ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 1 (7)

UDC 004.032.26 DOI: https://doi.org/10.30837/2522-9818.2019.7.016

YE. BODYANSKIY, I. PEROVA, P. ZHERNOVA

ONLINE FUZZY CLUSTERING OF HIGH DIMENSION DATA STREAMS BASED ON NEURAL NETWORK ENSEMBLES

The **subject** matter of the article is fuzzy clustering of high-dimensional data based on the ensemble approach, provided that a number and shape of clusters are not known. The **goal** of the work is to create the neuro-fuzzy approach for clustering data when the data stream is fed for online processing and a number and shape of clusters are unknown. The following **tasks** are solved in the article – the input feature space is compressed in the online mode; the model of neural network ensembles for data clustering is built; the ensemble of neuro-fuzzy networks for clustering high-dimensional data is developed; the approach for clustering data in the online mode is worked out. The following **results** are obtained – the main idea of the proposed approach is based on a modification of the fuzzy C-means algorithm. To reduce the dimension of the input space, the modified Hebb-Sanger network is suggested to be used; this net is characterized by the increased speed and is built on the basis of the modified Oja neurons. A speed-optimized learning algorithm for the Oja neuron is proposed. Such a network implements the method of principal components in the online mode with high speed. **Conclusions.** In the event the reduction-compression procedure cannot be used due to the probability of losing the physical meaning of the original space, a new clustering criterion was introduced; this criterion contains both a well-known polynomial fuzzifier and the weighment of individual components of the deviations of presented images from cluster centroids. The recurrent modification based on the algorithms proposed in this article is introduced. A mathematical model is developed to determine the quality of clustering with the use of the Xi-Beni index, which was modified for the online mode. The experimental results confirm the fact that the proposed system enables solving a wide range of Data Mining tasks when data sets are processed online, provided that a number and shape of clusters are not known and there is a large number of observations as well.

Keywords: clustering; fuzzy C-means method; sequential analysis of principal components; the ensemble of neuro-fuzzy networks; T. Kohonen's neural network; self-learning.

Introduction

The task of multidimensional observations clustering when observations are sequentially fed to processing is an important area within Data Stream Mining, and for its solution a sufficiently large number of different methods have been proposed. The most popular approaches here are based on prototypes-centroids [1–4], in which K-means, K-medians, K-medoids, etc. can be used. It should be noted that clustering neural networks of T. Kohonen [5], are the best suited for processing information in the online mode. In this case, apriori it is assumed that the number of clusters into which the analyzed data array has to be divided is known in advance. If the number of clusters is not known apriori, the X-means method [6, 7], which is based on rather strict statistical assumptions, can be used. In addition, this method can be implemented only in batch mode. If the information for processing is received sequentially, the alternative based on clustering ensembles [8–11] can be used as an X-means, with each of the members of the ensemble being designed for a different number of possible clusters. If T. Kohonen's neural networks (SOM) can be members of the ensemble, each of which operates in conditions of a different number of classes in the data, such a system can operate effectively in real time.

Thus, if data sample is a set (possibly growing) $X = \{x(1),...,x(2),...,x(k),...,x(N),...\} \subset R^n$, $x(k) = (x_1(k),...,x_i(k),...,x_n(k))^T$, which is fed to the inputs of an ensemble formed by $M-1$ parallel-connected SOMs so that the first one works in conditions that the number of possible clusters in the data is $m = 2$, and the last assumes that $m = M$, the best results will be obtained using SOM with $2 \le m^* \le M$ neurons in the Kohonen layer, where m^* determines the true number of classes in the sample under processing [12].

The situation becomes much more complicated if the formed clusters overlap in the features space. Such problems are solved using fuzzy clustering methods [4, 13], the most popular of which is the fuzzy C-means algorithm (FCM). Fuzzy Kohonen`s clustering networks [14] can be successfully used to work in online mode.

It should be remembered that the effectiveness of fuzzy clustering procedures is limited by the so-called concentration of norms effect – CoN [15, 16], when the results are unsatisfactory at high dimensions of the features space. The simplest approach to deal the problem of high-dimension feature space is preliminary data compression in online-mode.

Online data compression for reduction of initial feature space

When dataset is fed to processing in the form of data stream, Principal Component Analysis cannot be used for reduction of initial feature space, so solving of data reduction task can be performed using neural network technologies [17–21]. A neural network based on Oja's neuron [17] and constructed on its base T.Sanger's neural network [18] are the most popular systems that permit to perform the data reduction and information compressing sequentially in online mode.

The neural network based on Oja's neuron permits to calculate in sequential mode eigen vectors of correlation matrix $R(k)$ when dataset in fed to processing sequentially $x(1), x(2),..., x(k), x(k+1)$ without calculating of full correlation matrix. To find first principal component E. Oja proposed self-learning

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 1 (7) ISSN 2524-2296 (online)

algorithm for linear neuron presented in fig. 1. For previously centered data Oja's algorithm can be written in the form:

$$\begin{cases} w^1(k+1) = w^1(k) + \\ +\eta(k+1)\left(\tilde{x}(k+1) - \varphi^1(k+1)w^1(k)\right)\varphi^1(k+1); \\ \varphi^1(k+1) = \tilde{x}^T(k+1)w^1(k), \quad w^1(0) \neq 0, \\ \varphi^1(1) = \tilde{x}^T(1)w^1(0) \end{cases} \quad (1)$$

where $\eta(k+1)$ – learning rate parameter, its value must be chosen sufficiently small for stable algorithms work and needs to correspond to condition of stochastic approximation [22].

In such case neural network based on usual Oja's neurons characterized by low rate of convergency, that's why we have used modificated form of Sanger's neural network [19]. This system is characterized by high speed of synaptic weights tuning. In fig. 2 modified form of Sanger's neural network is presented.

