UDC 519.7

V.I. Lytvynenko

APPLICATION OF CLONAL NEGATIVE ALGORITHM TO CANCER CLASSIFICATION WITH DNA-MICROARRAY DATA

Abstract. In the paper, a classification method is proposed. It is based on Combined Clonal Negative Selection Algorithm, which was originally designed for binary classification problems. The accuracy of developed algorithm was tested in an experimental way with the use of microarray data sets. The experiments confirmed that direction of changes introduced in developed algorithm improves its accuracy in comparison to other classification algorithms.

Key words: Negative Selection Algorithm, Clonal Selection Algorithm, Classifier, DNA-Microarray Data, Principal Component Analysis, Wavelet transformation, Feature reduction, Feature selection.

1.Introduction

DNA microarray technology, introduced in 1995–1996, allows the measurement of thousands of gene expression values simultaneously, providing insight into the global gene expression patterns of cells (tissues) being studied [1,2,3]. Despite the need for further technological developments with microarray assays [4], the approach remains powerful for studying the myriad of transcription-related pathways involved in cellular growth, differentiation, and transformation in various organisms. In particular, the ability to measure thousands of gene expressions simultaneously using DNA microarrays has made it possible to investigate genome-wide objective approaches to molecular cancer classification[5].

Empirical microarray data produce large datasets having expression levels of thousands of genes with a very few numbers (upto hundreds) of samples which leads to a problem of “curse of dimensionality”. Due to this high dimension the accuracy of the classifier decreases as it attains the risk of overfitting. As the microarray data contains thousands of genes, hence a large number of genes are not informative for classification because they are either irrelevant or redundant. Hence to derive a subset of informative or discriminative genes from the entire gene set is necessary and a challenging task in microarray data analysis. The purpose of gene selection or dimension reduction is to simplify the

 ISSN 1562-9945

classifier by retaining small set of relevant genes and to improve the accuracy of the classifier. For this purpose, researchers have applied a number of test statistics or discriminant criteria to find genes that are differentially expressed between the investigated classes [15]

A typical DNA microarray data set in tumor tissue c classification studies consists of expression measurements on thousands of genes over

a small number of known tumor tissue samples ($p \gg N$). However, many

standard statistical methodologies for classification and prediction require more samples than predictors. For example, in regression, $N < p$ leads to an ill-posed problem because the ordinary least squares (OLS) solution is not unique. Another example is Fisher's discriminant analysis, wherethe covariance matrix is singular when $N < p$ [5].

It is challenging to use gene expression data for cancer classification because of the following two special aspects of gene expression data. First, gene expression data are usually very high dimensional. The dimensionality ranges from several thousands to over ten thousands. Second, gene expression data sets usually contain relatively small numbers of samples, e.g., a few tens. If we treat this pattern recognition problem with supervised machine learning approaches, we need to deal with the shortage of training samples and high dimensional input features. Recent approaches to solve this problem include artificial neural networks [7],an evolutionary algorithm[8],nearest shrunken centroids [9],and a graphical method [10].

A number of recent publications report on the successful application of support vector machines (SVMs) to the classification of high-dimensional microarray data [11-13].

Therefore, high-dimensional microarray data present a major challenge for these classifiers. However, the algorithms of Artificial Immune System (AIS) have not been widely explored for cancer classification with microarray data. Yet there exist in literature only very few studies in which AIS were applied to microarray classification. Therefore, this study introduced an artificial immune system approach for

cancer detection based on negative selection algorithm (NSA) and Clonal Selection Algorithm (CSA) named Clonal Negative Algorithm (HCNA).

2. Materials and methods

In this study, the microarray data classification was performed in three stages: dimensionality reduction using the Principal Component analysis, Feature extraction using the discrete wavelet transform and classification using hybrid clonal negative algorithm (HCNA).

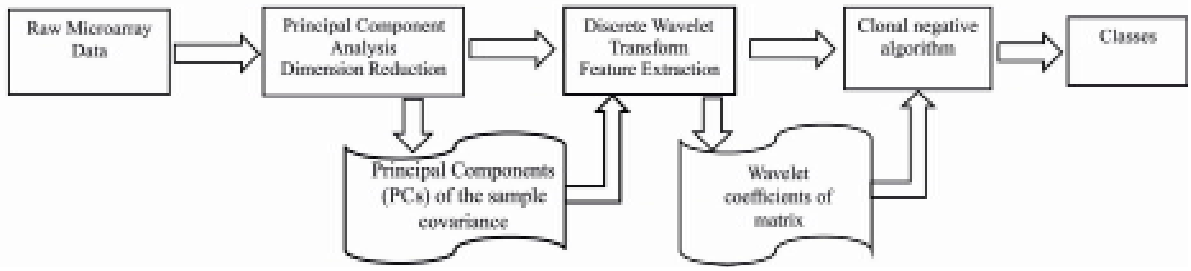

Figure 1 – Structure of the HCNA Classifier

2.1 Dataset

Microarray datasets take the form of expression data matrix where rows represent the genes and columns represent the samples. Each cell in this data matrix is a gene expression value which expresses the gene intensity in the corresponding sample. The expression data matrix will be finally dealt with in the form X_{ij} where; $0 < i \le n_g, 0 < j \le n_s$ and n_s, n_g are the total number of genes, total number of samples respectively as in figure 2. Each expression data matrix will be further divided into two matrices; training data matrix Y_{ik} and test data matrix Z_{ip} where k, p are the number of samples used in the training process, test process respectively and $p + k = n_s$. The training data matrix will be used to train all the used classifiers and their performance will be evaluated using the test data matrix only [16].

