

МОДИФІКОВАНИЙ АЛГОРИТМ ПОБУДОВИ FP-ДЕРЕВ ПРИ АНАЛІЗІ НАДЗВИЧАЙНИХ СИТУАЦІЙ НА ЗАЛІЗНИЦІ

Савчук Т.О., Юхимчук С.В., Вознюк А.В.

Анотація. В роботі запропоновано модифікований алгоритм побудови FP-дерев, який використовується при аналізі надзвичайних ситуацій, що виникають при перевезенні небезпечних вантажів залізничним транспортом. До модифікації належить: введення небінарних атрибутів, що збільшують покриття надзвичайних ситуацій кожним правилом, та використання більш складних перетинів множин атрибутів ситуації порівняно з правилами, що оперують бінарними атрибутами; введення характеристики інформативності правила.

Ключові слова: асоціативні правила, надзвичайні ситуації, пошук закономірностей, модифікований алгоритм, побудова FP-дерев, небінарний атрибут, інформативність правила.

Аннотация. В работе предложен модифицированный алгоритм построения FP-деревьев, который используется при анализе чрезвычайных ситуаций, возникающих при перевозке опасных грузов железнодорожным транспортом. К модификации принадлежат: введение небинарных атрибутов, которые увеличивают покрытие чрезвычайных ситуаций каждым правилом, а также использование более сложных пересечений множественных чисел атрибутов ситуации в сравнении с правилами, которые оперируют бинарными атрибутами; введение характеристики информативности правила.

Ключевые слова: ассоциативные правила, чрезвычайные ситуации, поиск закономерностей, модифицированный алгоритм, построение FP-дерев, небинарный атрибут, информативность правила.

Annotation. The modified algorithm of construction of FP-trees, which is used for the analysis of extraordinary situations which arise up at transportation of dangerous loads a railway transport is in-process offered. It is belonged to modification: introduction of unbinary attributes, which increase coverage of extraordinary situations every rule and the use of more difficult crossings of plurals of attributes of situation by comparing to rules which operate binary attributes; introduction of description of informing of rule.

Keywords: associative rules, extraordinary situations, search of conformities to law, modified algorithm, construction of FP-tree, unbinary attribute, informing of rule.

Вступ

Найбільш значущими і необхідними завданнями прогнозування надзвичайних ситуацій (НС) на залізничному транспорті є ймовірність виникнення кожного з шляхів розвитку таких ситуацій і, відповідно, їх масштабів, розмірів зон НС. Можливі тривалі наслідки при виникненні НС певних типів, масштабів, часових інтервалів або їх певних сукупностей; потреби сил і засобів для ліквідації прогнозованих надзвичайних ситуацій [1].

В цілому результати моніторингу та прогнозування є вихідною основою для розробки довгострокових, середньострокових і короткострокових прогнозів для прийняття відповідних рішень щодо попередження та ліквідації надзвичайних ситуацій [1, 2].

Значні обсяги сучасних баз даних викликали стійкий попит на нові масштабовані алгоритми аналізу даних. Одним з ефективних методів виявлення знань стали алгоритми пошуку асоціативних правил [3].

Доцільність застосування пошуку асоціативних правил до аналізу НС на залізничному транспорті

Безпека функціонування небезпечних об'єктів залежить від багатьох чинників: фізико-хімічних властивостей сировини, напівпродуктів і продуктів, від характеру технологічного процесу, від конструкції і надійності устаткування, умов зберігання і транспортування радіоактивних вантажів, стану контрольно-вимірювальних приладів і засобів автоматизації, ефективності засобів протиаварійного захисту і т.д. [3].

Потрібно аналізувати задачу та приймати необхідні рішення на основі неповних, нечітких даних, то традиційні методи не зможуть справитися з поставленою задачею. Виникає необхідність використання саме інтелектуальних систем та методів. Всі відхилення на протязі процесу розвитку надзвичайної ситуації, що можуть змінити його хід чи вплинути на якість вихідного результату знаходяться в відомих рамках, які також визначаються в процесі аналізу.

Таким чином, створюється статистична стабільність [3, 4], що обумовлює застосування технологій Data Mining при знаходженні закономірностей між зв'язаними подіями в базах даних та відповідає задачі пошуку асоціативних правил [5].

Постановка задачі

Перед побудовою асоціативних правил та подальшим їх аналізом необхідно визначити об'єкти, що будуть використовуватись.

Для побудови асоціативних правил достатньо бази даних, кожен кортеж якої описує конкретну надзвичайну ситуацію. Кортеж містить набір значень атрибутів надзвичайної ситуації, які описують наступні об'єкти: небезпечний вантаж ($HB\{pp_1, \dots, pp_m\}$), оточуюче середовище ($C\{c_1, \dots, c_p\}$), контейнер ($K\{k_1, \dots, k_n\}$), надзвичайна ситуація ($HC\{h_1, \dots, h_t\}$).

Усі атрибути можуть приймати як числові значення, так і представлятись у вигляді проміжку значень.

