Selecting Informative Genes from Microarray Data for Cancer Classification with Genetic Programming Classifier Using K-Means Clustering and SNR Ranking

Supoj Hengpraprohm and Prabhas Chongstitvatana
Department of Computer Engineering, Chulalongkorn University, Thailand
supojn@yahoo.com; prabhas@chula.ac.th

1. Problem Description

Microarray technique is a popular method in bioinformatics. This technique presents gene expression data in a different environment in the same organism, or a different expression of the same gene for different organism. In addition, it can investigate thousands of genes simultaneously.

The microarray data presently consists of a small and high dimensional data. It is very complex and difficult to analyze. There are many methods to analyze such data – clustering, classification and feature selection [1].

Clustering technique is an unsupervised learning method. It is usually used to discover some novel knowledge from data, while a classification task is a supervised method that is used to predict any unseen data. Feature selection is a technique used to improve the performance of both clustering and classification task, especially the classification accuracy.

Recently, cancer classification is a major challenge in microarray data analysis. Many researchers use microarray technique to specify and identify cancer [2-9]. As a result, there are large volume of data, and many researchers using these data for classification and clustering automatically with many learning algorithms. Such researches aim to improve effectiveness of the model derived from learning algorithms [10-14].

One difficulty in analyzing microarray data is that it has high-dimension. Any learning algorithm that deals with high dimensional data will consume a large computational resource. Also, performance and efficiency of the model may be decreased due to noise in data. To alleviate these problems, dimension of data should be reduced by feature selection. There are many researches that study feature selection methods [15-20]. Such methods aim to rank features by some scoring metric or finding subset of features with respect to classifiers. However, features (genes) selected by scoring metrics may contain set of redundant features.

2. Research Questions

To improve the accuracy of prediction, this work proposed a method that select informative features, and maximize an effective model for a classifier. The feature selection step composed of two steps: K-Means clustering algorithm and SNR ranking of features to select informative gene for the classifier.

3. Research Method and Evidence

Eight data sets of cancer microarray data from Biomedical Data Analysis web site [21] were used to test the proposed method. The details of each data set are shown in TABLE I. The classifier used in this work was Genetic Programming Classifier (GPC) introduced in [14]. The features of data were clustered by K-Means clustering [22] and ranked by SNR method. The best score feature in each cluster was then selected. After that, the data with these features were tested by the GPC. The overall process is shown in Figure 1.

SNR (Signal-to-Noise Ratio) is a statistical method that measures effectiveness of feature in identifying a class out of another class. The signal-to-noise ratio is defined as follows:

$$F = \left| \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \right| \qquad (1)$$

where μ_1 and μ_2 denote the mean expression level for the samples in class 1 and class 2 respectively. σ_1 and σ_2 denote the standard deviation for the samples in each class.

To evaluate the performance of a classifier, we used a method known as round robin or leave – one – out method [23]. There are *N* records of data, *N-1* records are used as training set and one record is used as a test. We exchange a test data through *N* records and evaluate an equation in terms of its accuracy defined as follows:

$$Accuracy = \frac{TP + TN}{N} \qquad (2)$$

where *N* is the total number of tested cases, *TP* is a total number of affected subjects correctly classified, *TN* is a total number of normal subjects correctly classified, and *TP+TN* is the total number of subjects correctly classified.

TABLE I.
THE DETAILS OF DATA SET

Data Set	No. of Gene	No. of Instance (Class)
Leukemia	7,129	38 (27 ALLs, 11 AMLs)
Breast Cancer	24,481	78 (34 relapses, 44 non-relapses)
Central Nervous System Embryonal Tumors: CNS	7,129	60 (21 survivors, 39 failures)
Colon Cancer	2,000	62 (40 cancers, 22 normals)
Ovarian Cancer	15,154	253 (162 cancers, 91 normals)
Prostate Cancer	12,600	102 (52 cancers, 50 normal)
Lung Cancer	12,533	32 (16 MPMs, 16 ADCAs)
Lymphoma	4,026	47 (24 GCBs, 23 ACBs)

