e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2021. № 1
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2021. № 1

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

UDC 004.8:004.032.26

ONLINE FUZZY CLUSTERING OF INCOMPLETE DATA USING CREDIBILISTIC APPROACH AND SIMILARITY MEASURE OF SPECIAL TYPE

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

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

Klymova I. N. – Assistant at the Department of System Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. In most clustering (classification without a teacher) tasks associated with real data processing, the initial information is usually distorted by abnormal outliers (noise) and gaps. 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 propose the procedure of fuzzy clustering of incomplete data using credibilistic approach and similarity measure of special type.

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

Method. The procedure of fuzzy clustering of incomplete data using credibilistic approach and similarity measure of special type based on the use of both robust goal functions of a special type and similarity measures, 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. The introduced methods are simple in numerical implementation and are free from the drawbacks inherent in traditional methods of probabilistic and possibilistic fuzzy clustering data distorted by abnormal outliers (noise) and gaps.

Conclusions. The conducted experiments have confirmed the effectiveness of proposed methods of credibilistic fuzzy clustering of distorted data operability and allow recommending it for use in practice for solving the problems of automatic clusterization of distorted data. The proposed method is intended for use in hybrid systems of computational intelligence and, above all, in the problems of learning artificial neural networks, neuro-fuzzy systems, as well as in the problems of clustering and classification.

KEYWORDS: fuzzy clustering, distorted data, credibilistic fuzzy clustering, similarity measure.

ABBREVIATIONS

NN is a neural network;
NFN is a neuro-fuzzy network;
FCM is a fuzzy c-means method;
CFC is a credibilistic fuzzy clustering.

NOMENCLATURE

X is a data set matrix;
$\tilde{X}$ is a distorted data set matrix;
X_F is a data set matrix that contain all components;
X_G is a data set matrix that contain components of observation vectors that are absent in $\tilde{X}$;
X_P is a data set matrix that contain the value of the components of the vectors-observation available in $\tilde{X}$;
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;
$\tilde{x}_i(k)$ is a value of the vectors-observation;
l, q is a number of cluster;
$U_q(k)$ is a membership level;
Cl is a cluster;
D is a Euclidean distance;
D_P is a partial distance;

DOI 10.15588/1607-3274-2021-1-10

E is a goal function;
w is a centriod of cluster;
$\eta(k)$ is learning step parameter;
$Cr_q(k)$ is fuzzy credibilistic membership level;
σ is a Cauchy distribution;
$\delta_i(k)$ is a Lagrange indefinite multipliers;
β is a fuzzyfier.

INTRODUCTION

The problem of clustering (classification without a teacher) is an integral part of the general problem of Data Mining [1], for the solution of which many approaches, methods, and algorithms have been developed. Within the framework of this task, a special place is occupied by the problem of fuzzy clustering [2–4] which considers the situation when the classes being formed overlap, i.e. each observation can simultaneously belong to several or all classes. Within the framework of this subtask, two main approaches have been formed today: probabilistic [2], when the probability of its belonging to each of the possible classes is estimated for each observation, and the possibilistic [5], where the possibility (not probabilisty) of belonging to some of the classes is estimated. Both of these approaches are associated with solving the optimization problem (nonlinear programming) of the adopted goal functions and, in the general case, can lead to different final results. Despite the rather serious mathematical basis of these approaches, they suffer from a number of significant drawbacks: so the probabilistic approach is very sensitive to "abnormal" observations, which are practically "blurred" with the same levels of membership in all clusters.

The possibilistic approach, in turn, is associated with the so-called coincidence problem, when some clusters merge together, which generally does not allow splitting the processed sample into homogeneous groups – clusters.

Both of these approaches process data in batch mode, i.e. it is assumed that the entire array of observations is given a priori and does not change during the analysis. If the data are fed online (Data Stream Mining task), the classical probabilistic and possibilstic algorithms of fuzzy clustering become unworkable. In this situation, the fore sequential algorithms based on gradient optimization of goal functions taken. Such online procedures have been developed both within the framework of probabilistic [6–9] and possibilistic [8, 10] approaches and have proven their efficiency.

In clustering problems related to the processing of real data, the initial information, as a rule, is distorted by abnormal outliers (noises) and gaps, and the number of these outliers and "holes" can be commensurate with the volume of "clean" data, while a situation is possible when all data are "dirty". It is clear that "classical" methods (both batch and online) are ineffective in this situation.

To combat anomalous outliers in fuzzy clustering problems, robust methods were proposed based on the use of both robust goal functions of a special type and similarity measures insensitive to outliers and designed to operate both in batch [11, 12] and online [8, 13] modes.

