# Кібернетика та системний аналіз

УДК 004.8:004.032.26 DOI: 10.30748/zhups.2021.68.10

Є.В. Бодянський, А.Ю. Шафроненко, І.М. Климова

*Харківский національний університет радіоелектроніки, Харків*

## МЕТОД АДАПТИВНОЇ ДОСТОВІРНОЇ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ДАНИХ НА ОСНОВІ ЕВОЛЮЦІЙНОГО АЛГОРИТМУ

*Методи обчислювального інтелекту широко використовуються для вирішення багатьох складних проблем, включаючи, звичайно, традиційні: видобуток даних та такі нові напрямки, як динамічний видобуток даних, видобуток потоків даних, видобуток великих даних, веб-видобуток, видобуток тексту, тощо. Одна з основних областей обчислювального інтелекту – це еволюційні алгоритми, які по суті представляють певні математичні моделі еволюції біологічних організмів. У роботі запропоновано адаптивний метод нечіткої кластеризації з використанням оптимізації еволюційних котячих зграй. Використовуючи запропонований підхід, можна вирішити завдання кластеризації в режимі он-лайн.*

***Ключові слова:*** *нечітка кластеризація, достовірна нечітка кластеризація, рівень nature належності, оптимізація, котячі зграї, режим пошуку, режим трасування*

## Вступ

**Постановка проблеми.** Проблема нечіткої кластеризації масивів даних розглядається в умовах, коли сформовані кластери довільно перекриваються в просторі ознак.

Вихідною інформацією для вирішення задачі є масив багатовимірних векторів даних, утворених набором векторних спостережень $X = (x(1), x(2), ..., x(k), ...x(N)) \subset R^n$, де $k$ – у загальному випадку номер спостереження в початковому масиві.

Результатом кластеризації є розділ цього масиву на $m$ накладених класів з прототипами-центроїдами $Cl_j \in R^n$, $j = 1, 2, ..., m$.

**Аналіз останніх досліджень і публікацій.** Завдання класифікації в режимі самонавчання (кластеризації) багатовимірних даних є важливою частиною традиційного інтелектуального аналізу, такого як видобуток даних, динамічний видобуток даних, видобуток потоків даних, видобуток великих даних, веб-видобуток [1–2].

Однією з основних областей обчислювального інтелекту є так звані еволюційні алгоритми, які є математичними моделями еволюції біологічних організмів.

Проблема, пов'язана з векторними спостереженнями, кластеризацією, часто виникає у багатьох додатках інтелектуального аналізу даних, і насамперед у нечіткій кластеризації даних при обробці векторного спостереження з різними рівнями ймовірності, достовірності, тощо може належати більше ніж до одного класу. Дуже ефективними є самоорганізуючі карти Кохонена [3] та еволюційні алгоритми, які можуть покращити кластеризацію даних у випадку, коли дані обробляються послідовно в режимі онлайн.

**Мета статті.** Метою роботи було запропонувати метод адаптивної достовірної нечіткої кластеризації даних на основі еволюційного алгоритму, який був би позбавлений недоліків традиційних підходів до кластеризації даних.

## Виклад основного матеріалу

### Адаптивний алгоритм достовірної нечіткої кластеризації даних

Достовірна нечітка кластеризація пов'язана з мінімізацією цільової функції (1):

$$E\left(Cr_q(k), w_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m} Cr_q^{\beta}(k) D^2\left(x_k, w_q\right), \qquad (1)$$

з обмеженнями

$$0 \le Cr_q(k) \le 1 \forall q, k;$$

$$\sup \mathrm{Cr}_q(\mathrm{k}) \ge 0,5 \forall k;\ Cr_q(k) + supCr_l(k) = 1,$$

де $Cr_q(k)$ – достовірність спостереження $x_k$, що належить кластеру $Cl_q$.

У цьому випадку рівень належності розраховується за допомогою функції належності [5–6]:

$$U_q(k) = \varphi_q\left(D\left(x_k, Cl_q\right)\right), \qquad (2)$$

де $\varphi_q(\cdot)$ – монотонно зменшується в інтервалі $[0, \infty]$, $\varphi_q(0) = 1$, $\varphi_q(\infty) \to 0$.

