Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis 48 (2005) 869 – 885

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

An extensive comparison of recent classification tools applied to microarray data

Jae Won Lee[a,*], Jung Bok Lee[a], Mira Park[b], Seuck Heun Song[a]

[a] Department of Statistics, Korea University, 5-1 Anam-dong, Sungbuk-gu, Seoul 136-701, South Korea
[b] Department of Pre-Medicine, Eulji Medical College, Taejeon 301-832, South Korea

Received 20 December 2003; received in revised form 4 March 2004

Abstract

Since most classification articles have applied a single technique to a single gene expression dataset, it is crucial to assess the performance of each method through a comprehensive comparative study. We evaluate by extensive comparison study extending Dudoit et al. (J. Amer. Statist. Assoc. 97 (2002) 77) the performance of recently developed classification methods in microarray experiment, and provide the guidelines for finding the most appropriate classification tools in various situations. We extend their comparison in three directions: more classification methods (21 methods), more datasets (7 datasets) and more gene selection techniques (3 methods). Our comparison study shows several interesting facts and provides the biologists and the biostatisticians some insights into the classification tools in microarray data analysis. This study also shows that the more sophisticated classifiers give better performances than classical methods such as kNN, DLDA, DQDA and the choice of gene selection method has much effect on the performance of the classification methods, and thus the classification methods should be considered together with the gene selection criteria.

Keywords: Microarray; Classification; Feature selection

1. Introduction

The recent development of microarray technologies to monitor gene expression in model organisms, cell lines, and human tissues has become an important part of biological research over the last several years. Microarrays allow the monitoring of expression

* Corresponding author. Tel.: +82-2-3290-2237; fax: +82-2-924-9895.
E-mail addresses: jael@korea.ac.kr (J.W. Lee), jungboky@korea.ac.kr (J.B. Lee), mira@emc.eulji.ac.kr (M. Park), ssong@korea.ac.kr (S.H. Song).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2004.03.017

levels of thousands to tens of thousands of genes simultaneously in a given cell type (Brown and Borstein, 1999; Lander, 1999). The cDNA microarray production process involves taking samples of messenger RNA(mRNA) from two distinct tissue samples, reverse-transcribing the mRNA into complementary DNA (cDNA), dye-labeling with different flours (Cy 3 and Cy 5) and hybridization onto each arrayed gene in a microarray slide. For oligonucleotide array, oligos are synthesized on the chip using a photolithography process similar to that used for making semiconductor chips. The relative fluorescence signal of the two dyes at a spot is a measure of the relative expression levels of the corresponding gene in the two samples of mRNA.

Any microarray experiment involves a number of stages such as design of experiment, image processing, normalization of red/green ratios, selection of differentially expressed genes, clustering of genes with similar expression profiles and classifying the different RNA sources. Numerical statistical issues are raised at each stage of the microarray data analysis, and we focus on the classification issues here. Classification is known as discrimination in the statistical literature and as supervised learning in the machine learning literature, and it generates gene expression profiles which can discriminate between different known cell types or conditions. There is a distinction between classification (or discrimination, supervised learning) and clustering (or unsupervised learning). If the classes are pre-existing, then classification analysis is more appropriate than clustering analysis.

While many authors have proposed the classification methods based on global gene expression analysis (Golub et al., 1999; Alizadeh et al., 2000; Ross et al., 2000), no systematic comparison of statistical methods with different pre-processing strategies is available yet for finding the most appropriate classification tool once the specific type of data is given. Since most classification articles have applied a single technique to a single gene expression dataset, it is crucial to assess the performance of each method through a comprehensive comparative study. Dudoit et al. (2002) have recently compared the performance of various classification methods for classifying tumors based on gene expression profiles. Nine methods were applied to three well-known datasets in their comparison, but these methods and datasets do not sufficiently cover the various situations in microarray experiments. For example, all the subjects in three datasets are cancer patients, and also more sophisticated classification methods have been recently proposed in microarray data analysis and need to be evaluated. Thus, more extensive comparative study is essential to provide the researchers a considerable insight for choosing the most appropriate classification methods in a given situation.

In this paper, we evaluate by extensive comparison study the performances of recently developed classification methods in microarray experiment, and provide the insights for finding the most appropriate classification tools in various situations. We extended the comparison by Dudoit et al. (2002) in three directions: more classification methods, more datasets and more gene selection techniques. In our comparison study, we applied 21 classification methods to seven various types of datasets. We also tried three different methods for selecting gene subset which are used for classification because some classification tools seem to be quite sensitive to the gene subset selection.

The paper is organized as follows. Section 2 briefly describes the recently developed classification tools in microarray data analysis, and explains some backgrounds on the datasets. Pre-processing scheme including gene selection, imputation of missing data, and standardization is also discussed in Section 3. In Section 4, we present and discuss the results, and compare the performances of the classification tools. Finally, we summarize and discuss our major findings in Section 5.

2. Classification methods and datasets

2.1. Classification methods

We briefly describe almost all currently available classification methods (21 methods) based on the gene expression profiles in microarray experiment. Included are (1) classical methods such as linear discriminant analysis, diagonal linear and quadratic discriminant analysis, k nearest neighbor, logistic regression and generalized partial least square method fitting logit models, (2) classification trees and aggregation classifiers such as CART, bagging, boosting, logit boosting and random forest, (3) machine learning approaches such as neural network algorithms and support vector machine, and (4) some generalized algorithms such as flexible discriminant analysis, penalized discriminant analysis, mixture discriminant analysis and shrunken centroid methods. A brief introduction of each method is given as follows:

1. Fisher's linear discriminant analysis (FLDA)

FLDA is a classification method that projects high-dimensional data onto a line and performs classification in this one-dimensional space. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class. Maximizing this criterion yields a closed form solution that involves the inverse of a covariance-like matrix. FLDA assumes (1) a normal (Gaussian) distribution of observations and (2) "equal group covariance". Additionally, variables cannot form linear expressions of one another. That is, they may not be perfectly correlated.

