BIOINFORMATICS

Vol. 20 no. 15 2004, pages 2429–2437
doi:10.1093/bioinformatics/bth267

A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression

*Tao Li, Chengliang Zhang and Mitsunori Ogihara**

Computer Science Department, University of Rochester, Rochester, NY 14627-0226, USA

Received on November 11, 2003; revised on March 31, 2004; accepted on April 4, 2004
Advance Access publication April 15, 2004

ABSTRACT

Summary: This paper studies the problem of building multiclass classifiers for tissue classification based on gene expression. The recent development of microarray technologies has enabled biologists to quantify gene expression of tens of thousands of genes in a single experiment. Biologists have begun collecting gene expression for a large number of samples. One of the urgent issues in the use of microarray data is to develop methods for characterizing samples based on their gene expression. The most basic step in the research direction is binary sample classification, which has been studied extensively over the past few years. This paper investigates the next step—multiclass classification of samples based on gene expression. The characteristics of expression data (e.g. large number of genes with small sample size) makes the classification problem more challenging.

The process of building multiclass classifiers is divided into two components: (i) selection of the features (i.e. genes) to be used for training and testing and (ii) selection of the classification method. This paper compares various feature selection methods as well as various state-of-the-art classification methods on various multiclass gene expression datasets.

Our study indicates that multiclass classification problem is much more difficult than the binary one for the gene expression datasets. The difficulty lies in the fact that the data are of high dimensionality and that the sample size is small. The classification accuracy appears to degrade very rapidly as the number of classes increases. In particular, the accuracy was very low regardless of the choices of the methods for large-class datasets (e.g. NCI60 and GCM). While increasing the number of samples is a plausible solution to the problem of accuracy degradation, it is important to develop algorithms that are able to analyze effectively multiple-class expression data for these special datasets.

Contact: ogihara@cs.rochester.edu

*To whom correspondence should be addressed.

1 INTRODUCTION

The fate and functions of cells are characterized by production of proteins, which consist of amino acids. The patterns of the amino acid sequences are many-to-one encoded in genes, which constitute a part of the genome. To produce the protein defined by a gene, a cell first transcribes its genetic sequence into its messenger RNA (mRNA) sequence. This is the molecule-wise copy of the DNA sequence into RNA (in eukaryotic organisms, parts of RNA sequences are eliminated). Then the mRNA sequence is translated in triplets into an amino acid sequence, which is the protein of the sequence. Gene expression refers to the level of production of protein molecules defined by a gene. Monitoring of gene expression is one of the most fundamental approach in genetics and molecular biology. The standard technique for measuring gene expression is to measure the mRNA instead of proteins, because mRNA sequences hybridize with their complementary RNA or DNA sequences while this property lacks in proteins.

The DNA arrays, pioneered in Chee *et al*. (1996) and Fodor *et al*. (1991), are novel technologies that are designed to measure gene expression of tens of thousands of genes in a single experiment. A DNA array consists of probe DNA sequences that are immobilized on a surface (gold or glass). To assess gene expression of tissues, their mRNA sequences are first extracted from them. The mRNA sequences are then amplified (copied at an exponential rate) and then reverse-transcribed to DNA sequences. The reverse-transcribed DNA sequences are fluorescently tagged. The probe sequences on the array are designed to hybridize with these reverse transcriptions (by virtue of DNA–DNA hybridization). After DNA–DNA hybridization, the array is scanned to quantify the fluorescent light dissipated from each probe. There exist two technologies for designing DNA arrays: using the complete sequence for each gene, called cDNA arrays (see Chee *et al*., 1996), and using a small number (usually one to two dozens) of short fragments of the mRNA sequence which, as a whole, uniquely capture the mRNA sequence, called DNA

microarrays (see Fodor *et al*., 1991). For an extensive survey of the technologies see, e.g. Eisen and Brown (1999).

The capability of measuring gene expression for a very large number of genes, covering the entire genome for some small organisms, raises the issue of characterizing cells in terms of gene expression, i.e. the question of whether gene expression can be used to determine the fate and functions of the cells. There are four major technical issues that confront the researchers who address such questions. First, with the current technology amplification of mRNA from a single cell is an extremely difficult task. So, tissues that seemingly share the same fate or the same functions are pooled to obtain a significant amount of mRNA. This implies that the expression levels calculated are the means of all the cells in the pool. Second, genetic variability affects gene expression, i.e. the expressions of two individuals can be different. Third, there is much room for noise to affect the outcome, at various points of experiments, e.g. at the time of tissue collection, at the time of mRNA amplification and at the time of hybridization on to the chip (see Dorris *et al*., 2002 for such discussions). Finally, the samples collected are small in numbers (not more than two dozens in many cases; rarely in the hundreds). Given that the dimensions of the data are very large (thousands to tens of thousands), the sample sizes are much too small.

