Applied Medical Informatics
Original Research Vol. 35, No. 4/2014, pp: 1-10

Performance Study of Cancer Selection/Classification Algorithms Based on Microarray Data

Hamidreza SABERKARI*, Mousa SHAMSI, Faegheh GOLABI, Mohammad Javad AMOSHAHY, Mohammad Hossein SEDAAGHI

Genomic Signal Processing Laboratory, Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran.
E-mails:{h_saberkari, shamsi, f_golabi, m_amoshahy, sedaaghi}@sut.ac.ir

* Corresponding Author; Tel. +98-911-2362692

Received: 20 October 2014/Accepted: 18 December 2014/ Published online: 26 December 2014

Abstract
Microarray data has an important role in detecting and classifying all types of cancer tissues. In cancer researches, relatively low number of samples in microarray has always caused some problems in designing classifiers. So, microarray data is preprocessed through gene selection techniques and the genes which contain no information is discarded. Basically, a proper gene selection method can effectively improve the efficiency of diseases (cancers) classification. The purpose of this article is to compare different extraction algorithms of informative genes and also their different classification algorithms. First, ReliefF algorithms, information gain and normalized mutual information are introduced as algorithms used in order to extract feature and their features are noted. Then three classification algorithms, two proposed Bayesian Linear Discriminate Analysis (BLDA), Modified Support Vector Machine (υ-support Vector Machine) algorithms and Probabilistic Neural Network are compared in terms of classification accuracy. Implementation results show that combinational algorithm of normalized mutual information and BLDA classifier has best performance among other raised methods. So that, with applying this algorithm, classification accuracy in blood cancer data base is 95.34 percent.

Keywords: DNA (deoxyribonucleic acid); Microarray; Support Vector Machine; Gene.

Introduction

Microarray technology was born in 1996 and has been known as Deoxyribonucleic Acid (DNA) [1] arrays, gene chips, and DNA chips. Important viewpoints of gene performance can be obtained from gene expression profile. The gene expression profile is a process that determines the time and location of the gene expression. Genes are turned on or off in particular situations. For example, a DNA mutation may change the gene expression, resulting in tumor or cancer growth Moreover, sometimes the expression of a gene affects the other genes' expression. Microarray technology is one of the latest developments in the field of molecular biology which permits supervision on the expression of hundreds of genes at the same time and just in one hybridization test. Using the microarray technology, it is possible to analyze the pattern and gene expression level of different types of cells or tissues. In addition to the scientific potential of this technology in the fundamental study of gene expression, namely gene adjustment and solidarity, it has an important application in medical and clinical research. For example, by comparing the gene expression in normal and abnormal cells, the microarray technique can be used to detect the abnormal genes for remedial

medicine or evaluating their effects [1].

A microarray has thousands of spots, each of them consisting of different identified DNA strands, named probes. These spots are printed on glass slides by a robotic arrayer. Two types of microarray have the most application; microarrays based on complementary DNA (cDNA) and Oligonucleotide array which are briefly named Oligo [1]. In cDNA array method, each gene is represented by a long strand (between 200-500 bps). cDNA is obtained from two different samples; test sample and reference one which are mixed in an array. Test and reference samples are denoted with red and green fluorescence, respectively (these two samples which have different wave lengths, are named Cy3 and Cy5) [2]. If the two cDNA samples consist of trails which are complement of the DNA probe, then the cDNA sample is mixed with spot. cDNA samples which are found their own complementary probe, are hybrid on array and the remainder of samples are washed and then the array is scanned by the laser ray for determining the scaling of sample joined to spot. Hybrided microarray is scanned in red and green wavelength and two images are obtained. Fluorescent intensity ratio in each spot demonstrates the DNA trail relative redundancy in two mixed cDNA samples on that spot. With surveying the gene expression levels ratio in two images, Cy3 and Cy5, gene expression study is done. Gene expression dimension can be the logarithm of the red to green intensity ratio [3].

Microarray data is similar to a matrix with thousands of columns and hundreds of rows, each row and column representing a sample and a gene, respectively. A gene expression level is related to the generated protein value. Gene expression provides a criterion for measuring the gene activity under the special biochemical situation. The gene expression is a dynamic process that can vary in transient or steady-state form. Thus, it can resound momentary and insolubility variations in the biologic state of cells, tissues and organisms [4]. Using the microarray technology, it is possible to analyze the pattern and gene expression level of different types of cells or tissues.

