e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2022. № 1
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2022. № 1

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.8:004.032.26

ШВИДКА НЕЧІТКА ПРАВДОПОДІБНА КЛАСТЕРИЗАЦІЯ НА ОСНОВІ АНАЛІЗУ ПІКІВ ЩІЛЬНОСТІ РОЗПОДІЛУ ДАНИХ

Бодянський Є. В. – д-р техн. наук, професор, професор кафедри штучного інтелекту, Харківський національний університет радіоелектроніки, Харків, Україна.

Плісс І. П. – канд. техн. наук, провідний науковий співробітник ПНДЛ АСУ, Харківський національний університет радіоелектроніки, Харків, Україна.

Шафроненко А. Ю. – канд. техн. наук, доцент, доцент кафедри інформатики, Харківський національний університет радіоелектроніки, Харків, Україна.

АНОТАЦІЯ

Актуальність. Проблема кластеризації (класифікації без вчителя), що часто зустрічається при обробці масивів даних різної природи, є досить цікавою і невід'ємною частиною штучного інтелекту. Для вирішення цього завдання існує безліч відомих методів та алгоритмів, які базуються на принципах щільності розподілу спостережень в даних, що аналізуються. Однак ці методи досить складні в програмній реалізації та не позбавлені недоліків, а саме: проблеми визначення значущих кластерів в наборах даних різної щільності, багатоепохове самонавчання, застрягання в локальних екстремумах цільових функцій, тощо. Слід зазначити, що методи, засновані на аналізі піків щільності розподілу даних, є за своєю природою чіткими, тому для розширення можливостей цих методів доцільно ввести їх нечітку модифікацію.

Мета. Мета роботи полягає у запровадженні швидкої нечіткої кластеризації даних з використанням піків щільності розподілу даних, яка може знаходити екстемуми (центоїди) кластерів, що перетинаються незалежно від кількості даних, що надходять.

Метод. Розглянуто задачу нечіткої кластеризації масивів даних на основі гібридного методу, заснованого на одночасному використанні правдоподібного підходу до нечіткої кластеризації і алгоритму знаходження типів щільності розподілу вихідних даних. Особливістю запропонованого методу є обчислювальна простота і висока швидкість, пов'язана з тим, що весь масив обробляється тільки один раз, тобто виключається необхідність в багатоепоховому самонавчанні, що реалізується в традиційних алгоритмах нечіткої кластеризації.

Результати. Особливістю запропонованого методу швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу даних є обчислювальна простота і висока швидкість, пов'язана з тим, що весь масив обробляється тільки один раз, тобто виключається необхідність у багатоепоховому самонавчанні, що реалізується в традиційних алгоритмах нечіткої кластеризації. Результати обчислювального експерименту підтверджують ефективність запропонованого підходу в задачах кластерзації в умовах, коли кластери перетинаються.

Висновки. Результати експерименту дозволяють рекомендувати розроблений метод для вирішення проблем автоматичної кластеризації та класифікації даних та максимально швидко знаходити центри кластерів. Запропонований метод швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу даних призначений для використання в системах обчислювального інтелекту, нейро-фаззі системах,в навчанні штучних нейронних мереж та у завданнях кластеризації.

КЛЮЧОВІ СЛОВА: нечітка кластеризація, правдоподібна кластеризація, піки щільності розподілу даних.

АБРЕВІАТУРА

DBSCAN density-based spatial clustering of applications with noise;

OPTICS ordering points to identify the clustering structure;

NMI нормалізована взаємна інформація;

FCDP метод швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу.

НОМЕНКЛАТУРА

X – матриця набору даних;

k – номер вектору-спостереження;

DOI 10.15588/1607-3274-2022-1-9

i – номер атрибуту вектора-спостереження;

j – номер класу;

$x(k)$ – вектор-спостереження;

l, q – номери кластерів;

m – кількість неперетинних класів;

μ_j – рівень нечіткої належності j-го кластеру;

D – матриця відстаней між спостереженнями;

d – відстань між спостереженнями;

ρ – вектор локальної щільності;

c – центроїд кластера;

δ_k^* – точка з максимальною щільністю;

Cr – рівень правдоподібності.

