УДК 681.518:004.93.1’

# ІНФОРМАЦІЙНО-ЕКСТРЕМАЛЬНЕ НАВЧАННЯ СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ З АДАПТИВНОЮ КЛАСТЕРИЗАЦІЄЮ ДАНИХ

***В. В. Москаленко,*** *аспірант,*
*Сумський державний університет, м. Суми*

*Розглядається метод адаптивної кластеризації даних на основі алгоритму Густафсона-Кесселя та інформаційно-екстремального навчання з гіпереліпсоїдними контейнерами класів розпізнавання, що відновлюються в радіальному базисі бінарного простору ознак розпізнавання.*

***Ключові слова:*** *кластер-аналіз, система підтримки прийняття рішень, оптимізація, навчання, інформаційний критерій, клас розпізнавання.*

## ВСТУП

Основним шляхом підвищення функціональної ефективності керування слабо формалізованими процесами, що функціонують за умов апріорної невизначеності та дії зовнішніх неконтрольованих факторів, є впровадження інтелектуальних інформаційних технологій, що базуються на ідеях і методах машинного навчання та розпізнавання образів [1-3]. При цьому важливим завданням є розроблення алгоритмів кластер-аналізу для автоматизації формування за результатами моніторингу керованого процесу вхідного математичного опису системи підтримки прийняття рішень (СППР), яка є основною складовою інтелектуальної АСК. Завдання трансформації апріорно нечіткого розбиття в чітке розбиття еквівалентності класів розпізнавання успішно вирішується в рамках інформаційно-екстремальної інтелектуальної технології (ІЕІ-технологія), в якій контейнери класів розпізнавання відновлюються в радіальному базисі бінарного простору ознак розпізнавання [5].

Традиційні алгоритми нечіткої кластеризації використовують як вхідні параметри задану кількість кластерів розбиття, а деякі з них і заданий показник нечіткості (розмитості) кластерів у просторі ознак [6, 7]. Різні значення вхідних параметрів алгоритму кластеризації обумовлюють різні розбиття. При цьому існує багато відомих критеріїв оцінки якості результату кластеризації (індекси валідації), які часто дають суперечливі між собою рішення, що призводить до неоднозначності у виборі оптимального розбиття [8, 9]. Проте інформаційний критерій функціональної ефективності (КФЕ) інформаційно-екстремального навчання доцільно вважати загальним критерієм якості розбиття [10, 11]. На базі інформаційного КФЕ можна реалізувати адаптивний механізм налаштування параметрів алгоритму кластеризації. Такий підхід дозволить обрати оптимальне в інформаційному розумінні розбиття даних на кластери і отримати чіткі вирішальні правила для оперативного прийняття рішень в робочому режимі СППР.

У статті розглядається процес формування апріорно нечіткої класифікованої навчальної матриці за алгоритмом нечіткої кластеризації Густафсона-Кесселя з метою побудови чітких вирішальних правил у процесі навчання інформаційно-екстремальної СППР з гіпереліпсоїдними контейнерами класів розпізнавання [12].

## ПОСТАНОВКА ЗАВДАННЯ

Розглянемо інтелектуальну АСК, в якій СППР навчається в режимі кластер-аналізу. Нехай відома апріорно 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 в режимі кластер-аналізу перетворити вхідну некласифіковану навчальну матрицю у нечітку класифіковану і в режимі навчання побудувати чітке розбиття простору ознак розпізнавання на класи розпізнавання $\{X_m^o \mid m = \overline{1, M}\}$, які характеризують функціональні стани керованого процесу, шляхом оптимізації координат структурованого вектора параметрів функціонування

$$g =< M, w, \delta, x_{m1}, x_{m2}, d_m >, \quad (1)$$

де $M$ – кількість кластерів для алгоритму кластер-аналізу (одночасно і потужність алфавіту класів розпізнавання); $w$ – показник нечіткості для алгоритму кластер-аналізу; $\delta$ – параметр поля контрольних допусків на ознаки розпізнавання; $x_{m1}, x_{m2}$ – двійкові вектори, що визначають координати першого та другого фокусів гіпереліпсоїдного контейнера класу $X_m^o$ в бінарному просторі ознак $\Omega_Б$; $d_m$ – велика піввісь контейнера класу $X_m^o$ в просторі $\Omega_Б$.

При цьому задано обмеження

$$\begin{cases} 2 \le M \le n / n_{\min}, n_m \ge n_{\min}; \\ w > 1; \\ d_m > c_m, c_m \le N / 2; \\ d(x_{c1} \oplus x) + d(x_{c2} \oplus x) - 2c_c > 0, \forall x \in \{x : d(x_{m1} \oplus x) + d(x_{m2} \oplus x) = 2d_m\}; \\ \delta \in [0; \delta_H / 2]; \end{cases}, \quad (2)$$

де $n_{\min}$ – мінімальний обсяг репрезентативної навчальної вибірки для кожного класу; $n_m$ – кількість реалізацій, що належать класу $X_m^o$; $c_m$ – фокальна відстань гіпереліпсоїдного контейнера класу $X_m^o$; $d(x_{c1} \oplus x)$, $d(x_{c2} \oplus x)$ – кодові відстані від першого та другого фокусів контейнера сусіднього класу $X_c^o$ до вектора-реалізації $x$ бінарного простору $\Omega_Б$ відповідно; $c_c$ – фокальна відстань гіпереліпсоїдного контейнера сусіднього класу $X_c^o$ у просторі $\Omega_Б$; $d(x_{m1} \oplus x)$, $d(x_{m,2} \oplus x)$ – кодові відстані від першого та другого фокусів контейнера класу $X_m^o$ до вектора-реалізації $x$ бінарного простору $\Omega_Б$ відповідно; $\delta_H$ – нормоване поле допусків, що визначає область значень параметра $\delta$.

Необхідно в процесі навчання СППР визначити оптимальні значення координат вектора параметрів функціонування (1), що забезпечують максимальне значення усередненого за алфавітом класів розпізнавання КФЕ:

$$\overline{E}^* = \frac{1}{M} \sum_{m=1}^{M} \max_{\{k\}} E_m, \quad (3)$$

де $E_m$ – інформаційний КФЕ навчання СППР розпізнавати реалізації класу $X_m^o$; $\{k\}$ – впорядкована множина кроків навчання (відновлення контейнерів класів розпізнавання).

