами набора данных, выраженные в терминах "если-то", "то-то". Например, в магазине продуктов питания ассоциативное правило может выглядеть так: "Если покупатель покупает хлеб, то с высокой вероятностью он также купит молоко".

Основные понятия:

- Подмножество (itemset): это набор элементов, например, набор товаров, купленных вместе.
- Поддержка (support): это частота появления определенного подмножества в базе данных. Например, поддержка 3% для набора {хлеб, молоко} означает, что эти два товара вместе в 3% всех произведенных транзакций. Математическая формула для поддержки:

$$\text{поддержка (x)} = \frac{\text{число транзакций, содержащих x}}{\text{общее число транзакций в базе данных}}$$

Поддержка является важным показателем при использовании алгоритма «Apriori», поскольку позволяет выявлять наиболее часто встречающиеся наборы элементов в базе данных. При поиске ассоциативных правил в базе данных заданный порог поддержки определяет, какие наборы считаются значимыми. Набор, чья поддержка превышает заданный порог, считается "часто встречающимся" и используется для дальнейшего поиска ассоциативных правил. На практике, порог поддержки выбирается пользователем в зависимости от конкретной задачи и характеристик данных. Более высокий порог поддержки приводит

к выделению более общих и широких ассоциаций, в то время как более низкий порог позволяет выявить более узкие и специфичные ассоциации.

• Доверие: это вероятность того, что если в транзакции присутствует элемент {хлеб}, то в транзакции также будет присутствовать элемент {молоко}. Математическая формула для доверия:

$$\text{доверие}(x \rightarrow y) = \frac{\text{поддержка}(x \cup y)}{\text{поддержка}(x)}.$$

Высокое доверие означает, что правило часто верно, что делает его более интересным и полезным с точки зрения анализа данных.

• Лифт: лифт используют для оценки взаимосвязи между двумя элементами в наборе данных, основываясь на их частотах появления вместе и независимо друг от друга. Простыми словами, "лифт" позволяет определить, насколько вероятно, что два элемента будут выбраны вместе, учитывая их появление вместе в сравнении с тем, сколько раз каждый из них выбирается независимо. Математическая формула для «лифта»:

$$x \rightarrow y = \frac{\text{поддержка}(x \cup y)}{\text{поддержка}(x) \times \text{поддержка}(y)}.$$

Если:

а) лифт > 1 , это означает, что события x и y чаще происходят вместе, чем если бы они были независимыми.

б) лифт < 1 , это означает, что события x и y реже происходят вместе, чем ожидалось при независимости.

с) лифт $= 1$, это означает, что вероятность события x и y вместе такая же, если бы они были полностью независимыми друг от друга.

• Уверенность: это мера того, насколько часто определенное правило ассоциации верно в данных. Простыми словами, уверенность говорит о том, насколько вероятно, что если клиент выбрал один набор, то он выберет и другой. Математическая формула для «уверенности»:

$$\text{уверенность} = \frac{\text{поддержка}(x \cup y)}{\text{поддержка}(x)}.$$

Таким образом, уверенность измеряет, насколько часто правило соблюдается, основываясь на данных, и это важный показатель для определения полезности правил ассоциаций в алгоритме Априори.

Теоретический пример: Рассмотрим пример использования алгоритма «Априори» в контексте анализа покупок в магазине. Предположим, у нас есть база данных транзакций, где каждая транзакция представляет собой список продуктов, купленных одним покупателем.

Список транзакций из базы данных:

Транзакции
Хлеб, молоко, яйца
Хлеб, сок
Молоко, яйца, сок
Хлеб, молоко
Хлеб, яйца

Шаг 1. Определение поддержки

Устанавливаем порог поддержки на 2 транзакции:

- a) Хлеб: 4 транзакции содержат хлеб (поддержка = 4).
- b) Молоко: 3 транзакции содержат молоко (поддержка = 3).
- c) Яйца: 3 транзакции содержат яйца (поддержка = 3).
- d) Сок: 2 транзакции содержат сок (поддержка = 2).

Шаг 2. Генерация 2-х элементных наборов

Генерируем комбинации 2-х элементов из товаров с учетом порога поддержки:

- a) Хлеб, молоко: 2 транзакции содержат хлеб и молоко (поддержка = 2).
- b) Хлеб, яйца: 2 транзакции содержат хлеб и яйца (поддержка = 2).
- c) Молоко, яйца: 2 транзакции содержат молоко и яйца (поддержка = 2).

