Efficient Cancer Classification using Fast Adaptive Neuro-Fuzzy Inference System (FANFIS) based on Statistical Techniques

K.AnandaKumar
Assistant Professor,
Department of Computer Applications
Dr.SNS Rajalakshmi College of Arts and Science
Coimbatore,TamilNadu

Dr.M.Punithavalli
Director
Department of Computer Studies
Dr.SNS Rajalakshmi College of Arts and Science
Coimbatore,TamilNadu

Abstract- **The increase in number of cancer is detected throughout the world. This leads to the requirement of developing a new technique which can detect the occurrence the cancer. This will help in better diagnosis in order to reduce the cancer patients. This paper aim at finding 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: 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. 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. The proposed method uses Fast Adaptive Neuro-Fuzzy Inference System (FANFIS) as a classification model. This classification model uses Modified Levenberg-Marquardt algorithm for learning phase. The experimental results suggest that the proposed method results in better accuracy and also it takes lesser time for classification when compared to the conventional techniques.**

Keyword- Gene Expressions, Cancer Classification, Neural Networks, Neuro-Fuzzy Inference System, Analysis of Variance, Modified Levenberg-Marquardt Algorithm

I. INTRODUCTION

MICRO array data analysis has been successfully applied in a number of studies over a broad range of biological disciplines including cancer classification [3, 10] 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) [13] and multivariate statistical analysis method such as principal component analysis (PCA), singular value decomposition (SVD) [9] 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 [8, 11, 12] 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 (SVDimpute) and weighted k-nearest neighbors imputation (KNNimpute) has been introduced. It has been shown that KNNimpute shows better performance on non-time series data or noisy time series data, whereas, SVDimpute 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 paper, Fast Adaptive Neuro-Fuzzy Inference System (FANFIS) is used along with gene ranking technique called Analysis of Variance (ANOVA). The learning technique used in this paper is Modified Levenberg-Marquardt algorithm.

II. RELATED WORKS

Isabelle *et al.*,[1] proposed the Gene Selection for Cancer Classification using Support Vector Machines. In this paper, the author address the problem of selection of a small subset of genes from broad patterns of gene expression data [4, 5], 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 [14] performance and are biologically relevant to cancer. Jose *et al.*,

[2] presents a Genetic Embedded Approach for Gene Selection [15, 16] and Classification of Microarray Data [7, 17].

Murat *et al*., [6] gives the early prostate cancer diagnosis by using artificial neural networks. 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 BMI (body mass index), 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, PSA (prostate specific antigen), 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).

III. METHODOLOGY

Cancer classification proposed in this paper 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. This paper uses Analysis of Variance (ANOVA) method for ranking. In the second step, the classification capability of all simple two gene combinations among the genes selected are tested in this step using Fast Adaptive Neuro-Fuzzy Inference System (FANFIS) in which the training is performed using Modified Levenberg-Marquardt algorithm.

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.

Adaptive Neuro-Fuzzy Inference System (ANFIS)

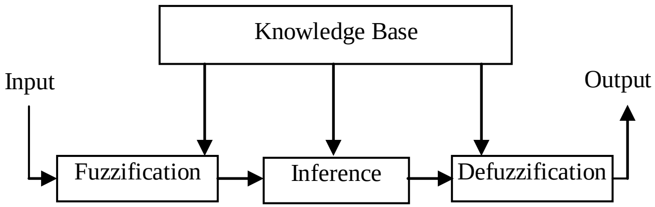

Fig.1. Adaptive Neuro-Fuzzy Inference System

The fuzzy inference system that we have considered is a model that maps

- Input characteristics to input membership functions,
- Input membership function to rules,
- Rules to a set of output characteristics,
- Output characteristics to output membership functions, and
- The output membership function to a single-valued output, or
- A decision associated with the output.

Architecture of ANFIS

The ANFIS is a framework of adaptive technique to assist learning and adaptation. This kind of framework formulates the ANFIS modeling highly organized and not as much of dependent on specialist involvement. To illustrate the ANFIS architecture, two fuzzy if-then rules according to first order Sugeno model are considered:

$$Rule\ 1: If\ (x\ is\ A_1) and\ (y\ is\ B_1) then\ (f_1 = p_1 x + q_1 y + r_1)$$
$$Rule\ 2: If\ (x\ is\ A_2) and\ (y\ is\ B_2) then\ (f_2 = p_2 x + q_2 y + r_2)$$

where x and y are nothing but the inputs, A_i and B_i represents the fuzzy sets, f_i represents the outputs inside the fuzzy region represented by the fuzzy rule, p_i, q_i and r_i indicates the design parameters that are identified while performing training process.

