Gene Expression Based Acute Leukemia Cancer Classification: a Neuro-Fuzzy Approach

B. B. M. Krishna Kanth bbkkanth@yahoo.com
Research Scholar S.R.T.M.University
Nanded, Maharastra, India

U. V. Kulkarni kulkarniuv@yahoo.com
Dean of Academics and Head Department
of Computer Science S.R.T.M.University,
Nanded,Maharastra, India

B. G. V. Giridhar murarihamlet@rediffmail.com
Assistant Professor Department of Endocrinology
Andhra Medical College Visakhapatnam,
A.P, India

Abstract

In this paper, we proposed the Modified Fuzzy Hypersphere Neural Network (MFHSNN) for the discrimination of acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) in leukemia dataset. Dimensionality reduction methods, such as Spearman Correlation Coefficient and Wilcoxon Rank Sum Test are used for gene selection. The performance of the MFHSNN system is encouraging when benchmarked against those of Support vector machine (SVM) and the K-nearest neighbor (KNN) classifiers. A classification accuracy of 100% has been achieved using the MFHSNN classifier using only two genes. Furthermore, MFHSNN is found to be much faster with respect to training and testing time.

Keywords: gene expression data, cancer classification, AAL/AML, membership function, hypersphere

1. INTRODUCTION

Microarrays [1], also known as gene chips or DNA chips, provide a convenient way of obtaining gene expression levels for a large number of genes simultaneously. Each spot on a microarray chip contains the clone of a gene from a tissue sample. Some mRNA samples are labeled with two different kinds of dyes, for example, Cy5 (red) and Cy3 (blue). After mRNA interacts with the genes, i.e., hybridization, the color of each spot on the chip will change. The resulted image reflects the characteristics of the tissue at the molecular level. Microarrays can thus be used to help classify and predict different types of cancers. Traditional methods for diagnosis of cancers are mainly based on the morphological appearances of the cancers; however, sometimes it is extremely difficult to find clear distinctions between some types of cancers according to their appearances. Hence the microarray technology stands to provide a more quantitative means for cancer diagnosis. For example, gene expression data have been used to obtain good results in the classifications of Lymphoma, Leukemia [2], Breast cancer, and Liver cancer etc. It is challenging to use gene expression data for cancer classification because of the following two special as-

pects of gene expression data. First, gene expression data are usually very high dimensional. The dimensionality ranges from several thousands to over ten thousands. Second, gene expression data sets usually contain relatively small numbers of samples, e.g., a few tens. If we treat this pattern recognition problem with supervised machine learning approaches, we need to deal with the shortage of training samples and high dimensional input features.

Recent approaches to solve this problem include unsupervised methods, such as Clustering [3] and Self-Organizing Maps (SOM) [4] and supervised methods, such as Support Vector Machines (SVM)[5], Multi-Layer Perceptrons (MLP) [6], Decision Trees (DT) [7] and K-Nearest Neighbor(KNN) [8, 9]. Su et al [10] employs modular neural networks to classify two types of acute leukemia's and the best 75% correct classification was reached. Xu et al [11] adopted the ellipsoid ARTMAP to analyze the AAL/AML data set and the best result was 97.1%. But most of the current methods in microarray analysis can not completely bring out the hidden information in the data. Meanwhile, they are generally lacking robustness with respect to noisy and missing data. Some studies have shown that a small collection of genes [12] selected correctly can lead to good classification results [13]. 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 a robust MFHSNN classifier which is an extension of Fuzzy Hypersphere Neural Network (FHSNN) proposed by Kulkarni et al [14] to the problem of cancer classification based on gene expression data. To reduce the dimensionality of genes correlation method such as Spearman Correlation Coefficient and statistical method such as Wilcoxon Rank Sum Test are used. The MFHSNN utilizes fuzzy sets as pattern classes in which each fuzzy set is a union of fuzzy set hyperspheres. The fuzzy set hypersphere is an n-dimensional hypersphere defined by a center point and radius with its membership function. We first experiment the classifier with 38 leukemia samples and test the classifier with another 34 samples to obtain the accuracy rate. Meanwhile, this study reveals that the classification result is greatly affected by the correlativity with the class distinction in the data set. The remainder of the paper is organized as follows. The gene selection methods for choosing effective predictive genes in our work are introduced in Section 2. Then Sections 3 gives a brief introduction for the architecture of the MFSHNN, followed by its learning algorithm in section 4. Section 5 examines the experimental results of the classifiers operated on leukemia data set. Conclusions are made in Section 6.