Неважко помітити, що функція (2) є, по суті, мірою подібності на основі відстані [7]. Як таку функ-

цію в [8] було запропоновано використовувати вираз:

$$U_q(k)=\left(1+D^2\left(x_k,Cl_q\right)\right)^{-1}. \quad (3)$$

Цікаво відзначити, що вираз (3) можна переписати у формі:

$$U_q(k)=\left(D^2\left(x_k,Cl_q(k)\right)\right)^{\frac{1}{1-\beta}}\left(\sum_{l=1}^{m}\left(D^2\left(x_k,Cl_l(k)\right)\right)^{\frac{1}{1-\beta}}\right)^{-1}=\left(D^2\left(x_k,Cl_k(k)\right)\right)^{\frac{1}{1-\beta}}\left(D^2\left(x_k,Cl_q(k)\right)\right)^{\frac{1}{1-\beta}}+$$

$$+\sum_{\substack{l=1\\l\neq q}}^{m}\left(D^2\left(x_k,Cl_l(k)\right)\right)^{\frac{1}{1-\beta}})^{-1}=\left(1+\left(D^2\left(x_k,Cl_q(k)\right)\right)^{\frac{1}{1-\beta}}\sum_{\substack{l=1\\l\neq q}}^{m}\left(D^2\left(x_k,Cl_l(k)\right)\right)^{\frac{1}{1-\beta}}\right)^{-1}, \quad (4)$$

що для евклідової метрики і $\beta=2$ приймає форму функції щільності розподілу Коші з параметром ширини $\sigma_q^2$ [9]:

$$U_q(k)=\left(1+\frac{\left\|x_k-Cl_q(k)\right\|^2}{\sigma_q^2}\right)^{-1}; \quad (5)$$

$$\sigma_q^2=\left(\sum_{\substack{l=1\\l\neq q}}^{m}\left\|x_k-Cl_l(k)\right\|^{-2}\right)^{-1}. \quad (6)$$

Неважко помітити, що функція приналежності (4) є особливим випадком (5) для $\sigma_q^2=1$.

Нарешті, пакетний алгоритм достовірної нечіткої кластеризації можна записати у вигляді [5–6]:

$$U_q^{(\tau+1)}(k)=\left(1+D^2\left(x_k,Cl_q^{(\tau)}\right)\right)^{-1}; \quad (7)$$

$$U*_q^{(\tau+1)}(k)=U_q^{(\tau+1)}(k)\left(\sup U_l^{(\tau+1)}(k)\right)^{-1}; \quad (8)$$

$$Cr_q^{(\tau+1)}(k)=\frac{1}{2}\left(U_q^{*(\tau+1)}(k)+1-\sup_{l\neq q}U_l^*(k)\right); \quad (9)$$

$$Cl_q^{(\tau+1)}=\sum_{k=1}^{N}\left(Cr_q^{(\tau+1)}(k)\right)^{\beta}x_k\left(\sum_{k=1}^{N}\left(Cr_q^{\tau+1}(k)\right)^{\beta}\right)^{-1}. \quad (10)$$

Виходячи з цих формул, ми можемо ввести до розгляду онлайн-версію методу достовірної нечіткої кластеризації у формі:

$$\sigma_q^2(k+1)=\left(\sum_{\substack{l=1\\l\neq q}}^{m}\left\|x_{k+1}-Cl_l(k)\right\|^{-2}\right)^{-1}; \quad (11)$$

$$U_q(k+1)=\left(1+\frac{\left\|x_{k+1}-Cl_q(k)\right\|^2}{\sigma_q^2(k+1)}\right)^{-1}; \quad (12)$$

$$\begin{cases}U_{(k+1)}^*=U_q(k+1)\left(\sup U_l(k+1)\right)^{-1}\\ Cr_q(k+1)=\frac{1}{2}\left(U_q^*(k+1)+1-\sup_{l\neq q}U_l^*(k+1)\right)\\ Cl_q(k+1)=Cl_q(k)+\eta(k+1)Cr_q^{\beta}(k+1)\times\\ \times\left(x_{k+1}-Cl_q(k)\right).\end{cases} \quad (13)$$

Тому з обчислювальної точки зору онлайн-алгоритм для достовірної нечіткої кластеризації не складніший, ніж періодичні версії FCM та PCM, зберігаючи при цьому переваги підходу довіри.

