Journal of Computer Science 7 (9): 1393-1399, 2011
ISSN 1549-3636

Cancer Classification using Support Vector Machines and Relevance Vector Machine based on Analysis of Variance Features

A. Bharathi and A.M. Natarajan
Bannari Amman Institute of Technology,
Sathyamangalam, Tamil Nadu State

Abstract: Problem statement: The objective of this study is, to find the smallest set of genes that can ensure highly accurate classification of cancer from micro array data by using supervised machine learning algorithms. The significance of finding the minimum subset is three fold: The computational burden and noise arising from irrelevant genes are much reduced; 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; 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 were chosen with the help of Analysis of Variance (ANOVA) ranking scheme. In the second step, the classification capability was tested for all simple combinations of those important genes using a better classifier. **Results:** The proposed method initially uses Support Vector Machine (SVM) classifier. Relevance Vector Machine (RVM) classifier was used for increasing the classification accuracy over SVM classifier. **Conclusion:** The experimental result shows that the proposed method performs the cancer classification with better accuracy when compared to the conventional methods.

Key words: Gene expressions, cancer classification, neural networks, Continuous Wavelet Transform (CWT), Support Vector Machine (SVM), Relevance Vector Machine (RVM), Principal Component Analysis (PCA), Generalized Singular Value Decomposition (GSVD), Singular Value Decomposition (SVD)

INTRODUCTION

MICRO array data analysis has been successfully applied in a number of studies over a broad range of biological disciplines including cancer classification by class discovery and prediction , identification of the unknown effects of a specific therapy , identification of genes relevant to a certain diagnosis or therapy and cancer prognosis. The multivariate supervised classification techniques such as Support Vector Machines (SVMs) (El-Naqa *et al*., 2002) and multivariate statistical analysis method such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD) (Alter *et al*., 2003.) and Generalized Singular Value Decomposition (GSVD) cannot be applied to data with missing values. The finding of missing value is an essential preprocessing step. Because of various reasons, there may be some loss of data in gene expression e.g., inadequate resolution, image corruption, dirt or scratches on the slides or experimental error during the laboratory process. Several algorithms have been developed for recovering data because it is costlier and time consuming to repeat the experiment. Moreover, estimating unknown elements in the given data has many potential applications in the other fields. There are several approaches for the estimating the missing values. Recently, for missing value estimation, the Singular Value Decomposition based method (SVD impute) and weighted k-nearest neighbors imputation (KNN impute) has been introduced. It has been shown that KNN impute shows better performance on non-time series data or noisy time series data, whereas, SVD impute works well on time series data with low noise levels. Considering as a whole, the weighted k-nearest neighbor based imputation offers a more robust method for missing value estimation than the SVD based method.

In this study, a simple yet very effective method using SVM (Shutao *et al*., 2008) and RVM classifier (Carin and Dobeck, 2003) 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

Corresponding Author: A. Bharathi, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu State

section 3. The experimental results are shown in section 4 and this study concludes in the section 5.

Related works: Proposed the Gene Selection for Cancer Classification using Support Vector Machines. In this study, the author address the 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.

Cun-Gui *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.

(Mingjun 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 and Li, 2009; Chen *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 .

Murat *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), Broaden-Fletcher-Goldfarb-Shannon (BFGS) and Liebenberg-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; (Shutao *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.

Relevance Vector Machine (RVM) for classification: RVM (Chen and Harris, 2001; Tipping, 2001) classifier is used for classification with better accuracy. Consider a two-class problem with training points X= $\{x_1,...,x_N\}$ and corresponding class labels t= $\{ t_1,\dots,t_N\}$ with $t_i \in \{0,1\}$. Based on the Bernoulli distribution, the likelihood (the target conditional distribution) is expressed as:

$$p(t|w) = \Pi_{\downarrow}(i=1)^{\uparrow} N \equiv [\![\sigma\left\{(y(x_{\downarrow}i))\right\} \\ {}^{\uparrow}(t_{\downarrow}i)\left[1-\sigma\left\{(y)x_{\downarrow}i))\right\}\right]^{\uparrow}(1-t_{\downarrow}i)]\!] \tag{1}$$

