УДК 681.3.06

А. Ю. Краковецкий, аспирант

МЕТОД ПОИСКА АССОЦИАТИВНЫХ ПРАВИЛ НА ОСНОВЕ СИЛЬНЫХ НАБОРОВ ДАННЫХ И FP-ДЕРЕВА

В роботе предлагается концепция сильных наборов данных, решающих проблему генерации большого количества кандидатов при решении задачи поиска знаний в виде ассоциативных правил, метод поиска непересекающихся сильных наборов данных, а также метод поиска ассоциативных правил на основе сильных наборов данных.

Ключевые слова: *ассоциативные правила, FP-дерево, Data Mining.*

Актуальность

Одной из проблем поиска знаний в виде ассоциативных правил является генерация кандидатов, количество которых может быть очень большим [1, 2]. Для сокращения их количества в разных методах используют разные подходы, среди которых правило антимонотонности [2, 3], экспертные знания, определенные предположения и т. д. Одним из методов сокращения количества кандидатов является построение дерева транзакций (FP-tree) [4]. Этот метод позволяет находить наборы данных, которые часто повторяются, однако он имеет ряд недостатков, среди которых неучитывание значения достоверности набора, отсутствие интуитивно понятной логической связи между элементами в наборах. Это приводит к тому, что большое количество правил, генерирующихся на основе данных наборов данных, отбрасываются на стадии проверки. Таким образом, решение проблемы генерации избыточных наборов данных является актуальной задачей.

В этой работе предлагается концепция *сильных* наборов данных, решающих поставленную проблему, метод поиска непересекающихся сильных наборов данных, а также метод поиска ассоциативных правил на основе сильных наборов данных.

Постановка задачи

Постановка задачи поиска ассоциативных правил: необходимо найти множество наборов данных, поддержка которых больше чем заданное значение минимальной поддержки $Supp_{\min}$, минимальной достоверности $Conf_{\min}$ и с улучшением больше единицы [1, 2]:

$$L = \{F \mid Supp(F) > Supp_{\min}, Conf(F) > Conf_{\min}, impr(F) > 1\}.$$

Понятие сильных наборов данных

Под *набором данных* X будем понимать непустое подмножество элементов общего множества элементов A:

$$X \neq \varnothing,\ X \subset A, \tag{1}$$

LSI ($left-side\ itemset$) – непустой набор данных, формирующий левую часть ассоциативного правила, соответственно RSI ($right-side\ itemset$) – непустой набор данных, который формирующий правую часть:

$$LSI \Rightarrow RSI,\ LSI \neq \varnothing,\ RSI \neq \varnothing,\ LSI \cup RSI = \varnothing. \tag{2}$$

Пусть $LSI = X$ и $RSI = Y$, $0 \le \sigma, \tau \le 100$, где σ, τ – соответственно заданные минимальная поддержка и достоверность [1, 2]. Тогда ассоциативное правило $X \Rightarrow Y$ будем называть *допустимым* для заданных σ и τ, если выполняются следующие условия:

$$\begin{aligned} &Support(X) \geq \sigma, \\ &Support(X \cup Y) \geq \sigma, \\ &\frac{Support(X \cup Y)}{Support(X)} \geq \tau. \end{aligned} \qquad (3)$$

Если любые наборы данных X и Y множества A ($X \cup Y = A, X \cap Y = \varnothing, |A| > 2$) образуют допустимые ассоциативные правила $X \Rightarrow Y$ для заданных σ и τ, то множество A будем называть *сильным набором данных*.

Рассмотрим это на примере. Пусть $A = \{a, b, c\}$ – множество, которое складывается с трех элементов. Множество всех возможных вариантов разделения A на подмножества, которые представляют собой шаблоны ассоциативных правил, состоит из 12 элементов (таблица 1).

Таблица 1

Наборы данных и соответствующие правила для множества $A = \{a, b, c\}$

Наборы данных	Правило	Наборы данных	Правило
$\{a\}, \{b\}$	$a \Rightarrow b$	$\{a\}, \{b, c\}$	$a \Rightarrow b, c$
$\{b\}, \{a\}$	$b \Rightarrow a$	$\{b, c\}, \{a\}$	$b, c \Rightarrow a$
$\{a\}, \{c\}$	$a \Rightarrow c$	$\{a, b\}, \{c\}$	$a, b \Rightarrow c$
$\{c\}, \{a\}$	$c \Rightarrow a$	$\{c\}, \{a, b\}$	$c \Rightarrow a, b$
$\{b\}, \{c\}$	$b \Rightarrow c$	$\{a, c\}, \{b\}$	$a, c \Rightarrow b$
$\{c\}, \{b\}$	$c \Rightarrow b$	$\{b\}, \{a, c\}$	$b \Rightarrow a, c$

Если *все* правила, которые образуют данные наборы, являются допустимыми, то множество A является сильным.

