ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Cancer Classification in Microarray Data Using Gene Expression with KNN and FNN

Dr. R. Mallika
Director, Department of Computer Applications,
Tamilnadu College of Engineering,
Coimbtore, India.
mallikapanneer@hotmail.com

A.K. Selvanayaki*
Research Scholar,
Sri Ramakrishna College of Arts and Science for Women,
Coimbatore, India.
akselvanayaki@gmail.com

Abstract : Classification is used for predicting class labels. In the accurate cancer classifications information gain is highly effective ranking scheme, gene subset selection using Euclidean distance metrics. K-Nearest Neighbor and Fuzzy Neural Network are used as good classifiers. Many other gene importance ranking schemes and classifiers may also be used in this approach. Classification involves four steps. In the first step, top genes are selected using a feature importance ranking scheme. In the second step, gene subset is generated using distance metrics. In the third step, the classification capability of all genes within the subset is classified by a good classifier. In the fourth step, the top genes selected using ranking scheme (without subset selection) is classified by a same classifier. Two data sets are used for classification, 1.Lymphoma dataset and 2.Liver dataset. In the two datasets, a small part of the data is missing. A k-nearest neighbor algorithm should be applied to fill the missing values. This research suggests a unified criterion for gene ranking and gene subset selection. In the micro array technology to find specific cancer –related genes that can be used to diagnose and predict cancer stage.

Key Words: Classification, KNN, FNN, Euclidean Distance, Information Gain.

I. INTRODUCTION

The microarray technology is used to classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. DNA microarrays have been used to profile gene expression in cancer and other diseases. Cancer microarray data normally contains a small number of samples which have a large number of gene expression levels as features. To select relevant genes involved in different types of cancer remains a challenge. In order to extract useful gene information from cancer microarray data and reduce dimensionality, feature selection algorithms were systematically investigated in this study [1].

This research aims at finding the gene subsets that can ensure highly accurate classification of cancers from micro array data by using supervised machine learning algorithms. The KNN and FNN classifier are used for cancer classifications [6]. The significance of finding the gene subsets is: 1) It greatly reduces the computational burden and “noise” arising from irrelevant genes. 2) It simplifies gene expression tests to include only a very small number of genes rather than thousands of genes, which can bring down the cost for cancer testing significantly. 3) It calls for further investigation into the possible biological relationship between these small numbers of genes and cancer development and treatment.

DNA micro array technology allows for simultaneous monitoring and measurement of thousands of gene expression activation levels in a single experiment, and is universally used in medical diagnosis and genetic analysis. Gene expression data typically has a high dimension and sample size [5]. Generally only small numbers of gene expression data are strongly correlated with certain phenotype. To analyze gene expression profiles correctly, feature selection is used. Feature selection has certain advantages, such as effective extraction of genes that influence classification accuracy, elimination of irrelevant genes, and improvement of the classification accuracy calculation [7].

Two data sets are used for classification, 1.Lymphoma dataset and 2.Liver dataset. In the two datasets, a small part of the data is missing [8][10]. A k-nearest neighbor algorithm should be applied to fill the missing values. A fuzzy neural network (FNN) is a connectionist model for fuzzy rules implementation and deduction. There is variety of architectures and functionalities of FNN. This approach could be able to produce more accurate results for the classification module. To evaluate the proposed algorithm, a series of experiments on a real gene data are conducted and the results are compared with the existing classification algorithms.

II. METHODOLOGY

A. Existing System:

In the existing approach, feature ranking algorithms is used to reduce the feature space of the DARPA data set from 41 features to the six most important features. Three ranking algorithms are used based on Support Vector Machines (SVMs), Multivariate Adaptive Regression Splines (MARSs), and Linear Genetic Programs (LGPs) to assign a weight to each feature. Experimental results showed that the classifier’s accuracy degraded by less than 1 percent when the classifier was fed with the reduced set of features. Sequential backward search was used to identify the important set of features: starting with the set of all features, one feature was removed at a time until the accuracy of the classifier was below a certain threshold. Even though these approaches provide an agreed result, the accuracy level and the time taken for the computation of the problem solution is only compromising.