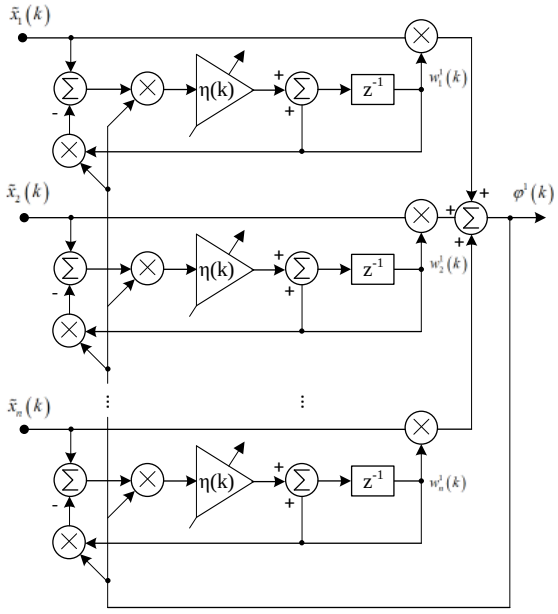

Fig. 1. Oja's neuron

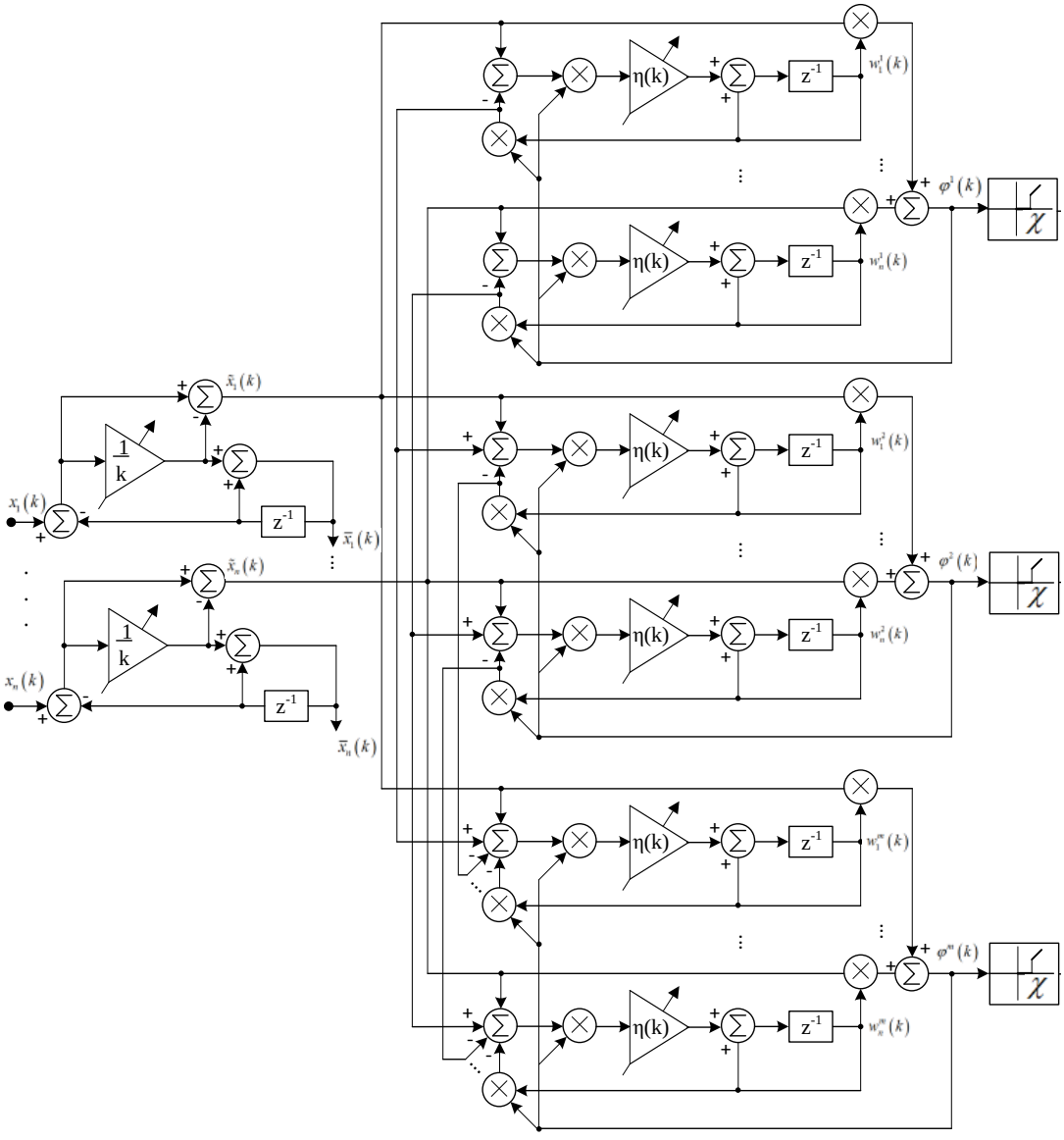

Fig. 2. Architecture of modified form of Sanger's neural network

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 1 (7)

To learn presented system we can use modified Hebb-Sanger algorithm in the form:

$$\begin{cases} w^l(k+1) = w^l(k) + \eta(k+1)\tilde{v}^l(k+1)\varphi^l(k+1); \\ \tilde{v}^l(k+1) = \tilde{v}^{l-1}(k+1) - \varphi^l(k+1)w^l(k); \\ \tilde{v}^0(k+1) = \tilde{x}(k+1), \quad l = 1,2,...,m; \end{cases} \quad (2)$$

$$\eta(k+1) = r^{-1}(k+1),$$

$$r(k+1) = \xi r(k) + \left\|\tilde{x}(k+1)\right\|^2, \quad 0 \le \xi \le 1,$$

where ξ – learning rate parameter.

It is easy to see, that first principal component is computed using Oja's algorithm, than projections of input vectors on w^1 subtract from inputs and all other data are processed by second neuron.

Previously all input data need to be centered upon mean in recurrent form using expression

$$\tilde{x}(k+1) = x(k+1) - \bar{x}(k+1),$$

$$\bar{x}(k+1) = \bar{x}(k) + \frac{1}{k+1}\left(x(k+1) - \bar{x}(k)\right).$$

Then signals $\tilde{x}(k)$ were processed by ensemble of m Oja's neurons. Output layer, based on linear elements with insensitivity zone χ permit to separate the most informative signal $\varphi^l(k)$ from noise.

Introduced neural network system permits to produce data reduction in online mode when data are fed to processing sequentially.