У режимі екзамену, тобто безпосередньо у робочому режимі СППР, необхідно прийняти рішення про належність реалізацій образу, що характеризують поточний функціональний стан процесу, до відповідного класу із сформованого на етапі навчання алфавіту $\{X_m^o \mid m = \overline{1, M}\}$ і таким чином дефазифікувати функціональний стан АСК.

## АЛГОРИТМ АДАПТИВНОЇ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ

Розглянемо основні узагальнені етапи реалізації адаптивної кластеризації даних на основі алгоритму нечіткої кластеризації Густафсона-Кесселя (Gustafson-Kessel) та алгоритму інформаційно-екстремального навчання:

1. Ініціалізація констант адаптивного алгоритму:
$M_{\min} := 2$; $M_{\max} := n / n_{\min}$; $w_{\min} := 1,5$; $w_{\max} := 4,6$; $\Delta w := 0,1$.
2. Ініціалізація оптимальних параметрів алгоритму Густофсона-Кесселя та максимального усередненого за алфавітом класів інформаційного КФЕ відповідно: $w^* := 0$, $M^* := 0$, $\overline{E}^*_{\max} = 0$.
3. Ініціалізація поточного значення кількості кластерів розбиття для алгоритму Густафсона-Кесселя: $M := M_{\min}$.
4. Ініціалізація поточного значення показника нечіткості для алгоритму Густафсона-Кесселя: $w := w_{\min}$.
5. Виконання алгоритму нечіткої кластеризації Густафсона-Кесселя із заданими параметрами $M$ та $w$.
6. Формування нечіткої класифікованої навчальної матриці для алгоритму інформаційно-екстремального навчання.
7. Обчислення в процесі виконання інформаційно-екстремального навчання усередненого інформаційного КФЕ $\overline{E}^*$ за формулою (3).
8. Якщо $\overline{E}^* > \overline{E}^*_{\max}$, то $\overline{E}^*_{\max} = \overline{E}^*$ і $M^* := M$, $w^* := w$.
9. $w := w + \Delta w$.
10. Якщо $w \le w_{\max}$, то перехід на пункт 4.
11. $M := M + 1$
12. Якщо $M \le M_{\max}$, то перехід на пункт 3.
13. $M := M^*$, $w := w^*$
14. Виконання пунктів 4-6.
15. ЗУПИН.

Особливістю алгоритму Густафсона-Кесселя є можливість виділяти кластери гіпереліпсоїдної форми різного розміру та орієнтації завдяки застосуванню для кожного кластера окремої нормоутворювальної матриці.

Розглянемо основні етапи реалізації алгоритму Густафсона-Кесселя для нечіткої кластеризації даних:

Ініціалізація вхідних некластеризованих даних у вигляді багатовимірної (векторної) матриці типу «об'єкт-властивість» $\{y_i^{(j)} \mid i = \overline{1, N}, j = \overline{1, n}\}$, де $N, n$ – кількість ознак розпізнавання і векторів-реалізацій образу відповідно.

1. Ініціалізація параметрів алгоритму: $M$ – кількість кластерів, $2 \le M \le \sqrt{n}$; $w$ – показник нечіткості (зважений показник), що регулює нечіткість розбиття, знаходиться в інтервалі $w \in (1, \infty)$, проте найчастіше обирають $w = 2$; $\varepsilon$ – точнісний параметр зупинки; $Q$ – максимальна кількість ітерацій.

2. Генерація матриці нечіткого розбиття (наприклад випадковим чином):

$$U = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ & & \dots & \\ u_{M1} & u_{M2} & \dots & u_{Mn} \end{bmatrix}, \text{ при } u_{mj} \in \{0,1\}; \quad \sum_{m=1}^{M} u_{mj} = 1; \quad 0 < \sum_{j=1}^{n} u_{mj} < n,$$

де $u_{mj}$ – ступінь належності $j$-го об'єкта $m$-му кластеру.

3. Ініціалізація лічильника кількості ітерацій $l = 0$.

4. $l = l + 1$.

5. Розрахунок центрів кластерів

$$y_m = \frac{\sum_{j=1}^{n} (u_{mj}^{(l-1)})^w \cdot y^{(j)}}{\sum_{j=1}^{n} (u_{mj}^{(l-1)})^w}.$$

6. Обчислення матриці коваріації для $m$-го кластера

$$A^{(m)} = \frac{\sum_{j=1}^{n} (u_{mj}^{(l-1)})^w \cdot (y^{(j)} - y_m)^T \cdot (y^{(j)} - y_m)}{\sum_{j=1}^{n} (u_{mj}^{(l-1)})^w}.$$

7. Обчислення відстані між векторами вхідних даних та центрами кластерів

$$d^2_{A^{(m)}}(y^{(j)}, y_m) = (y_m - y^{(j)})^T \cdot \left[ \left[ \sqrt{\det(A^{(m)})} \cdot \det(A^{(m)}) \cdot \right]^{\frac{1}{N}} \cdot (A^{(m)})^{-1} \right] \cdot (y_m - y^{(j)}).$$

8. Переобчислення елементів матриці нечіткого розбиття

$$u_{mj}^{(l)} = \frac{1}{\sum_{k=1}^{M} \left( \frac{d^2_{A^{(m)}}(y^{(j)}, y_m)}{d^2_{A^{(m)}}(y^{(j)}, y_k)} \right)^{\frac{1}{w-1}}}.$$

9. Якщо $l \le Q$, то перехід на крок 11, інакше – крок 12.

10. Якщо $\left\| U^{(l)} - U^{(l-1)} \right\| > \varepsilon$, то перехід на крок 5, інакше – крок 1.

11. ЗУПИН.

У результаті виконання алгоритму мінімізується цільова функція:

$$J = \sum_{m=1}^{M} \sum_{j=1}^{n} u_{mj}^{w} \cdot d^2_{A^{(m)}}(y^{(j)}, y_m),$$

Формування класифікованої нечіткої навчальної матриці відбувається шляхом віднесення векторів-реалізацій за максимумом функції

належності відповідним класам. Проте у випадку $\max\limits_{m} u_{mj} < 0,54$ реалізація вважається погано класифікованою і не належить до жодного класу.

