Online Credibilistic Fuzzy Clustering of Data Using Membership Functions of Special Type

Alina Shafronenko[1[0000-0002-8040-0279]], Yevgeniy Bodyanskiy[1[0000-0001-5418-2143]], Iryna Klymova[1[0000-0003-0455-6180]], Olexii Holovin[2[0000-0003-4662-4559]]

[1]Kharkiv National University of Radio Electronics, Nauky Ave., 14, Kharkiv, Ukraine
alina.shafronenko@nure.ua, yevgeniy.bodyanskiy@nure.ua,
iryna.klymova@nure.ua

[2]Central Research Institute of Weapons and Military Equipment of the Armed Forces of Ukraine, Povitroflotsky Ave., 28, Kyiv, Ukraine
a_a_golovin@ukr.net

Abstract. In the paper new online method of credibilistic fuzzy clustering of data was proposed. This algorithm is based on credibilistic approaches using on batch and online modes of information prosessing. Using proposed approach it's possible to solve clustering task in on-line mode when data are fed to processing sequentially, possible in real time.

Keywords: fuzzy clustering, learning rule, possibilistic fuzzy clustering, probabilistic fuzzy clustering, credibilistic fuzzy clustering, measure of similarity.

CEUR-WS.org/Vol-2608/paper56.pdf

1 Introduction

The task of clustering (classification in the self-learning mode) of multidimensional data is an important part of Data Mining, within which a number of directions and approaches have developed [1, 2]. One of these areas is formed by fuzzy clustering methods, that are based on the assumption that the generated clusters - classes mutually overlap so that each vector - observation with different levels of membership-probability-possibility can belong to several or to all classes.

Here, the algorithms of probabilistic fuzzy clustering and, first of all, the fuzzy c-means method (FCM) are most widely used [3, 4]. The possibilities of this approach are limited by probabilistic restrictions on membership levels so that observations "contaminated" with disturbances and outliers can be assigned to different classes with almost identical membership levels.

In this regard, in [5] was proposed the possibilistic fuzzy clustering method (PCM) that is more resistant to noise and disturbances. At the same time, PCM algorithms suffer from the so-called coincident problem, when during the processing of information some clusters begin to merge with each other, that leads to an incorrect estimate of the number of formed clusters.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Algorithms of credibilistic fuzzy clustering [6-8], based on the apparatus of the theory of credibility [9], are devoid of these shortcomings. As part of this approach in the calculate process to evaluate not only the fuzzy membership levels, but also credibility levels based on a membership measure of a special type [10]. The experimental results have shown [7, 8] that the credibilistic methods provide a higher quality of clustering comparively with probabilistic and possibilistic methods.

The initial information for solving the problem of fuzzy clustering is an array of n - dimensional observations - vectors $X = \{x_1, x_2, ..., x_N\} \subset R^n$, $x(k) \in X$, $k = 1, 2, ..., N$, that should be divided into m classes-clusters with a certain level of membership – probability – possibility $U_q(k)$ of k th vector x_k to q th cluster $(1 < m < N, 1 \le q \le m)$. It should also be noted that the initial data are pre-processed so that $-1 \le x_{ki} \le 1$ $(1 \le i \le n)$ where x_{ki} – i th component of vector x_k.

Thus, the clustering problem is solved in batch mode, when the entire data array is processed multiple times based on alternating cluster estimation [8]. If the data arrive for processing in the form of a stream or form big data, the batch mode does not allow to solve the problem under consideration effectively. In this situation, the most effective are the recursive fuzzy clustering procedures that allow to solve the problem online and refine the desired solution as each new observation that arrives. Thus, in the [11, 12] recurrent variants of FCM have been proposed that are essentially gradient optimization procedures adopted by the goal function, and in [13, 14] recurrent PCM modifications have been introduced designed for sequential data processing.

In this regard, it seems appropriate to develop a recurrent modification of the method of credibilistic fuzzy clustering, that allows clarifying the desired characteristics of the clusters as each new observation arrives.