The most fundamental of the characterization problem is the problem of identifying genes whose expression patterns either characterize a particular cell state or predict a certain forthcoming cell state. The first step in solving this problem is the development of tools for classifying samples according to their gene expression and of tools for clustering genes or samples. Significant progress has been made in the development. However, while binary classification and clustering are heavily studied, (see, e.g. Alizadeh *et al*., 2000; Ben-Dor *et al*., 2000; Brown *et al*., 2000; Der *et al*., 1998; Eisen *et al*., 1998; Friedman *et al*., 2000; Raychaudhuri *et al*., 2002; Tamayo *et al*., 1999; Welsh *et al*., 2001), only a small amount of work has been made on multiclass classification, i.e. classification involving more than two classes (Alizadeh *et al*., 2000; Golub *et al*., 1999; Khan *et al*., 2001; Ross *et al*., 2000; Tamayo *et al*., 1999).

This paper compares state-of-the-art machine learning techniques for multiclass classification with the goal of identifying a technique that is best suited for building multiclass sample classifiers based on gene expression. To achieve the goal, comprehensive experiments have been conducted, the programs for which will be made accessible to the community. Although no clear winner has been identified, some insights have been obtained from the experimental results classification.

2 PRIOR WORK ON MULTICLASS CLASSIFICATION

Classification problems aim at building an efficient, effective model for predicting class membership of data. The builder of a model (often called the learner) is given the training data, which consist of data points chosen from the input data space and their class label. A model built from the training data (often called a hypothesis) is expected not only to produce the correct label on the training data but to predict correctly the label for any unseen data. In the case when there are only two class labels, a classification problem is said to be binary. In the case when there are at least three class labels, a classification problem is said to be a multiclass classification problem. While binary classification problems are the simplest of all, but many real-world problems are multiclass problems.

Multiclass classification techniques can be roughly divided into two types. One type is the binary classification algorithms that can be naturally extended to handle multiclass problems directly. Discriminant analysis (Hastie *et al*., 2001; Li *et al*., 2003), regression and decision trees are of this type. The other type is the decompositions of multiclass problems into binary ones. One-versus-the-rest method (Scholkopf and Smola, 2002; Bottou *et al*., 1994), pairwise comparison (Kreeel, 1999; Hastie and Tibshirani, 1998; Friedman, 1996), error-correcting output coding (Dietterich and Bakiri, 1991, 1995; Allwein *et al*., 2000) and multiclass objective functions (Weston and Watkins, 1998; Lee and Lee, 2003) are of this type.

Scholkopf and Smola (2002) note that there is probably no multiclass method that outperforms everything else and that for practical purposes the choice of the method has to be made depending on the constraints, such as the desired level of accuracy, the time available for development and training, and the nature of the classification problems. However, choosing the best method is a very difficult task in practice. Even for a fixed multiclass approach, there are many details that can be fine tuned. For example, Crammer and Singer (2000) study various strategies for building codewords for error-correcting output coding, and find no clear winner.

This makes us to wonder whether there is one method that works the best for all classification problems based on gene expression. There are two pieces of prior work. Liu *et al*. (2002) present a comparative study of various feature selection heuristics using two datasets, Dudoit *et al*. (2002) compare various discrimination methods for tumor classification using three datasets. This paper combines and extends these two pieces of work. It provides comparison of various feature selection methods combined with various multiclass classification methods, include both of the two types of techniques, using a wider variety of datasets.

3 MULTICLASS LEARNING METHODS

This section briefly describes the multiclass classification methods that are studied in this paper.

3.1 Support vector machines and their reduction methods

Support vector machines (SVMs) (Vapnik, 1998) have exhibited superb performance in binary classification tasks. Intuitively, SVM aims at searching for a hyperplane that separates the two classes of data with largest margin (the margin is the distance between the hyperplane and the point closest to it). This paper studies four multiclass decomposition techniques for SVM. They are one-versus-the rest, pairwise comparison and error-correcting output coding (ECOC) with two code generation strategies: random coding and exhaustive coding. In the following, let $k \geq 3$ be the number of class labels.

In the one-versus-the-rest method, classifiers for discriminating one from all the other classes are assembled. For each i, $1 \leq i \leq k$, a binary classifier that separates class i from the rest is built. To predict a class label of a given data point, the output of each of the k classifiers is obtained. If there is a unique class label, say j, which is consistent with all the k prediction, the data point is assigned to class j. Otherwise, one of the k classes is selected randomly. In practical situations often arise in which consistent class assignment does not exist. The method is criticized because of this, and also for solving potentially asymmetric problems using a symmetric approach (Scholkopf and Smola, 2002). Although it is being used widely, experiments (Allwein *et al*., 2000) show that the method is easy to beat.

In the pairwise comparison method, a classifier is trained for each pair of classes, so there are $k(k-1)/2$ independently built binary classifiers. To predict a class label of a given data point, the prediction of each of the $k(k-1)/2$ classifiers is calculated, which is viewed as a vote. If there is a class, say j, which receives the largest number of votes, the data point is assigned to class j, where a tie is broken randomly. As in one-versus-the rest, the method can be criticized for solving asymmetric problems symmetrically. Also, the method is criticized for simplifying too much by removing the rest of the classes from consideration in training of pairwise clusters, which provides little overlap in the training sets between two classifiers. An advantage of using this method is that each classifier is easy to train since it is purely a binary problem to be solved. However, when individual training is time-consuming and the number of classes is large, the pairwise comparison method requires a large amount of time.