ВСТУП

Задача кластеризації (класифікації без вчителя) часто зустрічається при обробці масивів спостережень самої різної природи в рамках загальної проблеми Data Mining, Big Data Mining, Data Stream Mining, тощо. Для вирішення цієї задачі, існує безліч підходів від найпростіших (*k*-середніх, ієрархічних) алгоритмів до найбільш просунутих, заснованих на аналізі щільності розподілу даних

Слід зазначити, що методи, засновані на аналізі піків щільності розподілу даних, є за своєю природою чіткими, тому для розширення можливостей цих методів доцільно ввести їх нечітку модифікацію.

Об'єкт дослідження швидка нечітка кластеризація даних на основі піків щільності розподілу даних.

Предмет дослідження процедура аналізу піків щільності розподілу даних.

Мета роботи полягає у запровадженні швидкої нечіткої кластеризації даних з використанням піків щільності розподілу даних, яка може знаходити екстемуми (центри) кластерів, що перетинаються незалежно від кількості даних, що надходять.

1 ПОСТАНОВКА ЗАВДАННЯ

Процес нечіткої кластеризації на основі аналізу піків щільності розподілу даних зручно представити у вигляді послідовності кроків, при цьому вихідною інформацією як і в інших методах, заснованих на парадигмі самонавчання, є нерозмічена вибірка векторних спостережень $X = \{x(1), x(2), ..., x(k), ..., x(N)\}$, $x(k) = \{x_i(k)\} \in R^n$, при цьому для зручності розрахунків всі компоненти цих векторів попередньо закодовані в деякому обмеженому інтервалі, наприклад, $-1 \le x_i(k) \le 1 \forall i, k$.

2 ОГЛЯД ЛІТЕРАТУРИ

Для вирішення цієї задачі існує безліч підходів[1] від найпростіших типу *k*-середніх і ієрархічних алгоритмів до найбільш просунутих, заснованих на аналізі щільності розподілу даних [2–4]. Найбільш популярними з таких алгоритмів є DBSCAN [5], DENCLUE [6], OPTICS [7] та їм подібні, що дозволяють оцінити не тільки центроїди, але і їх кількість. У той же час алгоритми, які засновані на щільностях, досить складні з точки зору чисельної реалізації, а процедури градієнтної оптимізації, що лежать в їх основі, схильні до застрягання в локальних екстремумах цільових функцій.

Цих недоліків позбавлений метод, заснований на пошуку піків щільності розподілу даних [8]. Слід зазначити, що більш відомі алгоритми, які аналізують щільність розподілу даних, не виконують пошук екстремумів, а призначені для вирішення завдань чіткої кластеризації, тобто в умовах коли апріорно відомо, що кластери, які формуються, взаємно не перетинаються, а кожне спостереження з вихідного масиву може належати тільки одному кластеру.

У той же час в реальних задачах досить часто зустрічається задачі, коли будь-яке з спостережень із аналізованого масиву даних, може з різними рівнями належності одночасно належати до декількох класів-кластерів. Ця ситуація розглядається в рамках нечіткого кластерного аналізу [9], при цьому історично склалося два підходи до вирішення цієї проблеми: імовірнісний і можливісний.

Зауважимо також, що в останні роки з'явився, так званий, довірчий підхід [10, 11], що володіє перевагами як імовірнісного, так і можливісного підходів.

3 МАТЕРІАЛИ І МЕТОДИ

В процесі кластеризації на основі аналізу піків щільності розподілу даних аналізується два параметри: ρ_k – локальна щільність і δ_k – відстань до точки з більш високою щільністю. Окрім того вводиться єдиний вільний параметр d_c – відстань зрізу, яка задається і варіюється користувачем для отримання необхідної точності рішення задачі.

