International Journal of Computer Science and Informatics

Volume 1 | Issue 1 Article 6

July 2011

A Hybrid Method of Feature Extraction for Tumor Classification Using Microarray Gene Expression Data

Sitanshu Sekhar Sahu
National Institute of Technology, Rourkela, Odisha, India, sitanshusekhar@gmail.com

G. PANDA
School of Electrical Sciences , Indian Institute of Technology, Bhubaneswar, INDIA, ganapati.panda@gmail.com

Ramchandra Barik
cSambalpur University Institute of Information Technology, Odisha, India, ramchbarik@gmail.com

A Hybrid Method of Feature Extraction for Tumor Classification Using Microarray Gene Expression Data

Sitanshu Sekhar Sahu[a], Ganapati Panda[b], Ramchandra Barik[c]
[a]National Institute of Technology, Rourkela, Odisha, India
[b]Indian Institute of Technology, Bhubaneswar, Odisha, India
[c]Sambalpur University Institute of Information Technology, Odisha, India
E-mail: sitanshusekhar@gmail.com, ganapati.panda@gmail.com, ramchbarik@gmail.com

***Abstract*— Classification of disease phenotypes using microarray gene expression data faces a critical challenge due to its high dimensionality and small sample size nature. Hence there is a need to develop efficient dimension reduction techniques to improve the class prediction performance. In this paper we present a hybrid feature extraction method to combat the dimensionality problem by combining F-score statistics with autoregressive (AR) model. The F-score statistics preselect the discriminant genes from the raw microarray data and then this reduced set is modeled by the AR method to extract the relevant information. A low complexity radial basis function neural network (RBFNN) is also introduced to efficiently classify the microarray data. Exhaustive simulation study on six standard datasets shows the potentiality of the proposed method with the advantage of reduced computational complexity.**

Keywords- Microarray, F-score, AR modeling, RBFNN

I. INTRODUCTION

Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalousness occurring in the functioning of the human body. With the advanced statistical techniques, microarray analyses enable simultaneous study of the entire genome in one experiment quickly and in efficient manners which were once thought to be non-traceable. The main data analysis techniques used currently in biomedical applications related to microarrays are: gene selection, classification, clustering and understanding the mechanism of disease at molecular level and defining of drug targets. Among these diseases phenotype classification has gained a special interest. Gene's expressions are examined at different conditions or different cellular stages to reveal the functions of genes as well as their regulatory interactions. Gene expression of disease tissues may be used to gain a better understanding of many different diseases including heart disease, mental illness, infectious diseases. Study of different types of cancers with its classification through the gene expression profiles which has pulled the attention of many research communities as it is important for subsequent diagnosis and treatment.

Generally the microarray experiments produce large datasets having expression levels of thousands of genes with a very few numbers (up to hundreds) of samples which leads to the problem of "Curse of dimensionality". Due to the high dimension, the accuracy of the classifier decreases as it attains the risk of overfitting. Although the microarray data contains thousands of genes from the experiment, not all genes are informative for classification because they are either irrelevant or redundant. Hence to deduce a subset of informative or discriminative genes from the entire gene set is necessary and challenging task in microarray data analysis. The purpose of gene selection or dimension reduction is to simplify the classifier by retaining small set of relevant genes and to improve the accuracy of the classifier. For this purpose, researchers have applied a number of test statistics or discriminant criteria to find genes that are differentially expressed between the investigated classes which upshot to not only provide molecular and genomic understanding on how these genes are related to different classes of diseases but also reduce misclassification rates for prediction.

Various methods and techniques have been developed in recent past to perform the gene selection to reduce the dimensionality problem. The filter method basically use a criterion relating to rank and select key genes for classification such as Pearson correlation coefficient method [1], t-statistics method [2], signal-to-noise ratio method [3], the partial least square method. However, it does not account for interactions between features. Many transformation methods such as independent component analysis [4], linear discriminant analysis, principal component analysis [5] and wavelet analysis [6, 7] have also applied to reduce the dimension of the data. All these methods transform the original gene space to another domain providing reduced uncorrelated discriminant components. It requires a large matrix computation which increases computational complexity. In this paper we propose a hybrid method which combines both the feature selection and extraction to get the optimal relevant and discriminative genes for classification. A F-score statistics is used to preselect the discriminative genes from the

microarray data, then a model is developed by auto regressive (AR) method to extract the relevant information from the samples.

