УДК 621.8

V. Lytvynenko

Kherson National Technical University ,
Dept. of Informatics & Computer Science

SYNTHESIS OF THE WAVELET-NEURAL NETWORKS FOR THE CLASSIFICATION OF MASS SPECTRA USING CLONAL ALGORITHM

Abstract. The mass spectrometry spectra are recognized as a screening tool for detecting discriminatory protein patterns. However, the mass spectra represent high dimensional data that have a large number of local maxima (a.k.a. peaks) which have to be analyzed; to tackle this problem we have developeda new three-step strategy. After preprocessing for classification of mass spectra, we use analgorithm clonal selection for synthesis collective binary classifiers in the form of wavelet-neural networks. The results obtained by the analysis of a data set of tumor/healthy samples allowed us to correctly classify more than 99% of samples.

Keywords - mass spectra algorithm, clonal selection, binary classifiers, wavelet neural networks, MALDI-TOF , SELDI-TOF MATLAB, WEKA

1 . Introduction

SELDI-TOF (Surface-Enhanced Laser Desorption and Ionization Time-Of-Flight) technology is considered a modified form of MALDI-TOF (Matrix-Assisted Laser Desorption and Ionization Time-Of-Flight). According to these techniques, proteins are co-crystallizedwith UV-absorbing compounds, then a UVlaser beam is used to vaporize the crystals, and ionized proteins are then accelerated in an electric field. The analysis is then completed by the TOF analyzer. The differences in the two technologies, which reside mainly in the sample preparation, make SELDI-TOF more reliable for biomarkers discovery and other proteomic studies in biomedicine. The proteomic characterization by means of TOF (both SELDI and MALDI) technologies of samples from individuals is considered to carry information about the healthy or pathological state of the individual. In fact, such samples as serum, plasma, and other kinds of extracts contain proteins for which the covalent structure may be modified in specific pathologies, which may induce modifications as glycation or methylation, which imply the addition of a small molecule to the protein, or may alterateand prevent expected modifications. In any of these cases, the proteomes of samples by an healthy individual and an affected individual should be discernible, being their mass profile altered. Therefore, among the thousands of proteins and peptides present in a serum sample, which represent its proteome, few key signals may be significant markers of the pathological state, and their search within the proteome represents a still open field of research. Data produced by mass spectrometry (the spectra) are represented by a (typically) very large set of measures representing the quantity of biomolecules having specific mass-to-charge (m/z) ratio values Given the high dimensionality of spectra, given their different length and since they are often affected by errors and noise, preprocessing techniques are mandatory before any data analysis. After preprocessing (to correct noise and reduce dimensionality), several statistical and Artificial Intelligence based technologies could be used for mining these data.

2 Theoretical Part

Before SELDI or other protein data can be classified it has to go through several steps of what I will call pre-processing. Many different researchers used very different methods in order to process and classify SELDI data. However generalized approach is as follows (Fig 1):

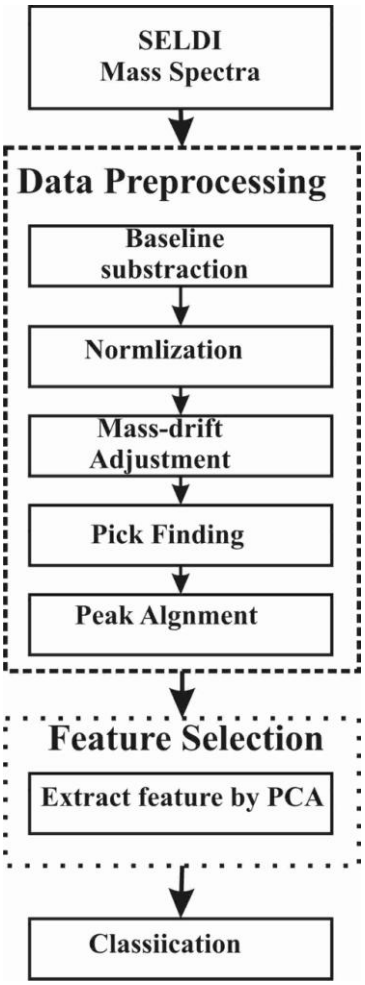

Fig. 1. The mass data analysis pipeline.

