IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Fuzzy Clustering of Distorted Observations Based On Optimal Expansion Using Partial Distances

Alina Shafronenko
Informatics Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
alina.shafronenko@nure.ua

Artem Dolotov
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
artem.dolotov@gmail.com

Yevgeniy Bodyanskiy
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine;
yevgeniy.bodyanskiy@nure.ua

Galina Setlak
Rzeszow University of Technology
Rzeszow, Poland;
gsetlak@prz.edu.pl

***Abstract*—The neural system that solves a problem of fuzzy clustering of distorted observations based on optimal expansion strategy using partial distance is proposed in this article. To solve this problem we propose the learning algorithm based on hybrid of rule "Winner-Takes-More" using modified self-organizing neuro-fuzzy Kohonen network. This modified system is characterized by basic characteristics, such as: high speed, simple numerical realization, processing of distorted information in online mode.**

Keywords— Kohonen self-organizing network, fuzzy clustering;, incomplete observations with gaps, partial distance, optimal expansion

I. Introduction

The problem of data sets clustering often occurs in many practical tasks, and for its solution has been successfully used artificial neural networks [1] and methods of fuzzy systems [2]. It is usually assumed that original array is specified a priori and processing is realized in batch mode. Currently, due to the widespread using of Dynamic Data Mining [3], which is associated with the processing of observations arriving sequentially, sometimes at a high frequency, the known methods of distorted data clustering become incompetent.

More effective in situation when the data are fed to the processing in on-line mode clustering is using of self-organizing Kohonen network [4], the use of which implies that the original vector data contain all the components. Here we attempt to hybridize the self-organizing maps and methods of fuzzy clustering of distorted observations with missing values that is based on the optimal expansion using the partial distances and nearest prototypes [2].

II. Problem Statetment

Let's present the distorted data arriving for processing in the form of the table "object-property" as shown on Table I.

The Table 1 contains information about N feature vectors of order n $X = \{x_1, x_2, ..., x_N\} \subset R^n$, $x_k \in X, k = 1, 2, ..., N$ that arrive for processing in online mode. Result of observations clustering is the partition of initial data into m classes $(1 < m < N)$ with some level of membership $U_q(k)$, where here k-th is a feature vector to the q-th cluster $(1 \le q \le m)$.

TABLE I. THE "OBJECT-PROPERTY" TABLE WITH DISTORED OBSERVATION

	1	…	*p*	…	*j*	…	*n*
1	x_{11}	…	x_{1p}	…	x_{1j}	…	x_{1n}
…	…	…	…	…	…	…	…
i	x_{i1}	…	x_{ip}	…	x_{ij}	…	x_{in}
…	…	…	…	…	…	…	…
k	x_{k1}	…	x_{kp}	…	x_{kj}	…	x_{kn}
…	…	…	…	…	…	…	…
N	x_{N1}	…	x_{Np}	…	x_{Nj}	…	x_{Nn}

Incoming data in a first stage are standardized and centered so that all observations belong to the hypercube $[-1,1]^n$. If there is an unknown number of missing values in the vector images $\tilde{x}_k$, that form the array $\tilde{X}$, let's introduce the sub-arrays:

$$X_F = \{\tilde{x}_k \in \tilde{X} \mid \tilde{x}_k \text{ - if vector containing all components}\}$$

$$X_P = \{\tilde{x}_{ki}, 1 \le i \le n, 1 \le k \le N \mid \text{ if all values } \tilde{x}_k, \text{ available in } \tilde{X}\}$$

$$X_G = \{\tilde{x}_{ki} = ?, 1 \le i \le n, 1 \le k \le N \mid \text{ if all values } \tilde{x}_k, \text{absent in } \tilde{X}\}$$

III. Overview of Partial Distances

The choice of the distance between objects is the focal point of the investigation, and the final variant of the partitioning of objects into classes depends on it for a given partitioning algorithm. The simplest way to calculate the distances between objects in a multidimensional space is to calculate the Euclidean distances, but the Euclidean metrics (and its square) is calculated from the source, rather than from the standardized data. In this case it is suggested to use the partial distance described by formula 1