Роботу методу можна сформулювати як наступну послідовність елементарних кроків:

1. На першому кроці на основі вихідної $(n \times N)$ матриці «об'єкт – властивість» вводиться $(N \times N)$-матриця відстаней між спостереженнями:

$$D = \{d_{kl}\},\ d_{kl} = \|x(k) - x(l)\| \forall k, l,$$

при цьому може бути використана будь-яка метрика, яка використовується в Data Mining і, зокрема, в кластерному аналізі.

2. На другому кроці розраховується $(N \times 1)$-вектор локальних щільностей $\rho = \{\rho_k\} \in R^N$:

$$\rho_k = \sum_{k=1}^{N} \chi(d_{kl} - d_c),$$

DOI 10.15588/1607-3274-2022-1-9

де $\chi(d) = \begin{cases} 1, \text{ якщо } d < 0, \\ 0, \text{ в іншому випадку.} \end{cases}$

Відстань зрізу обирається з суто емпіричних міркувань, при цьому автори методу [8] рекомендують вибирати його так, щоб в околі, який формується, потрапляло $0,01N - 0,02N$ спостережень вибірки, що оброблюється.

3. Розрахунок вектора мінімальних відстаней $\delta = \{\delta_k\} \in R^N$ до точок з більш високою щільністю

$$\delta_k = \min_{\forall l, \rho_l > \rho_k} \{d_{kl}\},$$

а для точки з максимальною щільністю δ_k^* розраховується:

$$\delta_k^* = \max_l \{d_{kl}\}.$$

4. Формування центроїдів кластерів $c_j, j = 1,2,...,m,$ при цьому в якості центроїдів $c_j = x(k)$ обираються точки з найвищою щільністю, тобто обираються деякі спостереження з вихідної вибірки X. До кожного з центроїдів c_j приписуються точки, найближчі до нього в сенсі

$$\min(d_{kl}) \equiv d_{jl}.$$

Зауважимо також, що в [12] в якості центроїдів пропонується використовувати значення $c_j = x(k)$ з максимальним значенням добутків $\rho_k \cdot \delta_k$.

Далі всі центроїди впорядковуються за зменшенням цього добутку $c_1,...,c_j,...c_m$, а якість одержуваного рішення оцінюється за допомогою будь-якого з критеріїв, прийнятих в чіткій кластеризації [1].

Якщо з точки зору використаного критерію якість кластеризації виявляється незадовільною, можна або зменшити значення d_c, або збільшити число можливих кластерів, тобто $j = 1,...,m, m+1, m+2,....$

Далі процедура нечіткої кластеризації повторюється, починаючи з першого кроку.

5. Починаючи з п'ятого кроку реалізується процедура нечіткої кластеризації. При цьому для кожної точки $x(k) \neq c_j$ розраховуються рівні нечіткої належності в стандартній формі [9]

$$\mu_j(k) = \frac{d_{jk}^{-2}}{\sum_{l=1}^{m} d_{lk}^{-2}}, \quad (1)$$

або на основі функції щільності розподілу Коші [13]

$$\mu_j(k) = \left(1 + \frac{d_{jk}^{-2}}{\sigma_j^2}\right), \quad (2)$$

де

$$\sigma_j^2 = \left(\sum_{\substack{l=1 \\ l \neq k}}^{m} d_{lk}^{-2}\right)^{-1}.$$

6. На основі оцінок ймовірнісної нечіткої належності (1), (2) розраховується рівень довіри до отриманих результатів на основі стандартного правдоподібного підходу [10, 11]

$$Cr_j(k) = \frac{1}{2}\left(\mu_j^*(k) + 1 - \sup \mu_j^*(k)\right), \quad (3)$$

де

$$\mu_j^*(k) = \frac{\mu_j(k)}{\sup \mu_l(k)}.$$

7. Завершення процедури нечіткої кластеризації шляхом оцінки якості результатів за допомогою будь-якого з критеріїв, що застосовуються в нечіткої кластеризації [9], хоча оцінка (3) вже сама по собі надає наскільки можна довіряти правдоподібності отриманих результатів.