Several Machine learning and statistical techniques have been applied to classify the microarray data. Tan and Gilbert [8] used the three supervised learning methods such as C4.5 decision tree, bagged and boosted decision tree to predict the class label of the microarray data. Dettling [9] have proposed an ensemble method of bag boosting approach for the same purpose. Many authors have used successfully the support vector machine (SVM) for the classification of microarray data [10]. Khan et al. [11] used the neural networks to classify the subcategories of small round blue-cell tumors. Also O'Neill and song [12] used the neural networks to analyze the lymphoma data and showed very good accuracy. B Liu et al. [13] proposed an ensemble neural network with combination of different feature selection methods to classify the microarray data efficiently. But the neural networks require a lot of computation and consume more time to train. In this paper we have introduced a new promising low complexity neural network known as radial basis function neural network (RBFNN) to efficiently classify the microarray data.

II. MATERIALS AND METHODS

A. Data sets

In this section, the cancer gene expression data sets used for the study are described. These datasets are also summarized below.

ALL/AML Leukemia Dataset [3]

This dataset consists of two distinctive acute leukemias, namely AML and ALL bone marrow samples with 7129 probes from 6817 human genes. The training dataset consists of 38 samples (27 ALL and 11 AML) and the test dataset consists of 34 samples (20 ALL and 14 AML).

SRBCT Dataset [11]

This dataset consists of four categories of small round blue cell tumors (SRBCT) with 83 samples from 2308 genes. The tumors are Burkitt lymphoma (BL), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS). There are 63 samples for training and 20 samples for testing. The training set consists of 8, 23, 12 and 20 samples of BL, EWS, NB and RMS respectively. The testing set consists of 3, 6, 6 and 5 samples of BL, EWS, NB and RMS respectively.

MLL Leukemia dataset [18]

This dataset consists of three types of leukemias namely ALL, MLL and AML with 72 samples from 12582 genes. The training dataset consists of 57 samples (20 ALL, 17 MLL and 20 AML) and the test data set consists of 20 samples (4 ALL, 3 MLL and 8 AML).

Prostate dataset [14]

This dataset consists of prostate tissue samples from 12,600 genes. The training dataset consists of 102 samples out of which 52 are from prostate tumor tissue samples and 50 are from normal tissue samples.

Lymphoma dataset [12]

This dataset consists of three most prevalent adult lymphoid malignancies. It consists of 62 samples from 4026 genes. This composes 42, 9 and 11 samples of DLBCL, FL and CL respectively.

Colon dataset [19]

This dataset consists of 62 samples of 2000 genes collected from colon cancer patients. Among these 40 samples are from tumor tissues and 22 are from healthy parts of the colons of the same patients.

B. F-score based feature selection method

F-score method is a statistical technique that measures the distinguishing power between two classes with real values [16]. In this method, a F-score value of each feature in the dataset is computed to show their discriminative power. The F-score value of k^{th} feature of a two class problem is defined as:

$$F(k) = \frac{(\bar{x}_k^{c1} - X)^2 + (\bar{x}_k^{c2} - X)^2}{(\sigma_k^{c1})^2 + (\sigma_k^{c2})^2} \qquad (1)$$

where X= average of the total samples of the k^{th} feature, $\bar{x}_k^{c1}, \bar{x}_k^{c2}$ are variances of the *c1* and *c2* class samples of the k^{th} feature and $(\sigma_k^{c1})^2, (\sigma_k^{c2})^2$ are variances of the *c1* and *c2* class samples of the k^{th} feature

The numerator of the Eq. 1 shows the discriminating power between the classes and the denominator reveals that within the individual classes. The larger is the F-score, the more likely the feature is significant. In order to select the efficient features from entire dataset, a threshold value is employed on the F-scores of all features. If the F-score value of any feature is bigger than threshold value, that feature is added to feature space. Otherwise, that feature is removed from feature space.

C. AR model based feature extraction method

The autoregressive (AR) model is a popular linear model used for modeling of time series data generated by a stochastic process in many applications such as speech processing, image processing and pattern recognition [15]. It

is a simple and robust method and requires no a priori knowledge of the sequence to be analyzed and also works well with a low signal–to-noise ratio. The parameters of the AR model comprise significant information of the system condition and can reflect the characteristics of a dynamic system. The auto regressive model can be viewed as a linear prediction filter. The coefficients of the linear filter can be used to model the microarray samples in gene space in terms of their global spectral characteristics.

