American Journal of Applied Sciences 8 (12): 1295-1301, 2011
ISSN 1546-9239

Efficient Classification of Cancer using Support Vector Machines and Modified Extreme Learning Machine based on Analysis of Variance Features

A. Bharathi and A.M. Natarajan
Bannari Amman Institute of Technology,
Sathyamangalam, TamilNadu State, India

Abstract: Problem statement: The primary objective is to propose efficient cancer classification techniques which provide reliable and significant classification accuracy. To achieve this primary research goal is to find the smallest set of genes that can ensure high accuracy in classification using supervised machine learning algorithms. The significance of finding the minimum subset is three fold: (a) The computational burden and noise arising from irrelevant genes are much reduced; (b) the cost for cancer testing is reduced significantly as it simplifies the gene expression tests to include only a very small number of genes rather than thousands of genes; (c) it calls for more investigation into the probable biological relationship between these small numbers of genes and cancer development and treatment. **Approach:** The proposed method involves two steps. In the first step, some important genes are chosen with the help of Analysis of Variance (ANOVA) ranking scheme. In the second step, the classification capability is tested for all simple combinations of those important genes using a better classifier. **Results:** The proposed method initially uses Support Vector Machine (SVM) classifier. Then Modified Extreme Learning Machine classifier is used for increasing the classification accuracy over SVM. **Conclusion:** The two datasets are used (Lymphoma and Liver cancer) in the experimental result shows that the proposed method performs the cancer classification with better accuracy when compared to the SVM methods.

Key words: Gene expressions, cancer classification, neural networks, support vector machines, modified extreme learning machine

INTRODUCTION

Cancer is one of the dreadful diseases found in most of the living being, which is one of the challenging studies for research in the 20th century. There has been lot of proposals from various researchers on cancer classification and detailed study is still on in the domain of cancer classification. Cancer (Alter *et al*., 2003) is fundamentally described by an abnormal, uncontrolled growth that may demolish and attack other healthy body tissues. There are billions of cells in the human body and most of the cells have an inadequate life-span and required to be replaced in a cyclic manner. Each cell is capable of duplicating themselves. Millions of cell divisions and replications occur daily in the body and it is amazing that the procedure occurs so accurately most of the time every cell division needs replication of the 40 volumes of genetic coding. But, sometimes, there is some fault in a division and it may lead to a rogue and potentially malignant cell. The immune system has the capability of identifying such events and is usually eliminates such abnormal cells before they have an opportunity to proliferate. Rarely, there is a failure of the mechanism and a potentially malignant cell survives, replicates and cancer is the result.

In this study, a simple yet very effective method using SVM (El-Naqa *et al*., 2002) and MELM classifier that leads to accurate cancer classification using expressions of two gene combinations in lymphoma data set is proposed. This study is organized as follows. Section 2 describes some related works for the proposed system. The methodology for the proposed system is provided in section 3. The experimental results are shown in section 4 and this study concludes in the section 5.

Related works: Guyon *et al*. (2002) proposed the Gene Selection for Cancer Classification using Support Vector Machines. In this study, the author address the

Corresponding author: A. Bharathi, Bannari Amman Institute of Technology, Sathyamangalam, TamilNadu State, India

problem of selection of a small subset of genes from broad patterns of gene expression data, recorded on DNA micro-arrays. Using available training examples from cancer and normal patients, the approach build a classifier suitable for genetic diagnosis, as well as drug discovery. Previous attempts to address this problem select genes with correlation techniques. The author proposes a new method of gene selection utilizing Support Vector Machine methods based on Recursive Feature Elimination (RFE). It is experimentally demonstrated that the genes selected by our techniques yield better classification performance and are biologically relevant to cancer.

Hernandez *et al*. (2007) presents a Genetic Embedded Approach for Gene Selection and Classification of Microarray Data Classification of microarray data requires the selection of subsets of relevant genes in order to achieve good classification performance. This article presents a genetic embedded approach that performs the selection task for a SVM classifier. The main feature of the proposed approach concerns the highly specialized crossover and mutation operators that take into account gene ranking information provided by the SVM classifier. The effectiveness of this approach is assessed using three well-known benchmark data sets from the literature, showing highly competitive results.

