Recursive Feature Addition for Gene Selection

Qingzhong Liu, Student Member, IEEE, Andrew H. Sung

Abstract—Class prediction and feature selection are two learning tasks that are strictly paired in the search of molecular profiles from microarray data [21]. In this paper, we present a scheme of recursive feature addition for gene selection combining classifiers for the purpose of classifying tumor tissues using DNA microarray data. Based on the highest train accuracy, the next gene is added into the feature set according to the measures of the correlation/mutual information between chosen genes and candidate genes. In comparison with the well-known gene selection methods of T-TEST and SVM-RFE using different classifiers, our method, on the average, performs the best regarding the classification accuracy under different feature dimensions.

I. INTRODUCTION

DNA microarrays are capable of profiling gene expression patterns of tens of thousands of genes in a single experiment. DNA targets are arrayed onto glass slides (or membranes) and explored with fluorescent or radioactively labeled probes [1]. Obtaining gene expression data from cancerous tissues gives insight into the gene expression variation of various tumor types, thus providing clues for cancer classification of individual samples. One of the key challenges of microarray studies is to derive biological insights from the unprecedented quantities of data on gene expression patterns. Partitioning genes into closely related groups has become an element of practically all analyses of microarray data [2]. However, identification of genes is faced with many challenges. The main challenge is the overwhelming number of genes compared to the smaller number of available training samples. In machine learning terminology, these data sets have high dimension and small sample size. Many of these genes are irrelevant to the distinction of samples. These irrelevant genes have a negative effect on the accuracies of the classifier. Another challenge is that DNA array data contain technical and biological noises. Thus, it is critical to identify a subset of informative genes from a large data pool that will give higher classification accuracy.

Machine learning techniques have been successfully applied to cancer classification using microarray data [3]. One of the earliest was a hierarchical algorithm developed by Eisen et al. [4]. Other popular algorithms, such as K-Nearest Neighbor (KNN) and Self-Organizing Maps (SOM) have also been widely used [5, 6].

Because DNA microarray data has high dimension and small samples, the gene selection is very important to the classification accuracy. T-TEST [7, 8] is one well-known gene selection in DNA microarray analysis. It ranks the significant genes according to the p-values. Unfortunately, it just considers the individual gene, not the interaction of genes. Because test-statistic gene selections just consider the individual gene data, the problem probably is that we might end up with many highly correlated genes. If there is a limit on the number of genes to choose we might not be able to include all informative genes. The method in [9] is to retrieve groups of similar genes, and apply a test-statistic to select genes of interest. SVM-RFE (Support Vector Machine Recursive Feature Elimination) is another well-known gene selection which refines the optimum feature set by using SVM in a wrapper approach [10]. The paper [11] presents a feature selection to achieve the max-dependency, max-relevance, and min-redundancy based on mutual information.

From the perspective of machine learning, we present a recursive feature addition based on the highest train accuracy and the measures of the correlation and mutual information between chosen genes and candidates. Five strategies are proposed. In comparison with the well-known gene selections, T-TEST and SVM-RFE using different classifiers, on the average, our method is the best regarding the classification accuracy under different feature dimensions.

II. RECURSIVE FEATURE ADDITION

DNA microarray data has high feature dimensions, choosing the genes with complete dialogistic information and minimum redundant information is of great value for clinical dialogistic and cancer research. On the other hand, the biological interaction of genes is complicated. From the perspective of machine learning, our scheme for gene selection combined with classifiers is presented as follows:

1. Insignificant or noise microarray gene data is filtered out according to test-statistical selection.

2. Each individual gene is ranked in the order from the highest classification accuracy to the lowest classification accuracy with specific classifier.

3. The gene with the highest classification accuracy is chosen as the most important feature, or the first feature. If there are many genes with the same highest classification, the most important feature is set to the gene with the lowest p-value measured by test-statistic. At this point the chosen
feature set, \(G_1\), consists of the first feature, \(g_1\), which corresponds to feature dimension one.

4. The \(N+1\) – dimension feature set, \(G_{N+1} = \{g_1, g_2, \ldots, g_N, g_{N+1}\}\) is produced by adding \(g_{N+1}\) into the chosen \(N\)-dimension feature set, \(G_N = \{g_1, g_2, \ldots, g_N\}\). The \(g_{N+1}\) is chosen according to the following method.

