Gene Extraction for Cancer Diagnosis by Support Vector Machines

An Improvement and Comparison with Nearest Shrunken Centroid Method

Te-Ming Huang and Vojislav Kecman

School of Engineering, The University of Auckland, New Zealand
v.kecman@auckland.ac.nz, huangjh@win.co.nz

Abstract. A cancer diagnosis by using the DNA microarray data faces many challenges the most serious one being the presence of thousands of genes and only several dozens (at the best) of patient's samples. Thus, making any kind of classification in high-dimensional spaces from a limited number of data is both an extremely difficult and a prone to an error procedure. The improved Recursive Feature Elimination with Support Vector Machines (RFE-SVMs) is introduced and used here for an elimination of less relevant genes and just for a reduction of the overall number of genes used in a medical diagnostic. The paper shows why and how the, usually neglected, penalty parameter C influence classification results and the gene selection of RFE-SVMs. With an appropriate parameter C chosen, the reduction in a diagnosis error is as high as 37% on the colon cancer data set. The results suggest that with a properly chosen parameter C, the extracted genes and the constructed classifier will ensure less over-fitting of the training data leading to an increase accuracy in selecting relevant genes.

1 Introduction

Recently, huge advances in DNA microarrays have allowed the scientist to test thousands of genes in normal or tumor tissues on a single array and check whether those genes are active, hyperactive or silent. Therefore, there is an increasing interest in changing the criterion of tumor classification from morphologic to molecular [1]. In this perspective, the problem can be regarded as a classification problem in machine learning, in which the class of a tumor tissue with a feature vector $\mathbf{x}$ is determined by a classifier. Each dimension, or a feature, in $\mathbf{x}$ holds the expression value of a particular gene which is obtained from DNA microarray experiment. The classifier is constructed by inputting l feature vectors of known tumor tissues into machine learning algorithms. To construct an accurate and reliable classifier with every gene included is not a straightforward task due to the fact that in the practice a number of tissue samples available for training is much less (a few dozens) than the number of features (a few thousands). In such a case, the classification space is nearly empty and it is difficult to construct a classifier that generalizes well. Therefore, there is a need

W. Duch et al. (Eds.): ICANN 2005, LNCS 3696, pp. 617–624, 2005.

to select a handful of most decisive genes in order to shrink the classification space and to improve the performance.

Support vector machines (SVMs) are one of the latest developments in statistical learning theory and they have been shown to perform very well in many areas of biological analysis including evaluating microarray expression, detecting remote protein homologies, and recognizing translation initiation sites. More recently, SVMs-based feature selection algorithms dubbed, Recursive Feature Elimination with Support Vector Machines (RFE-SVMs) have been introduced and applied to a gene selection for a cancer classification. In this work, we present the simulation results of the improved RFE-SVMs by tuning the C parameters on the popular colon cancer data set [2] and make comparison with the well-known nearest shrunken centroid method [3,4]. The C parameter plays an important role for SVMs in preventing an over-fitting but its effects on the performance of RFE-SVMs are still unexplored.

The paper is organized as follows: In section 2, we review SVM-RFE and some prior work in this area. The results on the influence of the C parameter on a correct selection of relevant features are presented in section 3. Section 4 shows the comparison between the improved RFE-SVMs and the nearest shrunken centroid on colon data set [2].

2 Prior Work

2.1 Support Vector Machines

The support vector machine classifier is based on the idea of margin maximization and it can be found by solving the following optimization problem [5]:

$$\text{Min } \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i^2 \tag{1a}$$

$$\text{s.t } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l \tag{1b}$$

The decision function for linear SVMs is given as $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$. In this formulation, we have the training data set (x_i, y_i) $i = 1, \ldots, l$ where $x_i \in \Re^n$ are the training data points or the tissue sample vectors, y_i are the class labels, l is the number of samples and n is the number of genes measured in each sample. By solving the optimization problem (1), i.e., by finding the parameters $\mathbf{w}$ and b for a given training set, we are effectively designing a decision hyperplane over an n dimensional input space that produces the maximal margin in the space. Generally, the optimization problem (1) is solved by changing it into the dual problem below,