4 ЕКСПЕРИМЕНТИ

Експериментальні дослідження методу швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу(FCDP) даних був реалізований на трьох масивах даних: табл. 1. Порівняльний аналіз проведено з відомими методами кластеризації які використовують параметр піків щільності розподілу даних, а саме: *k*-середніх, DBSCAN, який для заданої множини точок у деякому просторі відносить в одну групу точки, які розташовані найбільш щільно та розмічає точки, які лежать в областях з невеликою щільністю; DENCLUE, в якому кластери визначаються локальними максимумами оцінки щільності; OPTICS – алгоритм знаходження щільності на основі кластерів у просторових даних, що вирішує проблему визначення значущих кластерів в наборах даних різної щільності. За допомогою цих алгоритмів було проведено аналіз якості кластеризації на основі цих вибірок. Заздалегідь, для порівняльного аналізу із вибірок бралась частина спостережень і проводився аналіз якості кластеризації даних, яка виміряна показником нормалізованої взаємної інформації (приймає значення 1, якщо ідеальна кластеризація даних знайдена).

DOI 10.15588/1607-3274-2022-1-9

Таблиця 1 – Зразки даних

Назва вибірки	Кількість спостережень	Кількість атрибутів
iris	150	4
wine	178	13
ecoli	336	8

5 РЕЗУЛЬТАТИ

За результатами аналізу кластеризації була отримана інформація, яка представлена у табл. 2. Для кожної з вибірок перевірили, як розмір вибірки впливає на якість кластеризації.

Таблиця 2 – Значення показника нормалізованої взаємної інформації для різних даних та методів, перше число у трьох правих стовпцях показує розмір вибірки

Data		k-means	FCDP	DBSCAN	DENCLUE	OPTICS
iris	0,8	0,67±0,06	0,79±0,03	0,68±0,06	0,67±0,06	0,78±0,02
	0,4	0,65±0,07	0,68±0,18	0,60±0,06	0,67±0,06	0,72±0,08
	0,2	0,64±0,07	0,64±0,10	0,54±0,04	0,54±0,07	0,64±0,06
wine	0,8	0,70±0,11	0,78±0,02	0,66±0,01	0,68±0,01	0,76±0,04
	0,4	0,70±0,05	0,78±0,04	0,62±0,00	0,72±0,00	0,72±0,08
	0,2	0,58±0,21	0,69±0,11	0,48±0,01	0,70±0,01	0,59±0,01
ecoli	0,8	0,65±0,02	0,77±0,09	0,75±0,07	0,75±0,11	0,75±0,05
	0,4	0,65±0,04	0,75±0,05	0,63±0,01	0,73±0,05	0,73±0,07
	0,2	0,65±0,03	0,70±0,10	0,55±0,01	0,66±0,21	0,68±0,11

На рис. 1 продемонстрована залежність нормалізованої взаємної інформації (NMI) від розміру навчальної вибірки, що дає змогу говорити про те, що розмір вибірки не впливає на якість кластеризації, а NMI не є лінійним. Таким чином, якість кластеризації не втрачається навіть при 20% наявності вибірки.

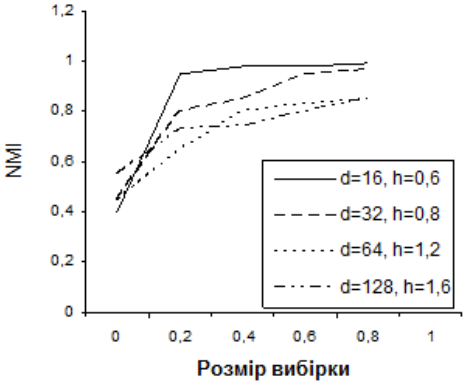

Рисунок 1 – Залежність показника нормалізованої взаємної інформації (NMI) від розміру навчальної вибірки

6 ОБГОВОРЕННЯ

За результатами експериментальних досліджень та аналізу отриманих результатів, можна зробити висновок, що запропонований метод швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу даних порівняно з методами, заснованими на використанні піків щільності, демонструє гарні результати роботи.