### Еволюційна оптимізація котячої зграї

Для пошуку глобального екстремуму функції достовірності (1) доцільно використовувати алгоритми оптимізації зграї біо-еволюційних частинок [10]. Серед алгоритмів зграї найшвидшими є алгоритми зграї котів [11–12], які виявилися ефективними у вирішенні широкого кола задач з обробки даних.

Модель поведінки котячої зграї (CS) передбачає, що кожна кішка $cat_p$ зграї, що складається з $Q$ осіб $(p=1,2,...,Q)$ може бути в одному з двох станів: Режим пошуку (SM) та Режим трасування (TM). У цьому випадку режим пошуку пов'язаний з повільними рухами з невеликою амплітудою поблизу початкової позиції (сканування простору в районі поточної позиції) та режимом трасування, який визначається швидкими стрибками з великою амплітудою і дозволяє коту $cat_p$ піти з місцевого екстремуму, якщо вона туди потрапила. Поєднання локального сканування та різких змін у поточному стані робить більш імовірним пошук глобального екстремуму порівняно з традиційними методами багатоекстремальної оптимізації.

У загальному випадку обидва ці режими для кожної із зграї котів можуть бути описані повторюваною процедурою оптимізації [13]:

$$cat_p(\tau+1)=cat_p(\tau)-\alpha(cat_p(\tau)-cat_p(\tau-1))-$$
$$-\eta\hat{\nabla}E_M(cat_p(\tau))+\eta_\xi\Xi(\tau), \quad (14)$$

де $cat_p(\tau+1)$ – стан $p$-кота зграї на $\tau$-ітерація пошуку;

$\alpha$ – параметр, який визначає властивості інерції режиму трасування. У випадку, коли $\alpha=0$ підходи оптимізації процесів до стандартного градієнтного пошуку;

$\eta$ – крок режиму пошуку;

$\hat{\nabla}E(cat_p(\tau))$ – оцінка градієнта цільової функції (1), сусідство точки $cat_p(\tau)$;

$\Xi(\tau)$ – випадкова складова, яка вносить додаткові стохастичні рухи в процес трасування;

$\eta_\xi$ – параметр, який визначає амплітуду цих рухів.

У цьому алгоритмі кожна кішка може мати два паралельних стани: режим пошуку та режим відстеження. Цей підхід забезпечує пошук глобального екстремуму в тому випадку, коли кількість котів у зграї достатня.

### Експериментальні дослідження

Нечітка кластеризація на основі еволюційної оптимізації зграї котів (CSO) проводилась на чотирьох різних вибірках даних: Cancer, Wine and Glass. Кожен із наборів даних має ряд параметрів, представлених у табл. 1.

Таблиця 1

Параметричні характеристики наборів даних

| Вибірка | Кількість кластерів | Кількість атрибутів | Кількість спостережень |
|---|---|---|---|
| Cancer | 2 | 9 | 683 |
| Wine | 3 | 13 | 178 |
| Glass | 6 | 8 | 214 |

Джерело: розроблено авторами.

Для налагодження алгоритму оптимізації котячих зграй, використовувались параметри, представлены в табл. 2

Таблиця 2

Параметри алгоритму оптимізації зграї котів

| Параметри | Значення |
|---|---|
| SRD | Випадково [0,1] |
| Seeking memory Pool (SMP) | 5 |
| Population size | Кількість кластерів |
| $r_1$ | Випадково |
| $c_1$ | Const |
| SPC | Випадково [0,1] |
| Кількість ітерацій | Вручну |

Джерело: розроблено авторами.

Результати обробки часу алгоритмів кластеризації, таких як нечіткий алгоритм с-середніх (FCM), оптимізація роя частинок (PSO), алгоритм Гауса - Зейделя (GSA), CSO та методу адаптивної достовірної нечіткої кластеризації даних на основі еволюційного алгоритму (ACFCSO) продемострировано в табл. 3. Порівняльний аналіз демонструє достаньо високу швидкість роботи, не поступаючись більш відомим на сьогодні алгоритмам нечіткої кластеризації даних. За допомогою оптимізаційних процедур, які містить основі еволюційної оптимізації зграї котів (CSO), збільшує швидкість роботи запропонованого методу в кілька разів.