The ANFIS architecture to execute these two rules is represented in figure 2, in which a circle represents a fixed node and a square represents an adaptive node.

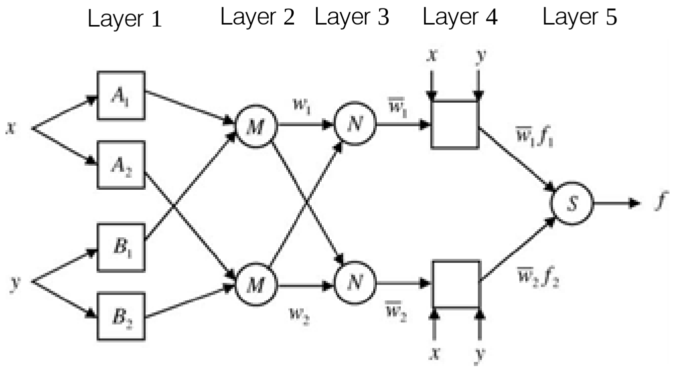

Fig.2. ANFIS Architecture

In the first layer, every node is adaptive node. The outputs of first layer are the fuzzy membership grade of the inputs that are represented by:

$$O_i^1 = \mu_{A_i}(x)\ \ i = 1,2 \tag{1}$$

$$O_i^1 = \mu_{B_{i-2}}(y)\ \ i = 3,4 \tag{2}$$

where $\mu_{A_i}(x), \beta_{B_{i-2}}(y)$, can accept any fuzzy membership function. For example, if the bell shaped membership function is employed, $\mu_{A_i}(x)$ is represented by:

$$\mu_{A_i}(x) = \frac{1}{1+\left\{\left(\frac{x-c_i}{a_i}\right)\right\}^{b_i}} \quad (3)$$

where a_i, b_i and c_i represents the parameters of the membership function, controlling the bell shaped functions consequently.

In layer 2, the nodes are fixed nodes. These nodes are labeled with M, representing that they carry out as a simple multiplier. The outputs of this layer can be indicated by:

$$O_i^2 = w_i = \mu_{A_i}(x)\mu_{B_i}(y) \;\; i = 1,2 \quad (4)$$

which are the called as firing strengths of the rules.

The nodes are fixed in layer 3 as well. They are labeled with N, representing that they are engaged in a normalization function to the firing strengths from the earlier layer.

The outputs of this layer can be indicated as:

$$O_i^3 = \bar{w}_i = \frac{w_i}{w_1+w_2} \quad i = 1,2 \quad (5)$$

which are the called as normalized firing strengths.

In layer 4, all the nodes are adaptive nodes. The output of the every node in this layer is merely the product of the normalized firing strength and a first order polynomial. Therefore, the outputs of this layer are provided by:

$$O_i^4 = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i) \qquad i = 1,2 \quad (6)$$

In layer 5, there exists only one single fixed node labeled with S. This node carries out the operation like summation of every incoming signal. Therefore, the overall output of the model is provided by:

$$O_i^5 = \sum_{i=1}^{2} \bar{w}_i f_i = \frac{\sum_{i=1}^{2} w_i f_i}{w_1+w_2} \quad (7)$$

It can be noted that layer 1 and the layer 4 are adaptive layers. Layer 1contains three modifiable parameters such as a_i, b_i, c_i that is associated with the input membership functions.

These parameters are called as premise parameters. In layer 4, there exists three modifiable parameters as well such as $\{p_i, q_i, r_i\}$, related to the first order polynomial. These parameters are called consequent parameters.

Learning algorithm of ANFIS

The intention of the learning algorithm is to adjust all the modifiable parameters such as$\{a_i, b_i, c_i\}$ and $\{p_i, q_i, r_i\}$, for the purpose of matching the ANFIS output with the training data.

