ISSN 1997-9266. Вісник Вінницького політехнічного інституту. 2015. № 4

УДК 681.518:004.93.1

С. А. Петров[1]
Н. В. Лисак[2]
Ю. В. Міронова[2]

ГІБРИДНИЙ АЛГОРИТМ КЛАСТЕР-АНАЛІЗУ ДЛЯ ФОРМУВАННЯ АПРІОРНОГО РОЗБИТТЯ ПРОСТОРУ ОЗНАК НА КЛАСИ ЗНАНЬ В СИСТЕМАХ ДИСТАНЦІЙНОГО НАВЧАННЯ

[1]Сумський державний університет;
[2]Вінницький національний технічний університет

Запропоновано модифікацію алгоритму k-means, ідея вдосконалення якого полягає у комбінованому використанні критерію оцінки помилки кластеризації та інформаційного критерію функціональної ефективності, що визначає рівень достовірності побудованих вирішальних правил визначення належності реалізацій до певного класу знань. При цьому використання комбінованого статистичного та інформаційного підходів дозволило включити такий параметр кластеризації як кількість кластерів в інтеграційну оптимізаційну процедуру та, базуючись на природній структурі розподілення векторів реалізацій результатів тестування слухачів в N-вимірному просторі ознак, розпізнавання дозволило знайти оптимальні геометричні параметри контейнерів класів, які характеризують рівні знань студенів в системах дистанційного навчання.

Ключові слова: кластеризація, k-means, критерій функціональної ефективності, критерій оцінки помилки кластеризації, системи дистанційного навчання.

Вступ

Під час виконання інформаційного аналізу і синтезу адаптивної системи керування дистанційним навчанням (СКДН) на етапі навчання системи виникає необхідність формування апріорного розбиття простору ознак розпізнавання на класи, яке у процесі побудови вирішальних правил корегується деяким оптимальним способом [1]. Фактично таку задачу розв'язує безпосередньо викладач, відносячи результати тестування до відповідної суб'єктивної та інтуїтивної оціночної шкали. Але оскільки за швидкого зростання кількості слухачів, набуття популярності заочно-дистанційної форми навчання, спеціалізованих навчальних сертифікованих курсів, збереження традиційної системи оцінювання знань вимагає постійного збільшення матеріальних витрат, підвищення тиску на професорсько-викладацький склад, то задача розробки алгоритмів машинного оцінювання знань студентів за методами сучасних прогресивних інтелектуальних технологій є актуальною [2].

Під час аналізу подібних процесів та систем виникає проблема їх формального описання, яка пов'язана з багатовимірністю вхідних параметрів та їх представлення. Найдієвішим інструментом для дослідження таких процесів є кластерний аналіз.

Окрім цього, в задачах з невеликою кількістю об'єктів, де важливішим є аналіз структури даних, а також існує окрема проблема визначення кількості кластерів, використовують ієрархічні методи, такі як: метод ближнього сусіда (single linkage), метод дальнього сусіда (complete linkage), метод середнього зв'язку (pair group average), центроїдний метод (метод медіан зв'язку) [3, 4]. У випадку, якщо можна апріорно визначити кількість кластерів або ця кількість є відомою, то для класифікації частіше за все використовують паралельні кластер-процедури, у яких розбиття виконується згідно з певним функціоналом якості.

Всі відомі методи кластер-аналізу об'єднуються такими властивостями:

— застосування дистанційних критеріїв схожості (евклідова відстань, зважена евклідова відстань та відстань Хемінга.), які є частковими випадками узагальненої метрики Махаланобіса [5]

$$d_{ij}^2 = \sum_{k=1}^{q} v_k \left(z_j^{(k)} - z_j^{(k)} \right)^2, \quad (1)$$

де $z_i^k = \left(U_k', X_i\right)$ — одновимірна проекція лінійних комбінацій вхідних змінних, $v_k = \phi(Q_k)$ — лінійна комбінація $U_i, \forall i$ в новому базисі коваріаційної матриці S; потужність множини кластерів $\left\{X_m^0 \mid m = 1...M\right\}$ відома апріорно. Тут X_{ij} — матриця вхідних даних, навчальна матриця.

Але до теперішнього часу теорія автоматичної класифікації не дає відповіді на такі запитання:

1. Наскільки вдалим є побудоване на етапі навчання СКДН розбиття простору ознак на класи, яке визначає вирішальні правила.