Таблиця 3

Порівняльні результати обробки часу алгоритмів кластеризації

| Вибірка | FCM | PSO | GSA | CSO | ACF CSO |
|---|---|---|---|---|---|
| Cancer | 0.009 | 0.138 | 0.204 | 0.026 | **0.01** |
| Glass | 0.010 | 0.431 | 0.431 | 0.021 | **0.02** |
| Wine | 0.009 | 0.282 | 0.098 | 0.076 | 0.02 |

Джерело: розроблено авторами.

## Висновки

Розглянуто проблему нечіткої кластеризації на основі імовірнісного, можливого та достовірного підходів в Інтернеті. Повторні модифікації відомих пакетних процедур, призначених для вирішення проблем видобутку потоку даних, дозволяють обробляти інформацію в режимі онлайн як послідовне доповнення до розглянутої системи. Оскільки цільові функції нечіткої кластеризації в загальному випадку є багатоекстремальними, було запропоновано вдосконалити рішення, використовуючи алгоритми еволюційної оптимізації зграї. Запропоновано модифікацію, внесену на основі процедури оптимізації котячих зграй з поліпшеними властивостями за допомогою стохастичної оцінки градієнта.

Експерименти підтвердили ефективність розробленого підходу, який характеризується простотою чисельного впровадження та досить високим коефіцієнтом конвергенції. Запропонований метод оптимізації будучи представником еволюційних алгоритмів призначений для використання в гібридних системах обчислювального інтелекту, і перш за все в задачах навчання штучних нейронних мереж, нейро-фаззі системах, а так само в задачах кластеризації та класифікації.

## Список літератури

1. Wunsch Xu.R. Clustering / Xu.R. Wunsch, D.C. Hoboken, N.J. John. – New York: Wiley & Sons, 2009. – 234 p.
2. Aggarwal C.C. Data Mining / C.C. Aggarwal. – Berlin: Springer, 2015. – 43 p.
3. Kohonen T. Self-Organizing Maps / T. Kohonen. – Berlin: Springer-Verlag, 1995. – 121 p.
4. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / J.C. Bezdek. – New York: Plenum Press, 1981. – 421 p.
5. Credibilistic clustering: the model and algorithms / J. Zhou, Q. Wang, C.C. Hung, X. Yi // Fuzziness and Knowledge-Based Systems. – 2015. – No. 4. – P. 545-564.
6. Zhou J. Credibilistic clustering algorithms via alternating cluster estimation / J. Zhou, Q. Wang, C.C. Hung // Journal of Intelligent Manufacturing. – 2017. – No. 5. – P. 727-738.
7. Young F.W. Theory and Applications of Multidimensional Scaling-Hillsdale / R.M. Hamer, F.W. Young. – New York: Erlbaum, 1994. – 26 p.

8. Zhou J. A generalized approach to possibilistic clustering algorithms / J. Zhou, C.C. Hung // Fuzziness and Knowledge-Based Systems. – 2007. – No. 15. – P. 117-138.

9. Fuzzy clustering of incomplete data by means of similarity measures / Zh. Hu, Ye. Bodyanskiy, O. Tyshchenko, A. Shafronenko // 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering. – Lviv, 2-6 Jul 2019. – P. 149-152.

10. Grosan C. Swarm intelligence in Data Mining / C. Grosan, A. Abraham, M. Chis. – Berlin: Springer, 2006. – 123 p.

11. Chu S.C. Cat swarm optimization / S.C. Chu, P.W. Tsai, J.S. Pan. – Berlin: Springer-Verlag, 2006. – 210 p.

12. Chu S.C. Computational Intelligence based on the behavior of cats / S.C. Chu, P.W. Tsai // International Journal of Innovative Computing, Information, and Control. – 2007. – № 1. – P. 163-173.

13. Fuzzy Clusterization of Distorte d by Missing Observations Data Sets Using Evolutionary Optimization / A. Shafronenko, Ye. Bodyanskiy, I. Pliss, K. Patlan // Proceedings "Advanced Computer Information Technologies. – Česke Budejovice, 5-7 June 2019. – P. 217-220. https://doi.org/10.1109/ ACITT.2019.8779888.

14. Shafronenko A. The Fast Modification of Evolutionary Bioinspired Cat Swarm Optimization Method / A. Shafronenko, Ye. Bodyanskiy, I.P. Pliss // IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL). – Sozopol, 6-8 September 2019. – P. 548-552. https://doi.org/10.1109 / CAOL46282.2019.9019583.

