UDC 004.8:004.032.26

CREDIBILISTIC ROBUST ONLINE FUZZY CLUSTERING IN DATA STREAM MINING TASKS

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

Kasatkina N. V. – Dr. Sc., Professor, Division of doctoral and post-graduate training, National University of Food Technologies, Kyiv, Ukraine.

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

Shafronenko Ye. O. – Assistant at the Department of Media Engineering and Information Radio Electronic Systems, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. The task of clustering-classification without a teacher of data arrays occupies an important place in the general problem of Data Mining, and for its solution there exists currently many approaches, methods and algorithms. There are quite a lot of situations where the real data to be clustered are corrupted with anomalous outliers or disturbances with non-Gaussian distributions. It is clear that "classical" methods of artificial intelligence (both batch and online) are ineffective in this situation. The goal of the paper is to develop a credibilistic robust online fuzzy clustering method that combines the advantages of credibilistic and robust approaches in fuzzy clustering tasks.

Objective. The goal of the work is online credibilistic fuzzy clustering of distorted data, using of credibility theory in data stream mining.

Method. The procedure of fuzzy clustering of data using credibilistic approach based on the use of both robust goal functions of a special type, insensitive to outliers and designed to work both in batch and its recurrent online version designed to solve Data Stream Mining problems when data are fed to processing sequentially in real time.

Results. Analyzing the obtained results overall accuracy of clustering methods and algorithm, proposed method similar with result of credibilistic fuzzy clustering method, but has time superiority regardless of the number observations that fed on clustering process.

Conclusions. The problem of fuzzy clustering of data streams contaminated by anomalous non-Gaussian distributions is considered. A recurrent credibilistic online algorithm based on the objective function of a special form is introduced, which suppresses these outliers by using the hyperbolic tangent function, which, in addition to neural networks, is used in robust estimation tasks. The proposed algorithm is quite simple in numerical implementation and is a generalization of some well-known online fuzzy clustering procedures intended for solving Data Stream Mining problems.

KEYWORDS: fuzzy clustering, distorted data, credibilistic fuzzy clustering, Data Stream Mining, robust function.

ABBREVIATIONS

FCM is a fuzzy c-means method;

SOM is self-organizing map;

CROFC is credibilistic robust online fuzzy clustering method.

NOMENCLATURE

X is a data set matrix;

N is number of observations;

R is space of input vectors;

n is number of attributes;

m is number of overlapping classes;

k is a number of the vectors-observation;

i is a number components of the vectors-observation;

$x(k)$ is a vector of observations;

$x_i(k)$ is a preprocessed original data;

l, j is a number of clusters;

μ is a membership level;

$\mu_j(k)$ is a membership level of k-th vector-observation to j-th cluster;

c is a centroid of cluster;

c_j is a centroid of j-th cluster;

d is a Euclidean distance;

d_p is a Minkowski distance;

J is a goal function;

L is Lagrange function;

$\eta(k)$ is learning-rate parameter;

$Cr_j(k)$ is fuzzy credibilistic membership level;

$\lambda(k)$ is indefinite Lagrange multiplier;

β is a fuzzifier;

β_i is parameter specifying the modification of function.

INTRODUCTION

The task of clustering-classification without a teacher of data arrays occupies an important place in the general problem of Data Mining, and for its solution there are currently many approaches, methods and algorithms [1–3].

A special place here is occupied by methods of fuzzy clustering, when it is a priori assumed that each observation can simultaneously belong to several or all classes at the same time with different levels of fuzzy membership, i. e. classes overlapping in the feature space [4, 5].

DOI 10.15588/1607-3274-2023-3-10

Fuzzy clustering methods can be conditionally divided into two large classes: probabilistic, among which the Fuzzy C-means algorithm (FCM) by J. Bezdek [4] was the most popular, and probabilistic. Each of these classes has its advantages and disadvantages, and to overcome these disadvantages, a so-called credibilistic approach was proposed [6, 7], which has already proved its effectiveness in solving a number of problems.