Cheng *et al*. (2007) put forward the Classification of FTIR Gastric Cancer Data Using Wavelets and SVM. In order to improve the accuracy to diagnose rate earlier stage gastric cancer with Fourier Transform Infrared Spectroscopy (FTIR), a novel method of extraction of FTIR feature using Continuous Wavelet Transform (CWT) analysis and classification using the Support Vector Machine (SVM) was developed. To the FTIR of gastric normal tissue, early carcinoma and advanced gastric carcinoma, 9 feature parameters were extracted with continuous wavelet analysis. With SVM, all spectra were classified into two categories: normal or abnormal, which included early carcinoma and advanced gastric carcinoma. The accurate rate of poly and RBF kernel was high in all kernels. The accurate rate of poly kernel in normal, early carcinoma and advanced carcinoma were 100, 96 and 100%, respectively. The accurate rate of RBF kernel in normal, early carcinoma and advanced carcinoma were 100, 96 and 100%, respectively. The research result shows the feasibility of establishing the models with FTIR-CWT- SVM method to identify normal, early carcinoma and advanced gastric carcinoma.

Song and Rajasekaran (2010) gives a greedy algorithm for gene selection (Lee and Lee, 2003) based on SVM and correlation. Microarrays serve scientists as a powerful and efficient tool to observe thousands of genes and analyze their activeness in normal or cancerous tissues. In general, microarrays are used to measure the expression levels of thousands of genes in a cell mixture. Gene expression data obtained from microarrays can be used for various applications. One such application is that of gene selection. Gene selection is very similar to the feature selection problem addressed in the machine-learning area. In a nutshell, gene selection is the problem of identifying a minimum set of genes that are responsible for certain events (for example the presence of cancer). Informative gene selection is an important problem arising in the analysis of microarray data. In this study, a novel algorithm is presented for gene selection that combines Support Vector Machines (SVMs) with gene correlations. Experiments show that the new algorithm, called GCI-SVM, obtains a higher classification accuracy using a smaller number of selected genes than the well-known algorithms in the literature.

Chen *et al*. (2001); Liao and Li (2007) and Liao *et al*. (2007) proposed a support vector machine ensemble for cancer classification using gene expression data in this study, the author propose a Support Vector Machine (SVM) ensemble classification method. Firstly, dataset is preprocessed by Wilcoxon rank sum test to filter irrelevant genes. Then one SVM is trained using the training set and is tested by the training set itself to get prediction results. Those samples with error prediction result or low confidence are selected to train the second SVM and also the second SVM is tested again. Similarly, the third SVM is obtained using those samples, which cannot be correctly classified using the second SVM with large confidence. The three SVMs form SVM ensemble classifier. Finally, the testing set is fed into the ensemble classifier. The final test prediction results can be got by majority voting. Experiments are performed on two standard benchmark datasets: Breast Cancer, ALL/AML Leukemia. Experimental results demonstrate that the proposed method can reach the state of-the-art performance on classification.

Cınar *et al*. (2009) gives the early prostate cancer diagnosis by using artificial neural networks and support vector machines. The aim of this study is to design a classifier based expert system for early diagnosis of the organ in constraint phase to reach informed decision making without biopsy by using some selected features. The other purpose is to investigate a relationship between Body Mass Index (BMI), smoking factor and prostate cancer. The data used in this study were collected from 300 men (100: prostate adenocarcinoma, 200: chronic prostatism or benign prostatic hyperplasia). Weight, height, BMI, Prostate Specific Antigen (PSA), Free PSA, age,

prostate volume, density, smoking, systolic, diastolic, pulse and Gleason score features were used and independent sample t-test was applied for feature selection. In order to classify related data, the author have used following classifiers; Scaled Conjugate Gradient (SCG), Broyden-Fletcher-Goldfarb-Shanno (BFGS) and Levenberg-Marquardt (LM) training algorithms of Artificial Neural Networks (ANN) and linear, polynomial and radial based kernel functions of Support Vector Machine (SVM). It was determined that smoking is a factor increases the prostate cancer risk whereas BMI is not affected the prostate cancer. Since PSA, volume, density and smoking features were to be statistically significant, they were chosen for classification. The proposed system was designed with polynomial based kernel function, which had the best performance (accuracy: 79%). In Turkish Family Health System, family physician to whom patients are applied firstly, would contribute to extract the risk map of illness and direct patients to correct treatments by using expert system such proposed.

METERIALS AND METHODS

Cancer classification proposed in this study comprises of two steps. In the first step, all genes in the training data set are ranked using a scoring scheme. Then genes with high scores are retained. In the second step, the classification capability of all simple two gene combinations among the genes selected are tested in this step using a better classifier such as Support Vector Machine and Relevance Vector Machine classifier.