As for the presence of the gaps in observations, there was also developed a number of techniques (through probabilistic and possibilistic approaches) as a batch [14, 15], and online [16]. And finally, in [17], a robust credibilisic procedure for fuzzy clustering of data distorted by both outliers and gaps based on a similarity measure of a special type was introduced.

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

The subject of study is procedure for fuzzy clustering of data distorted by both outliers and gaps based on a similarity measure of a special type.

The purpose of the work is to introduce robust credibilisic procedure for fuzzy clustering of distorted data.

1 PROBLEM STATEMENT

The initial information for solving the problem of fuzzy clustering using any of the known approaches is a sample of observations (batch) formed by N $(n\times 1)$ vectors-observations $X=\{x(1),x(2),...,x(k),...,x(N)\}\subset R^n$ where $x(k)=(x_1(k),...,x_i(k),...,x_n(k))^T\in R^n$ which in the self-learning mode has to be divided into mutually overlapping classes-clusters, while in the process of solving the problem for each observation $x(k)$ the sample should be determined by its fuzzy membership level $U_q(k)$ to each of the possible clusters $Cl_q(1\le q\le m)$. It is also usually assumed that the original data are preprocessed (normalized, centered) so that $-1\le x_i(k)\le 1$ or $\|x(k)\|=1$.

2 REVIEW OF THE LITERATURE

Alternatively to probabilistic and possibilistic procedures [18, 19] it was introduced credibilistic fuzzy clustering approach using as its basis the credibility theory [20], and is largely devoid of the drawbacks of known methods.

The most common approach within the framework of probabilistic fuzzy clustering is associated with minimizing the goal function [3]:

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

with constraints

$$\begin{cases}\sum\limits_{q=1}^{m}U_q(k)=1,\\ 0<\sum\limits_{k=1}^{N}U_q(k)<N.\end{cases} \qquad (2)$$

DOI 10.15588/1607-3274-2021-1-10

Solution of nonlinear programming problem using the method of Lagrange indefinite multipliers leads to the well-known result

$$\begin{cases} U_q(k)=\dfrac{\left(D^2\left(x(k),w_q\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(D^2\left(x(k),w_l\right)\right)^{\frac{1}{1-\beta}}}, \\ w_q=\dfrac{\sum\limits_{k=1}^{N}\left(U_q(k)\right)^{\beta}x(k)}{\sum\limits_{k=1}^{N}\left(U_q(k)\right)^{\beta}} \end{cases} \tag{3}$$

coinciding with $\beta=2$ with a popular method of Fuzzy C-Means of J. Bezdek (FCM) [2].

If the data are fed to processing sequentially, the solution of the nonlinear programming problem (1), (2) using the Arrow-Hurwitz-Uzawa algorithm leads to an online procedure [8]:

$$\begin{cases} U_q(k+1)=\dfrac{\left(D^2\left(x(k+1),w_q(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(D^2\left(x(k+1),w_l(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ w_q(k+1)=w(k)+\eta(k+1)U_q^{\beta}(k+1)* \\ *\left(x(k+1)-w_q(k)\right). \end{cases} \tag{4}$$

The goal function of credibilistic fuzzy clustering has the form [18, 19] close to (1)

$$E\left(Cr_q(k),w_q\right)=\sum_{k=1}^{N}\sum_{q=1}^{m}Cr_q^{\beta}(k)D^2\left(x(k),w_q\right) \tag{5}$$

with "softer" than (2) constraints:

$$\begin{cases} 0\le Cr_q(k)\le 1, \text{ for all } q \text{ and } k, \\ \sup Cr_q(k)\ge 0.5, \text{ for all } k, \\ Cr_q(k)+\sup Cr_l(k)=1, \\ \text{for any } q \text{ and } k, \text{for which } Cr_q(k)\ge 0.5. \end{cases} \tag{6}$$

It should be noted that the goal functions (1) and (5) are similar and that there are no rigid probabilistic constraints in (6) on the sum of the membership in (2).

In the procedures of credibilistic clustering, there is also the concept of fuzzy membership, which is calculated using the neighborhood function of the form [21]

$$U_q(k)=\varphi_q\left(D\left(x(k),w_q\right)\right) \tag{7}$$

monotonically decreasing on the interval $[0,\infty]$ so that $\varphi_q(0)=1, \varphi_q(\infty)\to 0$.

Such a function is essentially an empirical similarity measure of [22] related to distance by the relation

$$U_q(k)=\frac{1}{1+D^2\left(x(k),w_q\right)}. \tag{8}$$

Note also that earlier it was shown in [16] that the first relation (3) for $\beta=2$ can be rewritten as

$$U_q(k)=\left(1+\frac{D^2\left(x(k),w_q\right)}{\sigma_q^2}\right)^{-1} \tag{9}$$

where

$$\sigma_q^2=\left(\sum_{\substack{l=1\\ l\ne\varepsilon}}^{m}D^2\left(x(k),w_l\right)\right)^{-1} \tag{10}$$

which is a generalization of the function (8) (for $\sigma_q^2=1$ (8) coincides with (10)) and satifies all the conditions for (7).

