Cancer classification based on gene expression profile

Among the most powerful and versatile tools for functional genomic studies are high-density DNA microarrays. The main advantage of microarray data is that it is able to allow biologists to simultaneously monitor the expression of thousands of genes as to obtain quantitative information about the complete gene expression profiles. With the wealth of gene expression data from microarrays, prediction, classification and clustering techniques are used for analysis and interpretation of the data. One of the most important applications of DNA microarray techniques is cancer classification. This paper will summarize this emerging field, giving particular attention to how to select a fixed subset of genes as predictors for classifying human cancers and propose a new approach for achieving it.

Overview of tumor classification

Cancers are major challenges to the health of human populations, yet the techniques used to diagnose these diseases have changed little in decades [1]. New approaches have emerged from the collaborative efforts of biologists, physicians, mathematicians and computer scientists involved in DNA microarray-based research into these diseases. More accurate disease diagnosis and improvements in therapy will be the clinical benefit.

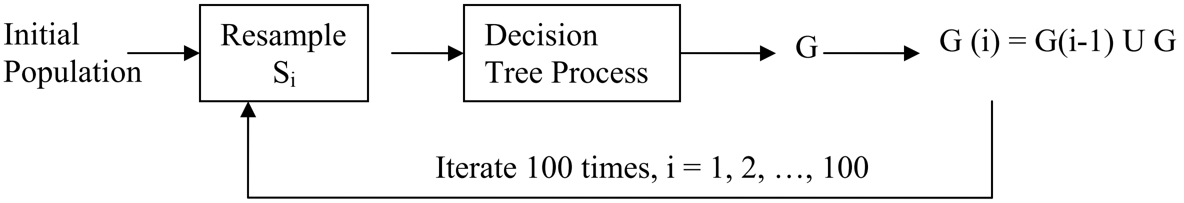
One of the most important applications of microarray data is cancer classification. Currently, cancer classification techniques rely on highly subjective judgments of tumor histology by pathologists, however, the strongest predictors for cancer sometimes fail to classify accurate tumors according to their clinical behavior [2]. Multiple studies have demonstrated that DNA microarrays are useful for classifying human cancers and have revealed that expression profiles are valuable both in cancer diagnosis and prognosis [3, 4]. This global quantitative approach to classifying cancers and predicting outcomes will become a valuable clinical tool for pathologists and oncologists. DNA microarray analysis allows physicians to follow the progression of disease even when there is no histological evidence of change. The information obtained through cancer classification will almost certainly contain valuable clues to cancer mechanisms and inspire new tactics to combat these diseases [1].

The approaches that are being used to classify cancer with DNA microarray expression profiles underscore the importance of collaboration between biology and computer science. Some important recent applications are in molecular classification of acute leukemia [3], cluster analysis to tumor and normal colon tissues [5], clustering and classification of human cancer cell lines [6], diffuse large B-cell lymphoma [4] and human mammary epithelial cells and breast cancer [7, 8, 9]. These techniques have also helped to identify previously undetected subtypes of cancer. However, we can see that DNA microarray data usually monitor thousands of genes expression per sample. This ability has resulted in data with the number of variables p (genes) far exceeding the number of samples N. Standard statistical methodologies in classification and prediction do not work well or even at all when $N < p$. In addition, a large number of genes increases the dimensionality, computational complexity and cost of data analysis, and introduces some undesired noise. In the clinical setting of testing or implementing a set of prognostic markers(genes), it would not be feasible to accurately measure and standardize measurement of an entire gene set in large numbers of patients. To improve the classification accuracy, an effective tool in machine learning is feature selection, that is, reduce the high-dimensional gene space to a lower dimensional gene component space. Feature selection is a process of selecting an optimal subset of features from a possibly enormous set of potentially useful features, for use in classifiers. The selected subset will provide an optimal separation of a population of patients into two or more groups with major difference in prognosis.

Computational approaches to feature selection

In recent years, some effort has been paid to discover analysis procedure for feature selection. PCA is one common method to achieve it, it tries to find the directions of greatest variance implied by the correlation matrix and to then ‘visualize’ the data in terms of their projection on these directions [10]. However, because PCA is an unsupervised learning technique, some researcher found that PCA does not take into account the class labels of the training set from microarray data, it is not reliable and cannot give better generalization [11]. Recently, Nguyen D. *et al.* used partial least square for dimension reduction that is superior to PCA in some conditions [12], also, Li W. and Xiong M. introduced a method combining Fisher's linear discriminant analysis and feature slelection and the results demonstrated that using only a subset of genes ranging from 3 to 10 can achieve high classification accuracy [13]. The details of these approaches will not be described here; instead, I will propose a new two-step approach to achieve feature selection in this paper. Briefly, this approach applies the random forest method as a first-step feature reduction method, and then it will apply the genetic algorithm to further select the features.