2. Чи є оптимальним (тут і далі в інформаційному розумінні) число класів розпізнавання?

Досліджуючи ці важливі питання, як правило, здійснюється розвідувальний аналіз результатів тестування шляхом оцінки їх статистичних характеристик і емпіричних закономірностей [6]. При цьому для оцінки апріорного розбиття часто використовують таку величину, як внутрішньо-класовий розкид [7]

$$W = \sum_{j=1}^{k} W_j, \quad W_j = \sum_{X_i \in G_j} d^2\left(X_i, \bar{X}\right), \quad (2)$$

де $\bar{X}$ — $\frac{1}{n}\sum_{i=1}^{n} X_i$ — загальний центр ваги; $X_j = \frac{1}{n_j} \sum_{x_i \in j} X_i$ — центр ваги j-ї групи; n_j — кількість об'єктів в групі G_j; k — кількість об'єктів, на які розбита множина $G = \left\{G_1, G_2, ..., G_k\right\}$.

Застосування дистанційних критеріїв (1) і статистичних показників типу (2) не виключає ситуацію, коли кластер можуть формувати геометрично близькі вектори-реалізації образу, але при цьому будуть суттєво відрізнятися між собою їх відповідні координати — значення ознак розпізнавання [1]. Тому, наприклад, у працях [7, 8] запропоновано підхід, що ґрунтується на виявленні у емпіричних даних об'єктивно існуючої функціональної закономірності. Такий підхід має низку недоліків:

— висока обчислювальна трудомісткість, пов'язана із необхідністю оброблення великих масивів емпіричних даних;

— наявність апріорно чіткого розбиття, що не є характерним для більшості практичних задач контролю та керування;

— лінійній вигляд функції регресії, яка не у всіх випадках адекватно описує зв'язок.

Таким чином, аналіз існуючих сучасних методів кластер-аналізу дозволяє зробити такі два основні висновки:

а) методи кластер-аналізу, що виключають процес машинного навчання, характеризуються низькою достовірністю прийняття рішень відносно методів із навчанням;

б) відомі методи кластер-аналізу, що навчаються, не дозволяють побудувати безпомилковий за навчальною матрицею класифікатор, оскільки ігнорують перетин класів розпізнавання, тобто носять модельний характер.

Однією із перспективних технологій аналізу та синтезу адаптивних систем керування слабо формалізованими процесами є інформаційно-екстремальна інтелектуальна технологія (ІЕІТ), що ґрунтується на реалізації принципу максимізації інформаційної спроможності системи шляхом введення в процесі оптимізації просторово-часових параметрів функціонування додаткових інформаційних обмежень з метою побудови в режимі навчання безпомилкового за навчальною матрицею класифікатора [9].

У статті розглядається алгоритм побудови у рамках ІЕІТ апріорного нечіткого розбиття класів розпізнавання (рівнів знань) студентів заочної і дистанційної форм навчання, що дозволяє сформувати вхідну навчальну матрицю для адаптивної СКДН.

Постановка задачі дослідження

Нехай дано масив вхідних даних $\left\{Y_{0,i}^{(j)} \mid i = \overline{1, N};\ j = \overline{1, n}\right\}$, де N — кількість ознак розпізнавання (результатів тестування в академічній групі, що атестується за поточний навчальний модуль дис-

танційного курсу); n — обсяг навчальної вибірки (кількість студентів в групі за списком). Відома потужність алфавіту класів розпізнавання $\left\{Y_m^0 \,|\, m=\overline{1,\,M}\right\}$, кожний рядок матриці Y_m утворює вектор кластеризації.

Ефективність кластеризації даних будемо оцінювати за інформаційним критерієм E функціональної ефективності (КФЕ), використовуючи ймовірність помилки або середній ризик, який розглядається у статті [10]. Слід зазначити, що найпридатнішим для задач такого класу є підхід, що ґрунтується на визначенні кількості інформації, що здатна отримати система за результатами первинної обробки вхідних даних [11].

