Magdalena Wietlicka-Piszcz
Małgorzata Ćwiklińska-Jurkowska
Department of Theoretical Backgrounds of Biomedical Science
and Medical Informatics, Collegium Medicum in Bydgoszcz,
Nicolaus Copernicus University

PERFORMANCE OF CLASSIFICATION METHODS FOR DIFFERENTIATION BETWEEN CIRRHOTIC TISSUES AND CIRRHOTIC TISSUE WITH CONCOMITANT HEPATOCELULAR CARCINOMA. CLASSIFICATION OF LIVER TISSUES

Abstract: This paper presents the comparison of various discriminant methods for differentiation between cirrhotic tissue and cirrhotic tissue with concomitant hepatocelular carcinoma on the basis of oligonucleotide microarray dataset. Four methods of dimensionality reduction by selection of features (genes) subsets: linear models with empirical Bayes methods [Smyth 2004], SAM, PAM and Wilcoxon statistic were implemented. For studied subsets of genes ranked by these methods the performance of seven discriminant procedures was estimated by test error, 10-fold CV and bootstrap 0.632 with 95% confidence intervals. The best performance was obtained for SVM, bootstrap aggregating trees as well as adaptive boosting trees.

Key words: classification, hepatocelular carcinoma, microarrays

Introduction

Microarray technology is a new promising tool allowing simultaneous investigation of expression levels of thousands or tens thousands of genes. The elaboration of data from microarray experiments involves the use of methods which enable application when the number of features (genes) is much larger than the number of samples (microarrays). The issue of microarray data classification had been reported in many studies (see, e.g., Boulesteix et al., 2008, Dupuy and Simon 2007, Dudoit et al., 2002; Lee et al., 2005; Statnikov et al., 2005; Van Sanden et al. 2008). The results indicate that, although a few methods like random forests or support vector machines seem to perform better than others, there is no single method that would be suitable for all applications.

ISBN 978–83–7431–266–0 ISSN 0860-150X

Because of the very high number of genes in one microarray the pre-selection of features-genes for inclusion into the classification rule may be an important issue. Lee et al. (2005) mention that various methods of active gene selection applied to the same set of microarray data may give different sets of genes and consequently lead to different discrimination results.

In the present paper the data concerning liver tissue with HCV infection are investigated. The purpose of this study was to investigate the performance of various statistical discrimination methods for classification of liver tissues such as: cirrhotic tissue and cirrhotic tissue from patients with hepatocellular carcinoma.

Materials and methods

Data from high density oligonucleotide arrays (Mas et al. 2009, public repository Gene Expression Omnibus, accession GSE1423) were used for investigation of differentially expressed genes in liver tissue samples.

The total number of 58 samples from liver tissue with HCV infection were examined. 41 cirrhotic tissue samples were from patients without hepatocellular carcinoma (Cirrhosis) and 17 cirrhotic tissue samples were from patients with hepatocellular carcinoma (CirrhosisHCC). Each microarray consisted of 2227 probe sets. All microarrays were divided into a learning set and a testing set, of respectively 40 and 18 microarrays.

The investigated data were previously preprocessed, so for each probe expression summaries were available. The analysis was performed with the use of R and Bioconductor package.

Genes Selection Methods

To identify genes differentially expressed, i.e. genes that exhibit a statistically significant difference across the examined tissue types, fours methods of gene selection were applied. T-statistic is widespread in assessing differential expression in microarray experiments, therefore two methods basing on t-statistic were used: a method based on linear models and moderated t-statistic reported as (*called*) Linear Models for Microarray Data (LimmaBH) introduced by Smyth (2004) and Significance Analysis of Microarrays (SAM) proposed by Thuser et al. (2001). Also a nonparametric test, a Wilcoxon rank sum test was applied to identify differentially expressed genes. Another approach to gene selection is presented by Prediction Analysis of Microarrays (PAM) introduced by Tibshirani et al. (Tibshirani, 2002). It is based on the nearest shrunken centroid classifier and provides a set of genes that best characterizes each class.

LimmaBH is a method based on the linear model and the empirical Bayesian method (Netwon et.al 2001) applied to estimate fold changes of differential expression and on moderated t-statistics to rank genes according to the probability of differential expression. The tests are also adjusted for multiplicity by using the Benjamini & Hochberg method (Benjamini and Hochberg, 1995) to control false discovery rate (the expected proportion of type I errors).