In batch form the algorithm of credibilistic fuzzy clustering in the accepted notation can be written as [18, 19]

$$\begin{cases} U_q(k)=\left(1+D^2\left(x(k),w_q\right)\right)^{-1}, \\ U_q^*(k)=U_q(k)\left(\sup U_l(k)\right)^{-1}, \\ Cr_q(k)=\dfrac{1}{2}\left(U_q^*(k)+1-\sup\limits_{l\ne q}U_l^*(k)\right), \\ w_q=\dfrac{\sum\limits_{k=1}^{N}\left(Cr_q(k)\right)^{\beta}x(k)}{\sum\limits_{k=1}^{N}\left(Cr_q(k)\right)^{\beta}} \end{cases} \tag{11}$$

and in the online mode, taking into account (9), (10) [23]:

$$\begin{cases} \sigma_q^2(k+1)=\dfrac{1}{\sum\limits_{\substack{l=1\\ l\ne q}}^{m}D^2\left(x(k+1),w_l(k)\right)}, \\ U_q(k+1)=\left(1+\dfrac{D^2\left(x(k+1),w_q(k)\right)}{\sigma_q^2(k+1)}\right)^{-1}, \\ U_q^*(k+1)=\dfrac{U_q(k+1)}{\sup U_l(k+1)}, \\ Cr_q(k+1)=\dfrac{1}{2}\left(U_q^*(k+1)+1-\sup\limits_{l\ne q}U_l^*(k)\right), \\ w_q(k+1)=w_q(k)+\eta(k+1)Cr_q^{\beta}(k+1)\left(x(k+1)-w_q(k)\right). \end{cases} \tag{12}$$

From the point of view of computational implementation, algorithm (12) is not more complicated than proce-

DOI 10.15588/1607-3274-2021-1-10

dure (4) and, in the general case, is its generalization to the case of credibilistic approach to fuzzy clustering.

3 MATERIALS AND METHODS

In situations when an array of initial data $\tilde{X} = \{\tilde{x}(1), \tilde{x}(2)..., \tilde{x}(k),..., \tilde{x}(N)\}$ contains gaps (missing observations), the approach considered above cannot be used and requires significant modification. Thus, in [14], a modification of the FCM procedure based on the partial distance strategy was proposed. Within the framework of this strategy, three subarrays of data are introduced into consideration:

$X_F = \{\tilde{x}(k) \in \tilde{X}$

where vector $\tilde{x}(k)$ containg all components$\}$,

$X_P = \{\tilde{x}_i(k), 1 \le i \le n, 1 \le k \le N$

where $\tilde{x}_i(k)$ – the value of the components of the vectors-observation available in $\tilde{X}\}$,

$X_G = \{\tilde{x}_i(k) = none, 1 \le i \le n, 1 \le k \le N$

where $\tilde{x}_i(k)$ – components of observation vectors that are absent in $\tilde{X}\}$.

Further, the partial distance is introduced into consideration in the form

$$D_P^2\left(\tilde{x}(k), w_q\right) = \frac{n}{\delta_\Sigma(k)} \sum_{i=1}^{n} \left(\tilde{x}_i(k) - w_{qi}\right)^2 \delta_i(k) \qquad (13)$$

and instead of (1) – the goal function

$$\tilde{E}\left(U_q(k), w_q\right) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^\beta(k) \frac{n}{\delta_\Sigma(k)} * \\ * \sum_{i=1}^{n} \left(\tilde{x}_i(k) - w_{qi}\right)^2 \delta_i(k) \qquad (14)$$

where

$$\delta_i(k) = \begin{cases} 0 \ if \ \tilde{x}_i(k) \in X_G, \\ 1 \ if \ \tilde{x}_i(k) \in X_F, \end{cases}$$

$$\delta_\Sigma(k) = \sum_{i=1}^{n} \delta_i(k).$$

Using the method of Lagrange indefinite multipliers, we obtain [14]:

$$\begin{cases} U_q(k) = \dfrac{\left(D_P^2\left(\tilde{x}(k), w_q\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(D_P^2\left(\tilde{x}(k), w_l\right)\right)^{\frac{1}{1-\beta}}}, \\ w_{qi} = \dfrac{\sum\limits_{k=1}^{N}\left(U_q(k)\right)^\beta \delta_i(k)\tilde{x}_i(k)}{\sum\limits_{k=1}^{N}\left(U_q(k)\right)^\beta \delta_i(k)}. \end{cases} \qquad (15)$$