Step 1: Gene importance ranking: This step performs the computation of important ranking of each gene by means of Analysis of Variance (ANOVA) method.

Step 2: Finding the minimum gene subset: This step attempts to classify the data set with single gene after selecting several top genes in the important ranking list. Each selected gene is given as an input to the classifier. When good accuracy is not obtained, it is required to classify the data set with all possible 2 gene combination within the selected genes. Even if the good accuracy is not obtained, this procedure is repeated with all of the 3 gene combinations and so on until the good accuracy is obtained.

The following classifier is used to test 2-gene combinations in this study.

Support Vector Machines (SVMs): Support Vector Machines (SVMs) is a type of classifier that are a set of associated supervised learning methods used for classification.SVM will build a separating hyperplane in the space, one which maximizes the margin between the two data sets. To determine the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. In the case of support vector machines, a data point is sighted as a p dimensional vector and it is needed to know whether it can separate such points with a p-1-dimensional hyperplane. This is called a linear classifier.

As SVM are linear classifiers that are able to find the optimal hyper plane that maximizes the boundaries between patterns, this feature makes SVM a powerful tool for pattern recognition tasks. SVM have been previously in gene expression data analysis (Carin and Dobeck, 2003; Li *et al*., 2008). In this study, a group of SVMs with basic kernel functions are used. The 5 fold Cross Validation (CV) is carried out for SVM in the training data set to tune their parameters. This study includes CV accuracy for all of the data sets and selects the smallest CV error.

The procedure of cross validation is given in Fig. 1. Initially, the whole data set is randomly divided into training (F1) and testing (F2) data. The genes are ranked using samples of F1. The combination (FC1) is generated using 2 genes among 20. Then FC1 is randomly divided into 5 folds (fc1, fc2, fc3, fc4 and fc5). From these folds one fold id selected for testing. The other 4 folds are used as a classifier for SVM. This combination is generated until better accuracy is obtained. Finally with the fitted SVM, the prediction can be performed.

Modified extreme learning machine: A modified ELM technique which uses ELM and LM technique can be described as below.

Initially, the input weights and hidden biases are created by with the help of AHP technique.

Next, the equivalent output weights are analytically determined with the help of ELM algorithm only in first step and randomly produce the output hidden biases.

Then, the parameters (all weights and biases) are restructured with the help of LM algorithm. The processing of Hybrid Extreme Learning Machine is shown in Fig. 2.

The process for the Hybrid Extreme Learning Machine is described below:

Provided a training set $N = \{x_i,t_i\}|x_i \in R^n,\ t_i \in R^n,\ t_i \in R^m, I = 1, 2,\ldots,N\}$ activation functions $f_1(x)$ and $f_2(x)$ and hidden nodes namely $\tilde{N}$ and K of hidden first and second layer.

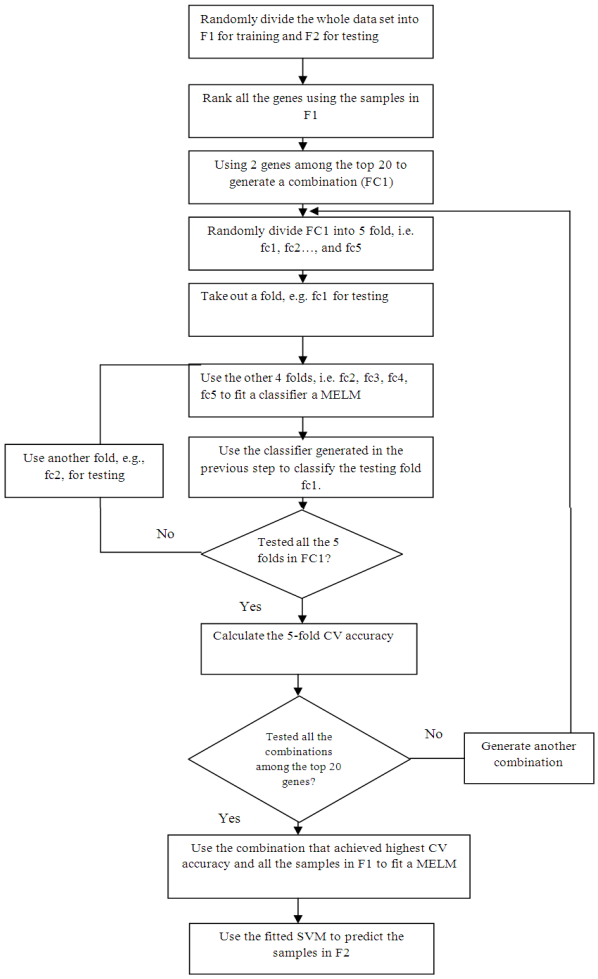

Fig. 1: Procedure for CV

Step 1: Randomly choose the starting values of input weight vectors w_1 and bias vector b_1 with the help of AHP technique and bias vector b_2 without using the AHP technique.