Порівняльний аналіз запропонованого методу продемонстровано в табл. 2, в якій наведені значення показника нормалізованої взаємної інформації для різних даних та методів, перше число у трьох правих стовпцях показує розмір вибірки. Аналізуючи табл. 2, можна зробити висновки, що якість кластеризації даних не втрачається від кількості наявних спостережень у вибірці, тобто, незалежно від 20%, 40% або 80% наявності вибірки, якість кластеризації не зменшується.

Як видно із порівняльної табл. 2, показник нормалізованої взаємної інформації (NMI), що приймає значення 1, якщо ідеальна кластеризація даних знайдена серед всіх запропонованих методів кластеризацій найкращій результат демонструє, коли даних все-ж таки більше. Якщо порівнювати якість кластеризації даних з відомими методами, можна зробити висновок, що запропонований метод швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу даних (FCDP) демонструє значно вищі показники ніж *k*-means, DBSCAN і DENCLUE та майже однаково з методом OPTICS. Так, якщо більш детально проаналізувати результат роботи цих двох методів по кількості спостережень, показник нормалізованої взаємної інформації у методі FCDP трошки вищій за OPTICS незалежно від виду вибірки, що подається на кластеризацію.

ВИСНОВКИ

Розглянуто задачу нечіткої кластеризації масивів даних на основі гібридного методу, заснованого на одночасному використанні правдоподібного підходу до нечіткої кластеризації і алгоритму знаходження типів щільності розподілу вихідних даних. Особливістю запропонованого методу є обчислювальна простота і висока швидкість, пов'язана з тим, що весь масив обробляється тільки один раз, тобто виключається необхідність в багатоепоховому самонавчанні, що реалізується в традиційних алгоритмах нечіткої кластеризації. Результати обчислювального експерименту підтверджують ефективність запропонованого підходу в задачах кластерзації в умовах коли кластери перетинаються.

Наукова новизна: вперше запропонована процедура швидкої нечіткої кластеризації даних з використанням піків щільності розподілу даних на основі правдоподібного підходу.

Практичне значення: результати експерименту дозволяють рекомендувати запропоновані методи для використання на практиці для вирішення проблем автоматичної кластеризації великих даних.

Перспективи подальших досліджень методи нечіткої кластеризації даних для широкого класу практичних проблем.

DOI 10.15588/1607-3274-2022-1-9

ПОДЯКА

Робота виконана в рамках науково-дослідного проєкту державного бюджету Харківського національного університету радіоелектроніки «Глибокі гібридні системи обчислювального інтелекту для аналізу потоків даних та їх швидке навчання» (номер державної реєстрації 0119U001403).