2 and 3. Diagonal linear and quadratic discriminant analysis (DLDA, DQDA)

DLDA and DQDA are simple Gaussian maximum likelihood discriminant rules for diagonal class covariance matrices with linear (DLDA) or quadratic (DQDA) discriminant function.

4. Logistic regression (LOGISTIC)

Logistic regression is a supervised method for the two- or multi-class classification problem (Hosmer and Lemeshow, 1989). Though a different model is used, it can be shown that logistic discrimination and Fisher discrimination are the same when the predictors are sampled from multivariate distributions with common covariance matrices.

5. Generalized partial least squares (GPLS)

Ding and Gentleman (2003) applied generalized partial least squares approaches. Their functionalties are based on and extended to Iteratively ReWeighted Least Squares (IRWPLS) by fitting logit models for all C classes vs. baseline class separately with an option of Firth's bias reduction procedure for two-group and multi-group classification proposed by Marx (1996).

6. k nearest neighbor (kNN)

For each feature in the input case, kNN is an intuitive method that classifies unlabeled examples based on their similarity with examples in the training set. It finds the k closest features in the training set and assigns to the class that appears most frequently within the k-subset.

7–11. CART and aggregating classifiers (BAG, BOOST, LogitBOOST, RandomForest)

Classification and regression tree (CART) is a tree-building technique which is unlike traditional data analysis methods (Breiman et al., 1984). CART analysis is a form of binary recursive partitioning. Aggregating means combining classifiers to improve accuracy of class prediction. In this study we consider CART based classification and CART with two aggregating systems: bagging (BAG) and boosting (BOOST) (Freund and Schapire, 1997). Recently Dettling and Buhlmann (2003) demonstrated that the generic boosting algorithm needs some modification to become an accurate classifier in the context of gene expression data. They built on the LogitBOOST which fits an additive logistic regression model by stagewise optimization of the binomial log-likelihood. Details can be found in Friedman et al. (2000). Random forests is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and the distributions are the same for all the trees in the forest (Breiman, 2001). Its accuracy is as good as that of Adaboost (Breiman, 1998) and is sometimes better. It is also relatively robust to outliers and noises and faster than bagging or boosting.

12 and 13. Single & multi layer neural network (NN-1, NN-3)

Artificial neural network is a well-known tool for unsupervised and/or supervised learning application (Zuruda, 1992). We performed neural network classifiers with both single and three layers.

14 and 15. Support vector machine (SVM-linear, radial)

Support vector machine is based on the structural risk minimization principle from statistical learning theory (Vapnik, 1998). It can be applied to regression, classification, and density estimation problems. The idea of structural risk minimization is to find a hypothesis for which one can guarantee the lowest probability of error. For SVM, Vapnik (1998) showed that this goal can be translated into finding the hyperplane with maximum margin for separable data. In this study, we used linear and radial kernel methods.

16 and 17. Flexible discriminant analysis (FDA-POL, FDA-MARS)

FDA is a generalization of linear discriminant analysis that casts the classification problem (Hastie et al., 1994) as one involving nonparametric regression procedures such as MARS (multivariate adaptive regression splines, FDA-MARS, Friedman, 1991) and polynomial regression model (FDA-POL, Hastie et al., 1994).

18. Penalized discriminant analysis (PDA)

PDA is a form of penalized LDA. It is designed for situations in which there are many highly correlated predictors (Hastie et al., 1995).

19 and 20. Mixture discriminant analysis (MDA-Linear, MDA-MARS)

MDA is a generalized LDA assuming that each observed class is a mixture of unobserved subclasses (Hastie and Tibshirani, 1996). MDA can be viewed as a smooth version of learning vector quantization (LVQ) which generalizes clustering to classification problems.

21. Shrunken centroids method (or Predictive Analysis of Microarrays (PAM))

PAM is an enhancement of nearest prototype (centroid) classifier whose prototype is shrunken by the method proposed by Tibshirani et al. (2002).

2.2. Datasets

All the twenty-one classification methods described above are applied to various types of datasets. Note that Leukemia, Lymphoma and NCI 60 data were applied in Dudoit et al. (2002).

1. Leukemia (LEU)

Leukemia dataset composed of 3571 gene expressions in three classes of leukemias: B-cell and T-cell acute lymphoblastic leukemia (B-cell ALL-38 patients, T-cell ALL-9 patients) and acute myeloid leukemia (AML-25 patients) (Golub et al., 1999). The data were obtained after three pre-processing (thresholding, filtering and logarithm transformation and standardization) described in Dudoit et al. (2002).

2. Lymphoma (LYM)

In order to examine the extent to which genomic-scale gene expression profiling can help our understanding of B cell malignancies of lymphoma, Alizadeh et al. (2000) studied gene expression of three prevalent adult lymphoid malignancies: B-cell chronic lymphocytic leukemia (B-CLL), follicular lymphoma (FL) and diffuse large B-cell lymphoma (DLBCL). Among 96 samples, we took 62 samples of 4026 genes in three classes (B-CLL-11, FL-9, and DLBCL-42).

