НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

УДК 004.93

Зайко Т. А.[1], Олійник А. О.[2], Субботін С. О.[3]

[1]Аспірантка, Запорізький національний технічний університет, Україна
[2]Канд. техн. наук, доцент, Запорізький національний технічний університет, Україна, E-mail: olejnikaa@gmail.com
[3]Д-р техн. наук, професор, Запорізький національний технічний університет, Україна

СКОРОЧЕННЯ РОЗМІРНОСТІ НАВЧАЛЬНОЇ ВИБІРКИ НА ОСНОВІ АСОЦІАТИВНИХ ПРАВИЛ

Розглянуто задачу скорочення навчальної вибірки. Розроблено метод редукції даних на основі асоціативних правил. Створено програмне забезпечення на основі запропонованого методу. Проведено експерименти з вирішення практичних задач, що дозволило дослідити ефективність запропонованого методу.

Ключові слова: асоціативне правило, вірогідність, модель, підтримка, скорочення, навчальна вибірка, терм.

ВСТУП

Вибірки даних, що використовуються для побудови діагностичних моделей складних об'єктів і процесів, містять, як правило, надлишкову інформацію [1, 2], представлену ознаками, що не впливають на вихідний параметр, і множиною подібних екземплярів. Використання надлишкових даних при синтезі діагностичних моделей може привести до побудови моделей, які характеризуються низькими узагальнювальними властивостями, а також високою структурною та параметричною складністю, що призводить до збільшення витрат пам'яті ЕОМ на зберігання моделей і збільшення часу обчислень при обробці великих масивів даних. Крім того, такі моделі, як правило, характеризуються низьким рівнем інтерпретабельності, а також не завжди забезпечують прийнятну точність розпізнавання, що ускладнює або унеможливлює їхнє застосування на практиці. Отже, перед здійсненням синтезу діагностичних моделей доцільним є скорочення навчальної вибірки шляхом виключення з неї надлишкової інформації.

Відомі методи редукції даних [1–5], як правило, призначені або для відбору ознак, або для відбору екземплярів і часто не враховують взаємозв'язки комбінацій деяких значень ознак, які також можуть бути виключені з вихідної вибірки. Тому актуальною є розробка нового методу скорочення навчальної вибірки, який дозволяє виконувати редукцію ознак, екземплярів, термів ознак і формувати множину даних з меншою кількістю елементів у порівнянні з вихідною вибіркою, а також будувати на її основі діагностичні моделі з високими значеннями показників узагальнення й інтерпретабельності.

Для редукції навчальної вибірки в цій роботі пропонується використовувати асоціативні правила [6–10], оскільки видобування таких правил з вибірок даних дозволяє суттєво скорочувати обсяги інформації та виконувати узагальнення даних, перетворювати значення ознак у деякі діапазони значень, оцінювати ступінь впливу ознак на вихідний параметр, а також рівень їх взаємозв'язку між собою, у т.ч. взаємозв'язки деяких значень ознак.

Метою роботи є створення методу скорочення розмірності навчальної вибірки на основі асоціативних правил.

1 ПОСТАНОВКА ЗАДАЧІ СКОРОЧЕННЯ РОЗМІРНОСТІ НАВЧАЛЬНОЇ ВИБІРКИ

Нехай задана навчальна вибірка D (1):

$$D = \left\{T_1, T_2, ..., T_{N_D}\right\}, \quad (1)$$

 DOI 10.15588/1607-3274-2014-1-15

у якій кожний елемент T_j, $j=1,2,...,N_D$ містить інформацію про деякі об'єкти або процеси, де $N_D=|D|$ – кількість екземплярів (елементів) у наборі даних D.

Елементи T_j являють собою множину значень вигляду (2):

$$T_j=\{\tau_{1j},\tau_{2j},...,\tau_{N_Ij},y_j\}, \quad (2)$$

де $\tau_{aj}=[\tau_{aj\min};\tau_{aj\max}]$ – значення a-ї ознаки τ_a для елементу T_j; τ_a – a-та ознака множини $I=\{\tau_1,\tau_2,...,\tau_{N_I}\}$, $a=1,2,...,N_I$; I – множина ознак, якими описуються елементи $T_j, j=1,2,...,N_D$ набору даних D; $N_I=|I|$ – кількість ознак у вибірці D; $\tau_{aj\min}$ та $\tau_{aj\max}$ – мінімальне та максимальне значення з діапазону можливих значень ознаки τ_a; y_j – значення вихідного параметра для елементу T_j.