Where, σ (y) is the logistic sigmoid function:

$$\sigma(y(x)) = \frac{1}{1+\exp(-y(x))} \tag{2}$$

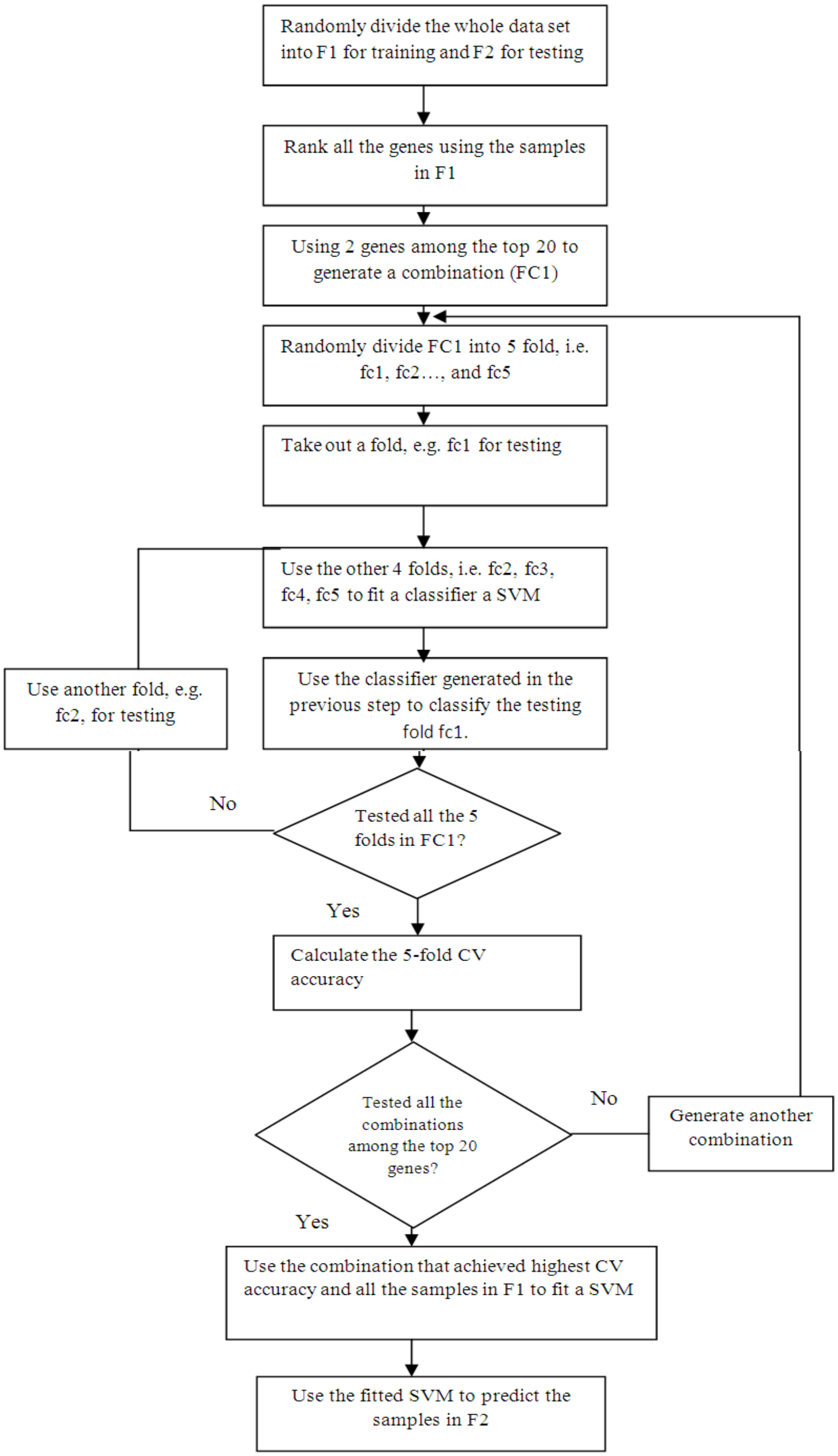
Randomly divide the whole data set into F1 for training and F2 for testing
Rank all the genes using the samples in F1
Using 2 genes among the top 20 to generate a combination (FC1)
Randomly divide FC1 into 5 fold, i.e. fc1, fc2…, and fc5
Take out a fold, e.g. fc1 for testing
Use the other 4 folds, i.e. fc2, fc3, fc4, fc5 to fit a classifier a SVM
Use another fold, e.g. fc2, for testing
Use the classifier generated in the previous step to classify the testing fold fc1.
No
Tested all the 5 folds in FC1?
Yes
Calculate the 5-fold CV accuracy
Tested all the combinations among the top 20 genes?
No
Generate another combination
Yes
Use the combination that achieved highest CV accuracy and all the samples in F1 to fit a SVM
Use the fitted SVM to predict the samples in F2