There are quite a lot of situations where the real data to be clustered are corrupted with anomalous outliers or disturbances with non-Gaussian distributions. This leads to the fact that traditional methods using quadratic metrics (Euclidian, Mahalanobis, etc.) do not provide the desired results. This led to the creation of robust clustering methods [8–10] resistant to these outliers and based on non-quadratic distances, while most of the known robust fuzzy clustering algorithms are based on a probabilistic approach.

It is appropriate to develop a credibilistic robust online fuzzy clustering (CROFC) method that combines the advantages of credibilistic and robust approaches in fuzzy clustering tasks and is designed to process data streams that arrive sequentially in real time.

The object of study is fuzzy clustering of data distorted by outliers.

The subject of study is procedure for fuzzy clustering of data distorted by outliers based on robust approaches in fuzzy clustering tasks.

The purpose of the work is to introduce robust online credibilistic method for fuzzy clustering of distorted data.

1 PROBLEM STATEMENT

The initial information for solving the clustering problem is an unlabeled sample of vector observations $X = \{x(1), x(2), ..., x(k), ..., x(N)\} \subset R^n$ where k-th observation number, in the sample when working in batch mode or index of the current discrete time, when solving Data Stream Mining tasks. The result of solving the problem of fuzzy clustering is the division of this sample into m overlapping classes-clusters with estimation of fuzzy membership levels $\mu_j(k)$ to each of the possible clusters c_j, $j = 1, 2, ..., m$.

2 REVIEWS OF THE LITERATURE

A special place in the general problem of Data Mining is occupied by tasks related to Data Stream Mining when the data to be processed, the tasks are not in the form of a batch, but are sequentially received for processing one by one, while the amount of this data is unknown a priori. The most characteristic example here is T. Kohonen's self-organizing maps (SOM) [11], which implements the traditional crisp algorithm of K-means clustering in an online version using the self-learning rule "Winner Takes All" (WTA). For fuzzy situations D.C. Park and J. Dagger [12] have proposed a recurrent version of FCM, and in [13] both probabilistic and probabilistic recurrent algorithms of online fuzzy clustering were considered.

Regarding credibilistic and robust approaches, in [14, 15] recurrent modifications of the credibilistic fuzzy clustering algorithm were proposed, and in [16, 17] recurrent robust procedures of fuzzy clustering designed for processing data streams in online mode where introduced.

3 MATERIALS AND METHODS

The most popular approach to solving this problem is related to the minimization of the objective function

$$J\left(\mu_j(k), c_j\right) = \sum_{k=1}^{N}\sum_{j=1}^{m}\mu_j^{\beta}(k) d_p^2\left(x(k), c_j\right)$$

with restrictions

$$\sum_{j=1}^{m}\mu_j(k) = 1,$$

$$0 < \sum_{j=1}^{m}\mu_j(k) < 1$$

where c_j – prototype-centroid of j-th cluster; β – parameter-fuzzifier (usually $\beta = 2$); $d_p^2\left(x(k), c_j\right)$ – the distance between $x(k)$ and c_j. Most often, this is the Minkowski distance

$$d_p\left(x(k), c_j\right) = \left\|x(k) - c_j\right\|_p = \left(\sum_{i=1}^{n}\left|x_i(k) - c_{ji}\right|^p\right)^{\frac{1}{p}},$$

the special case of which is the traditional Euclidean norm

$$d_2\left(x(k), c_j\right) = \left\|x(k) - c_j\right\|_2 = \left(\sum_{i=1}^{n}\left|x_i(k) - c_{ji}\right|^2\right)^{\frac{1}{2}}.$$

Using the standard procedure of non-linear programming, the Lagrange function is introduced for consideration

$$L\left(\mu_j(k), c_j, \lambda(k)\right) = \sum_{k=1}^{N}\sum_{j=1}^{m}\mu_j^{\beta}(k) d_p^2\left(x(k), c_j\right) +$$

$$+\sum_{k=1}^{N}\lambda(k)\left(\sum_{j=1}^{m}\mu_j(k) - 1\right) = \quad (1)$$

$$= \sum_{k=1}^{N}\left(\sum_{j=1}^{m}\mu_j^{\beta}(k) d_p^2\left(x(k), c_j\right) + \lambda(k)\left(\sum_{j=1}^{m}\mu_j(k) - 1\right)\right)$$