2. GENE SELECTION METHODS

Among the large number of genes, only a small part may benefit the correct classification of cancers. The rest of the genes have little impact 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. Correlation Analysis for Gene Selection

In order to score the similarity of each gene, an ideal feature vector [15] is defined. It is a vector consisting of 0's in one class (ALL) and 1's in other class (AML). It is defined as follows:

$ideal_i$ = (0,0,0,0,0,0,1,1,1,1,1,1) (1)

The ideal feature vector is highly correlated to a class. If the genes are similar with the ideal vector (the distance from the ideal vector and the gene is small), we consider that the genes are in-

formative for classification. The similarity of g_i and g_{ideal} using similarity measure such as the Spearman coefficient is defined as follows

$$SC= 1-\frac{6\sum_{i=1}^{n}\left(\text{ideal}_i - g_i\right)^2}{n\times\left(n^2-1\right)} \quad (2)$$

Where *n* is the number of samples; g_i is the i_{th} real value of the gene vector and ideal_i is the corresponding i_{th} binary value of the ideal feature vector.

2.2. Wilcoxon Rank-Sum Test (WRST) for Gene Selection

The Wilcoxon rank-sum test [16, 17] is a big category of non-parametric tests. The general idea is that, instead of using the original observed data, we can list the data in the value ascending order, and assign each data item a rank, which is the place of the item in the sorted list. Then, the ranks are used in the analysis. Using the ranks instead of the original observed data makes the rank sum test much less sensitive to outliers and noises than the classical (parametric) tests [18]. The WRST 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 WRST are as follows:

(i) Merge all observations from the two classes and rank them in value ascending order.

(ii) Calculate the Wilcoxon statistics by adding all the ranks associated with the observations from the class with a smaller number of observations.

3. MODIFIED FUZZY HYPERSPHERE NEURAL NETWORK CLASSIFIER

The MFHSNN 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 processing performed by each node of F_M layer is shown in Figure 1(b). The weights between F_R and F_M layer represent centre points of the hyperspheres. As shown in Figure 1(b), $C_j = \left(c_{j1}, c_{j2}, c_{j3}.....c_{jn},\right)$ 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. The 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 $0 \le \lambda \le 1$. The 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 $R \in \left\{R_h \middle| h = 1,2,.....P\right\}$, where $R_h = \left(r_{h1}, r_{h2}, r_{h3}.....r_{hn}\right) \in I^n$ is the h_{th} pattern the, membership function of the hypersphere node m_j is $m_j\left(R_h, C_j, \zeta_j\right) = 1 - f\left(l, \zeta_j, \gamma\right)$ (3)

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

$$f\left(l, \zeta_j, \gamma\right) = \begin{bmatrix} 0, \ if \quad (0 \le l \le \zeta_j) \\ (l - \zeta_j)\gamma, \ if \quad (\zeta_j \le l \le 1) \\ 1, \ if \quad (l \ge 1) \end{bmatrix} \quad (4)$$

and the argument l is defined as, $$l = \left(\sum_{i=1}^{n} \left(c_{ji} - r_{hi} \right)^2 \right)^{1/2} \quad (5)$$

The membership function returns m_j =1, if the input pattern R_h is contained by the hypersphere. The parameter γ, $0 \le \gamma \le 1$, is a sensitivity parameter, which governs how fast the membership value decreases R_h outside the hypersphere when the distance between R_h and C_j increases.