As we marked, algorithm based on stochastic approximation is characterized by low speed of convergency. That's why we have proposed to train Oja's neuron by modified procedure with tuned η parameter:

$$\begin{cases} w^1(k+1) = w^1(k) + \\ +\eta(k+1)\left(\tilde{x}(k+1) - \varphi^1(k+1)w^1(k)\right)\varphi^1(k+1); \\ \varphi^1(k+1) = \tilde{x}^T(k+1)w^1(k), \quad w^1(0) \ne 0, \\ \qquad\qquad \varphi^1(1) = \tilde{x}^T(1)w^1(0), \\ \eta(k+1) = r^{-1}(k+1), \\ r(k+1) = \xi r(k) + \left\|\tilde{x}(k+1)\right\|^2, \quad 0 \le \xi \le 1. \end{cases} \quad (3)$$

In fig. 3 modified Oja's neuron is presented. This system is characterized by only one tuning parameter ξ that permits to realize learning procedure in simple form.

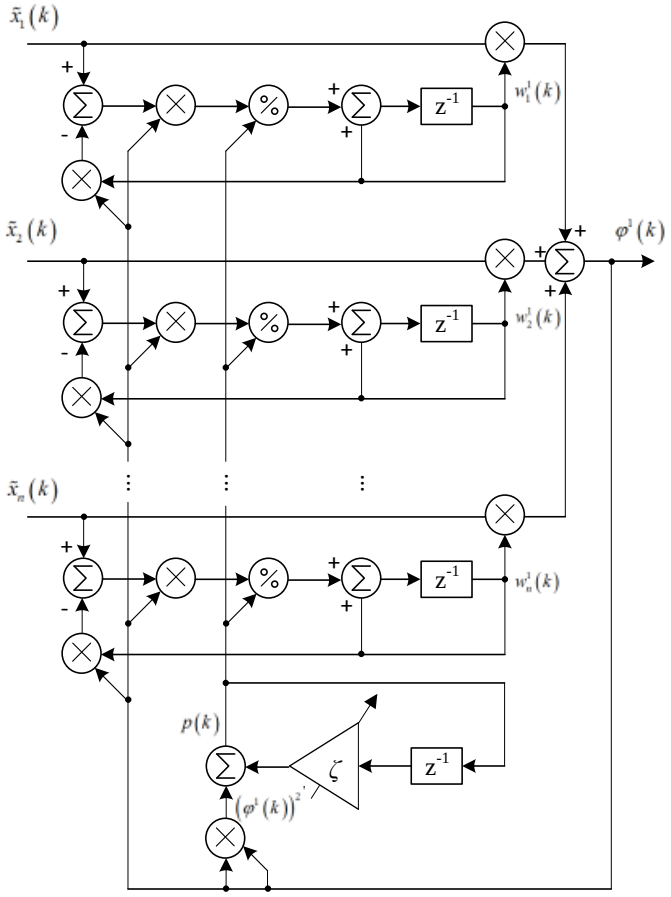

Fig. 3. Modified Oja's neuron

Using modified Oja's neuron we can to introduce modified Sanger neural network similar to previous combination, presented on fig. 1 and fig. 2. Modified Sanger neural network can be trained by the expression:

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 1 (7) ISSN 2524-2296 (online)

$$\begin{cases} w^l(k+1) = w^l(k) + \eta(k+1)\tilde{v}^l(k+1)\varphi^l(k+1); \\ \tilde{v}^l(k+1) = \tilde{v}^{l-1}(k+1) - \varphi^l(k+1)w^l(k); \\ \tilde{v}^0(k+1) = \tilde{x}(k+1), \quad l = 1,2,...,m; \\ \eta(k+1) = r^{-1}(k+1); \\ r(k+1) = \xi r(k) + \left\|\tilde{x}(k+1)\right\|^2, \quad 0 \le \xi \le 1. \end{cases} \tag{4}$$

However, it exist any situations when compression-reduction is impossible, since the physical meaning of the feature is lost. In such situations, it is proposed to perform clustering without compression.

In this regard, it seems appropriate to develop an online method for fuzzy clustering of high-dimensional data based on clustering ensembles under conditions of unknown number of classes in the stream of processed information.

Fuzzy clustering T. Kohonen`s neural network for high dimensional data stream processing

In the class of fuzzy clustering procedures from a mathematical point of view, the most rigorous are algorithms based on goal functions [4] and solving the problem of their optimization in the presence of certain constraints. Here the most popular is a fuzzy clustering probabilistic algorithm based on optimization of the goal function

$$E\left(u_j(k), c_j\right) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k)\left\|x(k) - c_j\right\|^2 = \\ = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k)\sum_{i=1}^{n}\left(x_i(k) - c_{ji}\right)^2, \tag{5}$$

subject to constraints

$$\sum_{j=1}^{m} u_j(k) = 1, \tag{6}$$

$$0 \le \sum_{k=1}^{N} u_j(k) \le N\,. \tag{7}$$

Here $u_j(k) \in [0,1]$ – the level of fuzzy membership of the observation $x(k)$ to the *j*-th cluster, c_j – centroid of *j*-th cluster, β – fuzzifier, which determines the blurring of the boundaries between clusters.

The solution of the optimization problem (5) in the presence of constraints (6), (7) with the help of Lagrange uncertain multipliers leads to the will known result

$$\begin{cases} u_j(k) = \dfrac{\left(\left\|x(k) - c_j\right\|^2\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(\left\|x(k) - c_l\right\|^2\right)^{\frac{1}{1-\beta}}}, \\ c_j = \dfrac{\sum\limits_{k=1}^{N} u_j^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N} u_j^{\beta}(k)} \end{cases} \tag{8}$$

which for $\beta = 2$ completely coincides with FCM of J. Bezdek.

The probabilistic algorithm of fuzzy clustering (8) is widely used in Data Mining, however, it loses its effectiveness in data processing tasks of high dimensionality due to the resulting effect of concentration of norms [23]. To overcome this drawback, in [15] it was proposed to use the so-called polynomial fuzzifier and a procedure known as fuzzy C-means with polynomial fuzzifier (PFCM). An adaptive online version of the PFCM was introduced in [24] for solving the tasks of Data Stream Mining.