In AR modeling the observed signal x(n) can be modeled as a linear combination of its 'p' past values x(n-k), defined as

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) \tag{2}$$

where a_k represents the coefficient of the model to be estimated. The coefficients can be estimated by the Yule-Walker method in least mean square sense which gives linear equations defined as

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i),\ 1 \le i \le p \tag{3}$$

This can be represented in matrix form as

$$\begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ R_2 & R_2 & R_0 & \cdots & R_{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_p \end{bmatrix} \tag{4}$$

Where R_p is the autocorrelation of the observed signal and a_p is the model order or parameters of the model. The parameters can be computed by the Levinson Durbin's recursive process [15].

III. Radial Basis Function Neural Network Classifier

The radial basis function network (RBFNN) is a kind of well studied neural network structure, suitable for function approximation and pattern classification problems because of their simple topological structure and their ability to learn in an explicit manner. In the classical RBF network, there is an input layer, a hidden layer consisting of nonlinear node function, an output layer and a set of weights to connect the hidden layer and output layer. Due to its simple structure it reduces the computational task as compared to conventional multi layer perceptron (MLP) network. In RBFNN the basis functions are usually chosen as Gaussian and the number of hidden units are fixed a priori using some properties of input data [17]. The structure of a RBF network is shown in Fig. 1.

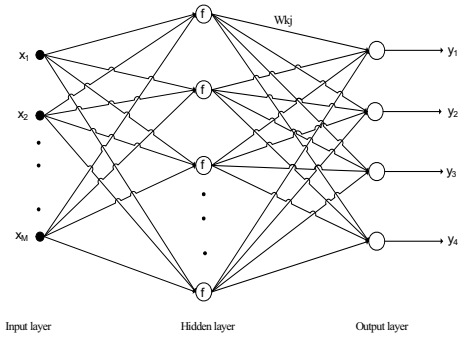

Fig. 1 The structure of the RBFNN

For an input feature vector x, the output of the jth output node is given as

$$y_j = \sum_{k=1}^{N} w_{kj}\varphi_k = \sum_{k=1}^{N} w_{kj} e^{-\frac{\|x(n)-C_k\|}{2\sigma_k^2}} \tag{5}$$

The error occurs in the learning process is reduced by updating the three parameters, the positions of centers (C_k), the width of the Gaussian function (σ_k) and the connecting weights (w) of RBFNN by a stochastic gradient approach as defined below:

$$w(n+1) = w(n) - \mu_w \frac{\partial}{\partial w} J(n) \tag{6}$$

$$C_k(n+1) = C_k(n) - \mu_c \frac{\partial}{\partial C_k} J(n) \tag{7}$$

$$\sigma_k(n+1) = \sigma_k(n) - \mu_\sigma \frac{\partial}{\partial \sigma_k} J(n) \tag{8}$$

where $J(n) = \frac{1}{2}|e(n)|^2$, e (n)=d(n) - y(n) is the error, d(n) is the target output and y(n) is the predicted output. μ_w μ_C and μ_σ are the learning parameters of the RBF network. The complete process of the proposed feature extraction based classification process is presented in Fig. 2.

IV. RESULT AND DISCUSSION

In order to compare the potentiality of the proposed method in predicting the class of the cancer microarray data, six standard datasets such as Leukemia, SRBCT, MLL-Leukemia, Colon, Prostate and Lymphoma have been used for the study. All the datasets are categorized into two