3. NCI 60 (NCI60)

This dataset was produced by The National Cancer Institute's anti-cancer drug screen project. The cell lines were derived from various tumor tissues: 7 breast, 5 central nervous system (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 nonsmall cell lung carcinoma (NSCLC), 6 ovarian, 2 prostate, 9 renal and 1 unknown. The full dataset composed of 60 samples and 9703 genes. Because the size of some classes was too small to perform discriminant analysis, we used a subset with 1375 genes and six classes which was also used in Ross et al. (2000). Based on hierarchical clustering depicted in Fig. 2 of Scherf et al. (2000), we assigned 6 classes and the size of each class is 8, 13, 9, 11, 10 and 8, respectively. Most of the samples in class 1 are leukemia patients, and CNS is predominant in class 6.

4. Colon cancer (COLON)

This dataset comes from a gene expression study of 40 tumor and 22 normal colon tissue samples which were analyzed with an Affymetrix oligonucleotide array complementary to more than 6500 human genes (Alon et al., 1999). A selection of 2000 genes with highest minimal intensity across the samples has been made by Table 1 of Alon et al. (1999). We used this gene expression data collected with size of 62 samples and 2000 genes.

5. Lung cancer (LUNG)

This dataset comes from gene expression study to find evidence that analysis of gene expression patterns can provide a basis for classification of lung cancer that recapitulates and extends the conventional division of lung tumors into four morphological subtypes (Garber et al., 2001). We selected 73 samples (Adeno-41, Normal-6, and Squamous-17, Small cell-5, Large cell-4) and 918 genes.

6. Small round blue cell tumor (SRBCT)

This dataset comes from SRBCT study of childhood (Khan et al., 2001). The data, consisting of expression measurements on 2308 genes, were obtained from glass-slide cDNA microarrays, which were prepared according to the standard of National Human Genome Research Institute. The tumors are classified as Burkitt lymphoma (BL): 22, Ewing sarcoma (EWS): 20, neuroblastoma (NB): 12, or rhabdomyosarcoma (RMS): 8. Since this data did not make public, we used training set of Tibshirani et al. (2002) with size of 63 samples and 2308 genes.

7. Yeast

Gene expression in the budding yeast *Saccharomyces cerevisiae* was studied during the diauxic shift, the mitotic cell division cycle, sporulation and temperature and reducing shocks by Eisen et al. (1998). The data matrix consists of 2467 genes by 79 slides. We assigned 8 classes to 79 slides according to the results of clustering analysis depicted in Fig. 2 of Eisen et al. (1998).

3. Data processing for classification analysis

After data processing described as follows, we performed 2:1 cross-validation (training set:test set) and observed classification error rates for seven standardized and imputed datasets (COLON, LEU, LYM, NCI60, LUNG, SRBCT and YEAST). This procedure was repeated 200 times.

3.1. Gene selection

To investigate the effect of gene selection methods, we applied BSS/WSS criterion (Dudoit et al., 2002), Wilcoxon rank-based statistics and soft-thresholding method (Tibshirani et al., 2002) to each data set. BSS/WSS method selects genes which maximize the ratio of between-group to within group sum of squares. For a gene j, the ratio is

$$\frac{\mathrm{BSS}(j)}{\mathrm{WSS}(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{*j})^2}{\sum_i \sum_k I(y_i = k)(\bar{x}_{ij} - \bar{x}_{kj})^2},$$

where $\bar{x}_{*j}$ denotes the mean expression level of gene j across all the samples and $\bar{x}_{kj}$ denotes the mean expression level of gene j across the samples belonging to class k. In our study, top 50 genes are selected in each data.

In order to apply the rank-based approach in gene selection, we compared each response class separately against all other classes and measure Wilcoxon statistic between one and all other groups. For K classes, the OVA (One vs. All) approach is used to form 2K sets of top-ranked genes based on Wilcoxon statistic to cover multi-group cases.

The third method used here is so-called 'soft-thresholding' method which was used in PAM (Tibshirani et al., 2002). It shrinks the class centroids toward the overall centroid after standardizing by the within-class standard deviation for each gene. This standardization has the effect of giving higher weight to genes whose expression is stable within samples of the same class. Let d_{jk} denote a t-statistic for gene j, comparing class k to the overall centroid, and each d_{jk} is reduced by the amount Δ in absolute value. Soft-thresholding is defined by

$$d'_{jk} = \mathrm{sign}(d_{jk})(|d_{jk}| - \Delta)_+,$$

where $+$ means positive part. We used soft-thresholding method for gene selection by choosing the optimal amount of shrinkage which minimizes test error rates under the restriction that the maximum number of selected genes is less than 200.

3.2. Missing values and standardization

All the datasets were standardized so that the observations (arrays) have mean 0 and variance 1 across genes. For the Lymphoma data, missing values of some arrays are imputed by k-nearest neighbor method with $k = 5$ (Troyanskaya et al., 2001).

4. Comparison results

We evaluated the performance of twenty-one classification methods applied to each of the seven datasets. For simplicity of the tables, we divided these classification methods into four categories: classical methods, tree methods, machine learning methods and recently proposed generalized LDA methods and PAM. For each category, we presented and compared the test set error rates of the classification method. Three different gene selection methods were also considered and compared in each category. Note that we extended the comparison of discrimination methods of Dudoit et al. (2002) in three directions: more classification methods, more gene selection methods and more datasets.

