УДК 004.93

А.О. ОЛІЙНИК*

ПОБУДОВА АСОЦІАТИВНИХ ПРАВИЛ НА ОСНОВІ ІНТЕЛЕКТУАЛЬНОГО СТОХАСТИЧНОГО ПОШУКУ

*Запорізький національний технічний університет, Запоріжжя, Україна

Анотація. *Вирішено задачу автоматизації побудови чисельних асоціативних правил на основі заданої множини спостережень. Запропоновано стохастичний метод побудови чисельних асоціативних правил, що враховує апріорну інформацію про значущість термів і ознак та використовує ймовірнісний підхід для перебору різних сполучень антецедентів і консеквентів асоціативних правил. Розроблено програмне забезпечення, що реалізує запропонований метод, а також проведено експерименти з його дослідження при вирішенні практичних завдань.*
Ключові слова: *асоціативне правило, діагностування, інформативність, модель, ознака, стохастичний підхід, терм, транзакція.*

Аннотация. *Решена задача автоматизации построения численных ассоциативных правил на основе заданного множества наблюдений. Предложен стохастический метод построения численных ассоциативных правил, который учитывает априорную информацию о значимости термов и признаков и использует вероятностный подход для перебора различных сочетаний антецедентов и консеквентов ассоциативных правил. Разработано программное обеспечение, реализующее предложенный метод, а также проведены эксперименты по его исследованию при решении практических задач.*
Ключевые слова: *ассоциативное правило, диагностирование, информативность, модель, признак, стохастический подход, терм, транзакция.*

Abstract. *The problem of automation of extracting quantitative association rules based on a set of observations is solved. The stochastic method to extracting quantitative association rules is proposed. It takes into account a priori information about the significance of the terms and features and uses a probabilistic approach for analysis of different combinations of antecedents and consequents of association rules. It was developed software implementing proposed method. The experiments with proposed method in practical problem solving were conducted as well.*
Keywords: *association rule, diagnostics, informativeness, model, feature, stochastic approach, term, transaction.*

1. Вступ

Для оброблення наборів даних, що містять велику кількість пропущених значень або представлених у виді баз транзакцій, де кожне спостереження (транзакція) містить значення деяких з можливих ознак досліджуваних об'єктів, доцільно використовувати методи побудови асоціативних правил [1, 2], оскільки вони дозволяють виявляти сховані залежності в даних, скорочувати розмірність даних, тим самим підвищуючи рівень узагальнення, а також знижуючи структурну і параметричну складність синтезованих на їх основі моделей. У цьому випадку розглядається варіант вхідних даних, де деякі значення ознак або вихідного параметра можуть бути не визначені. У результаті застосування методів виявлення асоціативних правил створюється множина $A = \{A_1, A_2, ..., A_{N_A}\}$ правил виду $A_r : P_r \to T_r$, де P_r – антецедент – ліва частина r-го правила A_r, що визначає набір умов виконання правила A_r, T_r – консеквент – права частина r-го правила A_r, що визначає значення вихідного параметра при виконанні умов P_r правила A_r, $N_A = |A|$ – кількість витягнутих правил [3–6].

ISSN 1028-9763. Математичні машини і системи, 2015, № 4

Відомі методи побудови асоціативних правил SCF, SETM, Apriori, DHP, Eclat та ін. [1–4] при формуванні наборів, що часто зустрічаються, у процесі синтезу правил використовують властивість антимонотонності підтримки (відповідно до якого підтримка набору елементів не перевищує значення підтримки кожної з його підмножин) або інші процедури [2–6], що обумовлює такі недоліки цих методів:

– аналогічно до жадібної стратегії (greedy strategy) пошуку аналізуються всі можливі комбінації з високими значеннями підтримки, що при великій кількості ознак P у вихідній множині S вимагає перевірки великої кількості комбінацій ознак P_r, виконуючи істотну кількість проходів по базі S і витрачаючи на це великі ресурси пам'яті та часу роботи ЕОМ;

– такий підхід не дозволяє у процесі пошуку генерувати правила A_r з наборів ознак, що містять комбінації з низькими значеннями підтримки (комбінацій ознак, що зустрічаються рідко);

– при використанні такого підходу виявляються тільки правила, синтезовані на основі наборів, що часто зустрічаються, внаслідок чого не витягаються цікаві правила $A_r : P_r \to T_r$ з високим рівнем вірогідності $\mathrm{conf}(A_r)$ при низькому рівні підтримки $\mathrm{supp}(A_r)$. Це істотно знижує апроксимаційні й узагальнюючі здібності синтезованої на основі виділеного набору $A = \{A_1, A_2, ..., A_{N_A}\}$ асоціативних правил моделі.

Крім того, більшість методів побудови асоціативних правил призначена для оброблення бінарних даних. У той же час більшість реальних задач розпізнавання образів, контролю якості, діагностування пов'язана з необхідністю оброблення чисельних даних, де більшість ознак приймають значення з деякого діапазону.

Потреба в усуненні зазначених недоліків обумовлює необхідність розробки нового методу побудови асоціативних правил.

Метою роботи є створення методу побудови асоціативних правил на основі стохастичного підходу.

2. Постановка завдання

Нехай задано множину спостережень $S =< P, T >$, де P – набір характеристик (ознак) спостережень, T – множина значень вихідного параметра, p_{qm} – значення m-го атрибута q-го спостереження ($m = 1, 2, ..., M$, $q = 1, 2, ..., Q$), t_q – значення вихідного параметра q-го спостереження, M – кількість атрибутів, Q – кількість спостережень.

