A Hybrid Feature Selection Approach for Microarray Gene Expression Data

Feng Tan, Xuezheng Fu, Hao Wang, Yanqing Zhang, and Anu Bourgeois

Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA
{ftan, xfu1, hwang10}@student.gsu.edu,
{yzhang, anu}@cs.gsu.edu

Abstract. Due to the huge number of genes and comparatively small number of samples from microarray gene expression data, accurate classification of diseases becomes challenging. Feature selection techniques can improve the classification accuracy by removing irrelevant and redundant genes. However, the performance of different feature selection algorithms based on different theoretic arguments varies even when they are applied to the same data set. In this paper, we propose a hybrid approach to combine useful outcomes from different feature selection methods through a genetic algorithm. The experimental results demonstrate that our approach can achieve better classification accuracy with a smaller gene subset than each individual feature selection algorithm does.

1 Introduction

As one of the important recent breakthroughs in the field of experimental molecular biology, microarray technology allows scientists to monitor changes in the expression levels of genes in response to changes in environmental conditions or in healthy versus diseased cells. It increases the possibility of disease classification and diagnosis at the gene expression level. However, due to a huge number of genes (features) and comparatively small number of samples, biologists are often frustrated by microarray gene expression data when they are trying to search for meaning from it. It is important for diagnostic tests to be able to remove redundant and irrelevant genes and find a subset of discriminative genes. Selecting informative genes is crucial to improving the diagnostic/classification accuracy [2, 7].

Many methods have been proposed for feature selection. Among them, feature ranking techniques are particularly attractive because of their simplicity, scalability, and good empirical success. A set of top ranked features can be selected for data analysis or building a classifier. Some feature ranking methods use information-theoretic functions, such as χ^2 -statistics [4], T-statistics [2], MIT correlation [5], information gain [1], and entropy-based measure [6]. Some other approaches utilize Support Vector Machines (SVMs) for feature ranking/selection [7, 8].

However, different feature selection methods result in different gene rankings or gene subsets. It is hard to decide which feature selection algorithm is best fit for a data set because the performance of an algorithm varies with different data sets. Our work

in this paper is to effectively combine information from various feature selection methods for more reliable classification. We propose a hybrid approach that combines valuable information from multiple feature selection methods through a genetic algorithm (GA). To evaluate our method, several feature selection algorithms, that is, entropy-based [6], T-statistics, and SVM-RFE (Recursive Feature Elimination) [7] are used to provide candidate features for the GA. We test our approach on two microarray data sets (Colon Cancer and Prostate Cancer). Experimental results show that our approach is effective and efficient in finding subsets of informative genes for reliable classification.

The rest of the paper is organized as follows. Section 2 describes our hybrid feature selection method. Section 3 introduces three existing feature selection algorithms used in later experiments. Section 4 compares our method with three existing feature selection algorithms on two data sets. Conclusion is drawn in Section 5.

2 Method

The idea of our hybrid approach (shown in Fig.1.) is to absorb useful information from different feature selection algorithms to find better feature subsets that can have smaller size or better classification performance than those individual algorithms. We use a genetic algorithm (GA) to fuse multiple feature selection criteria to accomplish this goal.

Gene/Feature Pool. The feature pool is a collection of candidate features to be selected by the genetic algorithm to find a feature subset. Instead of using all features from the original data, sets of features selected by multiple feature selection algorithms are input to the pool.

Representation of Feature Subset. Each feature subset (an individual) is encoded by n -bit binary vectors. The bits with value 1 in a vector represent the corresponding features being selected, while the bits with value 0 mean the opposite.

Induction Algorithm. The genetic algorithm is independent of the inductive learning algorithm used by the classifier. Multiple induction algorithms, such as Naïve Bayes, artificial neural network, and decision trees can be flexibly incorporated into our method. In this paper, we use SVM classifier [9] in our experiments.

Fitness Function. Our genetic algorithm is designed to maximize classification accuracy of the chosen classifier. So the accuracy obtained from the induction algorithm is used to evaluate each individual in a population.

Genetic Operators

- (1) Selection: Roulette wheel selection is used to probabilistically select individuals from a population for later breeding.
- (2) Crossover: We use single-point crossover operator. The crossover point i is chosen at random.
- (3) Mutation: Each individual has a probability p_m to mutate. We randomly choose a number of n bits to be flipped in every mutation stage.