2 Recurrent Method Of Credibilistic Fuzzy Clustering (RCCM)

The most popular method of probabilistic fuzzy clustering is associated with minimizing the goal function (1) [4]

$$E(U_q(k), w_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^{\beta}(k) D^2(x_k, w_q) \tag{1}$$

with constrains $\sum_{q=1}^{m} U_q(k) = 1,\ 0 < \sum_{q=1}^{m} U_q(k) < N.$ Solving the nonlinear programming problem using the method of indefinite Lagrange multipliers, we arrive to the known result (2) and (3)

$$U_q^{(\tau+1)}(k) = \left(D^2\left(x_{k,} w_q^{(\tau)}\right)\right)^{\frac{1}{1-\beta}} \left(\left(\sum_{l=1}^{m}\left(D^2\left(x_k, w_l^{(\tau)}\right)\right)\right)^{\frac{1}{1-\beta}}\right)^{-1}, \tag{2}$$

$$w_q^{(\tau+1)} = \sum_{k=1}^{N} \left(U_q^{(\tau+1)}(k) \right)^{\beta} x_k \left(\sum_{k=1}^{N} \left(U_q^{\tau+1}(k) \right)^{\beta} \right)^{-1}, \qquad (3)$$

where $U_q(k)$ – membership levels of vector observation x_k to q^{th} cluster $Cl_q (1 \le q \le m)$, w_q - prototype - centroid of q^{th} cluster, $\beta > 1$ fuzzifier, that defines the “blurring” of boundaries between classes, $D\left(x_k, w_q\right)$ – the distance between x_k and w_q in adopted metric, $\tau = 0,1,2,\ldots$ – index of the epoch of information processing in the alternate estimation mode. In this case, the calculation process continues until the conditions

$$w_q^{(\tau+1)} - w_q^{(\tau)} \le \varepsilon \forall 1 \le q \le m$$

is satisfied, where ε – reassigned threshold calculation accuracy.

In the case, when $\beta = 2$ and Euclidean metric $D^2\left(x_k, w_q\right) = \left\| x_k - w_q \right\|_2^2$ is used, we obtain to the popular fuzzy c-means algorithm (FCM) [15] in the form (4), (5)

$$U_q^{(\tau+1)}(k) = \left\| x_k - w_q^{(\tau)} \right\|^{-2} \left(\sum_{l=1}^{m} \left\| x_k - w_l^{(\tau)} \right\|^{-2} \right)^{-1}, \qquad (4)$$

$$w_q^{(\tau+1)} = \sum_{k=1}^{N} \left(U_q^{(\tau+1)}(k) \right)^{2} x_k \left(\sum_{k=1}^{N} \left(U_q^{\tau+1}(k) \right)^{2} \right)^{-1}. \qquad (5)$$

If the data are processed sequentially online, the nonlinear programming task can be solved using the Arrow-Hurwitz-Uzawa algorithm, which is essentially a gradient procedure for finding the saddle point of the Lagrange function based on criterion (1) with constraints on the sum of membership levels.

In this case relations (2), (3) can be rewritten in the form (6)

$$\begin{cases} U_q\left(k+1\right) = \left(D^2\left(x_{k+1}, w_q^{(k)}\right) \right)^{\frac{1}{1-\beta}} \left(\sum_{l=1}^{m} \left(x_{k+1}, w_l(k) \right)^{\frac{1}{1-\beta}} \right)^{-1} \\ w_q\left(k+1\right) = w\left(k\right) + \eta\left(k+1\right) U_q^{\beta}\left(k+1\right)\left(x_{k+1} - w_q\left(k\right)\right) \end{cases}, \qquad (6)$$

where $\eta(k)$ – learning rate parameter, and (4), (5) can be rewritten in the form (7)

$$\begin{cases} U_q\left(k+1\right) = \left\| x_k - w_q\left(k\right) \right\|^{-2} \left(\sum_{l=1}^{m} \left\| x_k - w_l\left(k\right) \right\|^{-2} \right)^{-1} \\ w_q\left(k+1\right) = w_q + \eta\left(k+1\right) U_q^{2}\left(k+1\right)\left(x_{k+1} - w_q\left(k\right)\right) \end{cases}, \qquad (7)$$

which are a generalization of the recurrent procedures of Park-Dagger [11] and Chung-Lee [12].