Тоді задача скорочення розмірності навчальної вибірки $D\to D'$ полягає в зменшенні кількості її екземплярів $N'_D<N_D$ та ознак $N'_I<N_I$, що їх описують, зі збереженням можливості побудови діагностичних моделей з прийнятними здатностями до апроксимації досліджуваних залежностей.

2 МЕТОД СКОРОЧЕННЯ РОЗМІРНОСТІ НАВЧАЛЬНОЇ ВИБІРКИ НА ОСНОВІ АСОЦІАТИВНИХ ПРАВИЛ

Для виявлення й усунення надлишкової інформації у вибірці пропонується метод скорочення розмірності навчальної вибірки, що реалізує послідовно етапи редукції екземплярів, редукції ознак і скорочення надлишкових термів.

У розробленому методі скорочення розмірності навчальної вибірки для редукції даних пропонується виявляти асоціативні правила. Інформація про цікавість виявлених правил використовується для оцінювання ступеню впливу ознак на вихідний параметр, а також взаємозв'язків деяких значень ознак між собою.

На початковому етапі для заданої вибірки D виконується редукція її екземплярів. Для цього дискретизуються значення числових ознак (діапазон значень $\Delta_a=[\tau_{a\min};\tau_{a\max}]$ кожної ознаки τ_a розбивається на $N_{\text{int}.a}$ інтервалів). Величина $N_{\text{int}.a}$ може задаватися користувачем як параметр методу та бути єдиною для всіх ознак вибірки D. Крім того, кількість $N_{\text{int}.a}$ інтервалів дискретизації a-ї ознаки τ_a може бути визначена шляхом кластеризації вибірки D і проектування границь кластерів на координатні осі в просторі ознак.

Після дискретизації виконується перетворення $D\to D'_1$, у результаті якого значення вихідних ознак τ_a заміняються номерами інтервалів значень ознак, виділених у процесі дискретизації (3):

$$\tau'_{aj}=n(\tau_{aj}), \quad (3)$$

де τ_{aj} та τ'_{aj} – значення a-ї ознаки для j-го екземпляру у вибірках D та D'_1, відповідно; $n(\tau_{aj})$ – номер інтервалу значень ознаки τ_a, у який попадає її значення τ_{aj} для j-го екземпляру.

Отримані в результаті перетворення $D\to D'_1$ екземпляри T'_j та T'_k з однаковими значеннями ознак τ'_{aj} та τ'_{ak}, $a=1,2,...,N_I$ вважаються еквівалентними й надлишковими. Тому у вибірці D'_1 послідовно для кожних двох еквівалентних екземплярів T'_j й T'_k слід залишити один екземпляр T'_j, а інший – виключити (4):

$$D'_1=D'_1\setminus T'_k. \quad (4)$$

Після виконання етапу редукції екземплярів відбувається виявлення неінформативних ознак з наступним їх виключенням з вибірки. Для редукції ознак τ_a з вибірки D'_1 будемо витягати асоціативні правила $\text{АП}_l\in\text{БП}$ (БП – база правил), оцінювати їх цікавість та цікавість кожного терму ознак, на основі чого будемо робити висновок про інформативність кожної ознаки. Для цього спочатку видобуваються чисельні асоціативні правила $\text{АП}_l: X_l\to Y_l$ [8, 11], потім виконується оцінювання цікавості $I_{\text{АП}_l}$ кожного з виявлених правил. У якості оцінок цікавості правил можна використовувати критерії (5)–(9) [6–10]:

$$I_{\text{АП}_l}=\text{supp}(X_l\to Y_l)+\text{supp}(\overline{X_l}\to\overline{Y_l}), \quad (5)$$

$$I_{\text{АП}_l}=\frac{\text{supp}(X_l\to Y_l)}{\text{supp}(X_l)\text{supp}(Y_l)}, \quad (6)$$

$$I_{\text{АП}_l}=\frac{\text{conf}(X_l\to Y_l)}{\text{conf}(\overline{X_l}\to Y_l)}, \quad (7)$$

$$I_{\text{АП}_l}=\frac{\text{supp}(X_l\to Y_l)\text{supp}(\overline{X_l}\to\overline{Y_l})}{\text{supp}(X_l\to\overline{Y_l})\text{supp}(\overline{X_l}\to Y_l)}, \quad (8)$$

$$I_{\text{АП}_l}=\text{supp}(X_l\to Y_l)-\text{supp}(X_l)\text{supp}(Y_l), \quad (9)$$