In [25], for solving problems of data of high dimensionality clustering, the modification of FCM was proposed with weighting each of the features $x_i(k)$ that form the vector-pattern $x(k) \in R^n$, $i = 1,2,...,n$.

By combining these two approaches, let's introduce into consideration the goal function of fuzzy clustering of the form

$$E\left(u_j(k), c_j, \alpha, \gamma_{ji}\right) = \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\left\|x(k) - c_j\right\|_{\Gamma_j^2}^2 = \\ = \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\sum_{i=1}^{n}\gamma_{ji}^2\left(x_i(k) - c_{ji}\right)^2, \tag{9}$$

with constraints (6), (7) and additional constraints

$$\sum_{i=1}^{n}\gamma_{ji} = Tr\Gamma_j = 1 \quad \forall\, j = 1,2,...m\,. \tag{10}$$

Here $0 < \alpha \le 1$ – polynomial fuzzifier, γ_{ji} – weight of *i*-th attribute in *j*-cluster, $\Gamma_j = diag\left(\gamma_{j1}, \gamma_{j2}, ..., \gamma_{jm}\right)$.

Optimization of the goal function (9) under the constraints (6), (7), (10) using the uncertain Lagrange multipliers leads to a result which is a generalization of (8) and coincides with it with $\alpha = 1$, $\gamma_{ji} = m^{-1}$.

$$\begin{cases} u_j(k) = \dfrac{\alpha - 1}{2\alpha} + \dfrac{1 - m\dfrac{\alpha - 1}{2\alpha}}{\sum\limits_{l=1}^{m}\dfrac{\left\|x(k) - w_j\right\|_{\Gamma_j^2}^2}{\left\|x(k) - w_l\right\|_{\Gamma_l^2}^2}}, \\ \gamma_{ji} = \left(\sum\limits_{h=1}^{n}\left(\dfrac{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\left(x_i(k) - c_{ji}\right)^2}{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\left(x_h(k) - c_{ji}\right)^2}\right)\right)^{-1}, \\ w_{ji} = \dfrac{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\gamma_{ji}^2 x_i(k)}{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\gamma_{ji}^2}. \end{cases} \tag{11}$$

The last expression of (11) for calculating centroids of clusters can be rewritten in recurrent form.

$$c_j(k) = c_j(k-1) + \theta(k)\left(\alpha u_j^2(k-1) + (1-\alpha)u_j(k-1)\right)\Gamma_j^2(k-1)\left(x(k) - c_j(k-1)\right), \tag{12}$$

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 1 (7)

which essentially coincides with the self-learning WTM-rule by T. Kohonen [5], where the factor $\left(\alpha u_j^2\ (k-1)+(1-\alpha)u_j(k-1)\right)\Gamma_j^2(k-1)$ describes the neighborhood function and $0<\theta(k)<1$ is the learning rate parameter.

Thus, the process of high-dimensional data clustering (11), (12) is conveniently implemented using the architecture shown in fig. 4 which is a modification of the neuro-fuzzy network of T. Kohonen [26, 27].

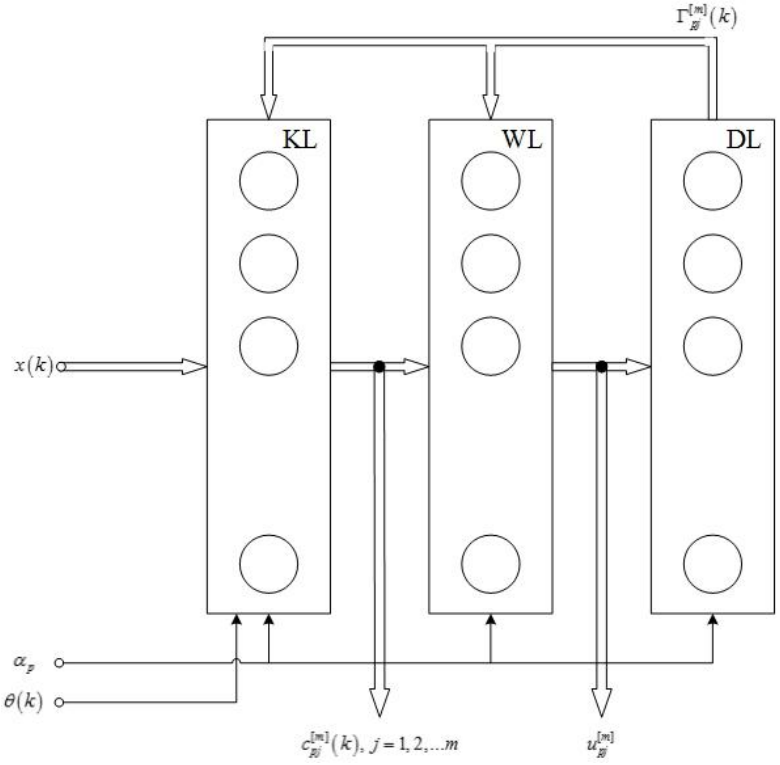

Fig. 4. Adaptive neuro-fuzzy Kohonen network – $FSOM_p^{[m]}$

Here, the first hidden kernel layer (KL) is essentially a standard SOM neural network [5], which contains m-neurons in the Kohonen layer, whose synaptic weights-centroids are tuned using the WTM learning rule (12), in the second hidden layer ML, the membership levels of k-th observation to j-th cluster $u_j(k)$ using the first relation (11) are estemated, and in the output layer WL weights γ_{ji} are calculated using the second relation of (11).

The values of learning rate parameter $\theta(k)$ and the polynomial fuzzifier α_p from a certain apriori given set $0<\alpha_1,\alpha_2,...,\alpha_p,...\alpha_q=1$ are fed to the additional inputs of the network.

Clustering Ensemble Architecture

To solve the problem of clustering in conditions where the number of clusters is unknown, we propose to use an ensemble of clustering neuro-fuzzy Kohonen`s networks, whose architecture is shown in fig. 5. This ensemble contains $(M-1)_q$ $FSOM_p^{[m]}$, where index $[m]$ means the number of clusters into which this network splits the sample to be processed – i.e. the number of neurons in the Kohonen`s layer KL, and p – is the index of a specific fuzzifier, taking q values. All elements are tuned using the same type of procedures (11), (12), which differ from each other only in the values of m and α.