(here $\lambda(k)$ – the unknown Lagrange multiplier) and the system of Kuhn-Tucker equations

DOI 10.15588/1607-3274-2023-3-10

$$\begin{cases} \dfrac{\partial L\left(\mu_j(k), c_j, \lambda(k)\right)}{\partial \mu_j(k)} = 0, \\ \dfrac{\partial L\left(\mu_j(k), c_j, \lambda(k)\right)}{\partial \lambda(k)} = 0, \\ \nabla_{c_j} L\left(\mu_j(k), c_j, \lambda(k)\right) = \vec{0}, \end{cases} \quad (2)$$

solving which we get the result in the form:

$$\begin{cases} \mu_j(k) = \dfrac{\left(d_p^2\left(x(k), c_j\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(d_p^2\left(x(k), c_l\right)\right)^{\frac{1}{1-\beta}}}, \\ \lambda(k) = -\left(\sum\limits_{l=1}^{m}\left(\beta d_p^2\left(x(k), c_l\right)\right)^{\frac{1}{1-\beta}}\right)^{1-\beta}, \\ c_j = \dfrac{\sum\limits_{k=1}^{N}\mu_j^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N}\mu_j^{\beta}(k)}, \end{cases}$$

if $\beta = 2$ turns into the classical FCM algorithm of J. Bezdek [4]:

$$\begin{cases} \mu_j(k) = \dfrac{\left\|x(k) - c_j\right\|_2^2}{\sum\limits_{l=1}^{m}\left\|x(k) - c_j\right\|_2^2}, \\ \lambda(k) = -\sum\limits_{l=1}^{m}\left(\dfrac{\left\|x(k) - c_j\right\|_2^2}{2}\right)^{-1}, \\ c_j = \dfrac{\sum\limits_{k=1}^{N}\mu_j^2(k)x(k)}{\sum\limits_{k=1}^{N}\mu_j^2(k)}. \end{cases}$$

For the online clustering procedure to find the saddle point of the Lagrange function (1), instead of directly solving the system of Kuhn-Tucker equations (2), the Arrow-Hurwitz-Uzawa algorithm [18] can be used, with the help of which we obtain a recurrent procedure [19]

$$\begin{cases} \mu_j(k) = \dfrac{\left(d_p^2\left(x(k), c_j(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(d_p^2\left(x(k), c_l(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ c_j(k+1) = c_j(k) - \eta(k)\nabla_{c_j} L\left(\mu_j(k), c_j(k), \lambda(k)\right) = \\ = c_j(k) - \eta(k)\mu_j^{\beta}(k) d_p\left(x(k+1), c_j(k)\right) * \\ *\nabla_{c_j} d_p^2\left(x(k+1), c_j(k)\right) \end{cases}$$

(here $\eta(k)$ – self-learning parameter) which $\beta = 2$, $p = 2$ turns into GBFC Park-Dagger algorithm [12]

$$\begin{cases} \mu_j(k) = \dfrac{\left\|x(k) - c_j(k)\right\|_2^{-2}}{\sum\limits_{l=1}^{m}\left\|x(k) - c_l(k)\right\|_2^{-2}}, \\ c_j(k+1) = c_j(k) + \eta(k)\mu_j^2(k)\left(x(k+1) - c_j(k)\right). \end{cases} \quad (3)$$

Here it is interesting to notice that the second relation (3) should completely match the self-learning rule "Winner Takes More" (WTM) by T. Kohonen [11] in terms of structure, while the role of the neighborhood function here is performed by a set $\mu_j^2(k)$.

All the algorithms discussed above, based on the Minkowski metric, do not ensure the process of clustering robust properties, since they only "amplify" the influence of anomalous emissions present in the data sample.

Therefore, it is advisable to use distances that have robust properties, "suppressing" these emissions.