Possibilistic algorithms for fuzzy clustering are based on minimizing the goal function (8) [5]

$$E\left(U_q\left(k\right),w_q,\mu_q\right)=\sum_{k=1}^{N}\sum_{q=1}^{m}U_q^{\beta}\left(K\right)D^2\left(x_k,w_q\right)+\sum_{q=1}^{m}\mu_q\sum_{k=1}^{N}\left(1-U_q\left(k\right)\right)^{\beta}, \qquad (8)$$

where $\mu_q \geq 0$ determines the distance at which the membership level takes the value 0.5, i.e. $U_q(k)=0$ if $\mathrm{D}^2(x_k,w_q)=\mu$.

Minimization of criterion (8) allows us to obtain analytical solution in the form (9) – (11)

$$U_q^{(\tau+1)}\left(k\right)=\left(1+\left(\frac{D^2\left(x_k,w_q^{(\tau)}\right)}{\mu_q^{(\tau)}}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \qquad (9)$$

$$w_q^{(\tau+1)}=\sum_{k=1}^{N}\left(U_q^{(\tau+1)}\left(k\right)\right)^{\beta}x_k\left(\sum_{k=1}^{N}\left(U_q^{\tau+1}\left(k\right)\right)^{\beta}\right)^{-1}, \qquad (10)$$

$$\mu_q^{(\tau+1)}=\sum_{k=1}^{N}\left(U_q^{(\tau+1)}\left(k\right)\right)^{\beta}D^2\left(x_k,w_q^{(\tau+1)}\right)\left(\sum_{k=1}^{N}\left(U_q^{(\tau+1)}\left(k\right)\right)^{\beta}\right)^{-1}, \qquad (11)$$

which in the quadratic case takes the form (12) – (14)

$$U_q^{(\tau+1)}\left(k\right)=\left(1+\frac{\left\|x_k-w_q^{(\tau)}\right\|^2}{\mu_q^{(\tau)}}\right)^{-1}, \qquad (12)$$

$$w_q^{(\tau+1)}=\sum_{k=1}^{N}\left(U_q^{(\tau+1)}\left(k\right)\right)^{2}x_k\left(\sum_{k=1}^{N}\left(U_q^{(\tau+1)}\left(k\right)\right)^{2}\right)^{-1}, \qquad (13)$$

$$\mu_q^{(\tau+1)}=\sum_{k=1}^{N}\left(U_q^{(\tau+1)}\left(k\right)\right)^{2}\left\|x_k-w_q^{(\tau+1)}\right\|^2\left(\sum_{k=1}^{N}\left(U_q^{(\tau+1)}\left(k\right)\right)^{2}\right)^{-1}. \qquad (14)$$

Online versions (9) – (14) at the same time have the form (15) [13, 14]

$$\begin{cases} U_q(k+1)=\left(1+\left(\dfrac{D^2\left(x_{k+1},w_{\varepsilon}\left(k\right)\right)}{\mu_q\left(k\right)}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ w_q\left(k+1\right)=w_q\left(k\right)+\eta\left(k+1\right)U_q^{\beta}\left(k+1\right)\left(x_{k+1}-w_q\left(k\right)\right), \\ \mu_q\left(k+1\right)=\sum_{p=1}^{k+1}U_q^{\beta}\left(p\right)D^2\left(x_p,w_q\left(k+1\right)\right)\left(\sum_{p=1}^{k+1}U_q^{\beta}\left(p\right)\right)^{-1} \end{cases} \qquad (15)$$

and in the case $\beta=2$ (16)

$$\begin{cases} U_q\left(k+1\right)=\left(1+\dfrac{\left\|x_{k+1}-w_q\left(k\right)\right\|^2}{\mu_q\left(k\right)}\right)^{-1}, \\ w_q\left(k+1\right)=w_q\left(k\right)+\eta\left(k+1\right)U_q^2\left(k+1\right)\left(x_{k+1}-w_q\left(k\right)\right), \\ \mu_q\left(k+1\right)=\sum\limits_{p=1}^{k+1}U_q^2\left(p\right)\left\|x_p-w_q\left(k+1\right)\right\|^2\left(\sum\limits_{p=1}^{k+1}U_q^2\left(p\right)\right)^{-1}. \end{cases} \tag{16}$$