General Process of Feature Selection and Microarray Data Classification

One of the main problems in microarray data analysis is gene selection process in which a few number of genes are selected before classification. Large dimensions (sizes?), relatively few numbers of samples and intrinsic variation in experimental and biological processes lead to some problems in analyzing the microarray data. Hence, the first important step to analyzing the microarray data is reducing the number of redundant genes. Figure 1 represents general steps of feature selection and microarray data classification. These steps are as below and will be discussed in more details:

- Preprocessing the gene expression data
- Selecting a set of informative genes
- Data classification
- Evaluation and cross validation of the results.

Preprocessing the Gene Expression Data

Before applying the gene selection and classification algorithms, the data need to be preprocessed. The utilizing data in this paper are recalled using Weka software in Attribute-Relation File Format (ARFF). Because of high data variation, it is necessary to discrete the values in order to reach the appropriate classification accuracy. Discretization is a process in which features and continuous variables are converted to discrete ones. Usually, during this process data are divided into k sections with same length (same ranges) or k% of whole data (same frequencies).

Selecting a set of Informative Genes for Microarray Classification

Research illustrated that the accurate identification of cancer can be done by microarray data classification. But the main problem in microarray data analysis is its large size. Although a large number of genes exist in microarray data, just a few of them have an important effect on the

classification accuracy. A lot of genes have same performance in the normal and cancer states. Also, some genes have a noisy role in the data. These noisy genes do not have any contribution in cancer outbreak and just have a negative effect on classification accuracy. So, microarray data should be preprocessed by gene selection techniques before the classification to eliminate the uninformative genes. Doing so will result in increasing the classification efficiency and also reducing the computational complexity [5].

