УДК 681.3.06

О. Ю. Краковецький, аспірант

МЕТОД ПОШУКУ АСОЦІАТИВНИХ ПРАВИЛ НА ОСНОВІ СИЛЬНИХ НАБОРІВ ДАНИХ І FP-ДЕРЕВА

В роботі запропоновано концепцію сильних наборів даних, що вирішують проблему генерації великої кількості кандидатів під час розв'язання задачі пошуку знань у вигляді асоціативних правил, метод пошуку сильних наборів даних, що не перетинаються, а також метод пошуку асоціативних правил на основі сильних наборів даних.

Ключові слова: *асоціативні правила, FP-дерево, Data Mining.*

Актуальність

Однією з проблем пошуку знань у вигляді асоціативних правил є генерація кандидатів, кількість яких може бути дуже великою [1, 2]. Для скорочення їх кількості в різних методах використовують різні підходи, серед яких правило антимонотонності [2, 3], експертні знання, певні припущення тощо. Одним із методів скорочення кількості кандидатів є побудова дерева транзакцій (FP-tree) [4]. Цей метод дозволяє знаходити набори даних, які часто повторюються, проте він має ряд недоліків, серед яких неврахування значення достовірності набору, відсутність інтуїтивно зрозумілого логічного зв'язку між елементами в наборах. Це призводить до того, що велика кількість правил, що генеруються на основі цих наборів даних, відкидаються на стадії перевірки. Отже, вирішення проблеми генерації надлишкових наборів даних, є актуальною задачею.

У цій роботі запропоновано концепцію *сильних* наборів даних, що вирішують поставлену проблему, метод пошуку сильних наборів даних, які не перетинаються, а також метод пошуку асоціативних правил на основі сильних наборів даних.

Постановка задачі

Постановка задачі пошуку асоціативних правил: необхідно знайти множину наборів даних, підтримка яких більша за наперед задане значення мінімальної підтримки $Supp_{\min}$, мінімальної достовірності $Conf_{\min}$ і покращення більше одиниці [1, 2]:

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

Поняття сильних наборів даних

Під *набором даних* X будемо розуміти непусту підмножину елементів загальної множини елементів A:

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

LSI ($left-side\ itemset$) – непустий набір даних, який формує ліву частину асоціативного правила, відповідно RSI ($right-side\ itemset$) – непустий набір даних, який формує праву частину:

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

Нехай $LSI = X$ і $RSI = Y$, $0 \le \sigma, \tau \le 100$, де σ, τ – відповідно задані мінімальна підтримка і достовірність [1, 2]. Тоді асоціативне правило $X \Rightarrow Y$ називатимемо *допустимим* для заданих σ і τ, якщо виконуються такі умови:

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

Якщо будь-які набори даних X і Y множини A ($X \cup Y = A, X \cap Y = \varnothing, |A| > 2$) утворюють допустимі асоціативні правила $X \Rightarrow Y$ для заданих σ і τ, то множину A будемо називати *сильним набором даних*.

Розглянемо це на прикладі. Нехай $A = \{a, b, c\}$ – множина, що складається з трьох елементів. Множина усіх можливих варіантів розділення A на підмножини, які представляють собою шаблони асоціативних правил, складається з 12 елементів (таблиця 1).

Таблиця 1

Набори даних і відповідні правила для множини $A = \{a, b, c\}$

Набори даних	Правило	Набори даних	Правило
$\{a\}, \{b\}$	$a \Rightarrow b$	$\{a\}, \{b, c\}$	$a \Rightarrow b, c$
$\{b\}, \{a\}$	$b \Rightarrow a$	$\{b, c\}, \{a\}$	$b, c \Rightarrow a$
$\{a\}, \{c\}$	$a \Rightarrow c$	$\{a, b\}, \{c\}$	$a, b \Rightarrow c$
$\{c\}, \{a\}$	$c \Rightarrow a$	$\{c\}, \{a, b\}$	$c \Rightarrow a, b$
$\{b\}, \{c\}$	$b \Rightarrow c$	$\{a, c\}, \{b\}$	$a, c \Rightarrow b$
$\{c\}, \{b\}$	$c \Rightarrow b$	$\{b\}, \{a, c\}$	$b \Rightarrow a, c$

Якщо *всі* правила, які утворюють дані набори, є допустимими, то множина A є сильною.

Кількість можливих варіантів вибору m елементів з N, які складатимуть ліву частину асоціативного правила, дорівнює C_N^m. Кількість можливих варіантів вибору n елементів з тих, що залишилися, тобто $N - m$, дорівнює C_{N-m}^n. Оскільки і ліва, і права частини повинні бути присутніми в правилах, то загальна кількість комбінацій буде дорівнювати $C_N^m C_{N-m}^m$. В нашому випадку m змінюється від 1 до $N-1$, а n – від 1 до $N - m$. Нехай $m = |LSI|$ і $n = |RSI|$ – кількість елементів відповідно в лівій і правій частинах асоціативного правила, $N = m + n$ – загальна кількість елементів в правилі. Тоді загальна кількість правил, які можна утворити з сильної множини $A = LSI \cup RSI$, що складається з N елементів, визначається за допомогою виразу:

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

Висновок: сильні множини не призводять до втрати інформативності даних і можуть компактно представити велику кількість асоціативних правил.