$$\text{Max } L_d(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \mathbf{x}_i^T\mathbf{x}_j \tag{2a}$$

$$\text{s.t } 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, l \tag{2b}$$

$$\text{and} \quad \sum_{i=1}^{l}\alpha_i y_i = 0 \tag{2c}$$

In this setting, one needs to maximize the dual objective function $L_d(\alpha)$ with respect to the dual variables α_i only. The equality constraint (2c) can be eliminated by adding a constant of 1 to all the entries of the kernel matrix as suggested in [6,7]. Hence, the dual objective becomes

$$\text{Max } L_d(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i^T \mathbf{x}_j + \mathbf{1}) \tag{3}$$

subject only to the box constraints $0 \leq \alpha_i \leq C$. The optimization problem can be solved by various established techniques for solving general quadratic programming problems with inequality constraints.

2.2 Recursive Feature Elimination with Support Vector Machines

The idea of using the maximal margin for gene selection was first proposed in [8] and it was achieved by coupling recursive features elimination with linear SVMs to find a subset of genes that maximizes the performance of the classifiers. In a linear SVM, the decision function is given as $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$ or $f(\mathbf{x}) = \sum_{k=1}^{n} w_k x_k + b$. For a given feature x_k, the size of the absolute value of its weight w_k shows how significantly does x_k contribute to the margin of the linear SVMs and to the output of a linear classifier. Hence, it is used as a feature ranking coefficient in RFE-SVMs. In the original RFE-SVMs, the algorithm first starts constructing a linear SVMs classifier from the microarray data with n number of genes, then the gene with the smallest w_k^2 is removed and another classifier is trained on the remaining $n-1$ genes. This process is repeated until there is only one gene left. A gene ranking is produced at the end from the order of each gene being removed and the most relevant gene will be the one that is left at the end. However, for computational reasons, the algorithm is often implemented in such a way that several features are reduced at a time. In such a case, the method produces a feature subset ranking, as opposed to a feature ranking. Therefore, each feature in a subset may not be very relevant individually, and it is the feature subset that is optimal in some sense [8].

2.3 Selection Bias and How to Avoid It

As shown in [8], the leave-one-out error rate of RFE-SVMs can reach as low as zero percent with only 16 genes on the well-known colon cancer data set from [2]. However, as it was later pointed out in [1], the simulation results in [8] did not take selection bias into account. The leave-one-out error presented in [8] was measured using the classifier constructed from the subset of genes that were selected by RFE-SVMs using the complete data set. It gives too optimistic an assessment of the true prediction error, because the error is calculated internally. To take the selection bias into account, one needs to apply the gene selection and the learning algorithm on a training set to develop a classifier, and only then to perform an external cross-validation on a test set that had not been seen during the selection stage on a training data set. As shown in [1], the selection

bias can be quite significant and the test error that is based on 50% training and 50% test can be as high as 17.5% for the colon cancer data set. Another important observation from [1] is that there are no significant improvements when the number of genes used for constructing the classifier is reduced: the prediction errors are relatively constant until approximately 64 or so genes. These observations indicate that the performance and the usefulness of RFE-SVMs may be in question. However, the influence of the parameter C was neglected in [1] which restricts the results obtained. As a major part of this work, we further investigate the problem by changing (reducing) the parameter C in RFE-SVMs, in order to explore and to show the full potentials of RFE-SVMs.

