A Fuzzy-Neural Approach for Leukemia Cancer Classification

Engineering

KEYWORDS : classification, fuzzy hypersphere neural network, microarray, gene

Dr.B.B.M.Krishna Kanth Principal, Hindu College of Engineering and Technology, Guntur, Andhra Pradesh, India.

ABSTRACT

The classification of cancers subtypes is essential for future clinical accomplishments of microarray based cancer diagnosis. In this paper, we use the Fuzzy Hyper sphere neural network (FHSNN) classifier for the discrimination of acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) subtypes present in the leukemia dataset. Prior to classification as the number of genes are larger in number compared to the samples available in the microarray datasets hence, to find the best features(genes) for classification dimensionality reduction methods such as Signal-to-Noise Ratio, Class-Separability, Wilcoxon rank sum statistic and Fisher Ratio are used. The experimental results show that our FHSNN is able to achieve 100% accuracy with much fewer genes than the previously published methods did. In particular, amongst various systematic experiments carried out, the best classification model is achieved using a subset of features chosen by Wilcoxon rank sum statistic gene selection method. Furthermore our FHSNN is found to be much faster with respect to training and testing time.

INTRODUCTION

DNA microarray is a technology that can assess the expression levels of thousands of genes in a single test. It is commonly used for comparing the gene expression levels in tissues under different conditions, such as healthy versus diseased. Some of the genes are expected to be differentially modulated in tissues under different conditions, with their expression levels increased or decreased to signify the experimental conditions. These discriminatory genes are very useful in clinical applications such as recognizing diseased profiles. However, due to high cost, the number of experiments that can be used for classification purpose is usually limited. This small number of experiments, compared to the large number of genes in an experiment, wakes up “the curse of dimensionality” and challenges the classification task and other data analysis in general. It is well-known that quite a number of genes are house-keeping genes and many others could be unrelated to the classification task [2]. Therefore, an important step to effective classification is to identify the discriminatory genes thus to reduce the number of genes used for classification purpose. This step of discriminatory gene identification is generally referred to as gene selection. Gene selection is a pre-requisite before classification. It should be noted that, often, the number of unrelated genes is much larger than the number of discriminatory genes.

At first Golub et al [1] employed a correlation metric to extract a small set of genes and developed a scheme named weighted voting to distinguish ALL and AML; the recognition rate they obtained was 94.1%. By using the same database, several algorithms have been proposed to deal with class prediction of acute leukemia to improve classification accuracy in the literature [2-10]. Toure et al [3] used the multilayer perceptron network (MLP) to predict the class of cancer and gave 58% accuracy on test data. Ryu et al [4] experimented with MLP, support vector machine (SVM)[12,17], and k-nearest neighbor(KNN)[11] as the classifiers, and the best classification rate they achieved was 97.1% if the gene is selected via Pearson’s correlation analysis and the MLP is used as classifier. Su et al [5] employes modular neural networks to classify two types of acute leukemia’s and the best 75% correct classification was reached. Xu et al [6] adopted the ellipsoid ARTMAP to analyze the ALL/AML data set [1] and the best result was 97.1%. Some studies [19] have shown that a small collection of genes selected correctly can lead to good classification results [18]. Therefore gene selection is crucial in molecular classification of cancer. Although most of the algorithms mentioned above can reach high prediction rate, any misclassification of the disease is still intolerable in acute leukemia’s treatment. Therefore the demand of a reliable classifier which gives 100% accuracy in predicting the type of cancer therewith becomes urgent.

In this paper, we apply robust FHSNN [15] classifier and four effective dimensionality reduction methods, i.e., Signal-to-Noise Ratio, Class-separability measure, Wilcoxon rank sum statistic and Fisher Ratio, to the problem of cancer classification based on gene expression data. The paper is organized as follows. Section 2 contains a brief description of various gene selection methods used to select the top ranked genes. The FHSNN classifier is described in Section 3. Experimental results with the leukemia dataset are presented in Section 4, followed by conclusions in Section 5.

2 Gene Selection Methods

The Microarray datasets usually contain large number of genes, but among them only a small portion of genes may help for the correct classification of cancers. The rest of the genes have little effect on the classification. Even worse, some genes may act as “noise” and undermine the classification accuracy. Hence, to obtain good classification accuracy, we need to pick out the genes that benefit the classification most. In addition, gene selection is also a procedure of input dimension reduction, which leads to a much less computation load to the classifier. Maybe more importantly, reducing the number of genes used for classification can help researchers put more attention on these important genes and find the relationship between the genes and the development of the cancer.

2.1 Signal-to-Noise Ratio

$$\mathrm{SNR}(g)=\frac{\mu_1(g)-\mu_2(g)}{\sigma_1(g)+\sigma_2(g)} \tag{1}$$

In SNR [17] the variables $\mu_1(g)$ and $\mu_2(g)$ are the means of the expression levels of gene g for the two classes ALL and AML respectively, and $\sigma_1(g)$ and $\sigma_2(g)$ are the standard deviations of gene g in classes ALL and AML respectively.