If the parameters such as a_i, b_i and c_i of the membership function are unchanging, the outcome of the ANFIS model can be given by:

$$f = \frac{w_1}{w_1+w_2} f_1 + \frac{w_2}{w_1+w_2} f_2 \quad (8)$$

Substituting Eq. (5) into Eq. (8) yields:

$$f = \bar{w}_1 f_1 + \bar{w}_2 f_2 \quad (9)$$

Substituting the fuzzy if-then rules into Eq. (15), it becomes:

$$f = \bar{w}_1(p_1 x + q_1 y + r_1) + \bar{w}_2(p_2 x + q_2 y + r_2) \quad (10)$$

After rearrangement, the output can be expressed as:

$$f = (\bar{w}_1 x)p_1 + (\bar{w}_1 y)q_1 + (\bar{w}_1)r_1 + (\bar{w}_2 x)p_2 + (\bar{w}_2 y)q_2 + (\bar{w}_2)r_2 \quad (11)$$

which is a linear arrangement of the adjustable resulting parameters such as p_1, q_1, r_1, p_2, q_2 and r_2. The least squares technique can be utilized to detect the optimal values of these parameters without difficulty. If the basis parameters are not adjustable, the search space becomes larger and leads to considering more time for convergence. A hybrid algorithm merging the least squares technique and the gradient descent technique is utilized in order to solve this difficulty. The hybrid algorithm consists of a forward pass and a backward pass. The least squares technique which acts as a forward pass is utilized in order to determine the resulting parameters with the premise parameters not changed. Once the optimal consequent parameters are determined, the backward pass begins straight away. The gradient descent technique which acts as a backward pass is utilized to fine-tune the premise parameters equivalent to the fuzzy sets in the input domain. The outcome of the ANFIS is determined by using the resulting parameters identified in the forward pass.

The output error is utilized to alter the premise parameters with the help of standard backpropagation method. It has been confirmed that this hybrid technique is very proficient in training the ANFIS.

Modified Levenberg-Marquardt algorithm

A Modified Levenberg-Marquardt algorithm is used for training the neural network.

Considering performance index is $F(w) = e^T e$ using the Newton method we have as:

$$W_{K+1} = W_K - A_K^{-1} . g_K \quad (12)$$

$$A_k = \nabla^2 F(w)\Big|_{w=w_k} \quad (13)$$

$$g_k = \nabla F(w)|_{w=w_k} \quad (14)$$

$$[\nabla F(w)]_j = \frac{\partial F(w)}{\partial w_j} = 2\sum_{i=1}^{N} e_i(w) . \frac{\partial e_i(w)}{\partial w_j} \quad (15)$$

The gradient can write as:

$$\nabla F(x) = 2J^T e(w) \qquad (16)$$

Where

$$J(w) = \begin{bmatrix} \frac{\partial e_{11}}{\partial w_1} & \frac{\partial e_{11}}{\partial w_2} \cdots & \frac{\partial e_{11}}{\partial w_N} \\ \frac{\partial e_{21}}{\partial w_1} & \frac{\partial e_{21}}{\partial w_2} \cdots & \frac{\partial e_{21}}{\partial w_N} \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ \frac{\partial e_{KP}}{\partial w_1} & \frac{\partial e_{KP}}{\partial w_2} \cdots & \frac{\partial e_{KP}}{\partial w_N} \end{bmatrix} \qquad (17)$$

$J(w)$ is called the Jacobian matrix.

Next we want to find the Hessian matrix. The k, j elements of the Hessian matrix yields as:

$$\begin{aligned} \left[\nabla^2 F(w)\right]_{k,j} &= \frac{\partial^2 F(w)}{\partial w_k \partial w_j} \\ &= 2\sum_{i=1}^{N} \left\{ \frac{\partial e_i(w)}{\partial w_k} \frac{\partial e_i(w)}{\partial w_j} \right. \qquad (18) \\ &\left. + e_i(w). \frac{\partial^2 e_i(w)}{\partial w_k \partial w_j} \right\} \end{aligned}$$

The Hessian matrix can then be expressed as follows:

$$\nabla^2 F(w) = 2J^T(W).J(W) + S(W) \qquad (19)$$

$$S(w) = \sum_{i=1}^{N} e_i(w).\nabla^2 e_i(w) \qquad (20)$$

If S(w) is small assumed, the Hessian matrix can be approximated as:

$$\nabla^2 F(w) \cong 2J^T(w)J(w) \qquad (21)$$

Using (13) and (21) we obtain the Gauss-Newton method as:

$$\begin{aligned} W_{k+1} &= \\ W_k - &\left[2J^T(w_k)\cdot J(w_k)\right]^{-1} 2J^T(w_k)e(w_k) \\ \cong W_k - &\left[J^T(w_k)\cdot J(w_k)\right]^{-1} J^T(w_k)e(w_k) \end{aligned} \qquad (22)$$

The advantage of Gauss-Newton is that it does not require calculation of second derivatives.