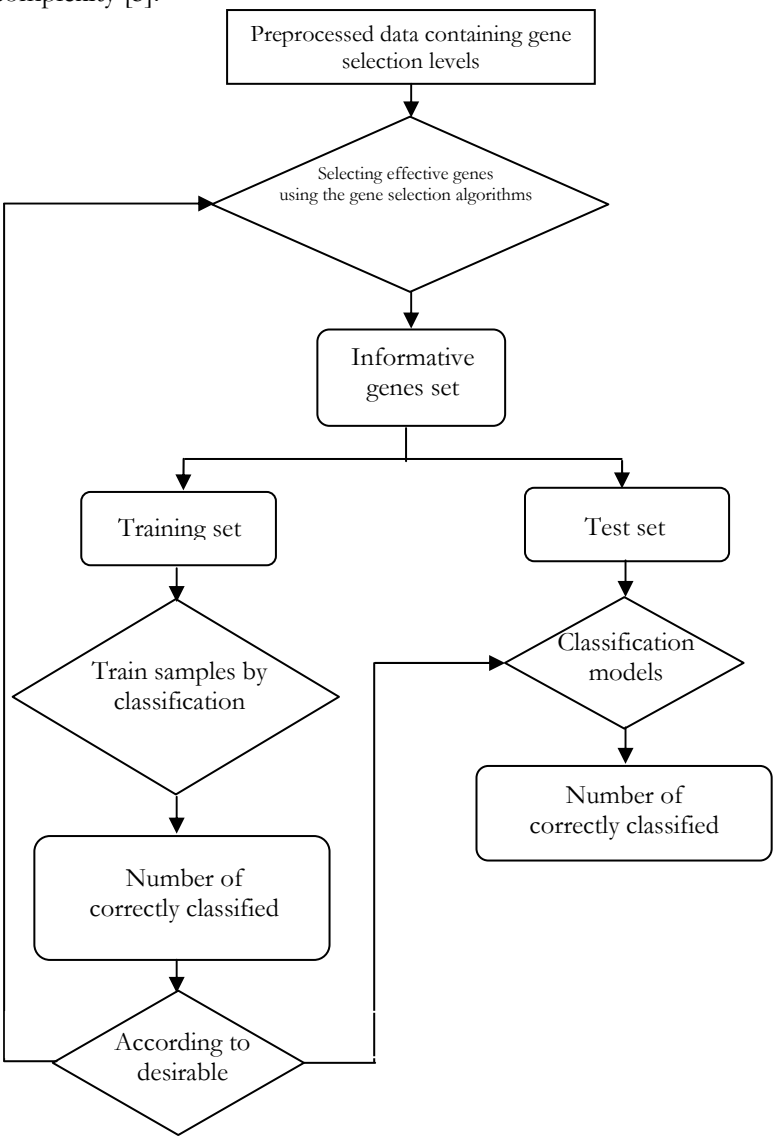

Figure 1. Block diagram of different steps of microarray data analysis

Generally, three features (gene) selection models exist [6]. The first model is filter model which carries out the features selection and classification in two separate steps. This model selects the genes as effective genes which have high discriminative ability. It is independent of classification or training algorithm and also is simple and fast. The second model is wrapper model which carries out the features selection and classification in one process. This model uses the classifier during the effective genes selecting process. In other words, the wrapper model uses the training algorithm to test the selected gene subset. The accuracy of wrapper model is more than filter one. Different methods are represented for selecting the appropriate subsets based on wrapper model in literature. In [7], evolutionary algorithms are used with K-neighborhood nearest classifier for this aim. Parallel genetic algorithms are extended by applying adaptive operations in [8]. Also in [9], genetic algorithm and support vector machine hybrid model are used to select a set of genes. Gene

selection and classification problem is discussed as a multi objective optimization problem in [10], in which the number of features and misclassified samples are reduced, simultaneously.

Finally in hybrid models, selecting a set of effective genes is done during the training process by a particular classifier. A sample of this model is using the support vector machine with recursive feature elimination. The idea of this method is eliminating the genes one by one and surveying the effect of this elimination on the expected error [11]. Recursive feature elimination algorithm is a backward feature ranking method. In other words, a set of genes that is eliminated at the last step, attains the best classification results, while these genes may do not have good correlation with the classes. Hybrid models can be considered as an extended form of wrapper model. Two other samples of hybrid model are mentioned in [12,13].

Evaluation and Cross Validation of the Results

The last step in analyzing the microarray data is evaluating the obtained results from applying the classification algorithms. In this paper, the results evaluation is done based on the k-fold cross validation method in which k mentions to numbers of repetition and is considered equal to 10. Hence, 10-fold cross validation is done as below: First, the samples are divided to 10 parts and 10% of the all data is considered as the test samples and the remainder data as the training ones, at each algorithm execution. This procedure is done 10 times with applying the different test and training data and the result is attained by calculating the mean value in 10 times test repetition.

Effective Genes Selection Algorithms

In this paper, three algorithms based on filter model are used in order to select an effective set of genes, which we will explain about them in this section:

A. Relief Algorithm

This algorithm is the generalized of relief standard algorithm and is proposed by Kononenkoin 1994 [14]. Its main idea is to estimate features (genes) which have greater weight than a threshold amount used for detecting a selected sample and two samples called Miss and Hit existing in their vicinity. This algorithm is as follows.

ReliefF Algorithm

Input: a vector space for training tests containing sample values and class values.
Output: a vector space for training tests containing W weight specified to each sample.
Initialize of weights: w[A]=0,
For i=1,…, m, following steps are performed:
R_i sample is selected randomly from class L, K the closest vicinity of class L is fined. These spots are called Hits and are shown by H_j,
For each C ≠ L:
K closest vicinity of class C is fined. These spots are called Misses and are shown by $M_j(C)$,
For A=1 to all samples:

$$W(A) = W(A) - \sum_{j=1}^{k} \frac{\text{diff}(A, R_j, H_j)}{m.k} + \sum_{C \neq L}^{k} [P(C) - P(L) \sum_{j=1}^{k} \text{diff}\left(A, R_j, M_j(C)\right)] / (m.k)$$

End.