One of these distances can be based on a function [20, 21] of the form:

$$d^R\left(x(k), c_j\right) = \sum_{i=1}^{n}\beta_i \ln\left(\cosh\left(\frac{x_i(k) - c_{ji}}{\beta_i}\right)\right), \quad (4)$$

where β_i – the parameter specifying the modification of this function is usually accepted $\beta_i = 2$, $i = 1, 2, \ldots n$.

Introducing the robust objective function further

$$J^R\left(\mu_j(k), c_j\right) = \sum_{k=1}^{N}\sum_{j=1}^{m}\mu_j^{\beta}(k) d^R\left(x(k), c_j\right) =$$
$$= \sum_{k=1}^{N}\sum_{j=1}^{m}\mu_j^{\beta}(k)\sum_{i=1}^{n}\beta_i \ln\left(\cosh\left(\frac{x_i(k) - c_{ji}}{\beta_i}\right)\right)$$

and the corresponding Lagrange function

$$L^R\left(\mu_j(k), c_j, \lambda(k)\right) = \sum_{k=1}^{N}\sum_{j=1}^{m}\mu_j^{\beta}(k)\sum_{i=1}^{n}\beta_i \times$$
$$\times \ln\left(\cosh\left(\frac{x_i(k) - c_{ji}}{\beta_i}\right)\right) + \sum_{k=1}^{N}\lambda(k)\left(\sum_{j=1}^{m}\mu_j(k) - 1\right),$$

it is possible to write the system of Kuhn-Tucker equations, which, however, due to the complexity of the distance (4), does not have an analytical solution.

Therefore, the only solution here is to use the same Arrow-Hurwitz-Uzawa algorithm, which leads to the result and

DOI 10.15588/1607-3274-2023-3-10

$$\begin{cases} \mu_j(k) = \dfrac{\left(d^R\left(x(k), c_j(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(d^R\left(x(k), c_l(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ c_j(k+1) = c_j(k) - \eta(k)\dfrac{\partial}{\partial c_{ji}} L^R\left(\mu_j(k), c_j(k), \lambda(k)\right) = \\ = c_j(k) + \eta(k)\mu_j^{\beta}(k)\tanh\left(\dfrac{x_i(k) - c_{ji}(k)}{\beta_i}\right), \end{cases} \quad (5)$$

where function $\tanh(\bullet)$, which is usually used as an activation in many neural networks is used to suppress specifically anomalous outliers in the data.

Next, taking the value of the fuzzifier $\beta = 2$ (5) can be rewritten in a somewhat simplified form:

$$\begin{cases} \mu_j(k) = \dfrac{\left(d^R\left(x(k), c_j(k)\right)\right)^{-1}}{\sum\limits_{l=1}^{m}\left(d^R\left(x(k), c_l(k)\right)\right)^{-1}}, \\ c_{ji}(k+1) = c_{ji}(k) + \eta(k)\mu_j^2(k)\tanh\left(\dfrac{x_i(k) - c_{ji}(k)}{\beta_i}\right), \end{cases} \quad (6)$$

which is essentially a robust Park-Dagger algorithm (3) that suppresses outliers in the data using the function $\tanh(\bullet)$.

Based on this algorithm, it is easy to consider its credibilistic modification by supplementing (6) with a simple relation [6, 7]:

$$\begin{cases} \mu_j^r(k) = \dfrac{\mu_j(k)}{\sup \mu_l(k)}, \\ Cr_j(k) = \dfrac{1}{2}\left(\mu_j^r(k) + 1 - \sup \mu_l(k)\right). \end{cases} \quad (7)$$

Ratios (6), (7) define the credibilistic robust online fuzzy clustering algorithm, intended for use in systems for processing data streams distorted by various types of disturbances and arriving online.

4 EXPERIMENTS

For test the method of credibilistic robust online fuzzy clustering in data stream mining tasks was conducted test data sets of Nursery from the UCI repository.

Nursery Database was derived from a hierarchical decision model originally developed to rank applications for nursery schools. It was used during several years in 1980's when there was excessive enrollment to these schools in Ljubljana, Slovenia, and the rejected applications frequently needed an objective explanation.

