Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach

Yevgeniy Bodyanskiy, Alina Shafronenko and Iryna Klymova

Kharkiv National University of Radio Electronics, Nauky ave 14, Kharkiv, 61166, Ukraine

Abstract

The problems of big data clustering is very interesting area of artificial intelligence nowadays. This task often occurs in many application, that related with data mining, deep learning, web mining etc. For solving these problems the traditional approaches and methods require that every vector – observation from processed data set is fed in batch form and does not change over the time and could belong more, than one cluster. In this situation more effective are fuzzy clustering methods that are synthesized under the assumptions of mutual overlapping of classes, based on credibility theory and adaptive goal function.

Therefore as alternative, to known clustering algorithms we propose adaptive recovery of distorted data algorithm based on credibilistic fuzzy clustering approach.

Keywords 1

Fuzzy clustering, credibilistic fuzzy clustering, adaptive goal function, distorted data, membership level, self-organizing neural network, machine learning

1. Introduction

To solve a wide class of Data Mining problems, machine learning technologies are effectively used, among which the most complex ones are the ones based on selflearning approach and are used in conditions of a priori information shortage. These technologies, combined with a neural network approach, provide effective solutions to many real-world problems. Real systems of image processing and computer vision, control of aerospace objects, technical and medical diagnostics, in economics and finance, in military application, motion control, energy, forensic science, signal analysis of various nature, etc. have been created, and this list is expanded almost daily.

At the moment there is a sufficient amount of information about the activities of enterprises, hospitals, firms, which reflects the activities of these objects. After analyzing this information, it is possible to find objective patterns, provided that the table generated reflects actual data reflecting cause-effect relationships.

This information has been collected over the years (for example, the rate of inflation, the income level of the population, the structure of household spending, the cost of housing and communal services, the state of industrial and agricultural production, the timeliness of wages and pensions, etc.). Of course, such data is difficult to analyze manually due to the large amount of information and complex non-linear cause-and-effect relationships. Therefore, it became necessary to develop new methods of analysis and forecasting, which include machine-based methods for the detection of patterns.

In many data mining tasks related to the processing of empirical quantitative observations, the data may be distorted by omissions. The task of restoring such observations was given sufficient attention, while approaches based on machine learning, above all, artificial neural networks and fuzzy systems that solve the problem of recovering these lost observations are very effective in this situation.

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: yevgeniy.bodyaskiy@nure.ua (Ye. Bodyanskiy); alina.shafronenko@nure.ua (A. Shafronenko); iryna.klymova@nure.ua (I. Klymova)
ORCID: 0000-0001-5418-2143 (Ye. Bodyanskiy); 0000-0002-8040-0279 (A. Shafronenko); 0000-0003-0455-6180 (I. Klymova)

At the same time, the described approaches to data recovery are workable only in cases when the initial data are set a priori, and the "object-property" table or time series has a fixed number of observations, i.e. do not change during processing. At the same time, there is a wide class of tasks, when the data are received for processing sequentially, as it happens when learning Kohonen's self-organizing maps [1] or their modifications [2-4].

One of the problem of artificial intelligence is clustering of big data distorted by omissions. For work with such data artificial neural networks, neuro-fuzzy systems, hybrid systems known for their universal approximating properties and ability to learn, seem to be the most effective. For the normal functioning of a neural network or a hybrid system distorted data have to be restored in some way. At the same time, most of these algorithms process information in batch form, that makes it difficult to use them in cases when data for processing are fed sequentially of time series. In the tasks when vectors-observation are fed sequentially in online mode, the number of missing values is doesn't known in advance, and in this case known approaches are ineffective.

Therefore, as an alternative this approaches, methods and algorithms have been proposed for adaptive recovery distorted data based on credibilistic fuzzy clustering approach.

2. Adaptive recovery of distorted data

When solving real problems related to the processing of data obtained as a result of either experiments or observations, quite often there arises a situation when the vectors - observations $x(\tau)$ contain missing observation that must be filled in the process before processing the source array. It is clear that such arrays with missing values can't be directly clustered, but must be preprocessed.

