УДК 681.518:004.93.1

ІНФОРМАЦІЙНО-ЕКСТРЕМАЛЬНА СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ У РЕЖИМІ КЛАСТЕР-АНАЛІЗУ

А.С. Довбиш, *д-р техн. наук, професор;*
В.О. Востоцький, *аспірант,*
Сумський державний університет, м. Суми

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

ВСТУП

Підвищення ефективності та оперативності керування виробничими процесами органічно пов'язане із розробленням та впровадженням інтелектуальних інформаційних технологій. Застосування здатних самонавчатися АСКТП 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 перетинання класів розпізнавання, який має місце в практичних задачах 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), головними завданнями якої є оцінка поточного функціонального стану технологічного процесу та вироблення відповідних керуючих команд для особи, що приймає рішення. У роботах [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 функціонує в режимі кластер-аналізу з навчанням. 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 в режимі навчання СППР 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, шляхом оптимізації координат структурованого вектора параметрів функціонування $g_m =< g_{m,1},\ldots,g_{m,\,q},\ldots,g_{m,Q} >$, для яких відомі обмеження $R_q(g_1,\ldots,g_Q) \le 0$. При цьому усереднений за алфавітом $\{X_m^o\}$ інформаційний критерій функціональної ефективності (КФЕ) навчання системи досягає глобального максимуму в робочій (допустимій) області визначення його функції:

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

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.

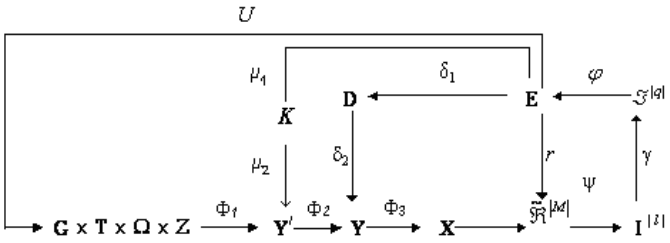
МАТЕМАТИЧНА МОДЕЛЬ

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

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

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.

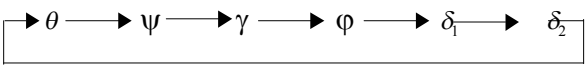
Категорійну модель процесу навчання СППР подамо у вигляді діаграми відображення множин, що застосовуються у режимі кластер-аналізу.



(2)

У діаграмі (2) 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: $\tilde{\Re}^{|M|} \to \mathrm{I}^{|l|}$, перевіряє основну статистичну гіпотезу про належність реалізацій $\{x_{(m)}^{(j)} \mid j = 1\ ,\ n\} \in X_{(m)}$ 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 розпізнавання реалізацій класу X_m^o замикається оператором r: $E \to \tilde{\Re}^{|M|}$. У діаграмі (2) терм-множина D складається із допустимих значень контрольних допусків на ознаки розпізнавання, які оптимізуються безпосередньо контуром операторів



.

Контур кластеризації в діаграмі (2) 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 початкової нечіткої класифікованої навчальної матриці Y здійснювалося шляхом поєднання дистанційних алгоритмів прямої кластеризації FOREL [6] і QUALITY TRESSPASS [5]. При цьому в процесі ітераційного пошуку глобального максимуму інформаційного КФЕ (3) 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 IEI-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 вирішальних правил.

Як критерій оптимізації розглянемо модифікацію критерію Кульбака, яка має такий вигляд [8]

$$E_m^{(k)} = 0,5 \log_2 \left(\frac{D_1^{(k)} + D_2^{(k)}}{\alpha^{(k)} + \beta^{(k)}} \right) \left[(D_1^{(k)} + D_2^{(k)}) - (\alpha^{(k)} + \beta^{(k)}) \right] = \\ = \log_2 \left(\frac{2 - (\alpha^{(k)} + \beta^{(k)})}{\alpha^{(k)} + \beta^{(k)}} \right) \left[1 - (\alpha^{(k)} + \beta^{(k)}) \right], \tag{3}$$

де α, β, D_1, D_2 -точнісні характеристики: помилки першого та другого родів , перша та друга достовірності.

Оскільки інформаційний критерій є функціоналом від точнісних характеристик, то при репрезентативному обсязі навчальної вибірки потрібно користуватися їх оцінками:

$$D_{1,m}^{(k)}(d)=\frac{K_{1,m}^{(k)}}{n_{\min}};\ \alpha_m^{(k)}(d)=\frac{K_{2,m}^{(k)}}{n_{\min}};\ \beta_m^{(k)}(d)=\frac{K_{3,m}^{(k)}}{n_{\min}};\ D_{2,m}^{(k)}(d)=\frac{K_{4,m}^{(k)}}{n_{\min}},\qquad(4)$$

де $K_{1,m}^{(k)}$ – кількість подій, які означають належність реалізацій образу контейнеру $K_{1,mk}^{o}$, якщо дійсно $\{x_1^{(j)}\}\in X_1^o$; $K_{2,m}^{(k)}$ – кількість подій, які означають неналежність реалізацій контейнеру $K_{1,m}^{o}$, якщо дійсно $\{x_1^{(j)}\}\in X_1^o$; $K_{3,m}^{(k)}$ – кількість подій, які означають належність реалізацій контейнеру $K_{1,m}^{o}$, якщо вони насправді належать класу X_2^o; $K_{4,m}^{(k)}$ – кількість подій, які означають неналежність реалізацій контейнеру $K_{1,m}^{o}$, якщо вони насправді належать класу X_2^o; $n_{\min}$ – мінімальний обсяг репрезентативної навчальної вибірки.