For the preprocessing of data and feature selection we used a "caMassClass" library [ ] and Matlab (Bioinformatic Toolbox).

2.1 Data pre-processing and feature selection

Baseline substraction. The baseline is an offset of the intensities of m/z ratios, which happens mainly at low m/z ratios, and varies between different spectra. Base line subtraction uses an iterative algorithm to attempt to remove the baseline slope and offset from a spectrum by iteratively calculating the best-fit straight line through a set of estimated baseline points.

Normalization. The second step of preprocessing is normalizing the multiple spectra. It usually involves cutting first low mass spectra where there is a lot of high frequency high volume noise, which can skew normalization. Afterwards, one finds mean intensity of each spectrum and scales all spectra in such a way as to match all mean intensities.

Normalization is done to make the data independent of experimental variations (like varying amounts of protein, degradation in the sample or variations in the detector sensitivity) and make different spectra comparable. To make this possible, the relative intensities of the spectrograms are normalized. We used in this paper a direct normalization of (1):

$$I_{j_{norb}} = \frac{I_j - I_{\min}}{I_{\max} - I_{\min}} \qquad 1)$$

Mass-drift Adjustment. At this stage an optional step of mass drift adjustment can be performed. Mass drift adjustment attempts to shift the whole spectrum one or more time steps forward or backward if that is going to improve that spectrumcorrelation with other samples. This step is especially useful in case of multiple copies of the same sample, which should have very high correlation. The process is done in the following way:

1. First we extract peak regions from all the spectra. That is done by finding a mean spectrum and identifying peak regions as the ones where mean spectrum is above average .

2. Then we create procedure for matching 2 spectra. First spectrum is not moved and the second is shifted one time step to the right or to the left as long as the correlation between two spectra improves.

3. Then we use the above procedure: first to match all spectra to their copies (if present) and than to match each spectrum to mean spectrum. Since mean spectrum will be changing due to those shifts, the procedure will probably have to be done two or three times before stabilizing. Most of the shifts are assumed to be a few time steps.

Peak finding. The next step would be peak finding and alignment in order to find biomarkers. However it seems like at this stage there are two major different approaches related to SELDI data classification: some research teams use peak alignment to reduce size of the data before classification.

Peak alignment. Peak alignment is used to find out which peaks among different spectra correspond to the same peak (protein, ...). The problem here is the machine measurement error of $0.03\% - 0.06\%$ which has to be considered by the decision if two peaks from different spectra are the same or not. This preprocessing is absolutely necessary because there exist a significant variability between samples in intensity, background and location of the m/z peaks which must be resolved before comparisons can be made across samples.

Feature selection. Feature selection can be also performed with goal of finding a good set of features instead of sets of good features. That is more time consuming approach but with potential high rewards. In our work, each spectrum is represented by a point in spectral-space. The set of all spectral points in spectral-space is dimensionality-reduced using Principal Components Analysis (PCA). In particular, PCA performs a transformation of spectral-space into a lower dimensional space with little or no information loss. A hyperplane, H, is then computed using Linear Discriminant Analysis (LDA). The PCA dimensionality reduced sample points are projected onto H. The hyperplane H maximizes the across-class variance while minimizing the within-class variance of the projected sample points. Thus, the LDA-computed hyperplane H satisfies our exactness criterion. As a result, classification is made easier in this projected space. Now suppose we wish to classify some new (test) spectra (that were not used in training). A test spectrum is first dimensionality-reduced by projecting onto the retained principal components. Next, it is projected onto the hyperplane H. Finally, if the classification confidence is above a threshold then the point is classified into the healthy or disease state.