In blocks $MEXB_p^{[m]}$, the quality of clustering provided by a particular FSOM is evaluated, and the output layer of the DM ensemble selects the best from $(M-1)q$ results of the previous layers, i.e. the number of clusters m* in the processed data, the centroids of the formed clusters $c_1^*, c_2^*,...,c_{m*}^*$ and the levels of each observation $u_1^*(k),u_2^*(k),...,u_{m*}^*(k)$ to the corresponding cluster membership.

To estimate the quality of clustering, each of the elements of the ensemble can be used in any of the fuzzy clustering indexes [2], where one of the most popular is the Xie-Beni index [28], which for the FCM procedure in the case of m clusters can be written in the form

$$XB^{[m]}=\frac{\left(\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^2(k)\left\|x(k)-c_j\right\|^2\right)\Big/N}{\min\limits_{l\neq j}\left\|c_j-c_l\right\|^2}=\frac{NXB^{[m]}}{DXB^{[m]}}. \quad (13)$$

For online processing, it is possible to enter the recurrent version of the XB-index in the form

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 1 (7) ISSN 2524-2296 (online)

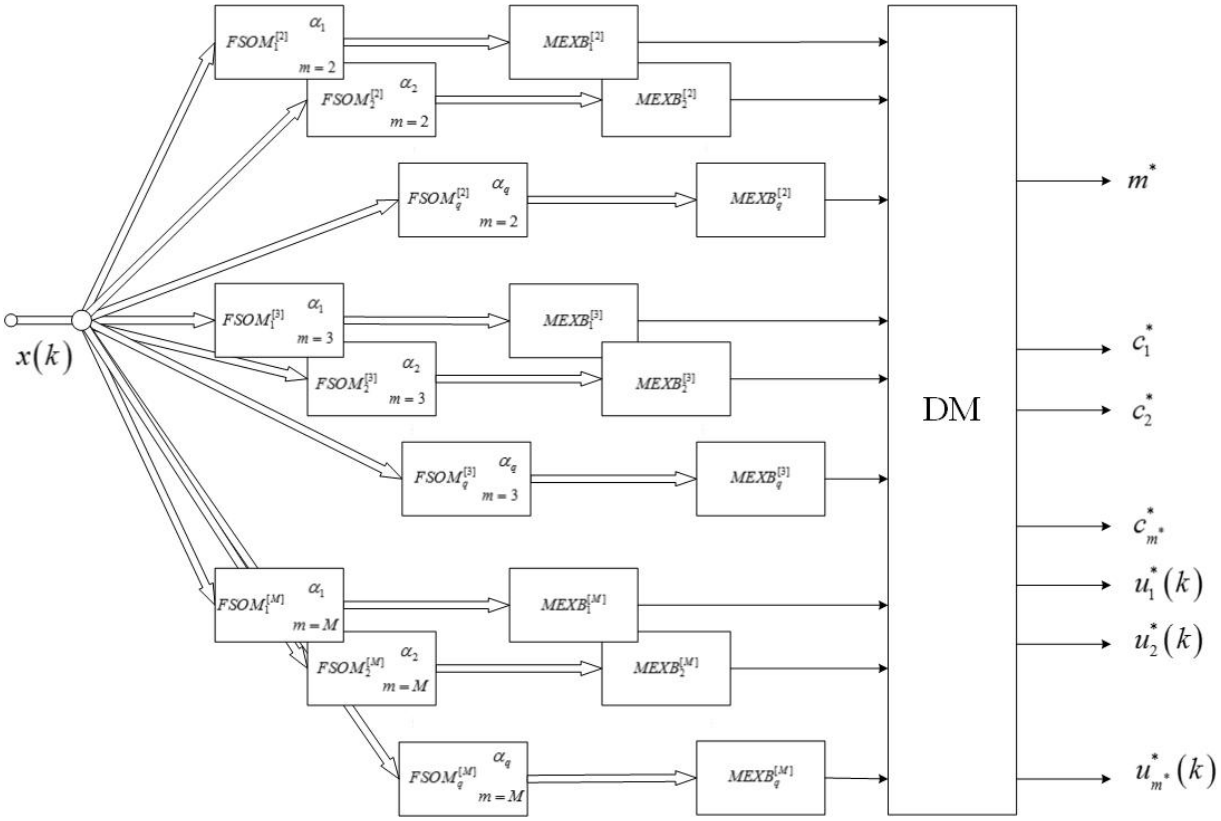

Fig. 5. The ensemble for online fuzzy clustering of high-dimensional data

$$XB^{[m]}(k)=\frac{NXB^{[m]}(k)}{DXB^{[m]}(k)}=\frac{NXB^{[m]}(k-1)+\frac{1}{k}\left(\sum\limits_{j=1}^{m}u_j^2(k)\left\|x(k)-c_j(k)\right\|^2-NXB^{[m]}(k-1)\right)}{\min\limits_{l\neq j}\left\|c_j(k)-c_l(k)\right\|^2}. \qquad (14)$$

The smaller is the value (13), (14), the higher is the quality of clustering. For procedure (8) the extended Xie-Beni index can be used [29]

$$EXB^{[m]}=\frac{\sum\limits_{k=1}^{N}\sum\limits_{j=1}^{m}u_j^{\beta}(k)\left\|x(k)-c_j\right\|^2\Big/N}{\min\limits_{l\neq j}\left\|c_j-c_l\right\|^2}, \qquad (15)$$

or its online version

$$EXB^{[m]}(k)=\frac{NEXB^{[m]}(k)}{DEXB^{[m]}(k)}=$$
$$=\frac{NEXB^{[m]}(k-1)+\frac{1}{k}\left(\sum\limits_{j=1}^{m}u_j^{\beta}(k)\left\|x(k)-c_j(k)\right\|^2-NEXB^{[m]}(k-1)\right)}{\min\limits_{l\neq j}\left\|c_j(k)-c_l(k)\right\|^2}. \qquad (16)$$