B. Information Gain Algorithm (IGA)

This algorithm is generally used as a means for approximation to the conditional distribution in classification process [15]. By knowing value of one feature, information gain measures the number of bits of achieved information In order to class prediction [16]. Information gain is defined as follows:

$$G = H(Y) - H(Y \mid X) \quad (1)$$

where X and Y are features and,

$$H(Y) = -\sum_{y \in Y} P(y) \log_2 (P(y)) \tag{2}$$

where Y feature entropy and representative of Y feature's uncertainty level. Conditional entropy H(Y|X) is defined as follows:

$$H(Y|X) = -\sum_{x \hat{I} X} P(x) \sum_{y \hat{I} Y} P(y|x) \log_2 (P(y|x)) \tag{3}$$

In information gain algorithm, numerical features require discretization. Hence, in this paper, discretization method based on entropy is used. More details of this algorithm are given in [17].

C. Normalized Mutual Information Algorithm (NMIA)

The purpose of this algorithm is to increase the effect of proper and informative genes and decrease the redundancy genes. To measure the relation between two random variables X and Y, mutual information with the following definition is used:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{4}$$

To provide the possibility of comparison, normalized information is used as follows:

$$U(X,Y) = \frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)} \tag{5}$$

Suitability of i^{th} gene (g_i) of labels vector (c) is shown with U(g_i, c), and the dependence between two genes is shown with U(g_i, c). If Sis a selective subset, then suitability of selective genes and also genes being extra are achieved with two following equations:

$$\begin{cases} J_1 = \sum_{i \in S} U(g_i, c) \\ J_2 = \sum_{i,j \in S} U(g_i, g_j) \end{cases} \tag{6}$$

The purpose is to select a set of genes which suitability level of genes is max and level of extra genes is min. In other words:

$$maxJ = J_1 - \beta J_2 = \sum_{i \in S} U(g_i, c) - \beta \sum_{i,j \in S} U(g_i, g_j) \tag{7}$$

where β is weights parameter which its amount in this paper is considered 0.7.

Microarray Data Classification Algorithms

In this article, three classification algorithms, two proposed Bayesian Linear Discriminant Analysis (BLDA), Modified Support Vector Machine (υ-SVM) Algorithms and Probabilistic Neural Network are evaluated and compared in order to microarray data classification, which will be explained in this section.

A. Bayesian Linear Discriminant Analysis (BLDA)

BLDA is a regularly algorithm that is utilized to avoid over-fitting in the high dimension data. Using this algorithm, the regulation degree can be estimated automatically and quickly by training data and without needing to cross validation. This classifier is used for classification of noisy data and also features that are not classifiable, accurately. The basis of this classifier is that the regression is done in Bayes frame. More details of this algorithm are available in [18].

B. Modified Support Vector Machine (υ-SVM)

The main problem of SVM algorithm is constancy an uncontrollability of parameter *c* in its relation [18]. To resolve this problem, in this paper, υ-SVM algorithm has been proposed. This

algorithm was introduced by Scholkopf in 2000. In this algorithm, a pair of $\omega^T x + \omega_0 = \pm \varrho$, $\varrho \geq 0$, hyper-planes, and also a new parameter named $\upsilon \in (0,1)$ has been employed. More details of this algorithm are available in [18-19].

C. Probabilistic Neural Network

Probabilistic Neural Network is formed of three layers. Whenever an input vector is applied to network, first layer computes the distance of input vector from training inputs, and in this way it provides a vector that its elements are determinant of distance level between input and training input. Second layer determines a vector of probabilities as network output using output of first layer. Finally competitive transfer function existing in second layer selects maximum amount of probabilities from probabilities vector, and outputs 1 for it, and zero for other probabilities. This algorithm is as follows [20].

Data Set

Generally, achieved results consist of two feature extraction and classification sections. In both sections Leukemia data set has been used [21]. The working data set includes 145 different patients, of which 78 were classified during the clinical portion as "normal," and 67 with various stages or forms of leukemia. The samples from the patients were processed multiple times, resulting in 425 cases for the study. We included acute and chronic forms of lymphoma and myelogenous leukemia, as well as adult T-cell leukemia and other non-Hodgkin's lymphoma. There are a number of other patients included in the data set with other stages or types of leukemia, however, these subsets were not examined for this study. Multiple cases from the same sample are called replicates. Figure 2 is a heat map of the Leukemia data set. Each row of pixels represents the abundances for all the molecules (or peaks) found in a single spectrum (or case); each column is the abundance of a specific molecule for all cases. The color of the (i,j) pixel reflects the abundance of peak i in case j, where i runs along the horizontal axis. When sorted into classes, the heat map may be useful for searching for diagnostic portions of the spectrum by eye. The dotted line represents the division between the normal class, which is the top half of the cases, and the disease class. It should be possible to see the class difference in the rightmost variables (disease cases are slightly brighter) that we will later find to be diagnostic.