978-1-5386-2874-4/18/$31.00 ©2018 IEEE

$$D_P^2(\tilde{x}_k, w_q) = \frac{n}{\delta_{k\Sigma}} \sum_{i=1}^{n} (\tilde{x}_{ki} - w_{qi})^2 \delta_{ki} \qquad (1)$$

where w_{qi} -ith component of q -th prototype (centroid) of the corresponding cluster ($q = 1, 2, ..., m$),

$$\delta_{ki} = \begin{cases} 0 \mid \tilde{x}_{ki} \in X_G, \\ 1 \mid \tilde{x}_{ki} \in X_F, \end{cases} \qquad \delta_{k\Sigma} = \sum_{i=1}^{n} \delta_{ki} \ .$$

Easy to see that for $\tilde{x}_k \in X_F$ the partial distance (1) becomes an usual Euclidean metric. In the opposite case, the distance between $\tilde{x}_k$ and prototype w_q is estimated on the basis of the components available in the $\tilde{x}_k$ as shown in Fig.1.

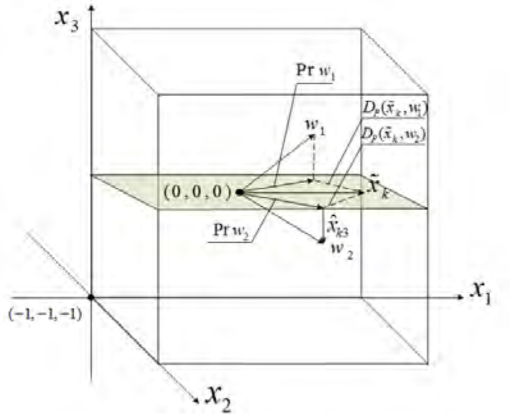


Fig. 1. The strategy of partial distances

Here, the three-dimensional vector $\tilde{x}_k$ lacks one component $\tilde{x}_{k3}$ so that the distance is measured on the plane x_1, x_2 , and instead of the prototypes w_1 and w_2 we use their projections onto this plane $\Pr w_1$ и $\Pr w_2$.

The competition process, which underlies the training of Kohonen map, is organized on the basis of estimating the partial distances, i.e. when a distorted (incomplete) vector of observations arrives $\tilde{x}_{k+1}$ first estimated distance between this vector and the centroids $w_1(k), w_2(k), ..., w_m(k)$ and then looking for the neuron-winner $w_q(k)$ such that

$$D_P^2(\tilde{x}_{k+1}, w_q(k)) = \arg\min_q \{D_P^2(\tilde{x}_{k+1}, w_1(k)), ..., \\ D_P^2(\tilde{x}_{k+1}, w_m(k))\}.$$

Further, the missing component $\tilde{x}_{k+1,i}$ is replaced by the corresponding component of the centroid of the winner neuron: $\hat{x}_{k+1,i} = w_{qi}(k)$. In Fig. 1, as missing component $\tilde{x}_{k3}$ its estimate $\hat{x}_{k3} = w_{23}$ is used. Further, the corresponding centroid is specified on the basis of the standard rule of self-learning of WTA ("The Winner Takes All") in the form

$$w_q(k+1) = w_q(k) + \eta(k+1)(\tilde{x}_{k+1} - w_q(k)) \qquad (2)$$

where $0 < \eta(k+1) < 1$ - parameter of the learning step.

To improve the quality of Kohonen networks learning, it’s possible by adjusting at each step not only the neuron-winner, but a whole group of neurons according to WTM - rule ("Winner Takes More") in the form

$$w_l(k+1) = w_l(k) + \eta(k+1)\varphi(q,l)(\tilde{x}_{k+1} - w_l(k)) \\ \forall l = 1, 2, ..., m \qquad (3)$$

where $\varphi(q,l)$ - a neighborhood function that depends on the distance between the centroid of the winner neuron w_q and arbitrary neuron w_l . As neighborhood functions, as a rule, kernel (bell-shaped) constructions with an extremum in w_q , i.e. $\varphi(q,q) = 1$ are used.

It is interesting to note that the use of Cauchian as neighborhood function

$$\varphi_l(k+1) = U_l^{\beta}(k+1) = \left(\frac{(D_P^2(\tilde{x}_{k+1}, w_l(k)))^{\frac{1}{1-\beta}}}{\sum_{r=1}^{m} (D_P^2(\tilde{x}_{k+1}, w_r(k)))^{\frac{1}{1-\beta}}} \right)^{\beta} \qquad (4)$$

(here $\beta > 1$ - the fuzzyfier), which is associated not with the winner, but with each of the prototype centroids, leads to the fact that relations (3), (4) are transformed into an adaptive algorithm for probabilistic fuzzy clustering of data with gaps [5-8] essentially FCM - clustering method for incomplete data [2]. Thus, within the WTM-rule of Kohonen self-learning network (3) it is possible to solve on-line problems of both crisp and fuzzy clustering using the standard architecture of the self-organizing network.