2.2 Classification

In solving the problem of constructing models using wavelet neural networks, focuses on setting the parameters of wavelet neurons located in the hidden layer networks, which, in essence, is a combinatorial problem.In this paper we propose to use AIS to optimize the parameters of wavelet neural network in the learning process. Architecture of wavelet neural network used in this paper for solving the problem of classification is shown in Figure 2. The neurons of hidden layer - that of the wavelet neurons containing the custom wavelet. As the parameters of a wavelet is used to scale (s) and shift (t) along the time axis. Wavelet neural networks using wavelet family, formed from a single parent by being subjected to operations of scaling and shift. Constructing an optimal family of derivatives of wavelets is the basic task of training the wavelet neural network.

Wavelet neural networks using wavelet family, formed from a single parent by being subjected to operations of scaling and shift. For example, the analytical expression widely used type of wavelet, namely, wavelet "Mexican hat" looks like this:

$$\phi(z) = (z^2 - 1)exp\left(\frac{-z^2}{2}\right). \quad (2)$$

On the basis of the mother wavelet can construct a family of derivatives using the expression:

$$\phi'(z) = \phi\left(\frac{z-t}{s}\right). \quad (3)$$

Constructing an optimal family of derivatives of wavelets is the basic task of training the wavelet neural network.

The output layer contains the usual linear or sigmoidnye neurons and by adjusting their weights (W^l) determines the output of the network.

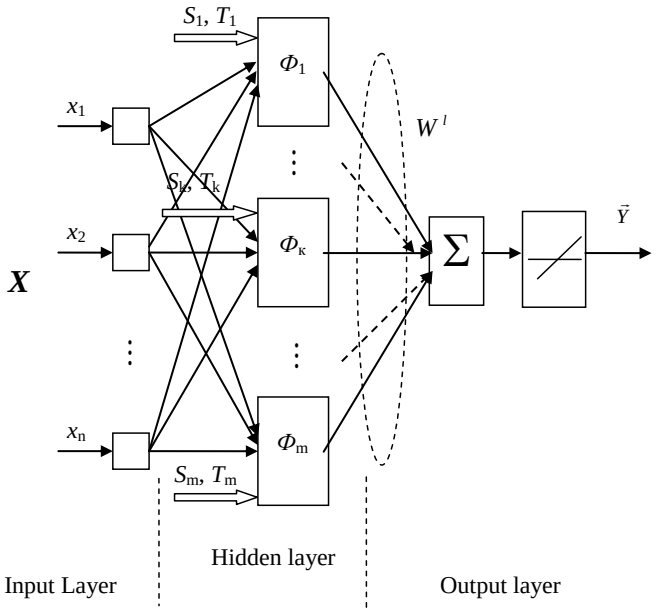

Fig. 2. Generalized architecture of wavelet neural network for solving classification

Основываясь на представленной архитектуре, выход нейронной сети будет определяться следующим образом:

Building on the architecture provided, the output neuronal network will determine the following way:

$$y = \sum_{k=1}^{m} w_k^l \Phi_k\left(x, S_k, T_k\right), \quad (4)$$

where, $x = (x_1, x_2, ..., x_n)$ $x \in X \subset \Re_n$-an arbitrary *n*-dimensional vector of input variables ; $w_k^l \in W^l$ - weight of the linear layer.

Since the simulated process contains the set of input variables, the wavelets are multidimensional. To construct a multi-dimensional wavelet we used the formula:

$$\Phi_k\left(x, S_k, T_k\right) = \prod_{i=1}^{n} \phi_k\left(\frac{x_i - t_{ik}}{s_{ik}}\right), \quad (5)$$

where - $t_{ik} \in T_k$, $s_{ik} \in S_k$ the elements of the displacement vectors and scale for each input variable network.

In the context of the classification problem for the network configuration is to find a function $y: \Re_n \to \Re$ satisfying the equation (1) at $p = 1$. Suppose there is a sample of training data points: $X_1, ..., X_S$, $X_i \in \Re_n$.

If you know the output values for each of these points $d_1, ..., d_S$, $d_i \in \Re$, then every basis function can be centered on one point X_i.

