Informatics and Information Technologies

DOI: https://10.15407/kvt198.04.003

UDC 519.7:004.8

PEROVA I.G., PhD, Associate Professor,
Associate Professor of Biomedical Engineering Department,
e-mail: rikywenok@gmail.com
BODYANSKIY Ye.V., DSc (Engineering), Professor,
Professor of Artificial Intelligence Department
e-mail: yevgeniy.bodyanskiy@nure.ua
Kharkiv National University of Radio Electronics,
14, Nauky av., Kharkiv, 61166, Ukraine.

ONLINE MEDICAL DATA STREAM MINING BASED ON ADAPTIVE NEURO-FUZZY APPROACHES

Introduction. *Data mining approaches in medical diagnostics tasks have a number of special properties that do not allow the use of such approaches in a classical form. That's why adaptive neuro-fuzzy systems for online medical data stream processing tasks and its learning algorithms have been developed. Proposed systems can process medical data streams in three modes: supervised learning, unsupervised learning and active learning.*

The purpose *of the paper is to develop approach, based on adaptive neuro-fuzzy systems to solve the tasks of medical data stream mining in online-mode.*

Methods. *The methods of computational intelligence are used for medical data stream processing and, first of all, artificial neural networks, neuro-fuzzy systems, neo-fuzzy systems, their supervised, unsupervised and active learning approaches, gradient methods of optimization, methods of evolving system.*

Results. *As a result, approbation of the developed approach in supervised learning mode using multidimensional neo-fuzzy neuron on medical data of patients with urological disease was investigated. Percentage of errors in system testingusing all feature space is 11.11 %, using the most informative features the error rate becomes 6.4 %. Also multidimensional neo-fuzzy neuron was used for diagnostic of the pharmacoresistant form of epilepsy, percentage of errors in system testing is 5.82 %. Approval of the developed approach in the mode of active training and association on the data of patients with pulmonary diseases was performed. For all approbation results performance criterion was calculated, its values are suitable for the tasks of medical diagnostics in data stream mode.*

Conclusions. *The proposed neuro-fuzzy approaches allow obtaining additional information about patients diagnosis in conditions of limited a priori information about patient.*

Keywords: *adaptive system, neuro-fuzzy system, medical data mining, medical data stream.*

INTRODUCTION

Currently, data mining approaches are widely used to solve a wide range of problems in industry, economics, finance, banking, agriculture etc. In the area of medical diagnostics these methods are called medical data mining approaches [1-4]. The specialty of this area is the inability to use the traditional methods of data mining in its pure form, which is associated with a number of circumstances:

- limited sampling to be classified;
- significant overlapping of classes related to various diseases;
- nonlinear nature of hypersurfaces that divide these classes;
- the presence of anomalous observations that can distort primary information;
- a significant role of the subjective human factor that does not provide accurate data;
- a need to process medical data sequentially in online mode;
- an ability to present medical data in the form of data streams.

All of these circumstances lead to the formation of non-convex and blurred classes, for which the suitable mathematical methods are methods of computational intelligence, above all, artificial neural networks, fuzzy inference systems and hybrid neuro-fuzzy systems. However, the systems listed are not a «panacea» in the tasks of medical diagnostic because they require large amounts of information for their training, which are often not available to physicians and are not adapted to the need to process data in a sequential online mode. It should also be emphasized that the compulsory stage is the preparation of medical data, covering the task of filling in missing values (if any), normalizing the data and reducing the number of features by compressing the data or selecting the most informative features. The available approaches do not ensure compliance with these online requirements for medical diagnostic tasks.

The eHealth system was introduced in Ukraine in 2018, with the ultimate goal of creating a database of medical records for all Ukrainians by 2020. At this stage, the processing of medical information in sequential mode using the medical data mining approaches will be relevant. The main tasks in this area are the problems of diagnosis, which are solved by means of pattern recognition based on paradigms of supervised learning and unsupervised one (self-learning), and they are reduced to solving problems of classification or clustering.

The goal of this work is to improve the effectiveness of medical diagnostics in online mode based on the intellectual analysis of medical datasets using hybrid neuro-fuzzy systems in conditions of limited a priori information about patient.

DATA STREAM PROCESSING

High-dimensional medical data form data streams — sequential feeding to processing at short time intervals. That is why all the approaches used to processing and analyzing such datasets must be adapted to sequential data stream processing. An example is the formation of a data streaming one of the departments of a hospital or one hospital as a whole, when different physicians form separate data about each patient and these data are sent to the central repository or hospital database. Thereafter, medical data from all hospitals are transferred to the eHealth Repository (Fig. 1).

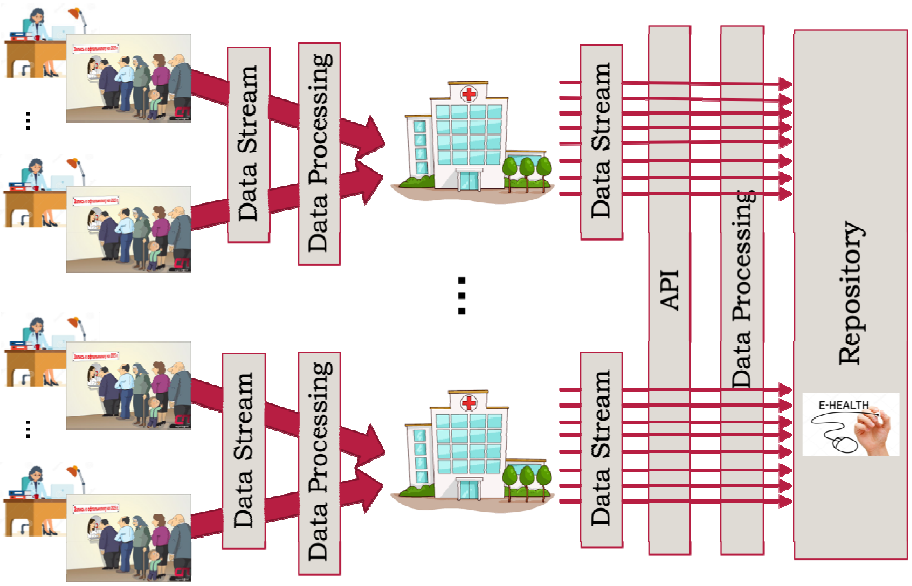

Fig. 1. Medical Data Stream

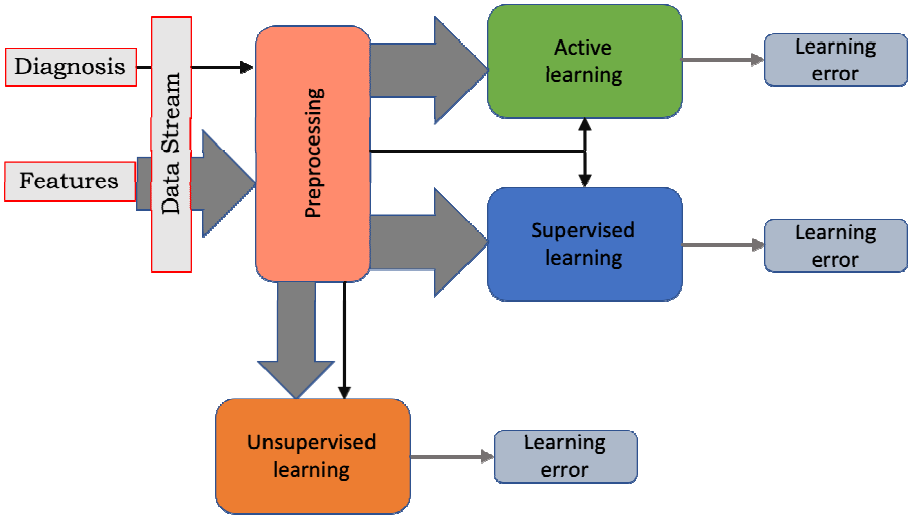

Fig. 2. System training process

It is understandable that data stream characteristics (such as density and time between patients) in a single hospital department will be less than those ones in the entire hospital and country.

Medical data stream processing should be performed in several steps, which depends on ability to know the diagnosis d_j of a sufficient number of patients to form training set for using supervised learning approaches or using unsupervised or active learning methods in other cases. The use of the mentioned approaches requires different initial data, depending on whether the system is trained or tested (Fig. 2, Fig.3).