3 Influence of the Parameter C in RFE-SVMs

The formulation in (1) is often referred to as the 'soft' margin SVMs, because the margin is softened and the softness of the margin is controlled by the C parameter. If C is infinitely large, or larger than the biggest α calculated, the

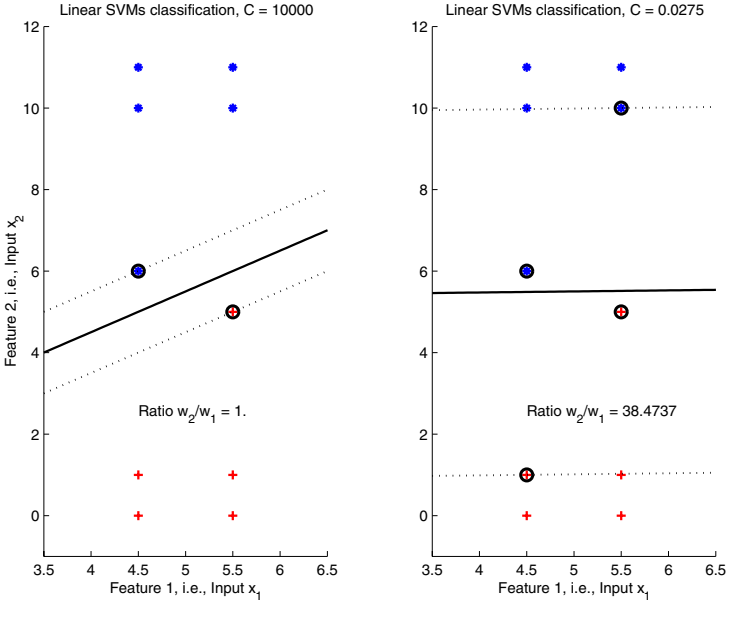

Fig. 1. A toy example shows how C may be influential in a feature selection. With C equal to 10000, both features seem to be equally important according to the feature ranking coefficients (namely, $w_1 = w_2$). With $C = 0.025$, a request for both a maximal and a 'hard' margin is relaxed and the feature 2 becomes more relevant than feature 1, because w_2 is larger than w_1 ($w_2/w_1 = 73$). While the former choice $C = 10000$ enforces the largest margin and all data to be outside it, the later one ($C = 0.025$) enforces the feature 'relevance' and gives better separation boundary because the two classes can be perfectly separated in a feature 2 direction only.

margin is basically 'hard', i.e., no points in the training data can be within or on the wrong side of the margin.

If C is smaller than the biggest original α_i, the margin is 'soft' one. As seen from all the $\alpha_j > C$ will be constrained to $\alpha_j = C$ and corresponding data points will be inside, or on the wrong side of, the margin. In the most of the work related to RFE-SVMs such as [8,9], the C parameter is set to a number that is sufficiently larger than the maximal α_i i.e., a hard margin SVM is implemented within such an RFE-SVMs model. Consequently, it has been reported that the performance of RFE-SVMs is insensitive to the parameter C. However, Fig.1 shows how C may influence the selection of more relevant features in a toy example where the two classes (stars * and pluses +) can be perfectly separated in a feature 2 direction only. In other words, the feature 1 is irrelevant for a perfect classification here. Note in the right hand side plot that a decrease in C i.e., a constraining of the dual variables $\alpha_i = C$, leads to a moving of some data within the margin. However, at the same time this helps in detecting the more relevant feature which is an input 2 here.

4 Gene Selection for the Colon Cancer and Comparison with the Nearest Shrunken Centroid

In this section, we present the selection of relevant genes for the colon data set which is well known in the gene microarray literature. The colon data set was analyzed initially in [2] and it is composed of 62 samples (22 normal and 40 cancerous) with 2000 genes' expressions in each sample. The training and the test sets are obtained by splitting the dataset into two equal groups of 31 elements, while ensuring each group has 11 normal and 20 cancerous tissues. The RFE-SVM is only applied on the training set to select relevant genes and to develop classifiers, and then the classifiers are used on the test set to estimate the error rate of the algorithms. 50 trails were carried out with random split for estimating the test error rate. A simple preprocessing step is performed on the colon data set to make sure each sample is treated equally and to reduce the array effects. Standardization is achieved by normalizing each sample to the one with zero mean and with a standard deviation of one. To speed up the gene selection process, 25% of the genes are removed at each step until less than 100 genes remained still to be ranked. Then the genes are removed one at a time. The simulation results for the colon data set are shown in Fig.2.

