Вирішено задачу автоматизації факторного аналізу в транзакційних базах даних. Метою роботи було створення еволюційного методу факторного аналізу для пошуку прихованих залежностей у транзакційних базах даних. Запропоновано метод факторного аналізу, в якому формування груп близьких ознак виконується на основі еволюційного підходу, оцінювання еквівалентності термів ознак здійснюється шляхом видобування асоціативних правил

Ключові слова: асоціативне правило, база правил, ознака, транзакція, еволюційний пошук

Решена задача автоматизации факторного анализа в транзакционных базах данных. Целью работы являлось создание эволюционного метода факторного анализа для поиска скрытых зависимостей в транзакционных базах данных. Предложен метод факторного анализа, в котором формирование групп близких признаков выполняется на основе эволюционного подхода, оценивание эквивалентности термов признаков осуществляется путем извлечения ассоциативных правил

Ключевые слова: ассоциативное правило, база правил, признак, транзакция, эволюционный поиск

УДК 004.93

ЕВОЛЮЦІЙНИЙ МЕТОД ФАКТОРНОГО АНАЛІЗУ ДАНИХ, ПРЕДСТАВЛЕНИХ У ВИГЛЯДІ БАЗ ТРАНЗАКЦІЙ

Т. А. Зайко
Аспірант*
E-mail: tzyakun@mail.ru
А. О. Олійник
Кандидат технічних наук, доцент*
E-mail: olejnikaa@gmail.com
С. О. Субботін
Кандидат технічних наук, професор*
E-mail: subbotin@zntu.edu.ua
*Кафедра програмних засобів
Запорізький національний технічний університет
вул. Жуковського, 64, м. Запоріжжя, 69063

1. Вступ

Складні технічні та медичні об'єкти, процеси й системи характеризуються, як правило, великою кількістю змінних, деякі з яких є взаємозалежними [1 – 3]. Для пошуку таких прихованих залежностей застосовують методи факторного аналізу [4 – 7], що дозволяє виявляти взаємозв'язки між різними ознаками, які характеризують досліджувані об'єкти або процеси, пояснюючи в такий спосіб їх внутрішню природу.

Однак застосування відомих методів факторного аналізу [4 – 7] є можливим при виконанні деяких умов [5, 6]: ознаки, які описують досліджувані об'єкти або процеси, повинні бути чисельними; кількість екземплярів вибірки повинна бути більшою не менш, ніж у два рази, ніж кількість ознак; однорідність навчальної вибірки; симетричність розподілу ознак навчальної вибірки.

Вибірки даних, що характеризують реальні технічні та медичні об'єкти, не завжди задовольняють наведеним вище умовам.

Крім того, ряд задач діагностування та класифікації пов'язаний з необхідністю обробки вибірок даних, представлених у вигляді баз транзакцій, у яких кожна транзакція описує деякий набір характеристик конкретного об'єкта або процесу, при цьому кількість ознак у різних транзакціях може бути різною [8 – 10].

Тому актуальною є розробка методу факторного аналізу, вільного від зазначених недоліків, що дозволяє обробляти дані, представлені у вигляді баз транзакцій.

Метою роботи є створення еволюційного методу факторного аналізу для пошуку прихованих залежностей у транзакційних базах даних.

2. Постановка задачі факторного аналізу в транзакційних базах даних

Нехай задана база транзакцій D:

$$D = \left\{ T_1, T_2, ..., T_{N_D} \right\},$$

у якій кожний елемент T_j, $j = 1,2,...,N_D$ містить інформацію про деякі взаємозалежні події, де $N_D = |D|$ – кількість елементів (транзакцій) у наборі даних D.