де $\text{supp}(A)$ – підтримка множини A, обчислена як відношення кількості елементів T_j, що містять A, до загальної кількості екземплярів N_D у наборі даних D; $\text{conf}(A)$ – вірогідність множини A, що розраховується як відношення підтримки імплікації A $(X\to Y)$ до підтримки її лівої частини X.

Використовуючи інформацію про цікавості $I_{\text{АП}_l}$ витягнутих асоціативних правил, виконується оцінювання цікавості термів $\Delta\tau_{ak}$, $k=1,2,...,N_{\text{int}.a}$ кожної ознаки τ_a, $a=1,2,...,N_I$. Цікавість термів $\Delta\tau_{ak}$ пропонується визначати за однією з наступних формул (10)–(12):

$$I_{\Delta\tau_{ak}}=\frac{1}{N_{\Delta\tau_{ak}}}\sum_{\substack{l:\text{АП}_l\in\text{БП},\\ \Delta\tau_{ak}\in\text{АП}_l}}I_{\text{АП}_l}, \quad (10)$$

$$I_{\Delta\tau_{ak}} = \min_{\substack{l:\text{АП}_l\in\text{БП},\\ \Delta\tau_{ak}\in\text{АП}_l}} \{I_{\text{АП}_l}\}, \quad (11)$$

$$I_{\Delta\tau_{ak}} = \max_{\substack{l:\text{АП}_l\in\text{БП},\\ \Delta\tau_{ak}\in\text{АП}_l}} \{I_{\text{АП}_l}\}, \quad (12)$$

де $N_{\Delta\tau_{ak}}$ – кількість асоціативних правил $\text{АП}_l \in \text{БП}$, що містять терм $\Delta\tau_{ak}$: $\Delta\tau_{ak} \in \text{АП}_l$.

Інформативність I_a ознак τ_a будемо оцінювати, виходячи з оцінок цікавостей термів, що входять у відповідну ознаку (13)–(15):

$$I_a = \frac{1}{N_{\text{int}.a}} \sum_{k=1}^{N_{\text{int}.a}} I_{\Delta\tau_{ak}}, \quad (13)$$

$$I_a = \max_{k=1,2,...,N_{\text{int}.a}} \{I_{\Delta\tau_{ak}}\}, \quad (14)$$

$$I_a = \min_{k=1,2,...,N_{\text{int}.a}} \{I_{\Delta\tau_{ak}}\}. \quad (15)$$

З метою приведення значень оцінок інформативності ознак до одного інтервалу $[0;1]$ виконаємо їх нормування (16):

$$I_a = \frac{I_a - \min\limits_{a=1,2,...,N_I}\{I_a\}}{\max\limits_{a=1,2,...,N_I}\{I_a\} - \min\limits_{a=1,2,...,N_I}\{I_a\}}. \quad (16)$$

Ознаки τ_a з низькими значеннями інформативності $I_a < I_{\text{р}}$ ($I_{\text{р}}$ – мінімально прийнятне значення інформативності) виключаються з вибірки D'_1. У результаті виключення з вибірки D'_1 неінформативних ознак можливою є поява надлишкових екземплярів, що містять однакові значення ознак і вихідного параметру. Такі екземпляри також виключаються. У результаті видалення неінформативних ознак і надлишкових екземплярів виконується перетворення $D'_1 \to D'_2$ й скорочення розмірності навчальної вибірки.

З метою виконання етапу скорочення надлишкових термів з вибірки D'_2 витягаються асоціативні правила та виявляються взаємозв'язки між різними інтервалами $\Delta\tau_{ak}$ й $\Delta\tau_{bm}$ ознак.