The ECOC method is due to Dietterich and Bakiri (1995). ECOC decomposes the original multiclass problem into a collection of binary classifiers that each solve a binary partition of the classes. Here the size, d, of the collection is determined by the 'coding strategy' to be used. Each binary classifier is designed to produce one of $+1$ and -1 as the class label. So, given a list of d classifiers, the outputs of them can be viewed as a (usually, a row) vector in $S = \{+1, -1\}^d$. Each class is assigned to a unique codeword in S. To predict the class label of an input x, the output 'word' of the d classifiers on input x is compared against the codeword of each class, and the class having the smallest Hamming distance (the number of disagreements) to the output 'word' is selected. Designing a good set of codewords for ECOC requires separation among classes and among classifiers and there are many design strategies (Dietterich and Bakiri, 1995). In this paper, two major ones are studied: random coding and exhaustive coding.

1. *Random coding*. In random coding, $[10 \log_2(k)]$ classifiers (i.e. columns) are used. The binary separation corresponding to each classifier is selected by assigning a value from $\{+1, -1\}$ uniformly at random (Allwein *et al*., 2000).
2. *Exhaustive coding*. In this strategy, each codeword starts out having length 2^{k-1}. The codeword for the first class is all $+1$. For i, $2 \leq i \leq k$, the codeword for the i-th class is constructed by repeating 2^{i-2} times a pattern, which is a length-2^{k-i} block of $+1$'s followed by a length-2^{k-i} block of -1's. The classifiers corresponding to the codewords are clearly pairwise distinct. The first classifier is supposed to assign $+1$ to every input, so unnecessary. Thus, the first $+1$ is dropped from every codeword. This reduces the codeword length to $2^{k-1} - 1$. It is easy to see that the minimum Hamming distance between any pair of codewords is $[2^{k-1} - 1/2]$ (see Dietterich and Bakiri, 1995). A limitation to this strategy is that the number of classifiers increases exponentially in the number of classes.

3.2 Other methods

Naive Bayes. Naive Bayes is one of the most successful learning algorithms for text categorization. Naive Bayes is based on the Bayes rule assuming conditional independence between classes. Based on the rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation.

K-nearest neighbor (KNN). KNN is a non-parametric classifier. KNN has been applied to various information retrieval problems. KNN uses an integer parameter, K. Given an input x, the algorithms finds the K closest training data points to x, and predicts the label of x based on the label of the K points. In this paper, the parameter for KNN is set to 1. It has been proven that the error of KNN is asymptotically at most two times the Bayesian error.

Decision Tree. Decision tree builds a binary classification tree. Each node corresponds to a binary predicate on one attribute, one branch corresponds to the positive instances of the predicate and the other to the negative instances. Thus, each node corresponds to a sequence of predicates and their values appearing on the downward path from the root to it. Each leaf is labeled by a class. To predict the class label of an input, a path to a leaf from the root is found depending on the value of the predicate at each node that is visited. The predicates are chosen from top to bottom by calculating the information

gain of each attribute, which is the expected reduction in entropy caused by partitioning of the samples according to the attribute. A post-pruning process is carried out to prevent overfitting. In our experiments, we use the J4.8 version of the decision tree algorithm, which is implemented in WEKA (Witten and Frank, 2000).

4 FEATURE SELECTION METHODS

Intuitively, one hypothesizes that a good feature set consists of those highly correlated with a class but are uncorrelated with other classes, which is confirmed in Hall (1999).

In gene expression data, the number of features is usually very high. The program Rankgene (Su *et al*., 2003)[1] is used for this study and feature selection is performed on the training set only. A total of eight feature selection methods are supported in the package: information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, one-dimensional SVM and t-statistics. The first six of these have been widely used either in machine learning (information gain and Gini index) or in statistical learning theory (twoing rule, max minority, sum minority and sum of variances). They quantify the effectiveness of a feature by evaluating the strength of class predictability when the prediction is made by splitting the full range of expression of a given gene into two regions, the high region and the low region. The split point is chosen to optimize the corresponding measure. The evaluation of the strength is different in each (Su *et al*., 2003). One-dimensional SVM measures the effectiveness of a future by calculating the accuracy of single-feature SVM classifiers. The t-statistics measure was first used in Golub *et al*. (1999)[2] to measure the class predictability of genes for two-class problems. Here, we compute t-statistics based on distinguishing one class from the rest.

5 EXPERIMENTAL RESULTS AND ANALYSIS

5.1 The datasets

The ALL/AML dataset consists of gene expression profiles of two acute cases of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML). The ALL part of the dataset comes from two types, B-cell and T-cell, while the AML part is split into two types, bone marrow samples and peripheral blood samples. The dataset was studied in the seminal paper of Golub *et al*. (1999). It is available at http://www-genome.wi.mit.edu. Golub *et al*. (1999) studied the data for binary classification between AML and ALL. However, due to the bipartition of each component, it can be treated both as a three-class dataset (B-cell, T-cell and AML) and as a four-class dataset (B-cell, T-cell, AML-BM and AML PB). Here, the three-class version is referred to as ALL-AML-3 and the four-class version as ALL-AML-4.

The ALL dataset (Yeoh *et al*., 2002) is one that covers six subtypes of ALL. The dataset is available at http://www.stjuderesearch.org/data/ALL1/. Using SVM with a set of discriminating genes selected by a correlation-based feature selection (CFS), Yeoh *et al*. (2002) achieve the accuracy of 96% on the test dataset.