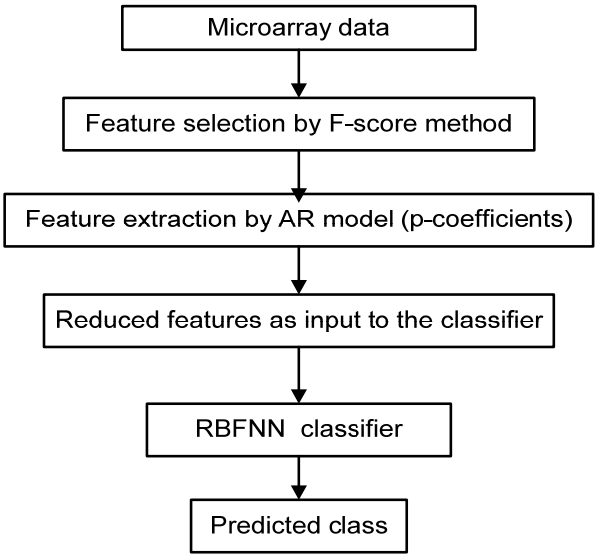

Fig. 2 The flow graph of the proposed feature based classification scheme

groups: binary and multi class. The Leukemia, Prostate and Colon dataset are binary class and SRBCT, lymphoma and MLL-Leukemia are multi class datasets. The proposed feature extraction process has two steps. First, the F-score method is employed on the gene space to choose the discriminant feature set. For example, the F-score values of the genes in Leukemia dataset is shown in Fig. 3

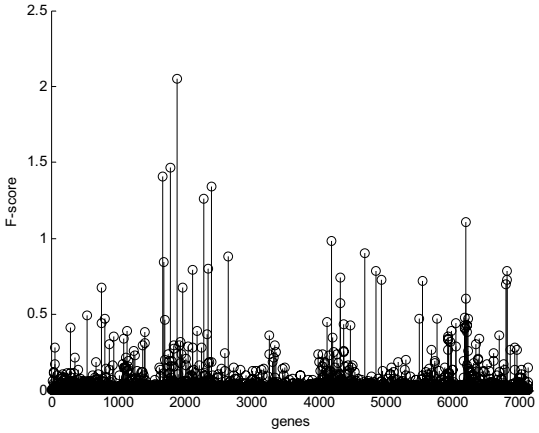

Fig. 3 The F-score values versus genes of Leukemia dataset

The average of the F-score is used as the threshold to select the discriminant genes. Then, the reduced feature set is modeled by the autoregressive modeling to capture the global characteristics of the samples. The parameters of the model contains the information regarding the classification of samples, hence constitutes the optimal features for class prediction. Through an empirical study, the model parameters are chosen 50 to achieve better accuracy. To evaluate the classification performance of the proposed hybrid feature vector, a leave one out cross validation (LOOCV) or Jackknife test is employed. The LOOCV is deemed the most objective and has been widely recognized and increasingly used by investigators to examine the accuracy of various predictors. A radial basis function neural network is employed to assess the performance of the proposed feature extraction method.

Table 1 Comparison of LOOCV predictive accuracy (%) of the proposed feature extraction method with three classifiers

Dataset	Method	LOOCV Accuracy
Leukemia	RBFNN	97.22
	MLP	97.22
	LDA	76.39
Colon	RBFNN	93.55
	MLP	90.32
	LDA	64.52
Prostate	RBFNN	93.13
	MLP	88.24
	LDA	71.57
MLL-Leukemia	RBFNN	93.02
	MLP	88.89
	LDA	73.61
Lymphoma	RBFNN	96.77
	MLP	93.55
	LDA	80.65
SRBCT	RBFNN	93.98
	MLP	84.34
	LDA	63.86

To have a comparative performance study, the proposed feature representation method is also analyzed with the well studied neural network classifiers, the multilayer perceptron and a statistical method, the linear discriminant analysis (LDA). The success rates of all the classifiers are evaluated with all the six benchmark datasets and are listed in Table 1 and also presented in Figs. 4 and 5. From the figures it is evident that the RBFNN network performs superior than the MLP and the LDA. Hence the proposed feature extraction scheme with radial basis function network can be used as an

efficient method to classify the microarray data samples with an advantage of reduced computational load.