У результаті видобування асоціативних правил з вибірки D'_2 синтезується база правил БП_2 виду $\text{АП}_l : X_l \to Y_l$ з рівнем вірогідності $\text{conf}(X_l \to Y_l)$, не нижче мінімально прийнятного minconfidence.

Тому з транзакцій (екземплярів) T'_{2j} вибірки D'_2 можна виключити терми $\Delta\tau_{ak} \in X_l$ при наявності в цих ж транзакціях термів $\Delta\tau_{bm} \in Y_l$, що входять у консеквенти Y_l правил АП_l бази БП_2 (17):

$$T'_{3j} = T'_{2j} \setminus \bigcup_{\substack{\Delta\tau_{ak}\in X_l,\\ \exists(\Delta\tau_{bm}\subset T'_{2j})\in Y_l,\\ (X_l\to Y_l)\in\text{БП}_2}} (\tau_a \in \Delta\tau_{ak}). \quad (17)$$

Шляхом виключення надлишкових термів з вибірки D'_2 виконується перетворення $D'_2 \to D'_3$ та формування вибірки D'_3 скороченої розмірності. У такий спосіб отримане розбиття простору ознак D'_3 містить суттєво меншу кількість елементів $\Delta\tau_{ak}$ у порівнянні з вихідною вибіркою D, характеризується більш високими узагальнюючими властивостями й дозволяє понизити структурну та параметричну складність синтезованих діагностичних моделей.

Запропонований метод скорочення розмірності навчальної вибірки на основі асоціативних правил передбачає виконання етапів редукції екземплярів, ознак і надлишкових термів, для оцінювання інформативності ознак використовує інформацію про витягнуті асоціативні правила й дозволяє формувати розбиття простору ознак з меншою кількістю екземплярів у порівнянні з вихідною вибіркою, що у свою чергу дозволяє синтезувати більш прості та зручні для сприйняття діагностичні моделі.

3 АНАЛІЗ ОБЧИСЛЮВАЛЬНОЇ СКЛАДНОСТІ

Обчислювальну складність методу скорочення розмірності навчальної вибірки визначимо як $O_\Sigma = O\left(\sum_{i=1}^{3} O_i\right)$, де кожний доданок O_i характеризує обчислювальну складність відповідного i-го етапу методу, а $O()$ – оператор нотації Ландау «о велике».

На етапі редукції екземплярів виконується дискретизація $N_I = |I|$ ознак з наступним пошуком для кожного елемента T_j ($j = 1,2,...,N_D$) еквівалентних екземплярів (таких, у яких значення відповідних ознак належить однаковим інтервалам $\tau'_{aj}, a = 1,2,...,|I|$). Отже, складність першого етапу може бути визначена в такий спосіб (18):

$$O_1 = O(N_D|I|). \quad (18)$$

Етап редукції ознак передбачає видобування чисельних асоціативних правил з наступним використанням відповідної інформації для виключення неінформативних ознак. На обчислення оцінок інформативності кожної з $|I|$ ознак буде потрібно $O_{\text{и}}(|I|)$ елементарних операцій. Оскільки обчислювальна складність виявлення чисельних асоціативних правил може бути оцінена як $O_{\text{АП}}\left(|I|\cdot N_D \log_2(N_D) + |I|^2\right)$, величину O_2 визначимо за формулою (19):

$$O_2 = O_{\text{и}}(|I|) + O_{\text{АП}}\left(|I|\cdot N_D \log_2(N_D) + |I|^2\right) = O\left(|I|\cdot N_D \log_2(N_D) + |I|^2\right). \quad (19)$$

Для виключення надлишкових термів необхідно проаналізувати кожний з N_D екземплярів на наявність у ньому термів $\Delta\tau_{ak}, a = 1,2,...,|I|$, які можуть бути виключені.