Classical methods such as FLDA, DLDA & DQDA, kNN, LOGISTIC and GPLS approach were evaluated. Table 1 displays the test set error rate of each classifier. Upper and middle table shows the results when the BSS/WSS and the rank-based selection criterion is used for gene selection while lower table uses the soft-thresholding method.

Dudoit et al. (2002) have applied some methods to Leukemia, Lymphoma and NCI 60 data, and used the BSS/WSS criterion for gene selection. They claimed that simple classifiers such as linear discriminant analysis (LDA) and kNN performed remarkably well compared to more sophisticated methods, and the same pattern can also be shown in Table 1. As stated by Tibshirani et al. (2003), in discriminating the abnormal group from the normal group, the variability of the average gene expression may differ between two groups. The variability in expression may be greater in the abnormal group, due to heterogeneity in the abnormal group. Fig. 1 of Tibshirani et al. (2003) showed that, in LYM data, the average expression of FL and CLL subgroups are much more variable than that of DLCL. Therefore, in our analysis both COLON and LYM data could be the good examples of heterogeneous

Table 1
Mean test set error rates of classical methods in various datasets

Gene selection	Methods	Datasets						
		COLON ($C=2$)	LEU ($C=3$)	LYM ($C=3$)	SRBCT ($C=4$)	LUNG ($C=5$)	NCI60 ($C=6$)	YEAST ($C=8$)
BSS/WSS	FLDA	0.30	0.21	0.18	0.24	0.33	0.31	0.27
	DLDA	0.21	0.14	0.15	0.18	0.06	0.24	0.14
	DQDA	0.14	0.15	0.14	0.19	0.12	0.32	0.18
	kNN	0.26	0.12	0.12	0.17	0.16	0.26	0.24
	LOG	0.29	0.21	0.22	0.37	0.31	0.46	0.40
	GPLS	0.31	0.20	0.18	0.20	0.19	0.40	0.23
Rank-based	FLDA	0.28	0.23	0.17	0.26	0.41	0.47	0.38
	DLDA	0.14	0.07	0.09	0.11	0.10	0.38	0.24
	DQDA	0.14	0.08	0.08	0.14	0.17	0.43	0.31
	kNN	0.13	0.11	0.10	0.11	0.19	0.47	0.36
	LOG	0.24	0.26	0.23	0.21	0.27	0.51	0.43
	GPLS	0.28	0.18	0.23	0.19	0.19	0.43	0.34
Soft-thresholding	FLDA	0.24	0.20	0.19	0.20	0.30	0.31	0.32
	DLDA	0.16	0.15	0.14	0.18	0.08	0.23	0.13
	DQDA	0.15	0.15	0.14	0.18	0.16	0.38	0.18
	kNN	0.20	0.06	0.13	0.16	0.17	0.24	0.27
	LOG	0.27	0.19	0.20	0.37	0.37	0.49	0.39
	GPLS	0.26	0.08	0.18	0.19	0.18	0.42	0.19

case. In heterogeneous dataset like COLON and LYM data, DQDA performs better than LDA and kNN. DLDA always performs better than FLDA in all the seven datasets, and kNN performs well especially when the number of classes is moderate as in LEU, LYM, SRBCT. DLDA has a good performance when the number of classes is relatively large as in LUNG (5 classes), NCI60 (6 classes) and YEAST (8 classes) data. Performance of most classical classification methods improves with the rank-based gene selection method when the number of classes is not large, whereas in LUNG, NCI60, YEAST data, the rank-based method gives higher error rate than the BSS/WSS method. The soft-thresholding method gives the similar pattern to the BSS/WSS method.

Classification trees and aggregating classifiers such as CART, bagging (BAG), boosting (BOOST), Logit boosting (LogitBOOST) and random forest (RandomForest) were evaluated. Table 2 displays the test set error rate of each classifier.

When the aggregating classifiers are applied, the soft-thresholding method for gene selection gives much better performances than the BSS/WSS selection method. The improvement seems to be notable in most datasets. Note that, in Dudoit et al. (2002), only the BSS/WSS criterion was used with BAG and BOOST.

It is also shown that the aggregating classifiers such as bagging, boosting, improve the performance of CART significantly in all the datasets. However, RandomForest is most excellent among the tree methods when the number of classes is moderate as in

Table 2
Mean test set error rates of tree methods in various datasets

Gene selection	Methods	Datasets						
		COLON ($C=2$)	LEU ($C=3$)	LYM ($C=3$)	SRBCT ($C=4$)	LUNG ($C=5$)	NCI60 ($C=6$)	YEAST ($C=8$)
BSS/WSS	CART	0.26	0.10	0.11	0.22	0.24	0.56	0.42
	BAG	0.22	0.07	0.08	0.11	0.14	0.41	0.28
	BOOST	0.21	0.12	0.21	0.09	0.13	0.43	0.26
	LBOOST	0.16	0.37	0.26	0.07	0.06	0.33	0.23
	RForest	0.16	0.04	0.04	0.01	0.12	0.32	0.19
Rank-based	CART	0.26	0.18	0.15	0.37	0.28	0.51	0.49
	BAG	0.19	0.10	0.13	0.21	0.18	0.47	0.39
	BOOST	0.21	0.14	0.10	0.20	0.16	0.48	0.38
	LBOOST	0.26	0.38	0.27	0.23	0.09	0.38	0.38
	RForest	0.16	0.10	0.07	0.18	0.16	0.37	0.28
Soft-thresholding	CART	0.33	0.15	0.08	0.20	0.29	0.53	0.42
	BAG	0.16	0.08	0.08	0.06	0.20	0.37	0.26
	BOOST	0.15	0.11	0.09	0.06	0.18	0.39	0.26
	LBOOST	0.14	0.16	0.29	0.19	0.10	0.31	0.20
	RForest	0.14	0.07	0.04	0.01	0.16	0.32	0.19

*LBOOST: LogitBOOST.
**RForest: RandomForest.