Елементи T_j можуть представлятися у вигляді:

$$T_j = \left(tid_j, item_j \right),$$

де tid_j – ідентифікатор j-ї транзакції T_j; $item_j = \left\{ t_{1j}, t_{2j}, ..., t_{N_{item_j} j} \right\} \subseteq I$ – список елементів, що входять у транзакцію T_j; t_{ij} – i-й елемент списку $item_j$, $i = 1,2,...,N_{item_j}$; $N_{item_j} = |item_j|$ – кількість елементів множини $item_j$; $I = \left\{ \tau_1, \tau_2, ..., \tau_{N_I} \right\}$ – множина можливих змінних (ознак), які можуть входити в список елементів $item_j$ кожної транзакції T_j, $j = 1,2,...,N_D$ набору даних D; τ_a – a-й елемент множини I, $a = 1,2,...,N_I$; $N_I = |I|$ – кількість елементів множини I.

У випадку, якщо база транзакцій D містить крім бінарних, ще й дійсні змінні, елементи t_{ij} транзакції T_j представляються кортежем:

$$t_{ij} = \left\langle \tau_{ij}; v\left(\tau_{ij}\right)\right\rangle,$$

де τ_{ij} – ознака із множини I, що відповідає елементу t_{ij}; $v\left(\tau_{ij}\right)$ – значення ознаки τ_{ij} в транзакції T_j, $v\left(\tau_{ij}\right) \in \Delta_{ij} = \left[\tau_{ijmin}; \tau_{ijmax}\right]$; τ_{ijmin} і τ_{ijmax} – мінімальне та максимальне значення з діапазону можливих значень Δ_{ij} ознаки τ_{ij}.

Тоді задача факторного аналізу полягає у виявленні набору $\Psi = \left\{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\right\}$ ($N_\Psi = |\Psi| \le N_I$), що складається з факторів Ψ_d, кожний з яких характеризує групу тісно пов'язаних ознак $\tau \in I$. Таким чином, у результаті факторного аналізу забезпечується стиснення інформації шляхом об'єднання множини ознак через деякий набір факторів.

3. Еволюційний метод факторного аналізу

Як відзначалося вище, відомі методи факторного аналізу [4 – 7] вимагають досить великої кількості екземплярів у навчальних вибірках, наявності виняткова кількісних ознак, однорідності навчальної вибірки, а також не призначені для обробки даних, представлених у вигляді баз транзакцій. З метою усунення зазначених недоліків і можливості виконання факторного аналізу в транзакційних базах даних у розробленому методі виконується видобування асоціативних правил, що дозволяє виконати оцінювання еквівалентності термів ознак, виключити з подальшого розгляду надлишкові ознаки, скоротивши тим самим простір пошуку й зменшивши час факторного аналізу. Після цього здійснюється формування груп $\Psi = \left\{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\right\}$ близьких ознак за допомогою оптимізаційних процедур. Оскільки пошук факторних груп необхідно виконувати в дискретній множині ознак, для розв'язання даної задачі доцільним є застосування методів еволюційної оптимізації [11 – 13], які є методами глобального пошуку, не висувають вимог до цільової функції та вхідних змінних, а також є методами дискретної оптимізації.

У запропонованому методі факторного аналізу на початковому етапі задається транзакційна база даних D, яка може містити як чисельні, так і бінарні або якісні ознаки.

Потім із заданої бази транзакцій D видобуваються асоціативні правила, використовуючи відомі методи пошуку таких правил [8 – 10, 14] $D \to БП$, у результаті чого виконується узагальнення даних, і, відповідно, виключення з подальшого розгляду надлишкових ознак, а також деяких термів ненадлишкових ознак. Це дозволяє скоротити простір пошуку та час виконання факторного аналізу.

Далі виконується спрощення синтезованої бази асоціативних правил БП [1, 12], поєднуючи по можливості деякі правила.