Credibilistic fuzzy clustering is associated with minimizing the goal function (17)

$$E\left(Cr_q\left(k\right),w_q\right)=\sum_{k=1}^{N}\sum_{q=1}^{m}Cr_q^{\beta}\left(k\right)D^2\left(x_k,w_q\right) \tag{17}$$

with constraints $0\leq Cr_q\left(k\right)\leq 1\forall q,k;$ $\mathrm{supCr}_q(\mathrm{k})\geq 0,5\forall k$; $Cr_q\left(k\right)+supCr_l\left(k\right)=1$ for any q and k for which $Cr_q\left(k\right)\geq 0.5$. Here $Cr_q\left(k\right)$ – credibility that observation x_k belongs to a cluster Cl_q. In this case, the membership level is calculated based on the membership function (18) [15]

$$U_q\left(k\right)=\varphi_q\left(D\left(x_k,w_q\right)\right) \tag{18}$$

where: $\varphi_q\left(\cdot\right)$ – monotonically decreases in the interval $[0,\infty]$, $\varphi_q\left(0\right)=1$, $\varphi_q\left(\infty\right)\rightarrow 0$.

It is easy to see that function (18) is essentially measure of similarity based on distance [16]. As such a function, it was proposed in [15] to use the expression (19)

$$U_q\left(k\right)=\left(1+D^2\left(x_k,w_q\right)\right)^{-1}. \tag{19}$$

It is interesting to note that expression (2) can be rewritten in the form (20)

$$\begin{aligned} U_q\left(k\right)&=\left(D^2\left(x_k,w_q\left(k\right)\right)\right)^{\frac{1}{1-\beta}},\left(\sum_{l=1}^{m}\left(D^2\left(x_k,w_l\left(k\right)\right)\right)^{\frac{1}{1-\beta}}\right)^{-1}= \\ &=\left(D^2\left(x_k,w_k\left(k\right)\right)\right)^{\frac{1}{1-\beta}}\left(D^2\left(x_k,w_q\left(k\right)\right)\right)^{\frac{1}{1-\beta}}+\sum_{\substack{l=1\\l\neq q}}^{m}\left(D^2\left(x_k,w_l\left(k\right)\right)\right)^{\frac{1}{1-\beta}}\ )^{-1}= \\ &=\left(1+\left(D^2\left(x_k,w_q\left(k\right)\right)\right)^{\frac{1}{1-\beta}}\sum_{\substack{l=1\\l\neq q}}^{m}\left(D^2\left(x_k,w_l\left(k\right)\right)\right)^{\frac{1}{1-\beta}}\right)^{-1}, \end{aligned} \tag{20}$$

and for the Euclidean metric and $\beta=2$ takes the form of a Cauchy distribution density function with a width parameter σ_q^2 (21), (22), [17]

$$U_q\left(k\right)=\left(1+\frac{\left\|x_k-w_q\left(k\right)\right\|^2}{\sigma_q^2}\right)^{-1}, \tag{21}$$

$$\sigma_q^2=\left(\sum_{\substack{l=1\\l\neq q}}^{m}\left\|x_k-w_l\left(k\right)\right\|^{-2}\right)^{-1}. \tag{22}$$

It is easy to see that the membership function (19) is a special case of (21) for $\sigma_q^2=1$.