Let's present the original data set in the form of a traditional table, which contain information about N objects, each of than is described by $(n\times 1)$ of feature-vector $x^T(\tau)=\left(x_1(\tau),...,x_i(\tau),...,x_n(\tau)\right)$. It is assumed that N_G rows contain a missing values and $N_F=N-N_G$ are full.

Table 1
"object-property" table

	1	....	p	....	j	....	n
1	x_{11}	....	x_{1p}	....	x_{1j}	....	x_{1n}
....	....	....	....	....	....	....	....
i	x_{i1}	....	x_{ip}	....	x_{ij}	....	x_{in}
....	....	....	....	....	....	....	....
τ	$x_{\tau 1}$	....	$x_{\tau p}$	....	$x_{\tau j}$	....	$x_{\tau n}$
....	....	....	....	....	....	....	....
N	x_{N1}	....	x_{Np}	....	x_{Nj}	....	x_{Nn}

That is, in fact, this table is the source array $X^T=\{x_i(\tau)\}, i=1,2...,n;\ \tau=1,2...N$ in which there are absent N_G elements. Next, it is assumed that between the columns of the table $x_j=\left(x_j(1),x_j(2)...x_j(\tau),...x_j(N)\right)^T$ there exists a linear correlation on the basis of which the missed values can be restored $x_j(\tau)$ using a regression equation

$$\hat{x}_j(\tau)=a_{j0}+a_{j1}x_1(\tau)+a_{j2}x_2(\tau)+a_{j,j-1}x_{j-1}(\tau)+a_{j,j+1}x_{j+1}(\tau)...+a_{jn}x_n(\tau) \qquad (1)$$

or

$$\hat{x}_j(\tau) = \underset{\rightarrow}{x}_j(\tau) a_j \tag{2}$$

where $a_j = (a_{j0}, a_{j1}, \ldots a_{jn})^T$ - $(n \times 1)$ - vector of parameters to be determined; $\underset{\rightarrow}{x}_j(\tau) = \left(1, x_1(\tau), \ldots x_{j-1}(\tau), x_{j+1}(\tau) \ldots x_n(\tau)\right)$ - $(1 \times n)$ - vector - features for τ-th object without j - th element and a unit in the first position.

Vector of unknown parameters a_j can be found using the standard least squares method, for which from the matrix X^T the τ -th row, *j*-th column are deleted and the column formed by units is added to the left. Next on the basis $(\mathrm{N}-1) \times n$ of matrix X_j^T calculate the desired vector of estimates $a_j = \left(\mathrm{X}_j^{\bullet} \mathrm{X}_j^T\right)^+ \mathrm{X}_j \vec{x}_j$, where $\vec{x}_j = \left(x_j(1), \ldots x_j(\tau-1), x_j(\tau+1), \ldots x_j(N)\right)^T$, $(\bullet)^+$ - a symbol of pseudo-inversion by Moore - Penrose.

If missing values are contained in N_G rows and in other columns, from the source table X^T these rows are removed and reduced on the basis of the $(N_F \times n)$ matrix n - times are vectors of parameters a_j for all $j = 1, 2 \ldots n$ and missing values are filled with the received estimates.

This approach can be extended to situation when object data are fed to the matrix - table sequentially row by row. When the $(N+1)$-th observation in the form of a completely filled row $x^T(N+1) = \left(x_1(N+1), \ldots, x_i(N+1), \ldots, x_n(N+1)\right)$ vector-estimate a_j can be used in addition to the recurrent method of least squares, or you can look at the online method of machine learning:

$$\begin{cases} a_j(N_F+1) = a_j(N_F) + \dfrac{P_j(N_F)\left(x_j^T(N+1) - \underset{\rightarrow}{x}_j(N+1) a_j(N_F)\right)}{1 + \underset{\rightarrow}{x}_j^T(N+1) P_j(N_F) x_j(N+1)} \underset{\rightarrow}{x}_j(N+1), \\ P_j(\mathrm{N}_F+1) = \mathrm{P}_j(\mathrm{N}_F) - \dfrac{P_j(N_F) \underset{\rightarrow}{x}_j(N+1) \underset{\rightarrow}{x}_j^T(N+1) \mathrm{P}_j(N_F)}{1 + \underset{\rightarrow}{x}_j^T(N+1) P_j(N_F) \underset{\rightarrow}{x}_j(N+1)}, \end{cases} \tag{3}$$

after that, can be specified value $\hat{x}_j(\tau)$.