Fig. 1: Procedure for CV

Unlike the regression case, however, the marginal likelihood p (t |a) can no longer be obtained analytically by integrating the weights and an iterative procedure has to be used.

Let ai^* denotes the Maximum a Posteriori (MAP) estimate of the hyperparameter α_i. W_{MAP} denotes the MAP estimate for the weights, which can be obtained by maximizing the posterior distribution of the class labels given the input vectors. This is equal to maximizing the following objective function:

$$J(w_1, w_2, ...w_N) = \sum_{i=1}^{N} \log p(t_i \,|w_i) + \sum_{i=1}^{N} \log p(w_i \,|ai^*) \qquad (3)$$

Where, the first summation term indicates the likelihood of the class labels and the second term corresponds to the prior on the parameters wi. In the resulting solution, only those samples associated with nonzero coefficients wi (called relevance vectors) will contribute to the decision function. The gradient of the objective function J with respect to w is given by:

$$\nabla J= -A^*w-\Phi^T (f-t)$$

Where, $f = [\sigma\ (y(x_1))\ldots\ \sigma(y\ (x_N))]^{T,}$ data Φ has elements $\Phi_{i,j}=K\ (x_i,x_j)$. The Hessian of J is:

$$H = \nabla^2(J) = - (\Phi^T b\ \Phi+a^*)$$

Where, $B = \text{diag}\ (\beta_1\ldots\beta_N)$ is a diagonal data with:

$$\beta_i=\sigma(y(x_i))\ [1-\sigma(y(x_i))]$$

The posterior is approximated around MAP by a Gaussian approximation with covariance:

$$\sum\ = -\left(H\big|_{WMAP}\right)^{-1}$$

And mean:

$$\mu = \sum \phi^T B_t$$

These results are similar to the regression case and the hyperparameters α_i is updated iteratively in the same manner as for the regression case.

RESULT

The experimentation on the proposed method is carried on lymphoma data set. 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.

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 RVM 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 genes tested for all possible combinations within the 20 genes.

DISCUSSION

Table 1 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). Table 2 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). As the Table 1 and 2 suggest, more combination are obtained for using the proposed method when compared to SVM method.

Table 3 shows the comparison of the proposed classifier with the existing SVM classifier for cancer classification by means of accuracy. From the table, it can be clearly observed that the proposed method achieves better accuracy when compared to SVM.

Table 1: 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	2.80	2.90	2.11	2.14	2.15	2.16	2.18
4.7	4.12	4.17					
7.8	7.90	7.14	7.18				
8.17							
9.12	9.17						
11.17							
12.14	12.18						
14.17							
17.18							
18.20							

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

1.4	1.6	1.9	1.15	1.16	1.17	1.18	1.19
2.6	2.80	2.90	2.12	2.14	2.15	2.17	2.18
4.7	4.80	4.15	4.17				
5.7	5.90	5.11	5.15				
7.8	7.90	7.12	7.15	7.19			
8.17	8.19						
9.11	9.15	9.17					
11.16	11.18						
12.13	12.17						
14.18							
17.19							
18.20							