Тоді задача побудови чисельних асоціативних правил полягає у побудові бази $A = \{A_1, A_2, ..., A_{N_A}\}$ правил виду $A_r : P_r \to T_r$, що задовольняють прийнятному рівню заданого критерію якості. Як такий критерій може використовуватися вірогідність правила [1–4], що обчислюється як відношення підтримки правила $\mathrm{supp}(P_k \cup T_k)$ до підтримки його антецедента $\mathrm{supp}(P_k)$ (1):

$$\mathrm{conf}(A_k) = \frac{\mathrm{supp}(P_k \cup T_k)}{\mathrm{supp}(P_k)}, \qquad (1)$$

де $\mathrm{supp}(P_k \cup T_k)$ – підтримка правила $A_k : P_k \to T_k$, яка визначається як відношення кількості екземплярів $N(P_k \cup T_k)$ вибірки S, що містять множину умов P_k і характеризуються значенням вихідного параметра T_k до кількості екземплярів $N(P_k)$ вибірки S, що задовольняють умовам антецедента P_k правила $A_k : P_k \to T_k$.

3. Метод побудови асоціативних правил на основі стохастичного підходу

Для побудови асоціативних правил $A_r : P_r \to T_r$ із заданих наборів чисельних даних S пропонується попередньо розбивати діапазони значень ознак P на інтервали, на основі яких визначати терми ознак, враховуючи при цьому ширину діапазону значень і частоту попадання ознак у кожний з термів, після чого за допомогою стохастичного підходу виявляти асоціативні правила $A_r : P_r \to T_r$, що характеризуються високим рівнем вірогідності $\mathrm{conf}(A_r)$.

У розробленому стохастичному методі побудови чисельних асоціативних правил на початковому етапі відбувається розбиття значень ознак P на інтервали. Для цього пропонується у множині значень p_{qm} кожної ознаки p_m провести кластерний аналіз [7–11], виділивши групи компактно розташованих екземплярів (транзакцій) в одновимірному просторі кожної ознаки. У результаті виділяється набір кластерів $Cl_m = \{Cl_{1m}, Cl_{2m}, ...Cl_{N_{\text{int}}m}\}$. При використанні методів кластерного аналізу, в яких потрібно задавати кількість кластерів N_{int} (наприклад, методу нечітких с-середніх [5, 6]), як параметр $N_{\text{int}\,m}$ можна задати зменшену в N_A разів кількість екземплярів $N(p_m)$, для яких визначено значення ознаки p_m. Параметр N_A роботи методу може бути визначений за формулою (2):

$$N_A = ceil\left(\frac{1}{M \cdot N_{\text{int}\,mean}} \sum_{m=1}^{M} N(p_m)\right), \tag{2}$$

де $N_{\text{int}\,mean}$ – очікуване середнє значення інтервалів розбиття (кластерів) кожної з ознак p_m, $m = 1, 2, ..., M$, $ceil(x)$ – функція, що повертає цілу частину числа x. Число $N_{\text{int}\,mean}$ повинно бути невеликим (наприклад, можна рекомендувати задавати $N_{\text{int}\,mean} = 10, N_{\text{int}\,mean} << Q$), що дозволить забезпечити невисокі вимоги до ресурсів пам'яті ЕОМ при виконанні відповідних обчислень, і в той же час таким, що дозволить забезпечити прийнятне розбиття значень ознак p_m на інтервали.

Потім на основі границь $Cl_{n\min m}$ та $Cl_{n\max m}$ ($n = 1, 2, ..., N_{\text{int}\,m}$, $m = 1, 2, ..., M$) виділених кластерів визначаються інтервали значень (терми) $[Cl_{n\min m}; Cl_{n\max m})$ ознак p_m.

Після цього генерується N_χ рішень для виконання стохастичного пошуку. Рішення χ_k при видобуванні асоціативних правил представляється у вигляді множини параметрів $\chi_k = \{g_{1k}, g_{2k}, ..., g_{N_g k}\}$, де g_{mk} – m-й параметр рішення, що містить інформацію про номер терму Δp_{nm} m-ї ознаки p_m (або її відсутності) у k-му асоціативному правилі $A_k : P_k \to T_k$ (рис. 1).

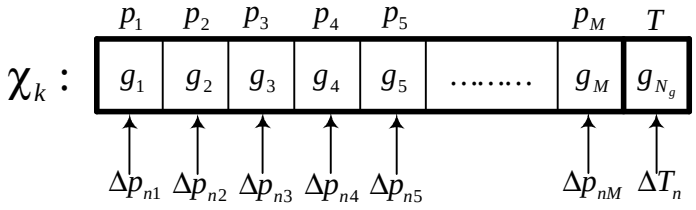

Рис. 1. Подання структури χ_k при побудові асоціативних правил

Як видно, у рішенні χ_k також присутня інформація про терм ΔT_n вихідного параметра T (у випадку, якщо він приймає дійсні значення). Якщо вихідний параметр T є бінарним, то останній ген g_{N_g} рішення χ_k відсутній.

