УДК 004.89+336.713

Т.О. Савчук, С.І. Петришин
Вінницький національний технічний університет

ІДЕНТИФІКАЦІЯ ПРОБЛЕМНИХ СИТУАЦІЙ ТА ЇХ СТАНІВ У СКЛАДНИХ ТЕХНІЧНИХ СИСТЕМАХ З ВИКОРИСТАННЯМ МОДИФІКОВАНОГО АЛГОРИТМУ ФОРЕЛ

В статті наведено особливості ідентифікації проблемних ситуацій та їх станів з використанням модифікованого алгоритму ФОРЕЛ кластеризації. Основною відмінністю модифікованого алгоритму від класичного є можливість користувача задавати значення показника якості розбиття, що робить алгоритм гнучким при проведенні аналізу проблемних ситуацій та їх станів в складних технічних системах.

Ключові слова: кластерний аналіз, алгоритм ФОРЕЛ.

The article describes the features of the identification of problem situations and their states using the modified algorithm FOREL of clasterization. The main difference of the modified algorithm from the classical one is the user's ability to set the quality score for clusters, which makes the algorithm flexible in analyzing problem situations and states in complex technical systems.

Key words: cluster analysis, algorithm FOREL.

Вступ

У будь-якій галузі людської діяльності може виникнути ситуація, яка містить протиріччя та не має однозначного вирішення відносно обставин і умов, в яких розгортається діяльність особистості, або групи, тобто, проблемна ситуація [1]. У зв'язку з швидким науково-технічним прогресом значно збільшилась кількість складних технічних систем, які сьогодні застосовуються повсюди, і в побуті, і на виробництві. У зв'язку з цим зростає ймовірність виникнення різного роду пошкоджень в таких системах, а, отже, і постає необхідність у розробленні таких методів та засобів, які б могли допомогти ідентифікувати проблемні ситуації та їх стан в складних технічних системах для збільшення ефективності виявлення неполадок у таких системах. Застосування кластеризації дасть змогу здійснювати таку ідентифікацію.

Метою дослідження є підвищення точності ідентифікації проблемних ситуацій і станів у складних технічних системах завдяки модифікації класичного алгоритму ФОРЕЛ.

Постановка задачі

Задачу ідентифікації проблемних ситуацій та їх станів в складних технічних системах можна сформулювати так.

Нехай $X_n = \{x_1, \mathbf{K}, x_n\} \in X\,(n = \overline{1,\infty})$ – вибірка проблемних ситуацій та їх станів, $w(X_i, X_j)$ – функція міри близькості між такими ситуаціями та їх станами.

Потрібно розбити вибірку проблемних ситуацій у складних технічних системах та їх станів на $k (k \le n)$ непересічних підмножин, що називаються кластерами так, щоб

– кожен кластер складався з ситуацій та їх станів, близьких за метрикою w,

– проблемні ситуації у складних технічних системах та їх стани, які знаходяться в різних кластерах, значно відрізнялися [2].

При цьому кожна проблемна ситуація $x_i \in X_n$ належать до одного з класерів; залежно від кластера, до якого вона віднесена, користувач отримує інформацію про стан, до якого було віднесено ситуацію, що і є метою ідентифікації.

Аналіз сучасних підходів до розв'язання задачі ідентифікації проблемних ситуацій та їх санів у складних технічних системах

Серед існуючих розв'язків задачі ідентифікації проблемних ситуацій та їх станів у складних технічних системах найпоширеніші такі.

Класифікації проблемних ситуацій та їх станів у складних технічних системах за її характеристиками. У цій задачі множина класів таких ситуацій, до яких може бути віднесений об'єкт дослідження, наперед відома. При цьому слід зазначити такі недоліки використання класифікації [3] для ідентифікації проблемних ситуацій та їх станів у складних технічних системах:

– навчальна вибірка має бути достатньо великою;

– у навчальну вибірку мають входити проблемні ситуації, які представляють всі класи, що є проблемним при аналізі таких ситуацій;

– проблема overfitting, сутність якої полягає в тому, що класифікаційна функція добре адаптується до даних і якщо серед них зустрічаються помилки або аномальні значення, то функція інтерпретує їх як частину внутрішньої структури даних, що є неприйнятним для аналізу проблемних ситуацій та їх станів у складних технічних системах;

– проблема underfitting, яка полягає в тому, що під час перевірки класифікатора виявляється велика кількість помилок, що є неприйнятним для предметної області, яка аналізується.