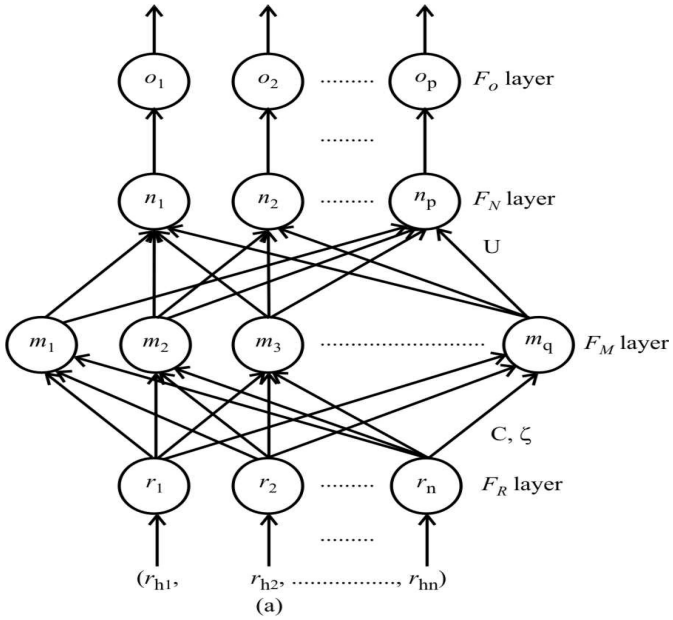

FIGURE 1: (a) Modified Fuzzy Hypersphere Neural Network **(b)** Implementation of Fuzzy Hypersphere

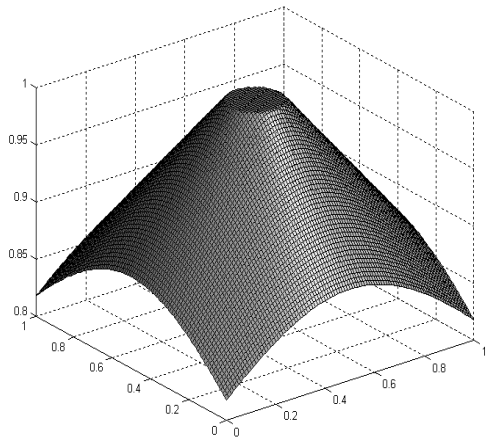

FIGURE 2: Plot of Modified Fuzzy Hypersphere Membership Function for γ = 1

The sample plot of membership function for MFHSNN with centre point [0.5 0.5] and radius equal to 0.3 is shown in Figure 2. It can be observed that the membership values decrease steadily with increasing distance from the hypersphere.

Each node of F_N and F_O layer represents a class. The F_N layer gives fuzzy decision and output of k_{th} F_N node represents the degree to which the input pattern belongs to the class n_k .The weights assigned to the connections between F_M and F_N layers are binary values that are stored in matrix U and updated during learning as

$$u_{jk} = \begin{cases} 1 \text{ if } m_j \text{ is a hypershere of class } n_k \\ 0 \text{ otherwise} \end{cases} \quad (6)$$

For $k = 1,2,3,..p$ and $j = 1,2,3,...q$

where m_j is the j_{th} F_M node and n_k is the k_{th} F_N node. Each F_N node performs the union of fuzzy values returned by the fuzzy set hyperspheres of same class, which is described by equation (7).

$$n_k = \max_{j=1}^{q} m_j u_{jk} \text{ for } k = 1,2,.....p \quad (7)$$

Each F_O node delivers non-fuzzy output, which is described by equation (8).

$$o_k = \begin{cases} 0 \text{ if } n_k \le T \\ 1 \text{ if } n_k = T \end{cases} \text{ for } k = 1,2,3,....p \quad (8)$$

Where $T = \max(n_k)$ for $k = 1,2,3,....p$

4. MFHSNN Learning Algorithm

The supervised MFHSNN learning algorithm for creating fuzzy hyperspheres in hyperspace consists of three steps

1. Creation of hyperspheres
2. Overlap test, and
3. Removing overlap.

These three steps are described below in detail.