$$\mathbf{x}_{ij} = \begin{pmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{1n_s} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & & \mathbf{x}_{2n_s} \\ & & & \\ \mathbf{x}_{n_g 1} & & & \mathbf{x}_{n_g n_s} \end{pmatrix}$$

Figure 2 – Expression data matrix

In this section, the cancer gene expression data sets used for the study are described. These datasets are also summarized below.

ALL/AML Leukemia Dataset. The dataset consists of two distinctive acute leukemias, namely AML and ALL bone marrow samples with 7129 probes from 6817 human genes. The training dataset consists of 3B8 samples (27 ALL and 11 AML) and the test dataset consists of 34 samples (20 ALL and 14 AML).

Colon Dataset. The dataset consists of 62 samples from 2000 genes. The training dataset consists of 42 samples where (30 class1, 12 class2) and the test data set consists of 20 samples (10 class1, 10 class2).

Prostate cancer. Prostate cancer data contains training set of 52 prostate tumor samples and 50 nontumor (labeled as "Normal") prostate samples with 12 600 genes. An independent set of test samples is also prepared, which is from a different experiment. The test set has 25 tumor and 9 normal samples.

2.2 Dimensionality reduction using the Principal Component analysis

Principal component analysis (PCA) is used to search new abstract orthogonal principal components (eigenvectors) which explain most of the data variation in a new coordinate system [17]. Classical PCA is based on the decomposition of a covariance/ correlation matrix (Geladi and Kowalski (1986)) by eigenvalue (spectral) decomposition (EVD) or by the decomposition of real data matrixes using SVD [18].

PCA is a multivariate procedure aimed at reducing the dimensionality of multivariate data while accounting for as much of the variation in the original data set as possible.

This technique is especially useful when the variables within the data set are highly correlated and when there is a higher than normal ratio of explanatory variables to the number of observation. Principal components seeks to transform the original variable to a new set of variables that are (1) linear combinations of the variables in the data set, (2) uncorrelated with each other, and (3) ordered according to the amount of variation of the original variables that they explain [17,19]

PCA is a well-known method of dimension reduction [20]. The basic idea of PCA is to reduce the dimensionality of a data set, while retaining as much as possible the variation present in the original predictor variables. This is achieved by transforming the p original variables $X = [x_1, x_2, \ldots, x_p]$ to a new set of K predictor variables, $T[t_1, t_2, \ldots, t_K]$, which are linear combinations of the original variables. In mathematical terms,

PCA sequentially maximizes the variance of a linear combination of the original predictor variables,

$$u_K = \arg\max_{u'u} \mathrm{Var}(Xu) \quad (1)$$

subject to the constraint $u_i' S_X u_j = 0$, for all $1 \le i \le j$. The orthogonal constraint ensures that the linear combinations are uncorrelated, i.e. $\mathrm{Cov}(Xu_i, Xu_j) = 0, i \ne j$. These linear combinations

$$t_i = Xu_i \quad (2)$$

are known as the principal components (PCs) [21]. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system. The new axes represent the directions with maximum variability and are ordered in terms of the amount of variation of the original data they account for. The first PC accounts for as much of the variability as possible, and each succeeding component accounts for as much of the remaining variability as possible. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem. The projection vectors (or called the weighting vectors) u can be obtained by eigenvalue decomposition on the covariance matrix S_X,

$$S_X u_i = \lambda_i u_i \quad (3)$$

where λ_i is the *i*-th eigenvalue in the descending order for $i = 1, \ldots, K$, and u_i is the corresponding eigenvector. The eigenvalue λ_i measures the variance of the *i*-th PC and the eigenvector u_i provides the weights (loadings) for the linear transformation (projection). The maximum number of components K is determined by the number of nonzero eigenvalues, which is the rank of S_X, and $K \le \min(n, p)$. The computational cost of PCA, determined by the number of original predictor variables p and the number of samples n, is in the order of $\min(np^2 + p^3, pn^2 + n^3)$. In other words, the cost is $O(pn^2 + n^3)$ when $p > n$ [22].