In recurrent online form (15) can be rewritten as [24, 25]

$$\begin{cases} U_q(k+1) = \dfrac{\left(D_P^2\left(\tilde{x}(k+1), w_q(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(D_P^2\left(\tilde{x}(k+1), w_l(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ w_{qi}(k+1) = w_{qi}(k) + \eta(k+1) U_q^\beta(k+1) * \\ *\left(\tilde{x}_i(k+1) - w_{qi}(k)\right)\delta_i(k). \end{cases} \qquad (16)$$

Similarly, using the partial distance strategy, a batch procedure of credibilistic fuzzy clustering can be introduced

$$\begin{cases} U_q(k) = \left(1 + D_P^2\left(\tilde{x}(k), w_q\right)\right)^{-1}, \\ U_q^*(k) = U_q(k)\left(\sup U_l(k)\right)^{-1}, \\ Cr_q(k) = \dfrac{1}{2}\left(U_q^*(k) + 1 - \sup\limits_{l \ne q} U_l^*(k)\right), \\ w_{qi} = \dfrac{\sum\limits_{k=1}^{N}\left(Cr_q(k)\right)^\beta \delta_i(k)\tilde{x}_i(k)}{\sum\limits_{k=1}^{N}\left(Cr_q(k)\right)^\beta \delta_i(k)} \end{cases} \qquad (17)$$

and its online version:

$$\begin{cases} \sigma_q^2(k+1) = \dfrac{1}{\sum\limits_{\substack{l=1 \\ l \ne q}}^{m} D_P^2\left(\tilde{x}(k+1), w_l(k)\right)}, \\ U_q(k+1) = \left(1 + \dfrac{D_P^2\left(\tilde{x}(k+1), w_q(k)\right)}{\sigma_q^2(k+1)}\right)^{-1}, \\ U_q^*(k+1) = \dfrac{U_q(k+1)}{\sup U_l(k+1)}, \\ Cr_q(k+1) = \dfrac{1}{2}\left(U_q^*(k+1) + 1 - \sup\limits_{l \ne q} U_l^*(k+1)\right), \\ w_{qi}(k+1) = w_{qi}(k) + \eta(k+1) Cr_q^\beta(k+1) * \\ *\left(\tilde{x}_i(k+1) - w_{qi}(k)\right)\delta_i(k). \end{cases} \qquad (18)$$

It is easy to see that algorithm (18) is a generalization of procedure (12) for the case of processing data not distorted by gaps.

4 EXPERIMENTS

To test the developed methods, as well as the analysis of translation over other more well-known approaches, the research was conducted using well-known test data sets of the UCI repository, such as Wine, Gas, Glass and Iris. Description of these data sets shown in Table 1.

Each of the data sets has its own of Attributes Number, Data Number, Cluster Number and Data Sourse.

DOI 10.15588/1607-3274-2021-1-10

Table 1 – Data set description Data set, Data number, Attributes number, Cluster number, Data source

Data set	Data Number	Attributes Number	Cluster Number	Data Source
Wine	178	13	3	Forina et al.(1988)
Gas	296	2	6	Box and Jenkins (1970)
Glass	214	9	6	Maskey and Glass (1977)
Iris	150	4	3	Fisher (1936)

To assess the quality of data clustering, we used Silhouette index, Calinski-Harabasz index and Davis-Baldwin index. The results of clustering Iris data set demonstrated Table 3.

5 RESULTS

Of course, the quality of proposed method should be estimated.

For this reason, we used the overall accuracy comparison of 100 experiments for different datasets and two clustering algorithms: fuzzy c-means method (FCM) and credibilistic fuzzy clustering (CFC).

Table 2 – A comparison of 100 experiments for the other data set

Data set	Clustering algorithm	Overall accuracy		
		Highest	Mean	Variance
Wine	FCM	68.54	68.54	0
	CFC	67.98	67.98	0
Glass	FCM	49.53	49.08	0.01
	CFC	44.86	44.86	0
Gas	FCM	79.05	77.33	11.33
	CFC	68.58	68.55	0.01
Iris	FCM	89.33	89.33	0
	CFC	91.33	90.06	0.04

Credibilistic fuzzy clustering algorithm works not only with complete data, but also with data that containing missing values. To conduct experimental studies, we artificially have introduced 10 missing values into the Iris data set. Figure 1 demonstrates credibilistic fuzzy clustering (CFC) Iris data set with 10 missing values.

