Prediction of Cancer Subtypes using Fuzzy Hypersphere Clustering Neural Network

B. B. M. Krishna Kanth[†], Dr. U. V. Kulkarni[†] and Dr. B. G. V. Giridhar[††]

[†]*S.G.G.S.I.E.T, Nanded, Maharashtra State, India*
[††]*Andhra Medical College, Visakhapatnam, Andhra Pradesh State, India*

Summary The classification of different tumor types is of great importance in cancer diagnosis and drug discovery. However, most previous cancer classification studies are clinical-based and have limited diagnostic ability. Cancer classification using gene expression data is known to contain the keys for addressing the fundamental problems relating to cancer diagnosis and drug discovery. The recent advent of DNA microarray technique has made simultaneous monitoring of thousands of gene expressions possible. We propose a new method of classification system namely, the fuzzy hypersphere clustering neural network (FHCNN) which combines clustering and classification inorder to differentiate cancer tissues such as acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Experimental results show that our FHCNN model using one outstanding gene, Zyxin achieves the best classification accuracy of 94.12% where as other state-of-art methods could reach the best accuracy of 91.18%. Moreover FHCNN is more stable, and contains less number of parameter adjustments compared to all the other classification methods.

Key words: ***classification, gene ranking, fuzzy sets, neural network,***

1. Introduction

Microarrays [1, 2] allow simultaneous measurement of tens of thousands of gene expression levels per sample. It has changed biomedical research in a profound way and has rapidly emerged as a major tool to obtain gene expression profiles of human cancers [3, 4]. Since the development of microarray technology, many data mining approaches [18, 19, 20, 21, 22, 24] have been developed to analyze microarray data. Because typical microarray studies usually contain less than one hundred samples, the number of features (genes) in the data far exceeds the number of samples. This asymmetry of the data poses a serious challenge for standard learning algorithms that can be overcome by selecting a subset of the features and using only them in the classification. Generally, the microarray data analysis includes two key procedures: gene selection and classifier construction. From biological and clinical points of view, finding the small number of important genes can help researchers to concentrate on these genes and investigate the mechanisms for cancer development and treatment. It may bring down the cost of laboratory tests, because a patient needs to be tested on only few genes, rather than thousands of genes. Furthermore, it may be possible to obtain simple rules for doctors to make diagnosis without even using a classifier or a computer. If we survey and examine the established reports in this field, we will find that almost all the accurate classification results are obtained using more than a single gene. Recently, Wang. X et al. [9] proposed a rough set based soft computing method to conduct cancer classification using single genes. However, multi-gene models suffer from the disadvantage that it is not easy to assess which gene is more important in the models, because they are run on the basis of a group of genes. As a result, the significant biomarkers of related cancers are hard to be detected. In addition, multi-gene models are prone to impart the difficulty in understanding the models themselves. In this article, we explore the classification of cancer on the basis of single genes with leukemia dataset using our proposed FHCNN model. We want to underscore that sufficiently accurate classification can be achieved, and important biomarkers can be found with ease and efficiently by using single-gene models.

2. Materials and methods

2.1 Microarray Data

The number of training samples in leukemia dataset [10] is 38 which of them contain 27 samples of ALL class and 11 samples of AML class; the number of testing samples is 34 where 20 samples belong to ALL and remaining 14 samples belongs to AML class respectively.

2.2 Gene Selection

In order to score the similarity of each gene, an ideal feature vector [5] 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:

$$ideal_i = (0,0,0,\ldots..,0,1,\ldots\ldots,1,1,1) \quad (1)$$

Manuscript received February 5, 2011
Manuscript revised February 20, 2011

The similarity of g_i and g_{ideal} using Spearman correlation coefficient (SCC) [6] is defined as follows

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

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**.**

3. Topology of Fuzzy Hypersphere Clustering Neural Network

The FHCNN consists of two layers as shown in the Fig 1. The F_R layer accepts an input pattern and consists of n processing elements, one for each dimension of the pattern. The F_C layer consists of q processing nodes that are constructed during training and each node represents fuzzy set hypersphere (HS) which is characterized by its membership function. The processing performed by HS node is shown in the Fig 2. The weights between the F_R and F_C layer represent centre points (CPs) of the HSs. As shown in the Fig 2, $C_j = \left(c_{j1}, c_{j2}, c_{j3}, ..., c_{jn}\right)$ represents CP of the HS m_j.