Step 2: Determine the hidden first layer output matrix a_1. With the help of ELM algorithm, determine the output weight:

$$w_2 = a_1^{-1}.t$$

Step 3: Determine the hidden second layer output matrix a_2, errors:

$$e_1 = t - a_2$$

And determine the sum of squared errors over all input.

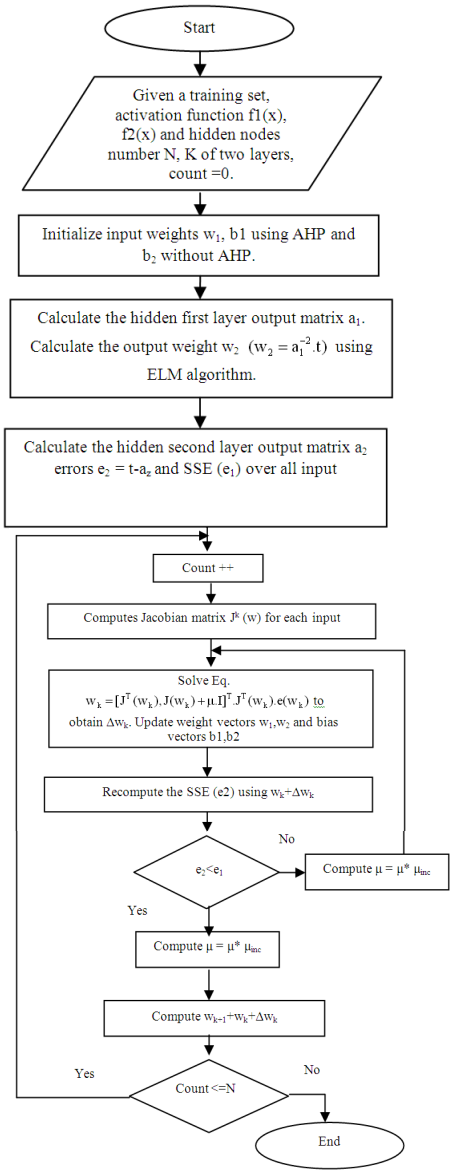

Fig. 2: Hybrid extreme learning machine

Step 4: Determine the Jacobian matrix. Calculate the sensitivities with the recurrence relations:

$$S_q^m = f^m(n_q^m)(w^{m+1})^T.S_q^{m+1}$$

After initializing with the following equation:

$$S_q^M = f^m(n_q^m)$$

Augment the individual matrices into the Marquardt sensitivities using the following equation:

$$S_q^m = [S_1^m, S_2^m, ..., S_Q^m]$$

Determine the elements of the Jacobian matrix with the equations:

$$[J]h,1 = S_{1,h}^m \times S_{j,k}^{m-1}]$$

And:

$$[J]h,1 = S_{1,h}^m$$

Step 5: Solve equation given below to determine Δ_{wk} and update weight vectors w_1, w_2 and bias vectors b_1, b_2.

$$\Delta w_k = [J^T(w_k), J(w_k) + \mu.I]^T.J^T(w_k).e(w_k)$$

Step 6: Recalculate the sum of squared errors with the help of $w_k+\Delta w_k$. If this new sum of squared error is lesser than the evaluated error value in step3, then multiply μ by μ_{dec}, let $w_{k+1}=w_k+\Delta w_k$ and process from step4. If the sum of squared error is not decreased, then multiply μ by μ_{inc} and process from step5.

The 5 fold Cross Validation (CV) is carried out for MELM in the training data set to tune their parameters. This study includes CV accuracy for all of the data sets and selects the smallest CV error.

The procedure of cross validation is given in Fig. 1. Initially, the whole data set is randomly divided into training (F1) and testing (F2) data. The genes are ranked using samples of F1. The combination (FC1) is generated using 2 genes among 20. Then FC1 is randomly divided into 5 folds (fc1, fc2, fc3, fc4 and fc5). From these folds one fold id selected for testing. The other 4 folds are used as a classifier for SVM. This combination is generated until better accuracy is obtained. Finally with the fitted MELM, the prediction can be performed.

