Combining Support Vector Machines and the *t*-statistic for Gene Selection in DNA Microarray Data Analysis

Tao Yang[1], Vojislave Kecman[2], Longbing Cao[1] and Chengqi Zhang[1]

[1] Faculty of Engineering and Information Technology,
University of Technology, Sydney, Australia
[2] Department of Computer Science,
Virginia Commonwealth University, Richmond, VA, USA

Abstract. This paper proposes a new gene selection (or feature selection) method for DNA microarray data analysis. In the method, the *t*-statistic and support vector machines are combined efficiently. The resulting gene selection method uses both the data intrinsic information and learning algorithm performance to measure the relevance of a gene in a DNA microarray. We explain why and how the proposed method works well. The experimental results on two benchmarking microarray data sets show that the proposed method is competitive with previous methods. The proposed method can also be used for other feature selection problems.

1 Introduction

The advent of DNA microarray technology, such as the cDNA arrays and the high density oligonucleotide chips, has revolutionized the field of molecular biology in recent years. This new technology allows scientists to study thousands of genes simultaneously in a single experiment. This is a significant improvement because in the past, only several specific genes in an organism could be investigated at a time.

While the revolution generates much hope, the large amount of data obtained from microarray experiments, along with the structures of the resulting data sets, also challenges the conventional ways of analysis and modeling. One particular obstacle for analyzing a microarray data set is that often the number of genes is much greater than the number of samples; typically, the number of samples is less than a hundred, while the number of genes is usually in the thousands. In this regard, modern machine learning techniques provide a valuable toolkit for gaining insights into such data sets and extracting useful information from them.

Out of a large number of genes that exist in a microarray data set, it is often the case that most of them are irrelevant for the diagnosis of a particular disease, say, cancer, and hence are redundant. It is well-known that the performance of a modeling procedure can be significantly degraded, when many redundant genes

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 55–62, 2010.

are included in the training. Finding relevant genes can not only improve the accuracy of the resultant classifier for diagnosis purposes, but can also narrow down the potential set of cancerous genes and help gain important discipline knowledge.

Several methods for gene selection are available in the literature. One state-of-the-art technique is the method of Support Vector Machine Recursive Feature Elimination (SVM-RFE), proposed in [1]. The ranking criterion of SVM-RFE is constructed not by the intrinsic property of the data, but by the feedback from the support vector machine (SVM) classifiers. Specifically, the magnitudes of weights found by linear SVMs are used to rank the genes. At each iteration of the algorithm, a linear SVM is fitted to the training data with the remaining genes, and one or several genes are eliminated for their least significance in terms of the ranking criterion.

In this paper, we propose an alternative SVM-based method for gene selection, and call it TSVM-RFE. In particular, we consider selecting genes by combining the classical t-statistic and the modern SVM-RFE method. By taking care of the problems that may exist in either the univariate or multivariate worlds, TSVM-RFE is more robust to noisy genes than SVM-RFE and other methods, as confirmed by simulation studies.

2 Methods

In this section, we describe our gene selection method, TSVM-RFE, and illustrate its strengths, as well as giving a brief introduction to the support vector machines and t-statistic as needed by our method.

2.1 Data

The results from the microarray experiments can be represented by a matrix of expression levels. For microarray experiments having n tissue samples and p genes, the results can be represented by a $p \times n$ matrix X. In this paper, we shall focus on the classification problems with two classes, labeled by 1 and 2, respectively, and let n_k denote the sample size for class k; i.e., $n_1 + n_2 = n$. The response variable y_j, $j = 1, \ldots, n$, takes on the values of $+1$ or -1 for the two classes, respectively. For gene i, we use x_{ki} to denote the vector of values on the ith row of X that belong the class $k \in \{1, 2\}$. The mean of the values in x_{ki} is denoted by $\bar{x}_{ki}$, and the sample standard deviation by s_{ki}.

2.2 Support Vector Machines

The objective of SVMs is to find a classifier with the largest margin between the observations belonging to two different classes, while minimizing the training error. Here, the principle is that the classifier with the maximal margin is more likely to have a better generalization ability. A remarkable feature of SVMs is that the classifier is determined by only a few training samples, known as

"support vectors". These vectors are borderline samples, in the sense that they are closest to the decision boundary or simply lie on the margin.

In the studies below, we shall simply use the linear support vector machines, following [1], [3]. From these and other studies, linear SVMs appear to work reasonably well for the purposes of gene selection; for nonlinear support vector machines, we refer the reader to [4]. In order to select genes, the method of support vector machine recursive feature elimination for gene selection uses the absolute weight value $|w_i|$ given in the vector of parameters $\mathbf{w}$ to rank genes.