If $(N+1)$-th row contains missing values, it is skipped and the algorithm (3) waits row that contain all observation, for example $\underset{\rightarrow}{x}(N+2)$, after that it is calculated $a_j(N_F+1)$ by dint of $(N+2)$-th observations and all values are adjusted $\hat{x}_j(\tau)$, including $\hat{x}_j(N+1)$.

At the initial stages of processing Table 1, when the number of completely filled rows N_F is comparable to the number of columns n, the estimates obtained using the least squares method are characterized by low accuracy. In this situation, for sequential processing, it is more efficient to use adaptive learning algorithms that have both filtering and tracking (for non-stationary situations) properties.

The quality of solving various problems of forecasting, pattern recognition, inverse modeling, control, data recovery, etc. can be enhanced through the use of neural network ensembles, in which the same data are processed simultaneously by several n-parallel artificial neural networks. The output signals are combined in some way into an overall assessment, which gives an idea of the quality of the results obtained using the local networks of the ensembles, as shown in Figure 1.

The most widespread approach to uniting ANNs into ensembles are two ways, such as modular and based on weighted averaging. These approaches are quite different from each other, but they also have similarities. They both use a linear combination of the outputs of their members in one form or another.

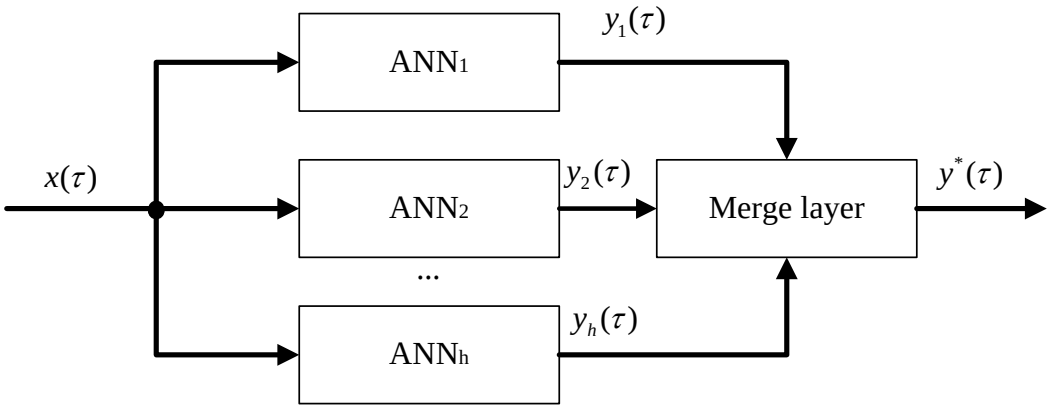

Figure 1: Ensemble of *h*- parallel organized artificial neural networks

The modular approach has a rather heuristic character in contrast to a more mathematically rigorous weighted averaging, although here remains an element of subjectivity associated with the choice of members of the ensemble. This problem is usually solved with the help of certain heuristics, although there are more or less rigorous results based on genetic programming or a gradual increase in the complexity of the networks - members of the ensemble.

It is convenient to organize information processing in the sequential data arrival mode on the basis of a neural network system, the main elements of which are parallel adaptive linear associators (ALA), trained using formula (3). Figure 3 shows the scheme of this system, which does not require additional explanations.