IV. Optimal Expansion Strategy Using Partitional Distance

Today, there are many situations when data are fed to processing sequentially as it occurs during training Kohonen self-organizing maps [4] or their modifications [9]. In this connection, an adaptive neuro-fuzzy Kohonen network is proposed, that is designed to solve the problem of clustering distorted data based on the strategy of partial distances. At the same time, in situations where the amount of distorted data is too large, the strategy of partial distances is ineffective. Thus, it may be necessary to solve the clustering problem simultaneously with restoring the gaps in the “object-property” table. In this situation, an approach based on the so-called optimal expansion strategy can be more effective. The optimal expansion strategy using the partition distances is that the elements of the submatrix X_G are considered as additional variables, which are estimated by minimizing the goal function E . Thus, in parallel with clustering, an evaluation of the gaps is made. The proposed fuzzy c-means algorithm, based on the optimal expansion strategy based on partition distances, consists of a sequence of steps [10]:

1 Step: Define the initial parameters for the algorithm: $\beta > 0$; $1 < m < N$; $\varepsilon > 0$; $w_q^{(0)}$; $1 \le q \le m$; $\tau = 0,1,2,...,Q$; $X_G^{(0)} = \{-1 \le \hat{x}_{ki}^{(0)} \le 1\}$.

2 Step: Calculation of membership levels:

$$U_q^{(\tau+1)}(k) = \underset{U_q(k)}{\arg\min}\, E(U_q(k), w_q^{(\tau)}, X_G^{(\tau)}) = \frac{(D_p^2(\hat{x}_k^{(\tau)}, w_q^{(\tau)}))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (D_p^2(\hat{x}_k^{(\tau)}, w_l^{(\tau)}))^{\frac{1}{1-\beta}}}.$$

3 Step: Calculation the cluster's centroids:

$$w_q^{(\tau+1)} = \underset{w_q}{\arg\min}\, E(U_q^{(\tau+1)}(k), w_q, X_G^{(\tau)}) = \frac{\sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta} \hat{x}_k^{(\tau)}}{\sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta}}.$$

4 Step: Stop, if $\left\| w_q^{(\tau+1)} - w_q^{(\tau)} \right\| < \varepsilon \ \forall \ 1 \le q \le m$ or $\tau = Q$

else go to step 5.

5 Step: Estimation of gaps:

$$\hat{x}_{ki}^{(\tau+1)} = \frac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k))^{\beta} w_{qi}^{(\tau+1)}}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k))^{\beta}}.$$

The processing of information using the optimal extension algorithm is organized in the form of sequences of recalculations:

$$w_q^{(0)} \to U_q^{(1)} \to \hat{x}_{ki}^{(1)} \to w_q^{(1)} \to U_q^{(2)} \to ...$$
$$\to w_q^{(\tau)} \to U_q^{(\tau+1)} \to \hat{x}_{ki}^{(\tau+1)} \to w_q^{(\tau+1)} \to ... \to w_q^{(Q)}.$$

Thus, the clustering algorithm can be rewritten in online mode

$$\left\{ \begin{array}{l} U_q^{(\tau+1)}(k+1) = \dfrac{(\left\| \hat{x}_{k+1}^{(\tau)} - w_q(k) \right\|^2)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (\left\| \hat{x}_{k+1}^{(\tau)} - w_l(k) \right\|^2)^{\frac{1}{1-\beta}}}, \\ \hat{x}_{k+1,i}^{(\tau+1)} = \dfrac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}(k)}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta}}, \\ w_q(k+1) = w_q(k) + \eta(k+1)(U_q^{(Q)}(k+1))^{\beta} * \\ \qquad\qquad * (\hat{x}_{k+1}^{(Q)} - w_q(k)). \end{array} \right. \quad (5)$$

The centroids of clusters can be recalculated in accelerated time:

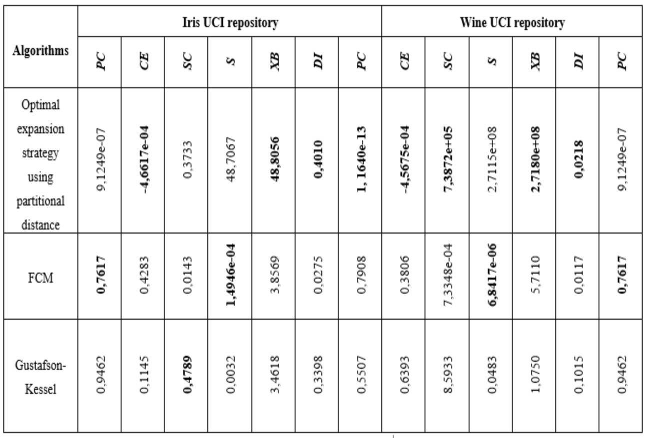
$$\left\{ \begin{array}{l} U_q^{(\tau+1)}(k+1) = \dfrac{(\left\| \hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1) \right\|^2)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (\left\| \hat{x}_{k+1}^{(\tau)} - w_l^{(\tau)}(k) \right\|^2)^{\frac{1}{1-\beta}}}, \\ w_q^{(0)}(k+1) = w_q^{(Q)}(k), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1) * \\ *(U_q^{(\tau+1)}(k+1))^{\beta} (\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)), \\ \hat{x}_{k+1,i}^{(\tau+1)} = \dfrac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}^{(\tau+1)}(k+1)}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta}}. \end{array} \right. \quad (6)$$

V. EXPERIMENTAL RESEACH

In experimental studies, the three main algorithms for clustering: FCM, Gustafson - Kessel and the proposed clustering algorithm based on the optimal expansion using partial distances were compared by the main clustering parameters: Classification Entropy (CE), Partition Coefficient (PC), Separation Index (S), Partition Index (SC), Dunn's Index (DI), Xie and Beni's Index (XB). We operated on data provided by the UCI repository data: Wine data set and Iris dataset. Each of the data sets has a certain number of observations with its attributes. For example, the Wine data set contains the results of a chemical analysis of three types of wines from different regions of Italy.

Table II and Table III show the results of clustering algorithms with different amounts of data distorted by gaps.

TABLE II. RESULTS OF EXPERIMENTS WITH 10 GAPS



Algorithms	Iris UCI repository							Wine UCI repository					
	PC	*CE*	*SC*	*S*	*XB*	*DI*	*PC*	*CE*	*SC*	*S*	*XB*	*DI*	*PC*
Optimal expansion strategy using partitional distance	9,1249e-07	**-4,6617e-04**	0,3733	48,7067	**48,8056**	**0,4010**	**1, 1640e-13**	**-4,5675e-04**	**7,3872e+05**	2,7115e+08	**2,7180e+08**	**0,0218**	9,1249e-07
FCM	**0,7617**	0,4283	0,0143	**1,4946e-04**	3,8569	0,0275	0,7908	0,3806	7,3348e-04	**6,8417e-06**	5,7110	0,0117	**0,7617**
Gustafson-Kessel	0,9462	0,1145	**0,4789**	0,0032	3,4618	0,3398	0,5507	0,6393	8,5933	0,0483	1,0750	0,1015	0,9462

As you can see from the obtained results of the algorithms, the proposed clustering method for many parameters of the data clustering quality is not inferior to the well-known algorithms and demonstrates quite good results of clustering data in online mode.

TABLE III. RESULTS OF EXPERIMENTS WITH 50 GAPS

Algorithms	**Iris UCI repository**							**Wine UCI repository**						
	PC	***CE***	***SC***	***S***	***XB***	***DI***	***PC***	***CE***	***SC***	***S***	***XB***	***DI***	***PC***
Optimal expansion strategy using partitional distance	9,1249e-07	**-4,6617e-04**	0,3775	48,7067	**48,8301**	**0,3365**	**7,5181e-12**	**-5,4992e-04**	**1,1834e+04**	4,1981e+06	**4,2024e+06**	**0,0240**	9,1249e-07
FCM	**0,7399**	0,4632	0,0174	**1,8345e-04**	4,4887	0,0355	0,7892	0,3838	7,6110e-04	**7,1760e-06**	8,8618	0,0237	**0,7399**
Gustafson-Kessel	0,9422	0,1177	**0,5219**	0,0035	3,4413	0,3341	0,5824	0,6010	4,7678	0,0268	1,1703	0,1030	0,9422