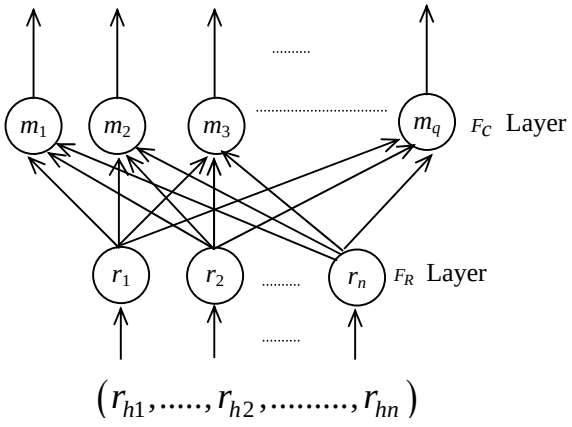

Fig 1. Fuzzy Hypersphere Clustering Neural Network**.**

The threshold input of HS, denoted by T is set to one and it is weighted by ζ that represents radius of the HS, which is created during the training process. The CPs of the HSs are stored in the matrix **C**. The radii of the HSs created during training process are bounded in the range $0 \le \zeta \le 1$. 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 puts maximum limit on the radius of the hypersphere.

Let the training set is, $R \in \left\{R_h \mid h = 1, 2, ..., P\right\}$, where $R_h = \left(r_{h1}, r_{h2}, r_{h3}, ..., r_{hn}\right) \in I^n$ is the h_{th} pattern, and the membership function of the hypersphere node m_j is defined as

$$m_j\left(R_h, C_j, \zeta\right) = 1 - f\left(l, \zeta, \gamma\right) \quad (3)$$

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

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

and the argument l is defined as,

$$l = \left(\sum_{i=1}^{n}\left(c_{ji} - r_{hi}\right)^2\right)^{1/2}. \quad (5)$$

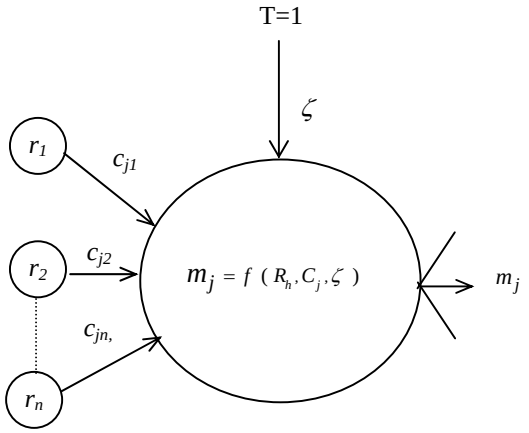

Fig. 2. Implementation of Fuzzy Hypersphere Clustering Neural Network.

The membership function returns m_j =1, if the input patterns 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 outside the hypersphere when the distance between R_h and C_j increases. The sample plot of membership function with centre point [0.5 0.5] and radius equal to 0.3 is shown in

Fig 3. It can be observed that the membership values decrease steadily with increase in distance from the hypersphere. Each node of F_C layer represents a cluster. The output of the j_{th} F_C node represents the fuzzy degree with which the input pattern belongs to the cluster m_j .

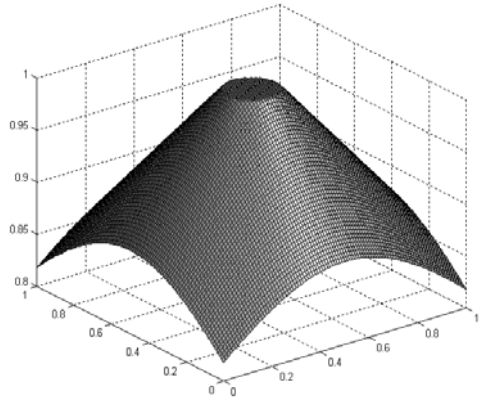

Fig. 3. Plot of Fuzzy Hypersphere Membership Function with center point CP [0.5 0.5], radius ζ =0.3 and sensitivity parameter γ =1.