B. Proposed Algorithm:

In proposed system, a new approach of handling the pattern recognition is proposed. Although there are many existing approaches in the case of diagnosing the cancer infected gene, a Fuzzy Neural Network based model is proposed for this problem [6]. The information gain measure is used for the gene ranking approach. Later the subset selection of the ranked data will be takes place. Euclidean distance function is used for the subset selection process. Next the classification of the K-Nearest Neighbor algorithm is takes place for the data sets taken. The experimental results show that the proposed approach performed better than the existing approach in terms of the accuracy.

The following steps are used for both Lymphoma and Liver dataset to find accuracy of each method.

Steps for Subset Selection

Step1: Preprocess the data using KNN impute.

Step2: Rank all the genes (M*N) using Information gain.

$$IG(Ex,a) = H(Ex) - \sum_{v \in values(a)} \frac{|\{x \in Ex | value(x,a) = v\}|}{|Ex|} \cdot H(\{x \in Ex | value(x,a) = v\})$$

Step3: For top selected gene pair, find the Euclidean distance.

$$\mathbf{d}(\mathbf{p},\mathbf{q}) \quad = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Step4: For a given threshold value, select the gene subsets.

Step5: Use KNN and FNN classification.

Steps for Without Subset Selection

Step1: Preprocess the data using KNN impute.

Step2: Rank all the genes (M*N) using Information gain.

Step3: Use KNN and FNN classification.

III. RESULTS AND DISCUSSION

The proposed approach is compared with the existing systems and evaluated. Here for the validation process, two datasets namely lymphoma and liver are used. Using those datasets proposed approach is compared with the existing approaches. In this case proposed approach is compared with existing in terms of accuracy. The subset selection and without subset selection are compared for both the approaches. The subset selection gives more accuracy compared to without subset selection in both the approaches. The Fuzzy Neural Network approach works better than K-Nearest Neighbor and other existing approaches in terms of accuracy.

Table 1: Accuracy results for Lymphoma Dataset using KNN and FNN

Number of genes	KNN		FNN	
	With Subset Selection	Without Subset Selection	With Subset Selection	Without Subset Selection
1000	**93.66**	70	**95**	83
1500	**93.75**	75	**98**	87
2000	**94.91**	84	**97**	87.50
2500	**90.58**	77	**96**	82
3000	**91.66**	87.50	**94**	87

In the table1, Lymphoma data set have with subset selection and without subset selection accuracy with both KNN and FNN as the classifier and searched for gene combinations among the top N genes value. The results show Fuzzy Neural Network algorithm is good and accuracy.

Table 2: Accuracy results for Liver Dataset using KNN and FNN

Number of genes	KNN		FNN	
	With Subset Selection	Without Subset Selection	With Subset Selection	Without Subset Selection
1000	**93.02**	65	**95**	83
1500	**90.54**	71.42	**96**	87
2000	**95.34**	77.14	**98**	86
2500	**94**	65.71	**97**	87
3000	**97**	68	**98**	88

In the table2, Liver data set have with subset selection and without subset selection accuracy with both KNN and FNN as the classifier and searched for gene combinations among the top N genes value. The results show Fuzzy Neural Network algorithm is good and accuracy.

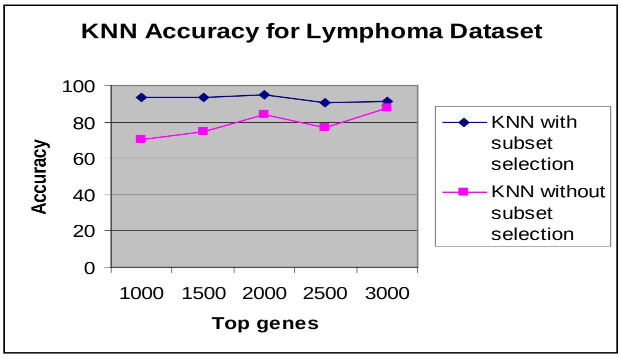

Figure 1: Accuracy Graph plotted with the of no. of Top Genes selected for Lymphoma Dataset using KNN

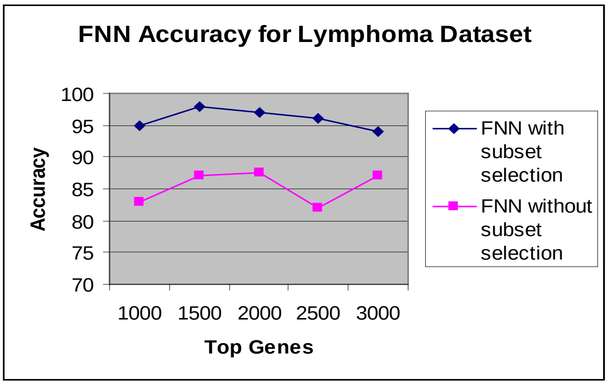

Figure 2: Accuracy Graph plotted with the no. Top Genes Selected for Lymphoma Dataset using FNN