Враховуючи також необхідність видобування асоціативних правил на цьому етапі, одержуємо наступну оцінку обчислювальної складності (20):

$$O_3 = O_{\text{т}}\left(N_D|I|\right) + O_{\text{АП}}\left(|I|\cdot N_D \log_2\left(N_D\right) + |I|^2\right) = \\ = O\left(|I|\cdot N_D \log_2\left(N_D\right) + |I|^2\right). \quad (20)$$

Отже, загальна оцінка обчислювальної складності методу скорочення розмірності навчальної вибірки може бути визначена за формулою (21):

$$O_\Sigma = O_1\left(N_D|I|\right) + O_2\left(|I|\cdot N_D \log_2\left(N_D\right) + |I|^2\right) + \\ + O_3\left(|I|\cdot N_D \log_2\left(N_D\right) + |I|^2\right) = O\left(|I|\cdot N_D \log_2\left(N_D\right) + |I|^2\right). \quad (21)$$

Як видно, оцінка O_Σ запропонованого методу є пропорційною до величини $N_D \log_2\left(N_D\right)$ та квадратично залежить від кількості ознак у вибірці D. Це дозволяє зробити висновок про те, що розроблений метод скорочення розмірності навчальної вибірки на основі асоціативних правил є обчислювально ефективним.

4 ЕКСПЕРИМЕНТИ Й РЕЗУЛЬТАТИ

Для виконання експериментального дослідження запропонованого методу скорочення розмірності навчальної вибірки на основі асоціативних правил він був програмно реалізований мовою C#. Навчальна вибірка для проведення експериментів містила інформацію про характеристики сировини й параметри технологічного процесу виготовлення кондитерської продукції для 3284 партій виробів (спостережень), що описуються за допомогою 43 ознак. Далі ця вибірка скорочувалася шляхом застосування запропонованого методу, а також різних методів скорочення навчальних множин (методи відбору ознак [2, 4, 5, 12] і методи відбору екземплярів [1–3, 13, 14]).

Для порівняння розробленого методу з аналогами використовувалися критерії, що враховують ступінь скорочення навчальної вибірки (зменшення кількості ознак, екземплярів), а також характеристики моделі, побудованої на основі скороченої вибірки:

– кількість екземплярів у вибірці після скорочення N'_D;

– коефіцієнт скорочення кількості екземплярів (22):

$$\alpha_{ND} = \frac{N'_D}{N_D}; \quad (22)$$

– кількість ознак у вибірці після скорочення $|I'|$;

– коефіцієнт скорочення кількості ознак (23):

$$\alpha_{|I|} = \frac{|I'|}{|I|}; \quad (23)$$

– коефіцієнт скорочення розмірності вибірки (24):

$$\alpha_D = \frac{N'_D|I'|}{N_D|I|} = \alpha_{ND}\alpha_{|I|}; \quad (24)$$

– помилка моделі, побудованої на основі навчальної вибірки ε_o;

– помилка моделі, побудованої на основі тестової вибірки ε_t;

– структурна складність синтезованої моделі β_s;

– параметрична складність синтезованої моделі β_p.

У якості моделі, синтезованої на основі вихідної та скорочених вибірок, була обрана нейро-нечітка мережа Мамдані [5, 15], яка будувалася шляхом відображення множини екземплярів у правила, використовувалася П-подібна функція належності [5, 10, 15]. Структурна складність β_s такої моделі визначалася як кількість використовуваних нейроелементів, параметрична β_p – як загальна кількість параметрів моделі (вагових коефіцієнтів, параметрів функції належності).

Результати експериментів зі скорочення навчальної вибірки для синтезу діагностичної моделі якості кондитерської продукції наведено в табл. 1.

Таблиця 1. Результати скорочення навчальної вибірки

Метод	N'_D	α_{ND}	$\|I'\|$	$\alpha_{\|I\|}$	α_D	ε_o	ε_t	β_s	β_p
1. Вихідна вибірка (скорочення вибірки не виконувалося)	3284	1	43	1	1	0	0,13	3510	4369
2. Методи відбору ознак									
2.1. Відбір з додаванням ознак [4, 5, 12]	3284	1	34	0,79	0,79	0,037	0,054	3465	4144
2.2. Відбір з видаленням ознак [4, 5, 12]	3284	1	31	0,72	0,72	0,042	0,061	3450	4069
2.3. Еволюційний метод відбору ознак [12]	3284	1	25	0,58	0,58	0,036	0,045	3420	3919
3. Методи відбору екземплярів									
3.1. Випадковий відбір [1, 2]	1642	0,50	43	1	0,50	0,041	0,059	1868	2727
3.2. Метод на основі переборного пошуку [13]	1193	0,36	43	1	0,36	0,036	0,056	1419	2278
3.3. Метод на основі еволюційного пошуку [14]	981	0,30	43	1	0,30	0,031	0,046	1207	2066
4. Метод скорочення навчальної вибірки на основі асоціативних правил	956	0,29	27	0,63	0,18	0,035	0,044	1102	1641