Значення m-го параметра g_m кожного k-го рішення χ_k визначаються в такий спосіб. Генерується випадкове число rnd з діапазону [0;1] ($rnd = rand[0;1]$, де $rand[0;1]$ – функція, що повертає випадково згенероване число з інтервалу [0;1]). Якщо згенероване число rnd не перевищує частоту ν_m появи m-ї ознаки p_m у вибірці S ($rnd \le v_m$), то значенню параметра g_m рішення χ_k привласнюється нульове значення ($g_{mk} = 0$), що характеризує відсутність m-ї ознаки в k-му асоціативному правилі χ_k. Величина ν_m обчислюється за формулою (3):

$$\nu_m = \frac{N(p_m)}{Q}. \tag{3}$$

У випадку, якщо $rnd > \nu_m$, виконується генерація цілого випадкового числа $rndc$ з діапазону $[1; N_{\text{int}\,m}]$, що визначає номер інтервалу (терму) ознаки p_m у правилі χ_k (4):

$$rndc = randc[1; N_{\text{int}\,m}]. \tag{4}$$

При виконанні умови (5)

$$rnd \in randc[\nu_m; \nu_m + \nu_{nm}(1-\nu_m)] \tag{5}$$

параметру g_m привласнюється значення $rndc: g_m = rndc$.

Умова (5) показує, що n-й терм Δp_{nm} m-ї ознаки p_m має тим більшу ймовірність увійти до k-го правила χ_k, чим вищою є частота ν_{nm} її наявності у транзакціях вибірки S, в яких визначені значення m-ї ознаки (6):

$$\nu_{nm} = \frac{N(p_{nm})}{N(p_m)}. \tag{6}$$

При числовому значенні вихідного параметра T величина ν_{nm} для k-го рішення χ_k обчислюється як частота наявності терму Δp_{nm} у транзакціях вибірки S, в яких визначені значення m-ї ознаки, а значення вихідного параметра T дорівнює $g_{M+1,k}$.

Якщо умова (5) не виконується, тоді вважається, що n-й терм Δp_{nm} ознаки p_m не може бути включений у правило χ_k. Після чого відбуваються повторна генерація випадкового числа $rnd = rand[0;1]$ і перевірка умови (5), і так доти, поки не буде визначено значення гена g_{mk}.

Аналогічним чином формуються M параметрів g_m для кожного з N_χ рішень початкової популяції $R^{(0)} = \{\chi_1^{(0)}, \chi_2^{(0)}, ..., \chi_{N\chi}^{(0)}\}$ стохастичного пошуку, що представляє собою множину асоціативних правил.

У випадку числового (не бінарного) вихідного параметра T вибірки S значення гена $g_{M+1,k}$, що визначає номер терму вихідного параметра у правилі χ_k, генерується випадковим чином у залежності від частоти появи термів у вибірці S. Для цього за формулою (7) визначається імовірність того, що значення вихідного параметра T екземплярів вибірки S потрапить до n-го інтервалу T_n діапазону його значень:

$$\rho(T_n) = \frac{N(T_n)}{Q}. \tag{7}$$

Після чого кожному терму T_n ставиться у відповідність інтервал $\rho i(T_n) \in (T_{n-1,\max}; T_{n-1,\max} + \rho(T_n)]$, при цьому $\rho i(T_1) \in [0; \rho(T_1)]$. Далі генерується випадкове число $rnd = rand[0;1]$. Значення параметра $g_{M+1,k}$ відповідає номеру інтервалу $\rho i(T_n)$, в який потрапило випадкове число *rnd*: $g_{M+1,k} = n,\ rnd \in \rho i(T_n)$.

Потім обчислюється вірогідність $\mathrm{conf}(\chi_k)$ кожного рішення. Для цього рішення χ_k перетворюються в асоціативні правила: $\chi_k \to A_k$. При цьому асоціативне правило A_k формується з ненульових параметрів g_{mk} рішення χ_k, у результаті чого створюється імплікація вигляду (8):

$$\bigcap (p_m \in p_{nm}) \to T. \tag{8}$$

При числових значеннях вихідного параметра *T* консеквент імплікації (8) може бути представлений у вигляді $T \in \Delta T(g_{M+1,k})$, де $\Delta T(g_{M+1,k})$ – терм T_n вихідного параметра *T*, що визначається за номером, представленим в останньому гені $g_{M+1,k}$ рішення χ_k.

Після виконання перетворень вигляду $\chi_k \to A_k$ обчислюються вірогідності $\mathrm{conf}(A_k)$ правил A_k.

Потім у множину A вводяться достовірні асоціативні правила A_k (правила з рівнем вірогідності $\mathrm{conf}(A_k)$, не нижче заданого *minconf*: $\mathrm{conf}(A_k) \ge minconf$). Крім критерію вірогідності (1), для більш детального дослідження асоціативних правил, що витягаються, при стохастичному пошуку також пропонується враховувати інші критерії, які характеризують частоту виконання правил, їхню інформативність та інтерпретовність (зручність для сприйняття людиною):

• підтримка правила $\mathrm{supp}(A_k)$ – дозволяє оцінити частоту виконання правила у вибірці S і може бути визначена за формулами (9) або (10):

$$\mathrm{supp}(A_k) = \mathrm{supp}(P_k \cup T_k) = N(P_k \cup T_k), \tag{9}$$

$$\mathrm{supp}(A_k) = \mathrm{supp}(P_k \cup T_k) = \frac{N(P_k \cup T_k)}{Q}. \tag{10}$$

При цьому формула (9) дозволяє визначити абсолютну величину підтримки як кількість екземплярів вибірки S, для яких виконується правило $A_k : P_k \to T_k$, а формула (10) призначена для визначення відносної величини підтримки як відношення $N(P_k \cup T_k)$ до загальної кількості екземплярів Q у вибірці S;