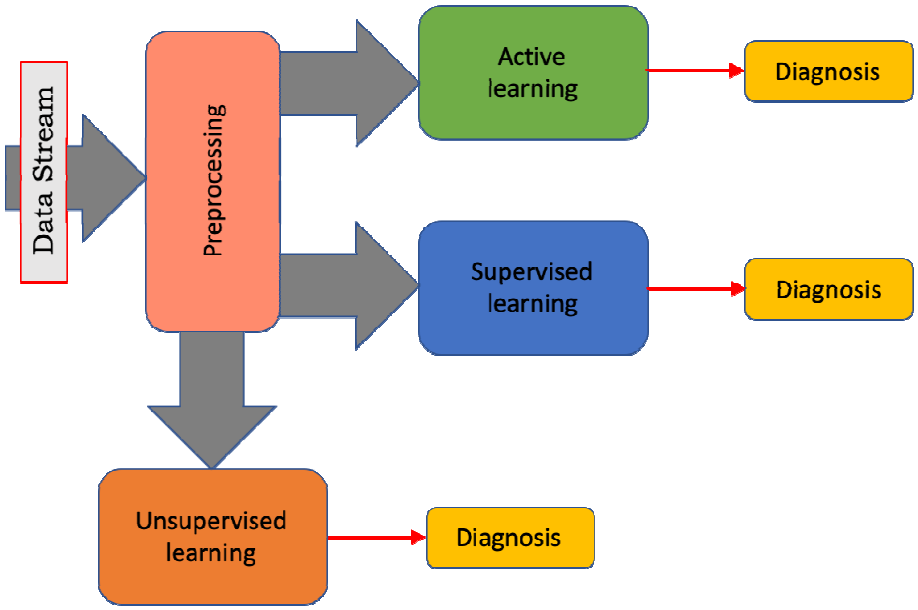

Fig. 3. System testing process

At first we need to find out input data characteristics such as parameters of one element of data stream, formed by N n – dimensional feature vectors

$$X = \{x(1), x(2), ..., x(N)\}, \tag{1}$$

where $x(k) \in X$; $k = 1, 2, ..., N$; N — total number of patients; n — total number of features for each patient.

In our case each element is a sample of observations

$$x(k) = (x_1(k), x_2(k), ...x_i(k), ..., x_n(k))^T \to d_j(k). \tag{2}$$

Further analysis assumes a fuzzy partitioning of the original dataset into m classes with some level of membership of k – th vector to the j – th cluster class (diagnosis) $d_j(k)$.

DATA PREPROCESSING STAGE

In the medical data stream mining tasks, the data of biomedical research are collected from different sources (different hospitals, hospital departments, etc.). That is why there is a problem of the absence of clearly defined parameters of the organism, which were measured for a particular nosology. This problem is expressed in gaps in the patient's features, in the presence of which it is not possible to process the whole table without filling them in advance. Approaches for filling gaps was described in [5-6], for its online version adapted for medical data processing see [7]. During data stream processing, the gaps will be filled in such a way that the recovered elements would in some sense be most "similar" or "close" to the a priori unknown patterns hidden in this table by improving the

method of spatial extrapolation (Fig.4) [8]. To fill gaps in the features of patient $x(k)$ at first we need to calculate the partial distances $dist(k,p)$ between $x(k)$ and all other patients $x(p)$, $k \neq p$ using formula

$$dist(k,p) = \frac{1}{n - g_k - g_p + g_{kp}} \sum_{i=1}^{n} \left| x_i(k) - x_i(p) \right|, \tag{3}$$

where g_k, g_p — number of gaps in the features of *k*-th and *p*-th patient, g_{kp} — number of gaps in common features of *k*-th and *p*-th patient.

Using partial distances values $dist(k,p)$ we can calculate membership function $mu(k,p)$:

$$mu(k,p) = \frac{dist^{-1}(k,p)}{\sum_{p=1}^{N-1} dist^{-1}(k,p)}. \tag{4}$$

At the final stage a gap value is filled using formula:

$$x_{\tilde{i}}(k) = \sum_{p=1}^{N-1} mu(k,p) \cdot x_{\tilde{i}}(p), \tag{5}$$

where $x_{\tilde{i}}(p)$ does not contain a gap.

After filling all gaps online medical data preprocessing tasks require to calculate basic sample statistics such as mean, variance, extreme values (maximum and minimum) also sequentially without the need for accumulating data [9].This step allows you to normalize the data stream to the necessary interval $[-1;1]^n$ or $[0;1]^n$ in online mode.

A final stage of data stream preprocessing is a step of feature selection-extraction. The need for such processing is due to the fact that medical datasets often contain too many features with a small number of patients, which significantly limits the possibilities of existing methods for further diagnosis. In order to select the most informative features from the available feature space, it is proposed to integrate the advantages of systems based on the combination of methods of compression of the original features space with the methods of finding the most informative features and to create a single adaptive hybrid method of evaluating the informativeness of features with the selection of the most informative ones [10–11]. A hybrid method of evaluating the informativeness of medical features is presented in Fig. 5.

It consists of a block of normalization and centering of input features, a block of calculating of first principal component using a modified Oja neuron [12-15], a block of definition of the "feature-winner" where a feature with minimal distance to the first principal component is defined.The distance in the sense of the Manhattan metric is calculated according to:

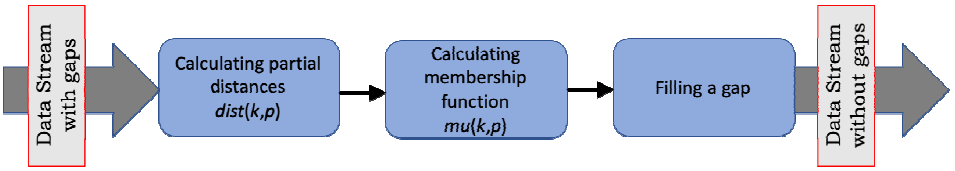

Fig. 4. Process of filling the missed values

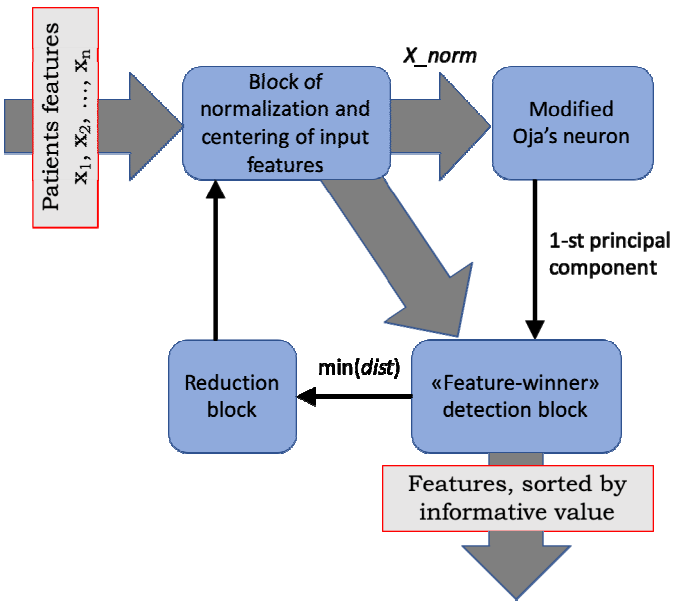

Fig. 5. Feature selection-extraction approach

$$dist\left(x\_norm(z), pr\_comp^{(1)}\right) = \sum_{i=1}^{N}\left|x\_norm_i(z) - pr\_comp^{(1)}\right| \quad (6)$$

in the reduction block, "feature-winner" is removed and the next informative feature begins to be searched.

SUPERVISED LEARNING MODE

Medical data mining approaches in a supervised learning mode based on adaptive hybrid neuro-fuzzy systems are appropriate to use in situations where there is a representative training sample that means that diagnosis of many patients is known. The first method in this approach is to modify a multidimensional neo-fuzzy-neuron (Fig. 6) [16–17].

