Investigations on Hybrid Learning in Anfis in Microarray Gene Expression Classification

C.Loganathan[1] & K.V.Girija[2]

[1]Dept of Mathematics, Maharaja Arts and Science College, Coimbatore-641407, Tamilnadu, India,
[2] Dept of Mathematics, Hindustan College of Engineering and Technology, Coimbatore -641032, Tamilnadu, India

Abstract- This paper discusses the performance of proposed learning method of Hybrid Back Propagation Neuro fuzzy Method (BPN) and Runge-Kutta Learning Method(RKLM). The proposed learning method is evaluated in the application of cancer classification. First, a classifier is trained with a part of samples in the cancer data set. Then one uses the trained classifier to predict the samples in the rest of the dataset to evaluate the effectiveness of the classifier. The analysis of the cancer classification using Hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) is explained in this paper.

Index Terms- Learning Method, Microarray Gene Expression, BPN, ANFIS, RKLM.

I. INTRODUCTION

In current years, DNA microarray technology has been established as an extremely powerful tool for the expression of genes. Analyzing the great amount of gene expression data from microarray chips it is understood that it can perform a very essential role in decease diagnosis, particularly in cancer diagnosis. It also offers an opportunity and a challenge for present machine learning research.

II. MICROARRAY GENE EXPRESSION

Cells are the main basic units of all organisms on earth, except for viruses, e.g., yeast has only one cell, while any of the mammals, hold tons of cells. Within a cell, there is a nucleus, and inside a nucleus, there are quite a number of divided long segments called chromosomes which is prepared by Deoxyribonucleic Acid (DNA). The basic units of DNA are nucleotides which consist of sugar phosphate backbone and four bases Adenine (A), Cytosine(C), Guanine (G), and Thymine (T). A pairs with T [3], while C pairs with G. DNA codes the inherited information through particular order of these base pairs on a double-stranded helix for reproduction of organisms (see Figure 1).

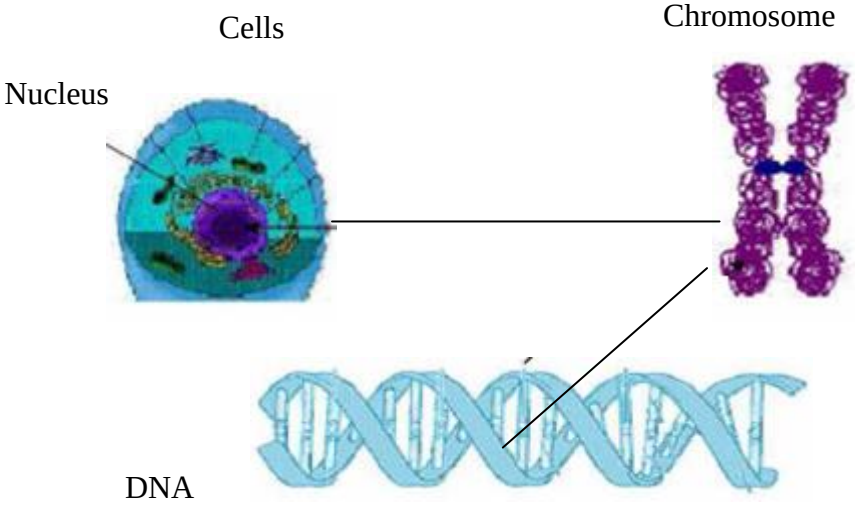

Figure.1 An overview of the relationship among cell, nucleus, chromosome, gene and DNA.

DNA has coding and non-coding segments, and the coding segments are called genes. The technique from genes to proteins includes two steps: First, DNA is recorded into messenger ribonucleic acid, via mRNA or RNA in short. Second, the mRNA transforms into proteins. Basically all cells in the related organism have similar genes, but these genes can be expressed in more than one way at different times and under different situations. The majority of molecular biology investigations focused on mRNA level due to the information that all most important differences in cell state or type are connected with changes in the mRNA levels of many genes [5].