З метою адекватного порівняння якості розбиття даних при різних параметрах $M$ та $w$ початкову ініціалізацію центрів кластерів у алгоритмі Густафсона-Кесселя доцільніше здійснювати за певним осмисленим правилом, а не традиційним шляхом генерації псевдовипадкових чисел у матриці нечіткого розбиття. Тому для алгоритму, що розглядається, вирішено застосувати правило рівномірного розподілу центрів кластерів уздовж найбільшої діагоналі гіперкуба вхідних даних. Таке початкове наближення повинне зменшити сумарну помилку кластеризації, збільшити сумарну відстань між центрами кластерів, і в кінцевому рахунку зменшити кількість ітерацій основного алгоритму. Реалізація алгоритму розрахунку початкових координат центрів кластерів за даним правилом виконується у два етапи:

1. Пошук двох найбільш віддалених один від одного векторів $y^{(j_1)}$ та $y^{(j_2)}$, що належать вхідній некласифікованій матриці $\{y_i^{(j)} \mid i = \overline{1,N},\ j = \overline{1,n}\}$, використовуючи Евклідову метрику:

$$d_{\max} = d(y^{(j_1)}, y^{(j_2)}) = \max_{j,l} d(y^{(j)}, y^{(l)}),\ j = \overline{1,n},\quad l = \overline{1,n},$$

де $d(y^{(j)}, y^{(l)}) = \sqrt{\sum_{i=1}^{N} (y_i^{(j)} - y_i^{(l)})^2}$ – Евклідова відстань.

2. Безпосередній розрахунок координат центрів кластерів

$$y_{m,i} = y_i^{(j_1)} \pm \frac{d_{\max}}{M-1},\ m = \overline{1,M},\quad i = \overline{1,N},$$

де знак (+) відповідає умові $y_i^{(j_1)} < y_i^{(j_2)}$.

Вхідними даними для алгоритму інформаційно-ектремального навчання є в загальному випадку дійсний масив векторів-реалізацій класів розпізнавання (навчальна матриця) $\{y_{m,i}^{(j)} \mid m = \overline{1,M};\ i = \overline{1,N};\ j = \overline{1,n}\}$ і система нормованих допусків на ознаки розпізнавання $\{\delta_{H,i} \mid i = \overline{1,N}\}$, що задає область значень відповідних контрольних допусків.

Основними завданнями базового алгоритму навчання є пошук глобального максимуму інформаційного КФЕ навчання СППР у робочій (допустимій) області визначення його функції та оптимізація геометричних параметрів гіпереліпсоїдних контейнерів класів розпізнавання. При цьому здійснюється перевірка умови непокриття контейнером одного класу фокусів та геометричних центрів контейнерів інших класів. Розглянемо основні етапи базового алгоритму побудови гіпереліпсоїдних вирішальних правил:

1. Ініціалізація максимальної фокальної відстані $c_{\max}$, $0 \le c_{\max} \le \frac{N}{2}$. При цьому умова $c_{\max} = 0$ відповідає побудові гіперсферичних контейнерів.

2. Обчислення нижнього $A_{KH,i}$ та верхнього $A_{KB,i}$ контрольних допусків для кожної ознаки розпізнавання за формулами

$$A_{KH,i} = y_{1,i} - \delta\,; \quad A_{KB,i} = y_{1,i} + \delta\,,$$

де $y_{1,i}$ – вибіркове середнє значення $i$-ї ознаки у векторах-реалізаціях базового класу $X_1^o$, відносно якого будується система контрольних допусків (СКД) на ознаки розпізнавання.

3. Формування бінарної навчальної матриці $\left\| x_{m,i}^{(j)} \right\|$ за правилом

$$x_{m,i}^{(j)} = \begin{cases} 1, & \text{if} \quad A_{KH,i} \le y_{m,i}^{(j)} \le A_{KB,i}. \\ 0, & \text{if} \quad \text{else}. \end{cases}.$$

4. Обчислення для класу $X_m^o$ двійкового еталонного вектора $x_m$ за правилом

$$x_{m,i} = \begin{cases} 1, & \text{if} \quad \dfrac{1}{\text{n}} \sum\limits_{j=1}^{n} x_{m,i}^{(j)} > \rho_m, \\ 0, & \text{if} \quad \text{else}\,. \end{cases}.$$

де $\rho_m$ – рівень селекції координат вектора $x_m$ ($\rho_m = 0,5$).

5. Ініціалізація двійкових координат фокусів контейнера класу $X_m^o$

$$x_{m1} := x_m\,; \quad x_{m2} := x_m\,.$$

6. Обнулення лічильника кількості класів: $m := 0$.

7. $m := m + 1$.

8. Обнулення лічильника кроків зміни фокального радіуса: $c_m = 0$.

9. Формування для еталонного вектора $x_m$ множини $\{x_m^{(\upsilon)} \mid \upsilon = \overline{1,N}\}$, що складається з $N$ оточуючих його двійкових векторів з кодовою відстанню $d(x_m \oplus x_m^{(\upsilon)}) = c_m$, шляхом послідовних $N$ зсувів вліво на один розряд операції інверсії над послідовно розміщеними $c_m$ розрядами в еталонному векторі $x_m$.

10. Розбиття множини векторів $\{x_m^{(\upsilon)} \mid \upsilon = \overline{1,V}\}$ на $P$ пар фокусів $\{X_{m,p}^{|2|} \mid p = \overline{1,P}\}$. При цьому для кожної пари $X_{m,p}^{|2|} =< x_{m1}^{(p)},\ x_{m2}^{(p)} >$ повинна виконуватись умова $d(x_{m1}^{(p)} \oplus x_{m2}^{(p)}) = 2c_m$.

11. Обнулення лічильника пар фокусів $p = \overline{1,P}$: $p := 0$.

12. $p := p + 1$.

13. Ініціалізація фокусів координатами пари векторів $< x_{m1}^{(p)},\ x_{m2}^{(p)} >$.

14. Обнулення лічильника кроків зміни великої півосі гіпереліпсоїдного контейнера $d_m := 0$.

15. $d_m := d_m + 1$.

16. Обчислення інформаційного КФЕ $E_m$ навчання СППР розпізнавати реалізації класу $X_m^o$.

17. Якщо виконуються умови (2), то здійснюється перехід до пункту 15, інакше – до пункту 18.

18. Визначається оптимальне значення великої півосі $d_m^* = \arg\max\limits_{\{d_m\}} E_m^*$.

19. Якщо $p < P$, то виконується пункт 12, інакше – пункт 20.

20. Визначення оптимальної пари фокусів контейнера класу $X_m^o$:

$$< x_{m1}, x_{m2} >^* = \arg\max_{\{p\}} E_m^*(< x_{m1}^{(p)}, x_{m2}^{(p)} >).$$

21. $c_m := c_m + 1$.