Triangular structures are usually used as a membership function $\mu_{li}(x_i)$, its value is determined by the distance between the value of the input feature x_i and the centers of these functions c_{li}. So the output of this layer can be presented in the form:

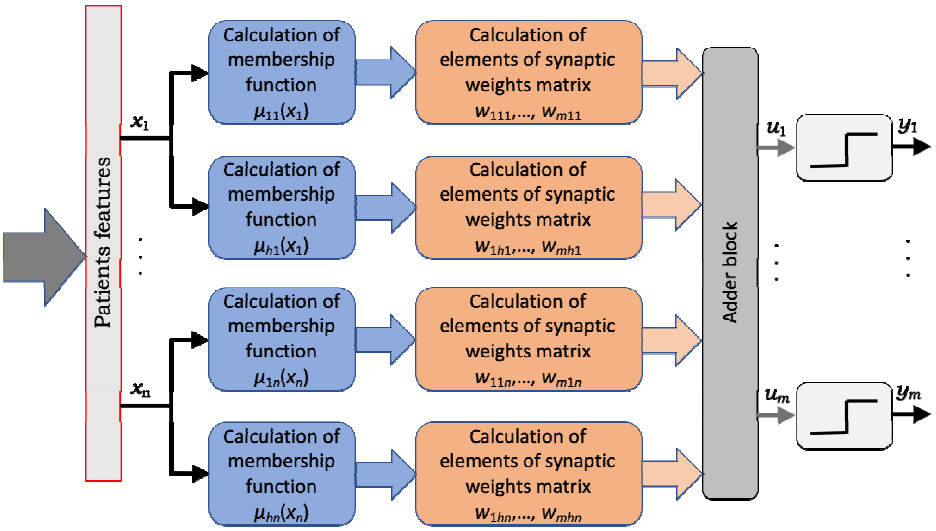

Fig. 6. Multidimensional neo-fuzzy-neuron

$$\mu(k)=\left(\mu_{11}(x_1(k)),\mu_{21}(x_1(k)),...,\mu_{h1}(x_1(k)),...\mu_{hn}(x_n(k))\right)^T. \quad (7)$$

The signal at the output of the neo-fuzzy-neuron can be calculated as:

$$u(k)=W(k)\mu(k),\ y(k)=\text{sign}\,u(k), \quad (8)$$

where $W(k)=\begin{pmatrix} w_{111}(k),..., & w_{1h1}(k),..., & w_{1li}(k),..., & w_{1hn}(k) \\ w_{211}(k),..., & w_{2h1}(k),..., & w_{2li}(k),..., & w_{2hn}(k) \\ \vdots\in & & & \\ w_{m11}(k),..., & w_{mh1}(k),..., & w_{mli}(k),..., & w_{mhn}(k) \end{pmatrix}$ — synaptic weights matrix, customizable using a modified procedure [18]

$$\begin{cases} W(k+1)=W(k)+r^{-1}(k)\left(d(k)-\text{sign}\,W(k)\mu(k)\right)\mu^T(k), \\ r(k)=\alpha r^{-1}(k-1)+\left\|\mu(k)\right\|^2, \quad 0\le\alpha\le 1 \end{cases} \quad (9)$$

that optimizes the learning criterion

$$E_j(k)=\frac{1}{2}e_j^2(k)=\frac{1}{2}\left(d_j(k)-y_j(k)\right)^2. \quad (10)$$

To enhance the capabilities of Medical Data Mining approaches in supervised learning mode, a hybrid neuro-fuzzy system (HNFS) has been provided [19] (Fig.7). A particularity of this system is the ability to further change its architecture in situation when number of features or diagnosis can be changed [19–21].The normalized feature vector is fed to the input layer of the HNFS, that consists of nh membership functions $\mu_{li}(x_i(k))$ and performs the fuzzyfication of feature vector:

$$\mu_{li}(x_i(k)) = \exp\left(-\frac{(x_i(k)-c_{li})^2}{2\sigma_i^2}\right), \tag{11}$$

where c_{li} — center of membership function, σ_i — width of membership function.

At the output of multiplication block values $\prod_{i=1}^{n}\mu_{li}(x_i(k))$ have been calculated. After we perform tuning of synaptic weights matrix, adder blocks (AB) calculate signals:

$$\tilde{x}_j(k) = \sum_{l=1}^{h} w_{jl}\prod_{i=1}^{n}\mu_{li}(x_i(k)). \tag{12}$$

In Rectified Linear Unit (ReLU) block calculates [22]:

$$\varphi(\tilde{x}_j(k)) = \begin{cases}\tilde{x}_j(k), & \text{if} \quad \tilde{x}_j(k) > 0,\\ 0, & \text{if} \quad \tilde{x}_j(k) \le 0.\end{cases} \quad j = 1,2,...,m, \tag{13}$$

which are transformed in normalizing block using adder block (AB) value $\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i(k))$ and provide calculating of $u_j(k)$:

$$u_j(k) = \frac{\varphi\left(\sum_{l=1}^{h} w_{jl}\prod_{i=1}^{n}\mu_{li}(x_i(k))\right)}{\sum_{j=1}^{m}\varphi\left(\sum_{l=1}^{h} w_{jl}\prod_{i=1}^{n}\mu_{li}(x_i(k))\right)}. \tag{14}$$

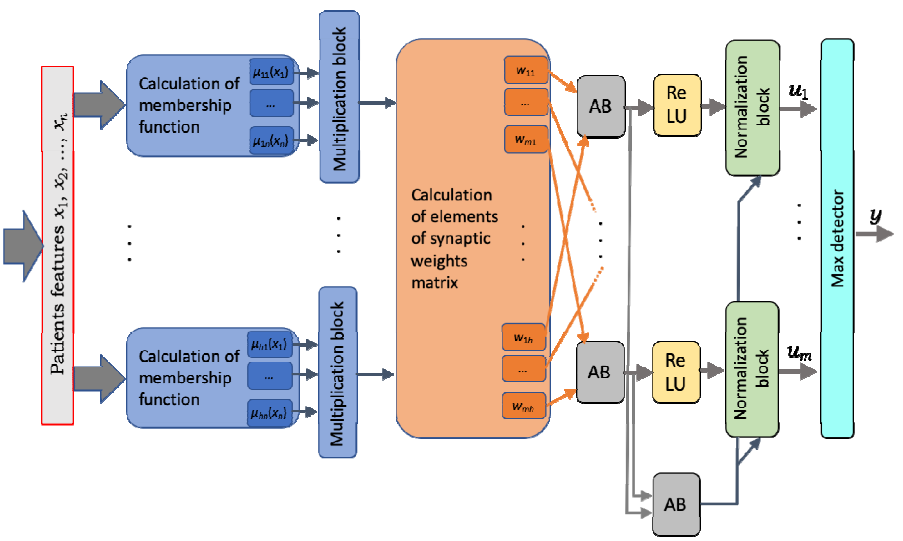

Fig. 7. Hybrid neuro-fuzzy system (HNFS)

Output signal of HNFS is calculated using maximum detector:

$$y(k) = \max\{u_1, u_2, ..., u_m\}. \tag{15}$$

Learning criterion for training proposed system can be presented in the form [22]:

$$E_j(k) = \frac{1}{2}\left\|d(k) - \tilde{x}(k)\right\|^2. \tag{16}$$

Modificated procedure that is used for tuning of synaptic weights matrix has been written in the form:

$$W(k+1) = W(k) + \frac{e(k)\left(\prod_{i=1}^{n}\mu_{li}(x_i(k))\right)^T}{\left\|\prod_{i=1}^{n}\mu_{li}(x_i(k))\right\|^2} = \tag{17}$$

$$= W(k) + \left(d(k) - W(k)\left(\prod_{i=1}^{n}\mu_{li}(x_i(k))\right)\right)\left(\prod_{i=1}^{n}\mu_{li}(x_i(k))\right)^{+},$$

where $W(k) = \begin{pmatrix} w_{11}(k) & w_{12}(k) & ... & w_{h1}(k) \\ w_{21}(k) & w_{22}(k) & ... & w_{h2}(k) \\ \vdots & \vdots & \vdots & \vdots \\ w_{m1}(k) & w_{m2}(k) & ... & w_{mh}(k) \end{pmatrix}$ — $(m \times h)$- synaptic weights matrix; $d(k) = (d_1(k), d_2(k), ..., d_m(k))^T$ — reference signal, which contains information about the patients diagnosis and can take only two values of 0 or 1, $e(k) = (e_1(k), e_2(k), ..., e_m(k))^T$ — vector of training errors.