III. CANCER GENE EXPRESSION CLASSIFICATION

Zhenyu Wang [16] proposed Neuro-fuzzy modeling for microarray cancer gene expression data. It is easy and very successful technique that applies to accurate cancer classification using expressions of only very few genes. The cancer classification is proposed by Niyue Tan [11]. The cancer type and phase are often extremely important to the assignment of the suitable treatments. It is identified that mutations in genes can lead to cancer. Standard cells can develop into malignant cancer cells through a series of mutations in genes that manage the cell cycle and genome integrity, to name only a few [7]. These mutations are missing in normal cells, and this attempt expects the expression levels of these genes, and genes regulated by these genes, to be different in usual and cancerous cells. By observing these differences, it is now achievable to classify cells as cancerous or usual by measuring the expression levels of a variety of genes present in the cells.

Initially, plan the microarray experiments according to a biological problem that need to be learnt. Normally, microarray experiments can be separated into two categories. One focuses on time series information which contains the gene expression data of a variety of genes under a range of research. An additional type of microarray experiment consists of gene expression data of a variety of genes taken from various tissue samples or under different experimental conditions, such as nutrition, temperature, chemical environment and etc.

The gene expression profiles from particular microarray experiments have been newly utilized for cancer classification. This advance promises to give an improved therapeutic measurement to cancer patients by diagnosing cancer variety with enhanced accuracy [13]. However, the feature of data produced by this new technology is more than one can manually investigate.

IV. GENE RANKING USING T-SCORE

To find out how the gene importance ranking scheme influences the classification result, T-Score (TS) ranking scheme was used [17]. In the datasets, the combinations were tested within the top 100 genes selected by the ranking scheme. Compute the importance ranking of each gene using a feature ranking measure, one of which is described below.

T-Test: The t-score (TS) of gene *i* is defined as follows:

[1] $$TS_i = \max\left\{\left|\frac{\bar{x}_{ik}-\bar{x}_i}{m_k s_i}\right|, k = 1,2,\dots K\right\}$$ [2] (4.1)

where

[3] $$\bar{x}_{ik} = \sum_{j\in C_k} \bar{x}_{ij}/n_k$$ [4] (4.2)

[5] $$\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$$ [6] (4.3)

[7] $$s_i^2 = \frac{1}{n-K}\sum_k \sum_{j\in C_k}(x_{ij} - \bar{x}_{ik})^2$$ [8] (4.4)

[9] $$m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}}$$ [10] (4.5)

There are k classes. $\max\{y_k, k = 1, 2, \dots K\}$ is the maximum of all y_k. C_k refers to class k that includes n_k samples, x_{ij} is the expression value of gene i in sample j, $\bar{x}_{ik}$ is the mean expression value in class k for gene I, n is the total number of samples, $\bar{x}_i$ is the general mean expression value for gene i and s_i is the joint within-class standard deviation for gene i. In fact, the TS used here is a t-statistic between the centroid of a specific class and the overall centroid of all the classes. Another possible model for TS could be a t-statistic between the centroid of a specific class and the centroid of all the other classes.

V. GENE SELECTION

The amount of features (usually in the range of 2000-30000) is greatly bigger than the amount of samples (usually in the range of 40-200). When such data is offered, various standard data analysis and machine learning methods are either unsuitable or become computationally infeasible [12]. A large amount of genes are not related to the presentation of the classification. Taking such genes into account during classification enlarge the dimension of the classification difficulty, poses computational complexity, and introduces unnecessary noise in the procedure. A main goal for diagnostic research is to develop diagnostic procedures based on inexpensive microarrays that have sufficient probes to discover diseases. Thus, it is crucial to identify whether a small amount of genes can be sufficient for good classification. This necessitates selection of several genes that are extremely related to particular classes for classification, which are called informative genes. This procedure is called gene selection, or feature selection in machine learning in common [8]. Some modern research has exposed that a small amount of genes are enough for accurate diagnosis of the major cancers, even though the amount of genes vary greatly among different diseases.

Therefore, the microarray cancer classification problems are classified as the combinational optimization problem with two major objectives: diminishing the amount of selected genes and maximizing the classification accuracy.

The difficulty of feature selection lies in a thorough treatment in pattern recognition and machine learning. The gene expression datasets are difficult to contain a huge amount of genes. Moreover, these datasets hold only a small amount of samples, so the discovery of irrelevant genes can suffer from statistical instabilities.