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Xu R. Clustering / R. Xu, D. C. Wunsch. – Hoboken N. J.: John Wiley & Sons, Inc., 2009. – 398 p.
2. Nadaraya E. A. On nonparametric estimates of density function and regression curves / E. A. Nadaraya // Theory of Probabilistic Application. – 1965. – № 10 – P. 186–190.
3. Epanechnikov V. A. Nonparametric estimation of multivariate probability density / V. A. Epanechnikov // Probability theory and its Application. – 1968. – 14, №2. – P. 156–161.
4. Fukunaga K. The estimation of the gradient of a density function with application in pattern recognition / K. Fukunaga, L. D. Hostler // IEEE Trans. on Inf. Theory, Jan., 1975. – IEEE. – 1975. – № 21 – P. 32–40. DOI: 10.1109/TIT.1975.10 55330.
5. Ester M. A density – based algorithm for discovering clusters in large spatial databases with noise / [M. Ester, H. Kriegel, J. Sandler, X. Xu] // Proc. 2nd Int. Conf. on Knowledge Discovering and Data Mining. – KDD96, N.Y.: AAAI Press, Aug. 2, 1996. – P. 226–231.
6. Hinneburg A. An efficient approach to clustering in large multimedia databases with noise / A. Hinneburg, D. Klein // Proc. 4th Int. Conf. in Knowledge Discovering and Data Mining. – KDD98, N.Y.: AAAI Press, Aug. 27, 1998. – Hinneburg, 1998. – P. 58–65.
7. OPTICS: Ordering points to identify the clustering structure / [M. Ankerst, M. Brening, H. Kriegel, J. Sander] // Proc. 1999 ACM SIGMOD Int. Conf. on Management of Data, Jun. 1, 1999. – Philadelphia, 1999. – P. 49–60.
8. Rodriguez A. Clustering by fast seach and find of density peaks / A. Rodriguez, A. Laio. – Science. – 2014. – № 34. – P. 1492–1496.
9. Fuzzy Clustering Analysis: Methods for Classification, Data Analisys and Image Recognition / [F. Höppner, F. Klawonn, R. Kruse, T. Runkler]. – Chichester : John Wiley &Sons, 1999. – 300 p.
10. Credibilistic clustering: the model and algorithms / [J. Zhou, Q. Wang, C.-C. Hung, X. Yi] // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. – 2015. – Vol. 23, № 4. – P. 545–564. DOI: https://doi.org/10.1142/S0218488515500245
11. Zhou J. Credibilistic clustering algorithms via alternating cluster estimation / J. Zhou, Q. Wang, C. C. Hung // Journal of Intelligent Manufacturing. – 2017. – Vol. 28. – P. 727–738. DOI: https://doi.org/10.1007/s10845-014-1004-6.
12. Begum N. Accelerating dynamic time warping clustering with a novel admissible pruning strategy / [N. Begum, L. Ulanova, J. Wamg, E. Klogh] // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 10, 2015. – Sydney, NSW, Australia. – P. 49–58. DOI: https://doi.org/10.1145/2783258.27 83286.
13. Online credibilistic fuzzy clustering of data using membership functions of special type [Electronic resource] / [A. Shafronenko, Ye. Bodyanskiy, I. Klymova, O. Holovin] // Proceedings of The Third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), April 27–1 May 2020. – Zaporizhzhia, 2020. – Access mode: http://ceur-ws.org/Vol-2608/paper56.pdf.

Стаття надійшла до редакції 28.10.2021.
Після доробки 25.12.2021.

УДК 004.8:004.032.26

БЫСТРАЯ НЕЧЕТКАЯ ПРАВДОПОДОБНАЯ КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ АНАЛИЗА ПИКОВ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ДАННЫХ

Бодянский Е. В. – д-р техн. наук, профессор, профессор кафедры искусственного интеллекта Харьковского национального университета радиоэлектроники, Харьков, Украина.

Плісс І. П. – канд. техн. наук, ведущий научный сотрудник ПНДЛ АСУ Харьковского национального университета радиоэлектроники, Харьков, Украина.

Шафроненко А. Ю. – канд. техн. наук, доцент, доцент кафедры информатики Харьковского национального университета радиоэлектроники, Харьков, Украина.

АННОТАЦИЯ

Актуальность. Проблема кластеризации (классификации без учителя) часто встречается при обработке массивов данных различной природы и является достаточно интересной и неотъемлемой частью искусственного интеллекта. Для решения этой задачи существует множество известных методов и алгоритмов основанных на анализе плотности распределения наблюдений в анализируемых данных. Однако эти методы достаточно сложны в программной реализации и не лишены недостатков, а именно: проблемы определения значимых кластеров в наборах данных различной плотности, многоэпоховое самообучение, застревание в локальных экстремумах целевых функций и тому подобное. Следует отметить, что методы, основанные на анализе пиков плотности распределения данных, являются по своей природе четкими, поэтому для расширения возможностей этих методов целесообразно ввести их нечеткую модификацию.