In situation when number of features of any patients or number of its possible diagnosis can be changed, HNFS architecture also changes (HNFS_evol — Fig. 8) [19].

If patient vector $x(k)$ is fed to the system input and it characterized by the same set of features as all previous patients and one more feature $n+1$:

$$x(k) = (x_1(k), ..., x_n(k), x_{n+1}(k))^T. \tag{18}$$

System evolution can be realized in the layer of membership function calculation, were new functions $\mu_{li}(x_i(k))$, $i = 1, 2, ..., n+1$, $l = 1, 2, ..., h$ appear. These functions are multiplied to h values $\prod_{i=1}^{n+1}\mu_{li}(x_i(k))$ and HNFS_evolworks according to the equations(11)–(15).

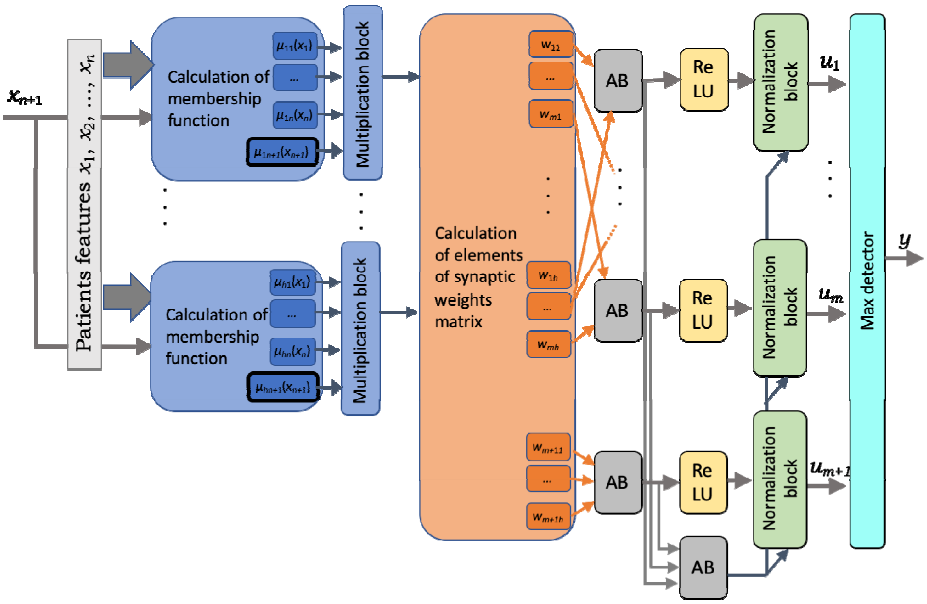

Fig. 8. Hybrid neuro-fuzzy evolving system (HNFS_evol)

In situations when a patient is diagnosed with a new diagnosis, the HNFS_evol architecturally evolves at the level of synaptic weights calculation: a new synaptic weights $w_{m+11},\ldots,w_{m+1h}$, a ReLU activation function and one more normalization block for defuzzyfication system results appear in the system architecture. The most probable diagnosis is formed by the maximum detection block a:

$$y(k)=\max\{u_1,\ldots,u_m,u_{m+1}\}. \tag{19}$$

For training of HNFS_evol criterion (16) is used. For synaptic weights tuning we should concatenate the pre-trained matrix W with a new weight-vector that corresponds to the new diagnosis:

$$W^*(k+1)=\begin{pmatrix}W(k+1)\\ ------\\ w_{m+1}(k+1)\end{pmatrix}=\begin{pmatrix}W(k)\\ ----\\ w_{m+1}(k)\end{pmatrix}+$$
$$+\frac{\left(\begin{pmatrix}d(k)\\ ----\\ d_{m+1}(k)\end{pmatrix}-\begin{pmatrix}\tilde{x}(k)\\ ----\\ \tilde{x}_{m+1}(k)\end{pmatrix}\right)\left(\prod_{i=1}^{n}\mu_{li}(x_i(k))\right)^T}{\left\|\prod_{i=1}^{n}\mu_{li}(x_i(k))\right\|^2}. \tag{20}$$

So synaptic weights matrixis transformed to the form of $\left((m+1)\times h\right)$-matrix:

$$W^{*}(k)=\begin{pmatrix} w_{11}(k) & w_{12}(k) & \ldots & w_{1h}(k) \\ w_{21}(k) & w_{22}(k) & \ldots & w_{2h}(k) \\ \vdots & \vdots & \vdots & \vdots \\ w_{m1}(k) & w_{m2}(k) & \ldots & w_{mh}(k) \\ w_{(m+1)1}(k) & w_{(m+1)2}(k) & \ldots & w_{(m+1)h}(k) \end{pmatrix}. \tag{21}$$

This matrix must be trained on $m+1$ diagnosis without retraining previously trained synaptic weights for m diagnosis.

UNSUPERVISED LEARNING MODE

The need of using unsupervised learning methods arises in mass health examination when the diagnosis of all patients is unknown or they are considered conditionally healthy. Structural diagram of a method of adaptive robust fuzzy clustering of a patient features using a Manhattan metricsis shown in Fig. 9 [23].

For calculating distances in Manhattan metrics between $x(k)$ and c_m we can use:

$$dist\left(x(k),c_m\right)=\left|x(k)-c_m\right|. \tag{22}$$

Calculation of position of cluster centers c_m and membership level of each patient to each of clusters μ_m is performed using self-learning algorithm [23]:

$$\begin{cases} \mu_m\left(x(k)\right)=\dfrac{\left\|x(k)-c_m(k)\right\|^{-1}}{\sum\limits_{l=1}^{m}\left\|x(k)-c_l(k)\right\|^{-1}}, \\ c_m(k+1)=c_m(k)+\eta(k)\,\mu_m^2\left(x(k)\right)sign\left(x(k)-c_m(k)\right). \end{cases} \tag{23}$$

Thus, the diagnostic physician receives the degree of membership of each patient to each of clusters-diagnosis.

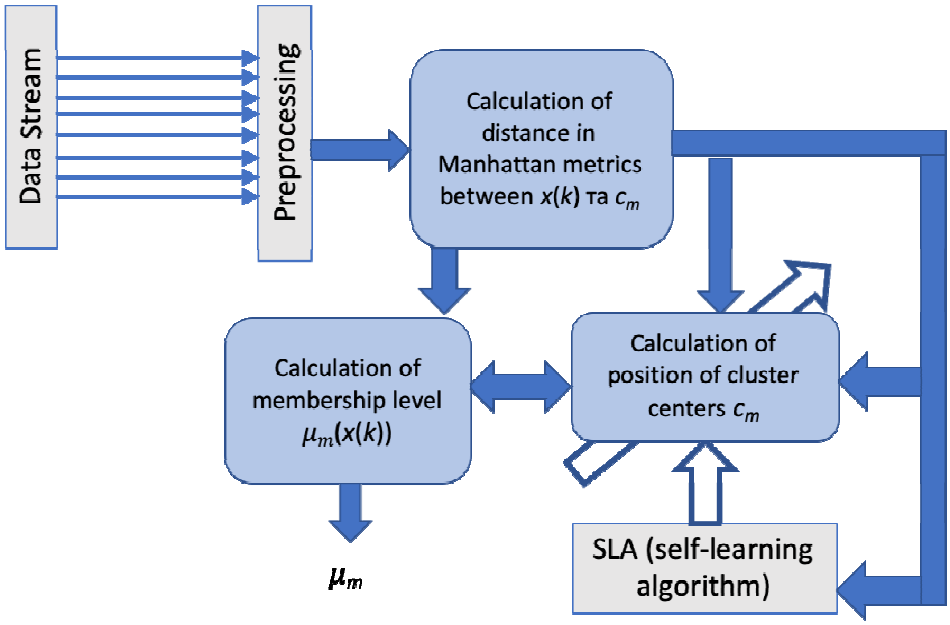

Fig. 9. Method of adaptive robust fuzzy clustering of a patient features using a Manhattan metric

ACTIVE LEARNING AND ASSOCIATIVE MODE

Approaches based on active learning are necessary in situations when a non-representative dataset is being processed, that is some patients diagnoses are known, but their numbers are insufficient for supervised learning mode [24–25].