Робоча модифікація критерію Кульбака після відповідної підстановки оцінок (5) у вираз (4) набирає вигляду

$$E_m^{(k)}=\frac{1}{n}\log_2\left\{\frac{2n+10^{-r}-\left[K_2^{(k)}+K_3^{(k)}\right]}{\left[K_2^{(k)}+K_3^{(k)}\right]+10^{-r}}\right\}\left[n-(K_2^{(k)}+K_3^{(k)})\right].\qquad(5)$$

Розглянемо узагальнену схему гібридного (дистанційно-інформаційного) алгоритму кластер-аналізу у рамках IEI-технології на i-й ітерації. Вхідними даними є: загальна кількість реалізацій n_Σ в мультимодальному розподілі, мінімальна і максимальна кількість реалізацій у таксоні, відповідно $n_{\min}$ і $n_{\max}$, та необхідна кількість реалізацій n_i у кластерах, яку потрібно досягти на поточному кроці формування та навчання системи. Параметри $delta_i$, ρi коригують дистанційні характеристики розподілу та формування кластерів , безпосередньо впливаючи на однорідність реалізацій у бінарному просторі під час реалізації процедури бінеаризації та формуючи центри кластерів під час реалізації алгоритму розрахунку центрів кластерів.

Процедура формування вхідного математичного опису на і-му кроці навчання для кожного зі створюваних кластерів має такий вигляд:

1. Вибір довільної початкової точки формування кластера. Радіус кластера вважати таким, що дорівнює нулю.
2. Формування еталонного вектора кластера за всіма реалізаціями утвореного кластера.
3. Формування центра контейнера за всіма реалізаціями даного кластера у вигляді еталонного вектора.
4. Збільшення радіуса контейнера, що описує кластер.
5. Приєднання до кластера реалізацій, що входять в утворений контейнер кластера.
6. Центрування кластера за доданими реалізаціями. Якщо кількість реалізацій у кластері менша за n_i, то перехід до п.2.
7. Оптимізація бінарного простору та перехід до алгоритму навчання на кроці 4. Якщо вибрано таке значення радіуса, що кількість реалізацій

дорівнює n_i, а центри кластерів, розраховані на кроках 2 та 6, збігаються, – перехід до пункту 8.

8. ЗУПИН. Формування нечіткого розбиття простору ознак за дистанційними критеріями.

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 алгоритму навчання СППР у рамках ІЕІ-технології.

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.

ПРИКЛАД ЗАСТОСУВАННЯ ІНФОРМАЦІЙНО-ЕКСТРЕМАЛЬНОГО АЛГОРИТМУ КЛАСТЕР-АНАЛІЗУ

Реалізація алгоритму здійснювалася за некласифікованою навчальною вибіркою, отриманою за результатами моніторингу технологічного процесу виробництва складних мінеральних добрив у ВАТ «Сумихімпром». Некласифікована навчальна матриця складалася з 400 структурованих векторів-реалізацій, кожна з яких містила 55 ознак. Оскільки ознаки мали різні шкали вимірювання, було виконано перетворення вихідної області визначення значень ознак в інтервал [0...255] за методом зведених шкал [11].

З метою підвищення функціональної ефективності СППР було застосовано такі параметри функціонування СППР: кількість реалізацій у кластері, мінімальна відстань від центра утвореного кластера до реалізації, що відбиралася з некласифікованої навчальної матриці, а також значення системи контрольних допусків на ознаки розпізнавання та рівнів селекції координат двійкових еталонних векторів-реалізацій образів, які оптимізували кодові відстані реалізацій та центр кластера за оптимальним значенням КФЕ. У процесі навчання СППР визначалися оптимальні значення параметрів функціонування шляхом багатоциклічної ітераційної процедури пошуку глобального максимуму усередненого значення КФЕ.

Графік залежності усередненого значення КФЕ $\overline{E}$ від кількості реалізацій n у створюваному кластері наведено на рис.1.

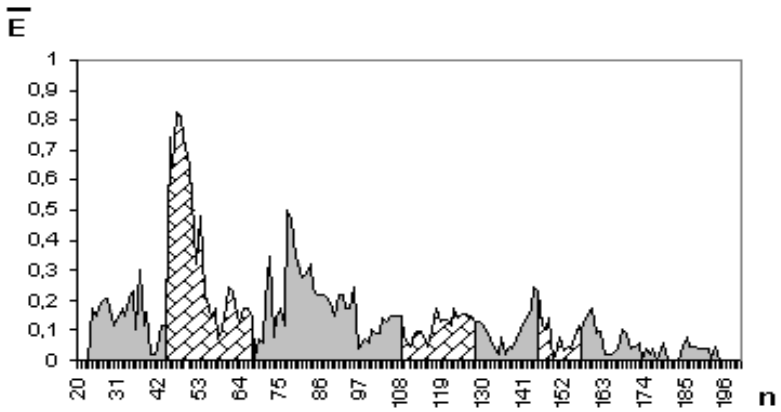


Рисунок 1 – Графік залежності усередненого КФЕ після навчання СППР від кількості реалізацій в утворюваному кластері

Аналіз рис. 1 показує, що максимальне середнє значення КФЕ в робочій області (заштриховано) було отримано при кількості реалізацій n = 43. При цьому тенденція до зменшення максимального значення КФЕ пояснюється розмиванням приєднаними реалізаціями утворених класів розпізнавання, що призводить до збільшення ступеня перетинання класів.

