Two-Class SVM Trees (2-SVMT) for Biomarker Data Analysis

Shaoning Pang, Ilkka Havukkala, and Nikola Kasabov

Knowledge Engineering & Discover Research Institute,
Auckland University of Technology,
Private Bag 92006, Auckland 1020, New Zealand
spang@aut.ac.nz

Abstract. High dimensionality two-class biomarker data (e.g. microarray and proteomics data with few samples but large numbers of variables) is often difficult to classify. Many currently used methods cannot easily deal with unbalanced datasets (when the number of samples in class 1 and class 2 are very different). This problem can be alleviated by the following new method: first, sample data space by recursive partitions, then use two-class support vector machine tree (2-SVMT) for classification. Recursive partitioning divides the feature space into more manageable portions, from which informative features are more easily found by 2-SVMT. Using two-class microarray and proteomics data for cancer diagnostics, we demonstrate that 2-SVMT results in higher classification accuracy and especially more consistent classification of various datasets than standard SVM, KNN or C4.5. The advantage of the method is its super robustness for class unbalanced datasets.

1 Introduction

In biomarker (predictive gene and protein) analysis tasks, based on microarray or proteomics data, a common approach is to find a limited set of genes or proteins from a small number of samples (patients) giving good prediction. Various feature selection procedures have been used to reduce the dimensionality of data before classification. However, feature selection is not guaranteed to give optimum gene sets, because often many equally well performing sets of genes or proteins can be found even from one dataset [2]. If features are selected from a subset of samples (the learning set), this subset can affect strongly the resulting genes or proteins selected for the classifier.

Reduction of features (genes or proteins) is normally necessary, because too many input variables cannot be handled by the algorithms efficiently, and because many redundant and correlated features confuse the classifier. Two main procedures have been used to reduce number of features before classifying: filtering and wrapping [3]. Filtering preselects a subset of genes by e.g. signal-to-noise ratio, correlation coefficient, t-test, F-test, etc. Wrapping selects genes for the classifier by directly evaluating each gene repeatedly for classifying accuracy. All these methods can introduce feature selection bias. Currently no standard

J. Wang et al. (Eds.): ISNN 2006, LNCS 3973, pp. 629–634, 2006.

widely accepted and used method for feature selection exists. This suggests it would be better to try to avoid this bias completely.

Another important source of bias is due to the unbalanced distribution of samples (patients) into the two classes, e.g. 20 tumour vs. 190 non-tumour samples. Resolving imbalance can be done by resampling the smaller class and/or subsampling the bigger sample [4]. However, reducing the number of samples in the bigger class loses information. Repeated subsampling with a desired ratio of two classes in subsamples can reuse information too much, leading to overfitting.

We avoid the above biases by not using gene selection and using overlapping recursive partitioning of search space by dividing the whole sample space into many subspaces. This is then followed by two-class support vector machine tree, 2-SVMT, a method originally developed for face recognition [6]. Partitioning leads to subspaces containing informative genes that the classifier can find more easily in a smaller subset of data. Overlapping of the partitions is especially useful for small datasets, where non-overlapping partitions might lead to too small or unbalanced numbers of samples in the subspaces.

The 2-SVMT method enables much more efficient computation, because the data is divided into smaller chunks for processing. The biggest advantage arises for classification of small and unbalanced datasets, because overlapping partitions can present more discriminatory information to the classification process from a limited number of samples, and reuse datapoints for the class with fewer samples. In this report we summarize our results on microarray and proteomics two-class data analysed by 2-SVMT and show the benefits of our approach compared to SVM, KNN, and C4.5.

2 2-Class SVM Tree

2.1 Principle

Consider a dataset G with two classes for classification, a principle of "divide and conquer" for constructing SVMT can be implemented as,

$$F(\mathbf{x}) = \begin{cases} 1 \quad if\ f_{g_i}(\mathbf{x}) = 1, \mathbf{x} \in g_i \ i = 1..L \\ -1\ if\ f_{g_i}(\mathbf{x}) = -1, \end{cases} \tag{1}$$

where the total set (G) is divided into subsets $\{g_1, g_2, \cdots, g_L\}$ by a data partitioning function, and the number of subgroups L is determined after the SVM tree model generation.

To minimize the loss function of F, each subset either contains only one class data, or is verified with an optimal classification for SVM. Hence, the whole classification system $F(\mathbf{x})$ consists of a number of local SVM classifiers, each input sample $\mathbf{x}$ is first judged to be the member of a certain subgroup, then the class of $\mathbf{x}$ is determined locally by the subset SVM or the class label of the subset. One SVM is taken as one node of the tree, and a set of SVMs encapsulated in a hierarchy into an SVM tree for 2-class data classification.

