

Gene Selection for Cancer Classification Using Bootstrapped Genetic Algorithms and Support Vector Machines

Xue-wen Chen

Electrical and Computer Engineering Department, California State University, Northridge, CA 91330

Abstract

The gene expression data obtained from microarrays have shown useful in cancer classification. DNA microarray data have extremely high dimensionality compared to the small number of available samples. In this paper, we propose a novel system for selecting a set of genes for cancer classification. This system is based on a linear support vector machine and a genetic algorithm. To overcome the problem of the small size of training samples, bootstrap methods are combined into genetic search. Two databases are considered: the colon cancer database and the leukemia database. Our experimental results show that the proposed method is capable of finding genes that discriminate between normal cells and cancer cells and generalizes well.

1. Introduction

The advent of DNA microarray technology has made it possible to analyze thousands or tens of thousands of genes simultaneously [1-3]. This hybridization based technology revolutionizes the traditional ways in molecular biology and has found applications in many different areas such as gene discovery, disease diagnosis, and drug discovery.

In this paper, we present a novel approach to select a subset of genes from microarray data for cancer classification using genetic algorithms (GA) [4] and support vector machines (SVM) [5]. The proposed bootstrapped GA/SVM algorithm is very efficient for selecting sets of genes in very high dimensional feature spaces for classification problems.

2. The proposed method

Our goal is to select a subset of predictive genes whose expressions can distinguish samples from different classes (e.g., tumor cells versus normal cells). The proposed method is based on genetic algorithms and support vector machines: the effectiveness of a subset of expressed genes is evaluated in terms of its discrimination power by SVMs; GA is applied to search in the combinational space of feature subsets in parallel to

identify the best subsets; and bootstrapped data are created to overcome the problem of small number of training samples.

Consider a two-class classification problem, where the training set is described as

$$(y_1, \mathbf{x}_1), \dots, (y_m, \mathbf{x}_m), \quad \mathbf{x}_i \in R^n, y_i \in \{-1, +1\} \quad (1)$$

where y_i are class labels. SVMs find an optimal hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ that produces large margins (where \mathbf{w} is the n -dimensional vector perpendicular to the hyperplane and b is the bias).

We first generate k groups of the bootstrapped samples from original training set. The GA algorithm is run on each bootstrapped data set. For each GA procedure, the fitness function used to evaluate a candidate solution is defined as

$$f = \sum_{i=1}^m \frac{1}{2} [1 + \text{sign}(y_i(\mathbf{w} \cdot \mathbf{x}_i + b))], \quad (2)$$

which is related to the number of correctly classified training samples using SVMs.

For microarray data sets, the number of training samples is typically small. Thus, more than one subset of genes that correctly classify training samples may exist. With the bootstrapped GA/SVM methods, we can obtain many such subsets. Similarly to the strategy used in Li et al. [6], the frequency of each gene selected in these near-optimal solutions is assessed and important genes are expected to have a high frequency to be selected.

3. Experiment results

3.1. Identify genes: colon cancer dataset

The colon cancer dataset [7] contains gene expression information extracted from DNA microarrays. This microarray dataset is used to distinguish tumor and normal colon tissues. There are 62 tissue samples, of which 22 are normal and 40 are cancer tissues, each having 2000 genes with highest minimal intensity across the 62 tissues. The

data set was divided into a training set with 32 samples and a test set with 30 samples.

A total number of 5157 subsets of genes that correctly classify all training samples are obtained using our bootstrapped GA/SVM algorithms. Each subset consists of five genes. Genes are then ordered based on the number of occurrences with which genes are selected. Figure 1 shows the number of occurrences for each gene.

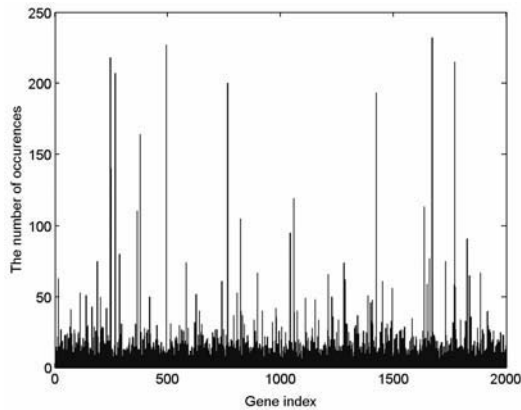


Figure 1: Genes and the number of their occurrences.

To test the discrimination ability of the genes selected, we classify the test samples using the top 25 genes. For comparison, we also includes results obtained with the combined forward selection and k nearest neighbor (FS/kNN) methods and the combined individual ranking and k nearest neighbor (IR/kNN) methods. With the IR/kNN-selected genes, six training samples are misclassified and 20 out of 30 test samples are misclassified. The FS/kNN algorithms produce better results than IR/kNN for training data set: all training samples are correctly classified. However, 20 out of 30 test samples are misclassified. Among the three algorithms, the proposed bootstrapped GA/SVM algorithm yields the best results: all training samples are correctly classified and only six out of 30 test samples are misclassified. This indicates that our bootstrapped GA/SVM algorithm is able to identify regulated genes that, when combined together, can discriminate cancer cells from normal cells.

3.2. Identify genes: leukemia dataset

The Leukemia dataset [8] consists of 72 samples, of which 47 are ALL and 25 are AML. The gene expression levels of all samples were extracted from microarray images. Each sample has 7129 features (i.e., 7129 genes). A dataset of 1800 genes is available after preprocessing. The data set was divided into a training set with 38 samples and a test set with 34 samples.

A total number of 2368 subsets of five genes that correctly classify all training samples were obtained using our GA/SVM algorithms. Similarly with colon cancer database, some genes were selected significantly more often than other genes. For classification, our results with the top 25 genes show that features selected by individual ranking methods perform poorly: six out of 38 training samples and 14 out of 34 test samples are misclassified. Forward selection performs better than individual ranking. Features selected from our GA/SVMs correctly classify all training samples and only misclassify one test sample.

4. Conclusions

The processing and exploitation of useful information from microarray gene data sets pose a challenging problem. In this paper, we propose a practical and efficient feature selection algorithm to select informative genes from very high-dimensional spaces. We perform the bootstrapped GA based gene selection in the context of SVMs. Thus, the selected genes are expected to generalize well. Experiment results demonstrate that the proposed bootstrapped GA/SVM algorithms are well suited for feature selection problems.

5. References

- [1] De Risi, J., Iyer, V., and Brown, P., "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, **278**, 680-686, 1997.
- [2] Cho, R., Campbell, J., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., and Lockart, D., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell*, **2**, 65-73, 1998.
- [3] Chu, S., Derisi, J., Eisen, M., Mullholland, J., Botstein, D., Brown, P., and Herskowitz, I., "The transcriptional program of sporulation in budding yeast," *Science*, **282**, 699-705, 1998.
- [4] Goldberg, D., *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison Wesley, 1989.
- [5] Vapnik, V., *Statistical Learning Theory*. Wiley, New York, 1998.
- [6] Li, L., Darden, T., Weinberg, C., Levine, A., and Pederson, L., "Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method," *Combinational Chemistry and High Throughput Screening*, vol. 4(8), pp. 727-739, 2001.
- [7] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A., "Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA*, **96**, 6745-6750, 1999.
- [8] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E., "Molecular classification of cancer: class discovery and class prediction by gene expression," *Science*, **286**, 531-537, 1999.