• загальна підтримка правила $\mathrm{suppG}(A_k)$ – враховує не тільки частоту виконання позитивних умов правил вигляду $P_k \to T_k$, але й негативних умов $\overline{P_k} \to \overline{T_k}$, визначається за формулою (11):

$$\mathrm{suppG}(A_k) = \mathrm{supp}(P_k \cup T_k) + \mathrm{supp}\left(\overline{P_k} \cup \overline{T_k}\right). \tag{11}$$

Чим вище значення даного критерію, тим більш значущим є правило $A_k : P_k \to T_k$ у вибірці S, оскільки велика кількість випадків одночасного виконання (одночасного невиконання) умов антецедента P_k і консеквента T_k свідчить про істотний зв'язок між P_k і T_k;

• загальна вірогідність правила $\mathrm{confG}(A_k)$ – аналогічно критерію $\mathrm{suppG}(A_k)$ враховує частоту виконання як позитивних, так і негативних умов правил $A_k : P_k \to T_k$, визначається за формулою (12):

$$\mathrm{confG}(A_k) = \frac{1}{2}\left(\mathrm{conf}\left(P_k \to T_k\right) + \mathrm{conf}\left(\overline{P_k} \to \overline{T_k}\right)\right) = \frac{1}{2}\left(\frac{\mathrm{supp}(P_k \cup T_k)}{\mathrm{supp}(P_k)} + \frac{\mathrm{supp}\left(\overline{P_k} \cup \overline{T_k}\right)}{\mathrm{supp}\left(\overline{P_k}\right)}\right); \quad (12)$$

• складність правила $VS(A_k)$ – визначається виходячи з відношення кількості ознак (умов) $N(P_k)$, визначених у лівій частині (антецеденті) правила $A_k : P_k \to T_k$, до загальної кількості ознак M у вибірці S (13):

$$VS(A_k) = 1 - \frac{N(P_k)}{M}. \quad (13)$$

Чим вище значення даного критерію, тим правило $A_k : P_k \to T_k$ є більш простим (інтерпретовним) і охоплює більш широкий спектр випадків, тим самим забезпечуючи високе узагальнення даних. Чим більше умов $N(P_k)$ в антецеденті правила (відповідно, чим менше значення критерію $VS(A_k)$), тим воно є більш специфічним;

• показник максимуму вірогідності правила $\mathrm{confG}(A_k)$ та індивідуальних оцінок інформативності ознак $VC(A_k)$ (14):

$$VC(A_k) = \mathrm{confG}(A_k) \sum_{m:\, p_m \in P_k} V_m, \quad (14)$$

де V_m – оцінка індивідуальної значимості m-ї ознаки p_m, що входить в антецедент P_k правила $A_k : P_k \to T_k$. Відзначимо, що окрім оцінки $\mathrm{confG}(A_k)$, у формулі (14) можна використовувати також значення вірогідності $\mathrm{conf}(A_k)$;

• показник максимуму вірогідності правила $\mathrm{confG}(A_k)$ – мінімуму кількості відібраних ознак $VM(A_k)$ (15):

$$VM(A_k) = \frac{1}{N(P_k)}\mathrm{confG}(A_k); \quad (15)$$

• інформативність правила $VI(A_k)$ – критерій, що дозволяє враховувати як вірогідність (значущість) усього правила $A_k : P_k \to T_k$, так і індивідуальну інформативність V_m кожної ознаки p_m, що входить у його антецедент P_k. Критерій $VI(A_k)$ пропонується розраховувати за формулою (16):

$$VI(A_k) = \frac{1}{N(P_k)}\mathrm{confG}(A_k) \sum_{m:\, p_m \in P_k} V_m. \quad (16)$$

Даний критерій забезпечує пошук правил $A_k : P_k \to T_k$ з максимальною вірогідністю $\mathrm{confG}(A_k)$, максимальними оцінками індивідуальних інформативностей V_m ознак і мінімальною кількістю ознак p_m, що входять в антецедент P_k правила.

Оцінку індивідуальної інформативності V_m m-ї ознаки p_m можна визначити як суму індивідуальних інформативностей V_{nm} термів цієї ознаки (17):

$$V_m = \sum_{n=1}^{N_{\text{int}\,m}} V_{nm}\,, \tag{17}$$

де V_{nm} – оцінка інформативності n-го терму Δp_{nm} m-ї ознаки p_m може бути розрахована за формулою (18):

$$V_{nm} = \frac{1}{1+e^{-\sum\limits_{l=1}^{N_{\text{int}}T} \rho(\Delta p_{nm}, T_l)\log\rho(\Delta p_{nm}, T_l)}}\,, \tag{18}$$

де $\rho(\Delta p_{nm}, T_l) = \dfrac{N(p_{nm}, T_l)}{N(p_{nm})}$ – умовна ймовірність того, що значення вихідного параметра T потрапить у l-й інтервал T_l за умови, що m-а ознака p_m потрапить у n-й терм Δp_{nm}; $N(p_{nm}, T_l)$ – кількість екземплярів вибірки S, значення вихідного параметра T яких належать l-му інтервалу діапазону його зміни T_l за умови, що значення їх m-ї ознаки належить n-му інтервалові p_{mn}; $N_{\text{int}}(T)$ – кількість інтервалів, на які розбивається діапазон значень вихідного параметра T.