3.1 The FHCNN Learning Algorithm

The HSs created in the F_C layer represent clusters. The ζ is a radius of the HSs created. It is user defined and bounded by $0 \le \zeta \le 1$. The number of clusters or hyperspheres constructed depends on the parameter ζ . The value of ζ is problem dependent and should be moderately low so that HSs created will include the patterns which are close to each other and possibly fall in the same cluster. The FHCNN learning algorithm consists of two steps for the creation of the HSs during training process. The steps in the learning algorithm are given by (1) Finding hypersphere centre points. (2) Removal of patterns grouped in the hypersphere.

3.2 Finding Hypersphere Centre Points

To determine the centre points of the cluster all the patterns are applied to each of the pattern and the patterns with euclidean distance less than or equal to ζ are counted for all the patterns. The pattern with maximum count is selected as a centroid or CP of the hypersphere. The process of selecting CP of the cluster is described below.

$$\text{If } \left(\sum_{j=1}^{P} \left| R_i - R_j \right| \le \zeta \right) \text{ then } D_i = D_i + 1 \text{ for } i = 1, 2, 3, ..., P, \quad (6)$$

where R_i and R_j are the i_{th} and j_{th} patterns respectively in the dataset R , D is a P -dimensional vector and D_i is i_{th} element of vector D which contains number of patterns falling around i_{th} pattern whose euclidean distance is less than or equal to ζ . To find the pattern with the maximum count the equation (7) is used in which D_{max} is the maximum value in the row vector D , and D_{ind} is an index of the maximum value.

$$[D_{max}\ D_{ind}] = max[D] \quad (7)$$

The pattern R_{ind} from the dataset R is the most appropriate and chosen as a CP of the first hypersphere m_1 . The hypersphere m_1 returns fuzzy membership value equal to one for the patterns which fall around the selected centre point R_{ind} with the distance less than or equal to ζ . Hence, these patterns are grouped in a cluster and the pattern R_{ind} acts as CP of the created cluster. The weight assigned to the synapses of the created hypersphere is described using equation (8).

$$C_1 = R_{ind} \quad (8)$$

3.3 Removal of Grouped Patterns in the Hypersphere

The clustered patterns in the previous step are eliminated and the next pass uses the remaining unclustered patterns to create new hyperspheres.

Let R_p , R_c and R_n represent set of patterns used in the current pass, set of patterns clustered in the current pass and set of patterns that will be used in the next pass, respectively. Then R_n can be described as,

$$R_n = R_p - R_c = \left\{ R_n \mid R_n \in R_p \text{ and } R_n \notin R_c \right\} \quad (9)$$

The R_n calculated in the current pass becomes R_p for the next pass. Above two steps are repeated until all the patterns are clustered. Each node of F_C layer constructed during training represents a cluster and gives a soft decision. The output of k_{th} F_C node represents the degree to which the input pattern belongs to the cluster m_k .

4. RESULTS AND DISCUSSION

We evaluated the proposed approach on the leukemia dataset, which consists of 72 samples (38 training samples, 34 testing samples) each described by 7129 attributes (genes). The pathological classes (targets) to be predicted are ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia). As a preprocessing step, we ranked all the 7129 genes using SCC scoring approach. We picked out 10 genes with the largest SCC values from the training samples to do the classification. Table 1 shows these 10 genes. We input these genes one by one to the FHCNN according to their ranks. During all the experiments using the FHCNN, the parameter γ is set to 1 and ζ is adjusted to tune the performance to get maximum possible accuracy by varying number of created HSs. When we trained FHCNN with 38 patterns of the gene Zyxin with ζ equal to 0.2, it created four clusters. After that, the FHCNN performance is assessed on the independent 34 testing samples for classification. This process is repeated for all the remaining selected genes. Among the top 10 genes, the top four genes having Gene ids #4847, #1882, #1834 and #760 were among the biologically instructive genes identified earlier by many other approaches [7-12]. Moreover, when considering the performance of the selected genes and FHCNN with each class separately, the five genes with Gene ids #760, #1834, #4373, #6855 and #3252 showed 100% best classification accuracy for samples related to ALL class, and the Gene id #4377 attained 100% best classification accuracy for samples belonging to AML class in leukemia dataset respectively.