22. Якщо $c_m \le c_{\max}$, то виконується пункт 9, інакше – пункт 23.

23. Визначення оптимального значення фокальної відстані

$$c_m^* = \arg\max_{\{c_m\}} E_m^*.$$

24. Якщо $m < M$, то виконується пункт 7, інакше – пункт 25.

25. ЗУПИН.

Оптимізацію СКД на ознаки розпізнавання доцільно здійснювати за паралельно-послідовним алгоритмом, що забезпечує прийнятну оперативність та високу точність обчислення КФЕ. При цьому за алгоритмом паралельної оптимізації СКД визначаються квазіоптимальні контрольні допуски, які для послідовного алгоритму приймаються як стартові.

Розглянемо основні етапи реалізації алгоритму навчання з оптимізацією СКД на ознаки розпізнавання:

1. Реалізується процедура паралельної оптимізації системи контрольних допусків (СКД) на ознаки розпізнавання при гіперсферичних контейнерах класів розпізнавання ($c_{\max} = 0$) [5]

$$\delta^* = \arg\max_{G_\delta}\{\max_{G_E} \overline{E}\}, \qquad (4)$$

де $\overline{E}$ – усереднений за алфавітом класів КФЕ навчання СППР; $G_\delta$ – область допустимих значень контрольних допусків на ознаки розпізнавання; $G_E$ – область допустимих значень інформаційного КФЕ.

2. Одержані за процедурою (4) квазіоптимальні допуски приймаються як стартові для процедури послідовної оптимізації контрольних допусків на ознаки розпізнавання.

3. Реалізується ітераційна процедура послідовної оптимізації поля контрольних допусків на ознаки розпізнавання при гіперсферичних контейнерах класів розпізнавання ($c_{\max} = 0$)

$$\{\delta_{K,i}^*\} = \arg\{\max_{G_{\delta_i}}\{\max_{G_E}\left[\bigotimes_{s=1}^{S} \max_{G_d} \overline{E}^{(s)}\right]\}\},\ i = \overline{1,N}, \qquad (5)$$

де $\overline{E}^{(s)}$ – усереднений за алфавітом класів КФЕ навчання СППР на $s$-му прогоні послідовної процедури оптимізації; $G_{\delta_i}$ – область допустимих значень поля контрольних допусків для $i$-ї ознаки; $G_E$ – область допустимих значень критерію оптимізації; $G_d$ – область допустимих значень радіусів контейнерів; $\otimes$ – символ операції повторення.

4. При оптимальному полі СКД на ознаки $\{\delta^*_{K,i} \mid i = \overline{1,N}\}$ здійснюється запуск базового алгоритму з метою гіпереліпсоїдної корекції розв'язувального правила ( $c_{\max} = \frac{N}{2}$ ).

Як КФЕ навчання використаємо модифіковану інформаційну міру Кульбака, в якій розглядається відношення повної ймовірності правильного прийняття рішень $P_t$ до повної ймовірність помилкового прийняття рішень $P_f$ . Для двохальтернативних гіпотез модифікований критерій Кульбака має вигляд

$$E_m^{(k)} = \left[ P_{t,m}^{(k)} - P_{f,m}^{(k)} \right] \cdot \log_2 \frac{P_{t,m}^{(k)}}{P_{f,m}^{(k)}} = \left| \begin{array}{l} P_{t,m}^{(k)} = p(\mu_1) \cdot D_{1,m} + p(\mu_2) \cdot D_{2,m} \\ P_{f,m}^{(k)} = p(\mu_1) \cdot \alpha_m + p(\mu_2) \cdot \beta_m \\ p(\mu_1) = \frac{n_1}{n_1 + n_2}; p(\mu_2) = \frac{n_2}{n_1 + n_2} \end{array} \right| =$$

$$= \frac{\left[ (n_1 \cdot D_{1,m}^{(k)} + n_2 \cdot D_{2,m}^{(k)}) - (n_1 \cdot \alpha_m{}^{(k)} + n_2 \cdot \beta_m{}^{(k)}) \right]}{n_1 + n_2} \cdot \log_2 \left( \frac{n_1 \cdot D_{1,m}^{(k)} + n_2 \cdot D_{2,m}^{(i)}}{n_1 \cdot \alpha_m{}^{(k)} + n_2 \cdot \beta_m{}^{(k)}} \right) =$$

$$= \frac{\left[ n_2 - n_1 + 2 \cdot (n_1 \cdot D_{1,m}^{(k)} - n_2 \cdot \beta_m{}^{(k)}) \right]}{n_1 + n_2} \cdot \log_2 \left( \frac{n_2 + (n_1 \cdot D_{1,m}^{(k)} - n_2 \cdot \beta_m{}^{(k)})}{n_1 - (n_1 \cdot D_{1,m}^{(k)} - n_2 \cdot \beta_m{}^{(k)})} \right), \quad (6)$$

де $D_{1,m}^{(k)}$ – перша достовірність, обчислена на $k$-му кроці навчання для $m$-го класу; $D_{2,m}^{(k)}$ – друга достовірність; $\alpha_m{}^{(k)}$ – помилка першого роду; $\beta_m{}^{(k)}$ – помилка другого роду; $n_1$ , $n_2$ – кількість реалізацій, що розмежовуються гіперповерхнею контейнера класу $X_m^o$ .

Оскільки навчальна вибірка є обмеженою за обсягом, то замість точнісних характеристик на практиці використовуються їх оцінки у вигляді емпіричних частот

$$D_{1,m}^{(k)} = \frac{K_{1,m}^{(k)}}{n_1}; \quad \beta_m^{(k)} = \frac{K_{2,m}^{(k)}}{n_2}, , \quad (7)$$

де $K_{1,m}^{(k)}$ – кількість подій, що характеризують належність реалізацій образу до контейнера класу $X_m^o$ , якщо вони дійсно є реалізаціями цього класу; $K_{2,m}^{(k)}$ – кількість подій, що характеризують належність реалізацій до контейнера класу $X_m^o$ , якщо вони насправді належать іншому класу.