Іншим рішенням є визначення залежностей, які часто повторюються серед проблемних ситуацій з використанням пошуку асоціативних правил. Знайдені залежності подаються у вигляді правил і можуть бути використані як для кращого розуміння природи даних, що аналізуються, так і для прогнозування виникнення певних подій, що не є важливим при аналізі таких ситуацій [3].

На відміну від означених рішень, кластеризація проблемних ситуацій та їх станів полягає в пошуку незалежних кластерів у множині даних про такі ситуації та їх стани, які підлягають ідентифікації. Це дозволяє зрозуміти структуру даних. Крім того, групування однорідних даних дає змогу зменшити їх кількість для спрощення аналізу надалі [3]. Перевагами такого підходу є ітераційний пошук оптимального результату розбиття проблемних ситуацій та їх станів у складних технічних системах на кластери на підставі сукупності обраних показників та виявлення внутрішніх зв'язків між ситуаціями, які підлягають ідентифікації [3], вибору інформативних ознак та мір близькості між двома об'єктами, об'єктом і кластером, двома кластерами, що є актуальним при їх ідентифікації.

Отже, найбільш доцільним для розв'язання задачі ідентифікації проблемних ситуацій у складних технічних системах є використання методів та алгоритмів кластеризації. Серед алгоритмів кластерного аналізу особливої уваги заслуговує алгоритм ФОРЕЛ як такий, що характеризується чіткістю, збіжністю за скінченне число кроків, мінімальною кількістю характеристик та параметрів проблемних ситуацій у складних технічних системах, що аналізуються.

Розроблення модифікованого алгоритму ФОРЕЛ кластеризації проблемних ситуацій та їх станів у складних технічних системах.

До особливостей класичного алгоритму ФОРЕЛ кластеризації [4] проблемних ситуацій в складних технічних системах належать такі:

– його продуктивність є невисокою;

– необхідність завдання радіуса кластера R;

– алгоритм є збіжним за скінченне число кроків;

– в лінійному просторі центром кластера може бути як будь-яка точка, так і проблемна ситуація, яка виникає в складній технічній системі;

– на першому кроці алгоритму обирається одна із проблемних ситуацій в складних технічних системах як початковий об'єкт, від якого проводитиметься кластеризація, що, своєю чергою впливатиме на її якість;

– наявність апріорних знань про діаметри кластерів;

– можливість включення в кластер об'єктів з інших кластерів через неправильний вибір радіуса кластера R.

Але наведений алгоритм у класичному варіанті не є прийнятним для ідентифікації проблемних ситуацій та їх станів у складних технічних системах, оскільки він передбачає завдання радіусу кластера, що може бути наперед невідомим. Це можна усунути за рахунок модифікації класичного алгоритму, що передбачало б розрахунок радіуса кластера R залежно від якості кластеризації h , яка б задовольнила користувача. Для виключення можливості появи викидів всередині кластерів (таких проблемних ситуацій та їх станів, що за обраними параметрами та характеристиками є віддаленими від основного скупчення об'єктів у кластері) після формування кожного кластера проводиться його аналіз і вилучення проблемних ситуацій, які в процесі аналізу були визначені як викиди, а також віднесення їх до іншого кластера.

У формалізованому вигляді модифікований алгоритм ФОРЕЛ кластеризації проблемних ситуацій та їх станів в складних технічних системах має такий вигляд:

1. Формування множини некластеризованих проблемних ситуацій в складних технічних системах:

$$\mathrm{U}=\mathrm{X}^{\mathrm{n}}, \tag{1}$$

де U – множина некластеризованих проблемних ситуацій у складних технічних системах; n – кількість проблемних ситуацій у складних технічних системах, що підлягають кластеризації; X – множина проблемних ситуацій в складних технічних системах, що підлягають кластеризації;

2. Знаходження значень відстаней між некластеризованими точками:

мінімального

$$\min = \min_{\mathrm{i,j=1,\mathbf{K},n,i\neq j}} \mathrm{a}\left(\mathrm{x_i}, \mathrm{x_j}\right), \tag{2}$$

де $a(x_i, x_j)$ – відстань між i-ю та j-ю проблемними ситуаціями;

