Accurate Cancer Classification using Expressions of Very few Genes

N.Revathy
Assistant Professor
Department of Computer Applications,
Karpagam College of Engineering
Coimbatore, India

Dr.R.Amalraj
Associate Professor
Department of Computer Science,
Sri Vasavi College, Erode, India

ABSTRACT
Gene expression profiling by microarray technique has been effectively utilized for classification and diagnostic guessing of cancer nodules. Several machine learning and data mining techniques are presently applied for identifying cancer using gene expression data. Though, these techniques have not been proposed to deal with the particular needs of gene microarray examination. Initially, microarray data is featured by a high-dimensional feature space repeatedly surpassing the sample space dimensionality by a factor of 100 or higher. Additionally, microarray data contains a high degree of noise. The majority of the existing techniques do not sufficiently deal with the drawbacks like dimensionality and noise. Gene ranking method is later introduced to overcome those problems. Some of the widely used Gene ranking techniques are T-Score, ANOVA, etc. But those techniques will sometimes wrongly predict the rank when large database is used. To overcome these issues, this paper proposes a technique called Enrichment Score for ranking purpose. The classifier used in the proposed technique is Support Vector Machine (SVM). The experiment is performed on lymphoma data set and the result shows the better accuracy of classification when compared to the conventional method.

Keywords
Enrichment Scores, Support Vector Machine, Gene Ranking

1. INTRODUCTION
THE diagnosis of complex genetic diseases like cancer has conventionally been done based on the non-molecular characteristics like kind of tumor tissue, pathological characteristics and clinical phase. DNA microarray method has concerned great attention in both the scientific and in industrial areas. Numerous examinations have been presented on the usage of microarray gene expression examination for molecular categorization of cancer. Several machine learning techniques have been developed for the examination of microarray data [5, 16]. The grouping of gene microarray method and machine learning technique assures new approaches into mechanisms of living schemes. An application field where these methods are likely to create key contributions is the identification of cancers depends on clinical phase and biological activities. Such classifications have a huge contribution on diagnosis and treatment.

Generally, a classifier for this purpose must deal with the following problems:

- The classifier must offer an easy-to interpret measure of assurance for its judgments. Thus, the final diagnosis rests with the medical specialist who evaluates if the confidence of the classifier is highly sufficient.
- The classifier must consider asymmetrical wrong classification costs for false positive and false negative classifications.

To achieve this, the microarray gene supplied to the classifier should be consistent. This can be achieved by ranking the gene accordingly. This paper uses Enrichment Score for ranking the gene and the classifier used in this paper is Support Vector Machine [6, 11].

2. RELATED WORKS
There are different techniques proposed by different authors for the prediction of cancer regions. Every technique has its own advantages and disadvantages. Some of the existing techniques are presented in this section.

Rui *et al.,* [1] proposed a multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data [7, 9]. It is critical for cancer prediction and treatment to perfectly categorize the site of origin of a cancer. With huge progress of DNA microarray techniques, creating gene expression profiles [8] for various cancer kinds has previously turn out to be a capable way for cancer classification [10]. In addition to research on binary classification like normal versus tumor samples that focuses on various issues from a mixture of disciplines, the discrimination of multiple tumor kinds is also essential. In the meantime, the choosing of genes that are appropriate to definite cancer kinds not only enhances the performance of the classifiers, but also offers molecular insights for treatment. Here, the author utilizes the semisupervised ellipsoid ARTMAP (ssEAM) for multiclass cancer discrimination and particle swarm optimization for informative gene selection. ssEAM is a neural network technique [14] embedded in adaptive resonance theory and applicable for classification purpose. ssEAM characterizes fast, stable, and finite learning and generates hyperellipsoidal clusters, containing complex nonlinear decision boundaries. PSO is an evolutionary algorithm-based method for global optimization. A discrete binary version of PSO is used to represent whether genes are selected or not. The effectiveness of ssEAM/PSO for multiclass cancer diagnosis is illustrated with the help of testing it on three publicly existing multiple-class cancer data sets.

Huilin *et al.,* [2] presents the optimized kernel machines for cancer classification using gene expression data. This technique enhances the performances of the classifiers in

classifying gene expression data [15]. Intending to enhance the class separability of the data, the author uses a highly flexible kernel function model, the data-dependent kernel, as the objective kernel to be optimized.