In situations where the diagnosis of only a few patients are known, the use of associative neuro-fuzzy memory methods is appropriate. The associative clustering method based on neuro-fuzzy auto-associative memory realizes in few stages (Fig.10). At first stage a membership function is introduced, which can be described by (11) and its centers correspond to features of patients with known diagnosis with which it is necessary to associate. Width parameter of Gaussian functionis adjusted using level Δ that corresponds to a given level of intersection of two neighboring membership functions.These membership functions are fed to the multiplication blocks, in which values $\prod_{i=1}^{n}\mu_p(x_i)$ are calculated. These values are summarized $\sum_{p=1}^{m}\prod_{i=1}^{n}\mu_p(x_i)$ at adder block. A normalizing procedure is realized at the system output:

$$u_p(x)=\frac{\prod_{i=1}^{n}\mu_p(x_i)}{\sum_{p=1}^{m}\prod_{i=1}^{n}\mu_p(x_i)},\ p=1,2,...,m\,. \tag{24}$$

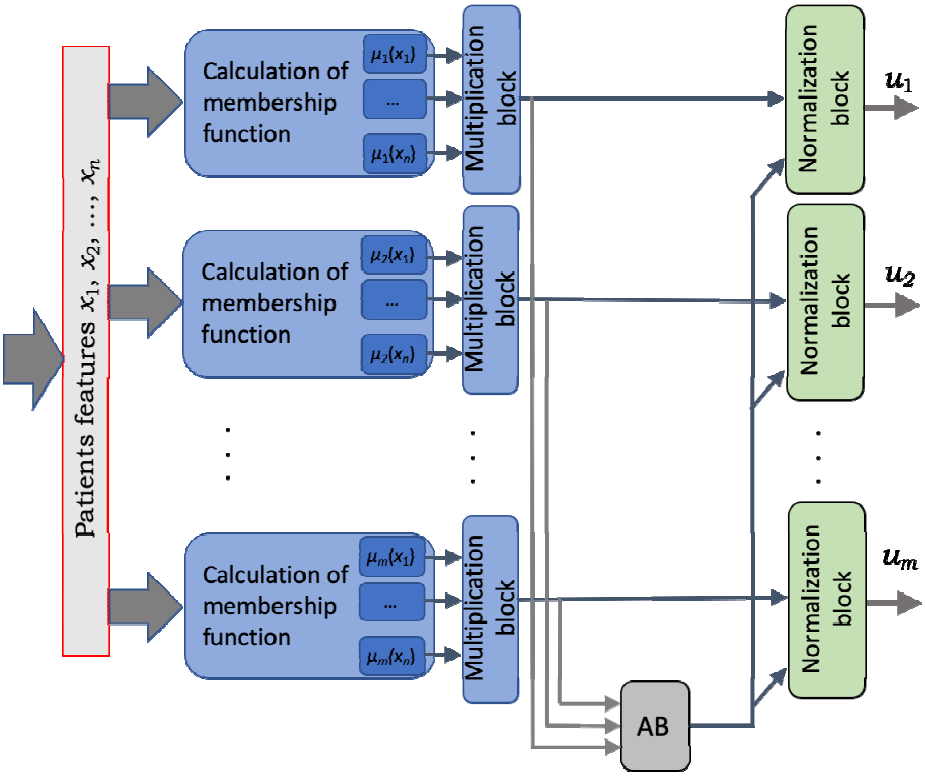

Fig. 10. Associative neuro-fuzzy memory

Thus, when patient's feature-vector is fed to the system input, signals $u_p(x)$ appear at the system output which correspond to levels of association with patients with known diagnosis.

As a part of the active learning mode, the neuro-fuzzy network with active learning is proposed. The system switches between self-learning and supervised learning modes, depending on whether the patient has a known diagnosis or not. In situations where the diagnosis is unknown, the system switches to self-learning mode and the first stage is the competition process when for each centroid-diagnosis cd_j the distances $dist$ are calculated:

$$dist^2\left(cd_j(k), x(k)\right) = \left\|x(k) - cd_j(k)\right\|^2. \tag{25}$$

Then the next step is the processes of cooperation and synaptic adaptation, presented in the form:

$$\begin{cases} cd_p(k+1)=\dfrac{cd_p(k)+\eta(k)u_p^2(k)\left(x(k)-cd_p(k)\right)}{\left\|cd_p(k)+\eta(k)u_p^2(k)\left(x(k)-cd_p(k)\right)\right\|} \quad \forall\ p=1,2,...,m, \\ 0\le u_p(k)=\left\|x(k)-cd_p(k)\right\|^{-2}\left(\sum\limits_{l=1}^{m}\left\|x(k)-cd_l(k)\right\|^{-2}\right)^{-1}\le 1, \\ \eta(k)=r^{-1}(k), \quad r(k)=\alpha r(k-1)+1, \quad 0\le\alpha\le 1. \end{cases} \tag{26}$$

In situations where the patient's diagnosis is known, the system switches to supervised learning, in which two situations may occur. If the feature vector falls into one of the specified Voronoi cells, the procedure for pulling up this particular centroid $cd_p(k)$ to $x(k)$ is implemented:

$$\begin{cases} cd_p(k+1)=\dfrac{cd_p(k)+\eta(k)\left(x(k)-cd_p(k)\right)}{\left\|cd_p(k)+\eta(k)\left(x(k)-cd_p(k)\right)\right\|}, \ \forall\ p=1,2,...,q, \\ \eta(k)=r^{-1}(k), \quad r(k)=\alpha r(k-1)+1, \quad 0\le\alpha\le 1. \end{cases} \tag{27}$$

In a situation when $x(k)$ falls in a Voronoi cell with a centroid winner $cd_j^*(k)$ that does not relate to particular diagnosis, the center of gravity is pushed away from $x(k)$:

$$\begin{cases} cd_j(k+1)=\dfrac{cd_j^*(k)-\eta(k)\dfrac{\left(1-\cos\left(cd_j(k),x(k)\right)\right)^{-1}}{\sum\limits_{l=1}^{m}\left(1-\cos\left(cd_l(k),x(k)\right)\right)^{-1}}\left(x(k)-cd_j^*(k)\right)}{\left\|cd_j^*(k)-\eta(k)\dfrac{\left(1-\cos\left(cd_j(k),x(k)\right)\right)^{-1}}{\sum\limits_{l=1}^{m}\left(1-\cos\left(cd_l(k),x(k)\right)\right)^{-1}}\left(x(k)-cd_j^*(k)\right)\right\|}, \\ \eta(k)=r^{-1}(k), \quad r(k)=\alpha r(k-1)+1, \quad 0\le\alpha\le 1. \end{cases} \tag{28}$$

PERFORMANCE CRITERION

To compare the effectiveness of online medical diagnosis methods, the quality information criterion (Performance Criterion *PerCr*) was provided, taking into account the *Fault* diagnosis and the time for decision making (*Time*) required to process one patient:

$$PerCr = \lambda_1 Fault + \lambda_2 Time\,, \tag{29}$$

where λ_1, λ_2 — weighting coefficients that are chosen under the conditions that $\sum \lambda = 1$.

By analyzing the processing time for diagnostics of one patient, it was assumed that the processing time (Time) should be normalized [0;1] where 0 corresponds to a zero processing time and 1 to a possible maximal processing time equal to 1 ms. For diagnostic tasks in the data stream mode a value of performance criterion should be from 0 to 0.25, acceptable from 0.25 to 1 under the conditions of controlling the accuracy of diagnostics. All the results were obtained using MacBook (Retina 12-inch, Early 2016), processor 1,1 GHz Intel Core m3, 8 Gb 1867 MHz LPDDR3. Python 3.7 programming language in Spyder 3.3.2 was used for programming.

