УДК 681.518:004.93.1

ІНФОРМАЦІЙНО-ЕКСТРЕМАЛЬНИЙ АЛГОРИТМ АГЛОМЕРАТИВНОГО КЛАСТЕР-АНАЛІЗУ

***В. О. Востоцький**, аспірант;*
***С. А. Занченко**, аспірант,*
Сумський державний університет, м. Суми

У рамках інформаційно-екстремальної технології, що ґрунтується на максимізації інформаційної спроможності системи шляхом введення в процесі навчання додаткових інформаційних обмежень, розглядається категорійна модель та алгоритм навчання системи підтримки прийняття рішень, що функціонує в режимі кластер-аналізу.

***Ключові слова:** кластер-аналіз, навчання, оптимізація, інформаційний критерій функціональної ефективності, система підтримки прийняття рішень*

ВСТУП

Підвищення ефективності та оперативності керування виробничими процесами органічно пов'язане із розробленням та впровадженням інтелектуальних інформаційних технологій. Застосування здатних самонавчатися АСКТП (автоматизованих систем керування технологічним процесом) у виробництві дозволяє здійснити перехід від застарілих суб'єктивних методів ручного керування до методів класифікаційного керування, що базуються на ідеях і методах машинного навчання та розпізнавання образів [1-3]. При цьому важливого значення набуває розроблення здатних навчатися (самонавчатися) алгоритмів кластер-аналізу, що обумовлено необхідністю формування за результатами відносно тривалого моніторингу керованого технологічного процесу відкритого алфавіту класів розпізнавання, потужність якого апріорно є невідомою. Існуючі методи кластер-аналізу, побудовані на дистанційній метриці [4-7] носять в основному модельний характер, оскільки вони не враховують перетину класів розпізнавання, який спостерігається в практичних завданнях автоматизації виробничих процесів

Один із перспективних шляхів аналізу та синтезу здатних навчатися в режимі кластер-аналізу АСКТП полягає у використанні ідей і принципів інформаційно-екстремальної інтелектуальної технології (ІЕІ-технологія), що ґрунтується на максимізації інформаційної спроможності системи шляхом введення в процесі навчання додаткових інформаційних обмежень [8-9]. При цьому основною складовою АСКТП є інтелектуальна система підтримки прийняття рішень (СППР), основними завданнями якої є оцінка поточного функціонального стану технологічного процесу та вироблення відповідних керуючих команд для особи, що приймає рішення. У працях [10,11] розглянуто питання автоматичної класифікації технологічного процесу у рамках ІЕІ-технології, але для випадку, коли алфавіт класів був апріорно частково визначений.

У статті розглядається у рамках ІЕІ-технології алгоритм кластер-аналізу для формування апріорної нечіткої навчальної матриці з метою побудови в процесі навчання СППР чіткого розбиття простору ознак на класи еквівалентності.

ПОСТАНОВКА ЗАВДАННЯ

Розглянемо АСКТП, в якій СППР, що навчається, функціонує в режимі кластер-аналізу вхідних даних. Нехай відома некласифікована багатовимірна навчальна матриця $\| y_i^{(j)} \|$, $i = \overline{1,N}$, $j = \overline{1,n}$, де N, n –

кількість ознак розпізнавання і випробувань (спостережень) відповідно. Необхідно перетворити вхідну апріорно некласифіковану навчальну матрицю $\| y_i^{(j)} \|$ у нечітку класифіковану і побудувати чітке розбиття простору ознак на класи розпізнавання $\{X_m^o \mid m = \overline{1, M}\}$, які характеризують можливі допустимі функціональні стани технологічного процесу. При цьому для оптимізації координат структурованого вектора параметрів функціонування $g_m =< g_{m,1}, \ldots, g_{m,q}, \ldots, g_{m,Q} >$, для яких відомі обмеження $R_q(g_1, \ldots, g_Q) \le 0$, здійснити пошук глобального максимуму усередненого за алфавітом $\{X_m^o\}$ інформаційного критерію функціональної ефективності (КФЕ) навчання системи

$$E_{\max}^* = \frac{1}{M} \sum_{m=1}^{V} \max_{\{k\}} E_m \ , \qquad (1)$$

де G – область допустимих значень параметрів функціонування, що оптимізуються; $\{k\}$ – множина кроків навчання СППР розпізнавати реалізації класу X_m^o.

