УДК 004.9

РАЗРАБОТКА ВЕРОЯТНОСТНОГО АЛГОРИТМА ПОИСКА АССОЦИАТИВНЫХ ПРАВИЛ В СТАТИСТИКЕ ПОКУПОК

Максимова К. И.,
научный руководитель канд. физ.-мат. наук, доцент Баранова И. В.
Сибирский Федеральный Университет,
Институт математики и фундаментальной информатики

1 Введение

Интеллектуальный анализ данных (Data Mining) —мультидисциплинарная область, возникшая и развивающаяся на базе прикладной статистики, искусственного интеллекта, теории баз данных и др. Оригинальное англоязычное название Data Mining было предложено Григорием Пиатецким-Шапиро в 1989 году. Название происходит от двух понятий: поиска ценной информации в большой базе данных (Data) и добычи горной руды (Mining). Термин переводится как «добыча» или «раскопка» данных.

Интеллектуальный анализ данных – это процесс обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining представляет собой технологию, предназначенную для поиска в больших объемах данных неочевидных и полученных на практике закономерностей.

К методам интеллектуального анализа данных относятся всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, нечеткой логики. Также методами Data Mining являются статистические методы: дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ и анализ временных рядов.

Одним из самых востребованных методов интеллектуального анализа данных является метод поиска ассоциативных правил. Данный метод предназначен для выявления взаимосвязей между наборами данных из статистики. Поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно. Наиболее известным алгоритмом решения задачи поиска ассоциативных правил является алгоритм ***Apriori***.

Впервые задача поиска ассоциативных правил (association rule mining) была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis).

В работе приводится подробное описание алгоритма Apriori. Предлагается вероятностный аналог данного метода. Также в работе решается практический пример нахождения ассоциативных правил на основе реальной статистики покупок товаров.

2 Поиск ассоциативных правил в интеллектуальном анализе данных

Рассмотрим алгоритмы работы метода ассоциативных правил в интеллектуальном анализе данных. Приведем основные понятия, связанные с данным методом.

Транзакция – это множество событий, которые произошли одновременно.
Транзакционная (операционная) база данных представляет собой двумерную таблицу, состоящую из номера транзакции (TID) и перечня событий, происходящих во время этой транзакции.
Поддержка – количество или процент транзакций, содержащих определенный набор данных.

2.1 Алгоритм Apriori

Приведем обозначения, используемые в алгоритме:
L_k − множество k-элементных наборов, чья поддержка не меньше заданной пользователем.
C_k − множество потенциально частых k-элементных наборов.

Алгоритм поиска ассоциативных правил Apriori имеет следующий вид:
1. Присвоить k = 1 и выполнить отбор всех 1-элементных наборов, у которых поддержка больше минимально заданной пользователем Suppmin.
2. k = k + 1.
3. Если не удается создавать k-элементные наборы, то завершить алгоритм, иначе выполнить следующий шаг.
4. Создать множество k-элементных наборов кандидатов из частых наборов. Для этого необходимо объединить в k-элементные кандидаты (k-1)-элементные частые наборы. Каждый кандидат будет формироваться путем добавления к (k-1) - элементному частому набору - p элемента из другого (k-1) − элементного частого набора - q. Причем добавляется последний элемент набора q, который по порядку выше, чем последний элемент набора p. При этом все k-2 элемента обоих наборов одинаковы.
5. Для каждой транзакции T из множества D выбрать кандидатов C_t из множества C_k, присутствующих в транзакции T. Для каждого набора из построенного множества C_k удалить набор, если хотя бы одно из его (k-1) подмножеств не является часто встречающимся т.е. отсутствует во множестве L_{k-1}.
6. Для каждого кандидата из C_k увеличить значение поддержки на единицу.
7. Выбрать только кандидатов L_k из множества C_k, у которых значение поддержки больше заданной пользователем Suppmin. Вернуться к шагу 2.
Результатом работы алгоритма является объединение всех множеств L_k для всех k.
Поиск закономерностей в классическом методе ассоциативных правил осуществляется между несколькими событиями, которые происходят одновременно.

В работе предлагается новый алгоритм поиска ассоциативных правил, являющийся вероятностным аналогом алгоритма Apriori.

3 Вероятностный аналог алгоритма поиска ассоциативных правил Apriori

Приведем некоторые основные понятия теории вероятностей и математической статистики, необходимые для понимания алгоритма.

Вероятностным пространством называется тройка (Ω,F,P), где Ω – *пространство элементарных событий*, F – *алгебра событий* и P – *вероятность*, определенная на элементах множества X.

Медиана является характеристикой распределения значений случайного события X. Медиана представляет собой такое число m, что X принимает с вероятностью 0.5 как значения больше m, так и меньше m.

Мода — значение во множестве наблюдений, которое встречается наиболее часто.

Квантиль в математической статистике — значение, которое случайное событие не превышает с фиксированной вероятностью.

Перечислим обозначения, используемые в алгоритме:

Конечное множество событий $X = \{x_1, x_2, \ldots, x_n\}$, выбранных из алгебры вероятностного пространства и состоящее из n=|X| событий, называется ***множеством случайных событий***.

Под некоторым событием x_i понимается покупка некоторого товара *i*. Всего рассматривается n различных товаров.