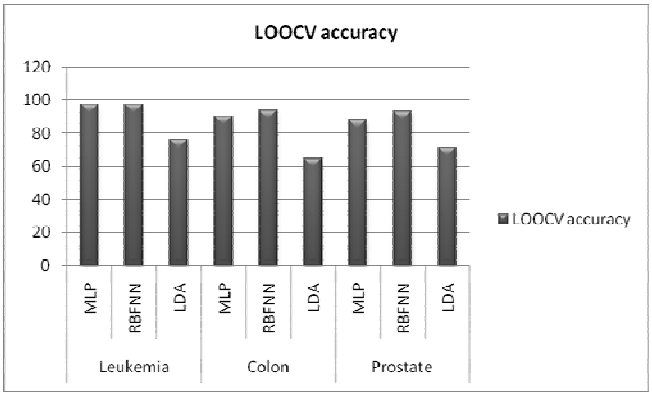

Fig. 4 The LOOCV accuracy for the binary class datasets

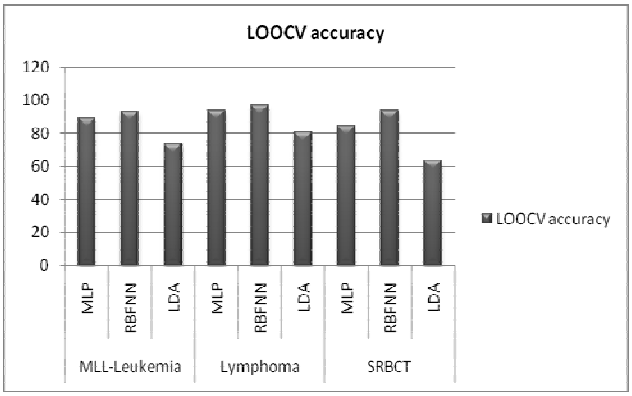

Fig. 5 The LOOCV accuracy for the multi class datasets

V. CONCLUSION

In this paper an efficient hybrid feature extraction method is presented by embedding the F-score statistics and the AR model. The proposed hybrid method effectively reduces the dimension of the samples in capturing the features relevant to classes. The results of the LOOCV test using the standard datasets shows the potential of the proposed method with the advantage of reduced computational complexity. Hence it can be used as an efficient approach for class prediction of microarray samples.

REFERENCES

[1] M. Xiong, L. Jin, W. Li, and E. Boerwinkle, "Computational methods for gene expression-based tumor classification," BioTechniques, vol. 29, no. 6, pp. 1264–1268, 2000.

[2] P. Baldi and A.D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," Bioinformatics, vol. 17, no. 6, pp. 509–519, 2001.

[3] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S.," Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", Science, 286(5439), pp.531-537, 1999

[4] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," Bioinformatics, vol. 22, no. 15, pp. 1855–1862, 2006.

[5] K.Y. Yeung, W.L. Ruzzo, "Principal component analysis for clustering gene expression data", Bioinformatics, 17, pp.763–774, 2002.

[6] Yihui Liu," wavelet feature extraction for high-dimensional microarray data," Neurocomputing, Vol. 72, pp. 985-990, 2009

[7] Yihui Liu," Detect Key Gene Information in Classification of Microarray Data", EURASIP Journal on Advances in Signal Processing, pp.1-10.2007

[8] Tan AC, Gilbert D," Ensemble machine learning on gene expression data for cancer classification", Applied Bioinformatics, 2, pp.75-83, 2003.

[9] M. Dettling, "Bag Boosting for tumor classification with gene expression data," Bioinformatics, vol. 20, no. 18, pp. 3583–3593, 2004.

[10] Guyon I, Weston J, Barnhill and Vapnik V," Gene selection for cancer classification using support vector machines", Mach. Learn, 46, pp. 389- 422, 2002

[11] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., Meltzer, P. S., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Medicine, 7(6), pp.673-679, 2001

[12] O'Neill MC and Song L.,"Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect", BMC Bioinformatics, 4:13, 2003

[13] Bing Liu, Qinghua Cui, Tianzi Jiang and Songde Ma," A combinational feature selection and ensemble neural network method for classification of gene expression data", BMC Bioinformatics, 5:136, pp. 1-12, 2004

[14] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A.V. DAmico and J.P. Richie, "Gene expression correlates of clinical prostate cancer behavior", Cancer Cell,vol. 1, pp. 203-209, 2002.

[15] J. Makhoul," Linear prediction : A Tutorial review", Proceedings of the IEEE,vol. 63, pp. 562-580, 1975.

[16] Salih Günes, Kemal Polat , S_ebnem Yosunkaya ," Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome", Expert systems with applications, 37, pp. 998-1004,2010

[17] Samantray, S.R., Dash, P.K., Panda G." Fault classification and location using HS-transform and radial basis function neural network", Electric Power Systems Research, 76 (9-10), pp.897-905, 2006

18. Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., de Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., Korsmeyer, S. J.," MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia", Nature Genetics , 30(1),pp. 41-47, 2002.

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mach andA. J. Levine, " Broad patterns of gene expression revealed by clusteringanalysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proc Natl Acad Sci., vol. 96, pp. 6745-6750, 1999.