Як видно з табл. 1, нейро-нечітка мережа, побудована на основі вихідної (не скороченої) вибірки, характеризується високими значеннями показників структурної та параметричної складності ($\beta_s = 3510$, $\beta_p = 4369$), оскільки в мережі міститься велика кількість правил і термів. Це, з одного боку, дозволяє досягнути нульової помилки моделі, визначеної на основі навчальної вибірки, а з іншого боку, не дозволяє забезпечити прийнятне значення помилки, розрахованої на основі тестової вибірки ($\varepsilon_t = 0{,}13$). Крім того, високі значення критеріїв β_s, β_p та ε_t характеризують таку нейро-нечітку мережу як модель з низькими показниками інтерпретабельності й узагальнення.

Використання методів відбору ознак [4, 5, 12] дозволило незначно скоротити структурну та параметричну складність моделі, підвищивши її інтерпретабельність, і збільшити її узагальнюючі здатності (значення критерію ε_t для різних методів становило 0,054, 0,061, і 0,045, що є прийнятним для даної предметної області).

Методи відбору екземплярів [1, 2, 13, 14] дозволили зменшити навчальну вибірку на 50–70 % (значення коефіцієнту α_D для різних методів склало 0,3–0,5). Скорочення екземплярів забезпечило зменшення кількості правил у синтезованої нейро-нечіткої мережі, що у свою чергу дозволило скоротити структурну й параметричну складність (для еволюційного методу скорочення екземплярів $\beta_s = 1207$, $\beta_p = 2066$), забезпечивши при цьому прийнятні значення показника (від 0,046 до 0,059).

Запропонований метод скорочення навчальної вибірки на основі асоціативних правил дозволив суттєво зменшити розмірність навчальної вибірки ($\alpha_D = 0{,}018$), що забезпечується за рахунок виконання етапів редукції екземплярів, редукції ознак і скорочення надлишкових термів. Синтезована на основі вибірки, скороченої за допомогою розробленого методу, нейро-нечітка модель характеризується прийнятними значеннями показників ε_o та ε_t ($\varepsilon_o = 0{,}035$, $\varepsilon_t = 0{,}044$), а отже, і прийнятними апроксимаційними й узагальнюючими властивостями. Низькі значення показників β_s та β_p ($\beta_s = 1102$, $\beta_p = 1641$) досягаються за рахунок суттєвого скорочення кількості правил і ознак у синтезованій нейро-нечіткій мережі. Такі значення критеріїв β_s і β_p показують, що моделі, побудовані на основі вибірок, скорочених за допомогою запропонованого методу, є більш простими та зручними для сприйняття (тобто є більш інтерпретабельними).

Таким чином, результати експериментів показали, що запропонований метод скорочення розмірності навчальної вибірки на основі асоціативних правил дозволяє формувати множину даних з меншою кількістю елементів у порівнянні з вихідною вибіркою, а також будувати на її основі діагностичні моделі з високими значеннями показників узагальнення й інтерпретабельності.

ВИСНОВКИ

У роботі вирішено актуальну задачу редукції навчальних вибірок для побудови діагностичних моделей.

Наукова новизна роботи полягає в тому, що запропоновано метод скорочення розмірності навчальної вибірки на основі асоціативних правил, який передбачає виконання етапів редукції екземплярів, ознак і надлишкових термів, для оцінювання інформативності ознак використовує інформацію про витягнуті асоціативні правила та дозволяє формувати розбиття простору ознак з меншою кількістю екземплярів у порівнянні з вихідною вибіркою, що, у свою чергу, дозволяє синтезувати більш прості та зручні для сприйняття діагностичні моделі.

