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 clustering and classification.

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

ABBREVIATIONS

NN is a neural network;  
NFS 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;  
$	ilde{X}$ is a distorted data set matrix;  
$X_r$ is a data set matrix that contain all components;  
$X_o$ is a data set matrix that contain components of observation vectors that are absent in $	ilde{X}$;  
$X_p$ is a data set matrix that contain the value of the components of the vectors-observation available in $	ilde{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;  
$	ilde{x}(k)$ is a preprocessed original data;  
$ar{x}(k)$ is a value of the vectors-observation;  
$l, q$ is a number of cluster;  
$U_x(k)$ is a membership level;  
$Cl$ is a cluster;  
$D$ is a Euclidean distance;  
$D_p$ is a partial distance;
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 probability) 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 possibilistic algorithms of fuzzy clustering become unworkable. In this situation, the foreknowledge of both robust goal functions of a special type and similar-

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 credibilistic 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 \( X = \{x(1), x(2), \ldots, x(k), \ldots, x(N)\} \subset \mathbb{R}^n \) where \( x(k) = (x_1(k), x_2(k), \ldots, x_j(k), \ldots, x_j(k)) \in \mathbb{R}^n \) 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 \( C_q(1 \leq q \leq m) \). It is also usually assumed that the original data are preprocessed (normalized, centered) so that \(-1 \leq x_i(k) \leq 1\) 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]:

\[
E(U_q(k), w_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^\beta(k)D^2(x(k), w_q)
\]  

with constraints

\[
\sum_{q=1}^{m} U_q(k) = 1, \\
0 < \sum_{k=1}^{N} U_q(k) < N.
\]
Solution of nonlinear programming problem using the method of Lagrange indefinite multipliers leads to the well-known result

\[
\begin{align*}
U_q(k) &= \left(D^2 \left(x(k), w_q \right) \right)^{1-\beta}, \\
\sum_{i=1}^{m} \left(D^2 \left(x(k), w_i \right) \right)^{1-\beta}, \\
w_q &= \frac{\sum_{k=1}^{N} \left(U_q(k) \right)^{\beta} x(k)}{\sum_{k=1}^{N} \left(U_q(k) \right)^{\beta}}
\end{align*}
\]  

(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]:

\[
\begin{align*}
U_q(k+1) &= \left(D^2 \left(x(k+1), w_q \right) \right)^{1-\beta}, \\
\sum_{i=1}^{m} \left(D^2 \left(x(k+1), w_i \right) \right)^{1-\beta}, \\
w_q(k+1) &= w(k) + \eta(k)U_q^\beta(k+1)^* \\
 &= \left(x(k+1) - w_q(k) \right). 
\end{align*}
\]  

(4)

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

\[
E \left(C_q, w_q \right) = \sum_{k=1}^{N} \sum_{q=1}^{m} C_q^\beta(k)D^2 \left(x(k), w_q \right) 
\]  

(5)

with “softer” than (2) constraints:

\[
\begin{align*}
0 \leq C_q(k) &\leq 1, \text{ for all } q \text{ and } k, \\
\sup_q C_q(k) &\geq 0.5, \text{ for all } k, \\
\sup_q C_q(k) + \sup_q C_q(k) &= 1, \\
\text{for any } q \text{ and } k, \text{ for which } C_q(k) \geq 0.5.
\end{align*}
\]  

(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]

\[
U_q(k) = \varphi_q \left(D \left(x(k), w_q \right) \right) 
\]  

(7)

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

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

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

(8)

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

\[
U_q(k) = \frac{1 + D^2 \left(x(k), w_q \right)}{\sigma_q^2} 
\]  

(9)

where

\[
\sigma_q^2 = \frac{\sum_{l=1}^{m} \left(1 + D^2 \left(x(k), w_l \right) \right)}{m}, 
\]  

(10)

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

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

\[
\begin{align*}
U_q(k) &= \left(1 + D^2 \left(x(k), w_q \right) \right)^{-1}, \\
U_q^*(k) &= U_q(k)(\sup_{l \neq q} U_q^*(k))^{-1}, \\
C_r(k) &= \frac{1}{2} \left(U_q^*(k) + 1 - \sup_{l \neq q} U_q^*(k) \right), \\
w_q &= \frac{\sum_{k=1}^{N} \left(C_r(k) \right)^\beta x(k)}{\sum_{k=1}^{N} \left(C_r(k) \right)^\beta}.
\end{align*}
\]  

(11)