4.1 Creation of Hyperspheres

Given the h_{th} training pair (R_h, d_h) find all the hyperspheres belonging to the class d_h. These hyperspheres are arranged in ascending order according to the distances between the input pattern and the center point of the hyperspheres. After this following steps are carried sequentially for possible inclusion of input pattern R_h.

Step 1: Determine whether the pattern R_h is contained by any one of the hyperspheres. This can be verified by using modified fuzzy hypersphere membership function defined in equation (3). If R_h is contained by any of the hypersphere then it is included, therefore in the training process all the remaining steps are skipped and training is continued with the next training pair.

Step 2: If the pattern R_h falls outside the hypersphere, then the hypersphere is expanded to include the pattern if the expansion criterion is satisfied. For the hypersphere m_j to include R_h the following constraint must be met defined as:

$$\left(\sum_{i=1}^{n} \left(c_{ji} - r_{hi} \right)^2 \right)^{1/2} \leq \lambda \qquad (9)$$

If the expansion criterion is met then the pattern R_h is included as

$$\zeta_j = \left(\sum_{i=1}^{n} \left(c_{ji} - r_{hi} \right)^2 \right)^{1/2} \qquad (10)$$

Step 3: If the pattern R_h is not included by any of the above steps then new hypersphere is created for that class, which is described as

$$C_{new} = R_h \text{ and } \zeta_{new} = 0 \qquad (11)$$

4.2 Overlap Test

The learning algorithm allows overlap of hyperspheres from the same class and eliminates the overlap between hyperspheres from different classes. Therefore, it is necessary to eliminate overlap between the hyperspheres that represent different classes. Overlap test is performed as soon as the hypersphere is expanded by step 2 or created in step 3.

(a)Overlap test for step 2: Let the hypersphere m_u is expanded to include the input pattern R_h and expansion has created overlap with the hypersphere m_v, which belongs to other class. Suppose $C_u = (x_1, x_2 \ldots\ldots x_n)$ and ζ_u represents center point and radius of the expanded hypersphere and $C_v = \left[x_1^{'}, x_2^{'} \ldots\ldots x_n^{'} \right]$ and ζ_v, are centre point and radius of the hypersphere of other class as depicted in Figure 3(a). Then if

$$\left(\sum_{i=1}^{n} \left(c_{ui} - c_{vi} \right)^2 \right)^{1/2} \leq \zeta_u + \zeta_v \qquad (12)$$

means those hyperspheres from separate classes are overlapping.

(b) Overlap test for step 3: If the created hypersphere falls inside the hypersphere of other class means there is an overlap. Suppose m_p represents created hypersphere to include the input pattern R_h and m_q represents the hypersphere of other class as shown in Figure 4(a). The presence of overlap in this case can be verified using the membership function defined in the equation

(3). If $m_p\left(R_h, C_p, \zeta_p\right) = m_q\left(R_h, C_q, \zeta_q\right) = 1$ means two hyperspheres from different classes are overlapping.

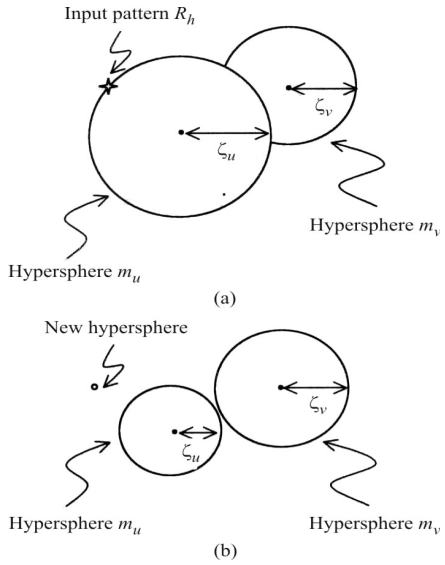

FIGURE 3: (a) Status of the hyperspheres before removing an overlap in step 2. **(b)** Status of the hyperspheres after removing an overlap in step 2