Consequently, in the limiting case, the number of centers, and therefore the hidden layer neurons will be equal to the number of data points of the training sample: $m = S$.

In this paper we propose an approach to determining the number and location of centers of RBF network using clonal selection algorithm (CSA).

In addition, following the idea of an integrated approach to solving the settings of the neural network, we used the CSA as a single (global) tool for searching the optimal values of all configurable parameters. Below is a description of the elements of CSA, which must be adapted to solve this problem [14, 69, 71].

Synthesis of collective neural networks, where each neural network to recognize only a single class in accordance with the algorithm shown in Figure 3.

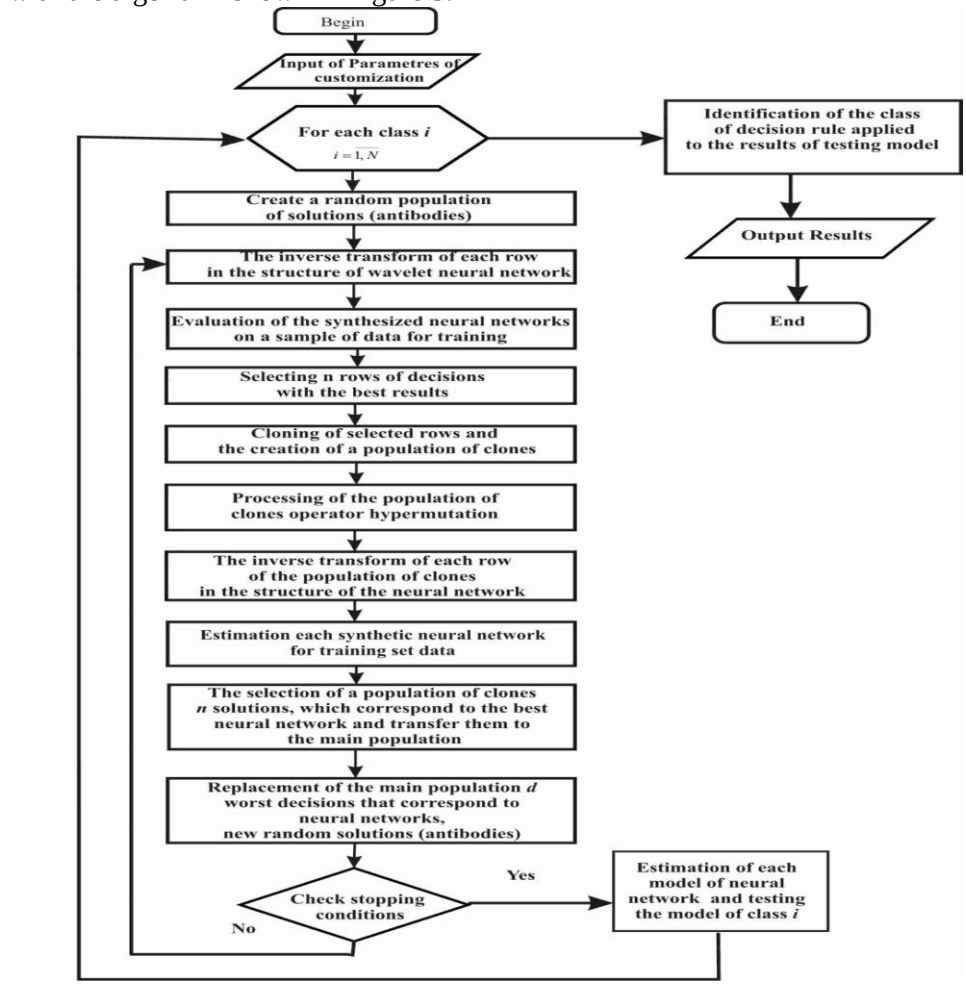

Fig. 3 The synthesis algorithm of wavelet neural network using the algorithm of clonal selection

Based on the architecture of the neural network (Fig. 4), as adjustable parameters are the following:

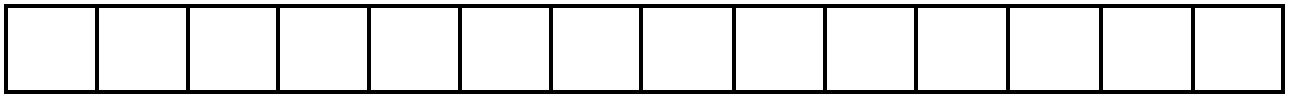

Fig. 4. Coding of adjustable parameters of wavelet neural network in the form of antibodies (chromosomes), where: n - number of inputs; m - the number of neurons; f - on / off neuron (0 or 1); t - setting shifts to the hidden layer neurons; w -synaptic weights of output layer; s - parameter scale for the hidden layer neurons; a - setting output neuron activation function.

In this paper the synthesis of classifying wavelet neural networks, with the classifier in the general case is a function, which vector features of the object makes a decision, precisely what class it belongs to:

$$F : \Re^n \to Y \tag{6}$$

The function ***F*** maps the space of feature vectors in the space of class labels ***Y***. In the case of two classes $Y = \{0,1\}$, '1' corresponds to the case of detection of the desired event '0' - the event is not revealed.

We are considering the option of training with a teacher (supervised learning), when training a classifier is available to us a certain set of vectors $\{x\}$, for which their true identity known to one of the classes.

The binary classification of identifiers classes can be interpreted as the state of the system (active or passive, normal or abnormal), which represented a number of properties.

Definition 1. System state space. A state of the system is represented by a vector of features $x^i = (x_1^i, \ldots, x_n^i) \in [0,1]^n$ Each state is represented by a set $U \subseteq [0,1]^n$. It includes the feature vectors corresponding to all possible states of the system.

Definition 2. Normal subspace (crisp characterization). A set of feature vectors, $Positiv \subseteq U$ represents the normal states of the system. Its complement is called *Negativ* and is defined as *Negativ* = *U* – *Positiv*. In some cases, we will define the *Positiv* (or *Negativ*) set using its characteristic function $\chi_{Positiv}: [0,1]^n \to \{0,1\}$:

–

$$\chi_{Positiv}(\vec{x}) = \begin{cases} 1 & if \quad \vec{x} \in Positiv \\ 0 & if \quad \vec{x} \in Negativ \end{cases} \tag{7}$$

Definition 3. The task of binary classification of wavelet neural network. Given a set of normal samples, $Positiv' \subseteq Positiv$, build a good estimate of the normal space characteristic function ($\chi_{Positiv}$),which, as a result, should have the ability to decide whether the observed state of the system of normal or abnormal.

The entire set of Δ neural networks can be divided into a countable set of subclasses A, determined by the selected topology of wavelet neural network (the number of wavelet neuron network Γ).

Within each class of $WNet_i \subset WNet$ neural networks will be characterized by an additional set of parameters: the number of inputs n, a set of synaptic weights of output layer $W = \{w^i, i = 1, \ldots, p\}$, the number of wavelet neuron network $\Gamma = \{\gamma^i, i = 1, \ldots, p\}$, the parameters of the displacement of neurons $\mathrm{T} = \{\tau^i, i = 1, \ldots, p\}$, the scale parameter for neurons $\Sigma = \{\sigma^i, i = 1, \ldots, p\}$, setting output neuron activation function, the first network $\mathrm{A} = \{\alpha^i, i = 1, \ldots, p\}$.

Thus forming a vector of customizable settings RBF neural network $\theta = \{\Gamma, W, T, \Sigma, \mathrm{A}\}$. A natural criterion for selecting the wavelet neural network is a function defined standard deviation for arbitrary input.

Consequently, the task of synthesis wavelet neural network can be reduced to an optimization problem of type

$$F^* = F(\theta^*) = min F(\theta), \quad a_1 \le x_1 \le b_1, \ldots, a_n \le x_n \le b_n, \tag{8}$$

where the function of F does not impose any restrictions, such as differentiability, Lipschitz condition, continuity.