2.2 Algorithms

The structure of 2-class SVM tree is constructed by a recursive procedure of data partitioning and a 2-class SVM classification. For modelling each node of 2-class SVM tree we have two iterated steps. First, the data is split into two subsets by the selected partition method. Next, any partition that contains only one class becomes a leaf node of the tree. For any other partition containing two-class data, a 2-class SVM classifier is trained for decision making over this partition. If the partition is not linear separable by SVM, the partitioning is repeated and the procedure is iterated. Partitioning is stopped when all partitions have been resolved.

To classify a new input x, first we decide in which partition the test input data belongs to by executing the partition method $P(x)$ at the root node in the SVM tree. Depending on the decision made by the root node, we will go down to one of the children nodes. This procedure is repeated until a leaf node or a SVM node is reached. We assign the class label to the testing sample x depending on the label of the reached leaf node, or the output of the SVM node. Therefore, the final decision of classifying new data can be cooperatively made by a hierarchical set of 2-class SVMs, with a suitable kernel chosen to suit the dataset.

2.3 Partitioning Method

Due to the loss function of F in Equation (1) being proportional to size of the subset (i.e. number of samples in the subset), it is desirable to seek a bigger size subset optimal for classification in the above 2-Class SVMT algorithm, which means a scalable data partitioning function would be beneficial. For such a scalable method, we chose the Evolving Clustering Method (ECM) which a fast one-pass algorithm for dynamic clustering of an input stream of data. This algorithm is a distance-based clustering method where the cluster centers (called "prototypes") are determined online such that the maximum distance, *MaxDist*, between an input $\boldsymbol{x}_i$ and the closest prototype cannot be larger than a threshold value, *Dthr*. The details of ECM algorithm can be found in [5]

ECM partitioning helps classification of gene expression data because ECM enables splitting the data in a multi-scaled way through parameter *Dthr* adjustment. Particularly, ECM allows overlapping partitions, which overcomes the common problem of few samples (patients) for biomarker data.

3 Application to Biomarker Data

3.1 Datasets and Processing

For validation of 2-SVMT approach with recursive partitioning for biomarker data analysis, we used one proteomics and seven microarray datasets with a two-class classification problem in cancer bioinformatics. The datasets showed a range of bias ratio, from 1.0 (equal number of samples in classes 1 and 2) to 0.33. Target classes were used as reported in the original references listed in Table 1.

Table 1. Comparison of classification accuracies of KNN, C4.5, SVM and 2-SVMT for eight biomarker (gene and protein) datasets, based on 10-fold cross validation, values are means of 10 runs. Numbers in boldface indicate the best results.

Cancer Dataset/Ref	Genes	Bias ratio class 1/2 patients	KNN	C4.5	SVM	2-SVMT
Lymphoma(1)/1	7129/	19/58=0.33	75.3%	**92.2%**	84.4%	77.9%
Leukemia*/2	7219/	11/27=0.41	**82.4%**	58.8%	64.7%	78.3%
CNS Tumour/3	7129/	21/39=0.53	58.3%	63.3%	50.0%	**72.0%**
Colon Cancer/4	2000/	22/40=0.55	79.0%	75.8%	71.3%	**80.7%**
Ovarian Cancer/5	15154/	91/162=0.56	91.7%	†	**97.3%**	75.0%
Breast Cancer */6	24482/	34/44=0.77	63.2%	†	52.6%	**73.7%**
Lymphoma(2)/1	6431/	26/32=0.81	53.5%	58.6%	51.7%	**60.3%**
Lung Cancer */7	12533/	16/16=1.0	**87.9%**	†	64.4%	75.0%

* Independent validation dataset was used for the accuracy evaluation. †not determined due to memory shortage of our computing system/

Data were divided into overlapping partitions using ECM as described above. Then a hierarchical 2-SVMT (second order polynomical kernel) was built using a classifier for each partition, validated by 10-fold cross-validation with random sampling. The average classification accuracies were calculated and the 2-SVMT results compared to other standard classifiers without partitioning (standard KNN (K=1, Euclidean Distance), C4.5, SVM (the second order polynomial kernel). Algorithms were implemented in Matlab Version 6.5, run on Pentium 4 PC, 3.0GHZ 512Mb RAM.

3.2 Results and Discussions

Cross-validated classification accuracies for data without gene selection in Table 1 show that 2-SVMT method with overlapping recursive partitioning performed best in the majority of cases (numbers in bold). Calculations of the average classification accuracy are in Table 2. The 2-SVMT was clearly the best for average accuracy and notably also most consistent, having a much narrower range of accuracies (60.3% to 80.7%) than other methods (50.0% to 97.3%) over the eight datasets.

Interestingly, the classification accuracy of SVMT seemed to increase with bias in distribution of patients to the two classes. This suggests unbalanced datasets

Table 2. Classification accuracy (%) of the three methods, based on the 8 datasets in Table 1. Boldface indicates the best result.