RESULTS OF RESEARCH IN SUPERVISED LEARNING MODE

Approbation of the developed approach in supervised learning mode for patients with urological diseases on the basis of evaluating the information content of the symptom complex was conducted. The medical sample contained information about features of patients with six urological diagnoses. The number of patients who participated in the study was 188, each described by 106 features. At first step Feature Selection-Extraction method was used to evaluate the informativeness of the features, number of features was reduced to 12. At second step all 188 patients were divided into training and testing sets (126 patients were in the training set, 62 patients were in the testing set). Table 1 lists the training and testing errors of multidimensional neo-fuzzy neuron and the value of the performance criterion.

It is easy to see that error percentage and performance criterion were less when input data were represented by 12 features compared to initial 106 features.

A fuzzy diagnostics of the pharmacoresistant form of epilepsy using a multidimensional neo-fuzzy neuron was performed. Examination of 309 patients with epilepsy was used as input data, the experience of the disease in each of the patients was at least one year. The total number of features was 25. All patients were divided into two groups: the first group with efficacy of drug treatment and the second group were patients with pharmacological resistance. A visualization of position of the classes-diagnosis in the space of the first three principal components is presented in Fig. 11.

Table 1. **Results for multidimensional neo-fuzzy neuron**

	Errors		*PerCr*
	Training set	Testing set	
All features (106)	7.19 %	11.11 %	0,085
The most informative features (12)	5.3 %	6.4 %	0,0545

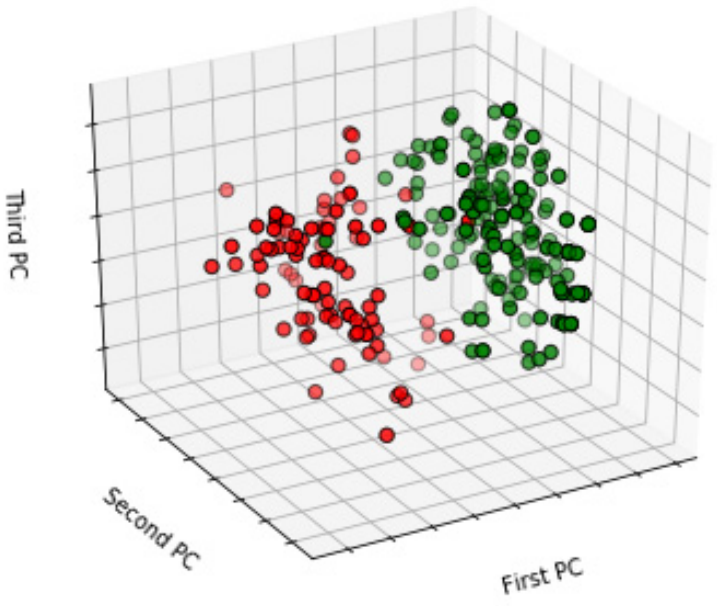

Fig. 11. Dataset was divided into training and testing sets in a ratio of 206 patients to 103 patients. The results of multidimensional neo-fuzzy training and testing were presented in Table 2.

Table 2. **Results for multidimensional neo-fuzzy neuron**

	Errors on training set (206 patients)	Errors on testing set (103 patients)
Number of patients	2	6
% of patients	0,97	5,82
PerCr	0,0683	0,0501

Approbation of the hybrid neuro-fuzzy system (HNFS) and hybrid neuro-fuzzy evolving system (HNFS_evol) in supervised learning mode performed on datasets from UCI Repository [26-29]. In the first step, datasets were normalized at interval $x_i \in [0,1]$. The columns corresponding to the patient's health status (5^{th} for the Iris.data, 9^{th} for Pima-indian-diabetes.data, 35^{th} for dermatology.data and 17^{th} for parkinson.data) were moved to the vector $d(k)$. In the next step, one random attribute was removed from all datasets. The system was trained on n features, after this deleted attribute was returned. The results are shown in Table 3.

All values of the performance criterion are acceptable for Medical Data Mining Tasks (its value is more than 0.25) in situations where the data is received sequentially, including situations when it is necessary to work in the conditions of data stream.

In the next phase of the approbation, patients with one of the diagnosis classes were excluded from all datasets. The system was trained on $n+1$ attributes and m diagnoses. Further removed patients were added to the experiment. The classification results are shown in Table 4.

Thus, it should be noted that the HNFS, HNFS_evol and procedures of its training showed a high percentage of correct classification for medical diagnosis tasks. Testing the operation of a system on three medical datasets confirms its performance in overlapping diagnostic classes on small datasets with variable numbers of features and diagnoses.

Table 3. **Classification results for HNFS and HNFS_evol**

Dataset	% errors for n features / Time(*HNFS)*		% errors for $(n+1)$ features / Time(*HNFS_evol)*	
	Training set	Testing set	Training set	Testing set
Iris.data	1,4 % /130 μs	2,68 % /70 μs	3,57 % /128 μs	8,23 % / 70 μs
Dermatology.data	0,87 % /171 μs	6,5 % /91 μs	4,75 % /158 μs	16,55 % /90 μs
Pima-indian-diabetes.data	0,45 % /136 μs	7,33 % /79 μs	0,25 % /135 μs	7,94 % /77 μs
Parkinson.data	1,33 % /135 μs	4,66 % /81 μs	0,56 % /132 μs	7,92 % /76 μs
PerCr	< 0,0898	< 0,07615	< 0,0102	< 0,127

Table 4. **Classification results for HNFS and HNFS_evol**

Dataset	% errors for m diagnosis / Time (*HNFS*)		% errors for $(m+1)$ diagnosis / Time (*HNFS_evol)*	
	Training set	Testing set	Training set	Testing set
Iris.data	1 % / 820 μs	2 % / 640 μs	9 % / 850 μs	16,6 % / 630 μs
Dermatology.data	0,5 % / 980 μs	4 % / 690 μs	13,16 % / 990 μs	20 % / 720 μs
PerCr	< 0,492	< 0,365	< 0,56	< 0,46

RESULTS OF RESEARCH IN THE ACTIVE TRAINING AND ASSOCIATION MODE

Approval of the developed approach in the mode of active training and association on the data of patients with pulmonary diseases was tested. The medical data set consists of 132 patients, each is characterized by 104 features (gender, age, patient complaints — 24 features, medical history — 14 features, objective description of the patient's condition — 26 features, clinical blood test — 10 features, biochemical blood test — 8 features, clinical analysis of urine — 10 features, chest radiography — 6 features, ECG — 8 features, spirometry — 2 features).

The entire data set was divided into three groups-diagnosis: 46 patients with chronic obstructive pulmonary disease, 53 patients with bronchial asthma, and 33 patients with pneumonia. Data visualization in the space of the three principal components is presented in Fig.12.

Table 5. **The results of fuzzy association patients**

Patients ID	Δ -value	Errors	*PerCr* -value
ID25	0.2	18,15 %	0,469
	0.5	24,7 %	0,502
ID64	0.2	21,44 %	0,486
	0.5	35,8 %	0,558
ID18	0.2	26,7 %	0,512
	0.5	30,66 %	0,532
ID32	0.2	19,75 %	0,477
	0.5	25,78 %	0,507
ID65	0.2	17,89 %	0,468
	0.5	19,9 %	0,478
ID70	0.2	21,33 %	0,485
	0.5	30,02 %	0,529

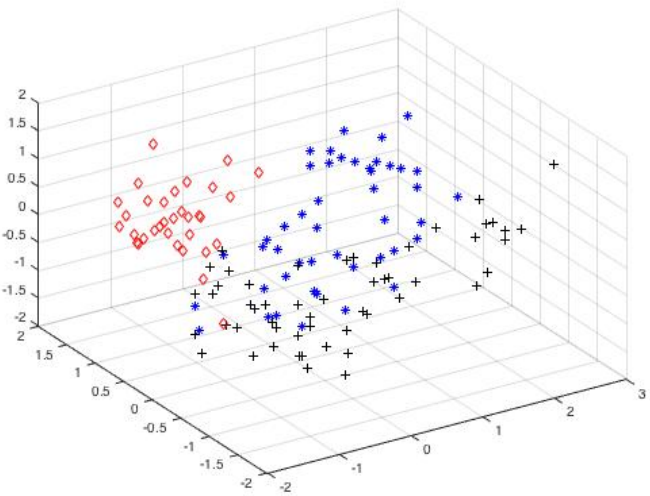

Fig. 12. Data visualization: ◊ — patients with pneumonia; * — patients with chronic obstructive pulmonary disease; + — patients with bronchial asthma.