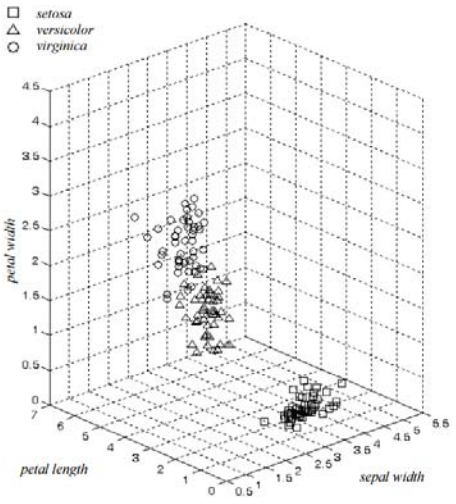

Figure 1 – Credibilistic fuzzy clustering Iris data set with 10 missing values

Table 3 – Result of clustering Iris data set with different algorithms

Clustering algorithm	Silhouette index	Calinski-Harabasz index	Davis-Baldwin index
Adaptive probabilistic fuzzy clustering data with missing values	0.2326	921.58	1.28
Adaptive possibilistic fuzzy clustering missing values	0.2325	922.01	1.25
Adaptive credibilistic fuzzy clustering missing values	0.3335	965.42	1.05
FCM	0.2354	986.39	1.23
K-means	0.3676	1419.28	1.09

6 DISCUSSION

The result of clustering data sets shown in Table 2. As the table shows, the prepositional credibilistic fuzzy clustering algorithm shows good results.

Comparative data analysis was performed with previously proposed clustering methods data that containing missing values such as adaptive probabilistic fuzzy clustering data with missing values, adaptive possibilistic fuzzy clustering missing values and classical algorithms FCM and K-means.

Thus, the silhouette index shows how the average distance to the objects of cluster differs from the average distance to the objects of other clusters. This value is in the range [–1, 1]. Values close to –1 correspond to "bad" (disparate) types of clustering. Values close to zero indicate that the clusters intersect and overlap. Values close to 1 correspond to "dense" clearly selected clusters. Thus, the larger the silhouette, the clearer the clusters and they are compact, densely grouped clouds of points. As can see from the silhouette index, the data recovery method works quite well. The higher the value of the Calinski-Harabasz index, the better is the solution. In the Davis-Baldwin index, values close to zero indicate the best section, i.e. as can see, with almost all missing data, the distribution is "good", so the method worked well.

CONCLUSIONS

The conducted experiments have confirmed the effectiveness proposed methods of credibilistic fuzzy clustering of distorted data operability and allow recommending it for use in practice for solving the problems of automatic clusterization of distorted data The proposed method is intended for use in hybrid systems of computational intelligence and, above all, in the problems of learning artificial neural networks, neuro-fuzzy systems, as well as in the problems of clustering and classification.

The scientific novelty of obtained results is that the method of credibilistic fuzzy clustering of distorted data based on the partial distance strategy, 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

DOI 10.15588/1607-3274-2021-1-10

practice for solving the problems of automatic clusterization of distorted data.

Prospects for further research methods of credibilistic fuzzy clustering of distorted datafor a broad class of practical problems.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of Kharkiv National University of Radio Electronics "Deep hybrid systems of computational intelligence for data stream mining and their fast learning" (state registration number 0119U001403).