The vast majority of known fuzzy clustering algorithms assume that the original data array X contains N observation and do not change in the process of analysis. Can be noted, there exists a fairly many Data Stream Mining tasks, where data are fed for processing sequentially and their volume is a priori unknown, and Big Data Mining when this volume is so large that it simply does not allow processing this data in batch mode. In such situations, recurrent fuzzy clustering algorithms come to the fore, with the help of which these data are analyzed sequentially vector by vector as they enter the system [1-4]. Therefore, it is advisable to analyze the well-known recurrent methods of fuzzy clustering and propose new ones that differ in broader functionality in comparison with the existing methods.

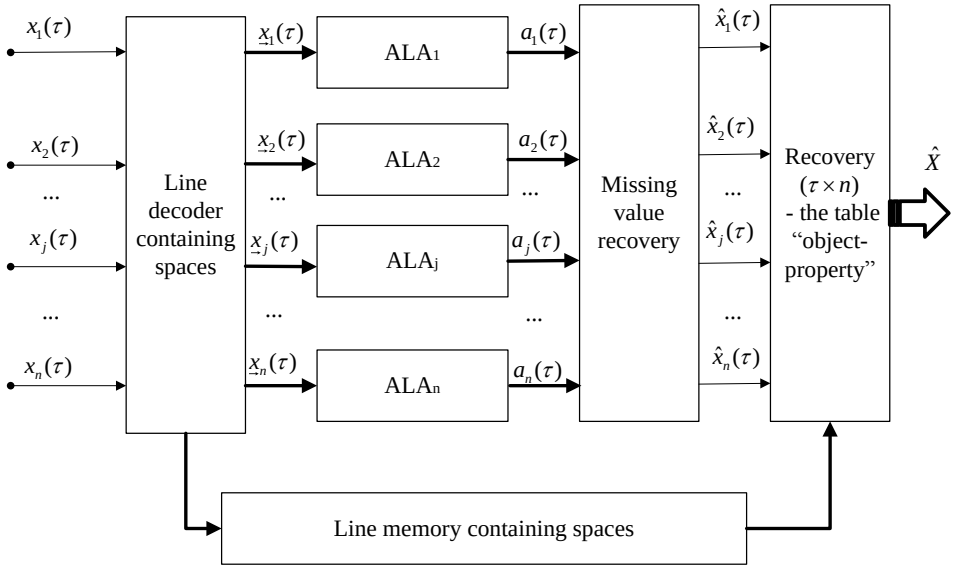

Figure 3: Adaptive neural network system for restoring missing values

3. Adaptive credibilistic fuzzy clustering

Credibilistic fuzzy clustering algorithms are connected with the goal function [5-8]:

$$E\left(Cred_q(\tau),c_q\right)=\sum_{\tau=1}^{N}\sum_{q=1}^{m}Cred_q^{\beta}(k)D^2\left(x(\tau),c_q\right) \tag{4}$$

in presence of constraints

$$\begin{cases}0\le Cred_q(\tau)\le 1\forall q,\tau,\\ \sup Cred_q(\tau)\ge 0,5\forall\tau,\\ Cred_q(\tau)+\sup Cred_l(\tau)=1\end{cases} \tag{5}$$

where $Cred_q(\tau)$ - level of observation $x(\tau)$ credibility.

In the procedures of credibilistic fuzzy clustering, the level of membership is determined by the membership functions [6]:

$$\mu_q(k)=\varphi_q\left(D\left(x(\tau),c_q\right)\right) \tag{6}$$

where φ_q - decreases monotonically on the interval $[0,\infty]$ and with condition $\varphi_q(0)=1,\varphi_q(\infty)\to 0$.