У режимі екзамену необхідно прийняти рішення про належність реалізації образу, що характеризує поточний функціональний стан технологічного процесу, до відповідного класу із заданого алфавіту.

МАТЕМАТИЧНА МОДЕЛЬ

Математична модель навчання СППР у режимі кластер-аналізу містить як обов'язкову складову частину вхідний математичний опис, який подамо у вигляді теоретико-множинної структури

$$\Delta_B =< G, T, \Omega, Z, W, Y, X; \Phi_1, \Phi_2, \Phi_3 >, \qquad (2)$$

де G – простір вхідних факторів; T – множина моментів часу зняття інформації; Ω – простір ознак розпізнавання; Z – простір можливих функціональних станів технологічного процесу; W – вибіркова множина – нечітка некласифікована навчальна матриця; Y – вибіркова множина – нечітка класифікована навчальна матриця; X – бінарна навчальна матриця; $\Phi_1 : G \times T \times \Omega \times Z \to W$ – оператор формування множини W на вході СППР; Φ_2 – оператор формування нечіткої класифікованої навчальної матриці Y; Φ_3 – оператор формування бінарної навчальної матриці X.

На рис. 1 показано категорійну модель навчання СППР у вигляді діаграми відображення множин, що застосовуються в процесі навчання.

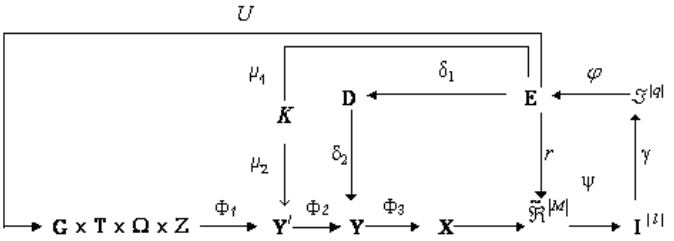

Рисунок 1 – Категорійна модель навчання СППР

У діаграмі (рис. 1) оператор $\theta : X_{(m)} \to \Re^{|M|}$ будує розбиття $\Re^{|M|}$ простору ознак на класи розпізнавання, яке у загальному випадку є нечітким, а оператор класифікації ψ: $\tilde{\Re}^{|M|} \to \mathbf{I}^{|l|}$, перевіряє основну статистичну гіпотезу про належність реалізацій $\{x_{(m)}^{(j)} \mid j = \overline{1,n}\} \in X_{(m)}$ нечіткому класу $X_{(m)}^{o}$. Тут l– кількість статистичних гіпотез. Оператор γ: $I^{|l|} \to \Im^{|q|}$ шляхом оцінки статистичних гіпотез формує множину точнісних характеристик $\Im^{|q|}$, де $q = l^2$. Оператор $\varphi : \Im^{|q|} \to E$ обчислює множину значень інформаційного КФЕ, який є функціоналом точнісних характеристик. Контур оптимізації геометричних параметрів нечіткого розбиття $\tilde{\Re}^{|M|}$ шляхом пошуку максимуму КФЕ навчання розпізнаванню реалізацій класу X_m^o замикається оператором r: $E \to \tilde{\Re}^{|M|}$. У діаграмі терм-множина D складається із допустимих значень контрольних допусків на ознаки розпізнавання, які оптимізуються операторами показаного на рис. 2 контура.

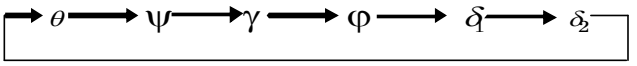

Рисунок 2– Контур оптимізації контрольних допусків

Контур кластеризації замикається через терм-множину K, яка складається із допустимих значень дистанційних критеріїв, операторами μ_1 і μ_2. Оператор U: $E \to G \times T \times \Omega \times Z$ регламентує процес кластеризації і дозволяє оптимізувати параметри плану навчання СППР.