2.3 The t-statistic

The t-statistic measures the separability between classes using a standardized distance for a single gene, which gives a relevance score for each gene. The ranking criterion is given as

$$t_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sqrt{s_{1i}^2/n_1 + s_{2i}^2/n_2}}, \tag{1}$$

where $\bar{x}_{ki}$, s_{ki} and n_k are defined in Section 2.1.

2.4 TSVM-RFE

The basic idea of TSVM-RFE is to combine the t-statistic and SVM-RFE. The recursive feature elimination (RFE) algorithm [1] is used as the search engine of TSVM-RFE. In order to combine two different gene selection criteria, each criterion is transformed into a comparable scale. In particular, denoting by v the statistic used by a ranking criterion, the linear transformation

$$\sigma(v) = \frac{v - \min(v)}{\max(v) - \min(v)}, \tag{2}$$

is employed. Since $\sigma(\min(v)) = 0$ and $\sigma(\max(v)) = 1$, the range of $\sigma(v)$ is $[0, 1]$ for the training data. It is also possible to use other transformations, such as the sigmoid function or the probability function of a distribution, say, Gaussian.

From (2), the ranking statistic used for TSVM-RFE is

$$r_i = \alpha\sigma(|w_i|) + (1 - \alpha)\sigma(|t_i|), \quad 0 \le \alpha \le 1 \tag{3}$$

where w_i the weight found by linear SVMs and t_i the t-statistic. When $\alpha = 1$, TSVM-RFE is equivalent to SVM-RFE; when $\alpha = 0$, TSVM-RFE is equivalent to using the t-statistic.

To use (3), one problem remains to be solved, i.e., a value for α needs to be provided. For this, we use the 10-fold cross-validation. Specifically, a grid of α_i values are tested on the training data by 10-fold cross-validation, and the one that gives the lowest cross-validation error is deemed as "optimal". In our observation, using 11 equally spaced points for α between 0 and 1 ($\alpha = 0, 0.1, \ldots, 1$) seems enough. It is also possible to choose a finer grid, at a higher

computational cost. When more than one α value produces the same cross-validation error, we use the smallest of them if $\alpha = 0$ gives a smaller cross-validation error than $\alpha = 1$; otherwise, we use the largest. In other words, the weight in this case is determined in such a way that TSVM-RFE is as close as possible to the better of the two individual methods.

2.5 An Illustration

The motivation for TSVM-RFE is to overcome the weaknesses of each individual criterion. The t-statistic is a great criterion in measuring the class separability for each individual gene. However, it can only summarize at most the patterns that exist in the univariate world. Those multivariate patterns that are common in microarray data, such as correlation among genes, may never be represented by it. By contrast, SVM-RFE is expected to capture multivariate patterns well due to its foundation in the maximal margin principle, and could outperform the t-statistic for a number of data sets. However, since support vector machines are prone to overfitting when there exist a large number of noisy genes, the t-statistic can have an advantage in such cases. It is typical in practice that both noisy genes and multivariate patterns exist in microarray data. Hence, a criterion that combines the information provided by both the t-statistic and SVM-RFE is likely to preform reasonably well: at least as well as the better of the two individuals.

The above consideration is illustrated in the following using two simple examples. As shown in the left panel of Figure 1, the two classes of a two-dimensional data set are completely linearly separable. Here, according to SVM-RFE (using $C = 1$), x_2 is more relevant than x_1, because $|w_1| < |w_2|$. From the t-statistic, x_1 is more relevant than x_2, because $|t_1| > |t_2|$. In this example, gene selection based on the t-statistic appears more reasonable than SVM-RFE. This is because statistically speaking, the variation of x_2 for separating the two classes is large, while x_1 has none. This is a situation where the maximal margin principle fails to work well.

The second example is shown in the right panel of Figure 1. In this example, the data are two-dimensional and x_1 and x_2 are positively correlated. The two classes are also linearly separable. Here, the magnitude of the t-statistic for the two features are very different: $|t_1| = 0$ and $|t_2| = 3.098$. Due to the t-statistic, x_2 is relevant and x_1 irrelevant. From SVM-RFE ($C = 1$), however, x_1 and x_2 are equally relevant, because $|w_1| = |w_2| = 0.500$. Since in this example x_1 and x_2 appear to be equally important for identifying the pattern, the t-statistic fails to select all relevant features while SVM-RFE performs well. This is a situation where a multivariate pattern exists and the support vector machines work well, but not the t-statistic.