Після цього на основі побудованої бази правил БП виділяються терми ознак $\tau_a \in I$. Для цього аналізується кожне асоціативне правило з бази БП ($AR_l \in БП$), у результаті чого формуються масиви термів кожної з ознак $\tau_a \in I$:

$$\Delta\tau_a = \left\{\Delta\tau_{a1}, \Delta\tau_{a2}, ..., \Delta\tau_{aN_{\Delta\tau_a}}\right\},$$

де $\Delta\tau_{ak} \in \left[\Delta\tau_{akmin}; \Delta\tau_{akmax}\right]$ – k-й терм (інтервал) a-ї ознаки; $\Delta\tau_{akmin}$ і $\Delta\tau_{akmax}$ – мінімальне й максимальне значення в k-му термі a-ї ознаки, відповідно; $N_{\Delta\tau_a}$ – кількість термів a-ї ознаки.

Важливо відзначити, що не обов'язково границі сусідніх інтервалів перетинаються (умова $\Delta\tau_{akmax} = \Delta\tau_{a(k+1)min}$ не повинна виконуватися для всіх k), оскільки раніше були синтезовані асоціативні правила й, відповідно, виключені надлишкові (неінформативні) терми деяких ознак.

Потім визначається еквівалентність термів ознак. Будемо вважати, що терми тим еквівалентніше, чим вище ймовірність (частота) того, що екземпляри (асоціативні правила), які потрапили в один терм $\Delta\tau_{ak}$ першої ознаки $\tau_a \in I$, потраплять в інший терм $\Delta\tau_{bm}$ другої ознаки $\tau_b \in I$. Тому для визначення еквівалентності термів ознак будемо розраховувати частоту попадання асоціативних правил у терми різних ознак:

$$\alpha_M = \frac{\sum_{l=1}^{N_{БП}} \beta_{Ml}}{N_{БП}},$$

де $M = \langle a, b, k, m\rangle$ – кортеж, який визначає взаємозв'язок k-го $\Delta\tau_{ak}$ терму a-ї ознаки $\tau_a \in I$ й m-го терму $\Delta\tau_{bm}$ b-ї ознаки $\tau_b \in I$; $N_{ВН}$ – кількість правил у базі БП; β_{Ml} – величина, що визначає наявність зв'язку між термами ознак кортежу M в l-му правилі $AR_l \in БП$ синтезованої бази правил БП:

$$\beta_{Ml} = \begin{cases} 1, \text{ якщо в l-му правилі бази правил БП} \\ \quad \text{містяться терми } \Delta\tau_{ak} \text{ та } \Delta\tau_{bm}; \\ 0, \text{ в іншому випадку.} \end{cases}$$

Після визначення еквівалентності термів α_M визначається еквівалентність ознак. Будемо вважати, що ознаки тим еквівалентніше, чим вони містять більше еквівалентних термів. Для оцінювання еквівалентності a-ї та b-ї ознак визначається величина γ_{ab}:

$$\gamma_{ab} = \frac{\sum_{m=1}^{N_{\Delta\tau_b}} \sum_{k=1}^{N_{\Delta\tau_a}} \alpha_{abkm}}{N_{\Delta\tau_b} N_{\Delta\tau_a}}.$$

Для формування груп $\Psi = \left\{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\right\}$ близьких ознак використовується еволюційна оптимізація. З метою застосування еволюційного пошуку для факторного аналізу визначимо спосіб подання розв'язку (множини факторних груп $\Psi = \left\{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\right\}$) у вигляді хромосом $H: \Psi \to H$. Оскільки розмір множини Ψ (кількість факторних груп N_Ψ) заздалегідь є не відомим, розмір хромосом N_{H_j} буде змінним: $N_{H_j^{(t)}} \ne N_{H_j^{(t+1)}}$, $N_{H_j^{(t)}} \ne N_{H_k^{(t)}}$, де $N_{H_j^{(t)}}$ – розмір j-ї хромосоми H_j на t-й ітерації еволюційного пошуку. Гени h_{ji} хромосоми H_j будуть відповідати i-му елементу множини $\Psi \to H_j$. Таким чином, кожна j-а хромосома

t-ї популяції $H_j^{(t)}$ буде відповідати j-му розв'язку на t-й ітерації еволюційного пошуку, що представляє собою j-у множину факторних груп:

$$H_j^{(t)} \to \Psi_j^{(t)} = \left\{ \Psi_{1j}^{(t)}, \Psi_{2j}^{(t)}, ..., \Psi_{N_{\Psi_j^{(t)}} j}^{(t)} \right\}.$$

Гени h_{ji} хромосом H_j являють собою вектори цілих чисел, що відповідають номерам ознак $\tau_a \in I$ із множини I.