2.2 Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test [14] organizes the observed data in value ascending order. Each data item is assigned a rank corresponding to its place in the sorted list. These ranks, rather than the original observed values, are then used in the subsequent analysis. The major steps in applying the Wilcoxon rank-sum test are as follows:

(1) Merge all observations from the two classes and rank them in value ascending order.
(2) Calculate the Wilcoxon statistics by adding all the ranks associated with the observations from the class with a smaller number of observations.

2.3 Fisher-Ratio

Fisher ratio [13] is a ratio of between-class distances to with-in class distances. If there are two classes in a data set, the Fisher Ratio (FR) for gene g is:

$$\mathrm{FR}(g)=\frac{\left(\mu_1(g)-\mu_2(g)\right)^2}{\sigma_1(g)^2+\sigma_2(g)^2} \tag{2}$$

Gene with highest FR value is most informative and the expression levels differ most on average in the two classes while also favoring those with small deviation in the respective classes.

Then the genes with high FR values are selected as the top features.

2.4 Class- Separability

Most popular method for gene selection is to measure the class-separability (CS)[16]. CS of gene *i* is defined as:

$$\text{CS} \quad \frac{\sum_i \sum_k \mathrm{I}(\mathrm{y}_i = k)\left(\bar{X}_{kj} - \bar{X}_{.j}\right)^2}{\sum_i \sum_k \mathrm{I}(\mathrm{y}_i = k)\left(X_{ij} - \bar{X}_{kj}\right)^2} \qquad (3)$$

where $\bar{X}_{.j}$ and $\bar{X}_{kj}$ denote the average expression level of gene *j* across all tumor samples and across samples belonging to class *k* only and X_{ij} is the gene expression of the *i*th gene.

3 Fuzzy HyperSphere Neural Network Classifier

The FHSNN consists of four layers as shown in Figure 1(a). The first, second, third and fourth layer is denoted as F_R , F_M , F_N and F_O respectively. The F_R layer accepts an input pattern and consists of *n* processing elements, one for each dimension of the pattern. The F_M layer consists of *q* processing nodes that are constructed during training and each node represents hypersphere fuzzy set characterized by hypersphere membership function. The weights between F_R and F_M layer represent centre points of the hyperspheres as shown in Figure 1(b). $C_j = (c_{j1}, c_{j2}, c_{j3} \ldots\ldots c_{jn})$ represents center point of the hypersphere m_j. In addition to this each hypersphere takes one more input denoted as threshold *T*, which is set to one and the weight assigned to this link is ξ_j. ξ_j represents radius of the hypersphere m_j, which is updated during training. The center points and radii of the hyperspheres are stored in matrix *C* and vector ξ respectively. The maximum size of hypersphere is bounded by a user defined value λ, where $0 \le \lambda \le 1$. λ is called as growth parameter that is used for controlling maximum size of the hypersphere and it puts maximum limit on the radius of the hypersphere. Assuming the training set defined as

$R \in \{R_h | h = 1, 2, \ldots.. P\}$, where

$R_h = (r_{h1}, r_{h2}, r_{h3} \ldots.. r_{hn}) \in I^n$ is the h_{th} pattern, the membership function of the hypersphere node m_j is defined as

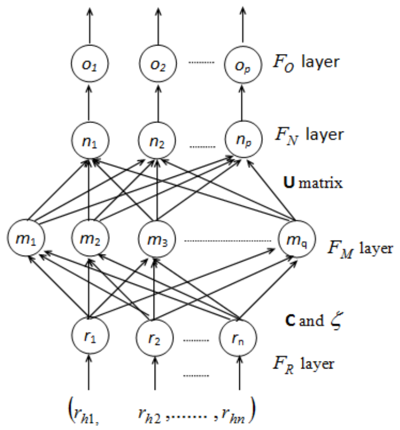
$$m_j\left(R_h, C_j, \zeta_j\right) = 1 - f\left(l, \zeta_j, \gamma\right) \qquad (5)$$

where *f* ( ) is three-parameter ramp threshold function defined as

$$f\left(l, \zeta_j, \lambda\right) = \begin{bmatrix} 0, \text{ if } (0 \le l \le \zeta_j) \\ l\gamma, \text{ if } (\zeta_j \le l \le 1) \\ 1, \text{ if } (l \ge 1) \end{bmatrix} \qquad (6)$$

and the argument l is defined as,

$$l = \left(\sum_{i=1}^{n} \left(c_{ji} - r_{hi}\right)^2\right)^{1/2} \qquad (7)$$