Шаг 3. Расчет лифта и выбор ассоциативных правил.

Рассчитываем лифт для каждой пары товаров и выбираем те, у которых лифт больше заданного порога, предположим, мы установили порог лифта = 1.

- a) Правило: Хлеб → Молоко (лифт = $\frac{2}{\frac{4}{5} * \frac{3}{5}} \approx 1,67$).
- b) Правило: Хлеб → Яйца (лифт = $\frac{2}{\frac{4}{5} * \frac{3}{5}} \approx 1,67$).
- c) Правило: Молоко → Яйца (лифт = $\frac{2}{\frac{3}{5} * \frac{3}{5}} \approx 2,22$).

Итог: таким образом, мы выявили несколько ассоциативных правил. Например, правило "Молоко → Яйца" имеет лифт более 1, что указывает на положительную ассоциацию. Это может быть использовано магазином для тактик маркетинга, таких как совместные предложения или размещение товаров рядом друг с другом на полках для стимулирования совместных покупок.

Вывод: использование алгоритма «Априори» позволяет извлекать ценные знания из данных, выявляя закономерности и связи между различными элементами.

Список литературы

1. Кот, К.А. Аффинитивный анализ данных потребительской корзины с помощью алгоритма Apriori / К.А. Кот, Н.В. Марковская // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня: сб. научных статей VIII Международ. науч.-практ. конф. (Республика Беларусь, Минск, 11-12 мая 2022 года): / редкол.: В.А. Богуш [и др.]. – Минск: Бестпринт, 2022. – С. 234-239.
2. Марковская, Н. В. Интеллектуальный анализ данных: выделение ассоциаций с помощью алгоритма априори / Н. В. Марковская, А. Э. Городок // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : в 3 ч. Ч.

2 : сб. материалов VI Междунар. науч.-практ. конф., Минск, 20-21 мая 2020 г. / ред. кол.: В. А. Богуш. – Минск : Бестпринт, 2020. – С. 350-361.

3. Залога, А. Ю. Аффинитивный анализ данных. Поиск ассоциативных правил / А. Ю. Залога, Н. В. Марковская // BIG DATA Advanced Analytics = BIG DATA и анализ высокого уровня : в 2 ч. Ч. 2 : сб. материалов V Междунар. науч.-практ. конф., Минск, 13-14 марта 2019 г. – Минск : БГУИР, 2019. – С. 20-26.

УДК 004.89

ОПРЕДЕЛЕНИЕ КОЛИЧЕСТВА БАКТЕРИЙ *BACILLUS THURINGIENSIS* НА ИЗОБРАЖЕНИЯХ ПОЛУЧЕННЫХ С ПОМОЩЬЮ ЦИФРОВОГО МИКРОСКОПА

Ю. С. Башаримов

*Гомельский государственный технический университет, Гомель, Беларусь
Научный руководитель: К. С. Курочка, канд. техн. наук, доцент*

Введение.

Bacillus thuringiensis var. *Kurstaki* – это бактерия, которая вырабатывает токсин, убивающий некоторых насекомых, особенно гусениц бабочек и молей. [1] Она используется как биологический пестицид для борьбы с вредителями на сельскохозяйственных культурах и лесах, который нуждается в количественной оценке.

Качество обработки напрямую связано с количеством бактерий, так как оно влияет на эффективность и безопасность биологического пестицида, а также на срок его хранения и применения. Не корректная концентрация бактерий для обработки может быть опасной для жизни и здоровья людей и животных, а также для окружающей среды.

Если концентрация бактерий слишком низкая, то она может не обеспечить достаточную эффективность биологического пестицида и не уничтожить вредителей, которые могут повредить урожай или распространить инфекции. Если концентрация бактерий слишком высокая, то она может привести к избыточному загрязнению почвы и воды, а также к риску интоксикации или аллергических реакций у людей и животных, которые контактируют с обработанными растениями или продуктами.

Поэтому важно соблюдать рекомендованные дозы и способы применения бактерий для обработки, а также контролировать их концентрацию с помощью специальных методов и оборудования.

Способы подсчёта бактерий. Существует несколько способов подсчёта бактерий, в зависимости от цели, типа и концентрации бактерий, квалификации персонала, а также доступности оборудования и реактивов. Ниже приведены наиболее распространённые способы подсчёта бактерий.