Таким чином, пропонується представляти хромосоми $H_j^{(t)}$ у вигляді векторів змінної довжини, кожний елемент (ген) яких також є масивом елементів змінної довжини. При чому гени у векторних хромосомах при такому способі кодування мають властивості негомологічності, тобто числа в кожному векторі можуть приймати значення із заданої множини та не повинні повторюватися [11 – 13].

Приклад хромосоми при розв'язанні задачі факторного аналізу наведено на рис. 1.

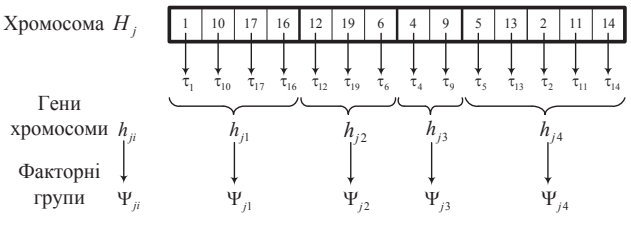

Рис. 1. Приклад хромосоми при розв'язанні задачі факторного аналізу

Як видно з рис. 1, деякі ознаки (τ_3, τ_7, τ_8 та ін.) не ввійшли в хромосому H_j, що свідчить про те, що дані ознаки не є членами ні однієї з чотирьох факторних груп $\Psi_j = \left\{ \Psi_{j1}, \Psi_{j2}, \Psi_{j3}, \Psi_{j4} \right\}$, визначених у хромосомі $H_j \to \Psi_j$ у вигляді генів $h_{j1}, h_{j2}, h_{j3}, h_{j4}$, відповідно. Множина ознак, що не ввійшли в хромосому H_j (τ_3, τ_7, τ_8 і ін.), є аналогією множині I′ при використанні жадібної стратегії. Гени h_{j1}, h_{j2}, h_{j3}, h_{j4} хромосоми H_j мають різну довжину, що відповідає кількості ознак у відповідній факторній групі. Так, наприклад, друга факторна група $\Psi_{j2} \leftarrow h_{j2}$ складається із трьох ознак $\Psi_{j2} = \left\{ \tau_{12}, \tau_{19}, \tau_6 \right\}$.

Для оцінювання хромосом H_j пропонується використовувати цільову функцію γ_{H_j}, що враховує еквівалентність γ ознак τ_a у кожному з генів h_{ji} (факторній групі):

$$\gamma_{H_j} = \frac{N_{H_j}}{1 + \sum_{i=1}^{N_{H_j}} \gamma_{h_{ji}}} \to \min,$$

де $\gamma_{h_{ji}}$ – середнє значення еквівалентності ознак в i-му гені h_{ji} j-ї хромосоми H_j:

$$\gamma_{h_{ji}} = \frac{\sum_{a=1}^{N_{h_{ji}}} \sum_{\substack{b=a+1 \\ a,b:\, \tau_a, \tau_b \in h_{ji}}}^{N_{h_{ji}}} \gamma_{ab}}{\frac{1}{2} N_{h_{ji}} \left(N_{h_{ji}} - 1 \right)},$$

де $N_{h_{ji}}$ – розмір i-го гену j-ї хромосоми.

Ініціалізація еволюційного пошуку в запропонованому методі відбувається шляхом генерації заданої кількості хромосом N_H, кожна з яких H_j відповідає деякій множині Ψ_j факторних груп $\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}$.