The physician identified the most typical representatives of each of the disease groups (five for each group). All other patients were fed to the system input unmarked (with unknown diagnosis). The neuro-fuzzy network with active learning showed the percentage of correctly classified patients at the level of 84.38 %. One patient's processing time is $Time = 36$ μs. Performance criterion value is $PerCr = 0.0961$.

Associative neuro-fuzzy memory methods were used for fuzzy diagnostics of hypertension and coronary heart disease.The following medical data of patients were obtained: age, gender, patient complaints coded from 0 to 1 (12 features), the value of clinical (8 features) and biochemical (22 features) analysis, heart rate. So, each of the patients can be represented as a vector containing 44 features. The total number of patients was 95. Physicians identified 6 patients who could be considered as the most

representative of each group (patients with ID18, ID25, ID64 for hypertension; patients with ID32, ID65, ID70 for ischemic heart disease). The results of fuzzy association at different values of centers of fuzzy membership functions are shown in Table 5 with different parameters Δ. The best percentages of association with value of $\Delta = 0.2$ has shown for patient with ID65 and ID25, as evidenced by the minimum value of the performance criterion.

CONCLUSION

Online medical data stream mining based on adaptive neuro-fuzzy approaches in the mode of supervised, unsupervised and active learning was considered.Special learning algorithm for neuro-fuzzy systems training was introduced. The proposed approaches allow obtaining additional information about patient diagnosis in conditions of limited a priori information about patient.Testing results on clinical medical data confirm the efficiency of the developed approaches.

REFERENCES

1. de Oliveira J.,Pedrycz W. Advances in Fuzzy Clustering and its Applications. 2007. 454p.
2. Berka P., Rauch J., Zighed D. Data mining and medical knowledge management cases and applications. New-York, 2009. 440p.
3. Giannopoulou E. Data mining in medical and biological research. New York, 2008. 331p.
4. Karahoca A. Data Mining Applications in Engineering and Medicine. *InTechOpen*. 2012. 336p.
5. Han J.,Kamber M. Data Mining: Concepts and Techniques. Amsterdam, 2006. 743p.
6. Gorban A., Kegl B., Wunsch B.,Zinovyev A. Principal Manifolds for Data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Engineering. Berlin-Heidelberg-New York, 2007.Vol. 58. 330p.
7. Mulesa P.,Perova I. Fuzzy Spacial Extrapolation Method Using Manhattan Metrics for Tasks of Medical Data Mining. Proceeding of Computer Science and Information Technologies CSIT'2015(Lviv, 14-17th of Sept, 2015). Lviv, 2015. P. 104–106.
8. Rastrigin L.A. Adaptation of complex systems. Riga,1981. 375 p. (in Russian).
9. Bodyanskiy E., Rudenko O. Artificial neural networks: architectures, training, applications. Kharkiv, 2004. 370p. (in Russian).
10. Fainzilberg L. Mathematical methods for evaluating the utility of diagnostic features. Kyiv, 2010. 152 p.
11. PerovaI., BodyanskiyY. Adaptive human machine interaction approach for feature selection-extraction task in medical data mining. *International Journal of Computing*. 2018. Vol. 17. № 2. P. 113–119.
12. Oja E., A simplified neuron model as a principal component analyzer. *J. of Math. Biology*. 1982.№ 15. P. 267–273.
13. Oja E. Neural Network, principal components and subspaces.*Int. J. of Neural Systems*. 1989. P. 61–68.
14. Oja E. Principal component, minor components, and linear neural networks. *Neural Networks*. 1992. no. 5.P. 927–935.
15. Bodyanskiy Ye., Perova I., Zhernova P. Online fuzzy clustering of high - dimensional data based on ensembles in data stream mining tasks.*The Current State of Research and Technology in Industry*. 2019. №1(7). P. 16–24.
16. YamakawaT., Uchino E., Miki T.,Kusanagi H. A neo fuzzy neuron and its applications to system identification and prediction of the system behavior. Proceeding 2nd Int. Conf. on Fuzzy Logic and Neural Networks. *July 1992. Iizuka, Japan*. 1992. P. 477–483.

17. Landim R., Rodrigues B., Silva S., Matos W. A neo-fuzzy-neuron with real- time training applied to flux observer for an induction motor. Proceeding V-th Brazilian Symp. on Neural Networks (Los Alamitos, CA 04th – 06th of Nov, 1998). Los Alamitos, 1998. P. 67–72.
18. MahmoudS., Perova I., Pliss I. Multidimensional neo-fuzzy-neuron for solving medical diagnostics tasks in online-mode. *Journal of Applied Computer Science*. 2017. Vol. 25. № 1. P. 39–48.
19. Perova I., PlissI.Deep hybrid system of computational intelligence with architecture adaptation for medical fuzzy diagnostics. *I.J. Intelligent System and Applications*. 2017. Vol. 7. P. 12–21.
20. Rizzo R. Computational Intelligence Methods for Bioinformatics and Biostatistics. In Lecture Notes in Bioinformatics: 7th International Meeting, CIBIB 2010. (Palermo, Italy, 16–18th of Sept, 2010). Palermo, 2010. 2011. 301 p.
21. Kountchev R. Advances in Intelligent Analysis of Medical Data and Decision Support Systems (Studies in Computational Intelligence). Berlin Heidelberg, 2013. 246 p.
22. Cichocki A.,UnbehauenR. Neural Networks for Optimization and Signal Processing. Stuttgart, 1993. 526p.
23. Perova I.,BodyanskiyY., Adaptive fuzzy clustering based on Manhattan metrics in medical and biological applications. *Newsletter of the National University "Lviv Polytechnic"*. Vol. 826. 2015. P. 8–12.
24. Lughofer E. Evolving Fuzzy Systems — Methodologies, Advanced Concept and Applications. Berlin-Heidelberg, 2011. 454p.
25. Lughofer E. Single pass active learning with conflict and ignorance. *Evolving Systems*. 2012. Vol.3. №4. P. 251-271.
26. Pima Indians Diabetes dataset. Available from URL: http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data. (Last accessed: 01.05.2008.)
27. Dermatology dataset. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/dermatology.data. (Last accessed:01.05.2008.)
28. Parkinson dataset. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.data. (Last accessed:01.05.2008.)
29. Iris dataset. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data. (Last accessed: 01.05.2008.)

Received 10.07.2019

ЛІТЕРАТУРА

1. de Oliveira J., Pedrycz W. Advances in Fuzzy Clustering and its Applications. 2007. 454p.
2. Berka P., Rauch J., Zighed D. Data mining and medical knowledge management cases and applications. New-York, 2009. 440p.
3. Giannopoulou E. Data mining in medical and biological research. New York, 2008. 331p.
4. Karahoca A. Data Mining Applications in Engineering and Medicine. *InTechOpen*. 2012. 336p.
5. Han J., Kamber M. Data Mining: Concepts and Techniques. Amsterdam, 2006. 743p.
6. Gorban A., Kegl B., Wunsch B., Zinovyev A. Principal Manifolds for Data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Engineering. Berlin-Heidelberg-New York, 2007.Vol. 58. 330p.
7. Mulesa P., PerovaI.Fuzzy Spacial Extrapolation Method Using Manhattan Metrics for Tasks of Medical Data Mining. Proceeding of Computer Science and Information Technologies CSIT'2015 (Lviv, 14-17th of Sept, 2015). Lviv, 2015. P. 104–106.
8. РастригинЛ.А. Адаптация сложных систем. *Рига: Зинатне*. 1981. 375с.
9. Бодянский Е.,Руденко О. Искусственные нейронные сети: архитектуры, обучение, применения. Харьков, 2004. 370с.
10. Файнзильберг Л. Математические методы оценки полезности диагностических признаков. Киев, 2010. 152 с.
11. PerovaI., BodyanskiyY. Adaptive Human Machine Interaction Approach for Feature Selection-Extraction Task in Medical Data Mining. *International Journal of Computing*, 2018. Vol. 17. № 2. P. 113–119.