To solve the problem of multiparameter optimization function of the form (8) use the corresponding operators clonal selection algorithm.

Types of basis functions and the activation function of the linear layer are defined as parameters to AIS. Learning and synthesis collectives of neural networks is performed according to the scheme shown in Fig. 6.

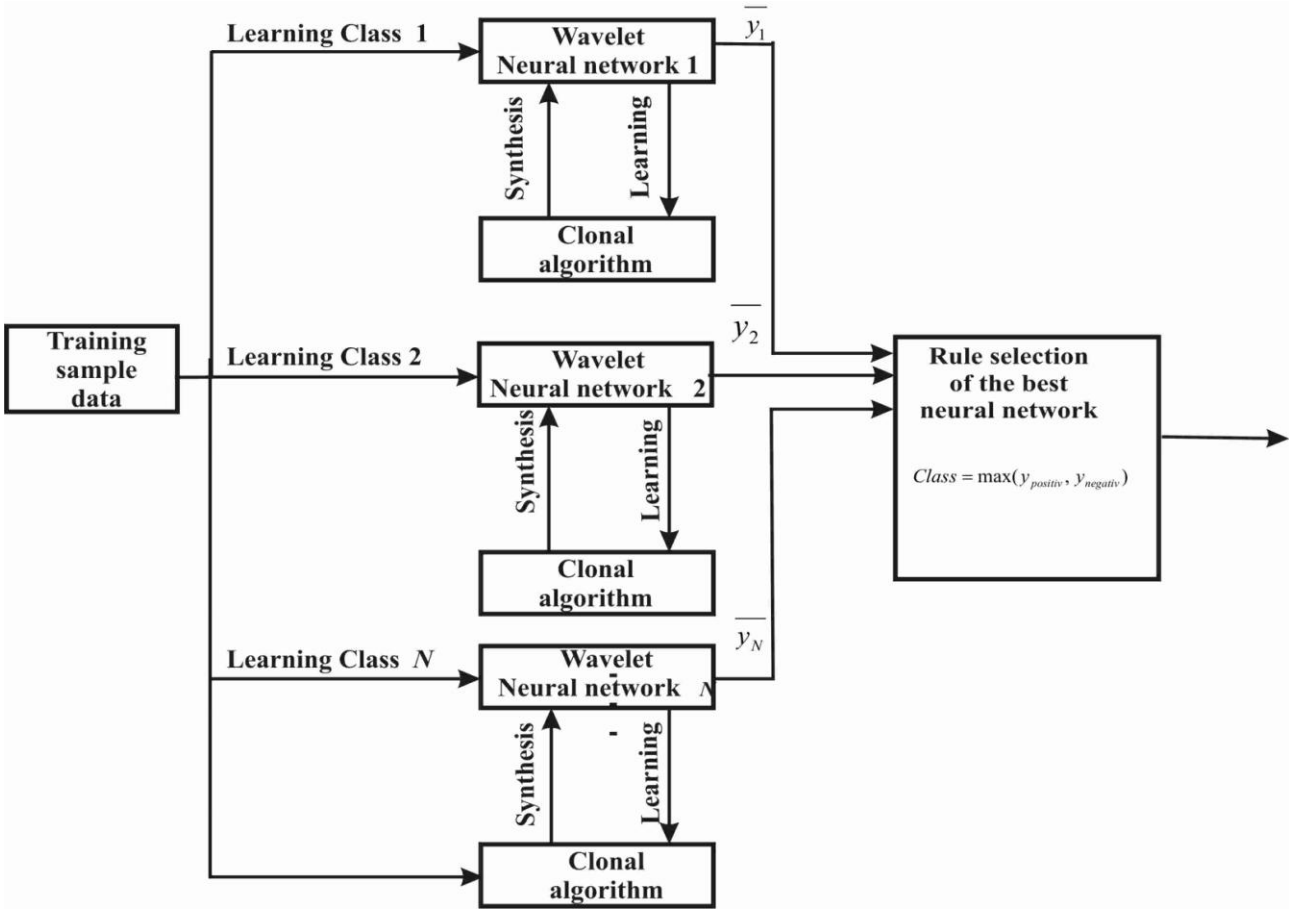

Fig. 6. Synthesis of collective binary classifiers wavelet neural networks

Identification of the system is carried out by the expression (9):