Finally, a batch algorithm of credibilistic fuzzy clustering can be written in the form (23) – (26) [7, 8]:

$$U_q^{(\tau+1)}\left(k\right)=\left(1+D^2\left(x_k,w_q^{(\tau)}\right)\right)^{-1}, \tag{23}$$

$$U*_q^{(\tau+1)}\left(k\right)=U_q^{(\tau+1)}\left(k\right)\left(\sup U_l^{(\tau+1)}\left(k\right)\right)^{-1}, \tag{24}$$

$$Cr_q^{(\tau+1)}\left(k\right)=\frac{1}{2}\left(U_q^{*(\tau+1)}\left(k\right)+1-\sup_{l\neq q}U_l^*\left(k\right)\right), \tag{25}$$

$$w_q^{(\tau+1)}=\sum_{k=1}^{N}\left(Cr_q^{(\tau+1)}\left(k\right)\right)^{\beta}x_k\left(\sum_{k=1}^{N}\left(Cr_q^{\tau+1}\left(k\right)\right)^{\beta}\right)^{-1}. \tag{26}$$

Based on (17), (21) – (26), we can write online version of the credibilistic fuzzy clustering method in the form (27)

$$\begin{cases}\sigma_q^2\left(k+1\right)=\left(\sum_{\substack{l=1\\l\neq q}}^{m}\left\|x_{k+1}-w_l\left(k\right)\right\|^{-2}\right)^{-1},\\ U_q\left(k+1\right)=\left(1+\frac{\left\|x_{k+1}-w_q\left(k\right)\right\|^2}{\sigma_q^2\left(k+1\right)}\right)^{-1},\\ U_{(k+1)}^*=U_q\left(k+1\right)\left(\sup U_l\left(k+1\right)\right)^{-1},\\ Cr_q\left(k+1\right)=\frac{1}{2}\left(U_q^*\left(k+1\right)+1-\sup_{l\neq q}U_l^*\left(k+1\right)\right),\\ w_q\left(k+1\right)=w_q\left(k\right)+\eta\left(k+1\right)Cr_q^{\beta}\left(k+1\right)\left(x_{k+1}-w_q\left(k\right)\right).\end{cases} \tag{27}$$

Therefore, from a computational point of view, the online algorithm for credibilistic fuzzy clustering is no more complicated than the recurrent versions of FCM and PCM, while retaining the advantages of a credibility approach.

3 Experiments

To check the performance efficiency of the developed methods as well as to prove their benefits over the analogs, experimental research was conducted with the help on two different databases. Also conducted a comparative analysis of the quality of clustering data on the main characteristics quality ratings, such as: Partition Coefficient (PC) defines "overlapping" between groups of points, Partition Index (SC) quantifies the ratio sum of compactness and separation of the clusters, Xie and Beni's Index (XB) gauges a ratio of the total variability inside clusters and their separation, of existing clustering methods and proposed method.

The experimental results are given in Table 1 and Table 2.

Such clustering tools as fuzzy c-means (FCM) and the algorithm by Gustafson-Kessel (GK), Gath-Geva (GG), Adaptive probabilistic fuzzy clustering, Adaptive fuzzy possibilistic data clustering and Adaptive fuzzy credibilistic data clustering were chosen for comparison with the developed procedure.

Table 1. Evaluation of the quality of fuzzy clustering methods using first data set

Data clustering methods	First Data Set		
	PC	SC	XB
Fuzzy C-means (FCM)	0.50	1.62	0.19
Gustafson-Kessel	0.27	1.66	1.62
Gath-Geva	0.25	1.54	1.35
Adaptive probabilistic fuzzy clustering	0.25	1.44	**0.01**
Adaptive fuzzy possibilistic data clustering	0.26	1.22	**0.01**
Adaptive fuzzy credibilistic data clustering	**0.21**	**1.13**	**0.01**

Table 2. Evaluation of the quality of fuzzy clustering methods using second data set

Data clustering methods	Second Data Set		
	PC	SC	XB
Fuzzy C-means (FCM)	0.48	1.60	0.19
Gustafson-Kessel	0.26	1.64	1.62
Gath-Geva	0.26	1.50	1.35
Adaptive probabilistic fuzzy clustering	0.25	1.42	**0.01**
Adaptive fuzzy possibilistic data clustering	0.37	**1.11**	0.18
Adaptive fuzzy credibilistic data clustering	**0.23**	1.22	**0.01**