REFERENCES

1. Aggarwal C. C. Data Mining. Switzerland : Springer, 2015, 727 p. DOI https://doi.org/ 10.1007 / 978-3-319-14142-8.
2. 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.
3. 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.
4. Xu R., D. C. Wunsch Clustering. Hoboken N. J., John Wiley & Sons, Inc., 2009, 398 p.
5. Krishnapuram R., Keller J. M. A Possibilistic Approach to Clustering, *IEEE Transactions on Fuzzy Systems, May 1993: proceedings*, IEEE, 1993, Vol. 1, pp. 98–110. DOI: 10.1109/91.227387.
6. Park D. C., Dagher 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.
7. Chung, F. L., Lee T. Fuzzy competitive learning, *Neural Networks*, 1994, Vol. 7, № 3, pp. 539–552. DOI: https://doi.org/10.1016/0893-6080(94)90111-2.
8. Bodyanskiy Ye. Computational intelligence techniques for data analisys, *Lecture Notes in Informatics*. Bonn, Gesellschaft für Informatik, 2005, pp. 15–36.
9. Hu Zh., Bodyanskiy Ye. V., Tyshchenko O. K. A deep cascade neuro-fuzzy system for high-dimensional online fuzzy clustering, *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 23–27 August, 2016: proceedings*. Lviv, IEEE, 2016, pp. 318–322. DOI: 10.1109/DSMP.2016.7583567.
10. Hu. Zh., Bodyanskiy Ye. V., Tyshchenko O. K. A cascade deep neuro-fuzzy system for high-dimensional online possibilistic fuzzy clustering, *2016 XI-th International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), Lviv, 6–10 September, 2016: proceedings.* Lviv, IEEE, 2016, pp. 119–122. DOI: 10.1109/STC-CSIT.2016.7589884.
11. Chintalapudi K. K., Kam M. A noise resistant fuzzy c-means algorithm for clustering, *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36228) 4–9 May 1998: proceedings.* Anchorage, IEEE, 1998, Vol. 2, pp. 1458–1463. DOI: 10.1109/FUZZY.1998.686334
12. Hathaway R. J., Bezdek J. C., Hu Y. Generalized fuzzy c-means clustering strategies using L/sub p/ norm distances, *IEEE Transactions on Fuzzy Systems*, IEEE, 2000, Vol. 8 (5), pp. 576–582. DOI: 10.1109/91.873580.
13. Marwala T. Computational Intelligence for Missing Data Imputation Estimation and Management: Knowledge Optimization Techniques. Hershey-New York, Information Science Reference, 2009, 326 p.
14. Hu Zh., Bodyanskiy Ye., Tyshchenko O., Shafronenko A. Fuzzy clustering of incomplete data by means of similarity measures/ Hu Zh., // *2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2–6 July 2019 Lviv 2019: proceedings*, IEEE, 2019. – Track 6. –Lviv, Ukraine, 2019. – P.149–152. DOI: 10.1109 /UKRCON. 2019.8879844
15. Bodyanskiy Ye., Shafronenko A., Mashtalir S. Online robust fuzzy clustering of data with omissions using similarity measure of special type, *Lecture Notes in Computational Intelligence and Decision. Waking-Cham, Springer*, 2020, Vol. 1020, pp. 637–646. DOI: https://doi.org/10.1007/978-3-030-26474-1_44.
16. 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
17. 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.
18. Liu B. A survey of credibility theory, *Fuzzy Optimization and Decision Making*, 2006, No. 4, pp. 387–408. DOI: https://doi.org/10.1007/s10700-006-0016-x.
19. Zhao F., Jiao L., Liu H. Fuzzy c-means clustering with nonlocals partial information for noisy image segmentation, *Frontiers of Computer Science*. China, 2011, Vol. 5(1), pp. 45–56. DOI: https://doi.org/10.1007/s11704-010-0393-8.
20. Yang Y. K., Shieh H. L., Lee C. N. Constructing a fuzzy clustering model based on its data distribution, *International Conference on Computational Intelligence for Modeling. Control and Automation (CIMCA 2004), Gold Coast 2004: proceedings*. Gold Coast, Australia, 2004.
21. Bodyanskiy Ye. V., Shafronenko A. Yu., Rudenko D. O., Klymova I. M. Online recurrent method of credibilistic fuzzy clustering, *Topical issues of the development of modern science. 5th International scientific and practical conference, Sofia 2020, proceedings. Sofia, Publishing House "ACCENT"*, 2020, pp. 37–40.
22. Bodyanskiy Ye., Shafronenko A., Volkova V. Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network, *Artificial Intelligence Methods and Techniques for Business and Engineering Applications, Rzeszow-Sofia*, ITHEA, 2012, pp. 287–296.
23. Bodyanskiy Ye., Shafronenko A. Online algorithm for possibilitic fuzzy clustering based on evolutionary cat swarm optimization, *Science and Education a New Dimension. Natural and Technical Sciences*, 2019, Vol. 193, pp. 86–88. DOI: 10.31174/SEND-NT2019-193VII23-22
24. 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.

Received 30.11.2020.
Accepted 11.01.2021.

DOI 10.15588/1607-3274-2021-1-10

УДК 004.8:004.032.26

ОНЛАЙН НЕЧІТКА КЛАСТЕРИЗАЦІЯ ДАНИХ З ПРОПУСКАМИ З ВИКОРИСТАННЯМ ДОСТОВІРНОГО ПІДХОДУ ТА МІРИ ПОДІБНОСТІ СПЕЦІАЛЬНОГО ВИГЛЯДУ

Бодянський Є. В. – д-р техн. наук, професор, професор кафедри штучного інтелекту Харківського національного університету радіоелектроніки, Харків, Україна.

Шафроненко А. Ю. – канд. техн. наук, доцент кафедри інформатики Харківського національного університету радіоелектроніки, Харків, Україна.

Клімова І. М. – асистент кафедри системної інжеренерії Харківського національного університету радіоелектроніки, Харків, Україна.

АНОТАЦІЯ

Актуальність. У більшості завдань кластеризації (класифікації без вчителя), пов'язаних з обробкою реальних даних, початкова інформація, у тому чи іншому випадку як правило, спотворюється через аномальні викиди (збурення) та пропуски. Зрозуміло, що «класичні» методи інтелектуального аналізу даних (як пакетні, так і онлайн) в цій ситуації неефективні. Метою роботи було запропонувати процедуру нечіткої кластеризації викривлених даних з використанням достовірного підходу та міри подібності спеціального типу, а також розробка метода достовірної нечіткої кластеризації спотворених даних із використанням теорії довіри, яка була би позбавлена недоліків імовірнісних і можливісних підходів кластеризації викривлених даних.