Количество возможных вариантов выбора m элементов с N, которые будут составлять левую часть ассоциативного правила, равняется C_N^m. Количество возможных вариантов выбора n элементов из тех, которые остались, то есть $N - m$, равняется C_{N-m}^n. Поскольку и левая, и правая части должны присутствовать в правиле, то общее количество комбинаций будет $C_N^m C_{N-m}^m$. В нашем случае m меняется в пределах от 1 до $N-1$, а n – от 1 до $N - m$. Пусть $m = |LSI|$ и $n = |RSI|$ – количество элементов соответственно в левой и правой частях ассоциативного правила, $N = m + n$ – общее количество элементов в правиле. Тогда общее количество правил, которые можно образовать из сильного множества $A = LSI \cup RSI$, состоящего из N элементов, определяется с помощью выражения:

$$\eta = \sum_{m=1}^{N-1} \sum_{n=1}^{N-m} C_N^m C_{N-m}^m . \qquad (4)$$

Вывод: сильные множества не приводят к потере информативности данных и могут компактно представить большое количество ассоциативных правил.

Необходимое и достаточное условие, при котором множество является сильным. Пусть имеем множество $A = \{a_1, a_2, ..., a_n\}$, $n \geq 2$, где $s_1, s_2, ..., s_n$ – поддержка элементов $a_1, a_2, ..., a_n$ соответственно, s_0 – поддержка множества A, σ и τ – соответственно минимальная поддержка и минимальная достоверность. Тогда A является сильным множеством, если выполняются условия:

$$mn \geq \sigma \text{ и } \frac{mn}{mx} \geq \tau , \qquad (5)$$

где $mn = \min\{s_0, s_1, ..., s_n\}$, $mx = \max\{s_0, s_1, ..., s_n\}$.

Доказательство. (Необходимое условие). Пусть $mn \ge \sigma$ и $\frac{mn}{mx} \ge \tau$. Множество A является сильным, если $X \Rightarrow Y$ – допустимое ассоциативное правило для любых непустых наборов данных X и Y, $X \cup Y = A$. Необходимо доказать, что *а)* $Support(X) \ge \sigma$, *б)* $Support(X \cup Y) \ge \sigma$, *в)* $\frac{Support(X \cup Y)}{Support(X)} \ge \tau$. Понятно, что $Support(X) \ge mn$. Так как $mn \ge \sigma$, то отсюда следует, что $Support(X) \ge \sigma$. Аналогично доказываем, что $Support(X \cup Y) \ge \sigma$. $\frac{Support(X \cup Y)}{Support(X)} \ge \frac{mn}{Support(X)} \ge \frac{mn}{mx}$, так как $\frac{mn}{mx} \ge \tau$, то все условия выполняются. Поэтому $X \Rightarrow Y$ – допустимое ассоциативное правило. Мы доказали, что A является сильным множеством.

(Достаточное условие). Пусть множество A является сильным. Нам необходимо доказать, что *а)* $mn \ge \sigma$, *б)* $\frac{mn}{mx} \ge \tau$. Так как $mn = \min\{s_0, s_1, ..., s_n\}$, $mx = \max\{s_0, s_1, ..., s_n\}$, пусть $X = \{a_k\}$ и $Support(X) = Support(\{a_k\}) = mx$. Пусть $Y = H \setminus X$ – множество всех элементов H кроме a_k. Так как H является сильным множеством, то для правила $X \Rightarrow Y$ выполняются условия *1)* $Support(X) \ge \sigma$, *2)* $Support(X \cup Y) \ge \sigma$, *3)* $\frac{Support(X \cup Y)}{Support(X)} \ge \tau$.

В наихудшем случае $Support(X \cup Y) = s_0 = mn$, поэтому отсюда следует, что $mn \ge \sigma$. $\frac{Support(X \cup Y)}{Support(X)} \ge \frac{mn}{mx}$, поэтому из *3)* делаем вывод, что $\frac{mn}{mx} \ge \tau$.

Метод поиска непересекающихся сильных наборов данных

Данный метод *(Disjoint Strong Itemsets Searching Method)* представляет собой метод поиска непересекающихся сильных наборов данных в базе транзакций.

Основная идея метода заключается в нахождении часто встречающихся наборов данных *(frequent patterns),* которые по формуле (5) проверяются, являются ли они сильными. Если сильный набор данных найден, он удаляется из оригинальной базы данных и аналогичный процесс поиска повторяется для новой базы.

