International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 7909-7922

http://www.ripublication.com

An Empirical Analysis of Gene Selection Using Machine Learning Algorithms For Cancer Classification

C.Devi Arockia Vanitha
Department of Computer Science,
The S.F.R.College for Women,
Sivakasi – 626123, Tamilnadu, India
vanima_c@yahoo.co.in

D.Devaraj
Department of Computer Science and Engineering, Kalasalingam University,
Krishnankoil – 626126, Tamilnadu, India
deva230@yahoo.com

M.Venkatesulu
Department of Computer Applications,
Kalasalingam University,
Krishnankoil – 626 126, Tamilnadu, India
m.venkatesulu@klu.ac.in

S.Krishnaveni
M.Phil Scholar,
Department of Computer Science,
Ayya Nadar Janaki Ammal College,
Sivakasi – 626124, Tamilnadu, India
krishnapow@gmail.com

Abstract

Gene expression analysis involves monitoring the expression levels of thousands of genes simultaneously under a particular condition using microarray technology. This is useful for the prediction and diagnosis of cancer. In order to precisely classify cancer, we have to select genes related to cancer from the noisy microarray data. Those genes are called informative genes and the process is called gene selection. In this paper, we systematically investigate four different approaches for the problem of gene selection. Two filter approaches (Mutual Information and t-test) and two wrapper approaches (Support Vector Machine – Recursive Feature Elimination and Random Forest) for gene selection have been applied on three benchmark gene expression datasets to examine their behavior in detecting group of genes that are strongly associated with cancer. To demonstrate the effectiveness of the selected genes, a classifier was developed using Artificial Neural Network. From the simulation study, it is observed that the wrapper approaches for gene selection choose genes that are biologically relevant and provide better classification accuracy.

Keywords: Gene Selection; Filter approach; Wrapper approach; Artificial Neural Network

Introduction

Gene expression analysis using microarray technology provides a way to measure the expression level of tens of thousands of genes simultaneously [1]. This technology can be useful in the classification of cancers. Cancer microarray data contains a small number of samples with a large number of gene expression levels as features. Knowing which genes are most relevant to the classification task is important. The reasons for selecting informative genes are 1) removal of noisy genes for improving the classification accuracy 2) a set of candidate genes are useful for further analysis of the disease 3) recording only a few genes in a clinical device is economical[2].

To identify the smallest possible set of genes that can achieve good predictive performance is an important issue in cancer classification [3]. Microarray data consist of large number of genes in small samples. Genes that are highly related with particular classes for classification are called informative genes [4]. This process is referred to as gene selection.

A new method of gene selection using Support Vector Machine methods based on Recursive Feature Elimination (RFE) is proposed by Guyon et al. The genes selected by SVM-RFE technique are biologically relevant to cancer and produce better classification performance and [5]. There are two general approaches to feature subset selection, namely wrappers and filters. Wrappers and filters differ in how they evaluate feature subsets.

The filter methods estimate the classification performance by evaluating the relevance of features by looking only at the intrinsic properties of the data. The wrapper methods are classifier-dependent. These methods evaluate the “goodness” of the selected feature subset directly from classifier feedback in terms of classification accuracy. Wrappers use classifiers to estimate the usefulness of feature subsets. The disadvantage of the wrapper approach is its computational requirement [2].

In this paper, an empirical analysis on gene selection using filter and wrapper approaches for cancer classification is done. Two filter approaches (Mutual Information and t-test) and two wrapper approaches (Support Vector Machine Recursive Feature Elimination and Random Forest) for gene selection are investigated and the results show that the wrapper approaches yield promising results than the filter approaches.

The primary objectives of this study are, (1) to identify and select the informative genes from gene expression data, (2) to optimize the performance of the classifier using the selected genes (3) to evaluate the performance of the four different gene selection algorithms using Artificial Neural Network and (4) to compare the filter and wrapper approaches for gene selection. All the four gene selection algorithms are evaluated on three microarray data sets: Colon cancer, Lymphoma and Ovarian data and the results are presented.

This paper is organized as follows. Section 2 deals with the importance of gene expression analysis. Section 3 describes the four different gene selection algorithms. Section 4 deals with the architecture of Artificial Neural Network used for classification. Section 5 presents the description about the data sets and experiment conducted and provides the results and discussions. Section 6 concludes the work.