2.3 Discrete wavelet transform-feature extraction

Suppose that the vector $\bar{\xi}_1$ has a sequence consisting of the 2^n points, for some integer $n > 0$. This sequence can be identified with the

 ISSN 1562-9945

next function in the space V^n of piecewise constant functions at equidistant intervals of length $1/2^n$:

$$f(t) = x_1\phi_{n,0}(t) + \ldots + x_{2^n}\phi_{n,2^n-1}(t) \quad (4)$$

where $\phi(t)$- scaling functions of space V^n. The first step in calculating the wavelet decomposition of the sequence $\{x_1, x_2, \ldots, x_{2^n}\}$ is the decomposition of $f(t)$ on the alternative basis of the space V^n, which constitute half of the wavelets $\psi(t)$:

$$f(t) = A_{n-1,0}\phi_{n-1,0}(t) + \ldots + A_{n-1,2^{n-1}-1}\phi_{n-1,2^{n-1}-1}(t) + D_{n-1,0}\psi_{n-1,0}(t) + \ldots + D_{n-1,2^{n-1}-1}\psi_{n-1,2^{n-1}-1}(t) \quad (5)$$

where A - approximation coefficients, defining coarse low-frequency component of the original signal, D - detail coefficients, defining the high-frequency component of the original signal. The next step of the conversion process is the use of the same basic conversion the members of (2), containing the approximation coefficients. Detail coefficients at the same time remain unchanged. Block diagram of the wavelet decomposition is presented in Figure 3.:

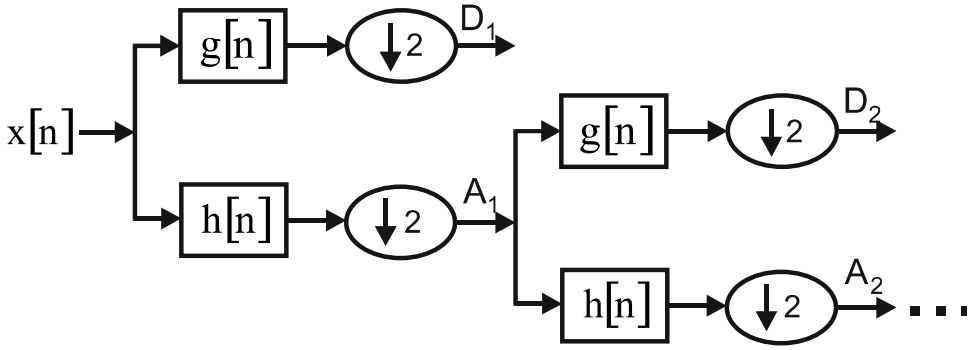

Figure 3 – Structural diagram of a discrete wavelet decomposition of the signal; $g[n]$ - high frequency transmit filter, $h[n]$- low frequency filter transmitting.

The Data used in this research were analyzed into the details D1-D2 and one final approximation, A2. Our previous studies [23] have shown that the the smoothing feature of the Daubechies wavelet of order 13 (db13) made it more suitable to detect changes of the microarray data. Hence in our research, we used the db13 to compute the wavelet coefficients of the microarray data.

For receiving approximation and detail coefficients using orthogonally and normalization the property. In the basic functions, all scaling functions, as well as the wavelet functions are orthogonal to each other, in addition, each $\psi(t)$ and each $\phi(t)$ are normal. Multiplying both parts of (2) on $\phi_{n-1,j}(t)$ and we will integrate on t from 0 to 1. As a result, we will receive

$$\int_0^1 f(t)\phi_{n-1,j}(t)dt = A_{n-1,j} \tag{6}$$

Now we substitute the right-hand side of equation (1) instead of f (t) in (3). If $j = 0$ at the left side of (3) will be equal to:

$$\int_0^{1/2^n} x_1\sqrt{2^n}\sqrt{2^{n-1}}dt + \int_{1/2^n}^{2/2^n} x_2\sqrt{2^n}\sqrt{2^{n-1}}dt = (x_1 + x_2)\left(\frac{1}{\sqrt{2}}\right)2^2\left(\frac{1}{2^n}\right) = \tag{7}$$

$$= \frac{x_1 + x_2}{\sqrt{2}}$$

Combining (3) and (4) at $j = 0$, we will receive:

$$A_{n-1,0} = \frac{x_1 + x_2}{\sqrt{2}} \tag{8}$$

The remaining coefficients $a_{n-1,j}, j = 0,\ldots,2^{n-1} - 1$ are computed in the same way:

$$A_{n-1,j} = \frac{x_{2j+1} + x_{2j+2}}{\sqrt{2}} \tag{9}$$

Similarly, using the properties of orthogonally and normalization of the functions $\psi_{n-1,j}$, we can calculate the detailing coefficients $D_{n-1,j}$ using the following formula:

$$D_{n-1,j} = \frac{x_{2j+1} - x_{2j+2}}{\sqrt{2}} \tag{10}$$

Ultimately, we obtain the matrix approximation and detail coefficients at a given level of decomposition.

The computed discrete wavelet coefficients provide a compact representation. In order to further decrease the dimensionality of the extracted feature vector, statistics over the set of the wavelet coefficients

 ISSN 1562-9945

are used. The following statistical features were used to represent the time-frequency distribution of the microarray data:

- Maximum of the wavelet coefficients in each subband
- Minimum of the wavelet coefficients in each subband
- Mean of the wavelet coefficients in each sub-band
- Standard deviation of the wavelet coefficients in each sub-band.