The random forest algorithm is a supervised feature space reduction method, which takes into account the class labels of the training set. The schema of this method is shown as follows:



where G(0) = Ø, and the threshold of classification accuracy of the decision tree is 80%.

After performing the random forest method to reduce the feature space, the genetic algorithm will be applied to further refine the features. The genetic algorithm is inspired by Darwin's theory about evolution. Basically, a population of potential solutions is set up at random. Each member of the population is encoded as a chromosome; the population of chromosomes is iteratively optimized. At each step, a point mutation may occur in a chromosome, or two chromosomes may *mate* to give a new offspring, if the end condition is satisfied, we stop iteration and return the best solution in current population. The general algorithm can be specified in the following steps for the feature selection:

1. Create an initial population $P=\{g_1, \ldots, g_N\}$
2. Evaluation: evaluate the fitness, $F(X_i)$ for each of the individuals in the population with a evaluate function, commonly used methods are correlation-based feature subset selection(CSE), decision tree, etc.
3. Compute the average fitness for the population, F_{avg}
4. Assign each individual the normalized fitness $F(g_i)/ F_{avg.}$
5. Assign each individual g_i a probability p_i proportional to its normalized fitness. Using this distribution, select N vectors from P to construct a subset S.

6. Pair all of the vectors in S at random forming N/2 pairs as parents.
7. Apply crossover with probability p_{cross} to each pair in S
8. Apply mutation with probability $p_{mutation}$ to some pairs in S
9. Check termination conditions. Terminate if solution achieved.
10. Otherwise, goto Step 2.

As we may obtain a fairly lower dimensional feature space (a subset of genes ranging from 3 to 10 is expected) by applying this approach, we can use neural network to perform the cancer classification and the task can terminate in a relatively short period of time. A feed forward multilayer neural network can be created applying back-propagation learning algorithm for performing classification based on the selected features. A commonly used statistical approach 10-fold cross-validation can be used to evaluate the classification accuracy.

Conclusions

There are a number of classification algorithms that can be used for DNA microarray data. Such qualitative properties can include diagnostic and prognostic indications. However, there is no a single algorithm that is suitable for all test cases. For genetic algorithm, it has the advantage of achieving parallelism and avoiding local optimization, besides, it is relatively easy to implement. Here, I propose a new approach that focuses on the feature selection problem to find a core subset of genes and then employs a feed forward neural network to perform the classification based on the selected features. After the method is implemented, if preliminary results suggest that the approach is feasible, and if clinical relevant partitions of the data set are found, further work aimed at the extension and enhancement of the method can be proceed.

References

1. Young R. (2000). Biomedical discovery with DNA arrays. *Cell* **102**: 9-15.

2. Goldhirsch A., Glick J., Gelber R., and Senn H. (2001). Adjuvant therapy for breast cancer. *J. Natl Cancer Inst.* **93**: 979-989.

3. Golub T. *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531-7.

4. Alizadeh A. *et al*. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503-511.

5. Alon U. *et al.* (2000). 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.

6. Ross D. *et al.* (2000). Sysmatic variation in gene expression patterns in human cancer cell lines. *Nature Genetics.* **24**: 227-235.

7. Perou C. *et al.* (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancer. *Proc. Natl. Acad. Sci. USA.* **96**: 9112-9217.

8. Perou C. *et al.* (2000). Molecular portrait of human breast tumors. *Natur*e **406**: 747-752.

9. Veer L. *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**: 530-535.

10. Raychaudhuri S. *et al*. (2000). Principal component of analysis to summarize micorarray experiments. *Pac Sym Biocomput.* 455-466.

11. Model F, Adorjan P, Olek A and Piepenbrock C. (2001) Feature selection for DNA methylation based cancer classification, *Bioinformatics*, 17, S157-S164

12. Nguyen D. and Rocke D. (2002). Tumor classification by partial least sure using microarray gene expression data. *Bioinformatics* **18:** 39-50.

13. Li W. and Xiong M. Tclass: tumor classification system based on gene expression profile. *Bioinformatics* **18:** 325-326.