ІНФОРМАЦІЙНО-ЕКСТРЕМАЛЬНИЙ АЛГОРИТМ КЛАСТЕР-АНАЛІЗУ

Формування вхідної нечіткої класифікованої навчальної матриці Y здійснювалося шляхом поєднання дистанційних алгоритмів прямої кластеризації типу FOREL [5] і QUALITY TRESSPASS [6]. При цьому в процесі ітераційного пошуку глобального максимуму інформаційного КФЕ (1) в робочій (допустимій) області визначення його функції здійснювалася цілеспрямована корекція як кількості реалізацій в таксоні, так і внутрішньокласові та міжкласові дистанційні критерії.

Алгоритм навчання СППР, що функціонує в режимі кластер-аналізу, має такі етапи:

1) формування алфавіту класів за вхідними дистанційними критеріями в бінарному парацептуальному просторі;

2) формування у рамках ІЕІ-технології вхідного математичного опису СППР за сформованим апріорним нечітким розбиттям простору ознак на класи розпізнавання;

3) оптимізація просторово-часових параметрів функціонування СППР з метою побудови оптимальних в інформаційному розумінні розв'язувальних правил.

При цьому як КФЕ навчання СППР розглянемо нормований ентропійний критерій (за Шенноном), модифікація якого для двохальтернативної системи оцінок (М = 2) і рівноймовірних гіпотез, що характеризують найбільш важкий у статистичному сенсі випадок прийняття рішень, має вигляд [8]

$$E_m^{(k)} = 1 + \frac{1}{2}\left(\frac{\alpha_m^{(k)}(d)}{\alpha_m^{(k)}(d) + D_{2,m}^{(k)}(d)} \log_2 \frac{\alpha_m^{(k)}(d)}{\alpha_m^{(k)}(d) + D_{2,m}^{(k)}(d)} + \right.$$

$$+ \frac{\beta_m^{(k)}(d)}{D_{1,m}^{(k)}(d) + \beta_m^{(k)}(d)} \log_2 \frac{\beta_m^{(k)}(d)}{D_{1,m}^{(k)}(d) + \beta_m^{(k)}(d)} + \qquad (3)$$

$$+ \frac{D_{1,m}(d)}{D_{1,m}^{(k)}(d) + \beta_m^{(k)}(d)} \log_2 \frac{D_{1,m}(d)}{D_{1,m}^{(k)}(d) + \beta_m^{(k)}(d)} +$$

$$\left. + \frac{D_{2,m}^{(k)}(d)}{\alpha_m^{(k)}(d) + D_{2,m}^{(k)}(d)} \log_2 \frac{D_{2,m}^{(k)}(d)}{\alpha_m^{(k)}(d) + D_{2,m}^{(k)}(d)} \right),$$

де $\alpha_m^{(k)}(d)$ – помилка першого роду прийняття рішення на k-му кроці навчання; $\beta_m^{(k)}(d)$ – помилка другого роду; $D_{1,m}^{(k)}(d)$ – перша достовірність; $D_{2,m}^{(k)}(d)$ – друга достовірність; d – дистанційна міра, яка визначає радіуси гіперсферичних контейнерів, побудованих у радіальному базисі простору Хеммінга.

Критерій (2) потрібно розглядати як функціонал від точнісних характеристики, оскільки вони є функціями відстані вершин еталонних векторів від геометричних центрів контейнерів відповідних класів розпізнавання.

Процедура формування вхідного математичного опису на i-му кроці кластеризації вхідних даних має такий узагальнений вигляд:

1) вибір довільної початкової точки формування кластера;
2) онулення лічильника кроків збільшення радіуса кластера: $r := 0$;
3) $r := r + 1$;
4) обчислення еталонного вектора реалізацій сформованого кластера, вершина якого визначає центр ваги кластера;
5) перенесення геометричного центра кластера в центр ваги;
6) корекція кластера за доданими реалізаціями. Якщо кількість реалізацій у кластері менша за обсяг репрезентативної вибірки ($n_{\min} \geq 40$), то виконується пункт 3, інакше – пункт 7;
7) ЗУПИН.