Практична цінність отриманих результатів полягає в тому, що на основі запропонованого методу вирішено практичну задачу скорочення навчальної вибірки для синтезу діагностичної моделі якості кондитерської продукції.

Роботу виконано в рамках держбюджетної науково-дослідної теми Запорізького національного технічного університету «Інтелектуальні інформаційні технології автоматизації проектування, моделювання, керування й діагностування виробничих процесів і систем» (номер державної реєстрації 0112U005350) за підтримки проекту «Centers of Excellence for young Researchers (CERES)» (№544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES) програми «Темпус» Європейської Комісії.

СПИСОК ЛІТЕРАТУРИ

1. *Chaudhuri, A.* Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p.
2. Encyclopedia of survey research methods / ed. P. J. Lavrakas. – Thousand Oaks : Sage Publications, 2008. – Vol. 1–2. – 968 p.
3. *Кокрен, У.* Методы выборочного исследования / У. Кокрен ; пер. с англ. И. М. Сонина ; под ред. А. Г. Волкова, Н. К. Дружинина. – М. : Статистика, 1976. – 440 с.
4. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken: John Wiley & Sons, 2008. – 339 p.
5. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов : монография / [С. А. Субботин, Ан. А. Олейник, Е. А. Гофман, С. А. Зайцев, Ал. А. Олейник ; под ред. С. А. Субботина]. – Харьков : ООО «Компания Смит», 2012. – 317 с.
6. *Gkoulalas-Divanis, A.* Association Rule Hiding for Data Mining / A. Gkoulalas-Divanis,V. S. Verykios. – New York : Springer-Verlag, 2010. – 150 p.
7. *Koh, Y. S.* Rare Association Rule Mining and Knowledge Discovery / Y. S. Koh, N. Rountree. – New York : Information Science Reference, 2009. – 320 p.
8. *Zhang, C.* Association rule mining: models and algorithms / C. Zhang, S. Zhang. – Berlin : Springer-Verlag. – 2002. – 238 p.
9. *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.
10. 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.
11. *Зайко, Т. А.* Извлечение численных ассоциативных правил с учетом значимости признаков / Т. А. Зайко, А. А. Олейник, С. А. Субботин // Східно-Європейський журнал передових технологій. – 2013. – № 5/4 (65). – С. 28–34.

12. *Олійник, А. О.* Інтелектуальний аналіз даних / А. О. Олійник, С. О. Субботін, О. О. Олійник : навчальний посібник. – Запоріжжя : ЗНТУ, 2012. – 271 с.
13. *Субботин, С. А.* Критерии индивидуальной информативности и методы отбора экземпляров для построения диагностических и распознающих моделей / С. А. Субботин // Біоніка інтелекту. – 2010. – № 1. – С. 38–42.
14. *Субботин, С. А.* Методы формирования выборок для построения диагностических моделей по прецедентам / С. А. Субботин // Вісник Національного технічного університету «Харківський політехнічний інститут» : зб. наук. праць. – Харків : НТУ «ХПІ», 2011. – № 17. – С. 149–156.
15. Гибридные нейро-фаззи модели и мультиагентные технологии в сложных системах : монография / [В. А. Филатов, Е. В. Бодянский, В. Е. Кучеренко и др. ; под общ. ред. Е. В. Бодянского]. – Дніпропетровськ : Системні технології, 2008. – 403 с.

Стаття надійшла до редакції 10.12.2013.
Після доробки 11.03.2014.

Зайко Т. А.[1], Олейник А. А.[2], Субботин С. А.[3]
[1]Аспирантка, Запорожский национальный технический университет, Украина
[2]Канд. техн. наук, доцент, Запорожский национальный технический университет, Украина
[3]Д-р техн. наук, профессор, Запорожский национальный технический университет, Украина

СОКРАЩЕНИЕ РАЗМЕРНОСТИ ОБУЧАЮЩЕЙ ВЫБОРКИ НА ОСНОВЕ АССОЦИАТИВНЫХ ПРАВИЛ

Рассмотрена задача сокращения обучающей выборки. Разработан метод редукции данных на основе ассоциативных правил. Создано программное обеспечение на основе предложенного метода. Проведены эксперименты по решению практических задач, что позволило исследовать эффективность предложенного метода.