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

\[
\begin{align*}
\sigma_q^2(k+1) &= 1, \\
U_q(k+1) &= \frac{1}{\sum_{l \neq q} D^2 \left(x(k+1), w_l(k) \right)}, \\
U_q^*(k+1) &= \frac{1}{\sup_{l \neq q} U_q(k+1)}, \\
C_r(k+1) &= \frac{1}{2} \left(U_q^*(k+1) + 1 - \sup_{l \neq q} U_q^*(k) \right), \\
w_q(k+1) &= w_q(k) + \eta(k)C_r(k+1) \left(x(k+1) - w_q(k) \right).
\end{align*}
\]  

(12)

From the point of view of computational implementation, algorithm (12) is not more complicated than proce-
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 \( \bar{X} = \{\bar{x}(1), \bar{x}(2), ..., \bar{x}(k), ..., \bar{x}(N)\} \) 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:

\[ X_F = \{\bar{x}(k) \in \bar{X}\} \]

where vector \( \bar{x}(k) \) containing all components,

\[ X_P = \{\bar{x}_i(k), 1 \leq i \leq n, 1 \leq k \leq N\} \]

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

\[ X_C = \{\bar{x}_i(k) = none, 1 \leq i \leq n, 1 \leq k \leq N\} \]

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

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

\[ D_P^2(\bar{x}(k), w_q) = \sum_{i=1}^{n} \left( \delta_i(k) - w_{qi}(k) \right)^2 \delta_i(k) \]

and instead of (1) – the goal function

\[ \bar{E}(U_q(k), w_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^0(k) \frac{n}{\delta_2(k)} \]

\[ \sum_{i=1}^{m} \left( \bar{x}_i(k) - w_{qi}(k) \right)^2 \delta_i(k) \]

where

\[ \delta_i(k) = \begin{cases} 0 & \text{if } \bar{x}_i(k) \in X_C, \\ 1 & \text{if } \bar{x}_i(k) \in X_F, \\ \sum_{i=1}^{n} \delta_i(k). 
\]

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

\[ U_q(k + 1) = \frac{\left( D_P^2(\bar{x}(k + 1), w_q(k)) \right)^{1-\beta}}{\sum_{i=1}^{m} \left( D_P^2(\bar{x}(k + 1), w_q(k)) \right)^{1-\beta}} \]

\[ \sum_{i=1}^{m} \frac{U_q^0(k) \beta \delta_i(k) \bar{x}_i(k)}{\sum_{k=1}^{N} \left( U_q^0(k) \right) \beta \delta_i(k)}. \]

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

\[ U_q(k + 1) = \left( 1 + D_P^2(\bar{x}(k + 1), w_q(k)) \right)^{-\beta} \]

\[ \sum_{i=1}^{m} \frac{D_P^2(\bar{x}(k + 1), w_q(k)) \beta \delta_i(k) \bar{x}_i(k)}{\sum_{k=1}^{N} \left( D_P^2(\bar{x}(k + 1), w_q(k)) \right) \beta \delta_i(k)}. \]

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

\[ U_q^*(k + 1) = \left( 1 + D_P^2(\bar{x}(k + 1), w_q(k)) \right)^{-\beta} \]

\[ \sum_{i=1}^{m} \frac{D_P^2(\bar{x}(k + 1), w_q(k)) \beta \delta_i(k) \bar{x}_i(k)}{\sum_{k=1}^{N} \left( D_P^2(\bar{x}(k + 1), w_q(k)) \right) \beta \delta_i(k)}. \]

and its online version:

\[ \sigma_q^2(k + 1) = \frac{1}{\sum_{i=1}^{m} D_P^2(\bar{x}(k + 1), w_q(k))} \]

\[ U_q^*(k + 1) = \frac{1}{\sum_{i=1}^{m} D_P^2(\bar{x}(k + 1), w_q(k))} \]

\[ \sum_{i=1}^{m} \frac{U_q^0(k) \beta \delta_i(k) \bar{x}_i(k)}{\sum_{k=1}^{N} \left( U_q^0(k) \beta \delta_i(k) \bar{x}_i(k) \right)}. \]

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 Source.
Table 1 – Data set description

| 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        |
| Glass    | CFC                  | 67.98   | 67.98 | 0.01     |
| Gas      | CFC                  | 49.08   | 49.08 | 0        |
| Iris     | CFC                  | 89.33   | 89.33 | 0        |

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](image)

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

| Clustering algorithm | Silhouette index | Calinski-Harabasz index | Davis-Baldwin index |
|----------------------|------------------|-------------------------|---------------------|
| 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
practice for solving the problems of automatic clusterization of distorted data.