Xiyi *et al.,* [3] given a cancer classification technique by sparse representation using microarray gene expression data. The author presents a novel technique is for diagnosis of cancer with the help of gene expression data by casting the classification difficulty as finding sparse representations of test samples in accordance with the training samples. The sparse representation is effective computed by lscr1-regularized least square.

Runxuan *et al.,* [4] proposed a multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis. The author used the newly created Extreme Learning Machine (ELM) for directing multicategory classification in the cancer diagnosis field. ELM neglects drawbacks such as local minima, improper learning rate and overfitting usually faced by iterative learning techniques and completes the training quickly. The author estimates the multicategory classification performance of extreme learning machine on three benchmark microarray data sets for cancer diagnosis, namely, the GCM data set, the Lung data set, and the Lymphoma data set.

3. METHODOLOGY

There are two phases included in the proposed technique. In the first phase, every gene in the training data are ranked with the help a scoring technique called enrichment scores. In the second phase, the classification ability of every simple combination among the selected geness is tested with the help of a classifier called Support Vector Machine.

Phase 1: Enrichment Scores for Gene Importance Ranking

This phase determine the importance ranking of all gene with the help of a feature ranking measure called enrichment score. Enrichment Scores technique will take the inputs as:

1. Genome-wide expression profiles containing p genes and n samples with every sample subsequent to one of two classes, C_1, C_2, the expression of the j^{th} gene in the i^{th} sample is x_j^i ;
2. A database consisting of m gene sets $\Gamma=\{\gamma_1,\ldots, \gamma_m\}$where every gene set γ_k is a list of genes (a subset of the p genes in the data set) belonging to a pathway or other functional or structural class;
3. A ranking method and correlation value that acquires the expression data set and labels as inputs and generates the correlation statistics for every sample that reveals the correlation of the p genes in that sample in accordance with the distribution of expression in the two categories. The correlation statistics for the i-th sample could be $c_i = \{c_1^i, \ldots, c_p^i\}$;

and generates the following as outputs:

1. An enrichment score for every sample in the data set in accordance with every gene set in the database, ES_i^k according to the enrichment of the i-th sample accordance with the k-th gene set;
2. The evaluation of confidence for every enrichment score is represented by a p-value with multiplicity by considering thr Family-Wise Error Rate (FWER) p-values and a False Discovery Rate (FDR) q-values.

Provided the correlation statistics for the i-th sample $c_i = \{c_1^i, \ldots, c_p^i\}$ and a gene set γ_k, the following discrete random walk over the indices of the rank-ordered correlation statistic is constructed

$$v(l) = \frac{\sum_{j=1}^{l} |c(j)|^r I(g(j) \in \gamma_k)}{\sum_{j=1}^{p} |c(j)|^r I(g(j) \in \gamma_k)} - \frac{\sum_{j=1}^{l} I(g(j) \notin \gamma_k)}{p - |\gamma_k|}$$

where $c_{(j)}$ represents the rank-ordered correlation statistic, r represents a parameter (usually r = 1), γ_k is the k-th gene set, $I(g_{(j)} \in \gamma_k)$ represents the indicator function on whether the j-th gene (the gene consequent to the j-th ranked correlation statistic) is in gene set γ_k, $|\gamma_k|$ represents the number of genes in the k-th gene set, and p represents the number of genes in the data set. The enrichment statistic for the i-th sample in accordance with the k-th gene set is the maximum variation of the random walk from zero.

$$ES_i^k = v[\arg\max_{l=1,\ldots p} v(l)]$$

Phase 2: Classification using Support Vector Machines

Support Vector Machines (SVMs) [12, 13] is a kind of classifier that are a set of associated supervised learning techniques especially for classification.SVM will create a separating hyperplane in the space, one that increases the boundary between the two data sets. To establish the boundary, two parallel hyperplanes are created, one on every side of the separating hyperplane between the two data sets. For SVM, a data point is represented as a p dimensional vector, and it is required to distinguish whether it can split such points with a p − 1-dimensional hyperplane. This is called a linear classifier.

As support vector machines are linear classifier that has the capability of finding the optimal hyper plane that increases the separation among patterns, this characteristic creates support vector machines as a potential means for gene expression data examination purposes. The 5 fold cross validation (CV) is performed for support vector machine in the training data set to adjust their constraints. First, the entire data set is split into training (F1) and testing (F2) data by random. The genes are ranked with the help of samples of F1. The combination (FC1) is produced with the help of 2 genes from 20. Then FC1 is arbitrarily split into 5 folds (fc1, fc2, fc3, fc4 and fc5). Among these folds one fold is chosen for testing. The other 4 folds are used as a classifier for SVM. This combination is produces continuously and stops only when the better accuracy is achieved. At last with the fitted SVM, the prediction can be carried out.