Рис. 2 ілюструє процес оптимізації системи контрольних допусків на ознаки розпізнавання для утвореного кластера із п'яти класів, кожний з яких містив по 43 реалізації.

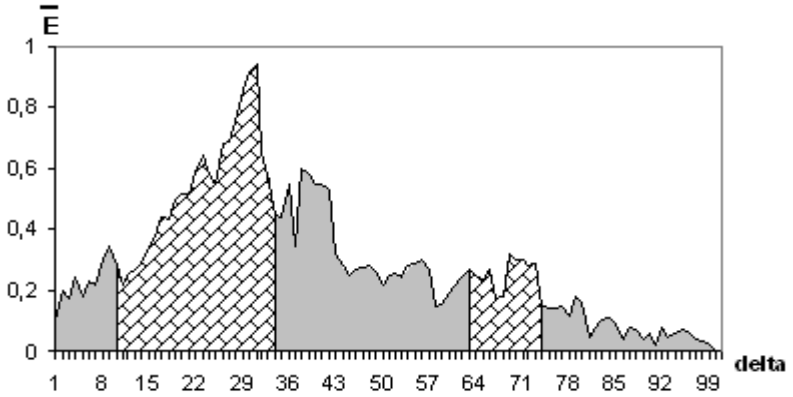


Рисунок 2– Графік залежності усередненого КФЕ від параметра поля контрольних допусків на ознаки розпізнавання в процесі навчання СППР

Аналіз рис. 3 показує, що максимальне середнє значення КФЕ в робочій області дорівнює $\overline{E}^{*} = 0,87$ при оптимальному параметрі поля допусків delta=31 відносних одиниць.

Таким чином, запропонований гібридний алгоритм, який містить як інформаційну міру схожості класів розпізнавання, так і дистанційні критерії, дозволяє у рамках IEI-технології формувати нечіткі апріорно класифіковані навчальні матриці та здійснювати у процесі навчання СППР оптимізацію параметрів функціонування, що впливають на достовірність розпізнавання класів.

ПЕРСПЕКТИВИ ЗАСТОСУВАННЯ

У загальному випадку запропонований алгоритм кластер-аналізу в рамках IEI-технології може використовуватися в задачах керування слабо формалізованими технологічними процесами в хімічній, металургійній, харчовій та інших галузях соціально-економічної сфери суспільства, які відбуваються за умов апріорної невизначеності. З метою подальшого удосконалення запропонованого алгоритму необхідно здійснювати оптимізацію інших параметрів функціонування СППР, що впливають на її функціональну ефективність, та перейти від лінійної структури алгоритму навчання в режимі кластер-аналізу до ієрархічної.

ВИСНОВКИ

1. Запропонований алгоритм автоматичної класифікації дозволяє будувати за апріорно некласифікованою багатовимірною вибіркою, одержаною у процесі функціонування АСКТП, нечіткі апріорно класифіковані навчальні матриці, що дозволяє використовувати їх у рамках IEI-технолгії для оптимізації просторово-часових параметрів функціонування СППР.

2. Шляхом фізичного моделювання доведено, що інформаційний критерій оптимізації параметрів функціонування СППР дозволяє оцінювати її функціональну ефективність як при збільшенні потужності алфавіту класів розпізнавання, так і навчальної матриці. При цьому для поточного кластера існує можливість побудови безпомилкових за навчальною вибіркою розв'язувальних правил.

SUMMARY

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.

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., Червоненкис А.Я. Теория распознавания образов: (статистические проблемы обучения). – 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. Мандель И.Д. Кластерный анализ. — М.: Финансы и статистика, 1988.
6. Загоруйко Н.Г., Елкина В.Н., Лбов Г.С. Алгоритмы обнаружения эмпирических закономерностей. – Новосибирск: Наука. –1985. – 110 с.
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., Козинець М.В., Котенко С.М. Оптимізація контрольних допусків на ознаки розпізнавання в інформаційно-екстремальних методах автоматичної класифікації // Вісник Сумського державного університету. Серія Техніка. - №1. - 2007. – С. 169-178.
11. Краснопоясовский А.С., Калюжная С.А. О выборе обобщённой шкалы для входных переменных при многофакторном эксперименте // Автоматизированные системы управления. - Харьков: Харьк. авиац. ин-т, 1984.– Вып. 5.– С. 114–118.

Надійшла до редакції 24 квітня 2010 р.