VI. FINDING THE MINIMUM GENE SUBSET

After selecting some top genes in the importance ranking list, it attempts to classify the dataset with only one gene. This work input each selected gene into our classifier. If no good accuracy is obtained, go on classifying the dataset with all the possible 2-gene combinations within the selected genes. If still

no good accuracy is obtained, then repeat this procedure with all the 3-gene combinations and so on, until obtain a good accuracy. This work used the following three learning methods in ANFIS classifier to test gene combinations.

i. BPN
ii. Hybrid (BPN and LSE)
iii. Hybrid (BPN and RKLM)

For Back propagation, Hybrid BPN and Least Square Estimator (LSE) and Hybrid BPN and RKLM carried out 5-fold Cross-Validation (CV) in the training dataset to tune their parameters.

VII. THE CANCER CLASSIFICATION PROBLEM

Cancer classification is central to cancer treatment. Further subdivision of morphologically similar tumors can be made at molecular level; traditionally cancer classification relied on specific biological insights, rather than on systematic and unbiased approaches. Cancer classification can be divided into two challenges: class detection and class prediction. Class detection refers to defining previously unrecognized tumor subtypes. Class prediction refers to the assignment of particular tumor samples to already-defined classes. To develop a more systematic approach to cancer classification based on the simultaneous expression monitoring of hundreds of genes using DNA microarrays as test cases [14,15].

The gene expression data from cancer samples, namely the cancer gene expression data allow comparison of gene expression levels between normal and cancer cells, so that further analysis work could be performed to find out the 'Internal pattern' which may serve as a classification technique.

Classification difficulty has been expansively studied by researchers in the area of statistics, machine learning and databases. Numerous classification algorithms have been developed in the history, such as the decision tree methods, the linear discrimination analysis, the Bayesian network and etc., [9]. For the last few years, researchers have started paying concentration to the cancer classification using machine learning. Studies have shown that gene expression changes are related with different types of cancers.

Most planned cancer classification techniques are from the statistical and machine learning areas, ranging from the old adjacent neighbor analysis, to the fresh approach. There is no single classifier that is greater over the rest. Some of the techniques only works well on binary-class problems and not extensible to multi-class problems, while others are more general and flexible [2]. One thing to note for most of those planned algorithms on gene classification is that the authors are only concerned with the precision of the classification and do not give much attention to the running time.

Cancer classification using gene expression data stands out from the additional preceding classification data due to its unique nature and application field. Through this study it is hoped to achieve some insight into the problem of cancer classification in aid of further increasing more efficient and capable classification algorithms.

7.1 *Training and testing of data*

The proposed approach computes the feature ranking score from a statistical analysis of weight vectors of multiple linear trained on subsamples of the original training data. Then the proposed method is tested on four gene expression datasets for cancer classification. The results show that the proposed feature selection method selects better gene subsets than the standard method and improves the classification accuracy. A Gene Oncology-based similarity assessment indicates that the selected subsets are functionally diverse, further validating proposed gene selection method. This investigation also suggests that, for gene expression-based cancer classification, average test error from multiple partitions of training and test sets can be recommended as a reference of performance quality.

A. Terminologies and Problem Statement

This chapter defines and introduces some terminologies and notations that will be used throughout the section for the problem of cancer classification using gene expression data, termed cancer classification, for briefness [10].

Let $X_1, X_2, \ldots, X_m$ be random variables for genes $G_1, G_2, \ldots, G_m$ respectively, where X_i has domain $dom(X_i)$ which is the range of expression values for gene G_i. Let C be the random variable for the class labels, and $dom(C) = \{1, \ldots, K\}$, where K denotes the total number of classes. Let $t = \{t.X_1, t.X_2, \ldots, t.X_m\}$ denotes a size m tuple of expression values for m genes. Let $T = \{(t_1, c_1), (t_2, c_2), \ldots, (t_n, c_n)\}$ denoting a training set of n tuples, where $i = \{1, 2, \ldots, n\}$, $c_i \in dom(C)$ is the class label of tuple t_i. Let the test set be $S = \{t_1, t_2, \ldots, t_l\}$ where l is the size of the test set.