The GCM dataset (Ramaswamy *et al*., 2001; Yeang *et al*., 2001) consists of 198 human tumor samples of 15 types. The prediction accuracy of 78% is reported in Ramaswamy *et al*. (2001) using one-versus-the rest SVM with all the genes.

SRBCT (Khan *et al*., 2001) is the dataset of small, round blue cell tumors of childhood and can be downloaded at http://research.nhgri.nih.gov/microarray/Supplement/. The training set of this dataset consists of 83 samples spanning four classes (excluding the five non-SRBCT samples). It is reported in Khan *et al*. (2001) that after excluding several samples a neural net achieved 100% accurate prediction.

The MLL-leukemia dataset consists of three classes and can be downloaded at http://research.dfci.harvard.edu/korsmeyer/Supp_pubs/Supp_Armstrong_Main.html. The dataset was first studied in Armstrong *et al*. (2002). The best reported performance is 95% with KNN.

The lymphoma dataset is a dataset of the three most prevalent adult lymphoid malignancies and available at http://genome-www.stanford.edu/lymphoma. The dataset was first studied in Alizadeh *et al*. (2000).

The NCI60 dataset was first studied in Ross *et al*. (2000). cDNA microarrays were used to examine the variation in gene expression among the 60 cell lines from the National Center Institute's anticancer drug screen. The dataset spans nine classes and can be downloaded at http://genome-www.stanford.edu/nci60/.

The HBC dataset consists of 22 hereditary breast cancer samples and was first studied in Hedenfalk *et al*. (2001). The dataset has three classes and can be downloaded at http://www.columbia.edu/xy56/project.htm.

In the experiments, the original partition of the datasets into training and test sets is used whenever information about the data split is available. In the absence of genuine test set, the different predictors are compared based on random divisions of the dataset into a training set and a test set. Popular choices for the split ratio are leave-one-out cross-validation

[1]The program can be downloaded at http://genomics10.bu.edu/yangsu/rankgene/.

[2]There are other ranking methods that are neither supported in RankGene nor tested in this paper. Bijlani *et al*. (2003) propose a binary classification algorithm that uses a pair of asymmetric measures. The method is not tested in this study because of its asymmetry. Bagirov *et al*. (2003) propose a method that uses the ranges of the values. Let $k = 2$ and for each $i \in \{1, 2\}$ let e_i^{max} and e_i^{min} respectively be the maximum and the minimum value for feature e in the i-th class. Suppose that $e_1^{max} \leq e_2^{max}$. Then the strength is measured by the ratio $e_1^{max} - e_2^{min} / e_2^{max} - e_1^{min}$. The quantity is at most 1. The genes are ranked in the decreasing order of their distance from 1. For each i, $e_i^{max} + e_i^{min} = 2\bar{e}_i$. So, $2(\bar{e}_2 - \bar{e}_1) = e_2^{max} + e_2^{min} - e_1^{max} - e_1^{min} = (e_2^{max} - e_1^{min}) - (e_1^{max} - e_2^{min})$. Thus, good discrimination based on the distance between the mean values of the two classes also requires $(e_2^{max} - e_1^{min})$ being large while $(e_1^{max} - e_2^{min})$ being small, which is essentially the same requirement as t-statistics.

Table 1. The breakdown of each datasets

Dataset	No. of training samples	No. of test samples	No. of genes	No. of classes
ALL-AML-3	72	—	7129	3
ALL-AML-4	72	—	7129	4
ALL	163	85	12 558	6
GCM	144	54	16 063	14
SRBCT	63	20	2308	4
MLL-leukemia	57	15	12 582	3
Lymphoma	62	—	4026	3
NCI60	60	—	1123	9
HBC	22	—	3226	3

and 10-fold cross-validation. The latter appears impossible due to smallness of the datasets. So, 4-fold cross-validation is used. The datasets and their characteristics are summarized in Table 1.

5.2 Experimental set-up

Our implementation of the various classifiers is based on the Weka (Witten and Frank, 2000) environment (http://www.cs.waikato.ac.nz/ml/weka/). The classification accuracy is used as the performance measures. All the experiments are performed on a P4 2 GHz machine with 512M memory running Linux 2.4.9-31. For experiments involving SVM, linear, polynomial and radius-based kernels are tested. The numbers reported are the best among these trials.

Data preprocessing is an important step for handling gene expression data. This includes two steps: filling missing values and normalization. For both training and test dataset, missing values are filled using the average value of that gene. Normalization is then carried out so that every observed gene expression has mean equal to 0 and variance equal to 1.

5.3 Deciding the number of genes

Deciding the number of genes to select is the first question for feature selection. Finding the optimal number of genes is generally very difficult. Many practical solutions are based on experience or some heuristics. A set of experiments are first conducted on the ALL dataset by varying the number of genes selected to investigate the effects of the number of genes using SVM with random coding (Fig. 1). It is observed that, when the number of selected genes is >150, the variation of the performance is small. The same test is then conducted on other methods, and the same property is observed. Hence, in all the rest of the experiments the top 150 genes are used.

5.4 Experimental results

Figures 2–9 show the results on various datasets. In each chart, No Rank presents the results obtained via various methods without feature selection, and the others, rank 1 through rank 8, correspond to the results obtained using the eight feature selection methods, numbered in the following order: information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, t-statistics and one-dimensional SVM.