**Prospects for further research** methods of credibilistic fuzzy clustering of distorted data for 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 Analysis 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 analysis, Lecture Notes in Informatics. Bonn, Gesellschaft für Informatik, 2005, pp. 15–36.
9. Hu Zh., Bodyanskiy Ye. V., Tsyshchenko 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., Tsyshchenko 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, 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. V., 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., Mashitali 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 nonlocal 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 possibilistic 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-NT2019-193VI123-22.
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.
ОЛІЙНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ ДАНИХ З ПРОПУСКАМИ З ВИКОРИСТАННЯМ ДОСТОВІРНОГО ПОДХОДУ ТА МЕРИ 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 інженерії 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 числі іншим випадками, як правило, спотворюється через аномальні випадки (збірні) та пропуски. Зрозуміло, що «класичні» методи інтелектуального аналізу даних (як пакетні, так і онлайні) в цій ситуації неефективні. Метою роботи було запропонувати процедуру нечіткої кластеризації викривлених даних з використанням достовірного підходу та мери подібності спеціального типу, а також розробка метода достовірної нечіткої кластеризації викривлених даних з використанням теорії довіри, яка була би позбавлена недоліків імовірнісних і можливих підходів кластеризації викривлених даних.

Метод. Процедура нечіткої кластеризації неповних даних із використанням достовірного підходу та мери схожості спеціального типу, заснована на використанні робочих цільових функцій спеціального типу, а також мір подібності, нечутливих до викривлень та при значних для роботи як у пакетні, 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 числовій реалізації та позбавлені недоліків, властивих традиційним методам імовірнісної та можливісті кластеризації.

Висновки. Проведені експериментальні дослідження підтвердили результативність та якість роботи запропонованих методів достовірної нечіткої кластеризації викривлених даних і дозволяють рекомендувати їх для використання на практиці для вирішення проблем автоматичної кластеризації викривлених даних. Запропонований метод призначений для вирішення 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, викривлени дани, достовірна нечітка кластеризація, мера подібності.

УДК 004.8:004.032.26

ОЛІЙНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ ДАНИХ С ПРОПУСКАМИ С ІСПОЛЗОВУЮЧЕМ ДОСТОВЕРНОГО ПОДХОДА ТА МЕРИ СХОЖЕСТІ СПЕЦІАЛЬНОГО ВИДА

Бодянський Е. В. – д-р техн. наук, професор, професор кафедри искусственного интеллекта Харьковского национального университета радиоэлектроники, Харьков, Украина.

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

Клімова І. М. – асистент кафедры системной инженерии Харьковского национального университета радиоэлектроники, Харьков, Украина.

АНОТАЦІЯ

Актуальність. В більшості задач кластеризації (класифікації без вчителя), зв’язаних з обробкою реальних даних, інформація в них іноді використовується як аномальні викривлення (шумові) та пропусками. Понятно, що «класичні» методи інтелектуального аналізу даних (як пакетні, так і онлайні) в даній ситуації неефективні. Цілью роботи було прочитати численно просту процедуру нечіткої кластеризації неповних даних з управлінням достовірного підходу та мери подібності спеціального типу, які були лишены недостатків вероятностних, численними подібностей и полносомптоностями кластеризації викривлених даних.

Метод. Розроблена процедура нечіткої кластеризації неповних даних з управлінням достовірного підходу та мери подібності спеціального типу, основана на використанні таких робочих цільових функцій спеціального типу, які були лишены недостатків вероятностних, численними подібностей и полносомптоностями кластеризації викривлених даних.

Результати. Представленные методы просты в численной реализации и лишены недостатков, присущих традиционным методам вероятностной и возможной нечіткої кластеризації даних.

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

© Bodiansky Ye. V., Shafronenko A. Yu., Klymova I. N., 2021
DOI 10.15588/1607-3274-2021-1-10
ЛІТЕРАТУРА

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 analysis / J. C. Bezdek. – New York : Springer, 1981. – 253 p. DOI https://doi.org/10.1007/978-1-4757-6581-3.
3. Fuzzy Clustering Analysis: Methods for Classification, Data Analyses 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 Possibleiti­Stic 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. Bodianskiy Ye. Computational intelligence techniques for data analysis / Ye. Bodianskiy // 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. Bodianskiy, O. K. Tsylshchenko // 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 possible fuzzy clustering / Hu Zh., Ye. V. Bodianskiy, O. K. Tsylshchenko // 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: 10.1109/91.875580.
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. Bodianskiy, O. Tsylshchenko, 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. Mashalite // 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/s10800-006-0016-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. Bodianskiy, 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 possibilistic 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. Bodianskiy, 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.

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