Метод. Процедура нечіткої кластеризації неповних даних із використанням достовірного підходу та міри схожості спеціального типу, заснована на використанні робастних цільових функцій спеціального типу, а також мір подібності, нечутливих до викидів та призначених для роботи як у пакетній, так і в онлайн версії для вирішення проблем Data Stream Mining, коли дані надходять на обробку послідовно в режимі реального часу.

Результати. Запропоновані методи є простими в чисельній реалізації та позбавлені недоліків, властивих традиційним методам імовірнісної та можливістної нечіткої кластеризації.

Висновки. Проведені експериментальні дослідження підтвердили результативність та якість роботи запропонованих методів достовірної нечіткої кластеризації спотворених даних і дозволяють рекомендувати їх для використання на практиці для вирішення проблем автоматичної кластеризації викривлених даних. Запропонований метод призначений для використання в гібридних системах обчислювального інтелекту і, перш за все, у проблемах навчання штучних нейронних мереж, нейро-фаззі систем, а також у завданнях кластеризації та класифікації.

КЛЮЧОВІ СЛОВА: нечітка кластеризація, викривлені дані, достовірна нечітка кластеризація, міра подібності.

УДК 004.8:004.032.26

ОНЛАЙН НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ ДАННЫХ С ПРОПУСКАМИ С ИСПОЛЬЗОВАНИЕМ ДОСТОВЕРНОГО ПОДХОДА И МЕРЫ СХОЖЕСТИ СПЕЦИАЛЬНОГО ВИДА

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

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

Климова И. Н. – асистент кафедры системной инжеренерии Харьковского национального университета радиоелектроники, Харьков, Украина.

АННОТАЦИЯ

Актуальность. В большинстве задач кластеризации (классификации без учителя), связанных с обработкой реальных данных, исходная информация в том или ином случае искажается аномальными выбросами (шумом) и пропусками. Понятно, что «классические» методы интеллектуального анализа данных (как пакетные, так и онлайн) в данной ситуации малоэффективны. Целью работы было предложить численно простую процедуру нечеткой кластеризации неполных данных с использованием достоверного подхода и меры подобия специального типа, которая была бы лишена недостатков вероятностных и возможностных подходов кластеризации искаженных данных.

Метод. Разработана процедура нечеткой кластеризации неполных данных с использованием достоверного подхода и меры подобия специального типа, основанная на использовании как робастных целевых функций специального типа, так и мер подобия, которые нечувствительны к выбросам и рассчитаны на работу как в пакетной, так и в онлайн - версии для решения проблем Data Stream Mining, когда данные поступают на обработку последовательно друг за другом, в режиме реального времени.

Результаты. Представленные методы просты в численной реализации и лишены недостатков, присущих традиционным методам вероятностной и возможностной нечеткой кластеризации данных.

Выводы. Проведенные экспериментальные исследования подтвердили работоспособность предложенных методов достоверной нечеткой кластеризации искаженных выбросами и пропусками данных и позволяют рекомендовать их к использованию на практике для решения задач автоматической кластеризации искаженных данных. Предлагаемый метод предназначен для использования в гибридных системах вычислительного интеллекта и, прежде всего, в задачах обучения искусственных нейронных сетей, нейро-фаззи системах, а также в задачах кластеризации и классификации.