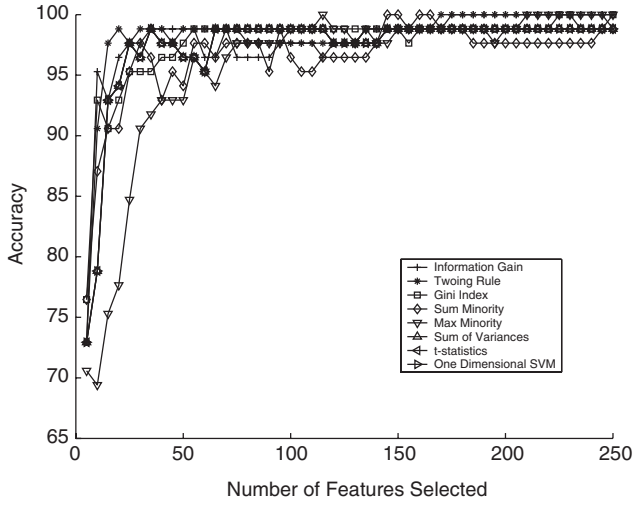

Fig. 1. SVM random.

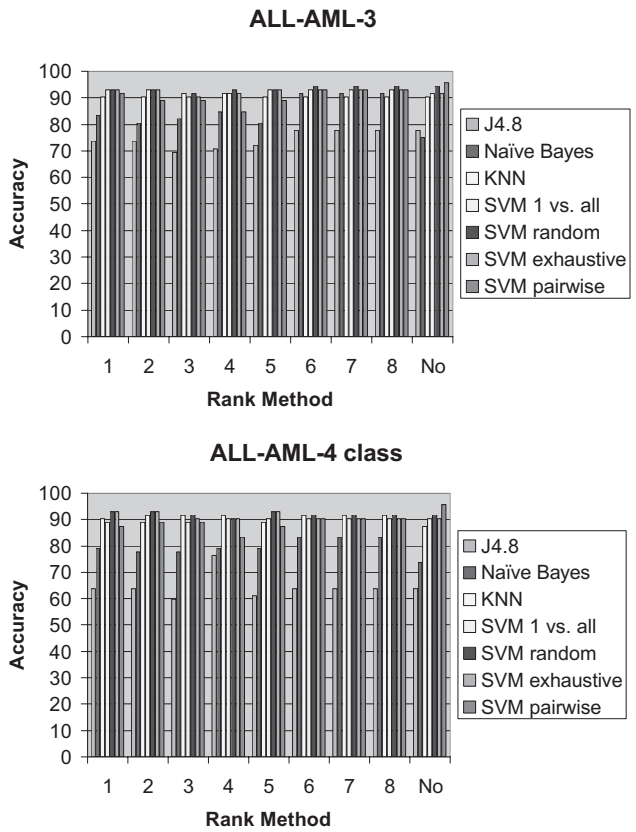

Fig. 2. ALL-AML-3 and ALL-AML-4 datasets.

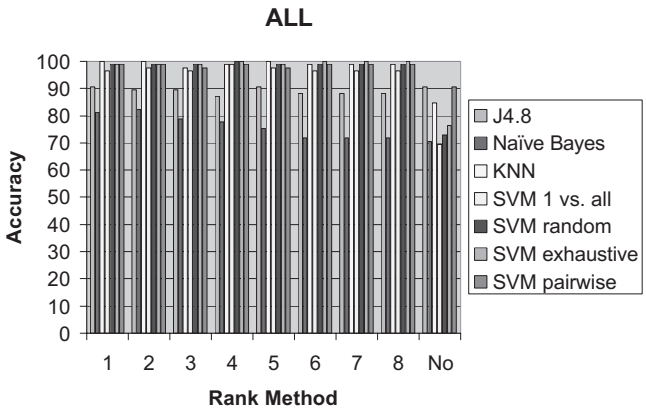

Fig. 3. ALL dataset.

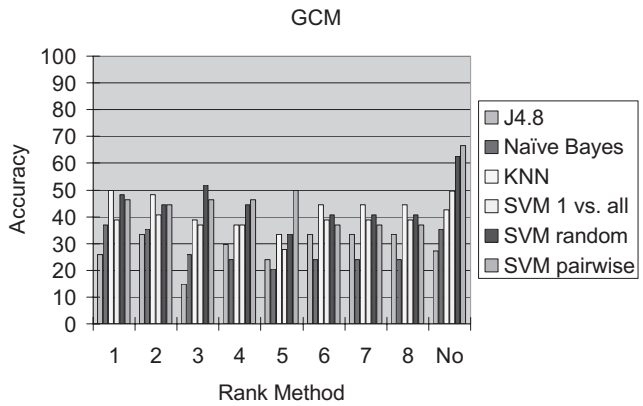

Fig. 4. GCM dataset.