A Classifier is a function class with two arguments, T and S, where T denotes the training samples and S is a testing sample. Function class returns a class prediction for sample s. The classification accuracy is defined as the number of correct predictions made by the classifier on a set of testing tuples using the function Class trained on the training tuples.

B. Main problem in cancer classification

Given a training set $T = \{(t_1, c_1), (t_2, c_2), \ldots, (t_n, c_n)\}$, where t_i s are independent m-dimensional random data tuples of gene expression values, m is the total number of genes, $t_i = (t_i.X_1, t_i.X_2, \ldots, t_i.X_m)$, $m > n$ and $c_i \in dom(C)$ is the class label of the i^{th} tuple. Given a test set $S = \{s_1, s_2, \ldots s_l\}$. Each s_i is a gene expression data tuple of length m. Each s_i is in the form of $\{s_i.X_1, s_i.X_2, \ldots, s_i.X_m\}$, where x_j is the expression value of gene j. Find a classification function class, which gives maximal classification accuracy on S.

The gathering of well-distributed, sufficient and accurately calculated input data is the basic condition to achieve an exact model [1]. Selection of the ANFIS inputs is the major important task of designing the classifier, since even the greatest classifier will carry out poorly if the inputs are not chosen sufficiently well. It is tricky for ANFIS to handle high dimensional problems, as this leads to a huge amount of input nodes, rules and hence resultant parameters.

VIII. EXPERIMENTAL RESULTS

The experiment was done using MATLAB 7 [4] under windows environment. There are various classifiers are studied in the literature but in this research only ANFIS is considered for classification, because the proposed learning model applied only with ANFIS classifier. Results that are achieved by using ANFIS classifier are encouraging. It analyzes the classification performance of the cancer using proposed learning methods in ANFIS classifier. In this assessment, the classification performance, together with the training error rate is considered as the primary comparison measures. From the result it is seen that the performance of ANFIS with RKLM gives the best in estimation.

The work, reported in this research, indicates that ANFIS structure is a good candidate for classification purposes. Additionally, it is a smart performance of the RKLM approach with on-line operation and with ANFIS

A. Datasets

There are many different benchmark microarray datasets, reported in published cancer gene expression studies, including leukemia cancer dataset, colon cancer dataset, lymphoma dataset, breast cancer dataset and ovarian cancer dataset [6]. In this research, the proposed learning models are tested on three datasets: leukemia cancer dataset, lymphoma cancer dataset and Small Round Blue Cell Tumour (SRBCT) cancer dataset.

The aim of testing on several different datasets is not only to show that proposed models are better or worse, but also to find out when new models performs better and why, what are the reasons causing the unsatisfying results, and how to solve the problems.

B. Accuracy, Error rate and Execution Time for Proposed Hybrid ANFIS

The average accuracy and the execution time is shown in the below tables and figures.

Table 1: Average accuracy

[11] Dataset	[12] No. of gene selected	[13] Average Accuracy (%)		
		[14] BPN	[15] Hybrid [16] (BPN and LSE) [17]	[18] Hybrid [19] (BPN and RKLM)
[20] Leukemia	[21] 2	[22] 89	[23] 93	[24] 97
	[25] 3	[26] 89	[27] 95	[28] 98
[29] Lymphoma	[30] 2	[31] 92	[32] 93	[33] 96
	[34] 3	[35] 93	[36] 93	[37] 96
[38] SRBCT	[39] 2	[40] 93	[41] 95	[42] 97
	[43] 3	[44] 95	[45] 96	[46] 97

Table 1 shows the average accuracy for ANFIS classifier with proposed Hybrid BPN and RKLM learning. Then the proposed learning model Hybrid BPN and RKLM is higher when compared with other two learning methods in ANFIS.