максимального

$$\max = \max_{i,j=1,\mathbf{K},n,i\neq j} a\left(x_i, x_j\right). \tag{3}$$

3. Знаходження значення радіуса кластера

$$R = \frac{\max - \min \times (100 - h)}{100} + \min, \tag{4}$$

де R – максимальне значення радіуса кластерів проблемних ситуацій у складних технічних системах; h –показник якості кластеризації $0 \le h \le 100$;

4. При умові $U \neq \varnothing$ (у вибірці є некластеризовані проблемні ситуації в складних технічних системах):

4.1. Обрати довільну проблемну ситуацію $x_0 \in U$ випадковим чином;

4.2. Сформувати кластер проблемних ситуацій у складних технічних системах – сферу з центром в x_0 і радіусом R :

$$K_0 = \{x_i \in U \mid a(x_i, x_0) \le R\}, \tag{5}$$

де K_0 – сформований кластер проблемних ситуацій у складних технічних системах; $a(x_i, x_0)$ - відстань від проблемної ситуації, яка має місце в складній технічній системі x_i до центра кластера x_0

4.3. Помістити центр кластера в його центр мас:

$$x_0 = \arg\min_{x_i \in K_0} \sum\nolimits_{x_j} a(x_i, x_j), \tag{6}$$

де x_0 – центр мас кластера;

4.4. Виконувати п.п. 4.1-4.3 ПОКИ центр x_0 не стабілізується;

4.5. Знайти

$$\overline{a_0} = \frac{\sum_{x_0, x_i \in K_0} a(x_i, x_0)}{|K_0|}, i = 1, \mathbf{K}, |K_0|, \tag{7}$$

де $\overline{a_0}$ – середнє значення відстані між центром кластера та проблемними ситуаціями в складних технічних системах, які йому належать;

4.6. Знайти

$$x_{\max} = \arg\max_{x_i \in K_0} a(x_i, x_0), i = 1, \mathbf{K}, |K_0|, \quad (8)$$

де $X_{\max}$ – максимально віддалена від центра кластера проблемна ситуація;

4.7. Якщо $\left(a(x_{\max}, x_0) \geq 2 \cdot \overline{a_0}\right)$ то

$x_{\max}$ вилучити з K_0

$$K_0 = K_0 \setminus \{x_{\max}\}, \quad (9)$$

та $x_{\max}$ додати до U :

$$U = U \cup \{x_{\max}\}. \quad (10)$$

4.8. Виконувати п.п. 4.6 4.7 , поки не буде вилучень $x_{\max}$ з K_0 .

4.9. Вилучити проблмні ситуації, що вкючені до кластера K_0 (як кластеризовані):

$$U = U \setminus K_0. \quad (11)$$

5. Виконувати п.4, поки $U = \varnothing$ (всі проблемні ситуації в складних технічних системах кластеризовані).

Тоді основними кроками модифікованого алгоритму ФОРЕЛ кластеризації проблемних ситуацій та їх станів в складних технічних системах (рисунок 1) є такі:

Крок 1. Знайти значення R на основі введеного користувачем параметра якості кластеризації η.

Крок 2. Помістити центр кластера в будь-яку з некластеризованих проблемних ситуацій у складних технічних системах та віднести до кластера проблемні ситуації, відстань до якої менше за радіус R.

Крок 3. Обчислити новий центр мас знайденого кластера. Центр кластера проблемних ситуацій у складних технічних системах перенести в знайдений центр мас.

Крок 4. Якщо новий центр мас відрізняється від попереднього, необхідно повернутися до кроку 2 і повторити цикл, поки центр мас не перестане зміщуватися. Отже, центр кластера переміщається в область локального згущення проблемних ситуацій у складних технічних системах.

Крок 5. Обчислити новий центр ваги і перенести в нього центр кластера.

Крок 6. Знайти середнє значення відстані між центром кластера та всіма проблемними ситуаціями у складних технічних системах з даного кластера.

Крок 7. Якщо відстань до найвіддаленішої проблемної ситуації більша ніж два середні значення відстані між центром кластера та всіма ситуаціями поточного кластера, то видалити цю проблемну ситуацію з кластера та перенести її в множину некластеризованих ситуацій та перейти до кроку 5, інакше перейти до кроку 8.