TABLE 1: TOP 10 GENES WITH THE BEST CLASSIFICATION ACCURACY USING FHCNN AND SCC GENE SELECTION METHOD

Gene id	#Correctly classified samples (ALL/AML)	Classification accuracy (%) (ALL/AML)
4847	32 (19/13)	**94.12** (95/92.86)
1882	32 (19/13)	**94.12** (95/92.86)
760	32 (20/12)	**94.12** (**100**/85.71)
1834	32 (20/12)	**94.12** (**100**/85.71)
2402	29 (18/11)	85.29 (90/78.57)
4373	31 (20/11)	91.18 (**100**/78.57)
6855	30 (20/10)	88.24 (**100**/71.43**)**
6041	31 (19/12)	91.18 (95/85.71)
3252	31 (20/11)	91.18 (**100**/78.57)
4377	30 (16/14)	88.24 (80/**100)**

The leukemia dataset has been well studied by many researchers. Regarding the leukemia dataset, the best classification accuracy results reported in our and some other works are shown in Table 2. If using single genes, our accuracy is the highest among all the methods, and the other methods must use far more genes to achieve our classification accuracy. Using one common gene Zyxin, until now all other previously published methods [9, 11, 12, 14, 20, 22] could reach the best classification accuracy of 91.18%, whereas our proposed method could achieve 94.12% best classification accuracy which is shown in Table 3.

TABLE 2: COMPARISON OF THE BEST CLASSIFICATION ACCURACY WITH LEUKEMIA DATA SET

Methods	# Genes	#Correctly-classified samples(Accuracy)
Proposed	1	**32** (94.1 %)
Wang. X. et al. [9]	1	32 (94.1 %)
Tong et al.[21]	2	31 (91.2 %)
Xu. R. et al.[24]	5	31 (91.2 %)
Sun et al. [12]	1	31 (91.2 %)
Banerjee et al. [13]	9	31 (91.2 %)
Li et al. [14]	1	31 (91.2 %)
Tan et al. [15]	1038	31 (91.2 %)
Wang. Y. et al. [12]	1	31 (91.2 %)
Cong. G. et al. [16]	10-40	31 (91.2 %)
Golub et al. [10]	50	29 (85.3 %)
Furey et al.[17]	25-1000	30-32 (88.2%-94.1%)

TABLE 3: COMPARISON OF THE BEST CLASSIFICATION ACCURACY WITH LEUKEMIA DATA SET USING ONE OUTSTANDING GENE ZYXIN

Methods	#Correctly-classified samples (Accuracy %)
Proposed	**32** (94.1 %)
Kulkarni. U.V. et al. [20]	31 (91.2 %)
Wang. X. et al. [9]	31 (91.2 %)
Wang. Y. et al. [11]	31 (91.2 %)
Sun. L. et al. [12]	31 (91.2 %)
Li .J. et al. [14]	31 (91.2 %)
Li. W. et al. [22]	31 (91.2 %)
Frank. et al. [23]	30-31 (88.24%-91.18%)

5. Conclusion

The proposed work mainly focused on classification of acute leukemia using only single genes, particularly using one outstanding top gene, Zyxin. A learning strategy which combines clustering and fuzzy classifier with fuzzy hypersphere membership function was used for predicting the class of cancer. Zyxin was selected as a top ranked gene using Spearman coefficient gene selection method

and using this single gene our method could achieve the best classification accuracy of 94.12%(2 out of 34 test samples are classified wrongly) where as all others could achieve only 91.18%(3 out of 34 test samples are classified wrongly). Our future work will look upon training the new patterns adaptively without retraining again along with already trained patterns and removing overlapping of hyperspheres of different classes so that it may help to increase the classification accuracy to a greater extent.

References

[1] Schena, M., Shalon, D., Davis, R.W. and Brown P.O. "Quantitative monitoring of gene expression patterns with a complementary dna microarray, " Science 270, 467-470, 1995

[2] DeRisi, J., Penland, L., Brown, P.O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y.A. and Trent J. M. 'Use of a cDNA microarray to analyze gene expression patterns in human cancer," Nature Genetic 14, 457-60, 1996.

[3] Shi, T., Liou, L. S., Sadhukhan, P., Duan, Z. H., Hissong, J., Almasan, A., Novick, A. and DiDonato, J.A. "Effects of resveratrol on gene expression in renal cell carcinoma," Cancer Biology and Therapy 3, pp. 882-888, 2004.

[4] Liou, L. S., Shi, T., Duan, Z. H., Sadhukhan, P., Der, S. D., Novick, A., Hissong, J., Almasan, A. and DiDonato, J.A. "Microarray gene expression profiling and analysis in renal cell carcinoma," BMC Urology, 4-9,2004.