Model	mean (%)	Max - Min (%)	Range
KNN	73.0	91.7-53.5	38.2
SVM	66.6	**97.3**-50.0	47.3
C4.5	69.7	58.6-92.2	33.6
2-SVMT	**76.0**	80.7-60.3	**20.4**

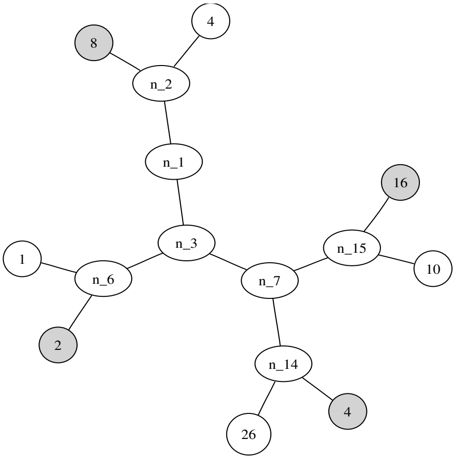

Fig. 1. An example of two-class support vector machine decision tree (2-SVMT) for CNS tumor. Ellipses: Partitioning and SVM decision nodes, numbering indicates the sequence of decisions. Open circles: samples (patients) assigned to class 1, filled circles: class 2, corresponding to Table 1. Numbers inside circles indicates the number of patients in the end of each decision path.

may be particularly suitable for analysis by 2-SVMT. This is in contrast to the KNN, C4.5, or SVM results, in which class bias did not correlate with accuracy.

An example of a decision tree for CNS tumor (Figure 1) illustrates the working of the method. Most of class 1 patients were classified into one node of 26 patients. Class 2 patients were classified into two main nodes with 16 and 8 individuals, suggesting a potential difference in these two subsets of patients. All the rest of the individuals were classified into nodes of only a few patients each, so the question arises, whether these are especially hard to classify patients, misclassifications or maybe represent some other types of cancer. This exemplifies the potentially valuable additional information that the 2-SVMT can provide, compared to other algorithms.

Other cancer datasets produced similar trees, some simpler, some more complex. It appears some cancers are inherently more difficult to classify than others, and this is reflected in the complexity of the classification trees.

4 Conclusions

The 2-SVMT method is a new useful classification tree method to analyse biomarker data. The recursive partitioning divides the problem to smaller subproblems solved in succession by a sequence of SVM nodes. Overlap of the partitions is useful in smaller and very unbalanced datasets, enabling the classifiers to utilize some of the data multiple times in more balanced subsets. Analogously, Ho [7] used random (but non-overlapping) subspace partitions to construct decision tree forests by combining multiple trees. Similarly, one can reuse information by creating ensembles of decision trees by bagging, boosting or randomization. Our 2-SVMT results with biomarker data thus confirm further the usefulness and utility of data partitioning to improve classifier performance.

The 2-SVMT method often outperformed KNN, C4.5 and SVM, producing both more accurate and, importantly, more consistent classification, safeguarding against occasional poor performance of other algorithms. Part of the reason for success may be in allowing overlaps in subspace partitions, though this would need to be verified by using 2-SVMT with various schemes of partitioning.

The advantages of the 2-SVMT are such that it would be worth to extend the method to multiclass classification case (work in progress). This might be most beneficial for small unbalanced datasets, where overlapping partitions might help to reuse the data in a way facilitating the correct classification of easily distinguishable subsets of data. This would help in many cases of biomarker datasets, which often focus on small numbers of patients in multigenic diseases having a variety of phenotypes to be classified simultaneously.

References

1. 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 Tumour and Normal Colon Tissues Probed by Oligonucleotide Arrays. Proc Natl Acad Sci U. S. A. 96(12) (1999) 6745-6750
2. Ein-Dor, L., Kela, I., Getz, G., Givol, D., Domany, E.: Outcome Signature Genes in Breast Cancer: Is There a Unique Set. Bioinformatics 21(2) (2005) 171-178
3. Deb, K., Agrawal, S., Pratap, A., and Meyarvian, T.: A Fast and Elitist Multi-objective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6(2) (2002) 182-197
4. Chawla, N.V., Bowyer, K.W, Hall, L.O. and Kegelmeyer, W.P.: SMOTE, Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16 (2002) 321-357
5. Kasabov, N.K.: Evolving Connectionist Systems, Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines. Springer-Verlag, (2002)
6. Pang, S., Kim D., and Bang, S.Y.: Face Membership Authentication Using SVM Classification Tree Generated by Membership-based LLE Data Partition. IEEE Trans. on Neural Networks 16(2)(2005) 436-446
7. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transaction on Pattern Analysis and Machine Intelligence 20(8) (1998) 832-844