COLON, LEU, LYM, SRBCT data. Although LogitBOOST gives higher error rates than BOOST in most cases, it surpasses BOOST when the number of classes is large.

Machine learning approaches such as neural network algorithms with single and three layers (NN-1 & NN-3) and support vector machine (SVM-Lin and SVM-Rad) were evaluated. Table 3 displays the test set error rate of each classifier.

Among the four methods, the SVM classifier performs the best in most datasets. And the choice of kernel do not affect the performance of SVM except in COLON data. The middle and lower table shows the similar pattern to the upper table which means that the gene selection method has no much effect on the performance of SVM.

Some recently proposed generalized LDA algorithms such as flexible discriminant analysis (FDA-POL, FDA-MARS), penalized discriminant analysis (PDA), and mixture discriminant analysis (MDA-POL, MDA-MARS) together with shrunken centroid method like PAM were evaluated. Table 4 displays the test set error rate of each classifier. Note that PAM internally uses the soft-thresholding gene selection technique only.

FDA-MARS performs much better than FDA-POL, but performs worse than PDA. There seems to be no preference between PDA and PAM when the soft-thresholding technique is used. PDA performs better than PAM in COLON, LEU and LYM data while it does worse in LUNG (5 classes), NCI 60 (6 classes) and YEAST (8 classes) data, and it seems that PDA is getting worse when the number of classes is getting larger. There seems to be no difference between these two methods in SRBCT data. However, the performance of generalized LDA is improved when the BSS/WSS

Table 3
Mean test set error rates of machine learning methods in various datasets

Gene selection	Methods	Datasets						
		COLON ($C=2$)	LEU ($C=3$)	LYM ($C=3$)	SRBCT ($C=4$)	LUNG ($C=5$)	NCI60 ($C=6$)	YEAST ($C=8$)
BSS/WSS	NN1	0.28	0.23	0.23	0.41	0.32	0.67	0.44
	NN3	0.22	0.07	0.15	0.12	0.20	0.47	0.25
	SVM-Lin	0.22	0.04	0.02	0.01	0.09	0.43	0.11
	SVM-Rad	0.14	0.05	0.02	0.01	0.11	0.46	0.10
Rank-based	NN1	0.21	0.18	0.22	0.29	0.31	0.60	0.45
	NN3	0.21	0.10	0.11	0.20	0.22	0.61	0.34
	SVM-Lin	0.22	0.06	0.06	0.08	0.13	0.44	0.28
	SVM-Rad	0.14	0.08	0.06	0.11	0.16	0.48	0.21
Soft-thresholding	NN1	0.17	0.22	0.23	0.37	0.29	0.67	0.45
	NN3	0.20	0.06	0.15	0.12	0.19	0.50	0.24
	SVM-Lin	0.19	0.05	0.03	0.01	0.12	0.45	0.13
	SVM-Rad	0.12	0.05	0.03	0.01	0.12	0.49	0.10

Table 4
Mean test set error rates of generalized LDA methods and PAM in various datasets

Gene selection	Methods	Datasets						
		COLON ($C=2$)	LEU ($C=3$)	LYM ($C=3$)	SRBCT ($C=4$)	LUNG ($C=5$)	NCI60 ($C=6$)	YEAST ($C=8$)
BSS/WSS	FDA-POL	0.46	0.63	0.69	0.69	0.54	0.80	0.15
	FDA-MARS	0.23	0.09	0.07	0.09	0.13	0.43	0.43
	PDA	0.18	0.05	0.04	0.01	0.12	0.31	0.12
	MDA-POL	0.24	0.09	0.12	0.23	0.39	0.49	0.22
	MDA-MARS	0.22	0.09	0.11	0.11	0.38	0.53	0.43
Rank-based	FDA-POL	0.42	0.66	0.32	0.64	0.55	0.67	0.28
	FDA-MARS	0.23	0.12	0.10	0.22	0.21	0.56	0.46
	PDA	0.19	0.08	0.07	0.09	0.17	0.47	0.23
	MDA-Pol	0.42	0.32	0.32	0.48	0.48	0.51	0.37
	MDA-MARS	0.24	0.47	0.15	0.15	0.18	0.44	0.49
Soft-thresholding	FDA-POL	0.40	0.54	0.62	0.46	0.49	0.81	0.21
	FDA-MARS	0.21	0.15	0.07	0.10	0.14	0.47	0.45
	PDA	0.16	0.06	0.04	0.01	0.13	0.40	0.23
	MDA-Pol	0.25	0.32	0.13	0.28	0.38	0.46	0.33
	MDA-MARS	0.22	0.28	0.13	0.18	0.38	0.46	0.45
	PAM	0.18	0.07	0.16	0.01	0.08	0.32	0.10

criterion is used. PAM performs the best in LUNG, SRBCT and YEAST data, but does not seem to be always good even with using the soft-thresholding gene selection technique.

5. Conclusion and discussion

In the previous section, we presented and compared the performances of the classification methods for each of the four categories. As mentioned above, we extended the comparison of discrimination methods of Dudoit et al. (2002) in three directions: more classification methods, more gene selection methods and more datasets. We now summarize the results and provide the guideline for choosing the most appropriate classification method in a given situation. Here we also summarize and discuss the major findings in our comparative study.