Нехай за результатами тестового контролю сформовано навчальну матрицю типу «об'єкт—властивість» $\| y_{0,i}^{(j)} \,|\, i=\overline{1,N};\;\; j=\overline{1,n}\|$, де N – кількість ознак розпізнавання (результатів тестування в академічній групі, що атестується за поточний навчальний модуль дистанційного курсу); n — обсяг навчальної вибірки (кількість студентів в групі за списком). При цьому відома потужність алфавіту класів розпізнавання $\left\{Y_m^0 \,|\, m=\overline{1,M}\right\}$. Сформовано структурований вектор просторово-часових параметрів функціонування СКДН: $g =< g_1,...,g_{\xi_1},...,g_{\Xi_1} >$ з відповідними обмеженнями: $R_{\xi_1}\left(g_1,...,g_{\xi_1},...,g_{\Xi_1}\right) \le 0$.

Треба у рамках ІЕІТ на етапі навчання СКДН, під яким будемо розуміти проектування та супроводження дистанційного курсу, за результатами машинного тестування сформувати апріорну нечітку навчальну матрицю для M класів розпізнавання і побудувати оптимальні вирішальні правила шляхом цілеспрямованої трансформації апріорного нечіткого розбиття простору ознак на класи розпізнавання у чітке розбиття еквівалентності за умови, що усереднений за алфавітом класів інформаційний КФЕ навчання СКДН, набуває глобального максимуму в робочій області визначення його функції [12]

$$E^* = \frac{1}{M}\sum_{m=1}^{M} \max_{G} E_m ,$$

де E_m — КФЕ навчання системи розпізнавати реалізації класу X_m^{o}; G — область допустимих значень параметрів функціонування СКДН.

Математична модель СКДН у режимі кластер-аналізу

Математичний вхідний опис СКДН, що здійснює кластерізацію результатів тестового контролю знань слухачів, представимо у вигляді теоретико-множинної структури:

$$\Delta_B < G,\, T,\, \Omega,\, Z,\, S,\, Y;\, \Phi >,$$

де G — множина вхідних сигналів (множина тестів); T — множина моментів зняття даних (результатів тестування); Ω — простір ознак розпізнавання (окремі відповіді на запитання); Z — множина функціональних станів системи (рівнів знань слухача); S — множина оцінок знань; Y — вибіркова множина (апріорно класифікована навчальна матриця); F: $G\times T\times\Omega\times Z\times S\to Y$ — оператор виходу, який формує навчальну матрицю.

На рис. 1 показано категорійну модель СКДН, що здійснює оптимізацію апріорного нечіткого розбиття рівнів знань у режимі кластер-аналізу у рамках ІЕІТ.

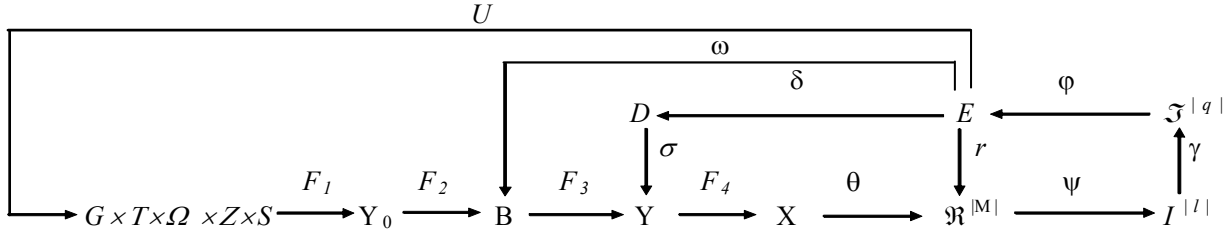


Рис. 1. Категорійна модель СКДН, що функціонує в режимі гібридного кластер-аналізу рівнів знань

На рис. 1 оператор формування навчальної матриці Y подамо як композицію операторів: $F = F_1 \circ F_2 \circ F_3$, де F_1 — оператор формування вхідної некласифікованої навчальної матриці Y_0; F_2 — оператор, що за таксонометричними методами визначає геометричні центри класів, які є вершинами еталонних векторів-реалізації образу, та здійснює попарне розбиття множин еталонних векторів і формує апріорну навчальну матрицю в евклідовому просторі.

Тут оператор $\theta : Y \to \Re^{|M|}$ будує нечітке розбиття $\Re^{|M|}$, яке в загальному випадку допускає перетин класів розпізнавання.