Fig. 1. Hybrid feature selection with genetic algorithm

3 Feature Selection Methods

To evaluate our method, three feature selection algorithms (i.e. entropy-based, T-statistics and SVM-RFE) are applied to microarray data sets. Then a number of top-ranked features from each algorithm are selected to form the feature pool for the GA. The idea of entropy-based method [6] is that removing an irrelevant feature would reduce the entropy more than that for a relevant feature. The algorithm ranks the features in descending order of relevance by finding the descending order of the entropy after removing each feature one at a time. T-statistics is a classical feature selection approach [2]. Each sample is labeled with $\{1, -1\}$. For each feature f_j , the mean μ_j^1 (resp. μ_j^{-1}) and standard deviation δ_j^1 (resp. δ_j^{-1}) are calculated using only the samples labeled 1 (resp. -1). Then a score $T(f_j)$ can be obtained by the equation defined in [2]. When making a selection, those features with the highest scores are considered as the most discriminatory features. Guyon et al. [7] proposed SVM-RFE, a backward feature elimination algorithm that removes one “worst” gene (i.e., the one changed the

objective/cost function J least after being removed) at one time. w_i^2 is taken as the ranking criterion in SVM-RFE.

4 Experiments

Leave-One-Out (LOO) cross-validation is used in Experiment 1 to obtain an estimate of the classification accuracy. Since a testing data set is available in Experiment 2, five-fold cross-validation is used to estimate the training accuracy. We use SVM with linear kernel as the classifier in both experiments. However, it is flexible to apply multiple induction algorithms to our hybrid method. Our experiments are implemented in a PC with Pentium 4(2.64MHz) and 512M RAM. All algorithms are coded in C++ and Matlab R14.

4.1 Experiment 1

Colon cancer data set [13] contains the expression of the 2000 genes across the 62 tissues/samples (40 tumor and 22 normal) collected from colon-cancer patients.

The experimental results of entropy, T-statistics, and SVM-RFE on Colon cancer data are presented in Table 1. We can see that in general, SVM-RFE gives the best performance among the three.

Table 1. LOO accuracy of entropy-based, t-statistics, and SVM_RFE on colon cancer data

Top Features	Entropy (%)	T-statistics (%)	SVM-RFE (%)
2	65.52	84.48	75.86
4	65.52	86.21	89.66
8	65.52	86.21	96.55
16	65.52	87.93	98.28
32	65.52	89.66	96.55
64	60.34	89.66	94.83
128	68.97	89.66	93.10
256	81.03	87.93	91.38
512	84.48	82.76	86.21
1024	82.76	81.03	84.48
2000	81.03	81.03	79.31

Table 2. Top-20 features from entropy-based, t-statistics, and SVM-RFE on colon cancer data

Feature Selection Algorithms	Top-20 Features
Entropy-based	169,1451,1430,1538,1660,375,1277,1150,445,1697,761,1170,825,609,1055,603,1882,1910,1341,808
T-statistics	493,765,377,1423,249,245,267,66,14,1772,625,822,622,411,137,1674,1771,111,1582,513
SVM-RFE	350,164,14,1378,43,1976,1325,353,44,16,250,175,159,115,458,24,988,47,33,30

Table 3. GA experiments on colon cancer data

Gene/Feature Pool			Feature Subsets Selected by GA	LOO Acc. (%)
Entropy-based	T-statistics	SVM-RFE		
Top 2 (169, 1451)	Top 2 (493,765)	Top 2 (350,164)	5 (164,169,350,493,765)	89.66
Top 2 (169, 1451)	Top 4 (493,765, 377,1423)	Top 4 (350,164,14, 1378)	6 (14,164,350,1378,1423,1451)	98.28
Top 4 (169, 1451,1430, 1538)	Top 4 (493,765, 377,1423)	Top 4 (350,164,14, 1378)	10 (14,164,350,377,765,1378, 1423,1430,1451, 1538)	96.56
Top 4 (169,1451,1430,1538)	Top 8 (493,765,377,1423,249,245,267, 66)	Top 8 (350,164,14,1378, 43,1976,1325,353)	12 (14,43,66,164,245,267,350,493,765,1325,1430,1451)	98.28

The numbers in bold are the genes selected by the GA from the feature pool.