12. OjaE. A simplified neuron model as a principal component analyzer. *J. of Math. Biology*. 1982. № 15. P. 267–273.
13. Oja E. Neural Network, principal components and subspaces. *Int. J. of Neural Systems*. 1989. P. 61–68.
14. Oja E. Principal component, minor components, and linear neural networks.*Neural Networks*. 1992. №5. P. 927–935.
15. Bodyanskiy Ye., Perova I., Zhernova P. Online fuzzy clustering of high - dimensional data based on ensembles in data stream mining tasks *Сучасний стан наукових досліджень та технологій в промисловості*. 2019. №1(7). С. 16–24.
16. YamakawaT., UchinoE., Miki T.,Kusanagi H. A neo fuzzy neuron and its applications to system identification and prediction of the system behavior. Proceedings 2nd Int. Conf. on Fuzzy Logic and Neural Networks. July 1992. Iizuka, Japan. 1992. P. 477–483.
17. Landim R., Rodrigues B., Silva S.,Matos W. A neo-fuzzy-neuron with real- time training applied to flux observer for an induction motor. *Proceeding V-th Brazilian Symp. on Neural Networks* (Los Alamitos, CA, $4^{th} – 6^{th}$ of Nov, 1998). Los Alamitos, 1998. P.67–72.
18. MahmoudS., Perova I.,Pliss I. Multidimensional neo-fuzzy-neuron for solving medical diagnostics tasks in online-mode. *Journal of Applied Computer Science*. 2017. Vol. 25. № 1. P. 39–48.
19. Perova I., PlissI.Deep hybrid System of Computational Intelligence with Architecture Adaptation for Medical Fuzzy Diagnostics. *I.J. Intelligent System and Applications*. 2017. Vol. 7. P. 12–21.
20. Rizzo R. Computational Intelligence Methods for Bioinformatics and Biostatistics. In Lecture Notes in Bioinformatics: 7th International Meeting, CIBIB 2010. (Palermo, Italy, 16-18^{th} of Sept, 2010). Palermo, 2010. 2011. 301 p.
21. Kountchev R. Advances in Intelligent Analysis of Medical Data and Decision Support Systems (Studies in Computational Intelligence). Berlin Heidelberg, 2013. 246 p.
22. Cichocki A., UnbehauenR. Neural Networks for Optimization and Signal Processing. Stuttgart, 1993. 526p.
23. Perova I., BodyanskiyY., Adaptive fuzzy clustering based on Manhattan metrics in medical and biological applications. *Вісник національного університету "Львівська політехніка"*. 2015. Vol. 826. P. 8–12.
24. Lughofer E. Evolving Fuzzy Systems — Methodologies, Advanced Concept and Applications.Berlin-Heidelberg, 2011. 454p.
25. Lughofer E. Single pass active learning with conflict and ignorance.*Evolving systems*. 2012. Vol.3. №4. P. 251–271.
26. Pima Indians Diabetes dataset. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data. (Last accessed: 01.05.2008.)
27. Dermatology dataset. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/dermatology.data. (Last accessed: 01.05.2008.)
28. Parkinson dataset. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.data. (Last accessed: 01.05.2008.)
29. Iris dataset. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data. (Last accessed: 01.05.2008.)

Отримано 10.07.2019

Перова І.Г., канд. техн. наук, доцент,
доцент кафедри біомедичної інженерії
e-mail: rikywenok@gmail.com
Бодянський Є.В,. д-р техн. наук, проф.,
професор кафедри штучного інтелекту
e-mail: yevgeniy.bodyanskiy@nure.ua
Харківський національний університет радіоелектроніки,
пр. Науки, 14, Харків, 61166, Україна.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ПОТОКІВ МЕДИЧНИХ ДАНИХ НА ОСНОВІ НЕЙРО-ФАЗЗІ ПІДХОДУ

Вступ. Підходи інтелектуального аналізу даних в завданнях медичного діагностування мають ряд особливих властивостей, які не дають змогу використовувати такі підходи в класичній формі. З цієї причини було розроблено адаптивні нейро-фаззі системи для завдань оброблення потоку медичних даних в режимі онлайн і алгоритми їхнього навчання. Запропоновані системи можуть обробляти потоки медичних даних в трьох режимах: у контрольованому навчанні, неконтрольованому навчанні та активному навчанні.

Метою статті є розроблення підходу, основаного на адаптивних нейро-фаззі системах, для розв'язання завдань оброблення потоків медичних даних в онлайн-режимі.

Методи. Для оброблення потоків медичних даних використовуються методи обчислювального інтелекту і, перш за все, штучні нейронні мережі, нейро-фаззі системи, нео-фаззі системи, їхнє контрольоване навчання, самонавчання і активне навчання, градієнтні методи оптимізації, методи еволюційних систем.

Результати. Проведено апробацію розробленого підходу в режимі контрольованого навчання за допомоги багатовимірного нео-фаззі нейрона з використанням медичних даних пацієнтів з урологічними захворюваннями. Відсоток помилок під час тестування системи з використанням всього простору ознак становить 11,11 %, з використанням найінформативніших ознак — 6,4 %. Також для діагностування фармакорезистентної форми епілепсії було використано багатовимірний нео-фаззі нейрон, відсоток помилки склав 5,82 %. Проведено апробацію розробленого підходу в режимі активного навчання та асоціації за даними пацієнтів із захворюваннями легень. Для всіх результатів апробації було розраховано критерій ефективності, його значення є задовільними для завдань медичного діагностування в режимі потоку даних.

Висновки. Запропонований підхід дає змогу отримати додаткову інформацію про діагноз пацієнта в умовах обмеженої апріорної інформації про пацієнта.

Ключові слова: *адаптивна система, нейро-фаззі система, інтелектуальне оброблення медичних даних, потік медичних даних.*

Перова И.Г., канд. техн. наук, доцент,
доцент кафедры биомедицинской инженерии
e-mail: rikywenok@gmail.com
Бодянский Е.В., д-р техн. наук, проф.,
профессор кафедры искусственного интеллекта
e-mail: yevgeniy.bodyanskiy@nure.ua
Харьковский национальный университет радиоэлектроники,
пр. Науки, 14, Харьков, 61166, Украина.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ПОТОКОВ МЕДИЦИНСКИХ ДАННЫХ НА ОСНОВЕ НЕЙРО-ФАЗЗИ ПОДХОДА

Введение. Подходы интеллектуального анализа данных в задачах медицинской диагностики обладают рядом особых свойств, которые не позволяют использовать их в классической форме. По этой причине были разработаны адаптивные нейро-фаззи системы для задач обработки потока медицинских данных в режиме онлайн и алгоритмы их обучения. Предлагаемые системы могут обрабатывать потоки медицинских данных в трех режимах: контролируемом обучении, неконтролируемом обучении и активном обучении.

Цель статьи — разработка подхода, основанного на адаптивных нейро-фаззи системах, для решения задач обработки потоков медицинских данных в онлайн-режиме.

Методы. Для обработки медицинских потоков данных используются методы вычислительного интеллекта и, прежде всего, искусственные нейронные сети, нейро-фаззи системы, нео-фаззи системы, их контролируемое, само- и активное обучение, градиентные методы оптимизации, методы эволюционирующих систем.

Результаты. Проведена апробация разработанного подхода в режиме контролируемого обучения с применением многомерного нео-фаззи нейрона при использовании медицинских данных пациентов с урологическими заболеваниями. Процент ошибок при тестировании системы с использованием всего пространства признаков составляет 11,11 %, с использованием наиболее информативных признаков — 6,4 %. Также для диагностики фармакорезистентной формы эпилепсии был использован многомерный нео-фаззи нейрон, процент ошибок составил 5,82 %. Проведена апробация разработанного подхода в режиме активного обучения и ассоциации по данным пациентов с заболеваниями легких. Для всех результатов апробации был рассчитан критерий эффективности, его значения являются удовлетворительными для задач медицинского диагностирования в режиме потока данных.

Выводы. Предложенный подход позволяет получить дополнительную информацию о диагнозе пациента в условиях ограниченной априорной информации о пациенте.

Ключевые слова: *адаптивная система, нейро-фаззи система, интеллектуальный анализ медицинских данных, поток медицинских данных.*