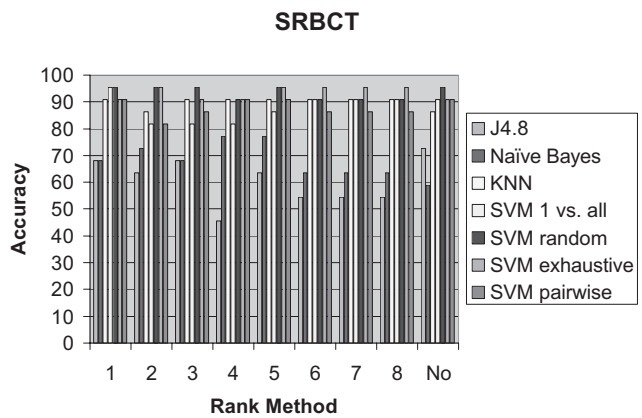

Fig. 5. SRBCT dataset.

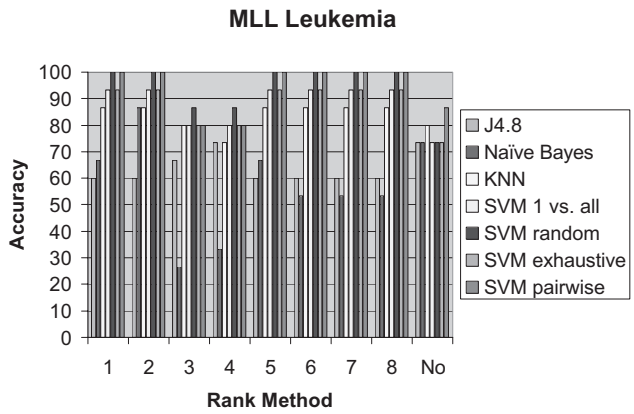

Fig. 6. MLL-leukemia dataset.

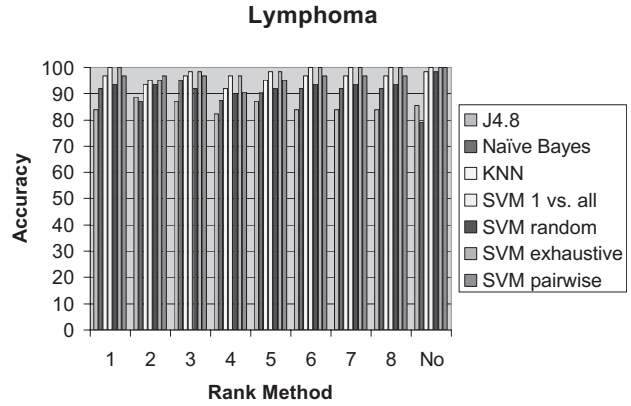

Fig. 7. Lymphomia dataset.

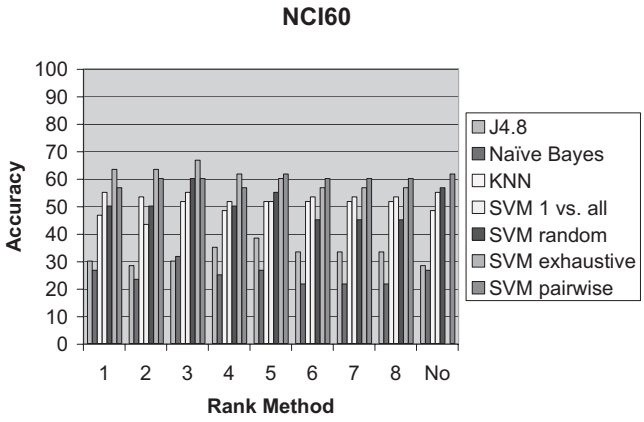

Fig. 8. NCI60 dataset.

We observe the following:

- SVM is shown to be the best classifier for tissue classification based on gene expression. They achieve better performance than any other classifiers on almost all the datasets. However, the best decomposition method for SVM appears to be problem-dependent, and there is is no clear overall winner.
- KNN achieves good performance on most of the datasets. Although its performance is not always as good

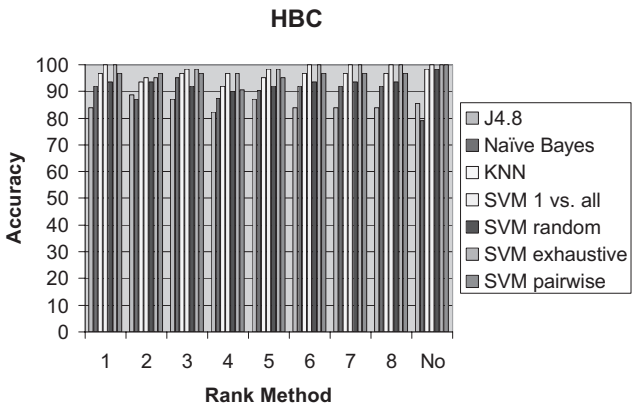

Fig. 9. HBC dataset.