Gene Expression Analysis

A gene is a segment of DNA that contains all the information necessary to create all sorts of proteins in our body. It is the unit of information that is transferred through transcription and translation. All cells have the same set of genes. [2]. A gene expression data set from a microarray experiment can be represented by a real-valued expression matrix.

Sample (S)

$$\text{Gene (G)}\begin{bmatrix} e_{11} & e_{12} & & e_{m1} \\ e_{21} & e_{22} & \cdots & e_{m2} \\ e_{31} & e_{32} & & e_{m3} \\ & \vdots & & \vdots \\ e_{m1} & e_{m2} & \cdots & e_{mn} \end{bmatrix}$$

$$\mathrm{MA} = \left\{ e_{ij} \middle| \, 1 \le i \le m, 1 \le j \le n \right\}$$

Where the rows represent the expression patterns of genes, the columns represent the expression profiles of samples, and each cell e_{ij} is the measured expression level of gene i in sample j.

Gene Expression data consists of large number of genes and limited number of samples. All genes measured by a microarray are not related to cancer classification. Some genes are irrelevant and some are redundant. The presence of irrelevant and redundant genes may affect the performance of the machine learning algorithms [2]. The analysis of gene expression data to pick out those genes whose expression patterns can distinguish phenotypes of samples is called Informative Gene Selection.

Gene Selection Algorithms

Gene selection focuses at identifying a small subset of informative genes from the initial data in order to obtain high predictive accuracy for classification. Better results of classification are obtained when increasing the number of genes. The proposed approaches for gene selection identify the discriminative genes from gene expression data, train the Artificial Neural Network classifier and then classify the test data using learned classifier. Figure 1 illustrates the schematic diagram of the proposed approach.

Two wrapper methods SVM-RFE (Support Vector Machine - Recursive Feature Elimination) and Random Forest and two filter methods Mutual Information and t-test are applied on three gene expression datasets to select the informative genes. A classifier using Artificial Neural Network (ANN) was developed to demonstrate the effectiveness of the selected genes in improving the classification accuracy.

In the filter approach, genes are selected according to the intrinsic characteristics. It works as a preprocessing step without the incorporation of any learning algorithm. In the wrapper approach, a learning algorithm is used to score the feature subsets based on the resultant predictive power, and an optimal feature subset is searched for a specific classifier [9]. The main disadvantage of the wrapper approaches is that during the feature selection process, the classifier must be called repeatedly to evaluate a subset.

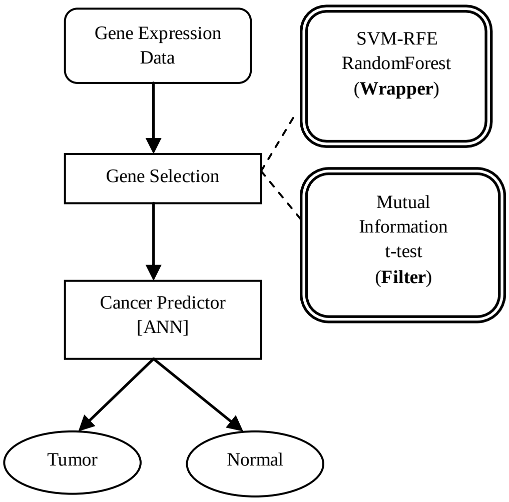


Figure 1: Schematic Diagram of the Proposed Approach

A. RandomForest

Random forest is an algorithm for classification developed by Leo Breiman [10]. In the standard random forest gene selection, the selection of the genes is done by using both the backward gene elimination and the selection based on the importance spectrum. Random forest gene elimination is carried out using the OOB (Out-Of-Bag) error as minimization criterion, by successfully eliminating the least important variables.

The procedure for building classification trees and obtaining OOB error is as follows:

(i) Make an empty bagger.
(ii) Grow and Train additional trees.

(iii) Add them to the bagger to ensemble.
(iv) Process inputs for the bagger.
(v) Prepare the OOB error values.
(vi) Finally, based on the OOB error values, the genes are sorted and ranked.

After fitting all the forests only, the OOB error rates are examined.