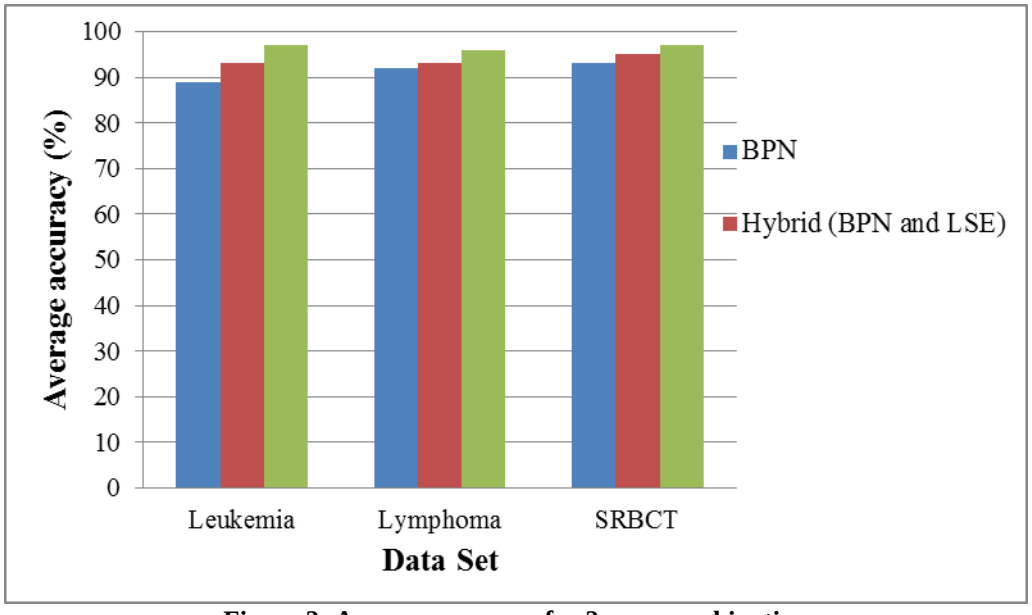

Figure 2: Average accuracy for 2-gene combination

Figure 2 shows the average accuracy for 2-gene combinations for leukemia, lymphoma and SRBCT datasets and proposed learning model. The proposed learning method of Hybrid BPN and RKLM has high accuracy when compared with other learning methods.

Table.2: Average error rate for 2-gene combinations

[47] **Hybrid Methods**	[48] Error rate
[49] BPN	[50] 1.750
[51] Hybrid (BPN and LSE)	[52] 1.573
[53] Hybrid (BPN and RKLM)	[54] 1.251