$$Class = \max(y_{positiv}, y_{negativ}) \tag{9}$$

3 EXPERIMENT AND RESULT

3.1 Description of data

The data in this example are from the FDA-NCI Clinical Proteomics Program Databank (http://ncifdaproteomics.com)/. The dataset was downloaded from clinical proteomics program databank (Ovarian Dataset 8-7-02, online data source). This low resolution data were produced by using the WCX2 protein chip and an upgraded PBSII surface enhance laser desorption ionization (SELDI) time of flight (TOF) mass spectrometer. The sample sets consist of 91 controls and 162 ovarian cancers. Preprocessing is an important step in the analysis of MS data. The data were preprocessed by the Matlab The data preprocessing includes calibration, spectral de-noising, baseline correction and normalization, peak detection, and peak quantification. After preprocessing, we obtained 326 discrete peaks (features) for each spectrum. Fig. 1 shows a comparison of two example spectra (one from control, and the other from cancer) before and after preprocessing. As shown in Fig. 1E&F, the differences of MS intensities of control and cancer are signals when m/z is close to zero, and the signals after preprocessing become stronger around the m/z value of 9000. The total 253 samples with 326 features per sample at same locations were used as training and testing data.

3.2 Estimation of efficiency of the algorithm

It is common practice in machine learning and data mining to perform k-fold cross-validation to assess the performance of a classification algorithm. K-fold cross validation is used among the researchers, to evaluate the behavior of the algorithm in the bias associated with the random sampling of

the training data. In k-fold cross-validation, the data is partitioned into k subsets of approximately equal size. Training and testing the algorithm is performed k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Thus, k different test results exist for the algorithm. However, these k results are used to estimate performance measures for the classification system. The common performance measures used in bioinformatic tasks are accuracy, sensitivity and specificity. Accuracy measured the ability of the classifier to produce accurate diagnosis. The measure of the ability of the model to identify the occurrence of a target class accurately is determined by sensitivity.

Specificity is determined the measure of the ability of the algorithm to separate the target class. The classification accuracies for the datasets are calculated as in Eq. 10:

$$Accuracy(Z) = \frac{\sum_{i=1}^{|z|} assess(z_i)}{|Z|} \tag{10}$$

where

$$Assess(z) = \begin{cases} 1, & if\ calssify(z) = z.c \\ 0, & otherwise \end{cases} \tag{11}$$

where:

z = The patterns in testing set to be classified

z.c = The class of pattern z and classify(z) returns the classification of z by classification algorithm

For analysis sensitivity and specificity, the following equations can be used:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{13}$$

where, TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives respectively.

3.2 Experimental Results

For this test were selected following settings of the algorithm (Table 1).

Table 1.

The settings of AIS for the classification problem of the mass spectra

№	Parameter Name	Parameter value
	General parametres	
1	The size of the main population of antibodies	50
2	The maximum number of generations	300
3	The coefficient of selecting the best antibody	0,7
4	Replacement ratio of antibody	0,3
5	The dependence of mutations on the affinity of antibodies	Yes
	Parameters of wavelet neural network	
6	The maximum number of elements of the hidden layer	10
7	Precision parameters	12 bits
8	The range of bias parameter	[-5; 5]
9	The range of the scale parameter	[0,1; 10,0]
10	The range of weights of output layer	[-2; 2]

There have been a comparative study of quality of classification with other known immune systems and algorithms in a software environment WEKA: Immunos-1, Immunos-2, Immunos-99 [6], AIRS1, AIRS2 [7-11], CLONALG [12-13], CSCA [14].