Необхідна і достатня умова, за якої множина є сильною. Нехай маємо множину $A = \{a_1, a_2, ..., a_n\}$, $n \geq 2$, де $s_1, s_2, ..., s_n$ – підтримка елементів $a_1, a_2, ..., a_n$ відповідно, s_0 – підтримка множини A, σ і τ – відповідно мінімальна підтримка і мінімальна достовірність. Тоді A є сильною множиною, якщо виконуються умови:

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

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

Доведення. (Необхідна умова). Нехай $mn \ge \sigma$ і $\frac{mn}{mx} \ge \tau$. Множина A є сильною, якщо $X \Rightarrow Y$ – допустиме асоціативне правило для будь-яких непустих наборів даних X і Y, $X \cup Y = A$. Необхідно довести, що *а)* $Support(X) \ge \sigma$, *б)* $Support(X \cup Y) \ge \sigma$, *в)* $\frac{Support(X \cup Y)}{Support(X)} \ge \tau$. Зрозуміло, що $Support(X) \ge mn$. Оскільки $mn \ge \sigma$, то звідси випливає, що $Support(X) \ge \sigma$. Аналогічно доводимо, що $Support(X \cup Y) \ge \sigma$. $\frac{Support(X \cup Y)}{Support(X)} \ge \frac{mn}{Support(X)} \ge \frac{mn}{mx}$, оскільки $\frac{mn}{mx} \ge \tau$, то всі умови виконуються. Тому $X \Rightarrow Y$ – допустиме асоціативне правило. Ми довели, що A є сильною множиною.

(Достатня умова). Нехай множина A є сильною. Нам необхідно довести, що *а)* $mn \ge \sigma$, *б)* $\frac{mn}{mx} \ge \tau$. Оскільки $mn = \min\{s_0, s_1, ..., s_n\}$, $mx = \max\{s_0, s_1, ..., s_n\}$, нехай $X = \{a_k\}$ і $Support(X) = Support(\{a_k\}) = mx$. Нехай $Y = H \setminus X$ – множина всіх елементів, H крім a_k. Оскільки H є сильною множиною, то для правила $X \Rightarrow Y$ виконуються умови *1)* $Support(X) \ge \sigma$, *2)* $Support(X \cup Y) \ge \sigma$, *3)* $\frac{Support(X \cup Y)}{Support(X)} \ge \tau$.

В найгіршому випадку $Support(X \cup Y) = s_0 = mn$, тому звідси випливає, що $mn \ge \sigma$. $\frac{Support(X \cup Y)}{Support(X)} \ge \frac{mn}{mx}$, тому з *3)* випливає, що $\frac{mn}{mx} \ge \tau$.

Метод пошуку сильних наборів даних, що не перетинаються

Disjoint Strong Itemsets Searching Method представляє метод пошуку сильних наборів даних, що не перетинаються між собою, в базі транзакцій.

Основна ідея методу полягає в знаходженні наборів даних, що часто зустрічаються *(frequent patterns),* які за формулою (5) визначаються, чи є дані набори сильними. Якщо сильний набір даних знайдено, він видаляється з оригінальної бази даних і аналогічний процес пошуку повторюється для нової бази.

Нехай D – база даних, яка складається з N транзакцій T, σ – мінімальна підтримка, τ – мінімальна достовірність.

Нехай $F = \{f_1, f_2, ..., f_n\}$ – множина елементів, які часто зустрічаються, тобто для них виконуються умови:

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

Зрозуміло, що сильні набори даних можуть складатися лише з елементів множини F:

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

Результатом роботи методу є множина сильних наборів даних S, які не перетинаються між собою, тобто:

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

а також множина елементів F', які не ввійшли в сильні набори даних:

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

Цей метод може використовуватися як альтернатива знаходження наборів даних, що часто повторюються. Проте для пошуку допустимих асоціативних правил описаний метод використовуватися не може, тому що він не враховує елементи, які часто зустрічаються, але

не входять в жоден із сильних наборів даних.

Проте цей метод може використовуватися як початковий етап для методу пошуку асоціативних правил на основі сильних наборів даних, який розглянуто нижче.

Метод пошуку асоціативних правил на основі сильних наборів даних

Цей метод є модифікацією методу Apriori [5]. Він дозволяє отримати асоціативні правила без генерування великої кількості кандидатів. Вхідними даними методу є множина сильних наборів даних, що не перетинаються $S = \{s_1, s_2, ..., s_n\}$, і множина елементів, що часто зустрічаються $F = \{f_1, f_2, ..., f_n\}$, які не входять в жодний сильний набір.

Результатом є множина сильних наборів даних, які можуть містити спільні елементи, і асоціативні правила, утворені на основі даних наборів.

У цьому методі кандидати довжиною k генерують на основі перетину сильних наборів даних і елементів, які часто повторюються:

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

де $m = \mid F \mid$ – кількість елементів, що часто повторюються.

Серед усіх кандидатів за допомогою (5) знаходять множину сильних наборів L_k, яку об'єднують з множиною S:

$$S = S \cup L_k,$$

а елементи, які входять в L_k, виключають з множини F:

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

Цей процес повторюють для кандидатів більших довжин доти, доки $L_k \ne \varnothing$. Заключним етапом є підрахунок покращення для отриманих асоціативних правил.

Цей метод дозволяє значно скоротити кількість кандидатів, що генеруються, а результуюча множина сильних наборів даних дає можливість показати зв'язки між окремими наборами і елементами в них.

Висновки

У цій роботі розглянуто концепцію сильних наборів даних, які вирішують проблему генерації великої кількості неефективних правил, запропоновано метод пошуку сильних наборів даних, що не перетинаються, а також метод пошуку асоціативних правил на основі сильних наборів даних. Розроблені методи можуть використовуватися для пошуку знань у вигляді асоціативних правил в економіці, біології, інженерних та наукових дослідженнях. Крім того метод пошуку сильних наборів даних може використовуватися самостійно для знаходження логічних зв'язків як між окремими наборами, так і між окремими елементами в них.

СПИСОК ЛІТЕРАТУРИ

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.

Краковецький Олександр Юрійович – аспірант кафедри комп'ютерні системи управління.
Вінницький національний технічний університет.