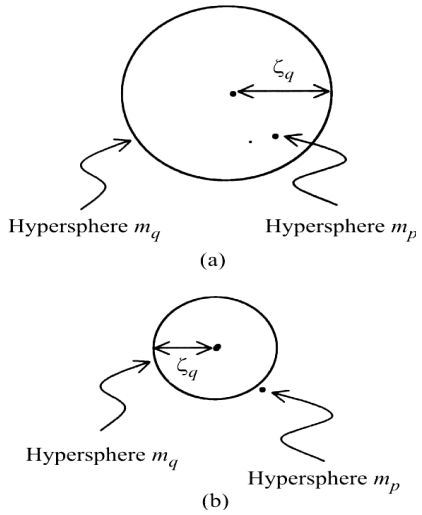

FIGURE 4: (a) Status of the hyperspheres before removing an overlap in step 3. **(b)** Status of the hyperspheres after removing an overlap in step 3

4.3 Removing Overlap

If step 2 has created overlap of hyperspheres from separate classes then overlap is removed by restoring the radius of just expanded hypersphere. Let, m_u be the expanded hypersphere then it is contracted as $\zeta_u^{new} = \zeta_u^{old}$ (13)

and new hypersphere is created for the input pattern as described by equation (11). This situation is shown in Figure 3(b). If the step 3 creates overlap then it is removed by modifying the hypersphere of other class. Let $C_p = \left(x_1, x_2 \ldots\ldots x_n\right)$ and ζ_p represents centre point and radius of the

created hypersphere, $C_q = \left[x_1^{'}, x_2^{'} \ldots\ldots x_n^{'} \right]$ and ζ_q are center point and radius of the hypersphere of other class. Then overlap is removed as

$$\zeta_q^{new} = \left(\sum_{i=1}^{n} \left(c_{pi} - c_{qi} \right)^2 \right)^{1/2} - \partial \qquad (14)$$

where ∂ is a small number selected just enough to remove the overlap. In our experiments the value of ∂ chosen is 0.0001. Hence, the hypersphere m_q is contracted just enough to remove the overlap as shown in Figure 4(b).

5. EXPERIMENTAL RESULTS

Dataset that we have used is a collection of expression measurements reported by Golub et al [2]. 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. The number of training samples in AAL/AML dataset is 38 which of them contain 27 samples of AAL class and 11 samples of AML class; here we randomly applied the training samples to the MFSHNN classifier. The number of testing samples is 34 where 20 samples belong to AAL and remaining 14 samples belongs to AML class respectively. This well-known dataset often serves as bench mark for microarray analysis methods. Before the classification, we need to find out informative genes (features) that are related to predict the cancer class out of 7129.