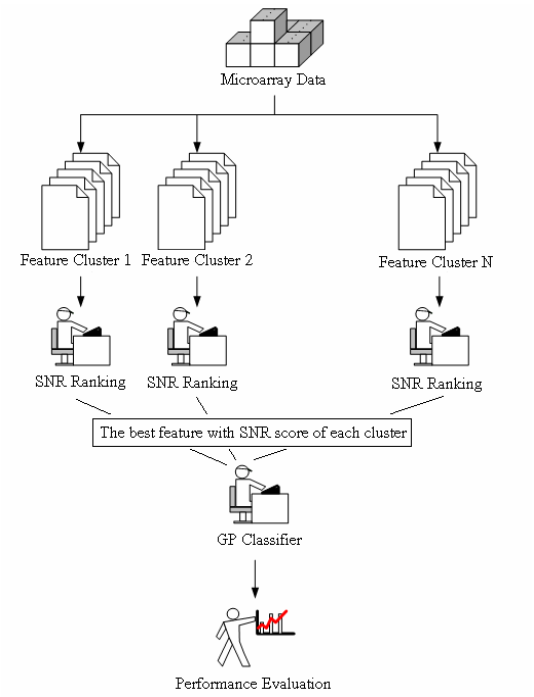

Figure 1: The overall process of the experiment

4. Outcome

In this paper, we applied SNR method with the best 30 features to all data sets described in TABLE I. The accuracy from three methods is compared: all genes, SNR and the proposed method (Clus). The result is reported from the average of 10 runs (using leave-one-out method, the total number of experiment in each data set is 10 multiply by the number of instance of the data set). The results are shown in Figure 2. (In this experiment, 30 clusters (K = 30) were used.)

The result shows that in some data set the SNR yields poorer performance against all genes such as Colon, Breast and Lung data set (see Figure 2 comparing between black and white bars). It indicates that using SNR method alone is inadequate.

By using K-Means clustering in conjunction with feature selection, genes expressing similarly are grouped together into the same cluster. After we applied SNR and selected a gene with the best SNR score in each cluster, we assure that selected genes have no redundant features. Therefore, a learning algorithm like GPC can use these features to obtain better performance than using all genes. For some data set such as Leukemia, Colon, Breast and Lung data set, the method proposed achieved the best performance.

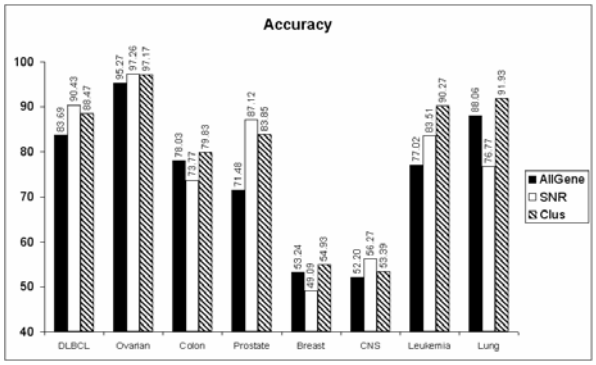

Figure 2: Comparing of GPC performance with SNR feature selection, all genes and the proposed method (Clus)

We compare the experimental results (Clus) with many feature selections and classifiers reported in [12-13] in 3 data sets (TABLE II). The feature selection methods are Pearson's and Spearman's correlation coefficients (PC, SC), Euclidean distance (ED), cosine coefficient (CC), information gain (IG), mutual information (MI) and signal to noise ratio (SNR). The classifiers are Multi-layer perceptron (MLP), k-nearest neighbour (KNN), support vector machine (SVM) and structure adaptive self–organizing map (SASOM).

In TABLE II, the values with highlight are better than our method. The comparison shows that most combination of feature selection and classifiers are good at only one or two data sets. Only k-nearest neighbour method using Pearson's coefficient correlation as the similarity measure (KNN Pearson) and using the information gain (IG) as feature selection shows better result than our method in all three data sets. This shows the promise of our proposal in selecting informative genes.

TABLE II.
Comparison of the accuracy of the proposed method with other methods. The value with highlight is better than our method.