Генерація j-ї хромосоми H_j починається з випадкового формування масиву цілих неповторюваних чисел з інтервалу $\left[1; N_I\right]$, кожне з яких відповідає певній ознаці $\tau_a \in I$. Потім створений масив розбивається на деяку випадковозгенеровану кількість $N_{H_j} << N_I$ генів h_{ji}, що представляють собою певні факторні групи змінної довжини. Таким чином, на початковому етапі еволюційного пошуку вважається, що кожна ознака $\tau_a \in I$ може входити в деяку факторну групу.

У якості оператора відбору пропонується використовувати пропорційний відбір [11 – 13], при якому до схрещування й мутації допускаються хромосоми H_j з рівнем пристосованості (значенням цільової функції γ_{H_j}), не нижче середньої по популяції.

Для схрещування хромосом пропонується використовувати модифікований з урахуванням розв'язуваної задачі оператор одноточкового схрещування. При використанні запропонованого оператора випадковим чином вибирається точка розриву tr:

$$tr = randc\left[1; \min\left(N_{H_1}; N_{H_2}\right) - 1\right],$$

де $randc[a; b]$ – випадково згенероване ціле число в інтервалі $[a; b]$; N_{H_1} і N_{H_2} – розмір хромосом H_1 і H_2, що схрещуються.

Потім відбувається обмін генами між хромосомами H_1 й H_2, у результаті чого формуються два нащадки H_{c1} та H_{c2}:

$$H_{c1} = \left\{ h_{11}, h_{12}, ..., h_{1tr}, h_{2(tr+1)}, h_{2(tr+2)}, h_{2N_{H_2}} \right\};$$

$$H_{c2} = \left\{ h_{21}, h_{22}, ..., h_{2tr}, h_{1(tr+1)}, h_{1(tr+2)}, h_{1N_{H_1}} \right\}.$$

Після формування хромосом-нащадків H_{c1} і H_{c2}, як правило, у них виникають неприпустимі гени (у різних генах однієї хромосоми містяться однакові елементи, що відповідають одним ознакам $\tau_a \in I$). Оскільки в схрещуванні брало участь дві хромосоми, у хромосомах-нащадках можуть міститися гени h_1 та h_2 з не більш, ніж двома однаковими елементами $\tau_a \in I$. Для приведення хромосом H_{c1} і H_{c2} до припустимого вигляду з кожної з них виключаються повторювані ознаки $\tau_a \in I$. При цьому якщо у двох генах h_{j1} і h_{j2} однієї хромосоми містяться однакові елементи, що відповідають одній ознаці $\tau_a \in I$, відбувається виключення відповідного елемента з гена h_{ji}, у якому він має менший вплив на загальну еквівалентність $\gamma_{h_{ji}}$. Для визначення впливу ознаки $\tau_a \in I$ в кожній з хромосом h_{j1} і h_{j2} обчислюються величини $\Delta\gamma_{h_{j1} \setminus \tau_a}$ й $\Delta\gamma_{h_{j2} \setminus \tau_a}$:

$$\Delta\gamma_{h_{j1} \setminus \tau_a} = \gamma_{h_{j1}} - \gamma_{h_{j1} \setminus \tau_a};$$

$$\Delta\gamma_{h_{j2} \setminus \tau_a} = \gamma_{h_{j2}} - \gamma_{h_{j2} \setminus \tau_a},$$

де $\gamma_{h_{j1} \setminus \tau_a}$ й $\gamma_{h_{j2} \setminus \tau_a}$ – середні значення еквівалентності без ознаки τ_a в генах h_{j1} і h_{j2}, відповідно.

Ознака $\tau_a \in I$ виключається з гена h_{ji} з меншим значенням величини $\Delta\gamma_{h_{ji}\setminus\tau_a}$. Таким чином, запропонований оператор схрещування дозволяє генерувати нову множину припустимих розв'язків шляхом передачі інформації від хромосом-батьків.

Після схрещування виконується мутація випадково обраних генів деяких хромосом, що дозволяє більш детально досліджувати простір пошуку й урізноманітнити його. Пропонується виконувати мутації вставки, обміну та видалення деяких ознак із хромосом.