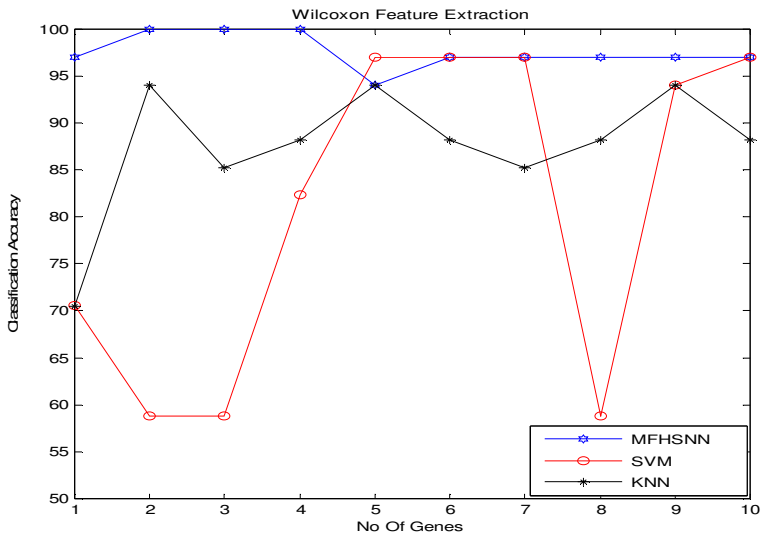

FIGURE 5: Comparison of classification accuracy among SVM, KNN(k= 5 neighbors) and MFHSNN classifiers with all the top 10 genes of Leukemia test data set selected by using Wilcoxon Rank Sum Test .

.

Figures 5 and 6 shows the comparison of the classification performance with respect to the features and the classifiers. Spearman correlation coefficient and Wilcoxon rank sum test gene selection techniques achieved 100% prediction accuracy on the test data set using MFHSNN classifier. It should also be noted that this high classification accuracy has been obtained using only two genes with Gene id's 4847 and 1882 which are selected by using Spearman correlation and Wilcoxon rank sum test gene selection methods.

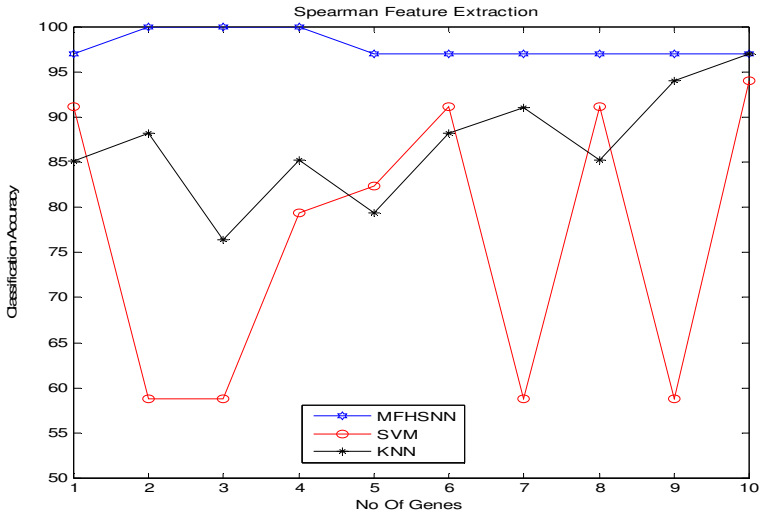

FIGURE 6: Comparison of classification accuracy among SVM, KNN (k= 5 neighbors) and MFHSNN classifiers of Leukemia test data set by using Spearman Correlation Coefficient**.**

But traditional classifiers such as Support vector machine and K-nearest neighbor produced the best accuracy of 97.1% using all the top 10 genes. As shown from Table 1 the average training time and testing time of MFHSNN classifier is in the range of 0.25 -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 MFHSNN classifier.

Classifier	Average Training time (seconds)	Average Testing time (seconds)
MFHSNN	0.25	0.39
KNN	2.60	2.65
SVM	3.20	3.50

TABLE 1: Comparison of training and testing time for the classifiers

The average classification accuracy of the three classifiers with all the 10 genes is shown in Table 2. The highest average classification accuracy achieved by MFHSNN is 97.94% which clearly dominates the other classifiers.

Gene selection\Classifier	MFHSNN	KNN	SVM
Wilcoxon Rank Sum Test	97.647	87.633	81.176
Spearman Coefficient	97.941	87.045	76.471

TABLE 2: Average classification accuracy

Gene Rank	Spearman Correlation Coefficient	Wilcoxon Rank Sum Test
1	4847	4847
2	1882	1882
3	3320	3320
4	6218	6218
5	1834	760
6	760	1834
7	2020	1745
8	5039	2020
9	1745	4499
10	4499	5039

TABLE 3: List of top 10 ranked genes (values are the Gene ids in the columns)

Table 3 shows the list of top 10 ranked genes that are chosen as the features of the input patterns to the classifiers. It is found that these top 10 genes selected by the gene selection methods are very informative features for the accurate prediction of cancer.