Для побудови такого розбиття використаємо алгоритм кластеризації *k*-середніх. Якщо можливо апріорно визначити кількість кластерів або їх кількість є відомою, то для класифікації частіше за все використовують паралельні кластер-процедури, у яких розбиття виконується згідного певного функціоналу якості. Саме до таких методів відноситься метод *k*-середніх. Серед неієрархічних методів цей метод є найрозповсюдженішим, завдяки швидкості, простоті, та прозорості й зрозумілості його алгоритму. При цьому вибір кількості кластерів може базуватися на попередніх дослідженнях, теоретичних міркуваннях або інтуїції. Після формування класів якість отриманого розбиття будемо оцінювати, знаходячи ймовірності помилки за формулою

$$R = 0,5\left\{1 - \exp\left(-\frac{m\ln\left(\frac{D}{d}\right)}{\frac{D^m}{d^m} - 1}\right) + \exp\left(-\frac{m\ln\left(\frac{D}{d}\right)}{1 - \frac{d^m}{D^m}}\right)\right\}, \tag{3}$$

де m — розмірність простору параметрів; d — найбільш ймовірне значення найближчої відстані, D — найбільш ймовірне значення міжкласової відстані. Ймовірність помилки R можна вважати критерієм якості кластеризації: чим менша ймовірність помилки — тим вища якість кластеризації [10]. З формули (3) бачимо, що з ростом відношення D/d ймовірність помилки відокремлення відстаней між реалізаціями всередині кластерів та відстаней між кластерами зменшується.

Далі оператор $\Psi : \Re^{|M|} \to I^{|l|}$ перевіряє основну статистичну гіпотезу $\gamma_1 : y_{m,i}^{(j)} \in X_m^0$, де $I^{|l|}$ — множина гіпотез, яка для $M = 2$ крім основної включає альтернативну гіпотезу $\gamma_2 : y_{m,i}^{(j)} \notin X_m^0$. Оператор γ визначає множину точністних характеристик (ТХ) $\Im^{|q|}$, де $q = l^2$ — кількість ТХ, а оператор ϕ обчислює множину E значень інформаційного критерію оптимізації, якій є функціоналом від ТХ[2]. Оператор r коректує розбиття $\Re^{|M|}$ в залежності від значень критерію. Визначення центрів класів здійснюється за формулою

$$d_0 = \frac{d\left(x_1 + x_2\right)}{2}. \tag{4}$$

Отриману матрицю В використовуючи рівень селекції $\rho_0 = 0,5$ перетворюємо у вхідну матрицю Y і запускаємо процес навчання системи.

Для визначення оптимальності розбиття введемо критерій ефективності розбиття Е. Як критерій будемо використовувати інформаційний критерій Шеннона, робоча формула якого має такий вигляд:

$$E_m = 1 + \frac{1}{2}\left(\frac{K_2}{K_2 + K_4}\log_2\frac{K_2}{K_2 + K_4} + \frac{K_1}{K_1 + K_3}\log_2\frac{K_1}{K_1 + K_3} + \frac{K_3}{K_1 + K_3}\log_2\frac{K_3}{K_1 + K_3} + \right.$$
$$\left. + \frac{K_4}{K_2 + K_4}\log_2\frac{K_4}{K_2 + K_4}\right), \tag{5}$$

де K_1, K_2 — кількість подій, які визначають належність або неналежність реалізацій класу X_m^0, якщо $\left\{x_m^n\right\} \in X_m^0$; K_3, K_4 — кількість подій, які визначають належність або неналежність реалізацій класу X_m^0, якщо вони дійсно належать цьому класу, які в свою чергу визначаються через точності характеристики та помилки першого і другого роду які обчислюються за формулами

$$D_1 = \frac{K_1}{n}; \quad \alpha = \frac{K_2}{n}; \quad \beta = \frac{K_3}{n}; \quad D_2 = \frac{K_4}{n}. \tag{6}$$

За навчальну матрицю, буде вибрана та матриця B, значення критерію функціональної ефективності якої буде максимальним.

$$\overline{E}^* = \max_{\{G_\xi\}} E_m^*. \tag{7}$$

При проведенні кластеризації методом *k*-means існує проблема вибору початкових центрів кластерів. Загальновживаними є такі підходи: вибір перших *k* реалізацій з вхідних даних, вибір випадкових *k* реалізацій з початкового набору даних, вибір реалізацій, найвіддаленіших одна від одної. В роботі пропонується здійснити цей вибір таким чином — обираються перших *k* відмінних між собою реалізацій (щоб не допустити можливість збігу центрів, а отже і уникнути ситуації, коли буде виконане розбиття на меншу кількість кластерів ніж задано).