Суми $K_{1,m}^{(k)}$ і $K_{2,m}^{(k)}$ обчислюються на $k$-му кроці навчання СППР за правилом

$$K_{1,m}^{(k)}[0] = 0; K_{2,m}^{(k)}[0] = 0; ,$$

$$if \ x_m^{(j)} \in X_m^o \ then \ K_{1,m}^{(k)}[j] := K_{1,m}^{(k)}[j-1] + 1; ,$$

$$if \ x_c^{(j)} \in X_m^o \ then \ K_{2,m}^{(k)}[j] := K_{2,m}^{(k)}[j-1] + 1,$$

де $x_c^{(j)}$ – $j$-я реалізація "чужого" класу $X_c^o$.

Визначення належності реалізації $x^{(j)}$, наприклад, класу $X_m^o$, для класифікатора з гіпереліпсоїдними контейнерами здійснюється за правилом

$$if\ \ d(x_{m1} \oplus x^{(j)}) + d(x_{m2} \oplus x^{(j)}) \le 2d_m \quad then\ \ x^{(j)} \in X_m^o \ \ else\ \ x^{(j)} \notin X_m^o,$$

де $d(x_{m1} \oplus x^{(j)})$, $d(x_{m2} \oplus x^{(j)})$ – кодові відстані між вектором $x^{(j)}$ і першим та другим фокусами контейнера класу $X_m^o$ відповідно; $d_m$ –значення великої півосі контейнера класу $X_m^o$; $\oplus$ – символ операції складання за модулем два.

Модифікація критерію Кульбака після відповідної підстановки оцінок (7) у вираз (6) набирає вигляду

$$E_m^{(k)} = \frac{\left[ n_2 - n_1 + 2 \cdot (K_1^{(k)} - K_2^{(k)}) \right]}{n_2 + n_1} \cdot \log_2 \left( \frac{n_2 + (K_1^{(k)} - K_2^{(k)}) + 10^{-r}}{n_1 - (K_1^{(k)} - K_2^{(k)}) + 10^{-r}} \right), \qquad (8)$$

де $10^{-r}$ – константа, що введена для виключення нескінченних піків у випадках нульових емпіричних частот при обчисленні критерію.

Нормовану модифікацію критерію (8) представимо у вигляді

$$E_m^{(k)} = \frac{E_m^{(k)}}{E_{\max}}, \qquad (9)$$

де $E_{\max}$ – значення критерію при $K_1^{(k)} = n_1 = n_{\min}$ та $K_2^{(k)} = 0$.

При цьому робоча область визначення функції інформаційного КФЕ обмежена як умовами (2), так і нерівностями $D_1 \ge 0,5$ і $D_2 \ge 0,5$.

Таким чином, алгоритм самонавчання інформаційно-екстремальної СППР полягає в ітераційній процедурі наближення глобального максимуму інформаційного КФЕ (10) до його граничного значення шляхом оптимізації параметрів функціонування СППР (1), що містять вхідні параметри алгоритму нечіткої кластеризації, СКД на ознаки розпізнавання та геометричні параметри контейнерів класів розпізнавання.

## ПЕРЕВІРКА РЕЗУЛЬТАТУ КЛАСТЕРИЗАЦІЇ ДАНИХ

Оцінка ефективності методу кластеризації та обґрунтованості отриманої структури кластерів традиційно вважається складною проблемою. Найбільш достовірні оцінки можна отримати шляхом порівняння результату кластеризації з ручним розбиттям даних на класи, що здійснюється експертами, які володіють додатковою інформацією прикладної області. Для цього існує багато так званих критеріїв зовнішньої валідації результату кластеризації, проте широкого застосування набула статистика Ренда [8,9].

Порівняння заданого розбиття вхідних даних $P$ із незалежною від нього структурою кластерів $C$, отриманою в результаті кластеризації вхідних даних, здійснюється шляхом підрахунку кількості пар точок даних $y^{(t)}$ та $y^{(l)}$ із множини вхідних даних $\{y^{(j)} \mid j = \overline{1,n}\}$, розглядаючи

чотири різних випадки залежно від способу розміщення $y^{(t)}$ та $y^{(l)}$ в $C$ та $P$:

1) $y^{(t)}$ і $y^{(l)}$ належать до одного кластера структури $C$ і до одного класу розбиття $P$;

2) $y^{(t)}$ і $y^{(l)}$ належать до одного кластера структури $C$, але до різних класів з розбиття $P$;

3) $y^{(t)}$ та $y^{(l)}$ належать до різних кластерів структури $C$, але до одного класу з розбиття $P$;

4) $y^{(t)}$ та $y^{(l)}$ належать до різних кластерів структури $C$ і до різних класів з розбиття $P$.

Число пар точок у чотирьох випадках позначають як $a$, $b$, $c$ та $d$ відповідно. При цьому статистику Ренда $R$ обчислюють за формулою

$$R = \frac{a+d}{a+b+c+d} . \qquad (10)$$

Як бачимо з визначення, чим більші значення індексів валідації, тим більш схожі між собою структура $C$ та розбиття $P$. При цьому значення індексу Ренда, коефіцієнта Жаккарда та FM-індекса знаходяться в діапазоні [0, 1], а значення Г-статистики лежить у діапазоні [-1, 1 ].

## ПРИКЛАД РЕАЛІЗАЦІЇ ЗДАТНОЇ САМОНАВЧАТИСЯ СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

Запропоновані алгоритми реалізовано при синтезі інтелектуальної СППР, яка є складовою частиною АСК процесом вирощування великогабаритних монокристалів за модифікованим методом Чохральського на установці типу "РОСТ 5" в науково-технічному комплексі "Інститут монокристалів" (м. Харків, Україна) [13,14].

Структурну схема здатної до самонавчання СППР в контурі АСК вирощування монокристалів показано на рис.1.