Table 3: Accuracy comparison

Knnimpute	No. of fold	No. of genes	No. of comb	CV Acc	Accuracy (SVM)	(Proposed) accuracy
(Data, 3)	5	20	2	91.70	96.77	97.21
(Data, 3)	5	20	3	93.97	97.60	99.16
(Data, 3)	5	10	2	92.11	96.77	98.45
(Data, 3)	5	10	3	93.31	99.70	100.00
(Data, 3)	10	20	3	93.42	97.30	98.56
(Data, 3)	10	20	2	91.26	96.77	97.12
(Data, 3)	10	10	2	91.25	96.77	98.11
(Data, 3)	10	10	3	92.47	99.87	100.00
(Data, 5)	5	20	2	93.11	98.39	99.60
(Data, 5)	5	20	3	96.40	98.40	98.93
(Data, 5)	5	10	2	94.62	98.38	96.20
(Data, 5)	5	10	3	97.15	99.23	100.00
(Data, 5)	10	20	2	93.41	98.38	99.76

By applying the data set to the proposed method, the accuracy obtained are 97.21, 99.16, 98.45, 100, 98.56, 97.12, 98.11, 100, 99.6, 98.93, 96.2, 100 and 99.76 whereas SVM technique achieves the accuracy as 96.77, 97.6, 96.77, 99.7, 37.3, 96.77, 96.77, 99.87. 98.39. 98.4, 98.38, 98.4, 98.38, 99.23 and 98.38 respectively.

CONCLUSION

For the intention of finding the smallest gene subsets for accurate cancer classification, both ANOVA and CV are highly effective ranking schemes, whereas SVM is sufficiently good classifiers. The disadvantages of SVM method is overcome by the proposed method. The usage of RVM classifier is much sparse when compared to SVM i.e., the number of relevance vectors can be much lesser than that of support vectors. The experimentation was conducted for the proposed technique using lymphoma dataset. In the lymphoma data set, the 20 selected genes are clustered using K-means method. The experimental results shows that the usage if RVM classifier helps in classifying the cancer more accurately than the conventional methods

REFERENCES

Alter, O., P.O. Brown and D. Botstein, 2003. Generalized singular value decomposition for comparative analysis of genome-scale expression datasets of two different organisms. National Academy of Sciences, 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. Proceeding of the OCEANS, Sept. 22-26, IEEE Xplore Press, USA, pp: 22-26. DOI: 10.1109/OCEANS.2003.178498

Chen, L. and S. Li, 2007. A support vector machine ensemble for cancer classification using gene expression data. Proceeding of the 3rd International Conference on Bioinformatics Research and Applications, ICBRA'07, Springer-Verlag Berlin, Heidelberg, pp: 488-495.

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

Chen, S.S.R.G. and C.J. Harris, 2001.The relevance vector machine technique for channel equalization application. IEEE Trans. Neural Net., 12: 1529-1532. PMID: 18249985

Cun-Gui, C., L. Cheng, R. Xu, 2007.Classification of FTIR gastric cancer data using wavelets and SVM", ICNC '07. Proceeding of the 3rd International Conference on Natural Computation, Aug. 24-27, IEEE computer society, China, pp: 543-547.

El-Naqa, I.Y., M.N. Yang, N.P.W. Galatsanos and R.M. Nishikawa, 2002. A support vector machine approach for detection of microcalcifications. IEEE Trans. Medical Imaging, 21: 1552-1563,

G.uyon, I., J. Weston, S. Barnhill and V. Vapnik, 2002. Gene selection for cancer classification using support vector machines. J. 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. Evolut. Comput. Mach. Learn. Data Min. Bioinformat. . 4447: 90-101. DOI: 10.1007/978-3-540-71783-6_9

Lee, Y. and C.K. Lee, 2003. Clasication of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics, 19: 1132-1139.

Mingjun, S. and S. Rajasekaran, 2010. A greedy algorithm for gene selection based on SVM and correlation. Int. J. Bioinformat. Res. Appl., 6: 296-307.

Murat, C., 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. Appl. Int. J., 36: 6357-6361. DOI: 10.1016/j.eswa.2008.08.010

Shutao, L., X. Wu and X. Hu, 2008. Gene selection using genetic algorithm and support vectors machines. Soft Comput. Fusion Found. Methodol. Appl., 12: 693-698. DOI: 10.1007/s00500-007-0251-2

Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector machine. J. Machine Learn. Res., 1: 211-244.