Запропонована система критеріїв (9)–(16), яка дозволяє враховувати різні характеристики асоціативних правил, що характеризують їхню вірогідність, частоту виконання, інформативність та інтерпретовність, може бути використана для автоматизації аналізу властивостей і порівняння моделей на основі асоціативних правил при вирішенні задач діагностування, розпізнавання образів і неруйнівного контролю якості.

Важливо відзначити, що при виборі правил $A_k : P_k \to T_k$ для внесення у множину $A = \{A_1, A_2, ..., A_{N_A}\}$, що представляє собою базу синтезованих асоціативних правил, можливо використовувати як один заданий критерій (наприклад, інформативність правила $VI(A_k)$), так і набір із декількох критеріїв із запропонованої системи (9)–(16). Крім того, також можливо обчислювати оцінки даних критеріїв і на тестовій вибірці, що дозволить враховувати узагальнюючі характеристики правил, що витягаються.

Після обчислення оцінок якості витягнутих асоціативних правил χ_k ($A_k : P_k \to T_k$), $k = 1, 2, ..., N_\chi$ і внесення кращих з них у множину $A = \{A_1, A_2, ..., A_{N_A}\}$ відбувається перевірка критеріїв завершення стохастичного пошуку: досягнення максимально припустимої кількості правил у множині A ($N_A \geq N_{Amax}$), перевищення максимально припустимої кількості ітерацій N_{It}, неможливість протягом заданої кількості ітерацій побудови правил $A_k : P_k \to T_k$, що характеризуються прийнятними значеннями критеріїв їхнього оцінювання.

У випадку невиконання критеріїв завершення стохастичного пошуку відбувається формування нових N_χ рішень χ_k. Для цього створюється множина $RP^{(i)}$ рішень χ_k, допущених до формування нової множини $R^{(i+1)}$. У множину $RP^{(i)}$ заносяться найбільш пристосовані структури χ_k (у залежності від значень критеріїв оцінювання асоціативних правил A_k) з множини рішень $R^{(i)} = \{\chi_1^{(i)}, \chi_2^{(i)}, ..., \chi_{N\chi}^{(i)}\}$ i-ї ітерації стохастичного пошуку.

Після цього на основі двох рішень $\chi_{parent1} = \{g_{1\,parent1}, g_{2\,parent1}, ..., g_{N_g\,parent1}\} \in RP^{(i)}$ і $\chi_{parent2} = \{g_{1\,parent2}, g_{2\,parent2}, ..., g_{N_g\,parent2}\} \in RP^{(i)}$ створюється нове рішення χ_{child}. Значення параметрів g_{mchild} нащадка χ_{child} визначаються за формулою (19):

$$g_{mchild1} = \begin{cases} g_{mparent1}, rnd \in \left[0; \dfrac{V_{mparent1}}{V_{mparent1} + V_{mparent2}}\right), \\ g_{mparent2}, rnd \in \left[\dfrac{V_{mparent1}}{V_{mparent1} + V_{mparent2}}; 1\right], \end{cases} \tag{19}$$

де $V_{mparent1}$ та $V_{mparent2}$ – інформативність $g_{mparent1}$-го і $g_{mparent2}$-го термів m-ї ознаки відповідно, $rnd = rand[0;1]$.

Наведена формула (19) дозволяє підсилювати ймовірність включення в нове рішення χ_{child} параметрів g_{mk}, що відповідають термам Δp_{nm} ознак з високими оцінками індивідуальної інформативності V_{nm}.

Таким чином, відбувається формування $N_{cross} = \beta N_{\chi}$ рішень вигляду $\chi_k^{(i+1)} = \{g_{1k}^{(i+1)}, g_{2k}^{(i+1)}, ..., g_{N_g k}^{(i+1)}\}$, де β – параметр, що визначає значущість створення нової множини рішень $R^{(i+1)}$ за допомогою запропонованої вище процедури схрещування.

Потім створюється $N_{mutation} = \gamma N_{\chi}$ рішень за допомогою оператора мутації, де γ – параметр, що визначає значущість формування нової множини рішень $R^{(i+1)}$ за допомогою процедури мутації. Для цього з множини $RP^{(i)}$ випадковим чином вибирається рішення χ_{parent}, в якому значення деяких параметрів $g_{mparent}$, що визначають номер терму Δp_{nm} m-ї ознаки p_m у відповідному асоціативному правилі, змінюються, у результаті чого створюється нове рішення χ_{child}. Нові значення змінюваних параметрів g_{mchild} визначаються стохастичним шляхом з урахуванням оцінок індивідуальних значущостей V_{nm} термів Δp_{nm} відповідної ознаки.