Classifier	Feature Selection	Data Set		
		Leukemia	Colon	Lymphoma
MLP	PC	***97.1***	74.2	64.0
	SC	82.4	58.1	60.0
	ED	***91.2***	67.8	56.0
	CC	***94.1***	***83.9***	68.0
	IG	***97.1***	71.0	***92.0***
	MI	58.8	71.0	72.0
	SN	76.5	64.5	76.0
SASOM	PC	76.5	74.2	48.0
	SC	61.8	45.2	68.0
	ED	73.5	67.6	52.0
	CC	88.2	64.5	52.0
	IG	***91.2***	71.0	84.0
	MI	58.8	71.0	64.0
	SN	67.7	45.2	76.0
SVM (linear)	PC	79.4	64.5	56.0
	SC	58.8	64.5	44.0
	ED	70.6	64.5	56.0
	CC	85.3	64.5	56.0
	IG	***97.1***	71.0	***92.0***
	MI	58.8	71.0	64.0
	SN	58.8	64.5	72.0
SVM (RBF)	PC	79.4	64.5	60.0
	SC	58.8	64.5	44.0
	ED	70.6	64.5	56.0
	CC	85.3	64.5	56.0
	IG	***97.1***	71.0	***92.0***
	MI	58.8	71.0	64.0
	SN	58.8	64.5	76.0
KNN (Cosine)	PC	***97.1***	71.0	60.0
	SC	76.5	61.3	60.0
	ED	85.3	***83.9***	56.0
	CC	***91.2***	***80.7***	60.0
	IG	***94.1***	74.2	***92.0***
	MI	73.5	74.2	80.0
	SN	73.5	64.5	76.0
KNN (Pearson)	PC	***94.1***	77.4	76.0
	SC	82.4	67.7	60.0
	ED	82.4	***83.9***	68.0
	CC	***94.1***	***80.7***	72.0
	IG	***97.1***	***80.7***	***92.0***
	MI	73.5	***80.7***	64.0
	SN	73.5	71.0	80.0
Our Method (GPC+Clus)		90.3	79.8	88.5

5. References

[1] S. Raychaudhuri, et al., "Basic microarray analysis: grouping and feature reduction", *TRENDS in Biotechnology*, 19(5): 189 – 193, May 2001.

[2] A. A. Alizadeh, et al., "Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, 403: 503-511, 2000.

[3] T. R. Golub, et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, 286: 531-537, 1999.

[4] L. J. Van't Veer, et al., "Gene expression profiling predicts clinical outcome of breast cancer", *Nature*, 415: 530-536, 2002.

[5] S. L. Pomeroy, et al., "Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression", *Nature*, 415:436-442, 2002.

[6] U. Alon, et al., "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays", *Proceedings of National Academy of Sciences of the United States of American*, 96:6745-6750, 1999.

[7] E. F. Petricoin III, et al., "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer", *The Lancet*, 359:572-577, 2002.

[8] D. Singh, et al., "Gene Expression Correlates of Clinical Prostate Cancer Behavior", *Cancer Cell*, 1:203-209, 2002.

[9] G. J. Gordon, et al., "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma", *Cancer Research*, 62:4963-4967, 2002.

[10] F. Azuaje, "Making Genome Expression Data Meaningful: Prediction and Discovery of Classes of Cancer Through a Connectionist Learning Approach", *Proceeding of the IEEE Symposium on Bio-Informatics and Biomedical Engineering (BIBE)*, 2000, pp. 208 – 213.

[11] F. Azuaje, "A computational neural approach to support the discovery of gene function and classes of cancer", *IEEE Transaction on Biomedical Engineering*, 48: 332 – 339, 2001.

[12] C. Park and S.-B. Cho, "Evolutionary Ensemble Classifier for Lymphoma and Colon Cancer Classification", *the 2003 Congress on Evolutionary Computation (CEC)*, 4: 2378 – 2392, 2003.

[13] S.-B. Cho and H.-H. Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification", *the Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003*, 19: 189 – 198, 2003.

[14] S. Hengpraprohm and P. Chongstitvatana, "Diffuse Large B-Cell Lymphoma Classification Using Genetic Programming Classifier", *Proceeding of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2005.

[15] J.-H. Hong and S.-B. Cho, "Lymphoma Cancer Classification Using Genetic Programming with SNR Features", *Proceeding of Genetic Programming: 7th European Conference*, 3003: 78 – 88, 2004.
[16] D.K. Slonim et al., "Class prediction and discovery using gene expression data", *Proceedings of the 4th Annual International Conference on Computational Molecular Biology*, 2000, pp.263 – 272.
[17] 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 Network*, 2002, pp.198 – 203.
[18] S.L. Pomeroy et al., "Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression", *Nature*, 415: 436 – 442, 2002.
[19] C.-J. Huang and W.-C. Liao. "A Comparative Study of Feature Selection Methods for Probabilistic Neural Networks in Cancer Classification". *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003.
[20] S. Hengpraprohm and P. Chongstitvatana, "Discovering an Optimal Feature Set of Microarray Data for Cancer Classification Using Perceptron Learning Rule with SNR Ranking", *Proceeding of International Conference on Software Knowledge Information Management and Applications (SKIMA)*, Chiangmai, Thailand, 2006.
[21] http://sdmc.lit.org.sg/GEDatasets/
[22] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297, 1967.
[23] F. Tourassi and C. Floyd, "The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis", *Medical Decision Making*, 17: 186-192, 1997.