It is easy to see that membership level (6) using the distance is based on similarity measure [8,11-15]. As such a measure in [8], it was proposed to use a function

$$\mu_q(k)=\left.1\right/1+D^2\left(x(\tau),c_q\right). \tag{7}$$

Thus, if the fuzzy clustering algorithm in a batch form can be can be written as [7,8]

$$\begin{cases}\mu_q(\tau)=\left.1\right/1+D^2\left(x(\tau),c_q\right),\\ \mu_q^*(\tau)=\left.\mu_q(\tau)\right/\sup\mu_l(\tau),\\ Cred_q(\tau)=\left.\mu_q^*(\tau)+1-\sup\mu_l^*(\tau)\right/2,\\ c_q=\dfrac{\sum_{\tau=1}^{N}Cred_q^{\beta}(\tau)x(\tau)}{\sum_{\tau=1}^{N}Cred_q^{\beta}(\tau)},\end{cases} \tag{8}$$

in the online mode this procedures (8) has the form (9):

$$
\begin{cases}
\sigma_q^2(\tau+1) = \sum\limits_{\substack{l=1 \\ l\neq q}}^{m} \left(D^2\left(x(\tau+1), c_l(\tau)\right)^{\frac{1}{1-\beta}} \right)^{-1}, \\
\mu_q(\tau+1) = \left(1 + \dfrac{\left(D^2\left(x(\tau+1), c_q(\tau)\right)\right)^{\beta-1}}{\sigma_q^2(\tau+1)} \right)^{-1}, \\
\mu_q^*(\tau+1) = \dfrac{\mu_q(\tau+1)}{\sup \mu_l(\tau+1)}, \\
Cred_q(\tau+1) = \dfrac{1}{2}\left(\mu_q^*(\tau+1) + 1 - \sup \mu_l^*(\tau+1)\right), \\
c_q(\tau+1) = c_q(\tau) + \eta(\tau+1) Cred_q^{\beta}(\tau+1)\left(x(\tau+1) - c_q(\tau)\right)
\end{cases}
\tag{9}
$$

or in case when $\beta = 2$ [9]

$$
\begin{cases}
\sigma_q^2(\tau+1) = \left(\sum\limits_{\substack{l=1 \\ l\neq q}}^{m} \left\| x(\tau+1) - c_l(\tau) \right\|^{-2} \right)^{-1}, \\
\mu_q(\tau+1) = \left(1 + \dfrac{\left\| x(\tau+1) - c_q(\tau) \right\|^2}{\sigma_q^2(\tau+1)} \right)^{-1}, \\
\mu_q^*(\tau+1) = \dfrac{\mu_q(\tau+1)}{\sup \mu_l(\tau+1)}, \\
Cred_q(\tau+1) = \dfrac{1}{2}\left(\mu_q^*(\tau+1) + 1 - \sup \mu_l^*(\tau+1)\right), \\
c_q(\tau+1) = c_q(\tau) + \eta(\tau+1) Cred_q^{2}(\tau+1)\left(x(\tau+1) - c_q(\tau)\right).
\end{cases}
\tag{10}
$$

It is easy to see that the recurrent fuzzy clustering algorithm is not more complex than the online modifications of probabilistic, possiblilistic, and robust procedures.

4. Experimental research

The proposed algorithm was tested in experimental research, that was conducted using well-known test data sets of the UCI repository: Abalone and Gas. The comparison results of mean error are demonstrated in Table 3. The mean error of the clusters centroids of proposed Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach (AdCFCA) was compared with another well known methods of FCM and Gustafson-Kessel (GK).

Table 2
Data set

Initial set	Amount of data	Amount of Attributes	Amount of Cluster
Abalone	4177	8	3
Gas	296	2	6

Table 3
Comparison results of mean error of the clusters centroids

Data set	FCM	GK	AdCFCA
Abalone	2.61	0.10	0.75
Gas	2.69	0.13	0.64

Table 4
Estimation of the quality of fuzzy clustering

Methods of Data clustering	PC	SC	XB
FCM	0.51	1.63	0.18
Gustafson-Kessel	0.26	**1.65**	1.63
Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach	**0.24**	0.64	**0.15**

Easy to see that the approach under consideration shows better results clustering quality.

To estimate the quality of Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach, we used the overall accuracy comparison of 50, 100 and 150 experiments for datasets and another clustering procedures such as: FCM and Gustafson-Kessel (GK).