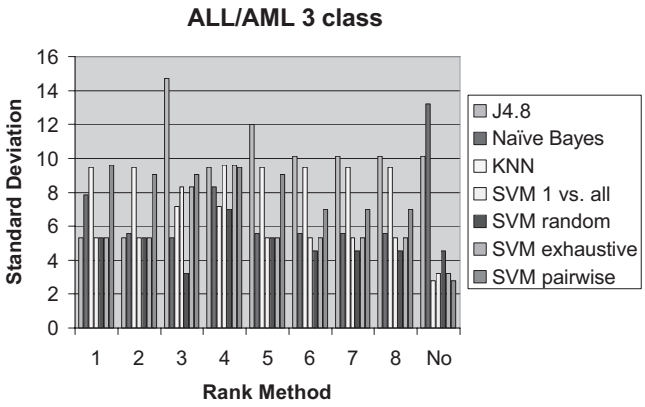

Fig. 10. ALLAML3 dataset SD.

as that of SVM, it outperforms decision tree and Naive Bayes on most datasets. On the ALL dataset, KNN with the twoing rule gives the perfect result. This indicates that after feature selection, the expression data can be well discriminated according to the distance. This also seems to suggest that the choice of feature selection method is very important for KNN. Although it has been widely used in text categorization, Naive Bayes does not appear to perform very well for tissue classification based on gene expression. This is not very surprising, since Naive Bayes is based on the assumption that the features are conditional independent given the class label, which may not be the case for gene expression data because of co-regulation.

- It is difficult to select the best feature selection method. There does not seem to exist a clear winner. For example, information gain has the superb performance on the ALL dataset; max minority performs the best on the SRBCT dataset; and sum of variances, t-statistic and one-SVM achieve the best result on the MLL-leukemia and Lymphoma datasets. Overall, the methods 6–8 (sum of variances, t-statistics and one-dimensional SVM) appear to have similar performance (by selecting almost identical set of features). In fact, on each of the datasets excluding GCM, the three methods produce exactly the same top-150 ranking.
- How the feature selection and the classification methods interact seems very complicated. On one hand, it is conceivable that feature selection lowers the accuracy since information may be lost by removing many features. This is indeed the case for decision tree, since the tree is built by dynamically selecting the most informative features. To wit, on all the datasets except NCI60 and GCM, feature selection actually downgrade the accuracy of decision tree. On the SRBCT dataset, the accuracy of decision tree with sum minority decreases by as much as 17%.

On the other hand, it is conceivable that feature selection raises the accuracy since it may eliminate noise and may reduce the number of insignificant dimensions, thereby overcoming the curse of dimensionality. This appears to be the case for KNN, which works based on geometric distance between samples. The accuracy of KNN is improved on all the datasets except the HBC and Lymphoma datasets (where the accuracy is lowered by 2–5%). The accuracy of Naive Bayes is also dramatically improved on almost all the datasets except the MLL-Leukemia and GCM datasets.

SVM does not seem to be either of the two cases. On the MLL-leukemia and ALL datasets, the accuracy of SVM is improved significantly (10–30%), while on the NCI60 and GCM datasets, the accuracy is lowered by 10–20%. Also, remarkably, only with the aid of feature selection SVM achieves the 100% accuracy on the MLL-leukemia and ALL datasets.

- The accuracy of classification is highly dependent on the choice of the classification method. The choice is more important than the choice of feature selection method.
- It is possible to achieve very high accuracy on most of the datasets studied here. For instance, on the ALL, MLL-Leukemia, Lymphoma and HBC datasets, SVM can achieve perfect prediction on the test datasets. On the SRBCT, ALL, ALL/AML-4-class and ALL/AML-3-class datasets, the best accuracy is >93%. However, the best performance on the NCI60 and GCM datasets is 66.66 and 63.33%, respectively. These two datasets have smaller sample sizes than the other datasets, so one may conclude that multiclass classification based on gene expression can be effectively solved when sample size is large.

Figure 10 shows the SD of the accuracy on the ALL/AML-3-class dataset by executing 4-fold cross-validation 20 times. It can be observed that the SD is always high for decision

tree and is consistently small for SVM. Similar results have been obtained (not included in this paper) with other datasets. However, one has to be careful, since estimation of SD can be very inaccurate for datasets with small sample size such as the HBC dataset.

6 CONCLUSION AND FUTURE WORK

This paper provides a comparative study on feature selection and multiclass classification for gene expression data. The study suggests that multiclass classification problems are more difficult binary one in general. The results are generally good for datasets with a small number of classes. The prediction accuracy is dramatically lower for the datasets with a large number of classes (e.g. NCI60 and GCM).

There are some natural future directions. First, are there better feature selection schemes? Most of the previously studied feature selection schemes rank features ignoring correlations between features (Dudoit *et al*., 2002). Is it possible to design a feature selection method that takes into consideration correlations between features? Second, can prediction strength, as those presented in Bijlani *et al*. (2003) and Golub *et al*. (1999), be taken into consideration to better estimate the predictive power of a feature? Finally, can ensemble methods in machine learning be applied to gene expression classification?

ACKNOWLEDGEMENTS

This work is supported in part by NSF grants EIA-0080124, DUE-9980943 and EIA-0205061 and in part by an NIH grant P30-AG18254.

REFERENCES

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Allwein,E.L., Schapire,R.E. and Singer,Y. (2000) Reducing multiclass to binary: a unifying approach for margin classifiers. *Proceedings of ICML*, Morgan Kaufmann, San Francisco, CA, pp. 9–16.

Armstrong,S.A., Staunton,J.E., Silverman,L.B., Pieters,R., den Boer,M.L., Minden,M.D., Sallan,S.E., Lander,E.S., Golub,T.R. and Korsmeyer,S.J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41–47.

Bagirov,A.M., Ferguson,B., Ivkovic,S., Saunders,G. and Yearwood,J. (2003) New algorithms for multi-class cancer diagnosis using tumor gene expression signatures. *Bioinformatics*, **19**, 1800–1807.