Кожному терму Δp_{nm} ($n = 1, 2, ..., N_{\text{int}\,m}$) m-ї ознаки p_m ставиться у відповідність інтервал $gI(\Delta p_{nm}) \in [gI_{\min}(\Delta p_{nm}); gI_{\max}(\Delta p_{nm}))$, де $gI_{\min}(\Delta p_{nm}) = gI_{\max}(\Delta p_{n-1,m})$ – мінімальне значення інтервалу $gI(\Delta p_{nm})$, $gI_{\max}(\Delta p_{nm}) = gI_{\min}(\Delta p_{nm}) + VN_{nm}$ – максимальне значення інтервалу $gI(\Delta p_{nm})$, VN_{nm} – нормоване значення оцінки індивідуальної значущості V_{nm} терму Δp_{nm}, розраховане за формулою (20):

$$VN_{nm} = \frac{\max\limits_{n=1,2,...,N_{\text{int}\,m}} V_{nm} - V_{nm}}{\max\limits_{n=1,2,...,N_{\text{int}\,m}} V_{nm} - \min\limits_{n=1,2,...,N_{\text{int}\,m}} V_{nm}}, \tag{20}$$

$gI_{\min}(\Delta p_{1m}) = 0$ – мінімальне значення в інтервалі $gI(\Delta p_{1m})$ першого терму Δp_{1m} ознаки p_m. Таким чином, чим вище буде значення величини V_{nm} (VN_{nm}), тим ширше буде діапазон значень $gI(\Delta p_{nm})$ терму Δp_{nm}.

Після обчислення границь інтервалів $gI(\Delta p_{nm})$ $(n = 1, 2, ..., N_{intm},\ m = 1, 2, ..., M)$ генерується випадкове число $rnd = rand[0;1)$. Нові значення параметрів g_{mchild} рішення χ_{parent}, обраного для мутації, відповідають номеру n інтервалу $gI(\Delta p_{nm})$, в який попадає число rnd (21):

$$g_{mchild} = n,\ rnd \in gI(\Delta p_{nm}). \tag{21}$$

Отже, чим ширше діапазон $gI(\Delta p_{nm})$, тим більшою є ймовірність терму Δp_{nm} бути включеним у рішення χ_{child}.

У нову множину $R^{(i+1)}$, крім $N_{cross} = \beta N_\chi$ і $N_{mutation} = \gamma N_\chi$ рішень, отриманих за допомогою процедур схрещування та мутації, заноситься також $N_{elite} = \alpha N_\chi$ найбільш пристосованих рішень $\chi_k^{(i)} \in R^{(i)}$, що характеризуються найкращими значеннями критеріїв оцінювання асоціативних правил A_k у популяції $R^{(i)}$, де α – параметр, що визначає значущість включення найкращих рішень у нову множину $R^{(i+1)}$.

Потім виконується обчислення вірогідності $\mathrm{conf}(\chi_k)$ й інших критеріїв оцінювання рішень χ_k з нової популяції $R^{(i+1)}$ з наступним внесенням кращих з рішень χ_k у множину $A = \{A_1, A_2, ..., A_{N_A}\}$, і при невиконанні критеріїв зупинення стохастичного пошуку відбувається створення нової множини рішень $R^{(i+2)}$.

У результаті стохастичного пошуку витягається набір $A = \{A_1, A_2, ..., A_{N_A}\}$ асоціативних правил вигляду $A_k : P_k \to T_k$, що характеризуються прийнятними значеннями заданих критеріїв оцінювання якості правил.

Запропонований стохастичний метод побудови чисельних асоціативних правил передбачає попереднє розбиття значень ознак на інтервали (терми), враховуючи при цьому ширину діапазону значень і частоту попадання ознак у кожний з термів, використовує ймовірнісний підхід для перебору різних сполучень антецедентів і консеквентів асоціативних правил, апріорну інформацію про значущість термів і ознак, що дозволяє обробляти чисельну інформацію при видобуванні асоціативних правил, не здійснювати істотну кількість проходів по заданій базі транзакцій, виявляти правила з високим рівнем вірогідності й інших критеріїв оцінювання їхньої якості.

4. Експерименти та результати

Виконаємо експериментальне дослідження запропонованого стохастичного методу побудови чисельних асоціативних правил. Для цього порівняємо його з відомими методами виявлення чисельних асоціативних правил – FARM [12], FWARM [13], методом синтезу асоціативних правил з урахуванням значущості ознак, запропонованих у [4]. Важливо відзначити, що вирішувалися задачі побудови правил з чисельних баз транзакцій, тому застосування відомих методів (Apriori, SETM та ін.) було ускладнено, оскільки такі методи дозволяють витягати асоціативні правила з бінарних даних. На мові C# було розроблено програмні модулі, що дозволяють витягати асоціативні правила з заданих баз транзакцій $S =< P, T >$ за допомогою запропонованого і відомого методів. За допомогою розроблених програмних модулів вирішувалися задачі прийняття рішень у технічному діагностуванні авіадвигунів.

У процесі випробувань авіадвигунів контролюються параметри, що характеризують якість їхньої роботи при різних режимах [14]. Однак процес випробувань є досить тривалим за часом, вимагає значної кількості випробувань (циклів) кожного виробу при різних режимах, а також істотних матеріальних витрат (палива) на іспити в кожному циклі. При цьому устаткування для проведення випробувань має обмежену пропускну здатність. Тому актуальним є скорочення часу, а також кількості режимів випробувань авіадвигунів, що дозволить скоротити матеріальні витрати на їхнє виготовлення. Для цього необхідно виявити залежності між характеристиками двигунів, що вимірюються або встановлюються у процесі випробувань. Виявлення таких залежностей дозволить скоротити кількість режимів випробувань.

Вибірка даних містить значення характеристик, вимірюваних у процесі випробувань для чотирьох режимів (зліт, номінальний, перший крейсерський, другий крейсерський) [14]: p_1 – кількість обертів турбіни компресора, об/хв; p_2 – температура газу перед турбіною, С; p_3 – витрати газу через турбіну; p_4 – температура на вході у двигун, С; p_5 – кількість ступенів; p_6 – кут установки лопаток вхідного направляючого апарата; p_7 – приведена потужність; p_8 – витрати повітря; p_9 – ступінь стиснення повітря; p_{10} – адіабатичний тиск, мм; p_{11} – p_{14} – прохідні перерізи соплового апарата першого, другого, третього та четвертого ступенів відповідно.