Table 5
A comparison of 50 experiments

Data	Algorithm of clustering	Accuracy	
		Highest	Mean
Abalone	Fuzzy C-means	63.34	63.34
	Gustafson-Kessel	64.45	64.40
	Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach	64.28	64.28
Gas	Fuzzy C-means	69.35	69.23
	Gustafson-Kessel	69.68	69.55
	Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach	64.68	64.55

Table 6
A comparison of 100 experiments

Data	Algorithm of clustering	Accuracy	
		Highest	Mean
Abalone	Fuzzy C-means	62.45	62.24
	Gustafson-Kessel	61.75	61.30
	Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach	61.68	60.98
Gas	Fuzzy C-means	65.45	57.33
	Gustafson-Kessel	64.68	55.55
	Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach	63.27	58.45

Table 7
A comparison of 150 experiments for the other data set

Data	Algorithm of clustering	Accuracy	
		Highest	Mean
Abalone	Fuzzy C-means	58.54	58.54
	Gustafson-Kessel	57.55	57.55
	Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach	57.48	57.48
Gas	Fuzzy C-means	69.05	67.33
	Gustafson-Kessel	68.88	65.55
	Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach	58.58	58.55

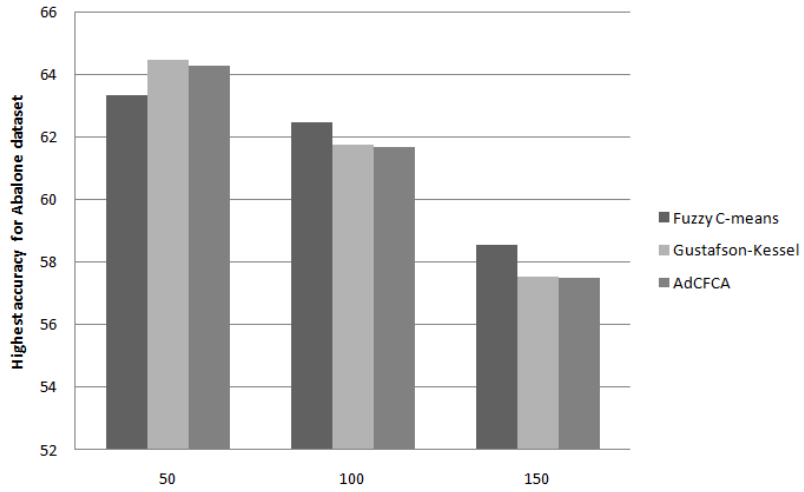

Figure 4: Comparison of high accuracy for Abalone data

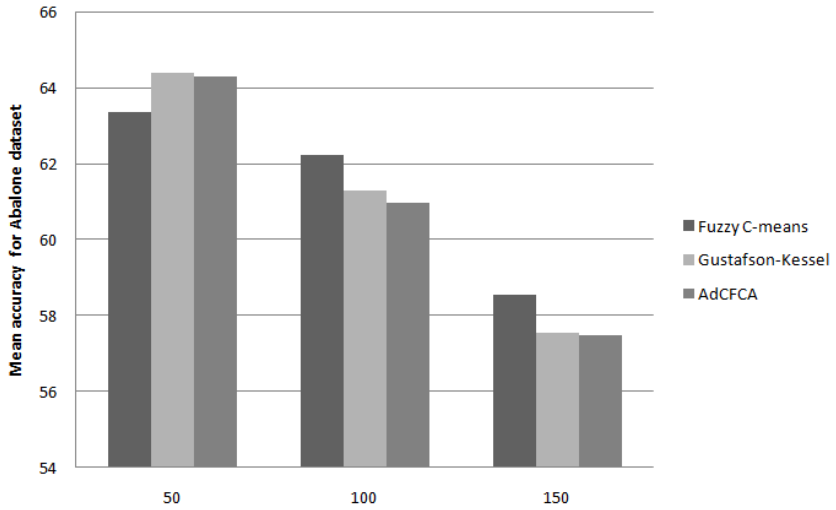

Figure 5: Comparison of mean accuracy for Abalone data

On Figure 4 and Figure 5 the comparison of overall accuracy of algorithms was presented. Easy to see that the proposed approach is better than the known algorithms.