*Надійшла до редколегії 11.03.2021*
*Схвалена до друку 13.04.2021*

***Відомості про авторів:***

**Бодянський Євгеній Володимирович**
доктор технічних наук професор науковий керівник проблемної НДЛ АСУ Харківського національного університету радіоелектроніки,
Харків, Україна
https://orcid.org/0000-0001-5418-2143

**Шафроненко Аліна Юріївна**
кандидат технічних наук доцент
доцент кафедри Харківського національного університету радіоелектроніки,
Харків, Україна
https://orcid.org/0000-0002-8040-0279

**Климова Ірина Миколаївна**
асистент кафедри Харківського національного університету радіоелектроніки,
Харків, Україна
https://orcid.org/0000-0003-0455-6180

***Information about the autors:***

**Yevgeniy Bodyanskiy**
Doctor of Technical Sciences Professor Scientific Head of Control Systems Research Laboratory
of Kharkiv National University of Radio Electronics,
Kharkiv, Ukraine
https://orcid.org/0000-0001-5418-2143

**Alina Shafronenko**
Doctor of Philosophy Associated Professor
Senior Lecturer of Department
of Kharkiv National University of Radio Electronics,
Kharkiv, Ukraine
https://orcid.org/0000-0002-8040-0279

**Iryna Klymova**
Assistant of Department
of Kharkiv National University of Radio Electronics,
Kharkiv, Ukraine
https://orcid.org/0000-0003-0455-6180

## МЕТОД АДАПТИВНОЙ ДОСТОВЕРНОЙ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ НА ОСНОВЕ ЭВОЛЮЦИОННОГО АЛГОРИТМА

Е.В. Бодянский, А.Ю. Шафроненко, И.Н. Климова

*Методы вычислительного интеллекта широко используются для решения многих сложных проблем, включая, конечно, традиционные: добыча данных, Data Mining, Dynamic Data Mining, Data Stream Mining, Big Data Mining, Web Mining, Text Mining. Одна из основных областей вычислительного интеллекта – это эволюционные алгоритмы, которые по сути представляют определенные математические модели эволюции биологических организмов. В работе предложен адаптивный метод нечеткой кластеризации с использованием оптимизации эволюционной кошачьей стаи. Используя предложенный подход, можно решить задачу кластеризации в режиме он-лайн.*

***Ключевые слова**: нечеткая кластеризация, достоверная нечеткая кластеризация, уровень принадлежности, оптимизация, кошачьи стаи, режим поиска, режим отслеживания.*

## ADAPTIVE CREDIBILISTIC FUZZY CLUSTERING METHOD BASED ON EVOLUTIONARY ALGORITHM

Ye. Bodyanskiy, A. Shafronenko, I. Klymova

*The task of fuzzy clustering data is very interesting and important problem and often found Data Mining, Dynamic Data Mining, Data Stream Mining, Big Data Mining, Web Mining, Text Mining, etc. One of the main areas of computational intelligence are evolutionary algorithms that essentially represent certain mathematical models of biological organisms evolution. The goal of the paper is to propose the procedure of the adaptive methods of credibilistic fuzzy clustering using evolutionary algorithm. The goal of the work is adaptive credibilistic fuzzy clustering of data, using of credibility theory and evolutionary algorithm. The procedure of adaptive credibilistic fuzzy clustering of data based on the used goal function of special type and evolutionary approach from cat swarm optimization algorithm. This method designed to work both in batch and online mode, when data are fed to processing sequentially in real time. Proposed approach characterized by simple numerical implementation and relatively high rate of convergence. Proposed optimization method as a representative of evolutionary algorithms is intended for use in hybrid systems of computational intelligence, and especially in the problems of learning artificial neural networks, neurophase systems, as well as in the problems of clustering and classification. The experement result has confermed effectiveness stability work of adaptive credibilistic fuzzy clustering method based on evolutionary algorithm. Proposed method of adaptive credibilistic fuzzy clustering method based on evolutionary algorithm designed for use in hybrid systems of computational intelligence, in the problems of learning artificial neural networks, neuro-fuzzy systems, clustering and classification.*

***Keywords:** Fuzzy clustering, credibilistic fuzzy clustering, membership level, optimization, cat swarm, seeking mode, tracing mode.*