Основні функції програми виконуються циклічно, до моменту виконання критерію зупинки ітераційного процесу. Таким критерієм може бути досягнення певного числа ітерацій або стабілізація центрів кластерів (тобто, якщо на певній ітерації, координати центрів кластерів не змінилися). У розробленій програмі за критерій зупинки було обрано саме стабілізацію центрів кластерів. До ітераційного процесу включено такі основні функції:

1) функція, що обчислює відстань між центрами кластерів та реалізаціями;

2) функція, яка визначає, до якого кластеру віднести кожну реалізацію;

3) функція, яка обчислює нові центри кластерів;

4) функція, яка перевіряє чи змінилися центри кластерів.

При реалізації першої функції за міру однорідності об'єктів було взято відстань Хемінга (block distance).

Таким чином, результатом виконання цієї функції є масив, що містить у собі відстань від кожної реалізації, до кожного з центрів кластерів. Цей масив обробляється функцією, в якій відбувається порівняння відстаней до центрів кластерів. Результатом виконання цієї функції, є віднесення кожної реалізації до певного класу. Реалізація буде належати тому класу, відстань до центра якої є найменшою.

Після цього обчислюються нові центри кластерів, шляхом усереднення значень за кожною ознакою серед реалізацій, що належать кожному відповідному кластеру. Для продовження чи зупинки ітераційного процесу реалізована функція, яка порівнює між собою, центри кластерів, що були на початку ітерації, та центри, отримані після перерахунку. Якщо центри відрізняються, то початковим центрам присвоюється значення перерахованих і починається нова ітерація, якщо ж центри кластерів не змінилися, то відбувається зупинка ітераційного процесу. Результатом роботи програми є сформовані текстові файли, у кількості, що дорівнює кількості кластерів, які містять у собі реалізації, віднесені до відповідного кластеру.

Проведено декілька експериментальних досліджень з даними, отриманими за результатом тестового контролю у студентів спеціальності «Інформатика» з дисципліни програмування протягом навчального року, у яких обчислювалися значення *d*, *D* та *R* за однакової розмінності простору, але з різною кількістю кластерів (рис. 2). Результати показані у таблиці.

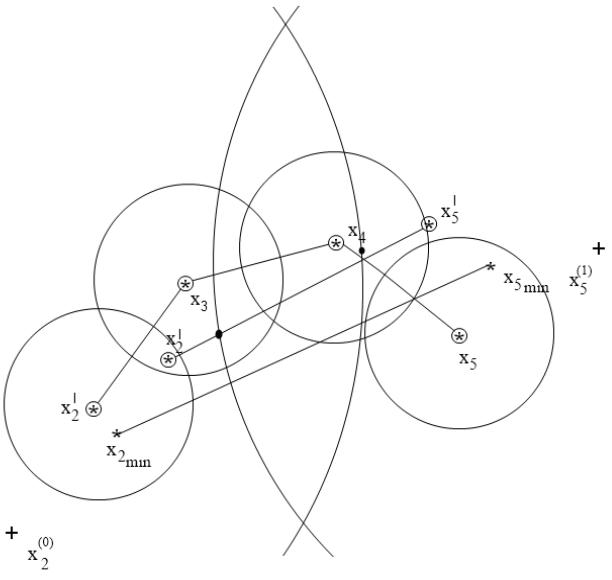
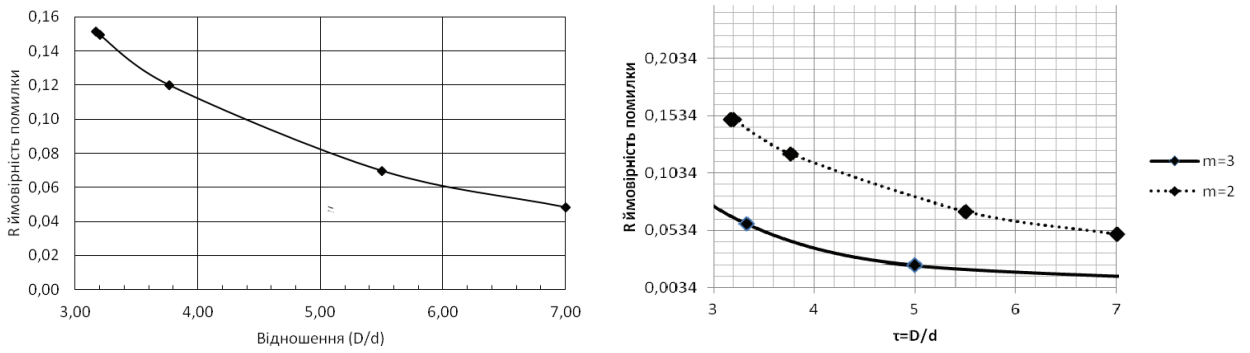


Рис. 2. Схематичне представлення класів знань студентів сформованих за результати тестового контролю

Результати дослідження залежності *R* від відношення τ = *D*/*d*.