There is a problem the Gauss-Newton method is the matrix $H = J^T J$ may not be invertible. This can be overcome by using the following modification.

Hessian matrix can be written as:

$$G = H + \mu I \qquad (23)$$

Suppose that the eigenvalues and eigenvectors of H are $\{\lambda_1, \lambda_2, \dots\dots, \lambda_n\}$ and $\{z_1, z_2, \dots\dots, z_n\}$.Then:

$$\begin{aligned} Gz_i &= [H + \mu I]z_i \\ &= Hz_i + \mu z_i \\ &= \lambda_i z_i + \mu z_i \\ &= (\lambda_i + \mu)z_i \end{aligned} \qquad (24)$$

Therefore the eigenvectors of G are the same as the eigenvectors of H, and the eigen values of G are $(\lambda_i + \mu)$.

The matrix G is positive definite by increasing μ until $(\lambda_i + \mu) > 0$ for all i therefore the matrix will be invertible.

This leads to Levenberg-Marquardt algorithm:

$$w_{k+1} = w_k - \left[J^T(w_k)J(w_k) + \mu I\right]^{-1} J^T(w_k)e(w_k) \qquad (25)$$

$$\Delta w_k = \left[J^T(w_k)J(w_k) + \mu I\right]^{-1} J^T(w_k)e(w_k) \qquad (26)$$

As known, learning parameter, μ is illustrator of steps of actual output movement to desired output. In the standard LM method, μ is a constant number. This paper modifies LM method using μ as:

$$\mu = 0.01e^T e \qquad (27)$$

Where e is a $k \times 1$ matrix therefore $e^T e$ is a 1×1 therefore $[J^T J + \mu I]$ is invertible.

Therefore, if actual output is far than desired output or similarly, errors are large so, it converges to desired output with large steps. Likewise, when measurement of error is small then, actual output approaches to desired output with soft steps. Therefore error oscillation reduces greatly.

IV. Experimental Result

Lymphoma Data Set

Lymphoma data set [18] contains 42 samples obtained from diffuse large B-cell lymphoma (DLBCL). Among these, 9 samples are from follicular lymphoma (FL), 11 samples are from chronic lymphocytic leukaemia (CLL). The whole data set contains the expression data of 4026 genes. In this data set, a small portion of data is missing. A k-nearest neighbor technique was utilized to fill those missing data. In the initial step, the 62 samples are randomly seperated into 2 groups in such a way those 31 samples for training, and 31 samples for testing. Next the complete 4026 genes are ranked with the help of ANOVA technique. Then 100 genes are taken from them with the highest rank. Then the proposed ANFIS technique is applied for classification.

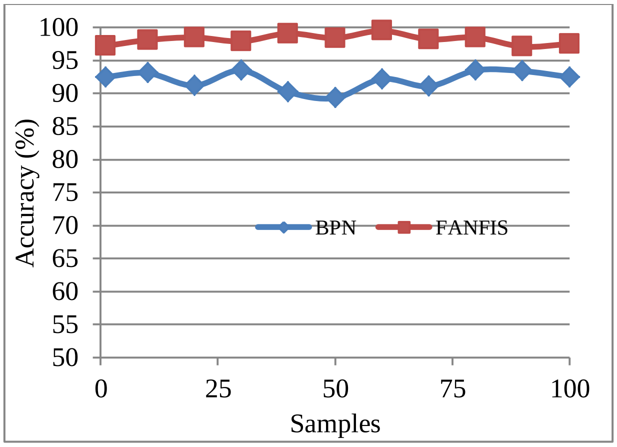

Fig.3.: Classification Accuracy for Lymphoma Data Set with ANOVA Ranking

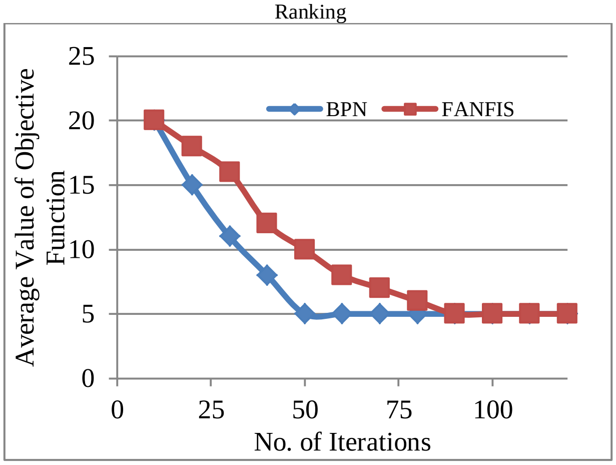

Fig.4. Convergence Behavior for Lymphoma Data Set with ANOVA Ranking

Figure 3 represents the resulted for classifying the lymphoma data set and figure 4 represents the convergence behavior of lymphoma data set.