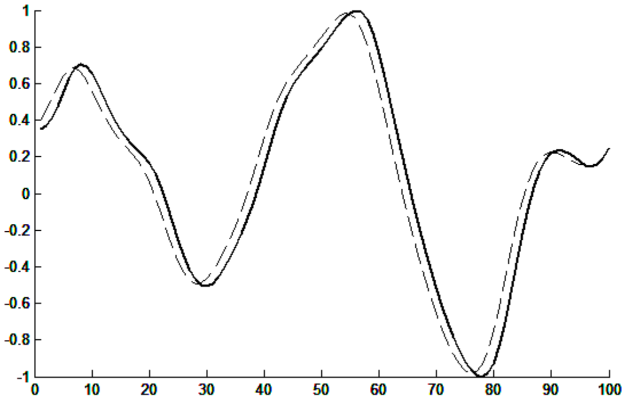


Fig. 2. Graph of estimates of gaps (dashed line) and real data (solid line)

Analyzing the results, we have plotted the recovered and original data. On the Fig.2 and Fig 3 Figures 1 and 2 show some of the results obtained, which demonstrates the work of the proposed algorithm. As can be seen from the graphs, the proposed algorithm well solves the problem.

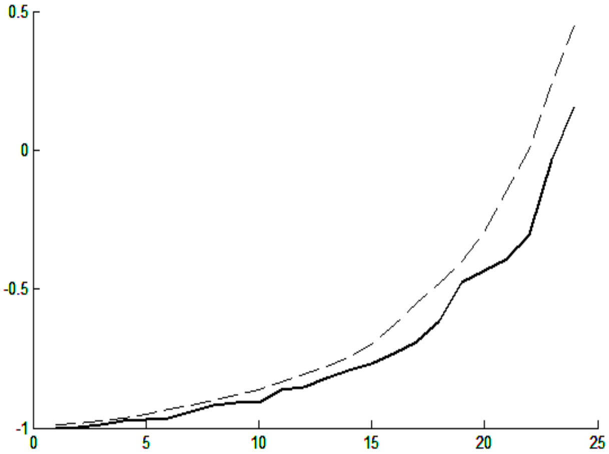


Fig. 3. Graph of estimates of gaps (dashed line) and real data (solid line)

VI. CONCLUSION

The neural systems that solves a problem of fuzzy clustering of distorted observations based on optimal expansion using partial distances is proposed in this article. To solve this problem we have proposed the learning algorithm based on hybrid of rule “Winner-Takes-More” using modified self-organizing neuro-fuzzy Kohonen network. This modified system is characterized by basic characteristic, such as: high rate, simple numerical realization, processing of distorted information in online mode.

REFERENCES

[1] T Marwala, “Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques,” Hershey-New York: Information Science Reference, 2009.

[2] J. C. Bezdek, “Pattern Recognition with Fuzzy Objective Function Algorithms,” Plenum Press, New York, 1981.

[3] E. Lughofer, “Evolving Fuzzy Systems. Methodologies, Advanced Concepts and Applications,” Berlin-Hagenberg, 2011.

[4] T. Kohonen, “Self-Organizing Maps,” Berlin: Springer-Verlag, 1995.

[5] A. Y. Shafronenko, V. V. Volkova, and Ye. Bodyanskiy, “Adaptive clustering data with gaps,” Radioelectronics, informatics, control, no. 2. pp. 115-119, 2011. (in Russian)

[6] Ye. Bodyanskiy, A. Shafronenko, and V. Volkova, “Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network. Artificial Intelligence Methods and Techniques for Business and Engineering Applications,” ITHEA, Rzeszow, Poland; Sofia, Bulgaria. pp. 287-296, 2012.

[7] Ye. Bodyanskiy, A. Shafronenko, and V. Volkova, “Adaptive fuzzy probabilistic clustering of incomplete data,” Int. J. “Information, models and analyses”, vol.2, no. 2, pp. 112-117, 2013.

[8] Ye. Bodyanskiy, A. Shafronenko, and V. Volkova, “Neuro fuzzy Kohonen network for incomplete data clustering using optimal completion strategy,” Proceedings 20th East West Fuzzy Colloquium 2013, Zittau, pp. 214-223, 25-27 September 2013.

[9] V. Kolodyazhniy, Ye. Bodyanskiy and Ye. Gorshkov, “New recursive learning algorithms for fuzzy Kohonen clustering network,” Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronc Systems, Rapperswil, Switzerland, pp. 58-61,, June 21-24, 2009.

[10] R. J. Hathaway, and J. C Bezdek, “Fuzzy c-means clustering of incomplete data,”. IEEE Trans. on Systems, Man, and Cybernetics, vol. 31, no. 5, pp. 735-744, 2001.