By analogy with (15), (16), we can to introduce modification EXB-index for the goal function (9)

$$MEXB_p^{[m]}=\frac{\left(\sum\limits_{k=1}^{N}\sum\limits_{j=1}^{m}\left(\alpha_p\left(u_{pj}^{[m]}(k)\right)^2+\left(1-\alpha_p\right)u_{pj}^{[m]}(k)\right)\left\|x(k)-c_{pj}^{[m]}\right\|^2_{\left(\Gamma_{pj}^{[m]}\right)^2}\right)\Big/N}{\min\limits_{l\neq j}\left\|c_{pj}^{[m]}-c_{pl}^{[m]}\right\|^2}=\frac{NMEXB_p^{[m]}}{DMEXB_p^{[m]}}, \qquad (17)$$

or its online version

$$MEXB_p^{[m]}(k)=\frac{NMEXB_p^{[m]}(k)}{DMEXB_p^{[m]}(k)}=$$
$$=\frac{NMEXB_p^{[m]}(k-1)+\frac{1}{k}\left(\sum\limits_{j=1}^{m}\left(\alpha_p\left(u_{pj}^{[m]}(k)\right)^2+(1-\alpha)u_{pj}^{[m]}(k)\right)\left\|x(k)-c_{pj}^{[m]}(k)\right\|^2_{\left(\Gamma_{pj}^{[m]}(k)\right)^2}-NMEXB_p^{[m]}(k-1)\right)}{\min\limits_{l\neq j}\left\|c_{pj}^{[m]}(k)-c_{pl}^{[m]}(k)\right\|^2}. \qquad (18)$$

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 1 (7)

In the course of data processing, the decision making (DM) unit finds $FSOM_p^{[m^*]}$ with the best value of $MEXB_p^{[m^*]}$ and the results of work of this particular neuro-fuzzy network determines the final result of clustering.

Experiments.

To solve the problem of determination the optimal number of clusters in datasets we have used proposed ensemble for online fuzzy clustering. We have choose Dermatology dataset [30] from UCI Machine Repository. It contains 366 instances, but some of attributes has missed values, so they need to be filled using, for example, fuzzy spatial extrapolation approach from [31]. Number of attributes is 34, 35-th attribute is class-diagnosis. Value 1 is interpreted as psoriasis (112 instances); value 2 – seboreic dermatitis (61 instances); value 3 – lichen planus (72 instances); value 4 – pityriasis rosea (49 instances); value 5 – cronic dermatitis (52 instances); value 6 – pityriasis rubra pilaris (20 instances).

It is easy to see that number of features is close to number of instances and we can talking about high-dimensional data.

In fig. 6 visualization using principal component analysis (PCA-analysis three principal components) is presented.

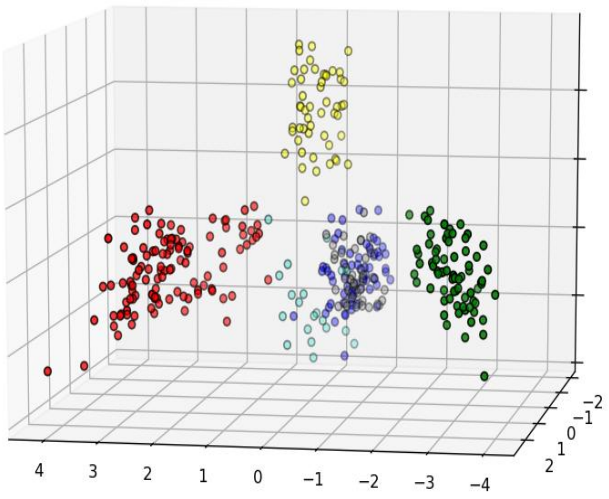

Fig. 6. The visualization of Dermatology dataset

In table 1 values of Xie-Beni indexes for different number of clusters are presented. Minimal value correspond to MHDFCM when parameter α is equal to 0,5 or 1. K-means algorithm shous very high values of Xie-Beni index, so we can talk about not effective clusterization results. Fuzzy c-means algorithm can not process proposed dataset because of CoN (concernatration of norm).

Table 1. *Values of parameter α and Xie-Beni index for Dermatology dataset*

Algorithms type	Xie-Beni	Number of Clusters
K-means	1,3524 x 1024	2
MHDFCM α = 0,5	0,00009	
MHDFCM α = 1,0	0,00006	
FCM	CoN	
K-means	9,401 x 1025	3
MHDFCM α = 0,5	0,00004	
MHDFCM α = 1,0	0,00002	
FCM	CoN	
K-means	2,63023 x 1025	4
MHDFCM α = 0,5	49266	
MHDFCM α = 1,0	1507297	
FCM	CoN	
K-means	4,5374 x 1025	5
MHDFCM α = 0,5	50909	
MHDFCM α = 1,0	404	
FCM	CoN	
K-means	1,6282 x 1026	6
MHDFCM α = 0,5	1244	
MHDFCM α = 1,0	1973697	
FCM	CoN	
K-means	7,3338 x 1025	7
MHDFCM α = 0,5	37252	
MHDFCM α = 1,0	1929	
FCM	CoN	
K-means	9,499 x 1025	8
MHDFCM α = 0,5	240	
MHDFCM α = 1,0	48135	
FCM	CoN	

Conclusions

The architecture and algorithm of the neuro-fuzzy self-learning system is proposed to solve the problem of online clustering of a high-dimensional data stream in conditions where the formed clusters can overlap and their number is not known apriori. The system under consideration is an ensemble of neuro-fuzzy T. Kohonen`s self-organizing maps, each of them differs from the others in the number of neurons and the value of the polynomial fuzzifier. Each member of the ensemble is tuned using the modified WTM self-learning rule, while in the process of tuning all components of the processed vectors are automatically weighed.

The proposed approach is a generalization of a number of known fuzzy probabilistic clustering procedures and can be used to solve Data Stream Mining tasks in online mode.

Experimental results on Dermatology datasets from UCI repository affect the performance of the proposed modified high-dimensional fuzzy c-means approach and ensemble for online fuzzy clustering of high-dimensional data streams based on it.