All implementation results of raised algorithms have taken place on a computer with 3.4GHz processor, and 2GHz RAM memory.

Probabilistic Neural Network (PNN) Algorithm

(1)Input Algorithm

Applying an unspecified pattern or a feature vector **x**.

(2)Pattern layer: x^i,is i^{th} feature vector.

For i=1:N

$y^i = x^i x^T - 0.5(xx^T + x^i(x^i)^T)$

$y^i = \exp(y^i/h^2)$

End.

(3)Collector Layer

For j=1:N

sum(j) = 0

For all is in % {1, …, N} all samples in the same group

$sum(j) += y^{(j,i)}$

End.

$sum(j) = sum(j)/(2\pi^{m/2} h^m N_j)$

End.

x pattern belongs to group j with following membership:

$$\text{membership}(j) = \frac{sum(j)}{\sum_{j-1}^{n} sum(j)}$$

For all j in [1,n].

(4) Output Layer

x pattern is dedicated to group j with maximum membership which in it:

$$s_j^* = \arg\max\{membership(j)\}$$

For all j∈{1, …,n}

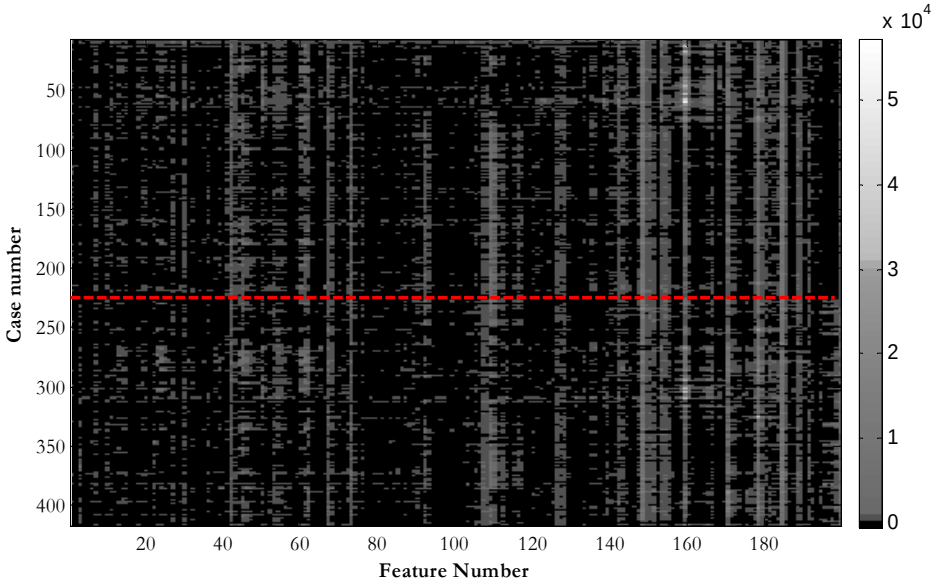

Figure 2. Leukemia data

This amount is 91.49 in PNN algorithm and 86.31 and 87.24 in υ-SVM algorithm with linear kernel and RBF, respectively. Also, as it is seen in Tables 1, 2 and 3, BLDA classification algorithm has better performance than two other classification algorithms. In Table 4, classification accuracy of BLDA, PNN and υ-SVM algorithms with selecting different number of features (genes) is shown. The number of selected genes to evaluate normalized mutual information algorithm is considered 10, 50, 100 and 200 genes. Based on gained results, BLDA algorithm's classification accuracy is 93.01 with selecting 50 and 100 genes, while it is 84.12 with selecting 10 and 200. Same results are achieved by applying two PNN and υ-SVM algorithm, which is shown in Table 4. Low accuracy obtained with selecting 10 genes is because of lack of informative genes as classifier training samples, which will result in test data's loss. Likewise, poor performance of the classifier with selecting 200 features is also due to existence of noisy and redundant genes among set of features. This is while with selecting 50 and 100 genes, sufficient information is applied to the classifier. Set of selected genes contain no redundant genes among themselves and this will improve the classification performance.