Цель. Цель работы заключается в введении быстрой процедуре нечеткой кластеризации данных с использованием пиков плотности распределения данных, которая может находить экстемумы (центры) кластеров, которые пересекаются независимо от количества поступающих данных.

Метод. Рассмотрена задача нечеткой кластеризации массивов данных на основе гибридного метода, основанного на одновременном использовании правдоподобного подхода к нечеткой кластеризации и алгоритма нахождения типов плотности распределения исходных данных. Особенностью предлагаемого метода является вычислительная простота и высокая скорость, связанная с тем, что весь массив обрабатывается только один раз, то есть исключается необходимость в многоэпоховом самообучении, реализуемом в традиционных алгоритмах нечеткой кластеризации.

Результаты. Особенностью предложенного метода быстрой нечеткой правдоподобной кластеризации на основе анализа пиков плотности распределения данных является вычислительная простота и высокая скорость, связанная с тем, что весь массив обрабатывается только один раз, то есть исключается необходимость в многоэпоховом самообучении, что реализуется в традиционных алгоритмах нечеткой кластеризации. Результаты вычислительного эксперимента подтверждают эффективность предложенного подхода в задачах кластеризации в условиях, когда кластеры пересекаются.

Выводы. Результаты эксперимента позволяют рекомендовать разработанный метод для решения проблем автоматической кластеризации и классификации данных, максимально быстро находить центры кластеров. Предложенный метод бы-

DOI 10.15588/1607-3274-2022-1-9

строй нечеткой правдоподобной кластеризации на основе анализа пиков плотности распределения данных предназначен для использования в системах вычислительного интеллекта, нейро-фаззи системах, в обучении искусственных нейронных сетей и в задачах кластеризации.

КЛЮЧЕВЫЕ СЛОВА: нечеткая кластеризация, правдоподобная кластеризация, пики плотности распределения данных.

UDC 004.8:004.032.26

FAST FUZZY CREDIBILISTIC CLUSTERING BASED ON DENSITY PEAKS DISTRIBUTION OF DATA BROAKYSIS

Bodyanskiy Ye. V. – Dr. Sc., Professor at the Department of Artificial Intelligence, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Pliss I. P. – PhD, Leading Researcher at Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Shafronenko A. Yu. – PhD, Associated Professor at the Department of Informatics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. The problem of clustering (classification without a teacher) is often occures when processing data arrays of various natures, which is quite an interesting and integral part of artificial intelligence. To solve this problem, there are many known methods and algorithms based on the principles of the distribution density of observations in the analyzed data. However, these methods are rather complicated in software implementation and are not without drawbacks, namely: the problem of determining significant clusters in datasets of different densities, multiepoch self-learning, getting stuck in local extrema of goal functions, etc. It should be noted that the methods based on the analysis of the peaks of the data distribution density are clear in nature, therefore, to expand the capabilities of these methods, it is advisable to introduce their fuzzy modification.

Objective. The aim of the work is to introduce fast fuzzy data clustering using density peaks distribution of the datasets, that can find the prototypes (centroids) of clusters that overlapping regardless of the amount of incoming data.

Method. The problem of fuzzy clustering data arrays based on a hybrid method that based on the simultaneous use of a credibilistic approach to fuzzy clustering and an algorithm for finding the types of distribution density of the initial data is proposed. A feature of the proposed method is computational simplicity and high speed, due to the fact that the entire array is processed only once, that is, eliminates the need for multi-era self-learning, implemented in traditional fuzzy clustering algorithms.