Мутація вставки виконується з метою додавання ознак у факторні групи. Для цього випадково обрана ознака $\tau_a \in I$, що не міститься в хромосомі H_j ($\tau_a \notin H_j$), додається в один з її генів. У випадку, якщо виконується умова:

$$\gamma_{h_{ji}\cup\tau_a} \geq \gamma_{h_{ji}},$$

додана ознака τ_a залишається в хромосомі H_j.

Мутація обміну передбачає вибір і заміну двох випадково відібраних ознак $\tau_a \in h_1$ і $\tau_b \in h_2$ з різних генів h_1 і h_2 однієї хромосоми H_j, обраної для мутації. Операція мутації обміну вважається успішною при виконанні умови:

$$\Delta\gamma_{h_1\cup\tau_b\setminus\tau_a} + \Delta\gamma_{h_2\cup\tau_a\setminus\tau_b} > 0,$$

де $\Delta\gamma_{h_1\cup\tau_b\setminus\tau_a} = \gamma_{h_1} - \gamma_{h_1\cup\tau_b\setminus\tau_a}$ та $\Delta\gamma_{h_2\cup\tau_a\setminus\tau_b} = \gamma_{h_2} - \gamma_{h_2\cup\tau_a\setminus\tau_b}$ – прирости середньої еквівалентності факторних груп, що відповідають генам h_1 і h_2 при обміні ознаками τ_a й τ_b.

Мутація видаленням спрямована на виключення з факторної групи $\Psi_{ji} \leftarrow h_{ji}$ ознаки $\tau_a \in h_{ji}$, що слабо корелює з іншими ознаками $\tau_b \in h_{ji}$ із цієї ж групи Ψ_{ji}. Для цього випадковим чином у відібраній для мутації хромосомі H_j вибирається мутуючий ген $h_{ji} \in H_j$, з якого виключається ознака $\tau_a \in h_{ji}$, що має найменший вплив на середню еквівалентність гена $\gamma_{h_{ji}}$:

$$\tau_a : \Delta\gamma_{h_{ji}\setminus\tau_a} = \min_{\substack{b=1,2,\ldots,N_{h_{ji}} \\ \tau_b \in h_{ji}}} \Delta\gamma_{h_{ji}\setminus\tau_b}.$$

Таким чином, операція мутації в деяких випадках змінює розмір хромосоми та дозволяє сформувати факторні групи з більш прийнятними оцінками еквівалентності.

Після схрещування й мутації відбувається формування нової множини розв'язків. На нову ітерацію переходять хромосоми з найкращими значеннями цільової функції, а також хромосоми, отримані в результаті схрещування й мутації.

Еволюційна оптимізація триває доти, поки не буде досягнута максимально припустима кількість ітерацій $N_{it.}$ або не знайдений розв'язок $H_j \rightarrow \Psi_j$ із прийнятним значенням цільової функції, що не перевищують мінімально припустиме значення $\gamma_{H_j} \leq \gamma_{Hmin}$.

Запропонований метод факторного аналізу на основі еволюційного підходу передбачає видобування правил із заданих баз транзакцій, що дозволяє здійснити оцінювання еквівалентності ознак та узагальнення даних, і, відповідно, виключення з подальшого розгляду надлишкових ознак, що дозволяє скоротити простір пошуку та час виконання факторного аналізу. У розробленому методі визначення еквівалентності ознак для формування факторних груп виконується виходячи із частоти їх спільного попадання в асоціативні правила синтезованої бази правил, що дозволяє оцінювати тісноту зв'язку між різними ознаками (якісними, кількісними), не висувати вимог до вхідних даних і виконувати факторний аналіз у транзакційних базах даних.

Використання еволюційного підходу для пошуку груп якісно близьких ознак дозволяє більш детально в порівнянні з жадібною стратегією досліджувати простір пошуку, а також формувати групи близьких ознак, що характеризуються більш прийнятними оцінками еквівалентності.