References

1. Gan, G., Ma, Ch., Wu, J. (2007), *Data Clustering. Theory, Algorithms and Application*, SIAM, Philadelphia, 489 p.
2. Xu, R., Wunsch, D. C. (2009), *Clustering*, IEEE Press Series on Computational Intelligence, John Wiley & Sons, Inc., Hoboken, NJ, 368 p.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 1 (7) ISSN 2524-2296 (online)

3. Bifet, A. (2010), *Adaptive Stream Mining. Pattern Learning and Mining from Evolving Data Streams*, Amsterdam, IOS Press, 224 p.
4. Bezdek, J. C. (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, N.Y., Plenum Press, 272 p
5. Kohonen, T. (1995), *Self-Organizing Maps*, Springer-Verlag, Berlin, 362 p.
6. Pelleg, D., Moor, A. (2000), "X-means: extending K-means with efficient estimation of the number of clusters", *Proc. 17th Int. Conf. on Machine Learning, Morgan Kaufmann, San Francisco*, P. 727–730.
7. Ishioka, T. (2005), "An expansion of X-means for automatically determining the optimal number of clusters", *Proc. 4th IASTED Int. Conf. Computational Intelligence*, Calgary, Alberta, P. 91–96.
8. Strehl, A., Ghosh, J. (2002), "Cluster Ensembles – A knowledge reuse framework for combining multiple partitions", *Journal of Machine Learning Research*, P. 583–617.
9. Topchy, A., Jain, A.K., Punch, W. (2005), "Clustering ensembles: models of consensus and weak partitions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 27, P. 1866–1881.
10. Alizadeh, H., Minaei-Bidgoli, B., Parvin, H. (2013), "To improve the quality of cluster ensembles by selecting a subset of base clusters", *Journal of Experimental & Theoretical Artificial Intelligence*, No. 26, P. 127–150.
11. Charkhabi, M., Dhot, T., Mojarad, S.A. (2014), "Cluster ensembles, majority vote, voter eligibility and privileged voters", *Int. Journal of Machine Learning and Computing*, No. 4, P. 275–278
12. Zhernova, P., Deyneko, A., Bodyanskiy, Ye., Riepin, V. (2018), "Adaptive kernel data streams clustering based on neural networks ensembles in conditions of uncertainty about amount and shapes of clusters", *IEEE Second International Conference on Data Stream Mining & Processing, August 21-25*, Lviv, Ukraine, P. 7–12.
13. Bezdek, J., Keller, J., Krisnapuram, R., Pal, N. (2005), *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Springer, 776 p.
14. Gorshkov, Ye., Kolodyazhniy, V., Bodyanskiy, Ye. (2009), "New recursive learning algoritms for fuzzy Kohonen clustering network", *In Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems, Rapperwil*, Switzerland, P. 58–61.
15. Höppner, F., Klawonn, F., Kruze, R. (1999), *Fuzzy Klusteranalyse*, Braunschweig, Vieweg, 280 p.
16. Höppner, F., Klawonn, F., Kruse, R. (1996), *Fuzzy-Klusteranalyse, Verfahren für die Bilderkennung, Klassifikation und Datenanalyse*, Braunschweig, Vieweg, 292 p.
17. Oja, E. (1989), "Neural Network, principal components and subspaces", *Int. J. of Neural Systems*, No. 1, P. 61–68.
18. Sanger, T. (1989), "Optimal unsupervised learning in a single-layer linear feedforward neural network", *Neural Networks*, No. 2, P. 459–473.
19. Bodyanskiy, Ye., Mihaliov, O., Pliss I. (2000), *Adaptive fault detection in control systems using artifitial neural networks*, Dniepropetrovsk : System Technologies, 140 p.
20. Überla, K. (1997), *Faktorenanalyse*, Springer Verlag, Berlin Heidelberg New York, 398 p.
21. Oja, E. (1982), "A simplified neuron model as a principal component analyzer", J. of Math. Biology, No. 15, P. 267–273.
22. Vazan, M. T. (1969), *Stochastic approximation*, Cambridge, Cambridge University Press, 289 p.
23. Shakhovska, N., Medykovsky, M., Stakhiv, P. (2013), "Application of algorithms of classification for uncertainty reduction", *Przeglad Elektrotechniczny*, No. 4, P. 284–286.
24. Kolchygin, B. V., Bodyanskiy, Ye. V. (2013), "Adaptive fuzzy clustering with a variable fuzzifier", *Cybernetics and Systems Analysis*, No. 3, P. 366–374.
25. Keller, A., Klawonn F. (2000), "Fuzzy Clustering with weighting of data variables", *Uncertainty, Fuzziness and Knowledge Based Systems*, No. 8, P. 735-746.
26. Bodyanskiy, Ye., Kolchygin, B., Pliss I. (2011), "Adaptive neuro-fuzzy Kohonen network with variable fuzzifier", *Inform. Theories and Appl*, No. 3, P. 215–223.
27. Bodyanskiy, Ye., Zhernova, P. (2018), "Kernel fuzzy data stream clustering based on neural networks ensemble", *Inovative Technologies & Scientific Solutions for Industries*, No. 4 (6), P. 42–49. DOI: https://doi.org/10.30837/2522-9818.2018.6.042.
28. Xie, X. L., Beni, G. A. (1991), "Validity Measure for Fuzzy Clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 13, P. 841–847.
29. Bodyanskiy, Ye. V., Tyshchenko, O. K., Kopaliani, D. S. (2017), "An Evolving Connectionist System for Data Stream Fuzzy Clustering and Its Online Learning", *Neurocomputing*, No. 262, P. 41–56.
30. "Dermatology dataset", available at: http://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/dermatology.data (last accessed: 1st of May, 2018).
31. Mulesa, P., Perova, I. (2015), "Fuzzy Spacial Extrapolation Method Using Manhattan Metrics for Tasks of Medical Data Mining", *Computer Science and Information Technologies*, CSIT'2015, Lviv, Ukraine, P. 104–106. DOI: https://doi.org/10.1109/STC-CSIT.2015.7325443.

Received 12.02.2019

Відомості про авторів / Сведения об авторах / About the Authors

Бодянський Євгеній Володимирович – доктор технічних наук, професор, Харківський національний університет радіоелектроніки, професор кафедри штучного інтелекту, науковий керівник ПНДЛ АСУ, Харків, Україна, e-mail: yevgeniy.bodyanskiy@nure.ua; ORCID ID: https://orcid.org/0000-0001-5418-2143.

Бодянский Евгений Владимирович – доктор технических наук, профессор, Харьковский национальный университет радиоэлектроники, профессор кафедры искусственного интеллекта, научный руководитель ПНДЛ АСУ, Харьков, Украина.