B. SVM-RFE

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik [11]. In SVMs the weights of the decision function are a function only of a small subset of the training examples, called "support vectors". Support vectors are closest to the decision boundary and lie on the margin. The existence of such support vectors is at the origin of the computational properties of SVM and their competitive classification performance. SVMs are suitable for the analysis of broad patterns of gene expression from DNA microarray data. They can easily deal with a large number of features and a small number of training patterns.

Recursive Feature Elimination

The idea is to compute the change in cost function caused by removing a given feature. This criteria estimate the effect of removing one feature at a time on the objective function. They become very sub-optimal when it comes to removing several features at a time, which is necessary to obtain a small feature subset. This problem can be overcome by using the following iterative procedure called Recursive Feature Elimination:

1. Train the classifier (SVM)
2. Compute the ranking criterion for all genes
3. Remove the gene with the smallest ranking criterion.

This iterative procedure is an instance of backward feature elimination [12]. SVM RFE is an application of RFE using the weight magnitude as ranking criterion.

Algorithm: SVM-RFE

```
Given training instances $X_{org} = [x_1 \cdots x_l]'$ and class labels Y = $[y_1 \cdots y_l]'$
initialize the subset of genes s = [1, 2, … ,n] and r = an empty array.
Repeat steps (i)-(v) until s has no genes.
(i)   Construct new training instances
          $X = X_{org}(:, s)$
(ii)  Train the classifier SVM(X,y)
(iii) Compute the gradient $w = \sum \alpha_i y_i X_i$
(iv)  Find the gene g with the smallest $w_j, j = 1, ... |s|$
          $g = argmin(|w_j|)$
(v)   Update r and eliminate the gene from s
          r = $[s(f), r]$ ; s = $s - \{ s(f)\}$
```

C. Mutual Information

In this work, Mutual information (MI) technique is used to select informative genes from the original gene expression profile. Mutual Information is the amount by which the knowledge provided by the feature vector decreases the uncertainty about the output [13]. The steps involved in computing the MI from the histogram of the training data are given below:

- The data set is arranged in the ascending order based on the output.
- The output class label (Y) is divided into two groups and the initial entropy H(Y) is calculated using

$$H(Y) = -\sum_{j=1}^{N_y} P(Y_i).\,\mathrm{LOG}(P(Y_i)) \qquad (1)$$

$where\ P(Y_i)$ is the probability of occurrence of the event Y = y_i.

- The input genes (X) are divided into ten levels and their conditional entropies H(Y/X) are evaluated using

$$H(Y/X) = -\sum_{i=1}^{N_x} P(X_i) \sum_{j=1}^{N_y} P\left(\frac{Y_j}{X_i}\right).\,\mathrm{LOG}\left(P\left(\frac{Y_j}{X_i}\right)\right) \qquad (2)$$

- Next, the mutual information of each gene with respect to the output is computed using

$$I(Y;X) = H(Y) - H(Y/X) \qquad (3)$$

The mutual information of all the genes is arranged in ascending order. The first ten genes that have high mutual information value are selected as features to train the artificial neural network.

D. T-TEST

A t-test [6] is any statistical hypothesis test in which the test statistic has a t distribution if the null hypothesis is true. It is applied when the population is assumed to be normally distributed but the sample sizes are small enough that the statistic on which inference is based is not normally distributed because it relies on an uncertain estimate of standard deviation rather than on a precisely known value. T-test is a parametric test for gene selection. It selects the genes that are most relevant to the disease. A ranking score is computed for each gene. It uses the following gene ranking criterion

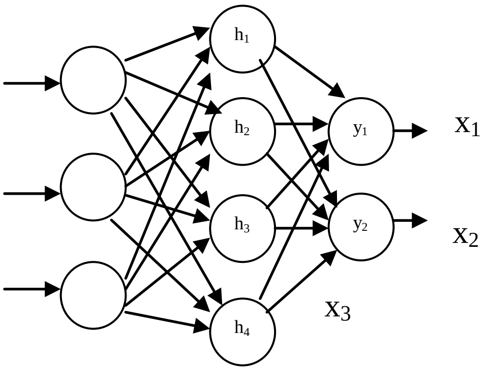
$$c_i = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}} - \qquad (4)$$

Where μ_i^+ and μ_i^- are the mean values of the ith gene respectively over cancerous and normal samples; σ_i^+ and σ_i^- are the corresponding standard deviations; n^+ and

n^{-} denote the number of cancerous and normal training samples. Eqn. (1) measures the normalized gene value difference between two classes.