4. Експерименти та результати дослідження розробленого методу

З метою експериментального дослідження розробленого методу факторного аналізу на основі еволюційного підходу за допомогою мови програмування C# було створено програмне забезпечення, що реалізує запропонований метод.

Для дослідження властивостей і характеристик запропонованого методу факторного аналізу, а також порівняння його з відомими аналогами використовувалися спеціально згенеровані тестові дані. Розроблений метод порівнювався з методом головних компонентів PCA (Principal Component Analysis) [4] і методом дискримінантного аналізу Фішера FDA (Fisher Discriminant Analysis) [7]. Існуючі методи пошуку груп взаємозалежних ознак дозволяють виділяти факторні груп на основі вибірок, які представляють собою прямокутні таблиці чисел, що містять значення всіх ознак для всіх екземплярів. Тому тестові вибірки генерувалися таким чином, щоб вони могли бути вихідними даними як для відомих методів факторного аналізу, так і для розробленого. Крім того, важливо відзначити, що при створенні тестових вибірок деякі ознаки залежали одна від одної, утворюючи, таким чином, групи еквівалентності.

На рис. 2. наведено графік залежності часу функціонування різних методів факторного аналізу від кількості ознак у вихідній вибірці. При цьому кількість екземплярів у вибірці була постійною і становила $N_D = 1000$ шт.

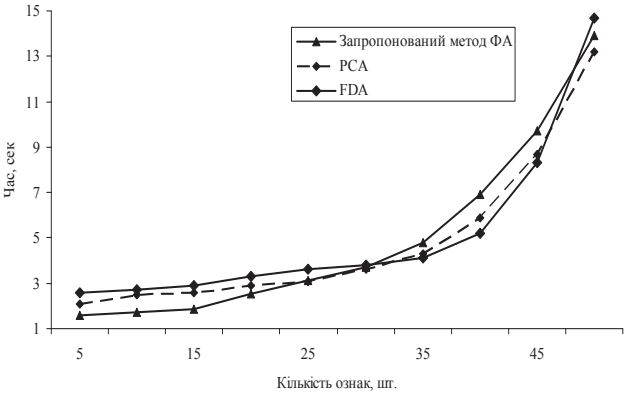

Рис. 2. Графік залежності часу функціонування методу від кількості ознак у вихідній вибірці

Графік, зображений на рис. 2, підтверджує квадратичну залежність оцінки обчислювальної складності O запропонованого методу від кількості ознак $|I|$ і характеризує його як обчислювально ефективний метод факторного аналізу. Крім того, з рис. 2 видно, що запропонований метод при невисоких значеннях кількості ознак швидше виконує факторний аналіз вхідних даних, при $|I| = 50$ час роботи методів несуттєво відрізняється.

На рис. 3 відображено результати експериментів по дослідженню залежності кількості згенерованих факторних груп від кількості ознак у вихідній вибірці (кількість екземплярів при проведенні експериментів становила $N_D = 5000$ шт.).

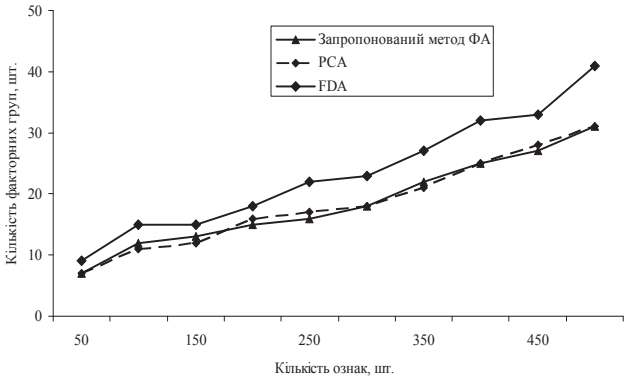

Рис. 3. Графік залежності кількості згенерованих факторних груп від кількості ознак у вихідній вибірці

Криві, зображені на рис. 3 і побудовані за результатами застосування різних методів факторного аналізу, показують, що кількість синтезованих факторних груп пропорційна кількості ознак у навчальній вибірці, що обумовлено внутрішніми взаємозв'язками між ознаками та структурою вибірки. Як видно, метод FDA синтезував більшу кількість факторних груп, виділяючи деякі взаємозалежні набори ознак у різні групи. Метод PCA і запропонований метод показали схожі результати.