The final decision depended on three subproblems: occupation of parents and child's nursery, family structure and financial standing, and social and health picture of the family. The Nursery Database contains 12958 examples with the structural information removed, i.e., directly relates NURSERY to the eight input attributes: parents, has_nurs, form, children, housing, finance, social, health.

5 RESULTS

We compared the obtained results with classical FCM algorithm, probabilistic, possibilistic and credibilistic fuzzy clustering methods.

Table 1 – The overall accuracy of clustering methods and algorithm

Clustering algorithm	Overall accuracy		
	Highest	Mean	Variance
FCM	68.54	68.54	0.01
Credibilistic fuzzy clustering	67.98	67.98	0
Possibilistic fuzzy clustering	68.55	68.54	0.01
Probabilistic fuzzy clustering	68.48	68.48	0.01
CROFC	67.68	67.65	0

A comparative analysis of the quality of the clustering data was carried out according to the main characteristics of the quality ratings, such as the speed of data clustering and the average error.

Table 2 show the results of the algorithms proposed for comparison with different numbers of observations.

Table 2 – Comparative characteristics of the average error with different number of observations in percentage

Algorithm	50	Time	100	Time	150	Time
FCM	1.62	1.19	1.35	2.55	0.98	3.03
Probabilistic fuzzy clustering	1.66	1.62	1.32	2.72	0.99	3.12
Possibilistic fuzzy clustering	1.22	1.15	1.02	2.02	0.75	2.10
Credibilistic fuzzy clustering	0.69	1.02	0.49	1.33	0.14	1.41
CROFC	0.68	1.00	0.45	1.25	0.12	1.33

Analyzing the obtained results, it can be concluded that regardless of the size of the initial information submitted for processing by the proposed method for comparing performance and efficiency, it is not inferior in speed and quality of clustering in comparison with known algorithms and methods.

The comparative analysis is demonstrated on the diagrams of the dependence of error and time on the number of observations on Fig. 1.

DOI 10.15588/1607-3274-2023-3-10

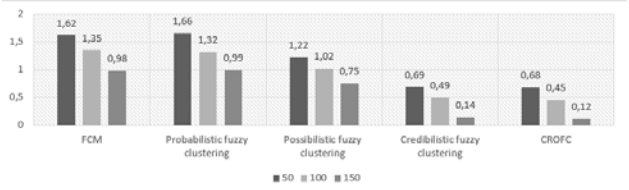

Figure 1 – Diagram of the dependence of the error on the number of observations (50, 100, 150)

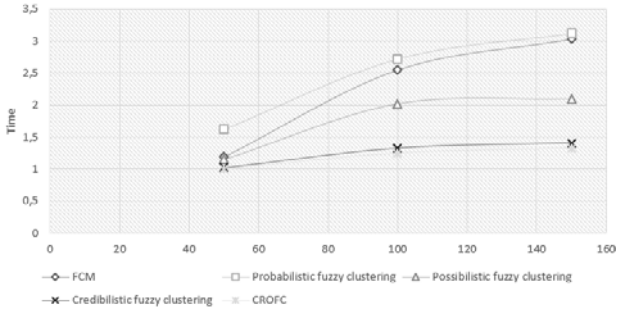

Figure 2 – Diagram of dependence of clustering time on the number of observations (50, 100, 150)

6 DISCUSSIONS

The result of clustering data set Nursery shown in Table 1 and Table 2. As the table shows, the prepositional credibilistic robust online fuzzy clustering in data stream mining tasks have shown good results.

As it can be seen in Fig. 1 the proposed method shows best result on diagram of the dependence of the error on the number of observations and Fig. 2, that demonstrate dependence of clustering time on the number of observations.

Analyzing the obtained results overall accuracy of clustering methods and algorithm, proposed method similar with result of credibilistic fuzzy clustering method, but has time superiority regardless of the number observations that fed on clustering process.

Due to its adaptability and robustness proposed method does not require a lot of time to process the data received in real time, and does not burden itself with intermediate calculations due to adaptability functions.