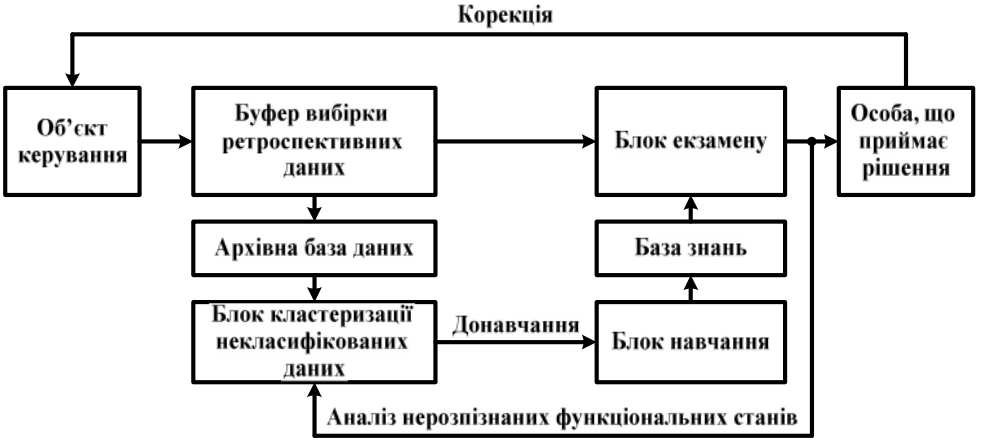


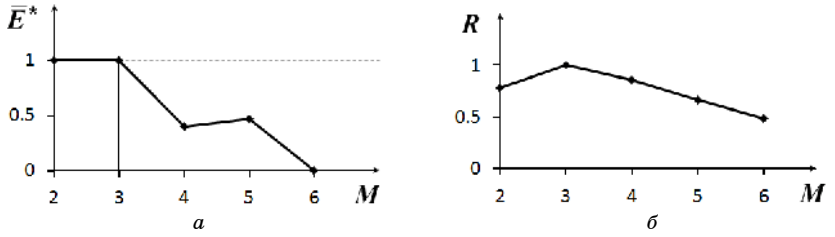

*Рисунок 1 – Інтелектуальна СППР в контурі АСК*

На рис. 1 показано, що інтелектуальна СППР отримує дані від об'єкта керування через буфер вибірки ретроспективних даних (OPC-сервер даних контролю за процесом) [15]. Кожна ретроспективна вибірка надходить у блок екзамену, де відбувається процес розпізнавання функціонального стану АСК, використовуючи вирішальні правила, що відновлюються за даними бази знань. Одночасно вибірки накопичуються в архівній реляційній базі даних. У разі нерозпізнаного функціонального стану і достатнього обсягу накопичених некласифікованих даних відбувається їх кластеризація для формування навчальних матриць нових класів розпізнавання. Блок навчання формує вирішальні правила і зберігає параметри їх відновлення в базі знань. При цьому технологічний цикл вирощування поділений на інтервали аналізу даних, для кожного з яких проводиться окреме навчання СППР. У прикладі розглянемо навчання СППР на часовому інтервалі від моменту досягнення довжини кристала 25 см і до моменту досягнення довжини 40 см.

Обсяг некласифікованої навчальної матриці становить $n = 270$, а розмірність структурованих векторів-реалізацій, що визначає кількість ознак розпізнавання, становить $N = 30$. При цьому 15 первинних ознак характеризують різні параметри теплових умов вирощування і стану локальних регуляторів, а як вторинні ознаки використовуються різниці першого та другого порядків над послідовностями найбільш інформативних трендів основних ознак.

Для зовнішньої валідації результатів кластеризації апріорна класифікація навчальної вибірки здійснювалась експертно за оцінками лабораторного контролю оптичних характеристик (рентгено-дефектоскопія), за вимірами відхилень діаметра монокристала від норми та за даними архівної історії аварійних ситуацій. У результаті вхідну некласифіковану матрицю було розбито на три класи по 90 векторів-реалізацій в кожному. Ці класи характеризували якість монокристала і відповідні функціональні стани АСК.

Вибір оптимальної кількості кластерів для алгоритму кластеризації Густафсона-Кесселя здійснювався за максимумом усередненого КФЕ навчання (10) (рис. 2 а), а зовнішня валідація такого вибору здійснювалася за максимумом індексу Ренда (10) (рис. 2 б).



*а* *б*

*Рисунок 2 – Залежність усередненого нормованого інформаційного КФЕ (а) та індексу Ренда (б) від заданої кількості кластерів розбиття*

Аналіз рис. 2 показує, що адаптивний алгоритм ефективно розв'язує задачу визначення кількості кластерів. При цьому за оптимальну кількість кластерів обирається максимальне число кластерів, що забезпечує максимум КФЕ навчання, і тому $M^* = 3$.

Вибір оптимального показника нечіткості для алгоритму кластеризації Густафсона-Кесселя при кількості кластерів $M = 3$ здійснюється за максимумом усередненого КФЕ навчання (9), що показано на рис. 3, де штрихова лінія відповідає навчанню з гіперсферичними контейнерами класів розпізнавання $(c_{\max} = 0)$, а суцільна лінія відповідає використанню гіпереліпсоїдних контейнерів ($c_{\max} = N/2$). Зовнішня

валідація результату кластеризації для кожного показника нечіткості здійснюється за максимумом індексу Ренда (10) (рис. 4).

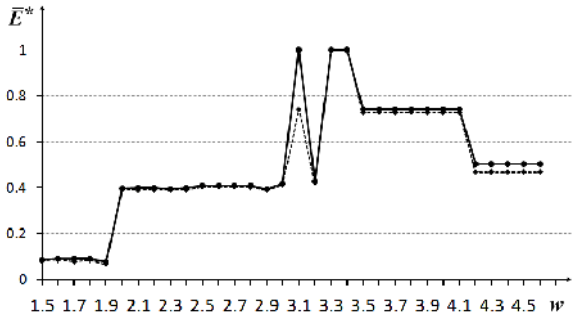



*Рисунок 3 – Залежність усередненого нормованого інформаційного КФЕ від заданого показника нечіткості в алгоритмі кластеризації Густафсоона-Кесселя при* $M = 3$

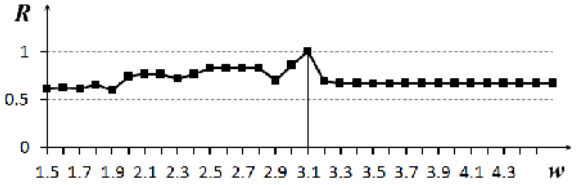



*Рисунок 4 – Залежність індексу Ренда від заданого показника нечіткості в алгоритмі кластеризації Густафсоона-Кесселя при* $M = 3$

Аналіз рис. 3-4 показує, що адаптивний алгоритм ефективно розв'язує і задачу визначення оптимального показника нечіткості, для вибору якого поки що не існує теоретично обґрунтованого правила, і на практиці обирають $w = 2$. При цьому в адаптивному алгоритмі за оптимальний показник нечіткості обирається мінімальне значення (але $w > 1$), що забезпечує максимум КФЕ навчання, і тому $w^* = 3{,}1$. До того ж, як бачимо з рис. 3-4, гіпереліпсоїдні розв'язувальні правила інформаційно-екстремального алгоритму ефективніші порівняно з гіперсферичними.