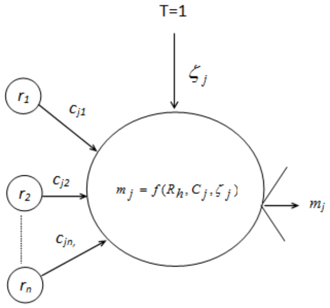


Figure 1 The proposed FHSNN, (a) Topology of FHSNN, (b) Implementation of fuzzy hypersphere

4 Experimental Results

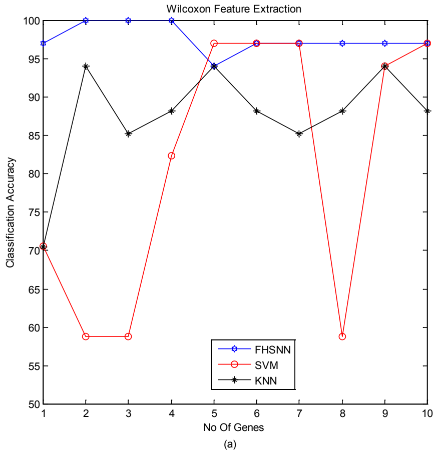
Dataset that we have used is a collection of expression measurements reported by Golub. Gene expression profiles have been constructed from 72 people who have either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML).

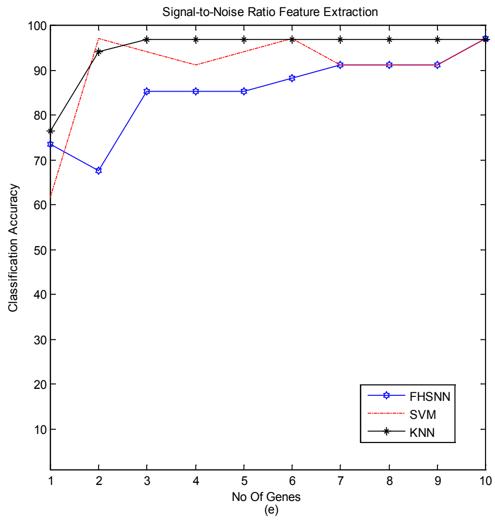
Each person has submitted one sample of DNA microarray, so that the database consists of 72 samples. Each sample is composed of 7129 gene expressions, and finally the whole database is a 7129 X 72 matrix. Before the classification, we need to find out informative genes (features) that are related to predict the cancer class out of 7129.In order to do this, each gene is scored based on the equations given by different feature (gene) selection methods. It is found that genes with Gene id's 4847, 3320 and 1745 appeared common in all the feature selection methods which indicate that they are very informative features for the accurate prediction of cancer.

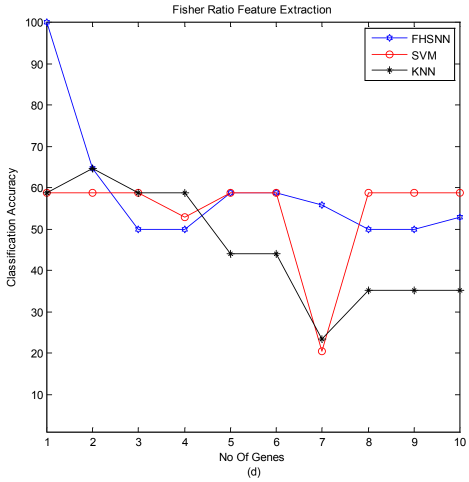
The high accuracy was obtained by using only two genes with gene id 4847 and 1882 which are selected by using Wilcoxon rank sum test gene selection method. Traditional classifiers such as Support vector machine and K-nearest neighbor produced the best accuracy of 97.1% using all top 10 genes. As shown from Table 1 the average training time and testing time of FHSNN classifier with ALL/AML dataset is in the range of 0.20 -0.39 seconds which is very fast compared to any other classifier published so far. Meanwhile the average training and testing time of SVM and KNN classifiers is around 2.60-3.5 seconds respectively which is very slow comparative to FHSNN classifier.

Classifier	Average Training time(seconds)	Average Testing time(seconds)
FHSNN	0.25	0.35
KNN	2.60	2.65
SVM	3.20	3.50

Table 1 Comparison of training and testing time for the three classifiers







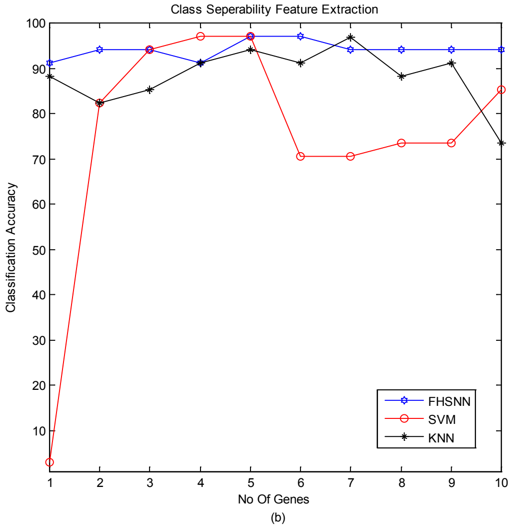


Figure 2 Comparison of classification accuracy among FHSNN ,SVM(linear kernel function), KNN(k= 5 neighbors) classifiers using the top 10 genes by (a) Wilcoxon Rank Sum Test (b) Signal to Noise Ratio (c) Class separability (d) Fisher Ratio