Розмірність простору *m*	Кількість кластерів *k*	Ймовірність похибки *R*	Між класова відстань *D*	Найближча відстань *d*	Відношення τ = *D*/*d*
2	3	0,12	49	13	3,77
	4	0,05	42	6	7,00
	5	0,07	22	4	5,50
	6	0,15	19	6	3,17
	7	0,15	16	5	3,20
3	3	0,0034	53	5	10,6
	4	0,059	10	3	3,33
	5	0,1293	14	6	2,3
	6	0,2430	10	6	1,66
	7	0,0229	10	2	5

Як видно з таблиці, найкращу якість кластеризації (тобто найменше значення *R*) отримали коли кількості кластерів дорівнювала 4. Графічно цю залежність показано на рис. 3. Отже, для отримання найкращого розбиття при заданих вхідних даних необхідно вибрати $k = 4$.



а) б)

Рис. 3. Залежність якості кластерів від їх кількості:
а — без урахування параметра *m*; б — з урахуванням параметра *m*

Висновки

Розроблено модифікацію *k*-means алгоритму та виконано дослідження його параметрів щодо кластеризації вхідних даних за умов їх апріорного природного впорядкування. З метою оцінки якості кластеризації оцінювання рівня знань студентів в системі дистацнійного навчання разом з критерієм функціональної ефективності використано розроблений новий критерій, який дозволив оцінити результативність кластерного аналізу в залежності від вхідних параметрів алгоритму таких як: прогнозована апостеріорна похибка прийняття рішення, кількість контейнерів класів розпізнавання, розмірність простору ознак розпізнавання. Окремо досліджено зв'язок параметрів алгоритму з КФЕ, який показав, що в залежності від розмірності простору ознак стає можливим суттєве зменшення похибки результатів кластеризації вхідних даних, та дозволяє досягти асимптотичного максимуму КФЕ системи в цілому. З використанням запропонованого гібридного підходу кластер-аналізу стає можливим зменшувати апостеріорні помилки на етапі функціонування СКДН в режимі екзамену за рахунок автоматичного визначення параметрів алгоритмів кластеризації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Петров С. А. Категориально-информационная модель адаптивной системы непрерывного обучения / С. А. Петров // Управляющие системы и машины. — 2009. — № 2. — С. 48—51.

2. Довбиш А. С. Машинна оцінка знань студентів у системах керування дистанційним навчанням / А. С. Довбиш, В. О. Любчак, С. О. Петров // Вісник Сумського державного університету. Серія «Технічні науки». — 2007. — № 1. — С. 167—178.

3. Факторный, дискриминантный и кластерный анализ / [Дж.-О. Ким, Ч. У. Миллер, У. Р. Клекк и др.] — М. : Финансы и статистика, 1989. — 215 с.

4. Jain A. K. Dataclustering : a review / A. K.Jain, M. N. Murty, P. J. Flynn // ACM Computeing Surveys(CSUR). — 1999. — Vol. 31. Issue 3—69 p.

5. Айвазян С. А. Прикладная статистика. Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер. — М. : Финансы и статистика. — 1989. 300—310 с.

6. Браверман Э. М. Структурные методы обработки эмпирических данных / Э. М. Браверман, И. Б. Мучник. — М. : Наука. Физматлит. — 464 с.

7. Алехин Е. И. Многомерные статистические методы / Е. И. Алехин. — Орел : Изд. центр ГОУ ВПО ОГУ, 2007 — 37 с.

8. Бабак О. В. Алгоритм решения некоторых задач кластерного анализа / О. В. Бабак, А. С. Касанов // Управляющие системы и машины. — 2001. — № 6. — 25—30 с.