The Ambroise and McLachlan's curve in Fig.2 is directly taken from [1] and it is unclear what C value is used in this paper. By comparing the error rates for various C parameters, it is clear that changing the parameter C has significant influence on the performance of RFE-SVMs in this data set. The error rate is reduced from previously 17.5% as reported in [1] to 11.16% (a reduction of 35%) when C is equal to 0.005. For $C = 0.01$, the gene selection procedure improves the performance of the classifier: this trend can be observed by looking at the error rate reduction from initially around 15% at 2000 genes to 11.9% with 26 genes. Similar trend can be observed when $C = 0.005$, but the error rate reduction is

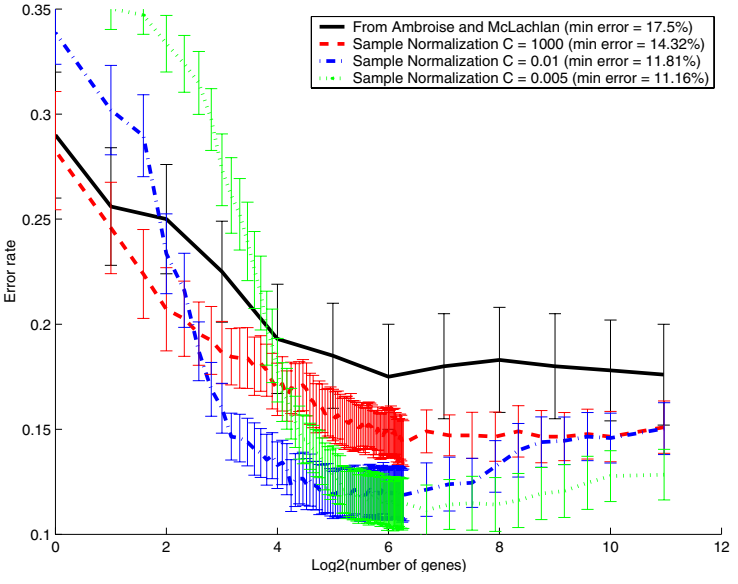

Fig. 2. Simulation result on the colon cancer data set with various C parameters. The error bar represents the 95% confidence interval.

not as significant as in the previous case. This is due to the fact that the error rate of the linear SVMs with $C = 0.005$ is already low, when all the genes are used. This also demonstrates that tuning the C parameter can reduce the amount of over-fitting on the training data even in such a high dimensional space with small number of samples. A preliminary comparison on the lowest leave-one-out error rate between RFE-SVMs and the well-known nearest shrunken centroid from [3] shows RFE-SVMs (8.0% at $C = 0.005$) is slightly better than nearest shrunken centroids (9.67%). The leave-one-out error rates presented here from both algorithms coincides with the suggestion in [1] that there are some wrongly labeled data in the training data set.

In order to further test the performance between the improved REF-SVMs and nearest shrunken centroid, we use again 50% of colon data for training and another 50% for testing. To make the comparison statistically more significant, we perform the experiment 100 times instead of 50 times as in Fig.2. Figure 3 shows the test errors and the corresponding 95% confidence interval of RFE-SVMs and the nearest shrunken centroid with various number of genes. As shown in the Fig.3, the performance of RFE-SVMs is superior to the nearest shrunken centroid in this test setting. It is interesting to point out that the error rate between the two algorithms is more significant in this more difficult setting (less training data) than in the leave-one-out setting. This may indicate that RFE-SVMs has more superior performance when the number of samples is low. The same, better, performance is observed in selecting the genes for CF (Cystic Fibrosis) diagnosis. (Due to the proprietary character of the CF data sets we