Liver Cancer Data Set

The liver cancer data set [19] contains two classes, i.e. the nontumor liver and HCC. The data set consista 156 samples and the expression data of 1648 important genes. In that, 82 are HCCs and the remaining 74 are nontumor livers.

The data is randomly separated into 78 training samples and 78 testing samples. In this data set, there are some missing values. K-nearest neighbor technique is utilized to fill those missing values. Initially, 100 important genes are chosen in the training data set. Next all possible 1-gene and 2-gene combinations are tested within the 100 important genes.

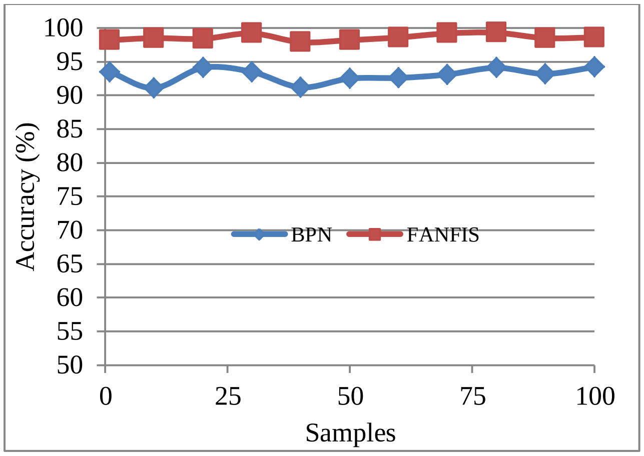

Fig.5.: Classification Accuracy for Liver Cancer Data Set with ANOVA Ranking

Figure 5 represents the resulted for classifying the liver cancer data set and figure 6 represents the convergence behavior of liver cancer data set. From these results, it can be observed that the proposed technique results in better accuracy of classification and it takes lesser time to converge.

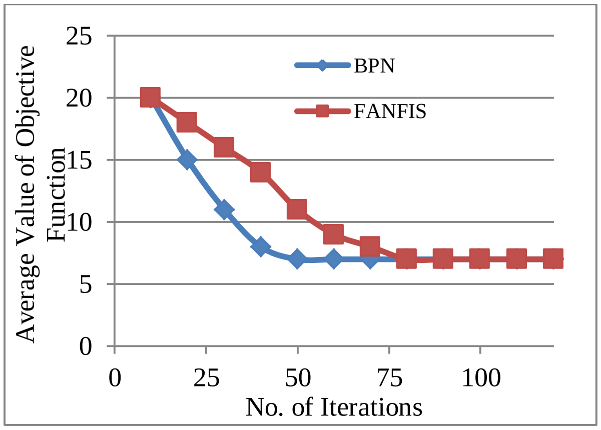

Fig.6. Convergence Behavior for Liver Cancer Data Set with ANOVA Ranking

V. Conclusion

This paper suggests a better technique for classification of cancer. In the proposed technique, the ANOVA ranking technique is initially applied to the dataset in order to find the higher ranked genes.

After ranking the genes, Adaptive Neuro-Fuzzy Inference System is used in used for classification which has both the advantages of neural network and fuzzy logic. But, it takes more time for classification.

To overcome this paper uses Fast Adaptive Neuro-Fuzzy Inference System (FANFIS). The learning is performed using the Modified Levenberg-Marquardt algorithm. The proposed technique is tested using two dataset namely, Lymphoma dataset and Liver cancer dataset. The experimental result shows that the proposed technique results in better accuracy of classification and also takes lesser time for convergence.

REFERENCES

[1] Isabelle Guyon, Jason Weston, Stephen Barnhill and Vladimir Vapnik," Gene Selection for Cancer Classification using Support Vector Machines", 2002.

[2] Jose Crispin Hernandez Hernandez, Béatrice Duval and Jin-Kao HaoGuyon, Jason Weston, Stephen Barnhill and Vladimir Vapnik," A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data", 2007.