This is quite clearly demonstrated by the diagrams of the dependence of the clustering time on the number of observations and the dependence of the error on the number of observations.

CONCLUSIONS

The problem of fuzzy clustering of data streams contaminated by anomalous non-Gaussian distributions is considered. A recurrent credibilistic online algorithm based on the objective function of a special form is introduced, which suppresses these outliers by using the hyperbolic tangent function, which, in addition to neural networks, is used in robust estimation tasks. The proposed algorithm is quite simple in numerical implementation and is a generalization of some well-known online fuzzy clustering procedures intended for solving Data Stream Mining problems.

The scientific novelty of obtained results is that the method of credibilistic robust online fuzzy clustering in data stream mining tasks, that shows good results in comparative analyses with another methods, that "worked" with distorted data sets.

The practical significance of obtained results is that analyze properties of the propose methods of credibilistic fuzzy clustering of distorted data. The experimental results allow to recommend the proposed methods for use in practice for solving the problems of automatic clusterization of distorted data.

Prospects for further research methods of online robust credibilistic fuzzy clustering of distorted data in tasks of stream data mining.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of Kharkiv National University of Radio Electronics "Development of methods and algorithms of combined learning of deep neuro-neo-fuzzy systems under the conditions of a short training sample".

REFERENCES

1. Gan G., Ma Ch., Wu J. Data Clustering: Theory, Algorithms and Applications. Philadelphia, Pennsylvania: SIAM: 2007. – 455 p. doi: https://doi.org/10.1137/ 1.9780898718348
2. Abony J., Feil D. Cluster Analysis for Data Mining and System Identification. Basel, Birkhouser, 2007, 303 p.
3. Xu R., Wunsch D. C. Clustering. Hoboken N.J., John Wiley & Sons, Inc., 2009, 398 p.
4. Bezdek J. C. Pattern recognition with fuzzy objective function algorithms. New York, Springer, 1981, 253 p. DOI https://doi.org/10.1007/978-1-4757-0450-1.
5. Höppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester, John Wiley &Sons, 1999, 300 p.
6. 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, № 4, pp. 545–564. DOI: https://doi.org/10.1142/S0218488515500245
7. 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.
8. Tsuda K., Senda S., Minoh M., Ikeda K. Sequential fuzzy cluster extraction and its robust against noise, *System and Computers in Japan*, 1997, 28, pp. 10–17.
9. Höppner F., Klawonn F. Fuzzy clustering of sampled functions, *19th Int. Conf. North American Fuzzy Information Processing Society (NAFIPS).* Atlanta, USA, 2000, pp. 257–255.
10. Georgieva O., Klawonn F. A clustering algorithm for identification of single clusters in large data sets, *Proc. 11th East – West Fuzzy Coll*. Zittau/Görlitz, FH, 2004, pp. 118–125.
11. Kohonen T. Self-Organizing Maps. Berlin, Springer, 1995, 362 p. DOI: 10.1007/978-3-642-56927-2.
12. Park D. C., Dagger I. Gradient based fuzzy c-means (GBFCM) algorithm, *IEEE International Conference on Neural Networks, 28 June – 2July,1984, proceedings.* Orlando, IEEE, 1984, pp. 1626–1631. DOI: 10.1109 / ICNN. 1994.374399.