9. Петров С. О. Вплив структури простору ознак розпізнавання в системах підтримки прийняття рішень / С. О. Петров // Інтернет-Освіта-Наука –2010 : Сьома міжнар. конф. ІОН-2010: 28 вер.—3 жов. 2010 р. : тези доп. — Вінниця : Вінницький національний технічний університет, 2010. — С. 71—72.

10. Коваль П. Н. Использование кластеризации при анализе данных / П. Н. Коваль // Управляющие системы и машины. — 2010. — № 6. — С. 32—34.

11. Куренков Н. И. Энтропийный подход к решению задач классификации многомерных данных / Куренков Н. И. Ананьев С. Н. // Информационные технологии. — 2006. — № 8. — С. 50—55.

12. Petrov S. Mathematical model of distance learning control system in framework of IEIT / Sergey Petrov // Internet Education Science: Proceedings of the Sixth International Conference, 7—11 October 2008. — Vinnytsia, Ukraine. — 2008. — Vol. 1. — P. 167—169.

Рекомендована кафедрою менеджменту та безпеки інформаційних систем ВНТУ

Стаття надійшла до редакції 14.07.2015

Петров Сергій Олександрович — канд. техн. наук, старший викладач кафедри комп'ютерних наук, e-mail: sergpet@gmail.com;

Сумський державний університет;

Лисак Наталія Володимирівна — канд. техн. наук, доцент кафедри менеджменту та безпеки інформаційних систем;

Міронова Юлія Володимирівна — канд. екон. наук, старший викладач кафедри менеджменту та безпеки інформаційних систем.

Вінницький національний технічний університет, Вінниця

S. O. Petrov[1]
N. V. Lysak[2]
Yu. V. Mironova[2]

Hybrid algorithm of cluster analysis forming a priori space division into classes of knowledge in the systems of distance educating

[1] Sumy State University;
[2] Vinnytsia National Technical University

There has been offered the modification of algorithm of k - means, the idea of improvement of which consists in the combined use of criterion of estimation of error of clusterization and informative criterion of functional efficiency, that determines authenticity of the built decision rules of determination of belonging of realization to some class of knowledge. Thus the simultaneous use of statistical and informative approaches allowed including such important parameter for the algorithms of clusterization as an amount of clusters in iterative optimization procedure. Having a priori information about distribution of N- measure vectors of realization, presenting the results of testing of knowledge of students, it also allows to define the optimal geometrical parameters of containers, describing the classes of knowledge of students in the systems controlled from distance education.

Key words: clustering, *k*-means, criterion of functional efficiency, clustering quality criteria, distance learning systems.

Petrov Sergii O. — Cand. Sc. (Eng.), Senior Lecturer of the Chair of Computer Science, e-mail: sergpet@gmail.com;

Lysak Natalia V. — Cand. Sc. (Eng.), Assistant Professor of the Chair of Management and Security of Information Systems;

Mironova *Yuliia* ***V.*** — Cand. Sc. (Econ.), Senior Lecturer of the Chair of Management and Information Systems Security

С. А. Петров[1]
Н. В. Лысак[2]
Ю. В. Миронова[2]

Гибридный алгоритм кластер-анализа для формирования априорного разбиения пространства признаков на классы знаний в системах дистанционного обучения

[1]Сумской государственный университет;
[2] Винницкий национальный технический университет

Предложена модификация алгоритма k-means, идея усовершенствования которого заключается в комбинированном использовании критерия оценки ошибки кластеризации и информационного критерия функциональной эффективности, который определяет достоверность построенных решающих правил определения принадлежности реализаций к некоторому классу знаний. При этому одновременное использование статистического и информационного подходов позволило включить такой важный параметр для алгоритмов кластеризации как количество кластеров в итерационную оптимизационную процедуру. Имея априорную информацию о распределении N-мерных векторов реализаций, представляющих результаты тестирования знаний студентов, определить оптимальные геометрические параметры контейнеров, описывающих классы знаний студентов в системах дистанционного образования.

Ключевые слова: кластеризация, *k*-means, критерий функциональной эффективности, критерий оценки качества кластеризации, системы дистанционного обучения.

Петров Сергей Александрович — канд. техн. наук, старший преподаватель кафедры компьютерных наук, e-mail: sergpet@gmail.com;

Лысак Наталья Владимировна — канд. техн. наук, доцент кафедры менеджмента и безопасности информационных систем;

Миронова Юлия Владимировна — канд. экон. наук, старший преподаватель кафедры менеджмента и безопасности информационных систем