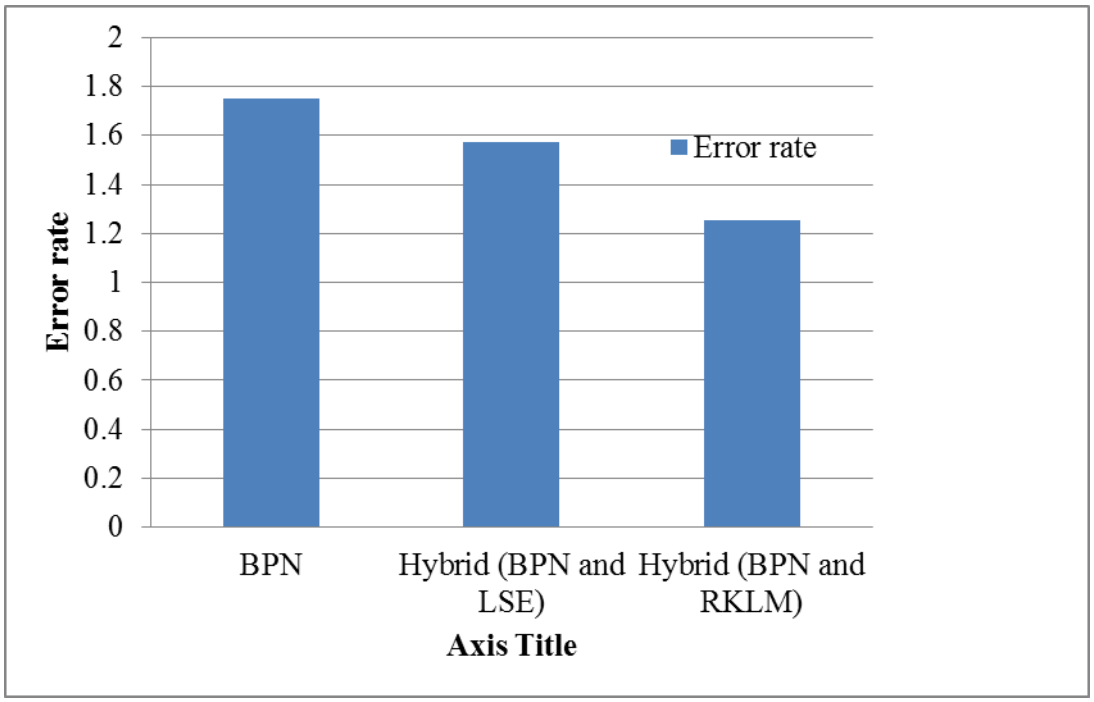

Figure 3: Error rate for 2-gene combinations

Table 2 and figure 3 show the error rate of the various learning model applied in ANFIS classifier. The proposed Hybrid learning of BPN and RKLM has less error rate when compared to other two standard learning methods.

Table 3: Execution time

[55] Dataset	[56] No. of gene selected	[57] Execution Time (Seconds)		
		[58] BPN	[59] Hybrid [60] (BPN and LSE)	[61] Hybrid [62] (BPN and RKLM)
[63] Leukemia	[64] 2	[65] 59	[66] 45	[67] 34
	[68] 3	[69] 75	[70] 70	[71] 41
[72] Lymphoma	[73] 2	[74] 39	[75] 31	[76] 25
	[77] 3	[78] 53	[79] 57	[80] 26
[81] SRBCT	[82] 2	[83] 28	[84] 19	[85] 10
	[86] 3	[87] 43	[88] 35	[89] 12

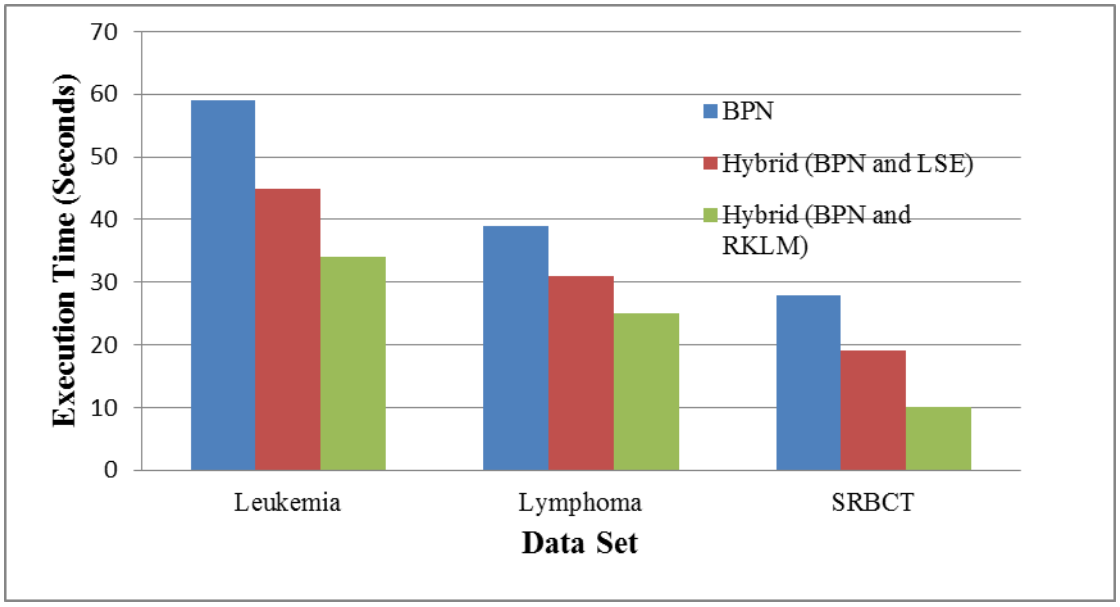

Figure 4 Execution time for the ANFIS methods

Figure 4 shows the execution time for proposed method of Hybrid BPN and RKLM. Hybrid BPN and RKLM have less execution when compared with ANFIS methods.