Results and Discussion

Table1 shows classification accuracy obtained from applying different feature extraction algorithms using BLDA classifier algorithm. As it is seen, normalized mutual information algorithm has better performance than other feature extraction algorithms, because of most extra genes cancellation and informative genes selection, which widely take part in determining class type. So that classification accuracy achieved using this algorithm is 95.34, while this amount is 90.21 and 85.36, in information gain and ReliefF algorithms, respectively. Also in Tables 2 and 3, classification accuracy with applying PNN classifier algorithm and υ-SVM algorithm, in two cases of using linear kernel and RBF1, is shown respectively. It is worth mentioning that the amount of υ parameter is considered 0.24 and also the amount of RBF kernel variance 0.002. Superiority of normalized mutual information algorithm in classification accuracy improvement in comparison to other

[1] Radial Basis Function

methods is clearly seen in these Tables.

Table 1. Comparison of Quantitative Results of Classification Accuracy by Employing Different Feature Extraction and BLDA Classifier Algorithms

Accuracy	Algorithms (Feature Selection + Classification)
95.34	**NMIA+BLDA**
90.21	IGA+BLDA
85.36	ReliefF+ BLDA

Table 2. Comparison of Quantitative Results of Classification Accuracy by Employing Different Feature Extraction and PNN Classifier Algorithms

Accuracy	Algorithms (Feature Selection + Classification)
94.19	NMIA+PNN
89.01	IGA+PNN
83.00	ReliefF+PNN

Table 3. Comparison of Quantitative Results of Classification Accuracy by Employing Different Algorithms of Feature Extraction and υ-SVM Classifier in Two Modes of RBF and Linear kernels

Accuracy	Algorithms (Feature Selection + Classification)
87.24	NMIA+υ-SVM (RBF Kernel)
84.08	IGA+υ-SVM (RBF Kernel)
78.46	ReliefF+υ-SVM (RBF Kernel)
86.31	NMIA+υ-SVM (Linear Kernel)
82.98	IGA+υ-SVM (Linear Kernel)
77.23	ReliefF+υ-SVM (Linear Kernel)

Table 4. Comparison of Quantitative Results of Different Algorithms Classification Accuracy by Applying Different Selected Genes using Normalized Mutual Information Algorithm

Accuracy	Gene Selected by Applying Normalized Mutual Information	Classification Algorithms
BLDA	10	84.12
	50	**93.01**
	100	**93.01**
	200	84.12
PNN	10	81.03
	50	**91.42**
	100	**91.42**
	200	81.03
υ –SVM (RBF Kernel)	10	70.14
	50	**75.18**
	100	**74.32**
	200	69.24
υ –SVM (Linear Kernel)	10	69.21
	50	**71.14**
	100	**70.42**
	200	69.21

As we mentioned in [22-26], DNA is the most important biological coding material in living cells, bacteria and human. Mutation in DNA could be caused many diseases such as cancers. In this paper, different extraction algorithms of informative genes in microarray data with large size and also these data's classification algorithms were compared. Results represented that normalized mutual information algorithm select many effective genes and greatly reduces genes containing