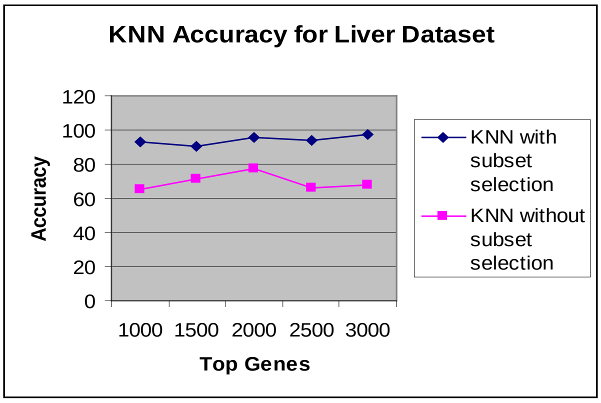

Figure 3: Accuracy Graph plotted with the no. of Top Genes selected for Liver Dataset using KNN

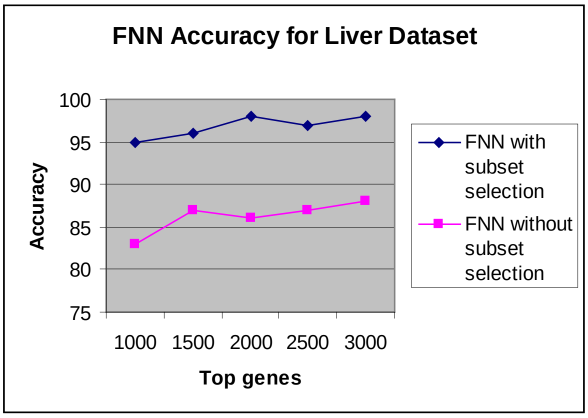

Figure 4: Accuracy Graph plotted with the no. of Top Genes Selected for Liver Dataset using FNN

IV. CONCLUSION

A novel approach to select the best features for the recognition of the cancer is presented. The proposed K-Nearest Neighbor and Fuzzy Neural Network approaches are used [7]. The performance evaluation of the proposed system was evaluated and compared with existing approaches. Information related to gene expression levels may have played an important role in the results. When compared to k-Nearest Neighbor, Fuzzy Neural Network predicts the cancer in terms of the accuracy from microarray data.

V. REFERENCES

[1] Alizadeh et al. Dstinct types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling. Nature 403:503-511(2000).

[2] Hughes, S.E. Differential expression of the fibroblast growth factor receptor (FGFR) multigene family in normal human audit tissues. J.Histochem. Cytochem. 45, 1005-1019(1997).

[3] Dave. S et al. prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-infiltrating Immune Cells New England Journal of Medicine, Vol 351,2159-2169 Nov 2004.

[4] Lucila Ohno-Machado, Staal Vinterbo, and Griffin Weber, Classification of Gene Expression Data Using Fuzzy Logic Decision Systems Group, Brigham and Women’s Hospital Division of Health Sciences and Technology, Harvard and MIT Decision Systems Group, Thorn 310.

[5] Ying Lu, Jiwei Han., Cancer classification using gene expression data, 2003.

[6] Hon Keung Kwan, Senior Member, IEEE and Yaling Cai, A Fuzzy Neural Network and its Application to Pattern Recognition, IEEE transactions on fuzzy systems, vol. 2. No. 3, august 1994.

[7] Rui-Ping Lia, Masao Mukaidonob, A fuzzy neural network for pattern classification and feature selection, aCaelum Research Corporation, Rockville, MD 20850, USA.

[8] Shipp et al. Diffuse Large B-Cell Lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nat. Med. 8, 68-74(2002).

[9] W. Penny and D. Frost, Neural networks in clinical medicine, Med. Decis. Making 16 (1996), pp. 386-398.

[10] X. Chen et al., Gene expression patterns in human liver cancers, Molecular biology of the cell, Vol 13, pp 1929-1939, 2002.