The results of a series of microarray experiments can be represented as a $m \times n$ matrix where m represents the number of genes and n the number of microarrays ($n = n1 + n2$). Let y_{gij} denote the expression level of gene g in array i from group j.

Significance Analysis of Microarrays generates a list of genes ranked according to the probability of differential expression on the basis of the modified t-statistic.

In this method the genes are ranked according to the probability of differential expression on the basis of the modified t-statistic:

$$t_g = \frac{\bar{y}_{1_g} - \bar{y}_{2_g}}{s_g + c}$$

where

$g = 1, \ldots, m$;

s_g is the standard deviation of repeated expression measurements.

This modified t-statistic has t-Student distribution with $n - 1$ degrees of freedom. When the variability measured by s_g is close to 0, the values of classical t-statistic can become too large. So s_g in the denominator is augmented by a small positive constant c. Its value is chosen to minimize the coefficient of variation of the test statistic. This constant ensures that the variance of the score t is independent of gene-expression.

All gene selection methods were applied for the learning set (LS), which consisted of 40 arrays.

Classification methods

The four sets of genes obtained from the gene selection methods were used in the construction of discrimination rules by applying different discrimination methods. The following methods were considered: support vector machines (SVM), diagonal linear discriminant analysis (DLDA), diagonal quadratic discriminant analysis (DQDA), k nearest neighbour (k-NN) and classification trees. Additionally, the methods based on ensemble classifiers such as adaptive boosting and bagging trees were applied (Duda et al. 2001, Webb 2002, Dettling 2004).

For each of the four sets of genes the discrimination methods were applied to subsequently enlarged sets of genes, which included $2, 3, 4, \ldots, 100$ of the highest-ranked genes.

Parametric discriminant methods DLDA and DQDA are special cases of classical linear and quadratic discriminant functions, created by the assumption that the features are independent within each class, so the within-class covariance matrix is diagonal. In nonparametric k-NN method the parameter k (number of neigbours) was chosen for each subset of genes according to the criterion of minimization of CV error.

Support Vector Machines (SVM) method [1] separates 2 data groups by the hyperplane defined on the basis of the criterion that maximize the margin between groups. The Support Vector Machines classification technique is based on mapping the data to represent observations in high dimensional space – usually much higher than the original feature space.

The main advantage of the Support Vector Machines is that complexity of the classifier is determined by the number of support vectors – observations lying on the margin – rather than the dimensionality of the transformed space. As a consequence, SVM have less often problems with overfitting than many other methods and is appropriate for high dimensionality.

The SVM method (Cortes and Vapnik 1995) was applied with the regularizing constant equal to 0.5. The best performance of SVM was obtained for linear kernels, so we present only linear kernel results. The outcomes are consistent with the conclusions of other authors. However, some authors like Noble (2004), find other outcomes – he stated that the SVM using third-degree polynomial kernel was the best performing method.

Combining classifiers based on resampling (randomly generating learning sets)

The single classifiers built on relatively small learning data set are often biased. Then, the methods based on randomly generating subsets of training data joined with combining classifiers built on them may be useful and may improve the performance.

Bagging (Bootstrap AGGregatING) Breiman [1996] is an ensemble based on bootstrap samples created by drawing n times from the learning set with possible replacements, where n means the size of the learning set. The classifier is trained on each bootstrap subsample. Resulting classifiers are then combined, e.g. by the averaging the posterior probability or the unweighted majority vote.

Boosting is also the method based on resampling the learning data set; however boosting is a deterministic procedure, because the selection of

subsequent subsamples depends on the results of combined classifier performance achieved in previous loops. In sequentially generated learning sets the weights of misclassified cases are increased so the ensemble creates the improved classifiers.

"Boosting" the performance of weak classifiers is originated from Freund & Schapire [1996] ARCing-Adaptive Resampling and Combining. The most popular boosting method is Adaptive Boosting (AdaBoost). AdaBoost allows the designer to continue adding classifiers until some desired low training error is achieved. Bagging and boosting are the methods most often used for weak base classifiers, and trees as constituent classifiers are used.