2.4 Artificial immune algorithms

In the 1990s, Artificial Immune System (AIS) emerged as a new computational research filed inspired by simulation of biological behavior of Natural Immune System (NIS). The NIS is a very complex biological network with rapid and effective mechanisms for defending the body against a specific foreign body material or pathogenic material called antigen .

The Artificial Immune Systems, as defined by de Castro and Timmis [24] are: "Adaptive systems inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving". However AIS are one of many types of algorithms inspired by biological systems, such as neural networks, evolutionary algorithms and swarm intelligence. There are many different types of algorithms within AIS and research to date has focused primarily on the theories of immune networks, clonal selection and negative selection. These theories have been abstracted into various algorithms and applied to a wide variety of application areas such as anomaly detection, pattern recognition, learning and robotics [25].

Negative selection algorithm. The negative selection of T-cells is responsible for eliminating the T-cells whose receptors are capable of binding with self-peptides presented by self-MHC molecules. This process guarantees that the T-cells that leave the thymus do not recognize any self-cell or molecule. Forrest et al. [26] proposed a change detection algorithm inspired by the negative selection of T-cells within the thymus. This procedure was named as negative selection algorithm and was originally applied in computational security. A single type of immune cell was modelled: T-cells were represented as bit strings of length L. The negative selection algorithm of Forrest and collaborators is simple [26]. Given a set of self-peptides, named self-set S, the T-cell receptors will have to be tested for their capability of binding the self-peptides. If

a T-cell recognizes a self-peptide – it is discarded, else it is selected as an immune-competent cell and enters the available repertoire A.

The idea of negative selection algorithm is to generate a set of detectors in a complementary set of **N** and then to use these detectors for binary classification as "Self" or "Non-Self". Formally, the negative selection algorithm can be represented as [27-28]:

$$\mathrm{NegA\,lg} = \left(\Sigma^{\mathrm{L}}, \mathrm{L}, \mathrm{S}, \mathrm{N}, \mathrm{r}, \mathrm{n}, \mathrm{s}, \mathrm{pr}\right) \qquad (11)$$

where Σ^{L} denotes shape-space; L is receptor length; **S** is "Self" detector set; **N** is "Non-Self" detector set; r denotes cross-reactive threshold; n is total number of appointed detectors; s is detector set size; pr denotes rule matching rows in adjacent positions.

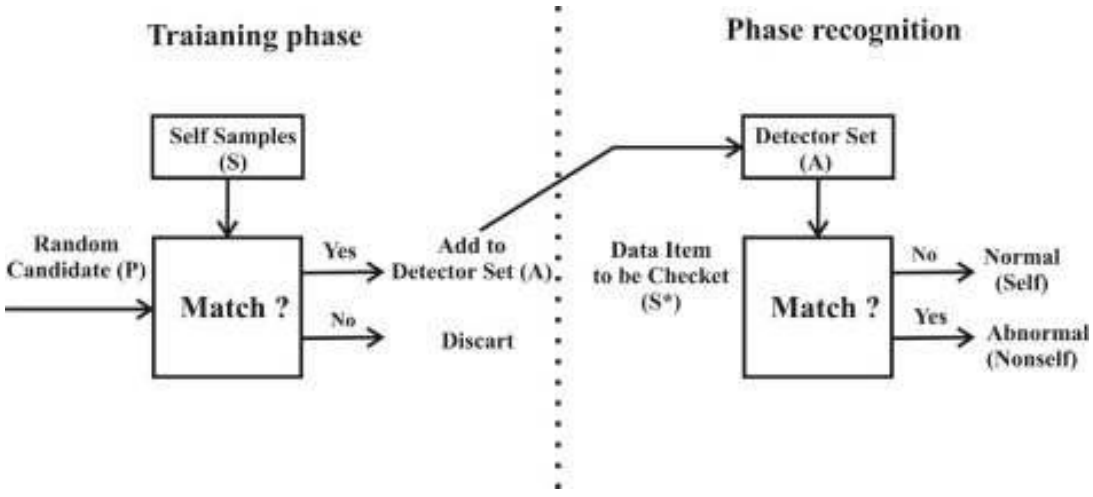

Figure 4 – Negative selection algorithm [29]

The negative selection algorithm can be summarized as follows:

• *Initialization:* randomly generate strings and place them in a set **P** of immature T-cells, assuming all the molecules (receptors and self-peptides) are represented as binary strings of the same length ***L***.

• *Affinity evaluation:* determine the affinity of all T-cells in V with all elements of the self set **S**.

• *Generation of the available repertoire:* if the affinity of an immature T –cell with at least one self-peptide is greater than or equal to a give cross reactive threshold, then the T-cell recognizes this self-peptide and has to be eliminated (negative selection); else the T-cell is introduced into the available repertoire **A**.