Отже, задачу аналізу НС, що базується на пошуку асоціативних правил. Можна сформулювати таким чином: на основі множини значень атрибутів надзвичайних ситуацій $HB\{pp_1, \dots, pp_m\}$, $K\{k_1, \dots, k_n\}$, $C\{c_1, \dots, c_p\}$, $HC\{h_1, \dots, h_t\}$ необхідно визначити асоціативні правила, що описують надзвичайні ситуації та застосувати їх для аналізу вхідних даних, якими є значення атрибутів конкретної НС, що аналізується.

Модифікований алгоритм побудови FP-дерев

Завдання знаходження асоціативних правил розбивається на дві підзадачі:

1. Знаходження всіх наборів елементів, які задовольняють порог minsupport . Такі набори елементів називаються часто зустрічаючими (частими наборами).

2. Генерація правил з наборів елементів, знайдених згідно п.1. з вірогідністю, що задовольняє порог minconfidence .

Для опису небінарних величин використано апарат алгебри кінцевих предикатів, тобто якщо ознака X_j об'єкту може приймати значення з множини $A = \{a_1, a_2, \dots, a_n\}$, то така ознака може описуватись за допомогою кінцевого предикату наступним чином:

$$X_j^{a_i} = \begin{cases} 1, & \text{якщо } X_j = a_i, \\ 0, & \text{якщо } X_j \neq a_i. \end{cases}$$

Зазначимо, що мають місце наступні тотожності:

$$X^{a_1} \vee X^{a_2} \vee \dots \vee X^{a_n} \equiv 1, \\ X^{a_i} X^{a_j} \equiv 0,$$

які дозволяють не використовувати окремі описи для асоціації з запереченням, тобто $X^{a_i} = X^{a_1} \vee \dots \vee X^{a_{i-1}} \vee X^{a_{i+1}} \vee \dots \vee X^{a_n}$.

У випадку бінарних ознак для їх опису використовується унарний предикат, що розрізняє два значення – 0 і 1: X^0 і X^1 .

Двох характеристик – підтримки й рівня довіри недостатньо для оцінки асоціації. У доповнення до підтримки й рівня довіри будемо використовувати таку характеристику, як рівень покращення, яка визначається таким чином:

$$\text{Imp}(X \rightarrow Y) = \frac{\text{Conf}(X \rightarrow Y)}{\text{Sup}(Y)}.$$

Розглянемо характеристики асоціативного правила з точки зору теорії ймовірностей:

$$\text{Sup}(X \rightarrow Y) = P(XY);$$

$$\text{Conf}(X \rightarrow Y) = \frac{\text{Sup}(X \rightarrow Y)}{\text{Sup}(X)} = \frac{P(XY)}{P(X)} = P_X(Y);$$

$$\text{Imp}(X \rightarrow Y) = \frac{\text{Sup}(X \rightarrow Y)}{\text{Sup}(X) \text{Sup}(Y)} = \frac{P(XY)}{P(X)P(Y)}.$$

Тоді обмеження, що накладаються на характеристики асоціації, визначаються як:

$$1) 0 \leq \text{Sup}(X \rightarrow Y) \leq 1;$$

$$2) \text{Sup}(X \rightarrow Y) \leq \text{Conf}(X \rightarrow Y) \leq 1;$$

$$3) \frac{(\text{Conf}(X \rightarrow Y))^2}{\text{Sup}(X \rightarrow Y) \cdot \text{Conf}(X \rightarrow Y) + \text{Conf}(X \rightarrow Y) - \text{Sup}(X \rightarrow Y)} \leq \text{Imp}(X \rightarrow Y) \leq \frac{\text{Conf}(X \rightarrow Y)}{\text{Sup}(X \rightarrow Y)}.$$

Після знаходження асоціативних правил необхідно визначити доцільність та можливість їх використання. Для цього використовуватимемо оцінку інформативності асоціації, що розраховується на основі стандартних характеристик асоціативного правила. Повна взаємна інформація у загальному випадку може бути представленою таким чином:

$$I_{X \leftrightarrow Y} = \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log_2 \frac{P_{ij}}{P_i P_j};$$

де $P_{ij} = P((X \sim x_i)(Y \sim y_j))$ – ймовірність того, що X знаходиться в стані x_i , а Y – в стані y_j ;

$p_i = P(X \sim x_i)$ – ймовірність того, що X знаходиться в стані x_i ;

$r_j = P(Y \sim y_j)$ – ймовірність того, що Y знаходиться в стані y_j .

Для асоціативних правил взаємна інформація буде визначатися як:

$$I_{X \leftrightarrow Y} = \sum_{i=1}^n \sum_{j=1}^m \text{Sup}(x_i \rightarrow y_j) \log_2 \text{Imp}(x_i \rightarrow y_j).$$

Правило процесу побудови FP-дерев: якщо для чергового об'єкта в дереві зустрічається вузол, ім'я якого збігається з іменем об'єкта, то об'єкт не створює нового вузла, а індекс відповідного вузла в дереві збільшується на 1. В іншому разі для цього об'єкта створюється новий вузол і йому присвоюється індекс 1.