5. Conclusion

The problem fuzzy clustering of data distorted by missing values and outliers is considered. The recovery in the mode of sequential data arrival for processing is investigated. As an alternative to the classic possibilistic and probabilistic approaches we use of the essence modify credibilistic one in a case when distorted data are fed in online mode. The modification consists in the introducing to the clustering system additional recovery block of data that are processing in real time. For solving the clustering problem we introduce the adaptive gradient algorithm for minimizing of goal function, related with credibilistic fuzzy clustering and based on similarity measure of special type. Experimental research confirms the effectiveness of evolving approach. This algorithm is characterized by easy numerical implementation, high speed and can to process information in online mode when the data are fed sequentially in real time.

6. Acknowledgement

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).

7. References

[1] T. Kohonen, Self-Organizing Maps, Springer, Berlin, 1995. doi: 10.1007/978-3-642-56927-2.
[2] Ye. Bodyanskiy, V. Kolodyazhniy, A. Stephan, Recurcive fuzzy clustering algorithms, in: Proc. 10th East West Fuzzy Coll. 2002, Zittau-Görlitz, HS, 2002, pp. 276–283.
[3] J. C. Bezdek, A convergence theorem for the fuzzy ISODATA clustering algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE 2(1) (1980) 1–8. doi: 10.1109/TPAMI.1980.4766964.
[4] R. Krishnapuram, J.M. Keller, A Possibilistic Approach to Clustering, IEEE Transactions on Fuzzy Systems, IEEE, 1(2) (1993) 98–110. doi: 10.1109/91.227387.
[5] F. Höppner, F. Klawonn, R. Kruse, T. Runker, Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition, John Wiley &Sons, Chichester, 1999.
[6] C.C. Aggarwal, Data Mining: Text Book, Springer, 2015.
[7] J. Zhou, Q. Wang, C.-C. Hung, X. Yi, Credibilistic clustering: the model and algorithms, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 23(4) (2015) 545–564. doi: 10.1142/S0218488515500245.
[8] J. Zhou, Q. Wang, C. C. Hung, Credibilistic clustering algorithms via alternating cluster estimation, Journal of Intelligent Manufacturing, 28 (2017) 727–738. doi: 10.1007/s10845-014-1004-6.
[9] A. Shafronenko, Ye. Bodyanskiy, I. Klymova, O.Holovin, Online credibilistic fuzzy clustering of data using membership functions of special type, in: Proceedings of The Third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), April 27-1 May 2020, Zaporizhzhia, 2020. URL: http://ceur-ws.org/Vol-2608/paper56.pdf.
[10] B. Liu, A survey of credibility theory, Fuzzy Optimization and Decision Making, 4 (2006) 387–408. doi: 10.1007/s10700-006-0016-x.
[11] Ye. Bodyanskiy, A. Shafronenko, S. Mashtalir, Online robust fuzzy clustering of data with omissions using similarity measure of special type, in: S. Babichev, V. Lytvynenko, W. Wójcik, S. Vyshemyrskaya (Eds.), Lecture Notes in Computational Intelligence and Decision Making-Cham, Springer, 2020, pp. 637-646.
[12] F. Zhao, L. Jiao, H. Liu, Fuzzy c-means clustering with nonlocals partial information for noisy image segmentation, 5(1) (2011) 45–56. doi: 10.1007/s11704-010-0393-8.
[13] J. J. Sepkovski, Quantified coefficients of association and measurement of similarity, Int. J. Assoc. Math, 2(6) (1974) 135–152.

[14] F. W. Young, R.M. Hamer Theory and Applications of Multidimensional Scaling-Hillsdale, Erlbaum, N.J., 1994.
[15] A. Shafronenko, Ye. Bodyanskiy, D. Rudenko, Online neuro fuzzy clustering of data with omissions and outliers based on completion strategy, in: Proceedings of The Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2019), Zaporizhzhia, 2019, pp. 18-27.