DOI 10.15588/1607-3274-2023-3-10

13. Bodyanskiy Ye. Computational intelligence techniques for data analysis, *Lecture Notes in Informatics*. Bonn, Gesellschaft für Informatik, 2005, pp. 15–36.
14. 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.
15. Shafronenko A., Bodyanskiy Ye., Pliss I., Klymova I. Online Credibilistic Fuzzy Clustering Method Based on Cauchy Density Distribution Function, *2021 11th International Conference on Advanced Computer Information Technologies (ACIT): proceedings*. Deggendorf, Germany, IEEE, 2021, pp. 704–707. DOI: 10.1109/ACIT52158.2021.9548572
16. Bodyanskiy Ye., Gorshkov Ye., Kokshenev I., Kolodyazhniy V. Robust recursive fuzzy clustering algorithms, *Proc. 12th East West Fuzzy Coll 2005*. Zittau-Grölitz, FH, 2005, pp. 301–308.
17. Bodyanskiy Ye. Gorshkov Ye., Kokshenev I., Kolodyazhniy V. Outlier resistant recursive fuzzy clustering algorithms, *Ed. By B. Reusch «Computational Intelligence Theory and Applications» – Advances in Soft Computing*, Vol. 38. Berlin Heidelberg, Springer Verlag, 2006, pp. 647–652.
18. Arrow K. J., Hurwitz L., Uzawa H. Studies in Linear and Nonlinear Programming. Stanford University Press, 1958, 242 p.
19. Bodyanskiy Ye., Kolodyazhniy V., Stephan A. Recurcive fuzzy clustering algorithm, *Proc 10th East West Fuzzy Coll*, 2002. Zittau-Görlitz, HS, 2002, pp. 276–283.

Received 21.06.2023.
Accepted 17.08.2023.

УДК: 004.8:004.032.26

ДОСТОВІРНА РОБАСТНА ОНЛАЙН НЕЧІТКА КЛАСТЕРИЗАЦІЯ В ЗАДАЧАХ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ПОТОКІВ ДАНИХ

Шафроненко А. Ю. – канд. техн. наук, доцент кафедри інформатики Харківського національного університету радіоелектроніки, Харків, Україна.
Касаткіна Н. В. – д-р техн. наук, професор, відділ докторантури та аспірантури Національного університету харчових технологій, Київ, Україна.
Бодянський Є. В. – д-р техн. наук, професор, професор кафедри штучного інтелекту Харківського національного університету радіоелектроніки, Харків, Україна.
Шафроненко Є. О. - асистент кафедри медіаінженерії та інформаційних радіоелектронних систем, Харківський національний університет радіоелектроніки, Харків, Україна.

АНОТАЦІЯ

Актуальність. Задача кластеризації-класифікації без вчителя масивів даних займає важливе місце у загальній проблемі Data Mining, а для її вирішення існує на цей час безліч підходів, методів та алгоритмів. Існує достатньо багато ситуацій, коли реальні дані, що підлягають кластеризації, забруднені аномальними викидами або збуреннями з не Гаусівськими розподілами. Це веде до того, що традиційні методи, що використовують квадратичні метрики не забезпечують бажані результати. Метою статті є розробка достовірного робастного методу нечіткої кластеризації онлайн, який поєднує в собі переваги теорії довіри та робастних підходів у задачах нечіткої кластеризації.

Метод. Процедура нечіткої кластеризації даних з використанням достовірного підходу, заснованого на використанні як робастних цільових функцій спеціального типу, нечутливих до викидів, так і призначених для роботи як у пакетному режимі, так і в його повторюваній онлайн-версії, призначеній для вирішення проблем Data Stream Mining, коли дані надходять на обробку послідовно в режимі реального часу.

Результати. Аналізуючи загальну точність отриманих результатів методів і алгоритму кластеризації, запропонований метод подібний до результату достовірного методу нечіткої кластеризації, але має перевагу в часі незалежно від кількості спостережень, які були використані в процесі кластеризації.

Висновки. Розглянута задача нечіткої кластеризації потоків даних, забруднених аномальними викидами. Введено у розгляд рекурентний достовірний онлайн алгоритм, заснований на цільовій функції спеціального вигляду, що придушує ці викиди за допомогою використання функції гіперболічного тангенса, що крім нейронних мереж використовується у задачах робастного оцінювання. Запропонований алгоритм є достатньо простим у чисельній реалізації і є узагальненням деяких відомих онлайн процедур нечіткої кластеризації призначених для вирішення задач Data Stream Mining.

КЛЮЧОВІ СЛОВА: нечітка кластеризація, викривлені дані, достовірна нечітка кластеризація, Data Stream Mining, робастна функція.