[5] Chen, Y., and Zhao, Y. "A novel ensemble of classifiers for microarray data classification," Applied Soft Computing 8(4):1664–1669,2008.

[6] Devore, J. L. "Probability and statistics for engineering and the sciences," 4th edition. Duxbury Press, California, 1995.

[7] Li, D., and Zhang, W. "Gene selection using rough set theory," In Proceedings of the 1st International Conference on Rough Sets and Knowledge Technology, pp. 778–785, 2006.

[8] Momin, B. F., and Mitra, S. "Reduct generation and classification of gene expression data," In Proceedings of the 1st International Conference on Hybrid Information Technology, 699-708, 2006.

[9] Wang, X., and Gotoh, O. "Cancer classification using single genes," Genome Informatics 23(1):176-88, 2009.

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

[11] Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K.F. and Mewes, H. "Gene selection from microarray data for cancer classification—a machine learning approach," Computational Biology and Chemistry 29(1):37-46, 2005.

[12] Sun, L., Miao, D., and Zhang, H. "Efficient gene selection with rough sets from gene expression data," In Proceedings of the 3rd International Conference on Rough Sets and Knowledge Technology, 164–171,2008.

[13] Banerjee, M., Mitra, S. and Banka, H. "Evolutionary-rough feature selection in gene expression data," IEEE Transaction on Systems, Man, and Cybernetics, Part C: Application and Reviews 37, pp. 622–632, 2007.

[14] Li, J., and Wong, L. "Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns," Bioinformatics 18(5):725-734, 2002.

[15] Tan, A. C., and Gilbert, D. "Ensemble machine learning on gene expression data for cancer classification," Applied Bioinformatics 2(3 Suppl):75-83, 2003.

[16] Cong, G., Tan, K.L., Tung, A. and Xu, X. "Mining top-k cover ing rule groups for gene expression data," In Proceedings of the ACM SIGMOD International Conference on Management of Data, 670-681, 2005.

[17] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D. "Support vector machine classification and validation of cancer tissue samples using microarray expression data," Bioinformatics 16(10):906-914,2000.

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

[19] Tan, A. H., and Pan, H. "Predictive neural networks for gene expression data analysis," Neural networks 18, 297-306,2005.

[20] Kulkarni, U. V., and Sontakke, T. R. "Fuzzy hypersphere neural network classifier," In 10th IEEE International conference on fuzzy systems, 1559-1562,2001.

[21] Tong, D. L., Phalp, K. T., Schierz, A. C. and Mintram, R. "Innovative hybridization of genetic algorithms and neural networks in detecting marker genes for leukemia cancer," In 4th IAPR International Conference in Pattern Recognition for Boinformatics, Sheffield, UK,2009.

[22] Li, W., and Yang, Y. "How many genes are needed for a discriminant microarray data analysis. Methods of Microarray Data Analysis," Kluwer Academic Publishers. pp 137-150,2002.

[23] Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten, I.H. "Data mining in bioinformatics using Weka," Bioinformatics 20, 2479–2481,2004.

[24] Xu, R., Anagnostopoulos. and Wunsch, D. "Tissue classification through analysis of gene expression data using a new family of ART architectures," In Proceedings of International Joint Conference on Neural Networks 1, pp. 300-304, 2002.

B.B.M.Krishna Kanth received the B.E. degree in Electronics and Communication Engineering from Andhra University, in the year 1999. He received the M.E. degree in Computer Technology from S.R.T.M. University, Nanded in the year 2002. He is currently Research Scholar under the guidance of Dr.U.V.Kulkarni in the Department of Computer Science Enginering, SGGS Institute of Engineering and Technology, Nanded, Maharashtra. His current research interests include various aspects of Neural Networks and Fuzzy Logic, DNA analysis and Bioinformatics.

Dr.U.V.Kulkarni received the PhD degree in Electronics and Computer Science Engineering from S.R.T.M. University, Nanded in the year 2003. He is currently working as Head of Computer Science Engineering Department and Dean Academics, SGGS Institute of Engineering and Technology, Nanded, Maharashtra.

Dr.B.G.V.Giridhar received the Doctor of Medicine (D.M) (post-doctorate) degree in Endocrinology and Doctor of Medicine (M.D) in General Medicine from Andhra Medical College, Visakhapatnam, in the year 2006 and 2000 respectively. He is currently working as Assistant Professor, Andhra Medical College, Visakhapatnam.