The process of generating the available repertoire in the negative selection algorithm was termed learning phase. The algorithm is also composed of a monitoring phase. In the monitoring phase, a set $\mathbf{S}^{*}$of

protected strings is matched against the elements of the available repertoire A. The set **S*** might be the own set **S**, a completely new set, or composed of elements of **S**. If recognition occurs, then a non-self pattern (string) is detected.

It is well known, that the algorithm of negative selection (NS) has the some restrictions and limitations [29]. When it is not appropriate, for example, the number of self samples is small and sparse.

Some limitations of the binary-string representation in NS algorithms are as follows:

- binary matching rules are not able to capture the semantics of some complex self/non-self spaces,
- it is not easy to extract meaningful domain knowledge,
- in some cases a large number of detectors are needed to guarantee better coverage (detection rate),
- it is difficult to integrate the NS algorithm with other immune algorithms,
- the crisp boundary of “self” and “non-self” may be very hard to define.

In real-valued representation the detectors are represented by hyper-shapes in n-dimensional space. The algorithms use geometrical spaces and use heuristics to distribute detectors in the non-self space.

Some limitations of the real-valued representation in NS algorithms are:

- the issue of holes in some geometrical shapes, and may need multi-shaped detectors,
- curse of dimensionality,
- the estimation of coverage,
- the selection of distance measure.

During our experiments it has been established that generation of set of detectors in at training phase occurs casually owing to what it is in advance impossible to define is minimum necessary quantity of detectors which will provide the maximum quality of recognition. The increase in quantity of detectors conducts to delay of a phase of recognition, and its reduction – to deterioration of work of algorithm since the probability of formation of the "cavities" which are areas in space of “Non-self” which are not distinguished by any of detectors increases.

Thus, a problem of the given research is working out of an advanced method of generation of the detectors, capable to adaptive selection of their options, quantity and an arrangement.

Clonal selection algorithm. Today the algorithm CLONALG exists in two forms [24]: (1) for optimization problems solving, and (2) for solving problems of classification and pattern recognition. Basic clonal selection algorithm [24]., named CLONALG, works as in Fig. 5

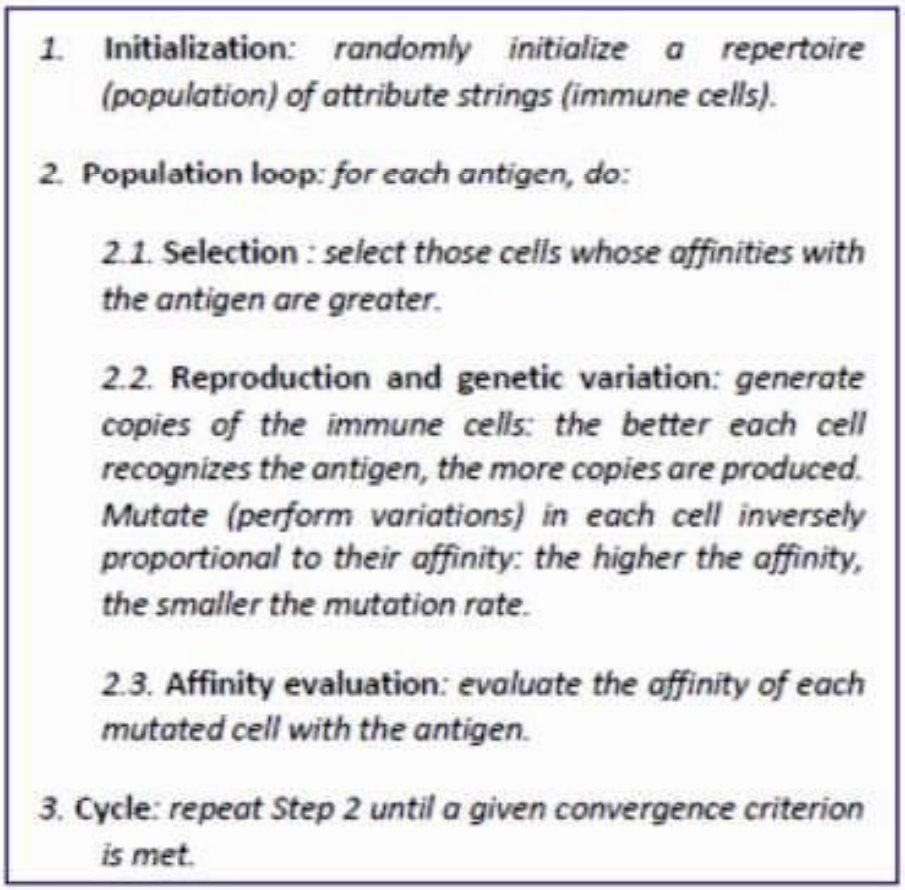

Figure 5 – Standard clonal selection algorithm

Formally algorithm of clonal selection can be represented as [30]:

$$\text{CLONALG} = (P^l, G^k, l, k, m_{Ab}, \delta, f, I, \tau, AG, AB, S, C, M, n, d) \quad (12)$$

where P^l is space of search (space of forms); G^k is space representation; l is the length of vector of attributes (dimension of space of search); k is the length of antibody receptor; m_{Ab} is dimension of population of antibodies; δ is the expression function; f is the affinity function; I is the function of initialization of the initial population of antibodies; τ is the condition of completion of algorithm work; AG is the subset of antigenes; AB is population of antibodies; S is the opera-

tor of selection; C is the operator of cloning; M is the mutation operator; n is the number of the best antibodies selected for cloning; d is the number of the worst antibodies subjected to substitution for new ones.