За результатами кластеризації формується вхідний математичний опис СППР для побудови чітких розв'язувальних правил. При цьому основний алгоритм навчання СППР складається з таких кроків:

1) ініціалізація системи. Формування вхідного математичного опису за поточними результатами роботи алгоритму кластер-аналізу;
2) реалізація інформаційно-екстремального алгоритму навчання СППР з оптимізацією системи контрольних допусків на ознаки розпізнавання [10];
3) коригування дистанційних критеріїв при максимальному значенні КФЕ з метою мінімізації середньої кодової відстані реалізацій образу від ядра його кластера та максимізації середньої міжцентрової відстані для сформованого алфавіту класів.

Таким чином, кластеризація вхідних даних у СППР, що навчається, відбувається шляхом поєднання агломеративного алгоритму, що базується на дистанційних мірах близькості, який дозволяє сформувати вхідну в загальному випадку нечітку класифіковану навчальну матрицю, з інформаційно-екстремальним алгоритмом навчання СППР, який

дозволяє побудувати безпомилкові за навчальною матрицею розв'язувальні правила.

ПРИКЛАД РЕАЛІЗАЦІЇ АЛГОРИТМУ КЛАСТЕР-АНАЛІЗУ

Реалізація алгоритму здійснювалася за некласифікованою навчальною вибіркою, отриманою за результатами моніторингу технологічного процесу виробництва складних мінеральних добрив *NPK* у ВАТ «Сумихімпром» для трьох класів, які характеризували функціональний стан АСКТП: X_1^o – азот, фосфор і калій у нормі, X_2^0 – азоту менше норми і X_3^0 – фосфору менше норми. Некласифікована навчальна матриця складалася зі 120 структурованих реалізацій, кожна з яких містила 55 ознак. Оскільки ознаки мали різні шкали виміру, було виконано їх нормалізацію за методом зведених шкал.

У процесі покрокової агломеративної композиції окремих кластерів оцінювалася функціональна ефективність СППР шляхом пошуку глобального значення усередненого КФЕ (2).

Графік залежності усередненого значення КФЕ $\overline{E}$ від кроків агломеративної кластеризації вхідних даних показаний на рис. 3.

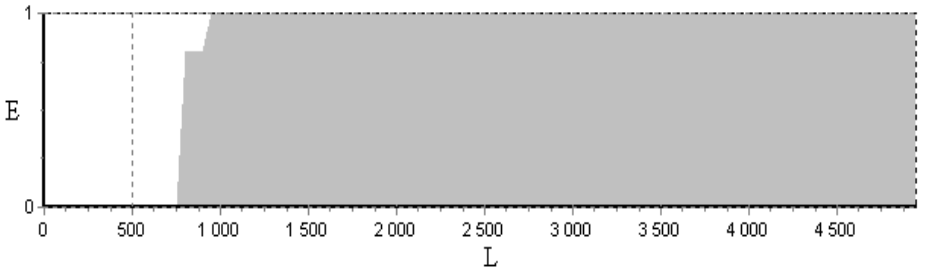

Рисунок 3 – Графік залежності усередненого КФЕ від кроків кластеризації даних

Аналіз рис. 3 показує, що максимальне середнє значення КФЕ було отримане при кількості реалізацій $n = 40$ на 1084-му кроці кластеризації. При цьому відсутність спаду значення КФЕ після досягнення його граничного максимального значення, що є підтвердженням побудови безпомилкових за навчальною матрицею розв'язувальних правил, свідчить про збіжність алгоритму навчання СППР у режимі кластеризації вхідних даних.

На рис. 4 показано графік зміни усередненого за алфавітом класів розпізнавання нормованого ентропійного критерію (2) від параметра δ(delta) поля допусків у процесі паралельної оптимізації системи контрольних допусків на ознаки розпізнавання. Тут і далі заштрихована ділянка графіка позначає робочу область визначення функції КФЕ, в якій перша і друга достовірності набувають значень більше 0,5 при двох альтернативних розв'язаннях.

Аналіз рис. 4 показує, що максимальне середнє значення КФЕ в робочій області дорівнює максимальному значенню $\overline{E}^* = 1$ при оптимальному параметрі поля допусків $\delta = \pm 64$ відносних одиниць.