Table 2

The obtained classification accuracy, sensitivity and specificity for signals classification

	Sensitivity	Specifity	Accuracy
Combined algorithm collective artificial Wavelet Network	99.92	98.36	99.15
Immunos-1	98.82	98.07	98.41
Immunos-2	98.71	98.00	98.12
Immunos-99	98.55	97.09	97.11
AIRS1	98.45	97.14	97.15
AIRS2	98.5	97.45	97.65
CLONALG	98.25	97.35	97.25
CSCA	98.25	97.24	97.35

4.Conclusions

The paper shows the results of research carried out by the authors of the combined classification algorithm based on group wavelet-networks for solving the problem of classification of mass spectra. Analysis of the problem solutions demonstrates the effectiveness of this algorithm that uses parallel-distributed organization of calculations. Feasibility of using it explains their high flexibility, the ability to search for parallel, resistant to noise, associative memory, selforganizing, structural flexibility and high adaptive capacity.

1. http://cran.r-project.org/web/packages/caMassClass. 2.M. Katajamaa, et al. Mzmine: Toolbox for processing and visualization of mass spectrometry based molecular protein data. Bioinformatics, 2006. 3. Alexandros Kalousis, J. P., Elton Rexhepaj and Melanie Hilario (2005). Feature Extraction from Mass Spectra for Classification. Knowledge Discovery in Databases: PKDD 2005: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, October 3-7, 2005, Porto, Portugal, Springer Berlin / Heidelberg. 4. Gentzel, M., T. Kocher, etal. (2003). "Preprocessing of tandem mass spectrometric data to support automatic protein identification." Proteomics 3(8): 1597-610.5. G. Ball et al., An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics, 18, 395 (2002). 6. H.W. Ressom , et al., Peak selection from MALDI-TOF mass spectra using ant colony optimization, Bioinformatics, Vol.23:5, 2007, 619:626. 7. Clinical Proteomics Programs National Cancer Institute, Center for Cancer Research.- http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. 8. PROcess R library by Xiaochun Li http://bioconductor.org/repository/devel/package/Source/PROcess_0.9.tar.gz 9. http://cran.r-project.org/web/packages/caMassClass. 10.V.L. Talrose, A.K. Ljubimova Secondary Processes in the Ion Source of a Mass Spectrometer (Reprint from 1952). J. Mass Spectrom. 1998, 33, 502—504. 11.Varmuza, W. Werther, J. Chem. Inf. Comput. Sci. 36 (1996) 323-333. Mass spectral classifiers for supporting systematic structure elucidation. 12.W. McLafferty, S. Y. Loh, D. B. Stauffer, in: H. L. C. Meuzelaar (Eds.), Computer-enhanced analytical spectroscopy, Plenum Press, New York, 1990, p. 163-181. Computer identification of mass spectra. 13. Brownlee, J. Immunos-81. The misunderstood artificial immune system; Technical Report No. 3-01; Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT), Victoria, Australia: 2005. 14 E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg,G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Use of proteomic patterns in serum to identify ovarian cancer, Lancet 359 (2002) 572– 577. 15. Melanie Hilario, Alexandros Kalousis, Markus Mller,and Christian Pellegrini. Machine learning approaches to lung cancer prediction from mass spectra. Proteomics, 3:1716–1719, 2003. 16. MichaelWagner, Dayanand Naik, and Alex Pothen. Protocols for disease classification from mass spectrometry data. Proteomics, 3:1692–1698, 2003. 17. Baolin Wu, Tom Abbott, David Fishman,Walter McMurray, Gil Mor3, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparisonof statistical methods for classication of ovarian cancer using mass spectrometry data. BioInformatics, 19(13), 2003. 18. J. D. Wulfkuhle, L. A. Liotta, and E. F. Petricoin.Proteomic applications for the early detection of cancer. Nature Reviewes, 3:267–275, 2003. 19 Hongtu Zhu, Chang-Yung Yu, and Heping Zhang. Tree-based disease classification using protein data. Proteomics, 3:1673–1677, 2003.