The Top-20 features given by the three algorithms on Colon data set are presented in Table 2. Only one gene/feature with column ID 14 is shared by T-statistics and SVM-RFE and entropy-based method has no common feature with the other two algorithms in Top-20 features. Besides the top 20 features, we notice that the ranks of features are very different in the three algorithms.

Table 3 shows the experimental results of applying our algorithm to Colon cancer data. To show the robustness of GA approach, we test several feature pools each of which contains a different number of top-ranked features chosen from the three methods. Our feature pools are relatively small because we assume a smaller informative gene subset (e.g., no more than 20 genes) is usually preferred for data analysis for a given accuracy due to the cost of performing the necessary clinical test and analysis. Features finally selected by the genetic algorithm are highlighted. In this experiment, the genetic algorithm is terminated with a maximum of 10 iterations. The size of the population is 30. It is possible to achieve better results if more iterations or a larger population size are allowed. As Table 3 shows, the genetic algorithm is capable of selecting a subset of 6 genes and achieves 98.28% accuracy while SVM-RFE needs 16 genes to get the same accuracy. In another case, our approach finds a subset of 12 genes from a different gene pool, which also reaches 98.28% accuracy.

4.2 Experiment 2

We further test the T-statistics, SVM-RFE and our approach on Prostate cancer data set [14]. Entropy-based algorithm is not used on this data set for the reason of time consumption. This data set consists of training data and testing data. The training data contain 52 prostate tumor samples and 50 non-tumor (normal) prostate samples with 12600 genes. The testing data contain 34 samples (25 tumor and 9 normal samples) obtained from a different experiment. In this experiment, feature selection algorithms first select features from training data. Then SVM classifies the testing data based on these selected features.

Table 4. Training and testing accuracy of t-statistic and SVM-RFE on prostate cancer data

Features	Training Accuracy (%)		Testing Accuracy (%)	
	T-statistics	SVM-RFE	T-statistics	SVM-RFE
2	76.47	84.31	97.06	73.53
4	78.43	86.27	97.06	70.59
8	86.27	96.08	88.24	73.53
16	83.33	100.00	88.24	85.29
32	89.22	100.00	88.24	94.12
64	90.20	100.00	76.47	91.18
128	91.18	99.02	91.18	91.18
256	93.14	95.10	82.35	91.18
512	93.14	95.10	82.35	91.18
1024	91.18	94.12	85.29	94.12
2048	91.18	93.14	88.24	94.12
4096	89.22	92.16	94.12	94.12
8192	90.20	91.18	97.06	94.12
12600	89.22	91.18	97.06	94.12

Table 5. Top-20 features from t-statistics and SVM-RFE on prostate cancer data

Feature Selection Algorithms	Top-20 Features
T-statistics	6185,10138,3879,7520,4365,9050, 205 ,5654,3649, 12153 ,3794,9172,9850,8136,7768,5462,12148,9034,4833,8965
SVM-RFE	10234, 12153 ,8594,9728,11730, 205 ,11091,10484,12495,49,12505,10694,1674,7079,2515,11942,8058,8658,8603,7826

Table 4. presents the results of T-statistics and SVM-RFE on Prostate cancer data. SVM-RFE again performs better than T-statistics in most of cases. However, its testing accuracy is much lower than that of T-statistics when using top 2 and top 4 features. Among the Top-20 features ranked by the two methods, there are only two common genes as shown in Table 5.

The results of our algorithm on the Prostate cancer data are presented in Table 6. Since this data set is relatively large with 12600 features, we terminate the genetic algorithm with 5 iterations. We adopt five-fold validation for training accuracy. For all of the three feature pools in Table 6, our approach can obtain the best testing accuracy (94.12%) with much smaller feature subsets compared to SVM-RFE. However, the testing accuracy is not as good as the one achieved by Top-2 or Top-4 genes selected by T-statistics. In the first case, our approach selects a subset of 3 genes which can achieve 93.14% training accuracy, while SVM-RFE and T-Statistics cannot even reach the same training accuracy with 4 genes. In the second case, a subset of 4 genes selected by our GA achieves 95.10% training accuracy, which is higher than SVM-RFE and T-statistics with the same size of gene subset. In the last case, we select a

subset of 8 genes, which achieves the same training accuracy as SVM-RFE does, but with higher testing accuracy.