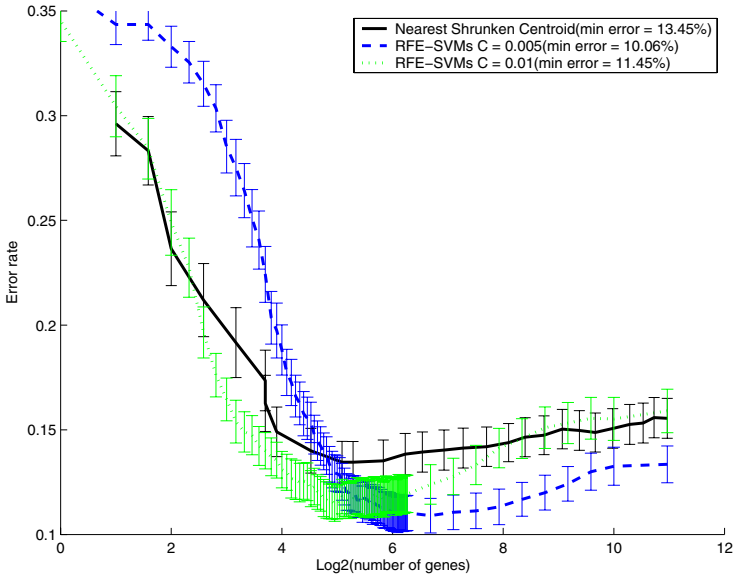

Fig. 3. Test errors on the colon cancer data set for two different methods

can't show the comparative results here). However, for a CF data set (with only 18 samples and approximately 12000 features), we would like just to mention that the RFE-SVM (having the error rate of 5%) performs again much better than the shrunken centroid (where the error rate is 33%).

5 Conclusions

We presented the performance of improved RFE-SVMs algorithm for genes extraction of DNA microarray data for diagnosing colon cancer. Why and how is this improvement achieved by using different values for the C parameter is discussed in details. With a properly chosen parameter C, the extracted genes and the constructed classifier will ensure less over-fitting of the training data leading to an increased accuracy in selecting relevant genes. The simulation results suggest that the classifier performs better in the reduced gene spaces selected by RFE-SVMs than in the complete 2000 dimensional gene space. This is a good indication that RFE-SVMs can select relevant genes, which can help in the diagnosis and in the biological analysis of both the genes' relevancy and their function. The comparison between the improved RFE-SVMs and nearest shrunken centroid on the colon data set suggested that the improved RFE-SVMs performs better when the number of data used for training is reduced. This phenomenal is also observed in the CF data set. Finally, the results in this work are developed from a more machine learning and data mining perspective, meaning unrelated to any valuable insight from a biology and medicine. Thus, there is

a need for a tighter cooperation between the biologists and/or medical experts and data miners in all the future investigations.

References

1. Ambriose, C., McLachlan, G.: Selection bias in gene extraction on the basis of microarray gene-expression data. In: PNAS. Volume 99. (2002) 6562–6566
2. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. In: Natl. Acad. Sci. USA, USA (1999) 6745–6750
3. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. In: National Academy of Sciences of the United States of America. Volume 99., USA (2002) 6567–6572
4. Black, M.: Statistical analysis of gene expression microarray data. Lecture note for Advanced Bioinformatics 2 (BIOSCI 744), Auckland, The University of Auckland (2004)
5. Kecman, V.: Learning and soft computing : Support vector machines, neural networks, and fuzzy logic models. Complex adaptive systems. MIT Press, Cambridge, Mass. (2001)
6. Huang, T., Kecman, V.: Bias b in svms again. In: 12th European Symposium on Artificial Neural Networks, Bruges, Belgium (2004)
7. Kecman, V., Vogt, M., Huang., T.: On the equality of kernel adatron and sequential minimal optimization in classification and regression tasks and alike algorithms for kernel machines. In: 11th European Symposium on Artificial Neural Networks, Bruges, Belgium (2003) 215–222
8. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46** (2002) 389–422
9. Rakotomamonjy, A.: Variable selection using svm-based criteria. Journal of Machine Learning (2003) 1357–1370