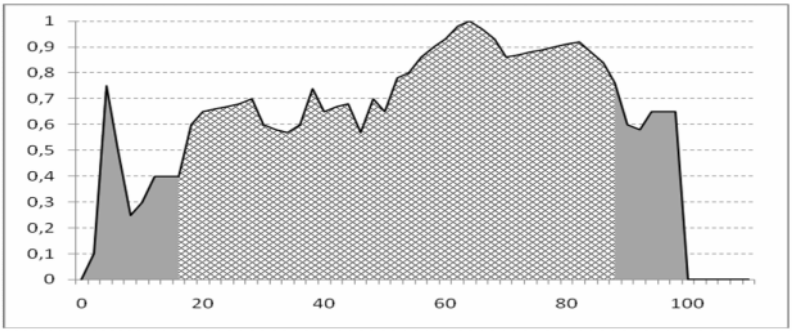

Рисунок 4– Графік залежності КФЕ від параметра поля контрольних допусків

На рис. 5 показано процес оптимізації радіусів контейнерів класів розпізнавання за запропонованим агломеративним алгоритмом кластеризації.

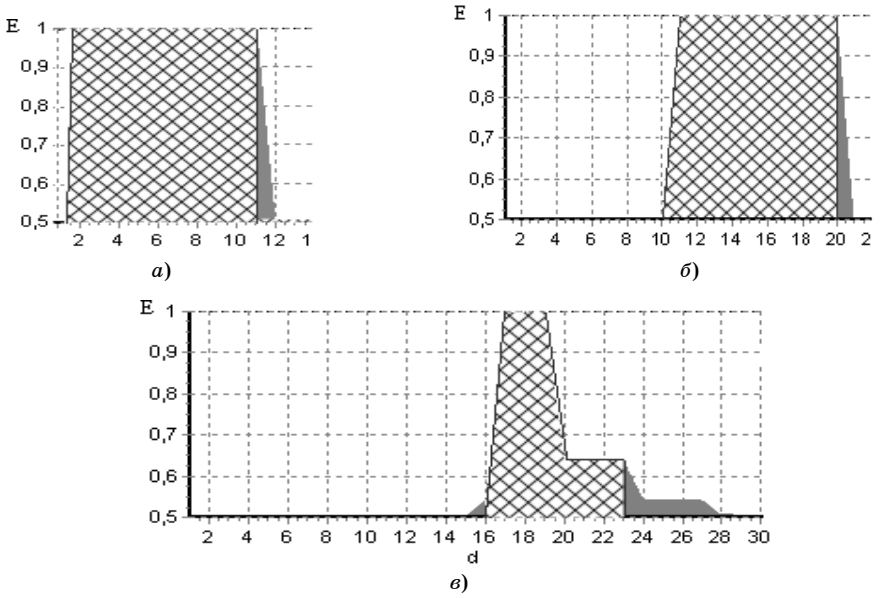

Рисунок 5 – Графіки залежності КФЕ від радіусів контейнерів класів розпізнавання: а) клас X_1^0; б) клас X_2^0; в) клас X_3^0

Результати аналізу рис. 5 і навчання СППР в режимі кластер-аналізу наведено в табл. 1. Тут d^* – оптимальне значення радіуса контейнера класу розпізнавання, d_c – міжцентрова кодова відстань.

Таблиця 1 – Результат оптимізації параметрів навчання СППР

	d^*	E	D_1	α	β	D_2	d_c
X_1^0	2	1	1	0	0	1	15
X_2^0	11	1	1	0	0	1	31
X_3^0	17	1	1	0	0	1	24

Таким чином, запропонований гібридний алгоритм містить як інформаційну міру схожості класів розпізнавання, так і дистанційні критерії, що дозволило побудувати у рамках IEI-технології безпомилкові за навчальною матрицею розв'язувальні правила.

ПЕРСПЕКТИВИ ЗАСТОСУВАННЯ

У загальному випадку запропонований алгоритм кластер-аналізу в рамках IEI-технології може використовуватися в задачах керування слабоформалізованими технологічними процесами в хімічній, металургійній, харчовій та інших галузях соціально-економічної сфери суспільства, які відбуваються за умов апріорної невизначеності. У випадках, коли не вдається побудувати безпомилкові за навчальною матрицею розв'язувальні правила, згідно з принципом відкладених рішень необхідно здійснювати оптимізацію інших параметрів функціонування СППР, що впливають на її функціональну ефективність. А при збільшенні потужності алфавіту класів розпізнавання перейти до ієрархічної структури алгоритму навчання СППР.

