

Эффективный алгоритм построение базиса ассоциативных правил

Бардашова Нина Викторовна

Санкт-Петербургский государственный университет
Математико механический факультет
Кафедра системного программирования
Научный руководитель - Панин Александр Андреевич.

15 мая 2008 г.

Data Mining и ассоциативные правила

Data mining

Процесс обнаружение в "сырых" данных нетривиальных интерпретация знаний. Решает задачи поиска закономерностей. Классификация, ассоциация, прогнозирование, визуализация, анализ связей и т.д.

ассоциативные правила

закономерности между связанными событиями в базах данных. Первое применение - задача оценки рыночной корзины.

- **Транзакция** множество событий, которые произошли одновременно.
- **Набор** множество элементов, событий.
- **Поддержка набора** Количество или процент транзакций, содержащих данный набор

Если В транзакции встретился набор элементов А, значит можно сделать вывод о том, что встретится набор элементов В, причем $A \cap B = \{\}$

Достоверность правила показывает, с какой вероятностью из события А следует событие В.

- **Транзакция** множество событий, которые произошли одновременно.
- **Набор** множество элементов, событий.
- **Поддержка набора** Количество или процент транзакций, содержащих данный набор

Если В транзакции встретился набор элементов А, значит можно сделать вывод о том, что встретится набор элементов В, причем $A \cap B = \{\}$

Достоверность правила показывает, с какой вероятностью из события А следует событие В.

Задача поиска ассоциативных правил

для заданной базы данных Найти множество ассоциативных правил с поддержкой и достоверностью выше заданной.

Задача поиска ассоциативных правил состоит из 2х подзадач:

Нахождение всех наборов элементов, которые удовлетворяют порогу минимальной поддержки.

Генерация правил из наборов элементов, найденных согласно первому пункту, с достоверностью, удовлетворяющей заданному порогу.

Задача поиска ассоциативных правил состоит из 2х подзадач:

Нахождение всех наборов элементов, которые удовлетворяют порогу минимальной поддержки.

Алгоритмическая сложность при нахождении часто встречающихся наборов. С ростом числе элементов количество потенциальных наборов растет экспоненциально.

Генерация правил из наборов элементов, найденных согласно первому пункту, с достоверностью, удовлетворяющей заданному порогу.

Задача поиска ассоциативных правил состоит из 2х подзадач:

Нахождение всех наборов элементов, которые удовлетворяют порогу минимальной поддержки.

Алгоритмическая сложность при нахождении часто встречающихся наборов. С ростом числе элементов количество потенциальных наборов растет экспоненциально.

Генерация правил из наборов элементов, найденных согласно первому пункту, с достоверностью, удовлетворяющей заданному порогу.

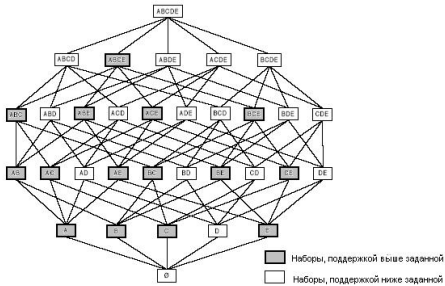
Зачастую количество полученных ассоциативных правил невероятно велико. Сложность анализирование полученных результатов и какое-либо прогнозирование.

Поставленные задачи

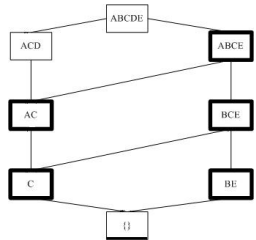
- Разработать алгоритм, который позволяет:
 - 1 Повысить эффективность алгоритма, путём сокращения множества искомых наборов, с поддержкой выше заданной
 - 2 Иметь возможность задавать дополнительные ограничения, шаблоны для искомых правил
 - 3 Без потери данных уменьшить множество получаемых ассоциативных правил. По сути - построить их базис.
- Реализовать прототип алгоритма
- Провести эксперименты, сравнить с существующими подходами

Поиск наборов

Решётка поиска Алгоритма Apriori



Решётка поиска с использованием подхода замкнутых множеств.



построение базиса

Базис ассоциативных правил

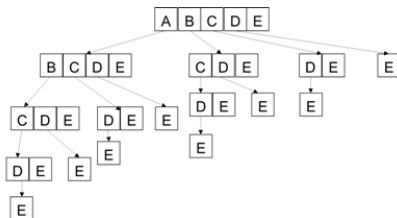
$$B = \{(r, \text{support}(r), \text{confidence}(r)) \mid r = l_1 \Rightarrow l_2, \\ l_1, l_2 \in FC, \\ l_1 \prec l_2, \text{confidence}(r) \geq \text{minconfidence}, \\ \text{support}(l_2) \geq \text{minsupport})\}$$

где FC - множество замкнутых наборов с поддержкой выше заданной.

Все возможные ассоциативные правила, их поддержки и достоверность могут быть получены (выведены) из базиса. Причем данный базис является минимальным.

структура данных

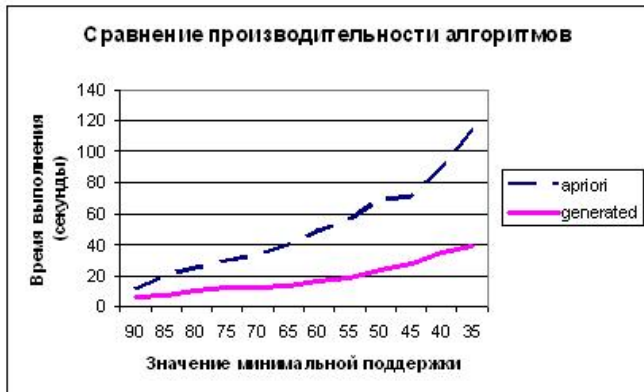
Префиксное дерево - используемая структура данных.



- Упрощается генерация нового поколения рассматриваемых наборов
- Упрощается задача поиска множества подмножеств. Её реализация сводится к задаче нахождения поддерев.

эксперименты

Время работы



Количество полученных ассоциативных правил



Заключение

Результаты работы

- разработан эффективный алгоритм построения базиса ассоциативных правил
- реализован прототип
- проведены эксперименты

Применение разработанного алгоритма

Эксперименты доказали преимущества разработанного алгоритма, и в ближайшее время будет применён вместо алгоритма Apriori в коммерческом приложении, в разработке которого автор принимает участие.