Процес навчання інформаційно-екстремальної СППР за навчальною матрицею, що отримана при оптимальних параметрах алгоритму Густафсона-Кесселя, відображено на рис. 5-7.

На рис. 5 показано динаміку зміни максимуму усередненого нормованого інформаційного КФЕ (9) при оптимізації СКД за паралельно-послідовним алгоритмом з побудовою гіперсферичних контейнерів класів розпізнавання в радіальному базисі бінарного простору ознак. Тут і далі заштрихована ділянка графіка позначає робочу область визначення інформаційного КФЕ.

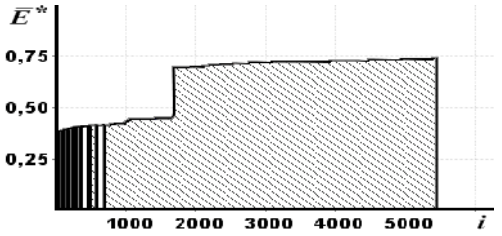
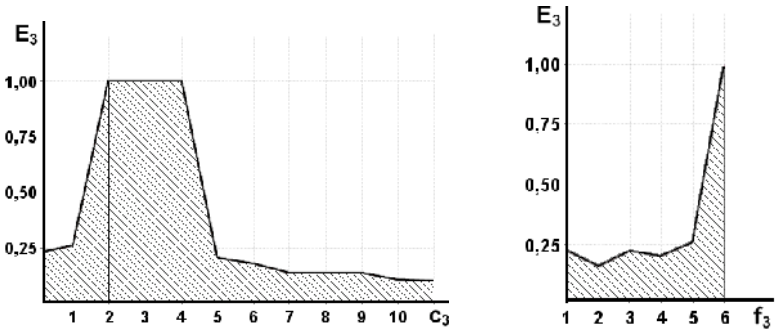



*Рисунок 5 – Графік зміни максимумів КФЕ при оптимізації СКД для гіперсферичного класифікатора*

Аналіз рис. 5 показує, що оптимальний вектор СКД був одержаний на 5457 кроці навчання, на якому досягнуто глобальний максимум усередненого інформаційного КФЕ $\overline{E^*} = 0,748$, що свідчить про наявність перетину контейнерів класів розпізнавання.

З метою зменшення перетину контейнерів класів розпізнавання після паралельно-послідовної оптимізації СКД додатково запускається базовий алгоритм для гіпереліпсоїдної корекції вирішальних правил. У результаті цього для всіх контейнерів класів розпізнавання були визначені оптимальні значення фокальних відстаней, які відповідно дорівнювали $c_1^* = 0$, $c_2^* = 0$ і $c_3^* = 2$ (у кодових одиницях). Це свідчить, що контейнери класів $X_1^0$ та $X_2^0$ залишилися гіперсферичними, а контейнер класу $X_3^0$ деформувався у гіпереліпсоїд обертання з ексцентриситетом $e_2^* = \frac{c_2^*}{d_2^*}$, де $d_2^*$ – оптимальна довжина великої півосі. Графік залежності нормованого інформаційного КФЕ (9) від фокальної відстані та вибору різних фокусів контейнера класу $X_3^0$ у процесі його обертанні в бінарному просторі ознак показано на рис. 6.

Аналіз рис. 6 показує, що значення фокальної відстані контейнера класу $X_3^0$ від 2 до 4 кодових одиниць забезпечує граничне значення КФЕ. При цьому обираємо значення рівне $c_2^* = 2$, що відповідає мінімальній корекції вирішального правила.

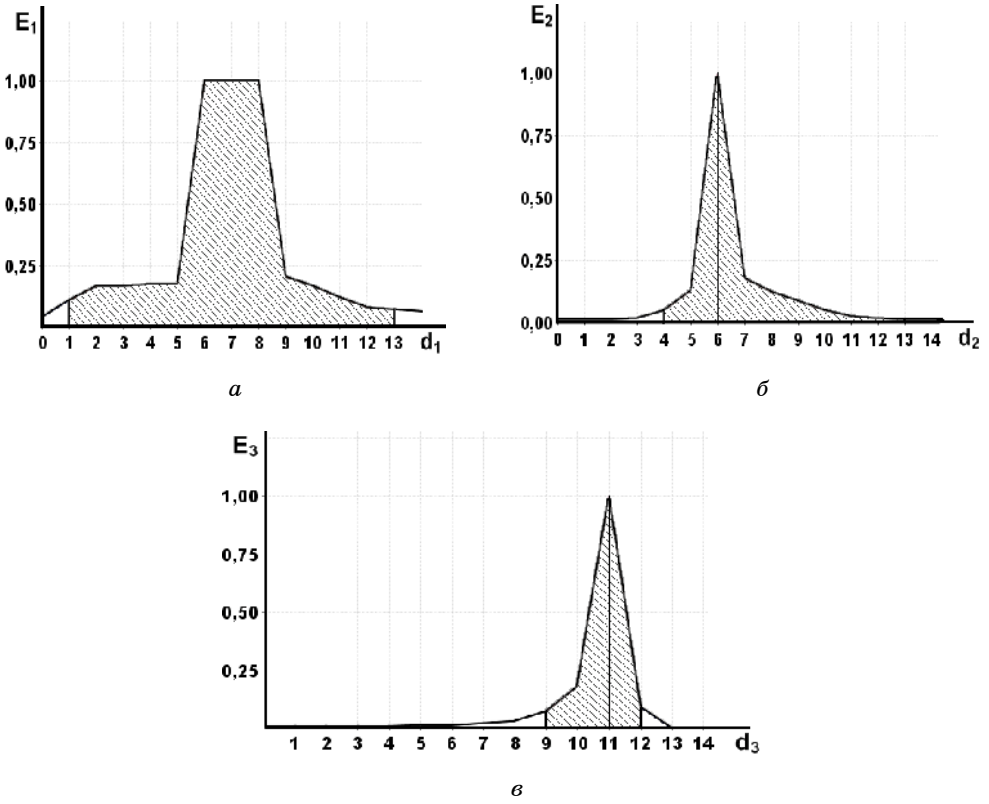



*а* *б*

*Рисунок 6 – Залежність інформаційного КФЕ для класу $X_3^0$: а) від фокальної відстані його контейнера; б) від вибору пар фокусів на оптимальній фокусній відстані*

На рис. 7 наведено графіки залежності нормованого КФЕ (9) від довжини великої півосі гіпереліпсоїдних контейнерів класів $X_1^0$, $X_2^0$ та $X_3^0$, що відновлюються у бінарному просторі ознак. Для класів $X_1^0$ та $X_2^0$ довжина великої півосі є радіусом їхніх гіперсферичних контейнерів.