5.1. Most recommendable methods

Fig. 1 shows the test error rates of the most recommendable methods in each dataset. Fig. 1(a) shows the best methods when the BSS/WSS selection criterion is used, and Fig. 1(b) and Fig. 1(c) depicts the minimum error rate method when the rank-based method and the soft-thresholding method is used, respectively. The best classifier in each dataset from our comparison study can be summarized as follows:

- Colon cancer (COLON): SVM-Rad with soft-thresholding, test error rate=0.12
- Leukemia (LEU): SVM-L with BSS/WSS criterion, test error rate=0.04
- Lymphoma (LYM):
  SVM-Lin & Rad with BSS/WSS criterion, test error rate=0.02
- Small round blue cell tumor (SRBCT):
  RandomForest, PDA, SVM-Lin & Rad with BSS/WSS and soft-thresholding,
  PAM with soft-thresholding, test error rate=0.01
- Lung cancer (LUNG): PAM with soft-thresholding, test error rate=0.08
- NCI 60 cells (NCI 60): DLDA with soft-thresholding, test error rate=0.23
- Yeast *Saccharomyces cerevisiae* (YEAST):
  SVM-R with BWSS/WSS and soft-thresholding, test error rate=0.10
  PAM with soft-thresholding, test error rate=0.10

5.2. Summary of major findings

Here we summarize the major findings in our comparative study as follows:

- Linear discriminant analyses (FLDA, DLDA) and kNN perform well compared to more sophisticated methods in homogeneous dataset, but DQDA performs better in heterogeneous dataset.
- DLDA always performs better than FLDA in all the seven datasets, and kNN performs well especially when the number of classes is moderate. DLDA performs well with large number of classes.
- Most classical classification methods perform better with the rank-based gene selection method compared to the BSS/WSS and the soft-thresholding method.
- Aggregating classifiers such as bagging, boosting and RandomForest improve the performance of CART significantly.

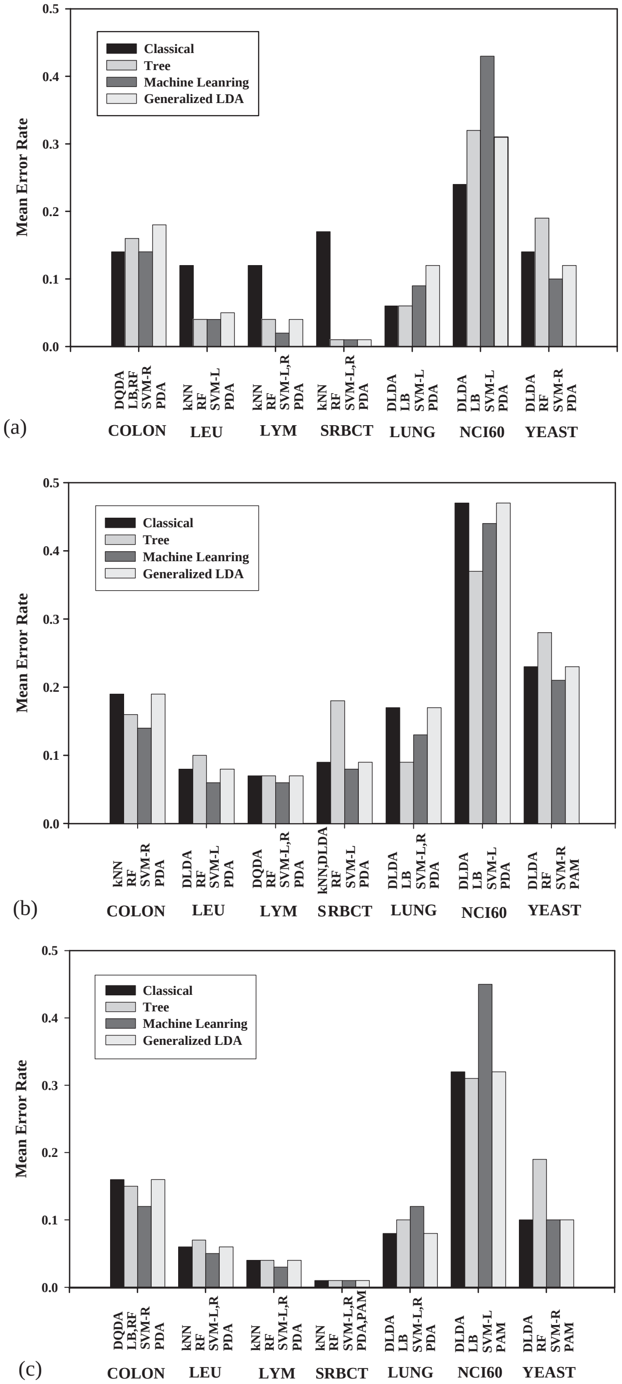

Fig. 1. Mean test set error rate of the best method in each of the four categories: (a) when the BSS/WSS criterion is used; (b) when the rank-based criterion is used; (c) when the soft-thresholding criterion is used.

- RandomForest is the most excellent method among the tree methods when the number of classes is moderate.
- SVM gives the best performance among the machine learning methods in most datasets regardless of the gene selection.
- The performance of PDA and PAM are comparable, but PDA with the BSS/WSS criterion seems to perform better than PAM in most cases.