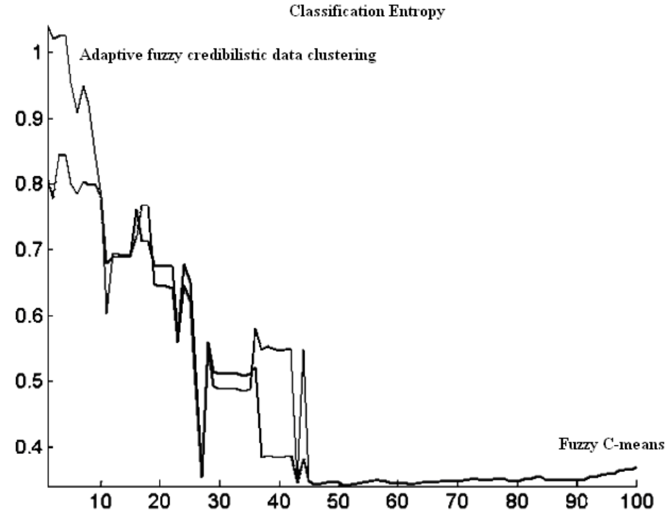

Fig. 1. Comparison CE of Adaptive fuzzy credibilistic data clustering and Fuzzy C-means (FCM) methods for First Data Set

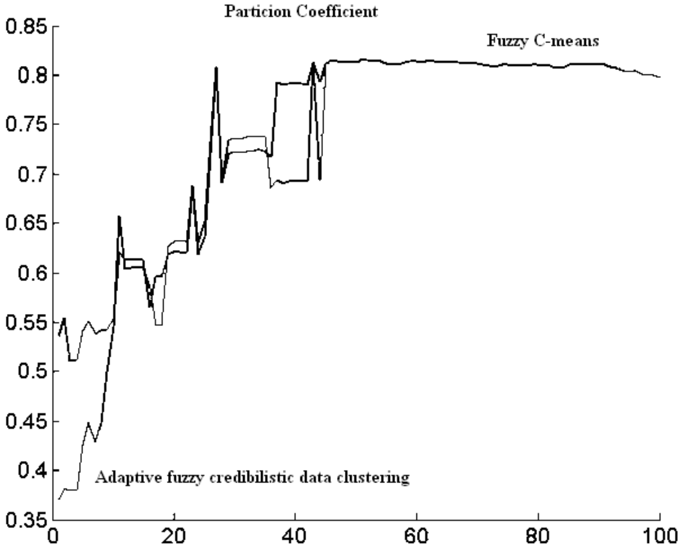

Fig. 2. Comparison PC of Adaptive fuzzy credibilistic data clustering and Fuzzy C-means (FCM) methods for First Data Set

In comparison with the well-known methods, this approach to data clustering demonstrates somewhat reliable results (Fig. 1, Fig. 2, Fig. 3, Fig. 4).

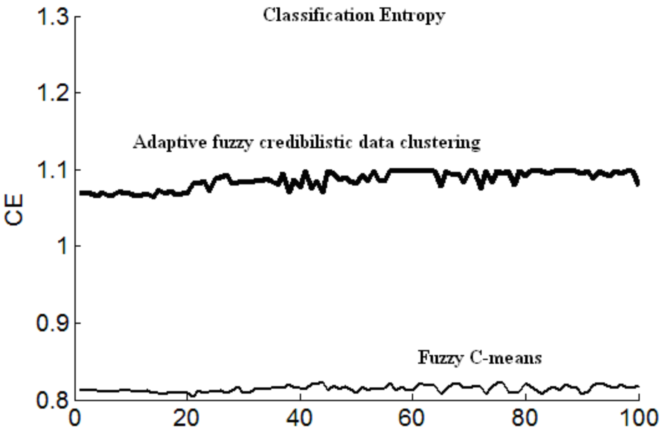

Fig. 3. Comparison CE of Adaptive fuzzy credibilistic data clustering and Fuzzy C-means (FCM) methods for Second Data Set