Крок 8. Проблемні ситуації в складних технічних системах, які належать новому таксону, виключаються з некластеризованих, та робота алгоритму розпочинається з кроку 2, поки всі ситуації не будуть виключені з множини некластеризованих.

Для визначення якості ідентифікації проблемних ситуацій та їх станів у складних технічних системах було проведено експериментальні дослідження стосовно розбиття множини таких ситуацій (потужність множини проблемних ситуацій та їх станів складає 100) за допомогою модифікованого та класичного алгоритму ФОРЕЛ. На основі отриманих результатів було розраховано значення відносного показника якості розбиття [5] множини проблемних ситуацій та їх станів в складних технічних системах, що склав для класичного алгоритму 80,7%, а для модифікованого алгоритму 85,3%. Це свідчить про доцільність використання запропонованого модифікованого алгоритму ФОРЕЛ при ідентифікації проблемних ситуацій та їх станів.

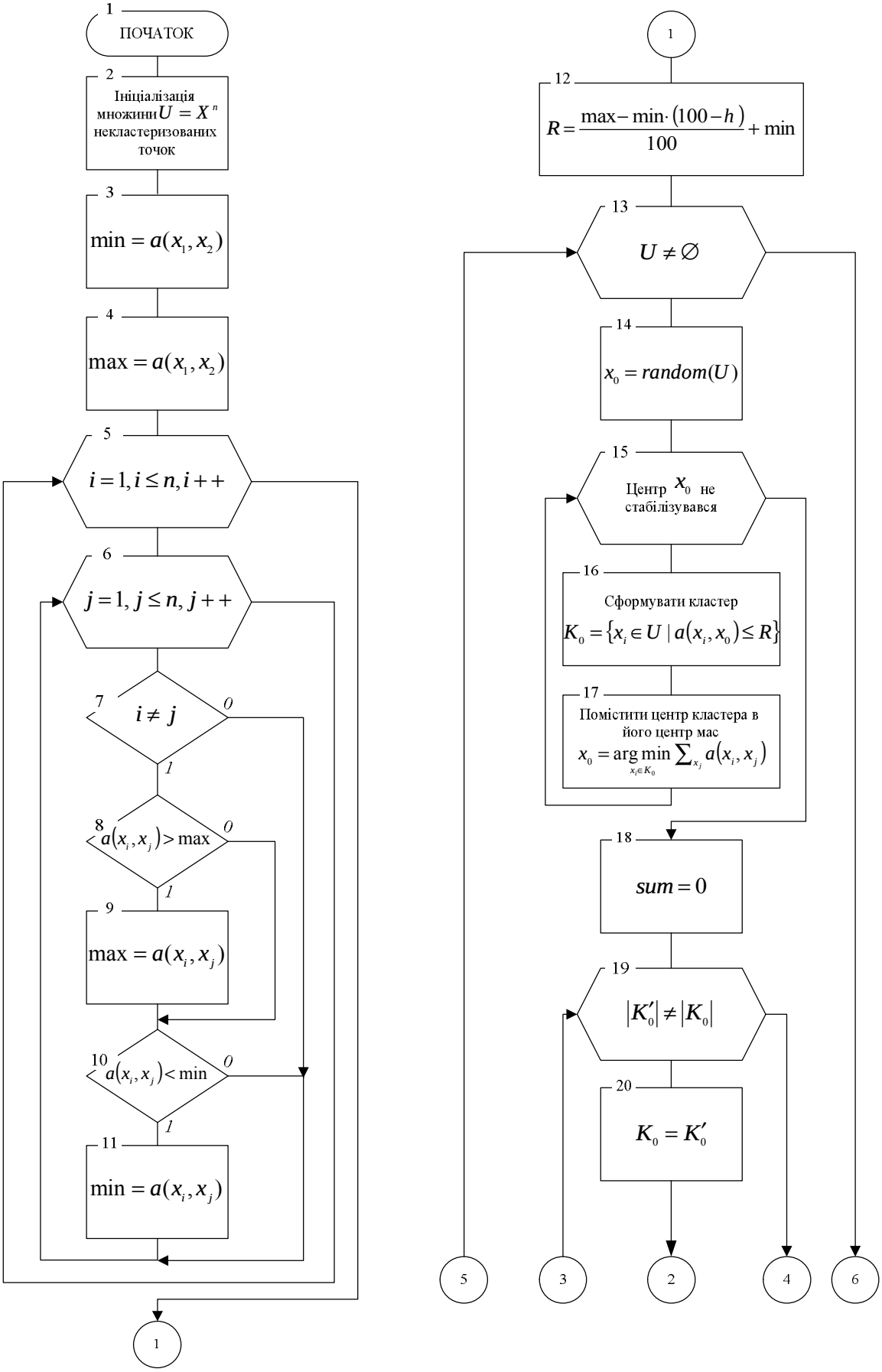

Рис. 1. Схема модифікованого алгоритму ФОРЕЛ кластеризації проблемних ситуацій та їх станів в складних технічних системах

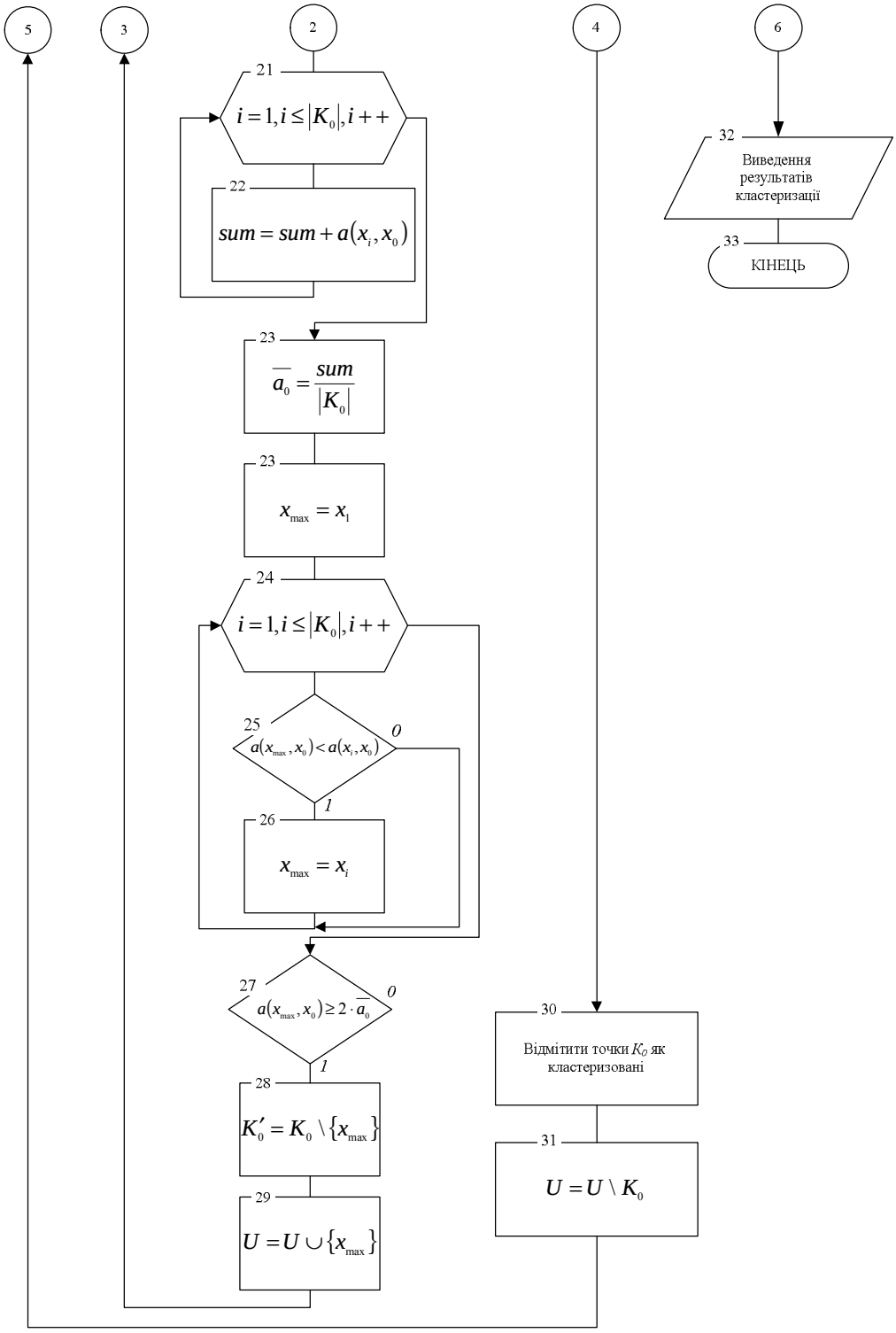

Рис. 1. (Продовження). Схема модифікованого алгоритму ФОРЕЛ кластеризації проблемних ситуацій та їх станів в складних технічних системах