Errors evaluation

In microarray experiments the appropriate classification error estimation is a very important issue. Although in the case when the amount of data is relatively small, which is typical for microarray experiments, the estimation of error rate is not straightforward. The evaluation of the error rate by the test sample error can be biased. Therefore, the assessment of the constructed discrimination procedures was performed by the error estimation on the test set, by 10–fold cross-validation (CV) [Mc Lachlan 1992] applied to the whole dataset and by the bootstrap 0.632 method [Efron 1983] based on 100 randomly generated samples. For 10 fold CV and bootstrap 0.632 mean error rates the 95% confidence intervals were estimated additionally.

Results and discussion

The identification of differentially expressed genes was performed by the use of the four features selection methods mentioned above and consequently four sets of differentially expressed genes were obtained. In each set genes were ordered according to the appropriate test statistic.

The first highest-ranked 100 genes produced by each of the gene selection methods were taken into further consideration, because the increased number of genes over the first 100 most important ones does not result in the improved classification performance. Therefore, further analysis was restricted to the first 100 genes obtained from each of the considered gene selection methods.

Fig. 1. presents the Venn diagrams illustrating all pair wise comparisons among the four genes sets. 27 genes are common across all the subsets. The highest overlapping of genes present sets produced by SAM (in further analysis called set2) and PAM (set3) methods (96 genes). The remaining

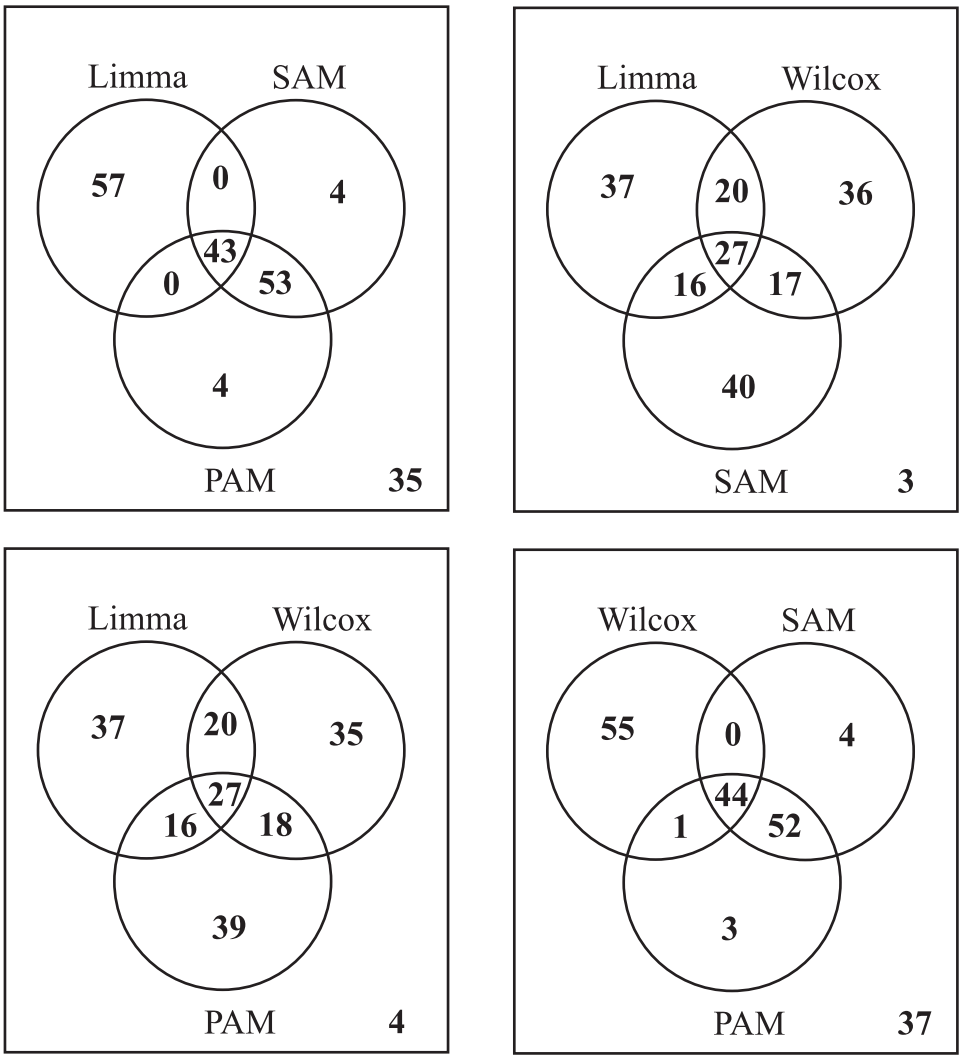

Fig. 1. Venn diagrams illustrate all pair wise comparisons of overlapping among the four gene sets produced by considered gene selection methods like LimmaBH, SAM, PAM and Wilcoxon