IX. Conclusion

Systematic and unbiased approach to cancer classification is of great importance to cancer treatment and drug discovery. Previous cancer classification methods were all clinical based and were limited in their diagnostic ability. It has been known that gene expressions contain the keys to the fundamental problems of cancer diagnosis, cancer treatment and drug discovery. The gene expression data classification using ANFIS may use list of datasets like Leukemia, Lymphoma and SRBCT for the proposed method. Thus the above table and chart show the proposed learning method of Hybrid BPN and RKLM have high accuracy with less time when compared with the other ANFIS methods.

References

[1] Belal. S, Taktak. A, Nevill. A, Spencer. S, Roden. D and Bevan. S, Automatic detection of distorted plethysmogram pulses in neonates and paediatric patients using an adaptive-network-based fuzzy inference system, Artificial Intelligence in Medicine, 149C165, (2002).

[2] Boardman, David and Alison Flynn, A gamma-ray identification algorithm based on fisher linear discriminant analysis, IEEE Transactions on Nuclear Science, 60(1), (2013), Pp: 270-277.

[3] Brandle. N, Bischof. H and Lapp. H, Robust DNA microarray image analysis, Machine Vision and Applications, (2003).

[4] Cho. S and Won. H, Machine learning in DNA microarray analysis for cancer classification, In Asia-Pacific Bioinformatics Conference, vol. 34, (2003), Pp 189–198.

[5] Dettling. M, Supervised learning in very high dimensional problems with application to microarray data, Ph.D thesis, Swiss Federal Institute of Technology, Zurich, (2004).

[6] Gabriella Rustici, Nikolay Kolesnikov, Marco Brandizi, Tony Burdett, Miroslaw Dylag, Ibrahim Emam and Anna Farne, Array Express update-trends in database growth and links to data analysis tools. Nucleic acids research, 41, no. D1, D 987-D990, (2013).

[7] Gregory. P. S and Pablo. T, Microarray data mining facing the challenges, SIGKDD Explorations, (2003).

[8] Inza, Larranaga. P, Blanco. R and Cerrolaza. A. J, Filter versus wrapper gene selection approaches in DNA microarray domains, Artificial Intelligence in Medicine, (2004), Pp: 91–103.

[9] Khashei, Mehdi, Ali Zeinal Hamadani and Mehdi Bijari, A fuzzy intelligent approach to the classification problem in gene expression data analysis-Knowledge-Based Systems, 27, (2012), Pp: 465-474.

[10] Mitchel. T, Machine Learning, McGraw-Hill, New York, (1997).

[11] Niyue Tan, Cancer Gene Expression Data Analysis: a Neuro-Fuzzy System Approach, Ph.D Thesis, University of Oxford, (2007).

[12] Rogers. S, Machine Learning Techniques for Microarray Analysis, Ph.D. thesis, University of Bristol, UK, (2004).

[13] Slonim. D.K, Tamayo. P, mesirov. J.P, Golub. T.R and Lander. E.S, Class Prediction and Discovery Using Gene Expression Data, Proceedings of Annual International conference on Research in Computational Molecular Biology (2000), Pp: 263–272.

[14] Xiong. M, Li. W, Zhao. J, Jin. L and Boerwinkle. E, Feature (gene) Selection in Gene Expression-based Tumor Classification, Molecular Genetics and Metabolism, (2001), Pp: 239–247.

[15] Yu. L and Liu. H, Redundancy based feature selection for microarray data, Technical Department of Computer Science and Engineering, Arizona State University, (2004).

[16] Zhenyu Wang, Neuro-fuzzy modeling for microarray cancer gene expression data, First Year Transfer Report, University of Oxford, (2005).

[17] Zou Bin, Luoqing Li, Zongben Xu, Tao Luo and Yuan Yan Tang, Generalization performance of Fisher linear discriminant based on Markov sampling, IEEE Transactions on Neural Networks and Learning Systems, 24(2), (2013), Pp: 288-300.

Authors

First Author – C.Loganathan, Dept of Mathematics, Maharaja Arts and Science College, Coimbatore-641407, Tamilnadu, India, E-mail Id: clogu@rediffmail.com
Second Author – K.V.Girija, Dept of Mathematics, Hindustan College of Engineering and Technology, Coimbatore -641032, Tamilnadu, India, E-mail Id: kvgirijamaths@gmail.com