Bodyanskiy Yevgeniy – Doctor of Sciences (Engineering), Professor, Kharkiv National University of Radio Electronics, Professor at the Department of Artificial Intelligence, Scientific Head at the CSRL, Kharkiv, Ukraine.

Перова Ірина Геннадіївна – кандидат технічних наук, с.н.с., доцент, Харківський національний університет радіоелектроніки, доцент кафедри біомедичної інженерії, Харків, Україна; e-mail: rikywenok@gmail.com; ORCID ID: https://orcid.org/0000-0003-2089-5609.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 1 (7)

Перова Ирина Геннадьевна – кандидат технических наук, с.н.с., доцент, Харьковский национальный университет радиоэлектроники, доцент кафедры биомедицинской инженерии, Харьков, Украина.

Perova Iryna – PhD (Engineering Sciences), Senior Researcher, Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor at the Department of Biomedical Engineering, Kharkov, Ukraine.

Жернова Поліна Євгеніївна – Харківський національний університет радіоелектроніки, асистент кафедри системотехніки, Харків, Україна, e-mail: polina.zhernova@gmail.com; ORCID ID: https://orcid.org/0000-0002-2154-4766.

Жернова Полина Евгеньевна – Харьковский национальный университет радиоэлектроники, ассистент кафедры системотехники, Харьков, Украина.

Zhernova Polina – Kharkiv National University of Radio Electronics, Assistant Lecturer at the Department of System Engineering, Kharkiv, Ukraine.

ОНЛАЙН НЕЧІТКА КЛАСТЕРИЗАЦІЯ ПОТОКІВ ДАНИХ ВЫСОКОЇ РОЗМІРНОСТІ НА ОСНОВІ АНСАМБЛІВ НЕЙРОННИХ МЕРЕЖ

Предметом дослідження в статті є нечітка кластеризація даних високої розмірності на основі ансамблевого підходу за умови, що кількість та форма кластерів невідомі. **Мета** роботи – створення нейро-фаззі підходу для кластеризації даних, коли потік даних подається на обробку в онлайн-режимі в припущенні, що кількість та форма кластерів невідомі. У статті вирішуються наступні **завдання**: компресія вхідного простору ознак в онлайн режимі, формування моделі ансамблів нейронних мереж для кластеризації даних, розробка ансамблю нейро-фаззі мереж для кластеризації даних високої розмірності, розробка підходу для кластеризації даних в онлайн режимі. Отримані наступні **результати**: основна ідея запропонованого підходу заснована на модифікації нечіткого алгоритму С-середніх. Для зниження розмірності вхідного простору пропонується використовувати модифіковану мережу Хебба-Сенгера, яка відрізняється підвищеною швидкодією та побудовану на основі модифікованих нейронів Ойя. Запропоновано оптимізований за швидкодією алгоритм навчання нейрона Ойя. Така мережа реалізує метод головних компонент в онлайн-режимі з високою швидкодією. **Висновки**: В тому випадку, якщо процедура редукції-компресії не може бути використана через можливість втрати фізичного сенсу вихідного простору, нами запропоновано новий критерій кластеризації, який містить в собі як відомий поліноміальний фаззіфікатор, так і зважування окремих компонент відхилень аналізованих образів від центроїдів кластерів. Введена рекурентна модифікація заснована на алгоритмах запропонованих в даній статті. Розроблено математичну модель для визначення якості кластеризації з використанням індекса Ксі-Бені, який був модифікований для онлайн режиму. Експериментальні результати підтвердили той факт, що запропонована система дозволяє вирішувати широкий спектр завдань Data Mining, коли набори даних обробляються в онлайн-режимі за умови, що кількість та форма кластерів не відомі, а також мають велику кількість спостережень.

Ключові слова: кластерування; метод нечітких С-середніх; послідовний аналіз головних компонент; ансамбль нейро-фаззі мереж; нейронна мережа Т. Кохонена; самонавчання.

ОНЛАЙН НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ ПОТОКОВ ДАННЫХ ВЫСОКОЙ РАЗМЕРНОСТИ НА ОСНОВЕ АНСАМБЛЕЙ НЕЙРОННЫХ СЕТЕЙ

Предметом исследования в статье является нечеткая кластеризация данных высокой размерности на основе ансамблевого подхода при условии, что количество и форма кластеров неизвестны. **Цель** работы – создание нейро-фаззи подхода для кластеризации данных, когда поток данных подается на обработку в онлайн-режиме в предположении, что количество и форма кластеров неизвестны. В статье решаются следующие **задачи**: компрессия входного пространства признаков в онлайн режиме, формирование модели ансамблей нейронных сетей для кластеризации данных, разработка ансамбля нейро-фаззи сетей для кластеризации данных высокой размерности, разработка подхода для кластеризации данных в онлайн режиме. Получены следующие **результаты**: основная идея предложенного подхода основана на модификации нечеткого алгоритма С-средних. Для снижения размерности входного пространства предлагается использовать модифицированную сеть Хебба-Сенгера, отличающуюся повышенным быстродействием и построенную на основе модифицированных нейронов Ойя. Предложен оптимизированный по быстродействию алгоритм обучения нейрона Ойя. Такая сеть реализует метод главных компонент в онлайн-режиме с высоким быстродействием. **Выводы**: В том случае, если процедура редукции-компрессии не может быть использована из-за возможности потери физического смысла исходного пространства, нами введен новый критерий кластеризации, содержащий в себе как известный полиномиальный фаззификатор, так и взвешивание отдельных компонент отклонений предъявляемых образов от центроидов кластеров. Введена рекуррентная модификация, основанная на алгоритмах, предложенных в данной статье. Разработана математическая модель для определения качества кластеризации с использованием индекса Кси-Бени, который был модифицирован для онлайн режима. Экспериментальные результаты подтвердили тот факт, что предлагаемая система позволяет решать широкий спектр задач Data Mining, когда наборы данных обрабатываются в онлайн-режиме при условии, что количество и форма кластеров не известны, а также содержат большое количество наблюдений.

Ключевые слова: кластеризация; метод нечетких С-средних; последовательный анализ главных компонент; ансамбль нейро-фаззи сетей; нейронная сеть Т. Кохонена; самообучение.