Однак деякі дані внаслідок людського фактора, збоїв і відмовлень вимірювального устаткування й інших причин у вибірці не зафіксовані. Крім того, для ряду авіадвигунів існує інформація про випробування лише при деяких режимах. Наявність пропущених значень у вихідній вибірці S обумовлює доцільність застосування апарата асоціативних правил для виявлення схованих залежностей у даних.

Результати експериментів по дослідженню різних методів побудови асоціативних правил при вирішенні задачі виявлення схованих залежностей між параметрами авіадвигунів при різних режимах випробувань наведено в табл. 1 (вихідна вибірка містила інформацію про 484 вироби). Як критерій оцінювання якості асоціативних правил при дослідженні стохастичного методу побудови чисельних асоціативних правил використовувалася інформативність правила $VI(A_k)$, оскільки цей критерій дозволяє враховувати вірогідність правила й індивідуальну інформативність V_m кожної ознаки p_m, що входить у його антецедент P_k.

Таблиця 1. Результати експериментів по побудові асоціативних правил

Метод	Supp, %	Conf, %	ConfG, %	N_A	$N(P_k)$	*VS*	*VC*	*VM*	*VI*
FARM [12]	6,2	82,3	73,7	121	6,78	0,52	3,05	10,87	6,63
FWARM [13]	5,4	87,7	77,5	87	6,32	0,55	3,09	12,26	7,73
Метод синтезу асоціативних правил з урахуванням значимості ознак (МУЗП) [4]	4,7	91,2	82,1	82	6,46	0,54	3,92	12,71	9,40
Стохастичний метод побудови чисельних асоціативних правил	4,2	90,1	89,2	133	5,06	0,64	3,61	17,63	14,10

У табл. 1 наведено середні значення параметрів supp, conf, conf, $N(P_k)$, *VS*, *VC*, *VM*, *VI*, що характеризують якість витягнутих асоціативних правил. У результаті досліджень виявлено залежності $A_k : P_k \to T_k$ між різними параметрами виробів, які описують якість їхньої роботи при різних режимах, що дозволило дати рекомендації щодо скорочення кількості випробувань виробів і, отже, зниження матеріальних витрат на їхнє виготовлення.

Як видно з табл. 1, значення середньої підтримки supp, виявлених за допомогою запропонованого методу асоціативних правил, трохи нижче $\text{supp} = 4{,}2$, ніж у наборів асоціативних правил, витягнутих відомими методами FARM [12] ($\text{supp} = 6{,}2$), FWARM [13] ($\text{supp} = 5{,}4$), МУЗП [4] ($\text{supp} = 4{,}7$), оскільки запропонований метод дозволив, крім достовірних правил, що часто зустрічаються, також виявити закономірності на основі наборів, що рідко зустрічаються. Про це також свідчить більша кількість витягнутих правил N_A (у

запропонованого методу $N_A = 133$, в інших методах кількість витягнутих правил є меншою: $N_A = 121$ для FARM, $N_A = 87$ для FWARM, $N_A = 82$ для МУЗП).

Значення середньої вірогідності conf виявлених правил на основі розробленого стохастичного методу побудови чисельних асоціативних правил ($\text{conf} = 90{,}1$) вище, ніж у методів FARM [12] ($conf = 82{,}3$) і FWARM [13] ($conf = 87{,}7$), це свідчить про те, що запропонований метод дозволяє виявляти більш достовірні правила (це досягається за рахунок використання стохастичного перебору різних сполучень антецедентів і консеквентів асоціативних правил, а також врахування апріорної інформації про значущість термів і ознак). У порівнянні з методом МУЗП значення критерію conf трохи нижче ($conf = 90{,}1$ і $conf = 91{,}2$ відповідно), оскільки при проведенні експериментів як критерій оцінювання асоціативних правил у запропонованому стохастичному методі побудови чисельних асоціативних правил використовувався критерій інформативності правил $VI(A_k)$, що враховує не тільки вірогідність conf, але й інші характеристики.

Запропонований стохастичний метод дозволив синтезувати базу $A = \{A_1, A_2, ..., A_{N_A}\}$ асоціативних правил $A_k : P_k \to T_k$, яка характеризується більш високою середньою загальною вірогідністю правил ($\text{confG} = 89{,}2$ у порівнянні з $\text{confG} = 73{,}7$, $\text{confG} = 77{,}5$ і $\text{confG} = 82{,}1$ для FARM, FWARM і МУЗП відповідно), що враховує частоту виконання не тільки позитивних умов $P_k \to T_k$, але й негативних умов $\overline{P_k} \to \overline{T_k}$ виконання правил.