5 Conclusion

In order to predict the class of cancer, we have demonstrated the usefulness of the FHSNN classifier using an informative genes extraction methods based on similarity measures and statistical analysis. Experimental results show that the FHSNN classifier is the most effective in classifying the type of leukemia cancer using only two of the most informative genes. The FHSNN is found to be superior compared to SVM and KNN classifiers with respect to prediction accuracy, training time and testing time respectively. FHSNN yields 100% recognition accuracy and is well suited for the ALL/AML classification in cancer treatment. Furthermore, our FHSNN is faster than any other classifiers published so far. The training and testing time is in the range of 0.2 to 0.39 seconds therefore our FHSNN classifier can be easily implemented in hardware for the diagnosis of cancer. In future work, a more sophisticated membership function for FHSNN classifier will be designed for improving accurate diversity. Moreover the parameters used in FHSNN can be changed accordingly and a new distance measure for gene selection should be developed.

REFERENCE

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” Science, vol. 286, pp. 531–537, 1999. | [2] D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub and E. S.Lander, “Class prediction and discovery using gene expression data,” Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, pp. 263–272, 2000. | [3] A. Toure and M. Basu, “Application of Neural Network to Gene Expression Data for Cancer Classification,” Proceedings of the 2001 International Joint Conference on Neural Networks, vol. 1, pp. 583–587, 2001. | [4] J. Ryu and S.-B. Cho, “Gene Expression Classification Using Optimal Feature/Classifier Ensemble with Negative Correlation,” Proceedings of the 2002 International Joint Conference on Neural Networks, vol. 1, pp. 198–203, 2002. | [5] Min Su, M. Basu and A. Toure, “Multi-Domain Gating Network for Classification of Cancer Cells Using Gene Expression Data,” Proceedings of the 2002 International Joint Conference on Neural Networks, vol. 1, pp. 286–289, 2002. | [6] R, Xu, G. Anagnostopoulos and D. Wunsch, “Tissue Classification Through Analysis of Gene Expression Data Using A New Family of ART Architectures,” Proceedings of the 2002 International Joint Conference on Neural Networks,vol. 1, pp. 300–304, 2002. | [7] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer and Z. Yakhini, “Tissue Classification with Gene Expression Profiles,” Proceedings of the Fourth AnnualInternational Conference on Computational Molecular Biology, pp. 54–64, 2000. | [8] M. Kuramochi and G. Karypis, “Classification Using Expression Profiles: A Feasibility Study,” Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering, pp. 191–200, 2001. | [9] F. Azuaje, “Gene Expression Patterns and Cancer Classification: A Self-Adaptive and Incremental Neural Approach,” Proceedings of the 2000 IEEE EMBS International Conference on Information Technology Applications in Biomedicine, pp. 308–313, 2000. | [10] F. Azuaje, “Making Genome Expression Data Meaningful: Prediction and Discovery of Classes of Cancer Through a Connectionist Learning Approach,” Proceedings of IEEE International Symposium on BioInformatics and Biomedical Engineering, pp. 208–213, 2000. | [11] M.B. Eisen, B.O. Brown, DNA arrays for analysis of gene expression, Methods Enzymol. 303 (1999) 179–205. | [12] S.-B. Cho, Exploring features and classifiers to classify gene expression profiles of acute leukemia, Int. J. Pattern Recogn. Artif. Intell. 16 (7) (2002) 1–13. | [13] K.Z. Mao, RBF neural network center selection based on Fisher ratio class separability measure, IEEE Trans. Neural Networks 13 (2002) 1211–1217. | [14] E.L. Lehmann. Non-parametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco, 1975. | [15] U V Kulkarni, T R Sontakke. ‘Fuzzy Hypersphere Neural Network Classifier.’ Electronics Letters, IEE, vol 85, May2004, p 23-28. | [16] S. Dudoit, J. Fridlyand and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” Technical Report 576, Department of Statistics, University of California, Berkeley, June 2000. | [17] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914. | [18]Wang X, Gotoh O. Microarray-Based Cancer Prediction Using Soft Computing Approach. Cancer Informatics. 2009; 7:123–39. | [19]Saeys Y, Iñza I, Larrañaga P: A review of feature selection techniques in bioinformatics. Bioinformatics 2007, 23(19):2507-2517. |