The process of converting a population of antibodies by clonal selection algorithm can be represented as a sequence of the following statements:

$$AB_t \xrightarrow{\text{Selection(S)}} G_S \xrightarrow{\text{Cloning(C)}} G_C \xrightarrow{\text{Mutation(M)}} G_M \xrightarrow{\text{Repeat mutation(S)}} G_S \xrightarrow{\text{Replacement(d)}} AB_{t+1}, \quad (13)$$

where t - is the number of generation, AB - is the population of antibodies (detectors), G_S - the subset of selected best antibodies, G_C - is the subset of clones, G_M - is the subset of clones after mutation.

Combined clonal and negative selection algorithm. The classifier presented in this paper is based on the hybridization process of negative selection with clonal selection, and was designed to solve problems of classification to many classes. Concept of classification is used in terms of supervised learning, which allows categorizing objects into known groups using training set prepared beforehand. The main task of every classifier based on supervised learning is to create an internal representation of classes (in the form of a function, set of rules or any other). It acquires it during training. When the training is completed the classifier is ready to produce an answer to any (known or unknown) pattern given subsequently.

In this study the efficiency of immune classifiersis researched, when as the classifier, in general, is a function that for attributes vector of object shall decide to which class it belongs [27]:

$$F : \Re^n \to Y. \quad (14)$$

The function *F* represents the space of sign vectors in the space of the class labels *Y*. In the case of two classes $Y = \{0,1\}$, '1' corresponding case of the detection event, '0' - the event is not detected. We consider the variant of training with a teacher (supervised learning), when the classifier training available to us a set of vectors $\{x\}$ for which is known their valid membership in one of the classes.

In developing this model treated the problem of developing an improved method of generation of detectors capable to adaptively select

their debugging and localization. This modification propose in a phase of training to optimize coverage by detectors set of "Non-self" via the mechanism of clonal selection. For the solution of the problem is introduced following submission of antibodies (Fig.6).

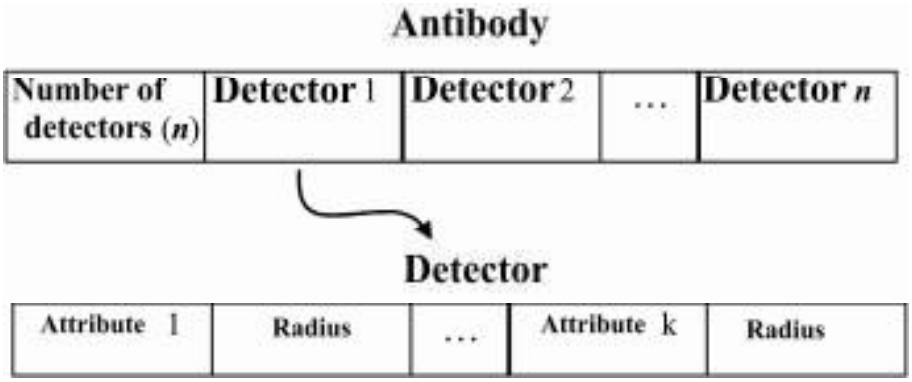

Fig. 6. View of clonal negative model antibody [28]

In this view, attributes are the coordinates of the center of the detector and the radius - the threshold sensitivity of the detector (cross-reactive threshold). Thus, each antibody encodes a possible alternative arrangement of detectors in space "Non-self", that option schemes covering. By manipulating the population of antibody-like structure, is the best option scheme covering. This algorithm is described in detail in [28].

3. Results and discussion

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 medical diagnosis 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 meas-

 ISSN 1562-9945

ure of the ability of the algorithm to separate the target class. The classification accuracies for the datasets are calculated as in Eq. 15:

$$\text{Accuracy}(Z) = \frac{\sum_{i=1}^{|z|} \text{Assess}(z_i)}{|Z|} \tag{15}$$

while

$$\text{Assess}(z) = \begin{cases} 1, & \text{if calssify}(z) = z.c \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

where z denotes the patterns in testing set to be classified, $z.c$ is the class of pattern z, $classify(z)$ returns the classification of z by classification algorithm.

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

$$\text{Sensitivit y} = \frac{TP}{TP + FN} \tag{17}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{18}$$

where TP, TN, FP i FN denote respectively true positive, true negative, false positive and false negative classification.