Більш прийнятні значення критеріїв $N(P_k)$, *VS*, *VC*, *VM*, *VI* (наприклад, середня складність $VS(A_k)$ витягнутих за допомогою розробленого методу правил склала $VS = 0{,}64$ в порівнянні з $VS = 0{,}52$, $VS = 0{,}55$ і $VS = 0{,}54$ для FARM, FWARM і МУЗП, відповідно) у запропонованого методу обумовлені також застосуванням інформативності $VI(A_k)$ (критерії $N(P_k)$, *VS*, *VC*, *VM* і *VI* є взаємозалежними) як критерію оцінювання асоціативних правил, що витягаються. Це дозволило забезпечити побудову більшої кількості N_A асоціативних правил $A_k : P_k \to T_k$, що є більш простими та інтерпретовними (такими, що характеризуються меншою кількістю умов $N(P_k)$ в антецеденті P_k), а також більш достовірними та інформативними (володіють більш прийнятними значеннями критеріїв conf, *VC*, *VM*, *VI*) у порівнянні з правилами, виявленими за допомогою відомих методів.

5. Висновки

У роботі вирішено актуальну задачу автоматизації побудови чисельних асоціативних правил.

Наукова новизна роботи полягає у тому, що запропоновано стохастичний метод побудови чисельних асоціативних правил, який передбачає попереднє розбиття значень ознак на інтервали (терми), враховує при цьому ширину діапазону значень і частоту попадання ознак у кожний з термів, використовує ймовірнісний підхід для перебору різних сполучень антецедентів і консеквентів асоціативних правил, використовує апріорну інформацію про значущість термів і ознак, що дозволяє обробляти чисельну інформацію при побудові асоціативних правил, не здійснювати істотну кількість проходів по заданій базі транзакцій, виявляти правила з високим рівнем вірогідності й інших критеріїв оцінювання їхньої якості.

Запропоновано систему критеріїв, яка дозволяє враховувати різні характеристики асоціативних правил, що характеризують їхню вірогідність, частоту виконання, інформативність та інтерпретовність. Розроблена система критеріїв може бути використана для ав-

томатизації аналізу властивостей і порівняння моделей на основі асоціативних правил при вирішенні задач діагностування, розпізнавання образів і неруйнівного контролю якості.

Практична цінність отриманих результатів полягає в тому, що мовою C# розроблено програмні модулі, які дозволяють будувати асоціативні правила з заданих баз транзакцій за допомогою запропонованого і відомого методів. За допомогою розроблених програмних модулів вирішено практичну задачі прийняття рішень у технічному діагностуванні авіадвигунів.

Роботу виконано в рамках держбюджетної науково-дослідної теми Запорізького національного технічного університету «Інтелектуальні інформаційні технології автоматизації проектування, моделювання, керування та діагностування виробничих процесів і систем» (номер державної реєстрації 0112U005350) за підтримки міжнародного проекту “Centers of Excellence for young RESearchers” (CERES) програми “Tempus” Європейської Комісії (реєстраційний номер 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES).

СПИСОК ЛІТЕРАТУРИ

1. Zhang C. Association rule mining: models and algorithms / C. Zhang, S. Zhang. – Berlin: Springer-Verlag, 2002. – 238 p.
2. Zhao Y. Post-mining of association rules: techniques for effective knowledge extraction / Y. Zhao, C. Zhang, L. Cao. – New York: Information Science Reference, 2009. – 372 p.
3. Gkoulalas-Divanis A. Association Rule Hiding for Data Mining / A. Gkoulalas-Divanis, V.S. Verykios. – New York: Springer-Verlag, 2010. – 150 p.
4. Олейник А.А. Синтез диагностических и распознающих моделей на основе гибридных нейронечётких технологий вычислительного интеллекта / Олейник А.А., Зайко Т.А., Субботин С.А.; под ред. С.А. Субботина. – Харьков: ООО “Компания Смит”, 2014. – 284 с.
5. Adamo J.-M. Data mining for association rules and sequential patterns: sequential and parallel algorithms / Adamo J.-M. – New York: Springer-Verlag, 2001. – 259 p.
6. Koh Y.S. Rare Association Rule Mining and Knowledge Discovery / Y.S. Koh, N. Rountree. – New York: Information Science Reference, 2009. – 320 p.
7. Encyclopedia of artificial intelligence / Eds. J.R. Dopico, J.D. de la Calle, A.P. Sierra. – New York: Information Science Reference, 2009. – Vol. 1–3. – 1677 p.
8. Encyclopedia of machine learning / Eds. C. Sammut, G.I. Webb. – New York: Springer, 2011. – 1031 p.
9. Intelligent fault diagnosis and prognosis for engineering systems / G. Vachtsevanos, F. Lewis, M. Roemer [et al.]. – New Jersey: John Wiley & Sons, 2006. – 434 p.
10. Bishop C.M. Pattern recognition and machine learning / Bishop C.M. – New York: Springer, 2006. – 738 p.
11. Abonyi J. Cluster analysis for data mining and system identification / J. Abonyi, B. Feil. – Basel: Birkhäuser, 2007. – 303 p.
12. Dubois D.A Systematic Approach to the Assessment of Fuzzy Association Rules / D. Dubois, E. Hullermeier, H. Prade // Data Mining and Knowledge Discovery. – 2006. – Vol. 13. – P. 167 – 192.
13. Khan M.S. Weighted Association Rule Mining from Binary and Fuzzy Data / M.S. Khan, M. Muyeba, F. Coenen // Lecture Notes in Computer Science. – 2008. – Vol. 5077. – P. 200 – 212.
14. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиационных двигателей / [А.В. Богуслаев, Ал.А. Олейник, Ан.А. Олейник и др.]; под ред. Д.В. Павленко, С.А. Субботина. – Запорожье: ОАО "Мотор Сич", 2009. – 468 с.

Стаття надійшла до редакції 24.11.2014