DOI 10.15588/1607-3274-2021-1-10

КЛЮЧЕВЫЕ СЛОВА: нечеткая кластеризация, искаженные данные, достоверная нечеткая кластеризация, мера сходства.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Aggarwal C. C. Data Mining / C. C. Aggarwal. – Switzerland : Springer, 2015. – 727 p. DOI https://doi.org/ 10.1007 / 978-3-319-14142-8.
2. 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.
3. 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.
4. Xu R. Clustering. / R. Xu, D. C. Wunsch. – Hoboken N.J.: John Wiley & Sons, Inc., 2009. – 398 p.
5. Krishnapuram R. A Possibilistic Approach to Clustering / R. Krishnapuram, J. M. Keller// IEEE Transactions on Fuzzy Systems, May 1993: proceedings. – IEEE, 1993. – Vol. 1. – P. 98–110. DOI: 10.1109/91.227387.
6. Park D. C. Gradient based fuzzy c-means (GBFCM) algorithm / D. C. Park, I. Dagher // IEEE International Conference on Neural Networks, 28 June – 2 July, 1984 : proceedings. – Orlando : IEEE, 1984. – P. 1626–1631. DOI: 10.1109 / ICNN. 1994.374399.
7. Chung F. L. Fuzzy competitive learning / F. L. Chung, T. Lee // Neural Networks. – 1994. – Vol. 7, № 3. – P. 539–552. DOI: https://doi.org/10.1016/0893-6080(94)90111-2.
8. Bodyanskiy Ye. Computational intelligence techniques for data analisys / Ye. Bodyanskiy // Lecture Notes in Informatics.–Bonn: Gesellschaft für Informatik, 2005. – P. 15–36.
9. Hu Zh. A deep cascade neuro-fuzzy system for high-dimensional online fuzzy clustering/ Hu Zh., Ye. V. Bodyanskiy , O. K. Tyshchenko // 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 23–27 August, 2016: proceedings. – Lviv : IEEE, 2016. – P. 318–322. DOI: 10.1109/DSMP.2016.7583567.
10. Hu. Zh. A cascade deep neuro-fuzzy system for high-dimensional online possibilistic fuzzy clustering / Hu. Zh., Ye. V. Bodyanskiy, O. K. Tyshchenko // 2016 XI-th International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), Lviv, 6–10 September, 2016 : proceedings. – Lviv : IEEE, 2016. – P. 119–122. DOI: 10.1109/STC-CSIT.2016.7589884.
11. Chintalapudi K. K. A noise resistant fuzzy c-means algorithm for clustering. / K. K. Chintalapudi, M. Kam // 1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36228) 4–9 May 1998: proceedings. – Anchorage: IEEE, 1998. – Vol. 2. – P. 1458–1463. DOI: 10.1109/FUZZY.1998.686334
12. Hathaway R. J. Generalized fuzzy c-means clustering strategies using L/sub p/ norm distances / R. J. Hathaway, J. C. Bezdek, Y. Hu // IEEE Transactions on Fuzzy Systems. – IEEE, 2000. – Vol. 8 (5). – P. 576–582. DOI: 10.1109/91.873580.
13. Marwala T. Computational Intelligence for Missing Data Imputation Estimation and Management: Knowledge Optimization Techniques / T. Marwala. – Hershey-New York : Information Science Reference, 2009. – 326 p.
14. 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 (UKRCON), 2–6 July 2019 Lviv 2019: proceedings. – IEEE: 2019. – Track 6. – Lviv, Ukraine, 2019. – P. 149–152. DOI: 10.1109 /UKRCON. 2019.8879844
15. Bodyanskiy Ye. Online robust fuzzy clustering of data with omissions using similarity measure of special type. / Ye. Bodyanskiy, A. Shafronenko, S. Mashtalir // Lecture Notes in Computational Intelligence and Decision. – Waking-Cham : Springer, 2020. – Vol. 1020. – P. 637–646. DOI: https://doi.org/10.1007/978-3-030-26474-1_44.
16. 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
17. 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.
18. Liu B. A survey of credibility theory / B. Liu // Fuzzy Optimization and Decision Making. – 2006. – № 4– P. 387–408. DOI: https://doi.org/10.1007/s10700-006-0016-x.
19. Zhao F. Fuzzy c-means clustering with nonlocals partial information for noisy image segmentation / F. Zhao, L. Jiao, H. Liu // Frontiers of Computer Science. – China : 2011. – Vol. 5(1). – P. 45–56. DOI: https://doi.org/10.1007/s11704-010-0393-8.
20. Yang Y. K. Constructing a fuzzy clustering model based on its data distribution / Y. K. Yang, H. L. Shieh, C. N. Lee // International Conference on Computational Intelligence for Modeling. Control and Automation (CIMCA 2004), Gold Coast 2004 : proceedings. – Gold Coast, Australia, 2004.
21. Online recurrent method of credibilistic fuzzy clustering / [Ye. V. Bodyanskiy, A. Yu. Shafronenko, D. O. Rudenko, I. M. Klymova] // Topical issues of the development of modern science. 5th International scientific and practical conference, Sofia 2020: proceedings.– Sofia : Publishing House “ACCENT”, 2020. – P. 37–40.
22. Bodyanskiy Ye. Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network / Ye. Bodyanskiy, A. Shafronenko, V. Volkova // Artificial Intelligence Methods and Techniques for Business and Engineering Applications, Rzeszow-Sofia : ITHEA, 2012. – P. 287–296.
23. Bodyanskiy Ye. Online algorithm for possibilitic fuzzy clustering based on evolutionary cat swarm optimization / Ye. Bodyanskiy, A. Shafronenko // Science and Education a New Dimension. Natural and Technical Sciences. – 2019. – Vol. 193. – P. 86–88. DOI: 10.31174/SEND-NT2019-193VII23-22
24. 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.

DOI 10.15588/1607-3274-2021-1-10