RESULTS AND DISCUSSION

The experimentation on the proposed method is carried on lymphoma data set and liver cancer dataset. In the lymphoma data set, there are 42 samples derived from Diffuse Large B-Cell Lymphoma (DLBCL), nine samples from Follicular Lymphoma (FL) and 11 samples from Chronic Lymphocytic Leukemia (CLL). The expression data of 4026 genes are included in the entire data set. Very few parts of data are missing in this data set. For filling those missing values k-nearest neighbor algorithm was applied.

Table 1: Maximum accuracy achieved by the following combinations (By MELM)

1 ,4	1,6	1,9	1,15	1,16	1,17	1,18	1,19
2 ,6	2,8	2, 9	2,12	2,14	2,15	2, 17	2,18
4 ,7	4,8	4,15	4,17	4,18			
5,7	5,9	5,11	5,15	5,18			
7, 8	7, 9	7,12	7,15	7,19			
8,17	8,19						
9,11	9,15	9,17	9,19				
11,16	11,18	11,19					
12,13	12,17						
14,18							
17,19							
18,20							

Table 2: Maximum accuracy achieved by the following combinations (By SVM)

1 ,4	1,8	1, 9	1,14	1,15	1,16	1,18	
2 ,4	2,8	2, 9	2,11	2,14	2,15	2, 16	2,18
4 ,7	4,12	4,17					
7, 8	7, 9	7,14	7,18				
8,17							
9,12	9,17						
11,17							
12,14	12,18						
14,17							
17,18							
18,20							

In the first step, the 62 samples are divided randomly into 2 parts: 31 samples for testing, 31 samples for training. According to the ANOVA in the training set, the complete sets of 4026 genes are ranked. Next, 20 genes with highest ANOVA is picked.

Then the proposed classifier is applied to classify the lymphoma micro array data set. Initially, the selected 20 genes are added one by one to the network according to their ANOVA ranks. That is, only a two gene that is ranked 1 is used as the input to the network. Then the network is trained with the training data set and subsequently, tested the network with the test data set.

The excellent performance of proposed MELM motivated to search for the smallest gene subsets that can ensure highly accurate classification for the entire data set. Initially, it attempted to classify the data set using two gene tested for all possible combinations within the 20 genes.

Table 1 shows the combination for achieving the maximum accuracy by usage of proposed method. The gene combination chosen by the proposed method for 1 gene are (1,4), (1,6), (1,9), (1,15), (1,16), (1,17), (1,18) and (1, 19). Table 2 shows the combination for achieving the maximum accuracy by usage of SVM classifier. Some of the combination choose by SVM for choosing 1 gene are (1,4), (1,8), (1,9), (1,14), (1,15), (1,16) and (1,18).As the Table 1 suggest, more combination is obtained for using the MELM method when compared to SVM method.

Table 3: Accuracy comparison for using ANOVA with No. of Folds = 5 (Lymphoma dataset)

No. of gene combinations	Accuracy	
	SVM	MELM
20,2	96.7742	100
20,3	98.7741	100

Table 4: Accuracy comparison for using ANOVA with No. of Folds = 10 (Lymphoma dataset)

No. of gene combinations	Accuracy	
	SVM	MELM
20,2	96.7742	100
20,3	98.7741	100

Table 5: Accuracy comparison for using correlation with No. of folds = 5 (Lymphoma dataset)

No. of gene combinations	Accuracy	
	SVM	MELM
20,2	96.7742	100
20,3	98.7741	100

Table 6: Accuracy comparison for using correlation with No. of folds = 10 (Lymphoma dataset)

No. of Gene Combinations	Accuracy	
	SVM	MELM
20,2	96.7742	100
20,3	98.7741	100

Table 7: Accuracy comparison for using ANOVA with No. of folds = 5 (Liver Cancer dataset)

No. of Gene Combinations	Accuracy	
	SVM	MELM
20,2	89.3221	100
20,3	90.3226	100

Table 8: Accuracy comparison for using ANOVA with No. of folds = 10 (liver cancer dataset)

No. of gene combinations	Accuracy	
	SVM	MELM
20,2	85.7741	100
20,3	87.0968	100

The resulted accuracy for using the lymphoma data set is presented in Table 3-6 which uses different number of folds.

The resulted accuracy for using the lymphoma data set is presented in Table 3-6 which uses different number of folds. From these observations, it can be suggested that the MELM method is better in classifying the cancer.