Now let us discuss our comparison results in more detail. Classical methods considered here are easily applicable in many softwares and have no strong theoretical restrictions. However, we observed the more sophisticated classifiers give better performances than these classical methods in many cases. kNN performs quite well among the classical methods in most datasets, but when the number of classes in the dataset is getting large, there are many other classifiers that perform better. We also found that both the number of classes and the variance–covariance structure of the dataset are important factors to consider for evaluating some classical methods such as kNN, DLDA, DQDA etc..

As stated by Dudoit et al. (2002), aggregating methods such as BAG and BOOST improve the accuracy of classification compared to the unstable single tree method. Among the aggregating methods considered here, RandomForest with 50 selected genes performs the best. It is known that RandomForest tends to work poorly when the size of subgroups is different, but we could not observe its weakness even when the size of the subgroups is quite different in our microarray data. In general, logitBOOST is expected be more robust than BOOST to the presence of mislabeled observations and heterogeneity of the learning dataset, but logitBOOST does not seem to be robust to the gene selection and the datasets in our study. Dettling and Buhlmann (2003) used the leave-one-out-cross-validation (LOOCV) to compare BOOST with logitBOOST, and claimed that logitBOOST is more accurate than BOOST. We used 2:1 cross-validation in our study, however, and showed that LogitBOOST gives higher error rates than BOOST in most cases and surpasses BOOST when the number of classes is large. It shows the possibility that the choice of cross-validation methods may have effect on performance of these classifiers.

Furey et al. (2000) demonstrated SVM for classifying microarray data had a good performance in COLON and LEU. When SVM was applied to the multi-class case, the error rates were quite low and similar to the results of Furey et al. (2000). We used the "libsvm" module (Chang and Lin, 2001) of R-library e1071, and it applied one-against-one technique by fitting all the binary sub-classifiers and finding the correct class by a voting mechanism for allowing for multi-class situation. Another strength of SVM is its robustness to gene selection methods, and we also observed that the choice of kernel has negligible effect in all the seven datasets.

For generalized LDA, all the methods except PDA are discouraged to use. PDA is quite attractive for microarray data. First, there are high throughput gene expressions produced by microarray experiments, which implies there are many predictor variables we need to consider and the selected genes are highly correlated. Second, in class prediction of microarray, the class boundaries in predictor space may not be always simple or linear. So, PDA is recommendable as an alternative to the classical methods.

In our analysis, PDA has better performance than the classical methods such as kNN, DLDA or DQDA.

We applied 2:1 cross-validation (two thirds for learning, one third for testing) to evaluate the performance of the classification methods. Although not tabulated here, we also tried to analyze the data based on LOOCV. In some cases, there were notable differences in test error rates between 2:1 CV and LOOCV, but two methods selected the same best classifiers in most cases. LOOCV often works well for continuous error functions such as the mean squared error, but it may perform poorly for discontinuous error functions such as the number or percent of misclassified cases. In our analysis, test error rates using LOOCV are similar among the classifiers, and thus it is difficult to find the best classifiers in various situations. For example, in LEU data, the error rates of most classical methods are 4.2% with BSS/WSS, and in SRBCT data, they are $0 \sim 1\%$ regardless of the gene selection. A possible standard choice is 10-fold (9:1) CV. However, in our study, test sets containing 10% of the data are not sufficiently large to provide adequate discrimination between the classifiers. Thus, we applied 3-fold CV to observe different performances of each classifier more clearly. In general, if n increases in n-fold CV, the error rate tends to decrease proportionally in most classifiers and the variance of error rate is large. It is also important to be aware of the optimal number of folds in practical analysis. Though there is a general strategy that the determination of the number of fold depends on the size of dataset, the choice of fold number tends to be determined empirically in practical analysis and the relationship between internal and external CV has become one of the key issues in machine learning field (Ambroise and McLachlan, 2002; McLachlan, 1992). One of the approaches is the introduction of generalization error which is broken down into three additive parts: noise variance, estimation variance, squared estimation bias (Nadeau and Bengio, 2003). But it is beyond our purpose to discuss detailed approaches, and further study is worthwhile.

For choosing the gene selection methods, some classifiers seem to fit well with specific gene selection method. For classical methods, the error rates based on the Wilcoxon statistic is slightly lower than those from the BSS/WSS or the soft-thresholding method. One possible interpretation of excellence in using the BSS/WSS method is due to the correlation structure among the selected genes. Fig. 2 shows the absolute value of the correlation matrix of LYM selected with (a) BSS/WSS and (b) Wilcoxon statistic method. The brightest color (white) represents the perfect correlation and no correlation is represented with the darkest color (red). By comparing two images, we can easily find a higher correlation among the genes with BSS/WSS method than the Wilcoxon statistic method. It implies that a higher correlation among the predictor variables plays an important role in the performance of the classical methods. For tree methods, the accuracies with the BSS/WSS method are better than those with the Wilcoxon statistics method. In general, recursive partitioning methods with CART use F-statistic as a measure of the separation when the predictor variables are continuous. In making a tree, the most significant gene satisfying the BSS/WSS criterion may contribute to a root node, and it might imply that the gene selection system built in the tree method seems to be theoretically similar to the BSS/WSS gene selection idea.