Artificial Neural Network For Classification

Artificial Neural Networks [14] are parallel and distributed information processing system which consists of a huge number of simple and massively connected processors. Fig. 2 shows a multilayer feed forward network with three layers. In this neural network architecture all neurons in a layer are connected to all neurons in adjacent layers through unidirectional branches. That is, the branches and links can only broadcast information in one direction, that is, the "forward direction".



Input layer Hidden layer Output layer

Figure 2: Architecture of Feed forward neural network

Feed forward neural network training is usually carried out using the back propagation algorithm. The standard back propagation algorithm for training the network is based on the minimization of an energy function representing the instantaneous error. In other words, we desire to minimize a function defined as

$$E(m) = \frac{1}{2}\sum_{q=1}^{n}\left[d_q - y_q\right]^2 \; - \qquad (5)$$

where d_q represents the desired network output for the qth input pattern and y_q is the actual output of the neural network. Each weight is changed according to the rule:

$$\Delta w_{ij} = -k\,\frac{dE}{dw_{ij}} \; - \qquad (6)$$

where, k is a constant of proportionality, E is the error function and w_{ij} represents the weights of the connection between neuron j and neuron i. The weight adjustment process is repeated until the difference between the node output and actual output are within some acceptable tolerance.

Training the network with back propagation algorithm results in a non-linear mapping between the input and output variables. Thus, given the input/output pairs, the network can have its weights adjusted by the back propagation algorithm to capture the non-linear relationship. After training, the networks with fixed weights can provide the output for the given input.

In this work, ANN is used to classify the sample tissues into cancerous or non-cancerous. Gene selection improves classification by searching for the subset of genes, which best classifies the training data. Ten significant genes extracted from the gene selection algorithm were fed as input parameters to the ANN for classification.

Simulation Results

This section presents the details of the simulation carried out on three datasets to demonstrate the effectiveness of the proposed feature selection algorithms. The proposed approaches are implemented in MATLAB and executed in a PC with Inter Core i3 processor with 2.40 GHz speed and 4 GB of RAM. The description of the data sets is presented below:

Lymphoma

Lymphoma is a broad term encompassing a variety of cancers of the lymphatic system. The lymphoma data set includes 45 tissues x 4026 genes with two distinct tumor subtypes germinal center B cell-like DLCL and activated B cell-like DLCL [15].

Lymphoma dataset consists of 23 samples of Germinal Centre B-like and 22 samples of activated B-like. 22 out of 45 samples were used as training data and the remaining were used as test data in this work.

Ovarian

The Ovary data set was generated by hybridizing randomly selected cDNAs from normal and neoplastic ovarian tissues to membrane arrays. The proteomic spectra were generated by mass spectroscopy and the data set provided includes a collection of samples from 121 ovarian cancer patients and 95 control patients [16].

Colon

Colon dataset consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains 2000 gene expression levels. 40 of 62 samples are colon cancer samples and the remaining are normal samples [17]. 31 out of 62 samples were used as training data and the remaining were used as test data in this paper. Table I shows the details of Gene Expression Datasets used in this study.

Table 1: Summary of Gene Expression Datasets

Dataset	# Genes	# Samples	#Classes
Lymphoma	4026	45	2
Ovarian	4000	216	2
Colon	2000	62	2

Selection of relevant genes for sample classification is a common task in most gene expression studies. The top ranking genes selected using the four different gene selection algorithms in the three gene expression datasets are given in Table II and III.