4. EXPERIMENTAL RESULTS

The experimentation on the proposed method is carried on lymphoma data set. In the lymphoma data set, there are 42 samples obtained from Diffuse Large B-cell Lymphoma (DLBCL), nine samples from Follicular Lymphoma (FL), and 11 samples from Chronic Lymphocytic Leukemia (CLL). The whole dataset contains the expression data of 4026 genes. Some data may be lost in the dataset because of some error.

For filling those lost values k-nearest neighbor technique is used.

Initially, the 62 samples are split randomly into 2 groups: 31 samples for testing, 31 samples for training. Based on the enrichment scores in the training set, the whole sets of 4026 genes are ranked. Then, 200 genes with highest rank are chosen. Finally, the genes are passed to the SVM classifier for classification.

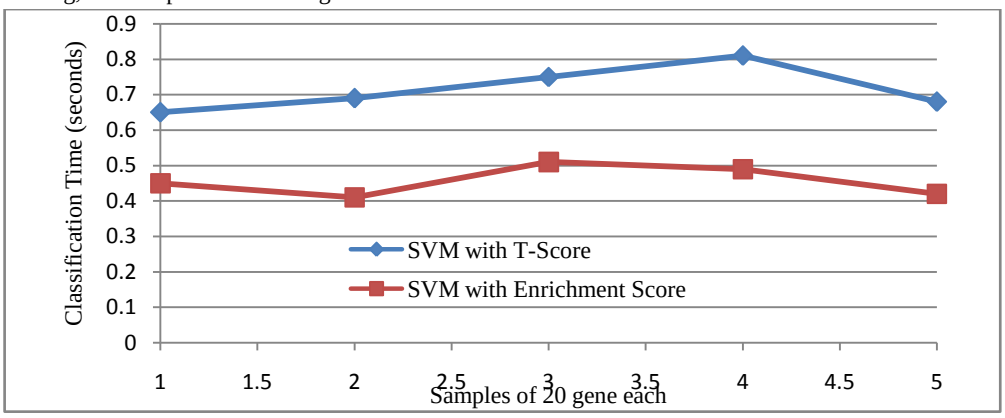

Figure 1 Classification Time for Different Gene Samples

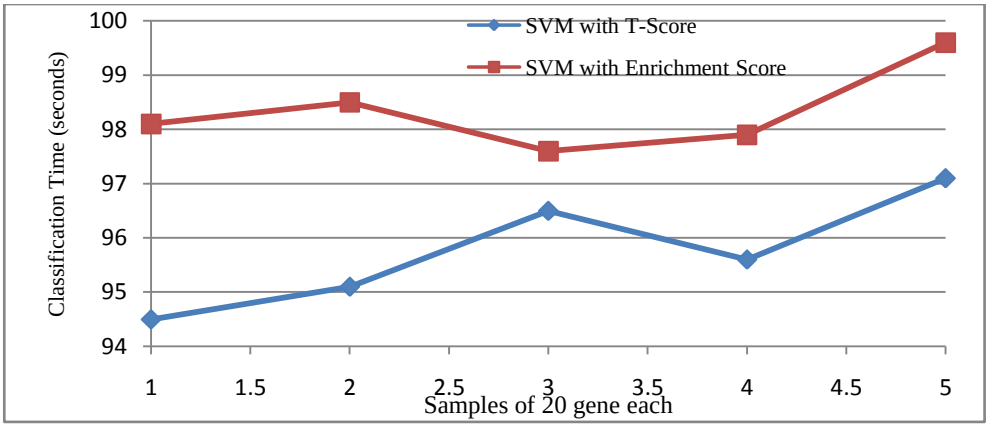

Figure 2 Accuracy of Classification for Different Gene Samples

Figure 1 shows the resulted classification time for different gene samples. It can be observed that the proposed technique with Enrichment Score takes lesser time for classification when compared to the existing technique with T-Score. Figure 2 shows the obtained accuracy for classification. It is clear from the figure that the proposed technique resulted in better accuracy for all the samples used for classification.