In applying BSS/WSS criterion for gene selection, we selected 50 most varying genes. Our choice of the number of genes is a kind of arbitrary and based on the

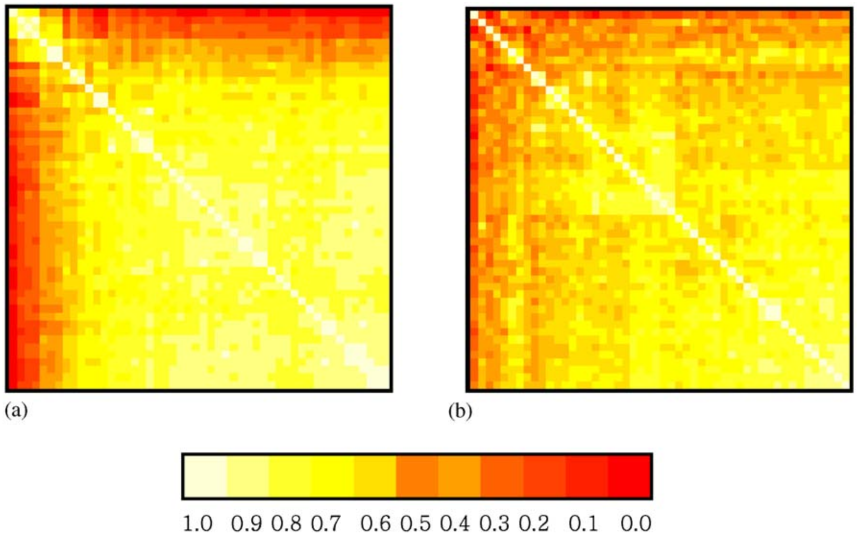

Fig. 2. Absolute value of correlation matrix in LYM with 50 selected genes: (a) BSS/WSS; (b) Wilcoxon rank-based statistic.

previous studies. Golub et al. (1999) found that 50 genes are adequate for leukemia data. Also, Dudoit et al. (2002) used BSS/WSS criterion and selected 50 genes for lymphoma, 40 genes for leukemia and 30 genes for NCI 60. As discussed in Section 6.3 of Dudoit et al. (2002), for the lymphoma or leukemia datasets, increasing the number of variables to 200 genes did not affect greatly the performance of the various classifiers. Throughout our study, we assumed that most of the classifiers are not very sensitive to the number of genes, although they improve slightly with the number of genes. However, determination of the number of genes is a big issue, and further study is also worthwhile.

Acknowledgements

This work was supported by Korea Research Foundation Grant (KRF-2001-015-DP0068).

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Bordrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T.Jr.J.H., Lu, L., Lwis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M., 2000. Distinct types of diffuse large B-cell lyumphoma identified by gene expression profiling. Nature 403, 503–511.

Alon, U., BarKai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sci. 96, 6745–6750.
Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS 99, 6562–6566.
Breiman, L., 1998. Arcing classifiers. Ann. Statist. 26, 801–824.
Breiman, L., 2001. Random forests. Mach. Learning 45, 5–32.
Breiman, L., Friedman, J.H., Olshen, R., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.
Brown, P.O., Borstein, D., 1999. Exploring the new world of the genome with DNA microarrays. Natur. Genetics (Suppl.) 21, 33–37.
Chang, C.C., Lin, C.J., 2001. LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/ ∼cjlin/libsvm.
Dettling, M., Buhlmann, P., 2003. Boosting for tumor classification with gene expression data. Bioinformatics 19, 1061–1069.
Ding, B., Gentleman, R., 2003. Classification Using Generalized Partial Least Squares. Department of Biostatistics, Harvard University.
Dudoit, S., Fridlyand, J., Speed, P., 2002. Comparison of discrimination methods for classification of tumors using gene expression data. J. Amer. Statist. Assoc. 97, 77–87.
Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Clustering analysis and display of genome-wise expression patterns. PNAS 95, 14863–14868.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. 55, 119–139.
Friedman, J., 1991. Multivariate adaptive regression splines (with discussion). Ann. Statist. 19, 1–141.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Statist. 28, 337–407.
Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906–914.
Garber, M.E., Troyanskaya, O.G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., de Rijn, M.V., Rosen, G.D., Perou, C.M., Whyte, R.I., Altman, R.B., Brown, P.O., Botstein, D., Petersen, I., 2001. Diversity of gene expression in adenocarcinoma of the lung. Proc. Nat. Acad. Sci. 98, 13784–13789.
Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.
Hastie, T., Tibshirani, R., 1996. Discriminant analysis by Gaussian mixtures. J. Roy. Statist. Soc. Ser. B 58, 158–176.
Hastie, T., Tibshirani, R., Buja, A., 1994. Flexible discriminant analysis by optimal scoring. J. Amer. Statist. Assoc. 89, 1255–1270.
Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. Ann. Statist. 23, 73–102.
Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, NY.
Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Natur. Medicine 7, 673–679.
Lander, E.S., 1999. Array of hope. Natur. Genetics (Suppl.) 21, 3–4.
Marx, B.D., 1996. Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics 38, 374–381.
McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.
Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Mach. Learning 52, 239–281.
Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Spellman, P., Iyer, V., Jeffrey, S.S., de Rijn, M.V., Waltham, M., Pergamenschikov, A., Lee, J.C.F., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., Brown, P.O., 2000. Systematic variation in gene expression patterns in human cancer cell lines. Natur. Genetics 24, 227–234.

Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O., Weinstein, J.N., 2000. A gene expression database for the molecular pharmacology of cancer. Nat Genet. 24, 236–244.

Tibshiranit, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Nat. Acad. Sci. 99, 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2003. Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. Statistical Science 18, 104–117.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P.O., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, Chichester, GB.

Zuruda, J.M., 1992. Introduction to Artificial Neural Systems. PWS Publishing Company, Boston, NY.