pair wise comparisons show overlapping of about half of the total amount of genes. Because of the high overlapping of genes subsets obtained from SAM and PAM methods, the classification results will be shown just for the SAM method. The set of genes obtained by LimmaBH method and ranked by Benjamini-Hochberg procedure will be called set1, and set4 will denote the set of genes ordered according to the nonparametric Wilcoxon statistic.

For each of the four gene sets the misclassification errors were estimated for subsequently enlarged gene sets, which included $2, 3, 4, \ldots, 100$ of the highest-ranked genes.

For each of the selection methods the results are presented in pairs of two figures: the first for the methods not connected with classification tress and the other one for trees.

Figures 2–3 show the misclassification errors estimated by using the test dataset for different discrimination methods applied to sequentially enlarged set of genes of *set1 (LimmaBH)*. Figures 4–5 show the misclassification error rates estimated by 10–fold cross-validation (CV) for different discrimination methods applied also to set 1 and fig 6–7 present for the same genes sets the error rates estimated by the bootstrap 0.632 method.

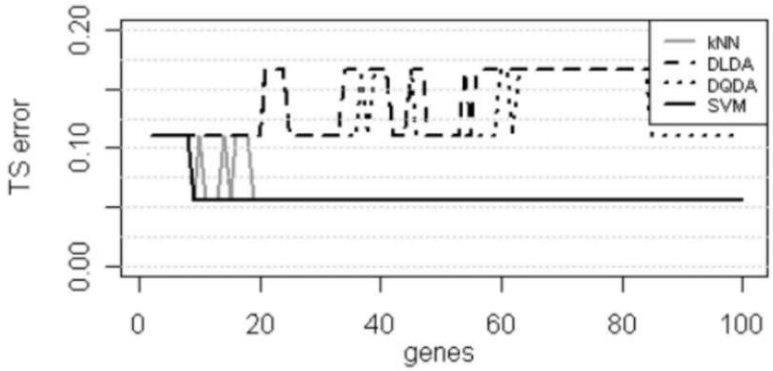

Fig. 2. Classification test errors of discriminant methods: k-NN, DLDA, DQDA, SVM for ascending subsets of *set1 (Limma)*, from 2 to 100 genes

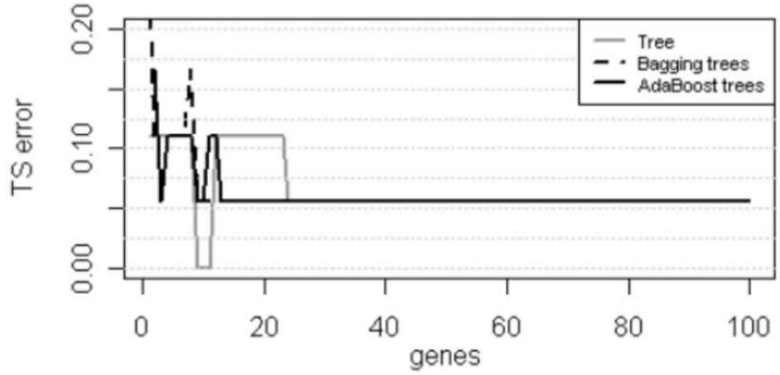

Fig. 3. Classification test errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set1* (Limma), from 2 to 100 genes

Test sample errors are between 0.056 and 0.167 for KNN, DLDA, DQDA and SVM (fig. 2). The smallest test errors equal 0.056 were obtained for both SVM and kNN for subsets containing 20 or more genes. Test errors for both bagging and AdaBoost ensemble tree and for single tree classifiers are also equal to 0.056 for more than 20 genes (fig. 3). For bagging and boosting no optimal subset of genes may be pointed. Though for tree classifier the minimum test error reached zero for about 15 genes, the tree is unstable and can be overtrained, so CV errors and bootstrap 0.632 do not confirm this result (fig. 5, 7).