Розглянемо генерацію агрегованих асоціативних правил на основі об'єднання гілок дерева покриттів, що полягає у наступному. Якщо потребується знайти узагальнені покриття, які містять значення ознаки P^{t_1} і P^{t_2} , то визначається список гілок дерева покриттів з даними значеннями ознаки P . Ці гілки групуються наступним чином: в одну групу входять ті, що містять ознаки на однакових ярусах дерева покриттів. В цьому списку існують гілки, які мають однакові префікси: $S_1\alpha_1, S_1\alpha_2, \dots, S_1\alpha_i$ (такі гілки обов'язково існують, тому що всі гілки містять хоча б один загальний префікс – нульову строку λ). У кожній групі можуть існувати гілки з різними загальними префіксами. Знаходимо $S = \max\{S_j\}$, ($j = \overline{1, k}$), де k – кількість різних однакових префіксів в даній групі гілок. Ці гілки об'єднуються по принципу: $S(\alpha_1 + \alpha_2 + \dots + \alpha_i)$.

Рядок $(\alpha_1 + \alpha_2 + \dots + \alpha_i)$ формується таким чином: між значеннями ознаки з одного ярусу ставиться знак $\vee$. Отримана агрегована гілка додається в свою групу. Далі цей процес продовжується ітеративно до тих пір, поки існують гілки з однаковими префіксами.

Рівень підтримки такого об'єднання дорівнює сумі рівнів підтримки гілок, що об'єднуються.

Запропонований метод розрахунку дозволяє розкласти агреговане асоціативне правило [4, 5] на прості й знаходити його рівень підтримки через рівні підтримки простих асоціацій, які легко знаходяться з побудованого дерева покриттів. Наприклад, правило $(Z^{c1} \vee Z^{c2}) \rightarrow X^{a3}(Y^{b1} \vee Y^{b2})$, що отримано з покриття $X^{a3}(Y^{b1} \vee Y^{b2})(Z^{c1} \vee Z^{c2})$ ($\text{Sup}=3$) може бути розкладено, як показано на рисунку 1.

Відмітимо, що розкладання агрегованого асоціативного правила може проводитися як зліва направо, так і навпаки. Даний підхід дозволяє не використовувати додаткове сканування БД для знаходження параметрів узагальнених асоціацій.

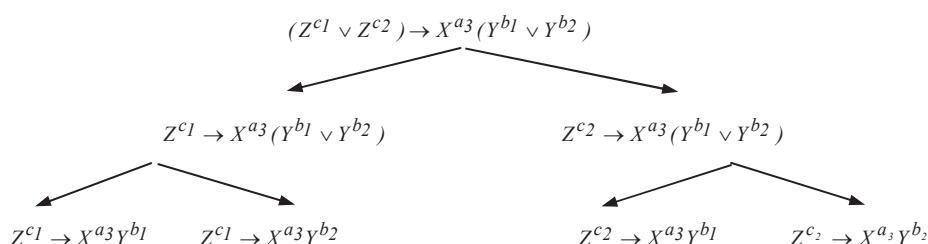


Рисунок 1 – Розкладання агрегованого асоціативного правила

Висновки

Таким чином, використання модифікованого алгоритму побудови FP-дерев, що базується на введенні інформативності небінарних асоціативних правил та подальша їх агрегація наведеним в роботі способом дозволяє, окрім розв'язання поставленої задачі, значно збільшити ефективність та швидкодію обробки надзвичайної ситуації. При цьому, одним з ефективних методів побудови асоціативних правил, що використовуватимуться для аналізу надзвичайних ситуацій є алгоритм побудови FP-дерев. При побудові небінарних асоціативних правил та розгляді їх з точки зору теорії ймовірностей вводить додаткова характеристика – “інформативність” правила.

Список літератури

1. С. В. Юхимчук, М. Д. Кацман «Моделі автоматизації вироблення рекомендацій керівнику гасіння пожежі на залізничному транспорті», Універсум-Вінниця, 2008 р.
2. М. Д. Кацман, Г. Б. Кононов, І. В. Діденко, Н. В. Огороднійчук. Ліквідація пожеж на залізничному транспорті. Київ, видавництво «Основа» 2006 р. 215 с.
3. 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 – 336с.
4. Nicolas Pasquier, Yves Bastide, Rafik Taouil, Lotfi Lakhal Discovering Frequent Closed Itemsets for Association Rules.
5. Игнатов Д, Методы бикластеризации для анализа интернет-данных.

Відомості про авторів

Савчук тамара Олександрівна – к.т.н., професор кафедри комп'ютерних наук, Україна, м. Вінниця, Вінницький національний технічний університет, savchtam@gmail.com.

Юхимчук Сергій Васильович – д.т.н., професор кафедри комп'ютерних наук, Україна, м.Вінниця, Вінницький національний технічний університет.

Вознюк Андрій Вікторович – магістрант кафедри комп'ютерних наук, Україна, м. Вінниця, Вінницький національний технічний університет, kenny@ukr.net.