Ключевые слова: ассоциативное правило, достоверность, модель, поддержка, сокращение, обучающая выборка, терм.

Zayko T.[1], Oliinyk A.[2], Subbotin S.[3]
[1]Postgraduate student, Zaporizhzhya National Technical University, Ukraine
[2]Ph.D., Associate Professor, Zaporizhzhya National Technical University, Ukraine
[3]Doctor of Science, Professor, Zaporizhzhya National Technical University, Ukraine

TRAINING SAMPLE DIMENSION REDUCTION BASED ON ASSOCIATION RULES

The problem of training sample reduction is considered. A method for data reduction based on association rules is developed. The proposed method of training sample dimensionality reduction includes stages of reduction of instances, features and redundant terms, to evaluate the informativety of features uses the information about the extracted association rules. The developed method allows to create a partition of the feature space with less examples than in the original sample, which in turn allows the synthesis of simpler and more convenient for the perception of the diagnostic model. The practical value of these results is that on the basis of the proposed method the practical problem of reducing the training sample for the synthesis of the diagnostic model of quality confectionery products is solved.

Keywords: association rule, confidence, model, support, reduction, training sample, term.

REFERENCES

1. Chaudhuri A., Stenger H. Survey sampling theory and methods. New York, Chapman & Hall, 2005, 416 p.
2. Encyclopedia of survey research methods / ed. P. J. Lavrakas. Thousand Oaks, Sage Publications, 2008, Vol. 1–2, 968 p.
3. Kokren U. per. s angl. I. M. Sonina ; pod red. A. G. Volkova, N. K. Druzhinina. Metody vyborochnogo issledovanija. Moscow, Statistika, 1976, 440 p.
4. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 339 p.
5. Subbotin S. A., Olejnik An. A., Gofman E. A., Zajcev S. A., Olejnik Al. A.; pod red. Subbotina S. A. Intellektual'nye informacionnye tehnologii proektirovanija avtomatizirovannyh sistem diagnostirovanija i raspoznavanija obrazov : monografija. Har'kov, OOO «Kompanija Smit», 2012, 317 p.
6. Gkoulalas-Divanis A., Verykios V. S. Association Rule Hiding for Data Mining. New York, Springer-Verlag, 2010, 150 p.
7. Koh Y. S., Rountree N. Rare Association Rule Mining and Knowledge Discovery. New York, Information Science Reference, 2009, 320 p.
8. Zhang C., Zhang S. Association rule mining: models and algorithms. Berlin, Springer-Verlag, 2002, 238 p.
9. Zhao Y., Zhang C., Cao L. Post-mining of association rules: techniques for effective knowledge extraction. New York, Information Science Reference, 2009, 372 p.
10. 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.
11. Zajko T. A., Olejnik A. A., Subbotin S. A. Izvlechenie chislennyh associativnyh pravil s uchetom znachimosti priznakov, *Shidno-Єvropejs'kij zhurnal peredovih tehnologij,* 2013, No. 5/4 (65), pp. 28–34.
12. Olijnik A. O., Subbotin S. O., Olijnik O. O. Intelektual'nij analiz danih: navchal'nij posibnik. Zaporizhzhja, ZNTU, 2012, 271 p.
13. Subbotin S. A. Kriterii individual'noj informativnosti i metody otbora jekzempljarov dlja postroenija diagnosticheskih i raspoznajushhih modelej, *Bionika intelektu,* 2010, No. 1, pp. 38–42.
14. Subbotin S. A. Metody formirovanija vyborok dlja postroenija diagnosticheskih modelej po precedentam, Visnik Nacional'nogo tehnichnogo universitetu «Harkivs'kij politehnichnij institut» : zb. nauk. prac'. Harkiv, NTU «HPI», 2011, No. 17, pp. 149–156.
15. Filatov V. A., Bodjanskij E. V., Kucherenko V. E. i dr. ; pod obshh. red. E. V. Bodjanskogo Gibridnye nejro-fazzi modeli i mul'tiagentnye tehnologii v slozhnyh sistemah : monografija. Dnipropetrovs'k, Sistemni tehnologiї, 2008, 403 p.