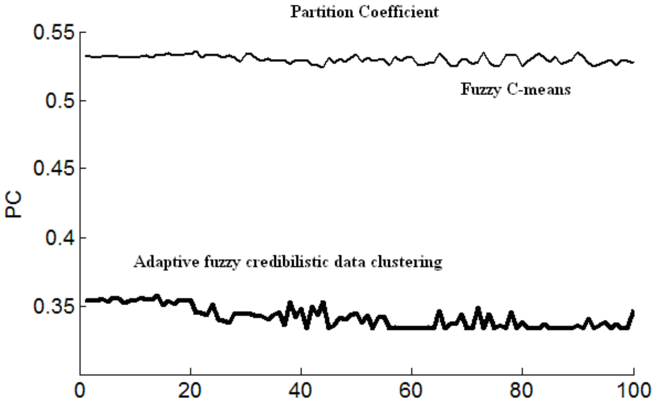

Fig. 4. Comparison PC of Adaptive fuzzy credibilistic data clustering and Fuzzy C-means (FCM) methods for Second Data Set

4 Conclusion

The problem of fuzzy clustering based on probabilistic, possibilistic and credibilistic approaches based online mode of information processing was considered. A recurrent version of credibilistic algorithm is introduced, which is essentially a gradient optimization procedure for the accepted criterion for fuzzy credibilistic clustering. A modification of the membership function is introduced, which is essentially a measure of similarity and a generalization of previously known functions. The considered recursive procedures are simple in numerical implementation and are intended to solve problems arising in the framework of Data Stream Mining.

References

1. Xu, R., Wunsch, D.C.: Clustering. Hoboken, N.J. John Wiley & Sons, Inc. (2009)
2. Aggarwal, C.C.: Data Mining: Text Book. Springer (2015).
3. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981).
4. Höeppner, F., Klawonn, F., Kruse, R., Runker, T.: Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester, John Wiley &Sons (1999)
5. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. Fuzzy Systems, №2, pp.98-110 (1993)
6. Chintalapudi, K. K., Kam, M.: A noise resistant fuzzy c-means algorithm for clustering. In: IEEE Conference on Fuzzy Systems Proceedings, vol. 2, pp. 1458-1463 (1998)
7. Zhou, J., Wang, Q., Hung, C.-C., Yi, X.: Credibilistic clustering: the model and algorithms. Int.J. of Uncertainty, Fuzziness and Knowledge-Based Systems, №4, pp.545-564 (2015).
8. Zhou, J., Wang, Q., Hung, C. C.: Credibilistic clustering algorithms via alternating cluster estimation. J. Intell. Manuf., pp.727-738 (2017).
9. Liu, B., & Liu, Y. Expected value of fuzzy variable and fuzzy expected value models. In: "IEEE Transactions on Fuzzy Systems", № 4, pp. 445–450 (2002).
10. Liu, B.: A survey of credibility theory. Fuzzy Optimization and Decision Making, №4, pp. 387–408 (2006).
11. Park, D.C., Dagher, I.: Gradient based fuzzy c-means (GBFCM) algorithm. In: Proc. IEEE Int. Conf. on Neural Networks, pp.1626-1631(1984).
12. Chung, F.L., Lee,T.: Fuzzy competitive clustering. Neural Networks, 7(3), pp.539-552. (1994).
13. Bodyanskiy, Ye, Kolodyazhniy, V., Stephan, A.: Recursive fuzzy clustering algorithms. In : Proc 10th East West Fuzzy Coll. Zittau-Görlitz, HS, pp.276-283 (2002).
14. Bodyanskiy, Ye.: Computational intelligence techniques for data analisys. In: Lecture Notes in Informatics. Bonn: V.P. – 72, GI, pp.15-36 (2005).
15. Zhou, J., & Hung, C.-C.: A generalized approach to possibilistic clustering algorithms. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems. 15. pp. 117–138. (2007).
16. Young, F.W., Hamer, R.M.: Theory and Applications of Multidimensional Scaling-Hillsdale, N.J.: Erlbaum (1994).
17. Hu, Zh., Bodyanskiy, Ye, Tyshchenko, O., Shafronenko, A.: Fuzzy clustering of incomplete data by means of similarity measures. In: 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering UKRCON - 2019, Conference Proceedings, July 2-6, 2019, Lviv, Ukraine, pp.149-152 (2019) doi: 10.1109/UKRCON.2019.8879844.