Аналіз рис. 7 показує, що оптимальні значення великих півосей контейнерів класів розпізнавання відповідно дорівнюють $d_1^* = 6$, $d_2^* = 6$ і $d_3^* = 11$.




*а* *б*

*в*

*Рисунок 7 – Графіки залежності КФЕ від значень великих півосей гіпереліпсоїдних контейнерів: а) клас $X_1^0$; б) клас $X_2^0$; в) клас $X_3^0$*

Таким чином, завдяки інформаційному КФЕ навчання вдалось організувати адаптивний механізм кластеризації за алгоритмом Густафсона-Кесселя з оптимізацією таких вхідних параметрів, як кількість кластерів розбиття та показник нечіткості. При цьому зовнішній критерій валідації результату кластеризації підтвердив ефективність такого підходу і підкреслив перевагу застосування в інформаційно-екстремальному алгоритмі гіпереліпсоїдної корекції вирішальних правил.

## ВИСНОВКИ

1. На основі алгоритму нечіткої кластеризації Густафсона-Кесселя та алгоритму інформаційно-ектремального навчання з гіпереліпсоїдною корекцією розв'язувальних правил розроблено адаптивний механізм самонавчання СППР, яка є складовою частиною АСК технологічного процесу вирощування великогабаритних сцинтиляційних монокристалів.

2. Фізичне моделювання за даними архівної історії вирощування сцинтиляційних монокристалів показало, що використання інформаційно-екстремального навчання для оцінки якості розбиття ефективно вирішує проблему вибору кількості кластерів розбиття та показника нечіткості кластерів у просторі ознак, що використовуються як вхідні параметри алгоритму кластеризації. При цьому гіпереліпсоїдна корекція вирішальних правил інформаційно-екстремального алгоритму забезпечила високий ступінь збігу ручного розбиття вхідних даних з розбиттям отриманим у результаті адаптивної кластеризації.

**ИНФОРМАЦИОННО-ЭКСТРЕМАЛЬНОЕ ОБУЧЕНИЕ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ С АДАПТИВНОЙ КЛАСТЕРИЗАЦИЕЙ ДАННЫХ**

***В. В. Москаленко,***
*Сумский государственный університет, г. Сумы*

*Рассматривается метод адаптивной кластеризации данных на основе алгоритма Густафсона-Кесселя и информационно-экстремального обучения с гиперэллипсоидальными контейнерами классов распознавания, восстанавливаемых в радиальном базисе бинарного пространства признаков распознавания.*

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

**ADAPTIVE SELF-LEARNING OF INFORMATION-EXTREME DECISION SUPPORT SYSTEM**

***V. Moskalenko***
*Sumy State University, Sumy*

*The method of adaptive data clustering based on Gustafson-Kessel algorithm and information-extreme learning algorithm with hyper-ellipsoidal containers of recognition classes that recovering in the radial basis of binary feature space is considered.*

***Key words:*** *cluster analysis, decision support system, optimization, learning, information criterion, class recognition.*

## СПИСОК ЛІТЕРАТУРИ

1. Симанков В. С. Адаптивное управление сложными системами на основе теории распознавания образов / В. С. Симанков, Е. В. Луценко. – Краснодар: Техн. ун-т Кубан. гос. технол. ун-та, 1999. – 318 с.
2. Евменов В. П. Интеллектуальные системы управления / В. П. Евменов. – М.: Книжный дом «ЛИБРОКОМ», 2009. – 304 с.
3. Турбович И. Т. Опознавание образов. Детерминированно-статистический подход / И. Т. Турбович, В. Г. Гитис, В. К. Маслов. – М.: Наука, 1971. – 246 с.
4. Ситник В. Ф. Системи підтримки прийняття рішень: навч. посіб. / В. Ф. Ситник – К.: КНЕУ, 2004. – 614 с.
5. Довбиш А. С. Основи проектування інтелектуальних систем: навчальний посібник / А. С. Довбиш. – Суми: Видавництво Сум ДУ. – 2009. – 171 с.
6. Everitt B. S., Landau S., Leese M. et al. (2011) Cluster Analysis. 5th ed. Wiley, 2011.
7. M. Ameer Ali, Gour C. Karmakar, Laurence S. Dooley. Review on Fuzzy Clustering Algorithms / // IETECH Journal of Advanced Computations. – 2008. - Vol. 2 (3). - P. 169-181.
8. Xu R., Wunsch II D.C. (2009) Clustering, Wiley and Sons.
9. Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis. On Clustering Validation Techniques // Journal of Intelligent Information Systems. - December 2001. – Volume 17, Issue 2-3. – P. 107–145.
10. Кузьмин И.В. Оценка эффективности и оптимизация автоматизиро-ванных систем контроля и управления / И. В. Кузьмин.– М.: Сов. радио, 1971.– 296 с.
11. Краснопоясовський А. С. Класифікаційний аналіз даних: навчальний посібник / А. С. Краснопоясовський. – Суми: Видавництво СумДУ, 2002.–159 с.
12. Довбиш А. С. Оптимізація словника ознак розпізнавання для інформаційно-екстремального гіпереліпсоїдного класифікатора / А. С. Довбиш, В. В. Москаленко // Вісник НТУ "ХПІ": збірник наукових праць. Тематичний випуск: Системний аналіз, управління та інформаційні технології. – Харків: НТУ "ХПІ", 2012. – № 30. – С. 65-77.
13. Суздаль В. С. Сцинтилляционные монокристаллы: автоматизированное выращивание / В. С. Суздаль, П. Е. Стадник, Л. И. Герасимчук, Ю. М. Епифанов. – Х. : ИСМА, 2009. – 260 с.
14. Горилецкий В. И. Рост кристаллов / В. И. Горилецкий, Б. В. Гринёв, Б. Г. Заславский, Н. Н. Смирнов, В. С. Суздаль. – Харків: Акта. – 2002.– 536с.
15. Денисенко В. В. Компьютерное управление технологическим процессом, экспериментом, оборудованием / В. В. Денисенко.– М.: Горячая линия-Телеком, 2009. – 608 с.

*Надійшла до редакції 11 жовтня 2012 р.*