Table 2: Selected Informative Genes (**Wrapper**)

Method **Dataset**	**SVM-RFE**		**RandomForest**	
Lymphoma	1276	1251	75	2077
	1279	1400	100	2200
	1280	1296	559	3348
	1264	1246	1284	2438
	1281	75	1302	1277
Ovarian	2450	2897	676	3014
	2644	2814	927	3048
	2643	2399	962	3155
	2398	2730	2333	3256
	2338	2731	2999	3320
Colon	1772	249	95	1263
	1582	138	269	1326
	513	515	625	1441
	1771	625	1223	1649
	780	1325	1258	1920

Table 3: Selected Informative Genes (**Filter**)

Method **Dataset**	**t-test**		**MI**	
Lymphoma	1276	1317	1317	1276
	1277	1291	1281	1264
	1279	1275	1279	2439
	1278	1280	1278	2438
	1281	75	1277	75
Ovarian	2814	2452	3034	2814
	2813	2644	3033	3039
	2650	2643	3032	3031
	2645	2398	3041	3025
	2642	2338	3040	2337
Colon	408	610	493	780
	179	515	1772	571
	513	625	1582	513
	1582	1325	1042	1671
	249	780	897	1423

Table IV lists the ten genes selected using MI for Lymphoma data set.

Table 4: Selected Genes using MI (Lymphoma Dataset)

Gene No.	**Gene ID**	**Gene Description**
1317	GENE3261X	Unknown; Clone=1353015
1281	GENE3332X	Unknown UG Hs.120716 ESTs; Clone=1334260
1279	GENE3330X	Unknown; Clone=825199
1278	GENE3329X	Unknown UG Hs.224323 ESTs, Moderately similar to alternatively spliced product using exon 13A [H.sapiens]; Clone=1338448
1277	GENE3328X	Unknown UG Hs.136345 ESTs; Clone=746300
1276	GENE3327X	Unknown UG Hs.169565 ESTs, Clone=825217
1264	GENE3315X	FMR2=Fragile X mental retardation 2=putative transcription factor=LAF-4 and AF-4 homologue; Clone=1352112
2439	GENE3968X	Deoxycytidylate deaminase; Clone=1302032
2438	GENE3967X	Deoxycytidylate deaminase; Clone=1185959
75	GENE3939X	Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene); Clone=1355435

Table V gives the detail of the genes selected using SVM-RFE for Colon data set.

Table 5: Selected Gene Description (Colon Dataset – SVM-RFE Method)

Gene No	Gene ID	Gene description
1772	H08393	Collagen Alpha 2(Xi) Chain (Homo sapiens)
1582	X63639	H.sapiens mRNA for p cadherin
513	M22382	Mitochondrial Matrix Protein P1 Precursor (Human)
1771	T48904	Heat Shock 27 Kd Protein (Human).
780	H40095	Macrophage Migration Inhibitory Factor (Human);
249	M63391	Human desmin gene, complete cds.
138	M26697	Human nucleolar protein (B23) mRNA, complete cds.
515	T56604	Tubulin Beta Chain (Haliotis discus)
625	X12671	Human gene for heterogeneous nuclear ribonucleoprotein core protein A1.
1325	T47377	S-100P Protein (Human).

The selected genes were used to train the Neural Network. Table VI shows the classification results obtained by using ANN as classifier.

Table 6: Classification Accuracy

Dataset / Approach	SVM-RFE	Random Forest	T-TEST	MI
Lymphoma	97.4%	100%	71.4%	95.4%
Ovarian	96.9%	93.8%	96.9%	93.5%
Colon	91.8%	77.8%	77.8%	90.3%

In general, filter approaches are much faster than wrapper approaches. However, as far as the final classification accuracy is concerned, wrappers normally provide better results. From Table 5, it is proven that the wrapper approaches SVM-RFE and Random Forest are providing promising results for Lymphoma and Ovarian datasets.

In the field of machine learning, confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Confusion matrix can be used to assess the performance of a classifier. All off-diagonal elements on the confusion matrix represent misclassified data. A good classifier will yield a confusion matrix that will look dominantly diagonal. Figure 3 shows the confusion matrix obtained after classifying the test data of ovarian gene expression data with gene selection using SVM-RFE.

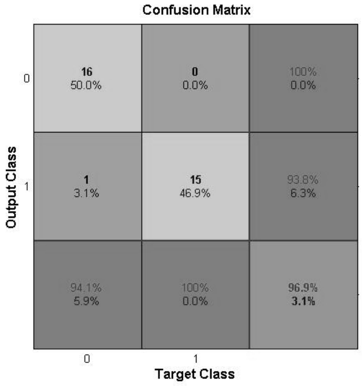


Figure 2: Confusion Matrix -Gene Selection by SVM-RFE (Ovarian Data)

To select the optimum number of genes for the neural network, the input variables are ranked using gene selection methods and the top ten genes are used to train the network. After training, the generalization performance of the network is evaluated with the test data. For comparison purpose, the number of genes selected through all the four approaches is fixed to ten; this number can be increased progressively until the maximum required accuracy is reached. The best validation performance obtained for Lymphoma data is shown in Figure 4.