In order to compare the efficiency of the proposed method in predicting the class of the cancer microarray data we have used three standard datasets such as All/AML Leukemia, Colon Dataset, Prostate cancer. All the datasets is binary class datasets. The feature selection process proposed in this paper has two steps. First the microarray data is decomposed by factor analysis optimally choose the discriminate feature set then using Discrete wavelet transform into level 4 using db13 wavelet to get the approximation coefficients as the extracted feature set. The performance of the proposed feature extraction method is analyzed with the well studied neural network classifiers such as MLP and RBFNN. The leave one out cross validation (LOOCV) test is conducted by combining all the training and test samples for both the classifiers with all the three datasets and the results are listed in Table 1. For this Data the performance of HCNA is comparable to RBFN and MLP.

Table 1

Comparison study of classification accuracy, sensitivity and specificity of HCNA with MLP and RBFN classifiers

Dataset	Method	Classification Accuracy	Sensitivity	Specificity
ALL/AML Leukemia	MLN	91.3%	98.21	97.01
	RBFN	98.4%	98.55	97.25
	HCNA	100%	99.06	99.20
Colon Dataset	MLN	94.5%	97.44	96.25
	RBFN	97.6%	98.10	97.10
	HCNA	99.7%	98.80	99.45
Prostate cancer	MLN	98.8%	98.03	97.21
	RBFN	97.2%	98.80	98.00
	HCNA	100%	99.10	99.60

The performance of the proposed method is also compared with those obtained by the recently reported methods and the results are listed in Table 2-4. The existing methods also used the cross validation test on the datasets. From Tables 2-4 it reveals that our method is equivalent to the counterparts with the advantage of reduced computational load. Weka [31]. Table 5 shows the decomposition stages upto 4th level by using db13 in discrete wavelet transform.

Table 2.

Comparison study of accuracy of Colon Dataset

Methods	Classification accuracy
Bayes Network	85.3%
Naive Bayes classifier	60.1 %
Multinomial logistic regression model	74.2%
Support Vector Classifier	94.3%
Class for doing classification using regression methods	91.8%
Simple Decision Table Majority Classifier	89.3%
1R classifier	73.9%
C4.5 decision tree	94.5%
Forest of Random Trees	97.4%
Factor Analysis + Wavelet + HCNA	99.7%

 ISSN 1562-9945

Table 3.

Comparison study of accuracy of ALL/AML Leukemia dataset

Methods	Classification accuracy
Bayes Network	87.6%
Naive Bayes classifier	64.2 %
Multinomial logistic regression model	77.4%
Support Vector Classifier	91.3%
Class for doing classification using regression methods	97.6%
Simple Decision Table Majority Classifier.	92.5%
1R classifier	70.9%
C4.5 decision tree	96.5%
Forest of Random Trees	97.6%
Factor Analysis + Wavelet + HCNA	100%

Table 4.

Comparison study of accuracy of Prostate cancer

Methods	Classification accuracy
Bayes Network	91.7%
Naive Bayes classifier	69.2 %
Multinomial logistic regression model	80.5%
Support Vector Classifier	99.4%
Class for doing classification using regression methods	91.7%
Simple Decision Table Majority Classifier.	89.4%
1R classifier	69.2%
C4.5 decision tree	97.1%
Forest of Random Trees	98.8%
Factor Analysis + Wavelet + HCNA	100%

Table 5

Reduction details of the dataset

Dataset	**Original Dimension**	**Factor Analysis**	**DWT Db 13 Level 4**
Colon	62×2000	62×700	62×180
ALL/AML Leukemia	72×7129	72×700	72×180
Prostatecancer	136×12600	136×700	136×180

4. Conclusion

In this paper we have presented a hybrid feature extraction method using the Factor analysis in conjunction with wavelet transform to effectively select the discriminative genes on microarray data. A simple HCNA based classifier has also been introduced to classify the microarray samples efficiently. The comparison results elucidated that the proposed approach is an efficient method which performs better than the existing methods. Besides it has reduced computational complexity.

REFERENCES

1. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E.L., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 14, 1675–1680.
2. Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.
3. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., Davis, R.W., 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. USA 93, 10614–10619.
4. Nguyen, D.V., Arpat, A.B., Wang, N., Carroll, R.J., 2002d. DNA microarray experiments: biological and technological aspects. Biometrics 58, 701–717.
5. D.V. Nguyen, D.M. Rocke On partial least squares dimension reduction
for microarray-based classi'cation: a simulation study / Computational Statistics & Data Analysis 46 (2004) 407 – 425
6 Jayakishan M., Barik R. C., Ranjani P. M., S. K. Pradhan C.F. G. Dash Analysis and Wavelet Transform Method for Tumor Classification Using Gene Expression Data/ I.J. Information Technology and Computer Science, 2012, 9, 73-79
7. J. Khan, J. S. Wei, M. Ringner, L. H. Saal and M. Ladanyi et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine 7 (2001) 673–679.
8. J. M. Deutsch, Evolutionary algorithms for finding optimal gene sets in microarray prediction, Bioinformatics19 (2003) 45–52.