D — множество транзакций D ={T1…Tm}.

Случайное множество событий под X определяется как измеримое отображение $K : (\Omega, F, P) \to \left(2^X, 2^{2^X}\right)$, где 2^X – множество всех подмножеств множества X.

supmin— минимальное значение P(X) - вероятности подмножества X.

L_k — множество k-элементных наборов, чья вероятность $P(L_k) > supmin$.

C_k — множество наборов, мощность которых равна k.

Перед работой алгоритма необходимо найти вероятности каждого подмножества множества $X \subseteq X$, т.е. вероятности покупки каждого набора. Они оцениваются из статистики покупок D.

Алгоритм поиска ассоциативных правил имеет следующий вид:

1. Присвоить k = 1 и выполнить отбор $X \leq \chi$, где |X| = 1, для которых P(X) > supmin, т.е. L_1.

2. Пользователь задает новое значение supmin. В качестве выбранного значения можно выбирать значение, соответствующее:

- медиане множества (т.е. supmin=0.5),
- моде (т.е. p_{max} − вероятность самого часто встречающегося набора),
- квантилю третьего порядка.(т.е. supmin=0.7)

Присвоить k = k + 1.

3. Если L_k пусто, то завершить алгоритм, иначе выполнить следующий шаг.

4. Пусть $X = \{x_1, x_2, \ldots x_{k-1}\} \leq \chi$, где |X| = k, $Y = \{y_1, y_2, \ldots y_{k-1}\} \leq \chi$, где |y| = 1. Необходимо создать $C_k \leq L_k$. При этом X=Y, при |X|=|Y|=k-2.

5. Для любого T≤D выбрать кандидатов $C_t \leq C_k$. Удалить набор C_t, если любое C_t не принадлежит L_{k-1},где $|C_t|=k-1$.

6. Выбрать $L_k \leq C_k$ у которых значение вероятности $P(L_k) > supmin$. Вернуться к шагу 2. Результатом работы алгоритма является объединение всех множеств L_k, для любого k.

Практический пример нахождения ассоциативных правил

Решим практическую задачу нахождения ассоциативных правил в статистике покупок продуктов. Одним из продуктовых розничных магазинов города Красноярска была предоставлена статистика покупок продуктов за октябрь 2012 года и февраль 2013 года. Всего статистика содержала около 1000 транзакций (чеков покупок). В статистике встречалось 50 различных наименований товаров: молоко, хлеб, сметана, творог, сыр, яйца, колбаса, сахар, различные виды мяса, полуфабрикаты и т.д. На основе статистики было необходимо найти закономерности между покупками в наборе данных.

Приведенные выше товары были обозначены переменными, в Таблице 1 приведены 13 самых часто встречающихся товаров.

Таблица 1.

Список товаров из статистики покупок

Наименование	обозначение	поддержка
молоко	a	100
сметана	b	31
сыр твердый	c	44
мясо куриное	d	26
яйца куриные	e	26
полуфабрикат	f	31
хлею	g	125
конфеты	h	33
газировка	i	27
овощи	j	34
фрукты	k	52
консервы	l	28
соусы	m	31
крупа	n	32

Был определен уровень поддержки для данной задачи Suppmin=25. На первом этапе происходило формирование одноэлементных кандидатов. Согласно алгоритму были подсчитаны поддержки одноэлементных наборов. Наборы с уровнем поддержки меньше установленного (25) отсекались. На следующем этапе выполнялось формирование двухэлементных кандидатов, подсчет их поддержки и отсечение неподходящих наборов. В результате данного процесса остались шесть двухэлементных наборов, принимающих участие в дальнейшей работе алгоритма. Это пары ac, ag, ak, bg, cg, gk.

Далее формировались трехэлементные наборы товаров. Были получены два набора с поддержкой более 25: acg и agk (с поддержкой 25 и 26, соответственно).

Формирование четырехэлементных наборов не удалось, поэтому работа алгоритма была прекращена. Итак, в результате работы алгоритма Apriori было получено два ассоциативных правила: «При покупке молока и сыра, покупатель, скорее всего, купит хлеб» и «При покупке молока и хлеба, покупатель, скорее всего, купит фрукты».

Теперь рассмотрим работу вероятностного алгоритма поиска ассоциативных правил на той же самой статистике покупок товаров.

Для полученных наборов подсчитаем условные вероятности по формулам. В результате получим следующие правила:

Молоко->сметана, сыр, мясо кур(100%)
Молоко->хлеб, сметана, сыр(100%)
Молоко->сметана, сыр, консервы. (100%)
Сыр -> сметана, молоко, соусы (82%)
Хлеб ->молоко, сметана, сыр(80%)
Сыр -> сметана, молоко, хлеб(80%)
Сметана ->молоко, сыр, мясо кур(74%)
Сметана ->молоко, сыр, соусы. (61%)

Полученные правила отсортированы по значению условной вероятности. Значения вероятности будут лежать на отрезке [0,1] (эквивалент в %: [0,100] и будут означать вероятность того, что при покупке набора товаров Xi также приобретут набор товаров Xj. Таким образом, получено множество ассоциативных правил разного вида. Такие результаты позволяют решать более широкий класс задач.