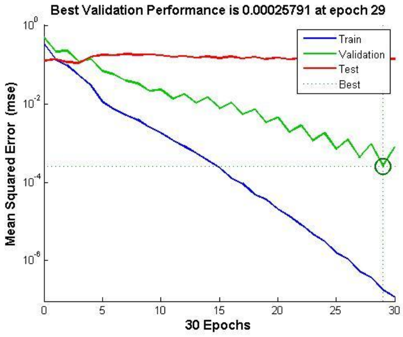


Figure 3: ANN Performance Chart on Lymphoma Data

Conclusion

This paper has presented four different approaches for gene selection in microarray data analysis. The problem of dealing with large number of features is eliminated by obtaining the feature subset for a given classifier. In order to select the informative genes from cancer microarray data and reduce dimensionality, four different feature selection algorithms were systematically investigated in this paper. The effectiveness of the proposed approaches has been demonstrated using three microarray datasets. From the simulation result, it is understood that the classification error estimated for all the datasets using wrapper approaches is minimum than the filter approaches.

References

[1] H. Chai and C. Domeniconi, 2004, “An evaluation of gene selection methods for multiclass microarray data classification”, In Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics, pages 3–10, Pisa, Italy.

[2] Y. Wang, I. V Tetko, M. A. Hall, E. Frank, A. Facius, K. F. X. Mayer, and H. W. Mewes, 2005, “Gene selection from microarray data for cancer classification — a machine learning approach”, vol. 29, pp. 37–46.

[3] R. Díaz-uriarte and S. A. De Andrés, 2006, “Gene selection and classification of microarray data using random forest”, BMC Bioinformatics, vol. 13, pp. 1–13.

[4] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, D.D. Bloomfield, and E.S. Lander, 1999, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” Science, vol. 286, no.15, pp.531-537.

[5] I. Guyon, J. Weston, S. Barnhill, M.D., and V. Vapnik, 2000, “Gene selection for cancer classification using support vector machines”, Machine Learning, 46, 389–422.

[6] J.G Thomas, J.M Olson, S.J Tapscott and L.P Zhao, 2001, “An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles”, Genome Res. 11, 1227–1236.

[7] C.A Tsai, Y.J Chen and J.J Chen, 2003, “Testing for differentially expressed genes with microarray data”, Nucl. Acids Res. 31, e52.

[8] Sung-Bae Cho and Hong-Hee Won, 2003, “Machine learning in DNA microarray analysis for cancer classification”, APBC '03 Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics, Volume 19 Pages 189-198.

[9] B. Ni and J. Liu, 2004, “A hybrid filter/wrapper gene selection method for microarray classification”, In Proceedings of International Conference on Machine Learning and Cybernetics, pages 2537–2542.

[10] L. Breiman, 2001, “Random forests”, Machine Learning, 45:5-32.

[11] B. E. Boser, I. M. Guyon, and V. N. Vapnik, 1992, “A training algorithm for optimal margin classifiers”, In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144–152, Pittsburgh, PA.

[12] R. Kohavi and G. John, 1997, “Wrappers for feature subset selection”, Artificial Intelligence, 97:12, 273–324.

[13] D. Devaraj, B. Yegnanarayana and K. Ramar, 2002, “Radial basis function networks for fast contingency ranking”, 24, 387–393.

[14] P.Ganesh Kumar and D. Devaraj, 2010, “Intrusion Detection using Artificial Neural Network with Reduced Input Features”, ICTACT Journal on Soft Computing, Issue:01, pp: 30-36.

[15] A. Alizadeh, et al., 2000, “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling”, Nature, 403(3), 503–511.

[16] T.P. Conrads, et al., 2004, "High-resolution serum proteomic features for ovarian detection", Endocrine-Related Cancer, 11, pp. 163-178.

[17] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M Schummer and N. Yakhini, 2000, “Tissue classification with gene expression profiles”, Journal of Computational Biology, 7:559-584.