 ISSN 1562-9945

9. R. Tibshirani, T. Hastie, B. Narashiman and G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, in Proc. Natl. Acad. Sci. USA, Vol. 99 (2002), pp. 6567–6572.
10. E. Bura and R. M. Pfeiffer, Graphical methods for class prediction using dimension reduction techniques on DNA microarray data, Bioinformatics 19 (2003) 1252–1258.
11. Ramaswamy S., Tamayo P., Rifkin R., Mukherjee S., Yeang C.H., Angelo M. Ladd C., Reich M., Latulippe E., Mesirov J.P., Poggio T., Gerald W., Loda M., Lander E.S., Golub T.R., Multiclass cancer diagnosis using tumor gene expression signatures, Proc. Natl. Acad. Sci. USA. 98(26):15149-15154, (2001).
12. Brown M.P.S., Grundy W., Lin D., Cristianini N., Sugnet C., Furey T., Ares M., Jr., Haussler D., Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. USA. 97(1):263-267, (2000).
13. Mukherjee S., Classifying microarray data using support vector machines, in Berrar D., Dubitzky W., Granzow M. (eds.), A Practical Approach to Microarray Data Analysis, Kluwer Academic Publishers, Boston, pp. 166-185, (2002).
14. Berrar, D.P., Downes, C.S., Dubitzky, W.: A Probabilistic Neural Network for Gene Selection and Classification of Microarray Data. ;In IC-AI(2003)342-352
15. Meher J., Barik R. C., Panigrahi M. R., Pradhan S.K., Dash G. Cascaded Factor Analysis and Wavelet Transform Method for Tumor Classification Using Gene Expression Data / MECS I.J. Information Technology and Computer Science, 2012, 9, 73-79
16. D. A. Salem, R.A. A. Seoud, H. A. Ali MGS-CM: A Multiple Scoring Gene Selection Technique for Cancer Classification using Microarrays /International Journal of Computer Applications (0975 – 8887) Volume 36– No.6, December 2011 p. 30 – 37.
17. Dong Hoon Lim Principal Component Analysis using Singular Value Decomposition of Microarray Data/ World Academy of Science, Engineering and Technology 81, 2013, p.56-58
18. Wall M.E., Dyck P.A., Brettin T.S.(2001). SVDMAN – singular valuedecomposition analysis of microarray data. Bioinformatics 17:566-68.
19. Everitt, B. and T Hothorn. (2011). An Introduction to Applied Multivariate Analysis with R (Use R!). Springer, New York, NY.
20. Jolliffe, I.T. (1986). Principal Component Analysis. Springer, New York.

21. Massey, W.F. (1965) Principal components regression in exploratory statistical research. Journal of American Statistical Association, 60, 234-246.
22. Jian J. Dai, Linh Lieu, David Rocke Dimension Reduction for Classification with Gene Expression Microarray Data/ Stat Appl Genet Mol Biol. 2006;5:Article6. Epub 2006 Feb 24
23. Babichev S.A., Babenko N.I., Dydik A.A. Lytvynenko V.I., Fefelov A.A., SV Shkurdoda Filtration chromatogram with pomoshchju veyvelet-analysis with Using Criteria entropyy (in Russion)/ Regional mizhvuzovskyy collection of scientific papers Issue 6 (71) Dnipropetrovsk 2010 C-17-3224.
24. De Castro, L.N. and J. Timmis, 2002. Artificial Immune Systems: A New Computational Intelligence Approach. Springer, London, England, pp: 357.
25. Smith, S.L. and J. Timmis, 2008. An Immune Network Inspired Evolutionary Algorithm for the diagnosis of Parkinson's disease. Biosystems, 94: 34-46. DOI: 10.1016/j.biosystems.2008.05.024
26. Forrest, S., A.S. Perelson, L. Allen and R. Cherukuri, 1994. Self-nonself discrimination in a computer. Proceedings of the Symposium IEEE Computer Society on Research in Security and Privacy, May 16-18, IEEE Xplore Press, Oakland, CA., USA., pp: 202-212. DOI: 10.1109/RISP. 1994.296580
27. Lytvynenko V. I.: Comparative experimental study of a modified negative selection algorithm and clonal selection algorithm negative for solving classification (in Russian), Vestnik Kherson National Technical University, nr 4(33), 2008, pp. 7-14.
28. Lytvynenko V. I.: Immune classifier for solving binary classification - Theoretical Aspects (in Russian), System technologies, nr 1(42), 2006, pp. 32-47.
29. D. Dasgupta. Advances in Artificial Immune Systems. IEEE Computational Intelligence Magazine. November, 2006. - P.40-49.
30. Bidyuk P.I. Litvinenko V.I., Fefelov A.O. Formalizatsiya metodiv pobudovi shtuchnih imunnih systems (in Ukrainian)// Naukovi Visti Natsionalnogo tehnichnogo universitetu Ukraine "Kiïvsky politehnichny institut" /, 2007, № 1, 2007,-C. 29-41.
31. http://www.cs.waikato.ac.nz/ml/weka/