Liver cancer dataset: The liver cancer data set (http://genome-www.standford.edu/hcc/) has two classes, i.e., the nontumor liver and HCC. The data set contains 156 samples and the expression data of 1, 648 important genes.82 are HCCs and the other 74 are nontumor livers. We randomly divided the data into 78 training and 78 testing samples

In Table 7 and 8, the accuracy resulted for using liver cancer dataset is presented. From these observations, it can be suggested that the MELM method is better in classifying the cancer.

CONCLUSION

This research focuses on the establishment of efficient classifiers for micro array data using statistical ranking techniques and machine learning algorithms. This research uses effective learning algorithm approaches such as SVM and MELM. In the first proposed approach, SVM algorithm with ANOVA ranking is proposed for the classification of cancer. The second proposed method uses MELM uses the AHP method. This proposed approach provides better accuracy than the SVM approach. The performance of the proposed approaches is evaluated based on the performance measures such as accuracy. The experiments are performed in two data sets namely lymphoma and liver cancer data set. The experimental results show that the proposed MELM approach shows significant performance in terms of classification accuracy. This is due to the salient features of the proposed MELM approach which provides better performance because of the advantages of SVM. Thus it is clear that, the proposed "Modified Extreme Learning Algorithm (MELM)" is very efficient in cancer classification.

REFERENCES

Alter, O., P.O. Brown and D. Botstein, 2003. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc. Natural Acad. Sci., 100: 3351-3356. DOI: 10.1073/pnas.0530258100

Carin, L. and G.J. Dobeck, 2003. Relevance vector machine feature selection and classification for underwater targets. Proc. OCEANS, 2: 1110-1110. DOI: 10.1109/OCEANS.2003.178498

Chen, S., S.R. Gunn and C.J. Harris, 2001. The relevance vector machine technique for channel equalization application. IEEE Trans Neural Netw., 12: 1529-1532. DOI: 10.1109/72.963792 PMID: 18249985

Cheng, C.G., L.Y. Cheng and R.S. Xu, 2007. Classification of FTIR gastric cancer data using wavelets and SVM. Proceedings of the 3rd International Conference on Natural Computation, Aug. 24-27, IEEE Xplore Press, Haikou, Hainan, China, pp: 543-547. DOI: 10.1109/ICNC.2007.299

Cınar, M., M. Engin, E.Z. Engin and Y.Z. Atesci, 2009. Early prostate cancer diagnosis by using artificial neural networks and support vector machines. Expert Syst. Appli., 36: 6357-6361. DOI: 10.1016/j.eswa.2008.08.010

El-Naqa, I., Y. Yang, M.N. Wernick, N.P. Galatsanos and R.M. Nishikawa, 2002. A support vector machine approach for detection of microcalcifications. IEEE Trans. Med. Imag., 21: 1552-1563. DOI: 10.1109/TMI.2002.806569 PMID: 12588039

Guyon, I., J. Weston, S. Barnhill and V. Vapnik, 2002. Gene selection for cancer classification using support vector machines. Mach. Learn., 46: 389-422. DOI: 10.1023/A:1012487302797

Hernandez, J.C.H., B. Duval and J.K. Hao, 2007. A genetic embedded approach for gene selection and classification of microarray data. Evol. Comput. Mach. Learn. Data Min. Bioinformatics, 4447: 90-101. DOI: 10.1007/978-3-540-71783-6_9

Lee, Y. and C.K. Lee, 2003. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics, 19: 1132-1139. DOI: 10.1093/bioinformatics/btg102

Li, S., X. Wu and X. Hu, 2008. Gene selection using genetic algorithm and support vectors machines. Soft Comput. Fusion Foundations, Method. Appli., 12: 693-698. DOI: 10.1007/s00500-007-0251-2

Liao, C. and S. Li, 2007. A support vector machine ensemble for cancer classification using gene expression data. Bioinformatics Res. Appli. Lecture Notes Comput. Sci., 4463: 488-495, DOI: 10.1007/978-3-540-72031-7_44

Liao, C., S. Li and Z. Luo, 2007. Gene selection using wilcoxon rank sum test and support vector machine for cancer classification. Comput. Intell. Security, 4456: 57-66. DOI: 10.1007/978-3-540-74377-4_7

Song, M. and S. Rajasekaran, 2010. A greedy algorithm for gene selection based on SVM and correlation. Int. J. Bioinformatics Res. Appli., 6: 296-307. DOI: 10.1504/IJBRA.2010.034077