Admittedly, the above two examples are rather crude, but they demonstrate the difficulties that, if used individually, the support vector machines and t-statistic may have for gene selection. It is not uncommon for microarray data that genes are correlated and a large number of them are irrelevant. Given this, our combined criterion is expected to perform better than each of the two individual methods; in the worst scenario, it is simply equivalent to the better one.

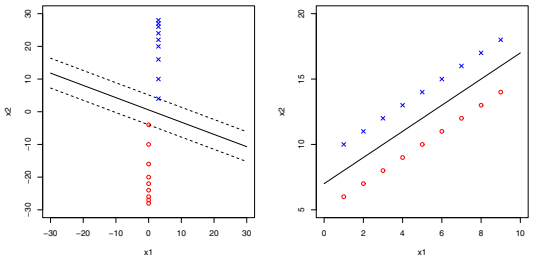

Fig. 1. Examples show the gene selection method using either the t-statistic or SVM-RFE may not be reliable. In the left panel, the gene selection method using t-statistic outperforms SVM-RFE. In the right panel, SVM-RFE outperforms the gene selection method using the t-statistic.

3 Experiments

Experiments were conducted to compare different gene selection methods and the results are given in this section. Two real data sets that are publicly available were used. A few points need to be clarified here. First, as part of pre-processing, we follow standardize each sample to mean zero and standard deviation one so as to treat each sample with an equal weight and thus to reduce array effects. Second, we follow [1] to select a fixed number of genes in the model *a priori*. In our experiments for the real data, the number of the genes retained are $10, 20, \ldots, 70$ for each method. Third, we use external cross-validation errors for comparison to avoid selection bias. Internal cross-validation errors are subject to selection bias, which are typically too optimistic [5]. Fourth, the SVM-RFE algorithm due to [3] was adopted. Fifth, a support vector machine classifier is constructed for each method after the genes are selected to assess its classification accuracy.

In the experiments, three gene selection methods, the t-statistic, SVM-RFE, and TSVM-RFE, are used to select genes. Classifiers based on linear SVM are then built using all training data, and subsequently examined using the test data.

3.1 Leukemia Data

The acute leukemia data consists of 72 samples and 7129 genes. They were obtained from Affymetrix oligonucleotide arrays. There are two types of leukemia: ALL (acute lymphocytic leukemia) and AML (acute mylogenous leukemia). We follow the procedure used in [5] to split the leukemia data set into a training set of size 38 and a test set of size 34 by sampling without replacement from all the samples, while ensuring that the training set has 25 ALL and 13 AML and the test set has 22 ALL and 12 AML. Different gene selection methods combined with linear SVMs are only applied to the training set, and then the methods are used on the test set to estimate their accuracies. Twenty such random partitions were carried out. Note that many proposed methods in the literature use a test set with size 34 only, whereas the testing procedure used here is equivalent to

use a test set with 680 samples, so it is much more reliable than using the independent test set of size 34 only. In our observation, $C = 1$ is a reasonable value for the penalty parameter of SVMs in this data set.

The results for the leukemia data are summarized in Figure 2. TSVM-RFE gives the smallest minimal error of 3.68%, and strictly smaller errors than SVM-RFE and the t-statistic-based method for $30, 40, \ldots, 70$ genes. The minimal test error obtained by TSVM-RFE is also smaller than the test errors obtained by using several other methods for this data set in the literature: the minimal test error 5.00% from SVM-RFE, obtained based on fifty similar random partitions [5]; the test error 7.00% from the nearest shrunken centroid method [7]; and the minimal test error 6.00% using soft-thresholding combined with kNN classifier obtained in [7]. Interestingly, all three methods give the lowest error when 60 genes are used. This provides a reasonable suggestion for the number of relevant genes that should be used for the leukemia data.

3.2 Colon Data

The colon cancer data consists of 62 samples and 2000 genes. They were also obtained from Affymetrix oligonucleotide arrays. The task is to distinguish between the normal and tumor samples. There are 22 normal samples and 40 tumor samples in the given data. We follow the procedure used in [6] to randomly split the colon data set into a training and test set by sampling without replacement from all the samples, while ensuring that the training set has 15 normal and 27 tumor samples and the test set has 7 normal samples and 13 tumor samples. Different methods are only applied to the training set, and the test set is used to estimate the classification accuracy. Twenty such random partitions were carried out. Note that it was suggested that there were some wrongly labeled data in the training data set [5]. We follow [3] and use $C = 0.01$ for this data set.