Ben-Dor,A., Bruhn,L., Friedman,N., Nachman,I., Schummer,M. and Yakhini,Z. (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–584.

Bijlani,R., Cheng,Y., Pearce,D.A., Brooks,A.I. and Ogihara,M. (2003) A biologically relevant classification approach to microarray data analysis: Independently Consistent Expression Discriminator (ICED). *Bioinformatics*, **19**, 69–80.

Bottou,L. *et al*. (1994) Comparison of classifier methods: a case study in handwriting digit recognition. In *Proceedings of ICPR-94*, IEEE Computer Society Press, Los Alamitos, CA, pp. 77–87.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M., Jr and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Nat. Acad. Sci.*, **97**, 262–267.

Chee,M., Yang,R., Hubbell,E., Berno,A., Huang,X.C., Stern,D., Winkler,J., Lockhardt,D.J., Morris,M.S. and Fodor,S.P. (1996) Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.

Crammer,K. and Singer,Y. (2000) On the learnability and design of output codes for multiclass problems. In *Proceedings of the Computational Learning Theory*, Morgan Kaufmann, San Francisco, CA, 35–46.

Der,S.D., Zhou,A., Williams,B.R.G. and Silverman,R.H. (1998) Identification of genes differentially regulated by interferon α, β, or γ using oligonucleotide arrays. *Proc. Natl Acad. Sci.*, **95**, 15623–15628.

Dietterich,T.G. and Bakiri,G. (1991) Error-correcting output codes: a general method for improving multiclass inductive learning programs. In Dean, T. L. and McKeown, K. (eds), *Proceedings of the Ninth AAAI National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, CA, pp. 572–577.

Dietterich,T.G. and Bakiri,G. (1995) Solving multiclass learning problems via error-correcting output codes. *J. Artif. Intell. Res.*, **2**, 263–286.

Dorris,D.R., Ramakrishnan,R., Trakas,D., Dudzik,F., Belval,R., Zhao,C., Nguyen,A., Domanus,M. and Mazumder,A. (2002) A highly reproducible, linear, and automated sample preparation method for DNA microarrays. *Genome Res.*, **12**, 976–984.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Methods Enzymol.*, **303**, 179–205.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci.*, **95**, 14863–4868.

Fodor,S.P., Read,J.L., Pirrung,M.C., Stryer,L., Lu,A.T. and Solas,D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–783.

Friedman,J. (1996) Another approach to polychotomous classification. *Technical report*, Department of Statistics, Stanford, Palo Alto, CA.

Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hall,M. (1999) Correlation-based feature selection for machine learning. PhD Thesis, Department of Computer Science, Waikato University, Waikato, NZ.

Hastie,T. and Tibshirani,R. (1998) Classification by pairwise coupling. In Jordan,M.I., Kearns,M.J. and Solla,S.A. (eds), *Advances in Neural Information Processing Systems*, Vol. 10. The MIT Press, Cambridge, MA.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, Prediction*. Springer.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Kallioniemi,O.P. *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **344**, 539–548.

Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Kreeel,U.H.-G. (1999) Pairwise classification and support vector machines. In Scholkopf,B., Burges,C. and Smola,A.J. (eds), Advances in Kernel Methods—Support Vector Learning, The MIT Press, Cambridge, MA, pp. 255–268.

Lee,Y. and Lee,C.-K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132–1139.

Li,T., Zhu,S. and Ogihara,M. (2003) Efficient multi-way text categorization via generalized discriminant analysis. In *Proceedings of Twelfth International Conference on Information and Knowledge Management (CIKM 2003)*, ACM Press, NY, pp. 317–324.

Liu,H., Li,J. and Wong,L. (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform.*, **13**, 51–60.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci., USA*, **98**, 15149–15154.

Raychaudhuri,S., Chang,J.T., Stuphin,P.D. and Altman,R.B. (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.*, **21**, 203–214.

Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., Van de Rijn,M., Waltham,M. *et al*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.

Scholkopf,B. and J.Smola,A. (2002) *Learning with Kernels*. MIT Press, Cambridge, MA.

Su,Y., Murali,T.M., Pavlovic,V. and Kasif,S. (2003) RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, 1578–1579.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci.*, **96**, 2907–2912.

Vapnik,V.N. (1998) *Statistical Learning Theory*. Wiley, New York, NY.

Welsh,J.B., Zarrinkar,P.P., Sapinoso,L.M., Kern,S.G., Behling,C.S., Monk,B.J., Lockhart,D.J., Burger,R.A. and Hampton,G.M. (2001) Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl Acad. Sci.*, **98**, 1176–1181.

Weston,J. and Watkins,C. (1998) Multi-class support vector machines. *Technical Report*, Department of Computer Science, Holloway, University of London, Egham, UK.

Witten,I.H. and Frank,E. (2000) *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco, CA.

Yeang,C.H., Ramaswamy,S., Tamayo,P., Mukherjee,S., Rifkin,R.M., Angelo,M., Reich,M., Lander,E., Mesirov,J. and Golub,T. (2001) Molecular classification of multiple tumor types. *Bioinformatics*, **11**, 1–7.

Yeoh,E.J., Ross,M.E., Shurtleff,S.A., Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.