ВИСНОВКИ

1. Запропоновано інформаційно-екстремальний алгоритм навчання СППР, що функціонує в режимі кластер-аналізу вхідних даних, який дозволяє будувати безпомилкові за навчальною матрицею вирішальні правила
2. Використання в рамках інформаційно-екстремального алгоритму навчання СППР агломеративного алгоритму, що базується на дистанційних мірах близькості, дозволяє автоматизувати формування вхідної нечіткої класифікованої багатовимірної навчальної матриці.

ИНФОРМАЦИОННО-ЭКСТРЕМАЛЬНЫЙ АЛГОРИТМ АГЛОМЕРАТИВНОГО КЛАСТЕР-АНАЛИЗА

А. Востоцкий, С. А. Занченко,
Сумский государственный университет, г. Сумы

В рамках информационно - экстремальной технологии, основанной на максимизации информационной возможности системы путем введения в процессе обучения дополнительных информационных ограничений, рассматриваются категориальная модель и алгоритм обучения системы поддержки принятия решений, функционирующая в режиме кластер-анализа.

Ключевые слова: *кластер-анализ, обучение, оптимизация, информационный критерий функциональной эффективности, система поддержки принятия решений.*

INFORMATIONAL EXTREME ALGORITHM OF AGGLOMERATIVE CLUSTER ANALYSIS

V. O. Vostotskyi, S. A. Zanchenko,
Sumy State University, Sumy

The categorical model and decision support system learning algorithm are considered in the article. Proposed algorithm allows to create decision support system, which is functioning in a cluster-analysis state. Synthesis of the decision support system is based on maximization of informational system ability due to making additional information restrictions in the learning process.

Key words: *cluster analysis, training, optimization, information criterion of functional efficiency, decision support system.*

СПИСОК ЛІТЕРАТУРИ

1. Цыпкин Я. З. Основы теории обучающихся систем. – М.: Наука, 1970. – 251 с.
2. Васильев В. И. Проблема обучения распознаванию образов – К.: Вища школа. Головное издательство, 1989 – 64 с.

3. Вапник В. Н. Теория распознавания образов: (статистические проблемы обучения) / В. Н. Вапник, А. Я. Червоненкс. – М.: Наука, 1974.– 416 с.
4. Сокал Р. Р. Кластер-анализ и классификация: предпосылки и основные направления // Классификация и кластер / под ред. Дж. Вэн Райзина. - М. : Мир, 1980. - С. 7-19.
5. Загоруйко Н. Г. Алгоритмы обнаружения эмпирических закономерностей / Н. Г. Загоруйко, В. Н. Елкина, Г. С. Лбов. – Новосибирск: Наука. –1985. – 110 с.
6. Мандель И. Д. Кластерный анализ. - М.: Финансы и статистика, 1988.
7. Методы анализа данных: Подход, основанный на методе динамических сгущений: пер. с фр. / кол. авт. под рук. Э. Дидэ / под ред. и с предисл. С. А. Айвазяна и В. М. Бухштабера.– М.: Финансы и статистика, 1985.–375 с.
8. Краснопоясовський А. С. Інформаційний синтез інтелектуальних систем керування: Підхід, що ґрунтується на методі функціонально-статистичних випробувань. – Суми: Видавництво СумДУ, 2004. – 261 с.
9. Довбиш А. С. Основи проектування інтелектуальних систем: навчальний посібник.– Суми: Видавництво СумДУ, 2009.–171 с.
10. Довбиш А. С. Оптимізація контрольних допусків на ознаки розпізнавання в інформаційно-екстремальних методах автоматичної класифікації / А. С. Довбиш, М. В. Козинець, С. М. Котенко // Вісник Сумського державного університету. Серія Технічні науки. – 2007. - №1.– С. 169-178.

Надійшла до редакції 10 жовтня 2012 р.