Table 6. GA experiments on prostate cancer data

Gene/Feature Pool		Feature Subsets Selected by GA	Training Acc. (%)	Testing Acc. (%)
T-statistics	SVM-RFE			
Top2 (6185 10138)	Top2 (10234 ,12153)	3 (6185,10138,10234)	93.14	94.12
Top4 (6185 ,10138, 3879 ,7520)	Top 4 (10234 ,12153, 8594 , 9728)	4 (3879,6185,8594, 10234)	95.10	94.12
Top8 (6185,10138, 3879,7520,4365 9050, 205 ,5654)	Top 8 (10234 , 12153 , 8594 , 9728 , 11730 , 2 05 , 11091 , 10484)	8 (205,8594,9728, 10234,10484,11091,11 730,12153)	96.08	94.12

The numbers in bold are the genes selected by the GA from the feature pool.

5 Conclusions

In this paper, we present a hybrid feature selection approach for microarray gene expression data. We use a genetic algorithm to combine valuable outcomes of multiple feature selection algorithms to try to find subsets of informative genes. We compare our method with entropy-based, T-statistics, and SVM-RFE on two data sets. Our method is able to robustly find a small-sized feature set for reliable classification. The experimental results show that our method is efficient for microarray gene selection and classification. In our experiments, we randomly choose top features to form the feature pool. However, it can be seen that top features cannot guarantee better classification. How to choose features from the outcomes of various algorithms is an important problem to be solved. The genetic algorithm can also be improved. We leave it for future work.

References

1. Liu,Y., A comparative study on feature selection methods for drug discovery, Journal of Chemical Information and Computer Sciences 44(5): 1823-1828,2004
2. Liu, H., J. Li, L. Wong, A comparative study on feature selection and classification methods using gene expression profiles and proteomic pattern, Genomic Informatics, 13, 51-60, 2002
3. J. Weston et al., Feature selection for SVMs, Advances in Neural Information Processing Systems 13 , 2000
4. H. Liu, R. Setiono, Chi2: Feature Selection and Discretization of Numeric Attributes, Proc. IEEE 7th International Conference on Tools with artificial Intelligence, 338-391, 1995
5. Golub, T.R. et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science, 286, 1999
6. Dash, M., Liu, H., Handling Large Unsupervised Data via Dimensionality Reduction, ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 1999

7. Guyon, I., Weston, J., Barnhill, S., Vapnik, V., Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, Vol. 46, No. 1-3, pp. 389-422, 2002
8. Weston, J., S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio and V. Vapnik: Feature Selection for SVMs. *Advances in Neural Information Processing Systems* 13, 2000
9. Burges, C.J.C., A tutorial on support vector machines for pattern recognition, *Data mining and Knowledge Discovery*, 2(2), 121-167, 1998
10. LeCun, Y., J.S. Denker, S.A.Solla, Optimum Brain Damage, *Advances in Neural Information Processing Systems II*, D.S.Touretzky, Ed. Mateo, CA: Morgan Kaufmann, 1990
11. W. S. Noble, Support vector machine applications in computational biology, *Kernel Methods in Computational Biology*. B. Schoelkopf, KTsuda and J.-P. Vert, ed. MIT Press, 71-92, 2004
12. B. Schölkopf, I. Guyon, and J. Weston, *Statistical Learning and Kernel Methods in Bioinformatics, Artificial Intelligence and Heuristic Methods in Bioinformatics* 183, (Eds.) P. Frasconi und R. Shamir, IOS Press, Amsterdam, The Netherlands, 1-21, 2003
13. Alon, U., et al. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor a Normal Colon Tissues Probed by Oligonucleotide Arrays, *PNAS*, 96:6745-6750, 1999
14. Dinesh Singh, et al. Gene Expression Correlates of Clinical Prostate Cancer Behavior. *Cancer Cell*, 1:203-209, March, 2002