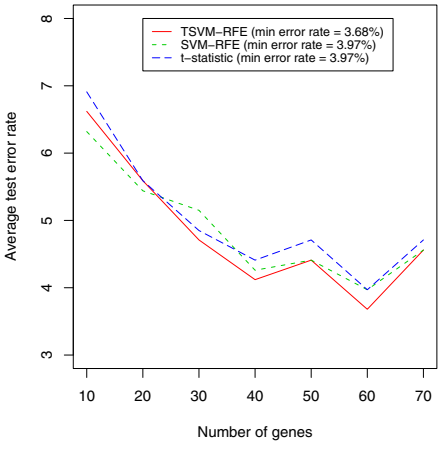

Fig. 2. Misclassification rates for the leukemia data

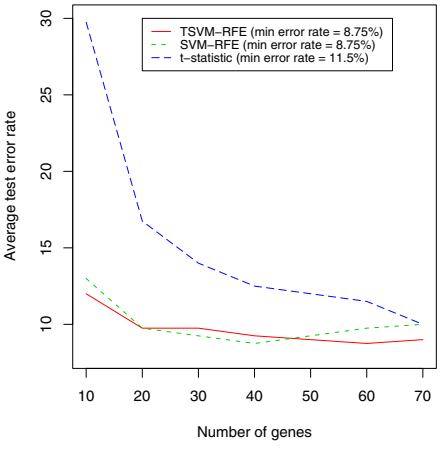

Fig. 3. Misclassification rates for the colon data

The results for the colon data are summarized in Figure 3. TSVM-RFE and SVM-RFE give the same minimal test error of 8.75%, and their performance is similar in this data set. In this data set, TSVM-RFE is always better than the method based on the t-statistic alone. The minimal test error obtained by TSVM-RFE here is also smaller than the test errors obtained by several other methods: the leave-one-out cross-validation error 9.68% obtained in [2], using correlation metric combined with SVMs; the minimal test error 17.50% from SVM-RFE obtained based on fifty similar random partitions [5]; the minimal jackknife error 12.50% obtained in [6] using weighted penalized partial least squares method; the minimal test error 11.16% from SVM-RFE with various values of C [3]; the test error 18.00% from the nearest shrunken centroid method [7]; the minimal test error 13.00% obtained in [7], using Wilcoxon statistic combined with kNN classifier; and the leave-one-out cross-validation error 8.90% obtained in [8], using the top scoring pair method.

4 Conclusions

We have proposed a new gene selection method, TSVM-RFE, for gene selection and classification. The criterion of TSVM-RFE combines the t-statistic and SVM-RFE, due to the consideration that the t-statistic only summarizes well the information in the univariate world and that SVM-RFE distinguishes the multivariate patterns better but is sensitive to noisy genes. We have chosen a linear transformation so that individual criteria are combined on a comparable scale, and the weight for each individual criterion is determined via cross-validation.

The proposed method was compared based on experiments with SVM-RFE and the t-statistic, using two practical data sets. The method presented in this paper gives competitive, if not better, results, compared to the other two. The improvement of the method proposed here upon the better-known SVM-RFE

method is significant. The method presented here seems to be rather suitable for microarray data analysis, where it is likely that a large number of irrelevant genes exist and the signal-to-noise ratio is fairly low. Experiments have shown that TSVM-RFE is better than both SVM-RFE and the t-statistic, in terms of reducing misclassification errors and lowering false discovery rates. It helps to identify more accurately the truly cancerous genes.

References

1. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002)
2. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906–914 (2000)
3. Huang, T.M., Kecman, V.: Gene extraction for cancer diagnosis by support vector machines - an improvement. Artif. Intell. Med. 35, 185–194 (2005)
4. Huang, T.M., Kecman, V., Kopriva, I.: Kernel Based Algorithms for Mining Huge Data Sets, Supervised, Semi-supervised, and Unsupervised Learning. Springer, Heidelberg (2006)
5. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA 99, 6562–6566 (2002)
6. Huang, X., Pan, W.: Linear regression and two-class classification with gene expression data. Bioinformatics 19, 2072–2078 (2003)
7. Lee, J.W., Lee, J.B., Park, M., Song, S.H.: An extensive comparison of recent classification tools applied to microarray data. Computational Statistics and Data Analysis 48, 869–885 (2005)
8. Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L., Geman, D.: Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21, 3896–3904 (2005)