6. CONCLUSIONS

In order to predict the class of cancer, we have demonstrated the effectiveness of the MFHSNN classifier on Leukemia data set using an informative genes extracted by methods based on their correlation with the class distinction, and statistical analysis. Experimental results show that the MFHSNN classifier is the most effective in classifying the type of leukemia cancer using only two of the most informative genes. MFHSNN yields 100% recognition accuracy and is well suited for the AAL/AML classification in cancer treatment. By comparing the performance with previous publications that used the same dataset, we confirmed that the proposed method provided the competitive, state-of-the-art results. Under the same context, it not only leads to better classification accuracies, but also has higher stability and speed. The training and testing time of MFSHNN is less than 0.4 seconds which will further drastically reduce if the proposed classifier is implemented in hardware. Our future work will focus on exploring unsupervised methods such as clustering combined with fuzzy classifier and the corresponding feature selection methods. Besides, we will further validate the performance of MFSHNN on more data sets.

REFERENCES

1 M. Schena, D. Shalon, R. W. Davis and P. O. Brown. “Quantitative monitoring of gene expression patterns with a complementary DNA microarray”, Science 267 (1995):pp. 467–470.

2 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.

3 R. Baumgartner, C. Windischberger, and E. Moser. “Quantification in functional magnetic resonance imaging: fuzzy clustering vs. correlation analysis”. Magn Reson Imaging, vol. 16, no. 2, pp. 115–125, 1998.

4 T. Kohonen, Ed. “Self-organizing maps”. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997.

5 T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. "Support vector machine classification and validation of cancer tissue samples using microarray expression data", Bioinformatics, vol. 16, pp. 906–914, 2000.

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

7 C. Shi and L. Chen. "Feature dimension reduction for microarray data analysis using locally linear embedding", APBC, 2005, pp. 211–217.

8 L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen. "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method", Bioinformatics, vol. 17, pp. 1131–1142, 2001.

9 T. Jirapech-Umpai and S. Aitken. "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes", Bioinformatics, vol. 6, pp. 168–174, 2005.

10 Min Su, M. Basu and A. Toure. "Multi-Domain Gating Network for Classification of Cancer Cells Using Gene Expression Data", In Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 286–289, 2002.

11 R Xu. G. Anagnostopoulos and D. Wunsch. "Tissue Classification Through Analysis of Gene Expression Data Using A New Family of ART Architectures", In Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 300–304, 2002.

.

12 Saeys Y, Inza I, Larranaga P. "A review of feature selection techniques in bioinformatics", Bioinformatics 2007, 23(19): 2507-2517.

13 Wang X, Gotoh O. "Microarray-Based Cancer Prediction Using Soft Computing Approach", Cancer Informatics, 2009, 123–39.

14 U V Kulkarni, T R Sontakke. "Fuzzy Hypersphere Neural Network Classifier", 10th IEEE int. conference on fuzzy systems, Dec 2001, 1559-1562.

15 S.-B. Cho, J. Ryu. "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features", Proc. IEEE 90 (11) (2002):1744–1753.

16 E.L. Lehmann. "Non-parametrics: Statistical Methods Based on Ranks". Holden-Day, San Francisco, 1975.

17 Deng Lin1, MAJinwen1 & PEI Jian2. "Rank sum method for related gene selection and its application to tumor diagnosis", Chinese Science Bulletin 2004. Vol. 49, No. 15, 1652-1657.

18 Devore, J. L. "Probability and Statistics for Engineering and the Sciences". 4th edition. California, Duxbury Press (1995).