noise. Also BLDA classification algorithm has the highest classification accuracy among other proposed classification methods. Thus, combination of normalized mutual information technique and BLDA classifier is considered a proper algorithm in order to analyze microarray data. As such, this algorithm improves classification accuracy in the amount of 1.2, 9.3 and 10.5, in comparison to normalized mutual information and PNN, normalized mutual information and υ-SVM (RBF kernel) and normalized mutual information and υ-SVM (linear kernel) combinational algorithms, respectively.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Wee A, Liew C, Yah H, Yang M. Pattern recognition techniques for the emerging field of bioinformatics: A review. Pattern Recogni 2005;38:2055-73.
2. Saberkari H, Shamsi M, Heravi H, Sedaaghi MH. A fast algorithm for exonic regions prediction in DNA sequences. J Med Signals Sens 2013;3:139-49.
3. Chu F, Wang L. Applications of support vector machines to cancer classification with microarray data. Int J Neural Syst 2005;15:475-84.
4. Lu Y, Han J. Cancer classification using gene expression data. InfSyst Data Manage Bioinform 2003;28:243-68.
5. Chen Y, Zhao Y. A novel ensemble of classifiers for microarray data classification. Appl Soft Comput 2008;8:1664-9.
6. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157-82.
7. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001;7:1131-42.
8. Jourdan L. Metheuristics for knowledge discovery: Application to genetic data. Ph.D. Thesis, 2003, University of Lille.
9. Peng S, Xu Q, Ling XB, Peng X, Du W, Chen L. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Letter 2003;555:358-62.
10. Reddy AR, Deb K. Classification of two-class cancer data reliably using evolutionary algorithms. Technical Report, 2003, KanGAL.
11. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., "Gene selection for cancer classification using support vector machines", Machine Learning, 2002, 46: 389-422.
12. Saeys Y, Aeyels Degroeve S, Rouze D, Van de Peer YP. Enhancing genetic feature selection through restricted search and Walsh analysis. IEEE Trans. on Systems, Man and Cybernetics, Part C 2004;34:398-406.
13. Goh L, Song Q, Kasabov N. A novel feature selection method to improve classification of gene expression data. In Proc. 2th Asia-Pacific Conference on Bioinformatics, 2004, pp. 161-166.
14. Kononenko I. Estimating Attributes: Analysis and Extensions of RELIEF. Proceeding of the European Conference on Machine Learning, pp. 171-182, 1994.
15. Cover T, Thomas J. Elements of Information Theory, New York: John Wiley and Sond, 1991.
16. Mitchell TM. Machine Learning, New York: McGraw-Hill, 1997.
17. Freund Y, Schapire RE. Experiments with a new boosting algorithm. Proceeding of the Thirteenth International Conference in Machine Learning, pp. 148-156, 1996.
18. Momennezhad A, Shamsi M, Ebrahimnezhad H, Saberkari H. Classification of EEG-P300 Signals Extracted from Brain Activities in BCI Systems using υ-SVM and BLDA Algorithms.

Appl Med Inform 2014;34(2):23-35.
19. Saberkari H, Shamsi, Joroughi M, Golabi F, Sedaaghi MH. Cancer Classification in Microarray Data using a Hybrid Selective Independent Component Analysis and υ-Support Vector Machine Algorithm. J Med Signals Sens 2014;4(4):291-9.
20. Saunders BiC, McPheron BA. Wing Pattern-Based Classification of the Rhagoletis pomonella Species Complex Using Genetic Neural Networks. Int J Comput Sci Appl 2007;4:1-14.
21. http:// datam.i2r.a-star.edu.sg/datasets/krbd.
22. Saberkari H, Shamsi M, Sedaaghi MH. Prediction of Protein Coding Regions in DNA Sequences using Signal Processing Methods. IEEE Symp on Industrial Electronics and Applications (ISIEA), Bandung, Indonesia September 2012, pp. 354-359.
23. Saberkari H, Shamsi M, Sedaaghi MH. A Punctual Algorithm for Small Gene Prediction in DNA Sequences Using a Time-Frequency Approach Based on the Z-Curve. GSTF Int J Eng Technol (JET) 2013;2(1):1-9.
24. Ahmadi F, Saberkari M, Abiri R, Mohammadi Motlagh H, Saberkari H. In Vitro Evaluation of Zn-Norfloxacin Complex as a Potent Cytotoxic and Anti-Bacterial Agent, Proposed Model for DNA Binding. Appl Biochem Biotech 2013;170:988-1009.
25. Saberkari H, Shamsi M, Sedaaghi MH. A Hybrid Anti-notch/Geortzel Model for Gene Prediction in DNA Sequences. Appl Med Inform 2014;34(2):13-22, 2014.
26. Saberkari H, Shamsi M, Heravi H, Sedaaghi MH. A Novel Fast Algorithm for Exon Prediction in Eukaryotic Genes using Linear Predictive Coding Model and Goertzel Algorithm based on the Z-Curve. Int J Comput Appl 2013;67(17):25-38.