Висновки

Отже, запропонований модифікований алгоритм кластеризації проблемних ситуацій та їх станів у складних технічних системах, що грунтується на алгоритмі ФОРЕЛ, дає змогу провести їх ідентифікацію з покращенням значення **в**ідносного показника якості розбиття на 4,6%.

1. Бим-Бад Б.М. Педагогический энциклопедический словарь / Б. Бим-Бад. – М., 2002.
2. Савчук Т.О. Оцінювання надзвичайних ситуацій на залізничному транспорті, що базується на

кластерному аналізі./ Т. Савчук, С. Петришин – м. Вінниця – 2010 р – (Тези XXXIX науково-технічної конференції професорсько-викладацького складу, співробітників та студентів університету з участю працівників науково-дослідницьких організацій та інженерно-технічних працівників підприємств м.Вінниці та області. ВНТУ). 3. Савчук Т.О. Порівняльний аналіз використання методів кластеризації для ідентифікації надзвичайних ситуацій на залізничному транспорті / Т. Савчук, С. Петришин – 2010. – Вип. 11(134). – С. 135–140 – (Наукові праці Донецького національного технічного університету. – Серія "Інформатика, кібернетика і обчислювальна техніка"). 4. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Н. Загоруйко – Новосибирск. – 1999. – с. 270. 5. Савчук Т.О. Оцінювання результатів моделювання процесу кластерного аналізу надзвичайних ситуацій на залізничному транспорті / Т.О. Савчук, С.І. Петришин. – 2012. – №1, С. 18–24 – (Інформаційні технології та комп'ютерна інженерія).

УДК 004.8

О.Ю. Сєдушев, Є.В. Буров
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

МЕТОДИ ВИДОБУВАННЯ ДАНИХ З БАЗ НЕЧІТКИХ ЗНАНЬ

Досліджено нечіткі методи видобування даних. Акцент при цьому робиться на інтелектуальному аналізі баз нечітких знань та задачах, які при цьому виникають. Описано найпопулярніші сьогодні методи, їхні переваги та отримані за їх допомогою результати. Наведено узагальнені варіанти використання таких методів.

Ключові слова: нечіткі методи видобування даних, база нечітких правил, нечітка логіка, база знань.

The paper aims to study the fuzzy data mining techniques. The emphasis is put on an intelligent analysis of fuzzy knowledge bases and problems that arise. Most popular methods are described, their advantages and results obtained with their assistance are highlighted. Generalized use cases of such methods are given.

Key words: fuzzy data mining methods, fuzzy rule base, fuzzy logic, knowledge base.

Вступ та постановка проблеми

Бази нечітких знань являють собою сукупність фактів, лінгвістичних змінних та відповідних функцій приналежності (сукупно трактуватимемо їх як знання), якими можна оперувати, та нечітких висловлювань "ЯКЩО–ТО", що мають назву нечітких продукційних правил виведення. Такі бази знань є цінним джерелом для опису нечітких понять, видобування даних та прийняття різнорідних рішень у різних галузях науки, бізнесу та виробництва, а також є ефективним засобом моделювання у багатьох задачах кібернетики та штучного інтелекту, що мають справу з нечіткостями, серед яких управління технологічними процесами, різного роду діагностики, розпізнавання образів та мови, прогнозування часових рядів тощо.

Чіткі та нечіткі бази знань використовуються сьогодні в багатьох напрямках застосування інформаційних технологій: для побудови експертних та інтелектуальних систем, систем дистанційного навчання та контролю знань тощо. Сьогодні поширені системи підтримки прийняття рішень, які використовують знання, отримані від експертів. Такі знання зберігаються у базах знань, які зазвичай слугують для різнорідного інтелектуального аналізу та виявлення (виведення) певних закономірностей. Усе частіше моделювання складних залежностей в економіці, медицині, будівництві та в інших областях здійснюється за допомогою саме нечітких баз знань. А тому