[3] Cun-gui Cheng, Lu-yao Cheng, Run-sheng Xu ,"Classification of FTIR Gastric Cancer Data Using Wavelets and SVM",ICNC '07: Proceedings of the Third International Conference on Natural Computation - Volume 01, 2007.

[4] Mingjun Song and Sanguthevar Rajasekaran, "A greedy algorithm for gene selection based on SVM and correlation", International Journal of Bioinformatics Research and Applications, 2010.

[5] Chen Liao and Shutao Li, "A support vector machine ensemble for cancer classification using gene expression data", ISBRA'07: Proceedings of the 3rd international conference on Bioinformatics research and applications, 2007.

[6] Murat Cınar, Mehmet Engin, Erkan Zeki Engin and Y. Ziya Atesci ,"Early prostate cancer diagnosis by using artificial neural networks and support vector machines", Expert Systems with Applications: An International Journal, April 2009.

[7] Kim H, and Park H, Multi class gene selection for classification of cancer subtypes based on generalized LDA

[8] Shipp M. A et al,.Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning, Nat. Med., 8,68-74, 2002.

[9] Alter O., Brown P.O. , and Botstein D., "Generalized Singular value decomposition for comparative analysis of genome-scale expression datasets of two different organisms", Proceedings of Natural academy of Science, USA, 100(6), 3351-3356, 2003.

[10] Y. Lee and C. K. Lee, "Clasication of multiple cancer types by Multicategory Support Vector Machines using gene expression data", Bioinformatics, 19, 1132-1139, 2003.

[11] S. Chen, S. R. Gunn and C. J. Harris, "The relevance vector machine technique for channel equalization application," IEEE Trans on Neural Networks, Vol. 12, No. 6, pp. 1529-1532, 2001.

[12] Tipping M. E. "Sparse Bayesian Learning and the Relevance Vector Machine", Journal of Machine Learning Research, pp. 211-244, 2001.

[13] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," IEEE Trans. on Medical Imaging, vol. 21, 1552-1563, 2002.

[14] L. Carin and G. J. Dobeck, "Relevance vector machine feature selection and classification for underwater targets," Proceedings of OCEANS 2003, Vol. 2, pp. 22-26, 2003.

[15] Shutao Li, Xixian Wu and Xiaoyan Hu, " Gene selection using genetic algorithm and support vectors machines", Springer-Verlsg, Soft Computing - A Fusion of Foundations, Methodologies and Applications, Feb 2008.

[16] Chen Liao, Shutao Li and Zhiyuan Luo," Gene Selection Using Wilcoxon Rank Sum Test and Support Vector Machine for Cancer Classification", Springer-Verlag, Computational Intelligence and Security, april 2007

[17] Chaoyang Zhang, Peng Li, Arun Rajendran and Youping Deng," Parallel Multicategory Support Vector Machines (PMC-SVM) for Classifying Microcarray Data", IMSCCS '06: Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences, 2006.

[18] http://llmpp.nih.gov/lymphoma

[19] http://genome-www.stanford.edu/hcc

AUTHORS PROFILE

Mr.K.AnandaKumar was born in TamilNadu, India on March 1975. He received the B.Sc Degree in Physics from Bharathiar University in 1995. He received his MCA Degree in Computer Applications from Bharathiar University in 1998. He received his M.Phil from Periyar University in 2006 and he is doing Doctor of Philosophy in Computer Science and Engineering from Bharathiar University, Coimbatore. He had 13 years of teaching experience. Currently he is working as HOD in Computer Applications Department, Dr.SNS Rajalakshmi College of Arts and Science College, Coimbatore. His Professional activities include…Guided Twenty PG projects and Ten M.Phil and guiding Ten PG and Six M.Phil projects. Published and presented 8 papers in International and National Conferences and 6 national and international journals.

Dr. M. Punithavalli received the Ph.D degree in Computer Science from Alagappa University, Karaikudi in May 2007. She is currently serving as the Director of the Computer Science Department, Sri Ramakrishna college of Arts and Science for Women, Coimbatore. Her research interest lies in the area of Data mining, Genetic Algorithms and Image Processing. She has published more than 10 Technical papers in International, National Journals and conferences. She is Board of studies member various universities and colleges. She is also reviewer in International Journals. She has given many guest lecturers and acted as chairperson in conference. Currently 10 students are doing Ph.D under her supervision