Results. A feature of the proposed method of fast fuzzy credibilistic clustering using of density peaks distribution is characterized by computational simplicity and high speed due to the fact that the entire array is processed only once, that is, the need for multiepoch self-learning is eliminated, which is implemented in traditional fuzzy clustering algorithms. The results of the computational experiment confirm the effectiveness of the proposed approach in clustering problems under conditions in the case when the clusters are ovelap.

Conclusions. The experimental results allow us to recommend the developed method for solving the problems of automatic clustering and data classification, as quickly as possible to find the centroids of clusters. The proposed method of fast fuzzy credibilistic clustering using of density peaks distribution of dataset is intended for use in computational intelligence systems, neuro-fuzzy systems, in training artificial neural networks and in clustering problems.

KEYWORDS: fuzzy clustering, credibilistic clustering, density peak of dataset.

REFERENCES

1. Xu R., Wunsch D. C. Clustering. Hoboken N.J., John Wiley & Sons, Inc., 2009, 398 p.
2. Nadaraya E. A. On nonparametric estimates of density function and regression curves, *Theory of Probabilistic Application*, 1965, No. 10, pp. 186–190.
3. Epanechnikov V. A. Nonparametric estimation of multivariate probability density, *Probability theory and its Application*, 1968, 14, No. 2, pp. 156–161.
4. Fukunaga K., Hostler L. D. The estimation of the gradient of a density function with application in pattern recognition, *IEEE Trans. on Inf. Theory, Jan., 1975, IEEE*, 1975, No. 21, pp. 32–40. DOI: 10.1109/TIT.1975.10 55330.
5. Ester M., Kriegel H., Sandler J., Xu X. A density – based algorithm for discovering clusters in large spatial databases with noise, *Proc. 2nd Int. Conf. on Knowledge Discovering and Data Mining – KDD96, N.Y.: AAAI Press, Aug. 2*, 1996, pp. 226–231.
6. Hinneburg A., Klein D. An efficient approach to clustering in large multimedia databases with noise, *Proc. 4th Int. Conf. in Knowledge Discovering and Data Mining – KDD98, N.Y.: AAAI Press, Aug. 27, 1998.* Hinneburg, 1998, pp. 58–65.
7. Ankerst M., Brening M., Kriegel H., Sander J. OPTICS: Ordering points to identify the clustering structure. *Proc. 1999 ACM SIGMOD Int. Conf. on Management of Data, Jun. 1, 1999*. Philadelphia, 1999, pp.49–60.
8. Rodriguez A., Laio A. Clustering by fast seach and find of density peaks, *Science,* 2014, № 34, pp. 1492–1496.
9. Höppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Clustering Analysis: Methods for Classification, Data Analisys and Image Recognition. Chichester, John Wiley &Sons, 1999, 300 p.
10. Zhou J., Wang Q., Hung C.-C., Yi X. Credibilistic clustering: the model and algorithms, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2015, Vol. 23, No. 4, pp. 545–564. DOI: https://doi.org/10.1142/S0218488515500245
11. Zhou J., Wang, Q., Hung C. C. Credibilistic clustering algorithms via alternating cluster estimation, *Journal of Intelligent Manufacturing,* 2017, Vol. 28, pp. 727–738. DOI: https://doi.org/10.1007/s10845-014-1004-6.
12. Begum N., Ulanova L., Wamg J., Klogh E. Accelerating dynamic time warping clustering with a novel admissible pruning strategy, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 10, 2015.* Sydney, NSW, Australia, pp. 49–58. DOI: https://doi.org/10.1145/2783258.27 83286.
13. Shafronenko A., Bodyanskiy Ye., Klymova I., Holovin O. Online credibilistic fuzzy clustering of data using membership functions of special type[Electronic resource], *Proceedings of The Third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), April 27–1 May 2020.* Zaporizhzhia, 2020. Access mode: http://ceur-ws.org/Vol-2608/paper56.pdf.

DOI 10.15588/1607-3274-2022-1-9