6. CONCLUSION

Cancer research is one of the key research fields in the medical science. Exact prediction of several tumor kinds has higher value in offering enhanced treatment and toxicity reduction on the patients. In the past, cancer categorization is generally depends on morphological and clinical analysis. These previous cancer classification techniques are stated to have many drawbacks in their diagnostic capability. To overcome those drawbacks in cancer classification, efficient technique in accordance with the global gene expression examination have been evolved. The expression level of genes holds the solutions to overcome basic drawbacks related to the prevention and treatment of cancer. The microarray gene data must be preprocessed for classification with good accuracy using the classifier. The gene ranking technique is used to support that task. This paper uses enrichment score for ranking the gene. Then the classifier is trained with that data. Finally, the classification of gene for identifying the cancer is performed. The classifier used in this paper is support vector machine. The experiment is performed with the help of lymphoma data set. The experimental result shows that the proposed technique results in better accuracy and consumes less time for classification when compared to the conventional techniques.

7. REFERENCES

[1] Rui Xu, Anagnostopoulos, G.C. and Wunsch, D.C.I.I., "Multiclass Cancer Classification Using Semisupervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.4, No.1, Pp. 65-77, 2007.

[2] Huilin Xiong and Xue-Wen Chen, "Optimized Kernel Machines for Cancer Classification Using Gene Expression Data", Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Pp. 1-7, 2005.

[3] Xiyi Hang, "Cancer Classification by Sparse Representation using Microarray Gene Expression Data", IEEE International Conference on Bioinformatics and Biomeidcine Workshops, Pp. 174-177, 2008.

[4] Runxuan Zhang, Huang, G.B., Sundararajan, N. and Saratchandran, P., "Multicategory Classification Using An Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No.3, Pp. 485 – 495, 2007.

[5] Brown, M., "Knowledge Based Analysis of Micorarray Gene Expression Data by using Support Vector Machines", In Proc. of the National Academy of Sciences, Vol. 97, Pp. 262–267, 2000.

[6] Xiaogang Ruan, Jinlian Wang, Hui Li and Xiaoming Li, "A Method for Cancer Classification Using Ensemble Neural Networks with Gene Expression Profile", The 2nd International Conference on Bioinformatics and Biomedical Engineering, Pp. 342-346, 2008.

[7] Berns, A., "Cancer: Gene Expression in Diagnosis", Nature, Pp. 491–492, 2000.

[8] Alizadeh, A., "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling", Nature, Pp. 503–511, 2000.

[9] Campbell, V., Li, Y. and Tipping, N. "An Efficient Feature Selection Algorithm for Classification of Gene Expression Data", 2001.

[10] Dubitzky, W., Granzow, M. and Berrar, D., "Comparing Symbolic and Subsymbolic Machine Learning Approaches to Classification of Cancer and Gene Indentification", Kluwer Academic, 2002.

[11] Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M. and Haussler, D. "Support Vector Machine Classification and Validation of Cancer Tissue Samples using Microarray Expression Data", Bioinformatics, 2001.

[12] Fujarewicz, K., Kimmel, M. and Rzeszowska-Wolny, J., "Improved Classification of Gene Expression Data using Support Vector Machines", Journal of Medical Informatics and Technologies, Vol. 6, Nov 2001.

[13] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., "Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning, 2000.

[14] Khan, J., Wei, J., Ringner, M. and Saal, L., "Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks", Nature Medicine, 2001.

[15] Ramaswamy, S., Tamayo, P. and Rifkin, R., "Multiclass Cancer Diagnosis using Tumor Gene Expression Signatures", PNAS, Pp. 15149–15154, 2001.

[16] Zhang, H., Yu, C., Singer, B. and Xiong, M., "Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data". PNAS, Pp. 6730–6735, 2001.

ABOUT AUTHORS

[1] Ms. Revathy N had completed B.Sc.,Computer Science in the year 2000 and Master of Computer Applications(MCA) in the year 2003 under Bharathiar university. Completed M.phil., Computer Science from Alagappa university in the year 2005. Currently pursuing Ph.d., and the area of research is Neural Network. Other areas of interest are Mobile Computing, Data Mining and Artificial Intelligence At present working as an Assistant professor in the Department of Computer Applications at Karpagam College of Engineering at Coimbatore-32.

[2] Dr. R. Amalraj had completed MCA, and PhD and his area of interest in Artificial Neural networks, Data Mining. He had published a lot of papers in International and National level journals. At present working as an Associate Professor at Sri Vasavi College, Erode, India.