ЛІТЕРАТУРА

1. Gan G. Data Clustering: Theory, Algorithms and Applications/ G. Gan, Ch. Ma, J. Wu. – Philadelphia, Pennsylvania: SIAM, 2007. – 455 p. DOI: https://doi.org/10.1137/1.9780898718348
2. Abony J. Cluster Analysis for Data Mining and System Identification / J. Abony, D. Feil. – Basel : Birkhouser, 2007. – 303 p.
3. Xu R. Clustering/ R. Xu, D. C. Wunsch. – Hoboken N. J. : John Wiley & Sons, Inc., 2009. – 398 p.
4. Bezdek J. C. Pattern recognition with fuzzy objective function algorithms / J. C. Bezdek – New York : Springer, 1981. – 253 p. DOI https://doi.org/10.1007/978-1-4757-0450-1.
5. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition/ [F. Höppner, F. Klawonn, R. Kruse, T. Runkler]. – Chichester : John Wiley &Sons, 1999. – 300 p.
6. 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.

DOI 10.15588/1607-3274-2023-3-10

– Vol. 23, № 4. – P. 545–564. DOI: https://doi.org/10.1142/S0218488515500245
7. 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.
8. Sequential fuzzy cluster extraction and its robust against noise / [K. Tsuda, S. Senda, M. Minoh, K. Ikeda] // System and Computers in Japan. – 1997. – 28. – P. 10–17.
9. Höppner, F. Fuzzy clustering of sampled functions / F. Höppner, F. Klawonn // 19th Int. Conf. North American Fuzzy Information Processing Society (NAFIPS) – Atlanta, USA, 2000. – P. 257–255.
10. Georgieva O. A clustering algorithm for identification of single clusters in large data sets / O. Georgieva, F. Klawonn // Proc. 11th East – West Fuzzy Coll. – Zittau/Görlitz : FH, 2004. – P. 118–125.
11. Kohonen T. Self-Organizing Maps/ T. Kohonen. – Berlin: Springer, 1995. – 362 p. doi: 10.1007/978-3-642-56927-2.
12. Park D. C. Gradient based fuzzy c-means (GBFCM) algorithm / D. C. Park, I. Dagger // IEEE International Conference on Neural Networks, 28 June – 2July,1984: proceedings. – Orlando : IEEE, 1984. – P. 1626–1631. DOI: 10.1109 / ICNN. 1994.374399.
13. Bodyanskiy Ye. Computational intelligence techniques for data analysis / Ye. Bodyanskiy // Lecture Notes in Informatics. – Bonn : Gesellschaft für Informatik, 2005. – P. 15–36.
14. 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.
15. Online Credibilistic Fuzzy Clustering Method Based on Cauchy Density Distribution Function / [A. Shafronenko, Ye. Bodyanskiy, I. Pliss, I. Klymova] // 2021 11th International Conference on Advanced Computer Information Technologies (ACIT): proceedings. – Deggendorf, Germany: IEEE, 2021. – P.704–707. DOI: 10.1109/ACIT52158.2021.9548572
16. Robust recursive fuzzy clustering algorithms / Ye. Bodyanskiy, Ye. Gorshkov, I. Kokshenev, V. Kolodyazhniy // Proc. 12th East West Fuzzy Coll 2005 – Zittau-Grölitz, FH, 2005. – P. 301–308.
17. Outlier resistant recursive fuzzy clustering algorithms / [Ye. Bodyanskiy, Ye Gorshkov, I. Kokshenev, V. Kolodyazhniy] // Ed. By B. Reusch «Computational Intelligence Theory and Applications» – Advances in Soft Computing – Vol. 38. – Berlin Heidelberg, Springer Verlag, 2006. – P. 647–652.
18. Arrow K. J. Studies in Linear and Nonlinear Programming / K. J. Arrow, L. Hurwitz, H. Uzawa. – Stanford University Press, 1958. – 242 p.
19. Bodyanskiy Ye. Recurcive fuzzy clustering algorithm/ Ye. Bodyanskiy, V. Kolodyazhniy, A. Stephan // Proc 10th East West Fuzzy Coll. 2002, – Zittau-Görlitz, HS, 2002. – P. 276–283.

DOI 10.15588/1607-3274-2023-3-10