Each gene \(g_f (f \neq 1, 2, \ldots, N)\) outside of \(G_N\) is added into \(G_N\); the classification accuracy of each feature set \(G_N + \{g_f\}\) is compared, the \(g_f\) with the highest classification accuracy is marked and put into the set of candidates, \(C\). Generally, the set of candidates consists of many genes, but only one gene will be the final winner, or \(g_{N+1}\). Five strategies are designed for choosing the winner, or \(g_{N+1}\) from the set of candidates:

The first strategy is to compare the individual classification accuracy of the candidates. The candidates with the highest accuracy enter into the final stunt; the lowest p-value in the final stunt is the final winner, or \(g_{N+1}\).

The second strategy is to measure the similarity of chosen genes and candidate genes. Pearson’s correlation [22] between the candidate \(g_c\) and \(g_n\) in \(C\) and the elements, \(g_n, g_n \in G_N (n = 1, 2, \ldots, N)\), is calculated. The sum of the square of the correlation (SC) is calculated to measure the relation, defined as follows:

\[
SC(g_c) = \sum_{n=1}^{N} \cos^2(g_c, g_n) \tag{1}
\]

Where, \(g_c \in C\), \(g_n \in G_N\), \(n = 1, 2, \ldots, N\). The \(g_c\) with the minimal value of \(SC(g_c)\) is chosen as \(g_{N+1}\).

The third strategy is to calculate the maximum value of the square of the correlation (MC),

\[
MC(g_c) = \max (\cos^2(g_c, g_n)) \tag{2}
\]

Where, \(g_c \in C\), \(g_n \in G_N\), \(n = 1, 2, \ldots, N\). The \(g_c\) with the minimum value of \(MC(g_c)\) is chosen as \(g_{N+1}\).

The forth strategy is same to the second strategy except that the relation of chosen genes and candidate genes is measured by the mutual information [23], not correlation. The sum of the square of the mutual information (SI) is calculated as follows.

\[
SI(g_c) = \sum_{n=1}^{N} I^2(g_c, g_n) \tag{3}
\]

Where, \(I(g_c, g_n) = \sum_{g \in G_n} \rho(g_c, g_n) \log \frac{\rho(g_c, g_n)}{\rho(g_c) \rho(g_n)}\). The \(g_c\) with the minimum value of \(SI(g_c)\) is chosen as \(g_{N+1}\).

The fifth strategy is to calculate the maximum value of the square of the mutual information (MI),

\[
MI(g_c) = \max (I^2(g_c, g_n)) \tag{4}
\]

Where, \(g_c \in C\), \(g_n \in G_N\), \(n = 1, 2, \ldots, N\). The \(g_c\) with the minimum value of \(MI(g_c)\) is chosen as \(g_{N+1}\).

In the methods mentioned above, a feature is recursively added into the set using a specific classifier. Similar to the name of SVM-RFE, we call the five strategies Classifier-Recursive Feature Addition (C-RFA), Classifier Minimal Sum of the square of Correlation-Recursive Feature Addition (CMSC-RFA), Classifier Minimal Maximal value of the square of Correlation-Recursive Feature Addition (CMMC-RFA), Classifier Minimal Sum of the square of Mutual Information-Recursive Feature Addition (CMSMI-RFA), Classifier Minim-maximal value of the square of Mutual Information-Recursive Feature Addition (CMMI-RFA), respectively. The first ‘C’ denotes classifier. For example, if the classifier is Naïve Bayes Classifier (NBC), we record the five strategies NBC-RFA, NBCMSC-RFA, NBCMMC-RFA, NBCMSMI-RFA, and NBCMMI-RFA, respectively.

III. EXPERIMENTS ON GENE SELECTIONS

A. Data Sets

The following benchmark datasets are available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.


2) The LYMPHOMA data set consists of 58 diffuse large B-cell lymphoma (DLBCL) samples and 19 follicular lymphoma (FL) samples. It was studied in [13]. The data file, lymphoma_8_lbc_fsc2_rr.nnres, and the class label file, lymphoma_8_lbc_fsc2.cls, are used in our experiments for identifying DLBCL and FL.

3) The PROSTATE data set in [14] contains 52 prostate tumor samples and 50 non-tumor prostate samples.