Пусть D – база данных, которая состоит из N транзакций T, σ – минимальная поддержка, τ – минимальная достоверность.

Пусть $F = \{f_1, f_2, ..., f_n\}$ – множество элементов, которые часто встречаются, то есть для них выполняются условия:

$$Support(f_i) \ge \sigma, Confidence(f_i) \ge \tau, f_i \in F, i = 0..|F|. \quad (6)$$

Понятно, что сильные наборы данных могут состоять только из элементов множества F:

$$f_i \in s_j, f_i \notin s_k, j \ne k, f_i \in F. \quad (7)$$

Результатом работы метода является множество сильных наборов данных S, которые не пересекаются между собой, то есть:

$$\begin{aligned} &\forall(s_i \cap s_j) = \varnothing,\ i \ne j, s_i, s_j \in S, \\ &Support(s_i) \ge \sigma,\ Confidence(s_i) \ge \tau, \end{aligned} \quad (8)$$

а также множество элементов F', которые не вошли в сильные наборы данных:

$$F' = \{f_i \notin (\forall s_j \in S)\}. \quad (9)$$

Этот метод может использоваться как альтернатива поиска часто повторяющихся наборов

данных. Однако для поиска допустимых ассоциативных правил данный метод использоваться не может, поскольку он не учитывает часто встречающиеся элементы, которые не входят ни в один из сильных наборов данных.

Однако данный метод может использоваться как начальный этап для метода поиска ассоциативных правил на основе сильных наборов данных, который рассматривается ниже.

Метод поиска ассоциативных правил на основе сильных наборов данных

Данный метод является модификацией метода Apriori [5]. Он позволяет получить ассоциативные правила без генерации большого количества кандидатов. Входными данными метода являются множество сильных наборов данных, которые не пересекаются $S = \{s_1, s_2, ..., s_n\}$, и множество часто встречающихся элементов $F = \{f_1, f_2, ..., f_n\}$, которые не входят ни в один сильный набор.

Результатом является множество сильных наборов данных, которые могут содержать совместные элементы, и ассоциативные правила, образованные на основе данных наборов.

В данном методе кандидаты длиной k генерируются в результате пересечения сильных наборов данных и часто повторяющихся элементов:

$$\begin{aligned} C_k = \{\{u, v\} \mid (1 \le i, j \le k, u \in s_i \wedge v \in s_j \wedge i \ne j) \vee \\ \vee (1 \le i \le m, 1 \le j \le k, u \in f_i \wedge v \in s_j) \vee \\ \vee (1 \le i, j \le m, u \in f_i \wedge v \in f_j \wedge i \ne j)\}, \end{aligned} \tag{10}$$

где $m = \mid F \mid$ – количество часто повторяющихся элементов.

Среди всех кандидатов с помощью (5) находят множество сильных наборов L_k, которое объединяют с множеством S:

$$S = S \cup L_k,$$

а элементы, которые входят в L_k, исключают из множества F:

$$F = F \setminus \{L_k\}.$$

Этот процесс повторяют для кандидатов большей длины до тех пор, пока $L_k \ne \varnothing$. Заключительным этапом является подсчет улучшения для полученных ассоциативных правил.

Данный метод позволяет значительно сократить количество кандидатов, которые генерируются, а конечное множество сильных наборов данных позволяет показать связь между отдельными наборами и элементами в них.

Выводы

В данной работе рассмотрена концепция сильных наборов данных, решающих проблему генерации большого количества неэффективных правил; предложен метод поиска непересекающихся сильных наборов данных, а также метод поиска ассоциативных правил на основе сильных наборов данных. Разработанные методы могут использоваться для поиска знаний в виде ассоциативных правил в экономике, биологии, инженерных и научных исследованиях. Кроме того метод поиска сильных наборов данных может использоваться самостоятельно для нахождения логических связей, как между отдельными наборами, так и между отдельными элементами в них.

СПИСОК ЛИТЕРАТУРЫ

1. Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.

2. Дюк В.А., Самойленко А.П. Data Mining: учебный курс. – СПб.: Питер, 2001.

3. Чубукова И.А. Data Mining БИНОМ. Лаборатория знаний, Интернет-университет информационных технологий - ИНТУИТ.ру, 2006.

4. Han, J., J. Pei, Y. Yin, “Mining Frequent Patterns without Candidate Generation”, Proc. ACM SIGMOD 2000.

5. Agrawal R., T. Imielinski, A. Swami, “Mining Associations between Sets of Items in Massive Databases”, Proc. ACM SIGMOD 1993. - p. 207 – 216.

Краковецкий Александр Юрьевич *– аспирант кафедры компьютерные системы управления.*

Винницкий национальный технический университет.