Таким чином, результати експериментів показали, що розроблений метод дозволяє виконувати факторний аналіз на основі баз транзакцій.

Порівняння запропонованого методу з існуючими аналогами підтвердило доцільність його застосування на практиці.

5. Висновки

У роботі вирішено актуальну задачу автоматизації факторного аналізу в транзакційних базах даних.

Наукова новизна роботи полягає в тому, що запропоновано еволюційний метод факторного аналізу, в якому формування груп близьких ознак виконується на основі еволюційного підходу, оцінювання еквівалентності термів ознак здійснюється шляхом видобування асоціативних правил, що дозволяє виконувати факторний аналіз в транзакційних базах даних, виключати надлишкові ознаки, скорочувати простір пошуку.

Практична цінність отриманих результатів полягає в тому, що було створено програмне забезпечення, що реалізує запропонований еволюційний метод факторного аналізу та дозволяє виділяти групи близьких ознак в транзакційних базах даних.

Литература

1. Encyclopedia of artificial intelligence [Text] / eds.: J. R. Dopico, J. D. de la Calle, A. P. Sierra. – New York : Information Science Reference, 2009. – Vol. 1–3. – 1677 p.
2. Зайченко, Ю. П. Основи проектування інтелектуальних систем : навчальний посібник [Текст] / Ю. П. Зайченко. – К.: Слово, 2004.– 352 с.
3. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиационных двигателей : монография [Текст] / [А. В. Богуслаев, Ал. А. Олейник, Ан. А. Олейник, Д. В. Павленко, С. А. Субботин]; под ред. Д. В. Павленко, С. А. Субботина. – Запорожье: ОАО "Мотор Сич", 2009. – 468 с.
4. Jolliffe, I. T. Principal Component Analysis [Text] / I. T. Jolliffe. – Berlin : Springer-Verlag. – 2002. – 489 p.
5. Rummel, R. J. Applied Factor Analysis [Text] / R. J. Rummel. – Evanston : Northwestern University Press. – 1988. – 617 p.
6. Иберла, К. Факторный анализ [Текст] / К. Иберла. – М. : Статистика. – 1980. – 398 с.
7. McLachlan, G. Discriminant Analysis and Statistical Pattern Recognition [Text] / G. McLachlan. – New Jersey : John Wiley & Sons. – 2004. – 526 p.
8. Субботін, С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навчальний посібник [Текст] / С. О. Субботін. – Запоріжжя : ЗНТУ, 2008. – 341 с.
9. Gkoulalas-Divanis, A. Association Rule Hiding for Data Mining [Text] / A. Gkoulalas-Divanis,V. S. Verykios. – New York : Springer-Verlag. – 2010. – 150 p.
10. Zhang, C. Association rule mining: models and algorithms [Text] / C. Zhang, S. Zhang. – Berlin : Springer-Verlag. – 2002. – 238 p.
11. The Practical Handbook of Genetic Algorithms [Text] / ed. L. D. Chambers. – Florida: CRC Press, 2000. – Vol. I: Applications. – 520 p.
12. Субботін, С. О. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: монографія [Текст] / С. О. Субботін, А. О. Олійник, О. О. Олійник; під заг. ред. С.О. Субботіна. – Запоріжжя : ЗНТУ, 2009. – 375 с.
13. Haupt, R. Practical Genetic Algorithms [Text] / R. Haupt, S. Haupt. – New Jersey: John Wiley & Sons, 2004. – 261 p.
14. Zhao, Y. Post-mining of association rules: techniques for effective knowledge extraction [Text] / Y. Zhao, C. Zhang, L. Cao. – New York : Information Science Reference. – 2009. – 372 p.