B. Experimental Setup

Our experiments are described as follows:

1. The training samples are chosen from the datasets at random. The remaining samples are used for test. The ratio of training sample to test sample is 3:2 in the same class.

2. The gene selections, T-TEST, SVM-RFE, and the Recursive Feature Additions are done for gene selection based on the training samples. Different gene sets are produced under the feature dimension 1 to 50.

3. Several classifiers are applied to the feature sets extracted from test samples.

4. Repeat steps 1-3 30 times. The average test accuracy is compared under feature dimension 1 to 50.

In our experiments, the classifiers for gene selection are the Naive Bayes Classifier (NBC) and uncorrelated normal based quadratic Bayes classifier that is recorded UDC. The classifiers for test are Fisher Linear Discriminate (FLD), NBC, UDC, Nearest Mean Scale Classifier (NMSC), and Support Vector Machine (SVM) [15, 16, 17, 18, 19, 20].

C. Comparison of Gene Selections

Figures 1 to 3 compare the average test accuracies of the gene selections with different classifiers. Our results also indicate that class prediction and feature selection are strictly paired in the search of molecular profiles from microarray data. On the average, SVM-RFE is better than T-TEST; and CMS-
RFA and CMMC-RFA are better than CMSMI, CMMI, and SVM-RFE in the most of the tests.

In the legends of Fig. 1-5, from top to bottom, the legends are C-RFA, CMSC-RFA, CMMC-RFA, CMMI-RFA, CMSMC-RFA, SVM-RFE, and T-Test, respectively.
Fig. 1. The comparison of gene selections for LEUKEMIA classification with different learning classifiers. On the average, the recursive gene selection based on correlation information is the best. The best classifications are the classifier NMSC with NBCMSC-RFA, UDCMSC-RFA, and UDCMMC-RFA; and the classifier UDC with NBC-RFA, NBCMSC-RFA, UDC-RFA, UDCMSC-RFA, and UDCMMC-RFA.
Fig. 2. The comparison of gene selections for LYMPHOMA classification with different learning classifiers. On the average, the recursive gene selection based on correlation information is the best. The best classifications are the classifier NMSC with NBCMSC-RFA, NBCMMC-RFA, UDCMSC-RFA, and UDCMMC-RFA.
Fig. 3. The comparison of gene selections for PROSTATE classification with different learning classifiers. On the average, the recursive gene selection based on correlation information is the best. The best classifications are the classifier NMSC with NBCMSC-RFA, and UDCMSC-RFA.
IV. DISCUSSION

On the average, the recursive feature addition is better than SVM-RFE and T-TEST. As a result, the recursive feature addition takes advantage of the train accuracy of train samples with specific classifier; in other words, the next gene addition to the feature set is closely related with the maximum classification accuracy. In the majority of the tests, NBCMSC-RFA, NBCMMC-RFA, UDCMSC-RFA, and UDCMMC-RFA are better than NBC-RFA and UDC-RFA, because the strategies of combining classifier with the correlation measures between candidate genes and chosen genes go towards less redundancy and more completeness in the whole feature set, while the first strategy doesn’t consider the interaction of candidate genes and chosen genes. In addition, our experimental results show that, on the average, the strategy of gene selection based on correlation is better than the strategy based on mutual information. As a result, the correlation directly measures the similarity of two variables and outperforms mutual information in the approach to reducing the redundancy in the feature set.

The computational complexity of our strategy for gene selection is $O(dn)$, where $d$ is the dimensionality of the chosen feature set and $n$ is the number of genes in the original dataset. It should be noted that our method of recursive feature addition is not optimal, but suboptimal. The optimal strategy would be reached in an exhausted way. But for DNA microarray data with high feature dimension, the computational cost is so high that it is impractical.

V. CONCLUSION AND FUTURE WORK

In this paper we propose five strategies of recursive feature addition for gene selection. In comparison with the well-known feature selection algorithms of T-TEST and SVM-RFE, our strategies, overall, perform the best. Additionally, the strategy based on correlation gene selection is better than the strategy based on mutual information gene selection. Our experiment also indicates that the class prediction and feature selection are strictly paired in the search of molecular profiles from microarray data.

Future work includes enhancing or optimizing the strategy of recursive gene addition.

REFERENCES