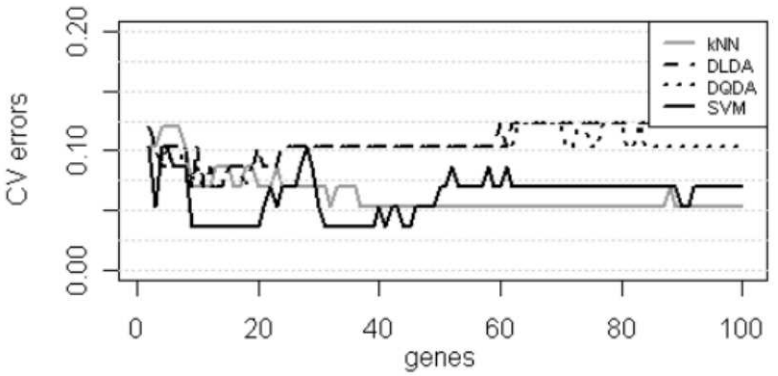

Fig. 4. Classification cross-validation errors of discriminant methods: k-NN, DLDA, DQDA, SVM for ascending subsets of *set1 (Limma)*, from 2 to 100 genes

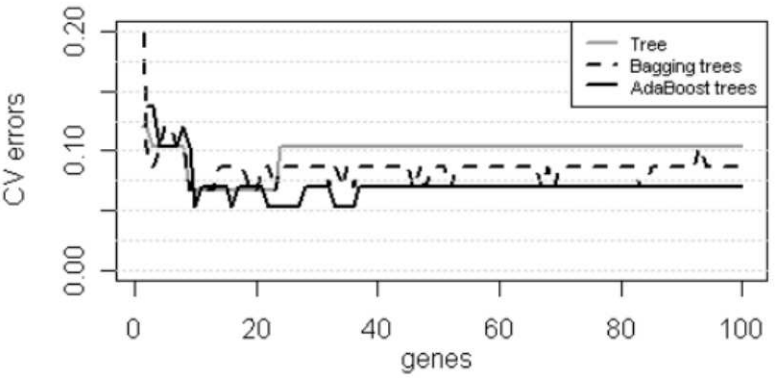

Fig. 5. Classification cross-validation errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set1 (Limma)*, from 2 to 100 genes

The representative set of genes, indicated by the smallest errors of SVM classifier, was between 8 and 20 and from 30 till 40 genes (fig. 5) – where the CV error reached the value equal to 0.037. For more then 50 probe sets the CV error for SVM is increasing, in the contrary to kNN classifier (Fig. 4). DLDA and DQDA gave higher CV errors.

CV errors of boosting trees are between 0.037 and 0.1 and for bagging are from 0.053 to 0.137. Bagging and boosting measured by bootstrap 0.632 error gave, similarly to CV error, better results (*smaller errors*) than single tree (comp. fig. 5 and 7). From Fig 7 the optimal subset for ensemble tree classifiers cannot be pointed.

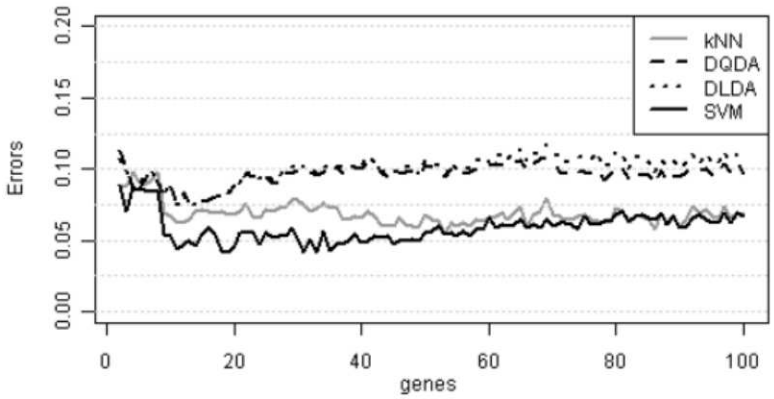

Fig. 6. Bootstrap 0.632 classification errors of discriminant methods: k-NN, DLDA, DQDA and SVM for ascending subsets of *set1* (Limma), from 2 to 100 genes

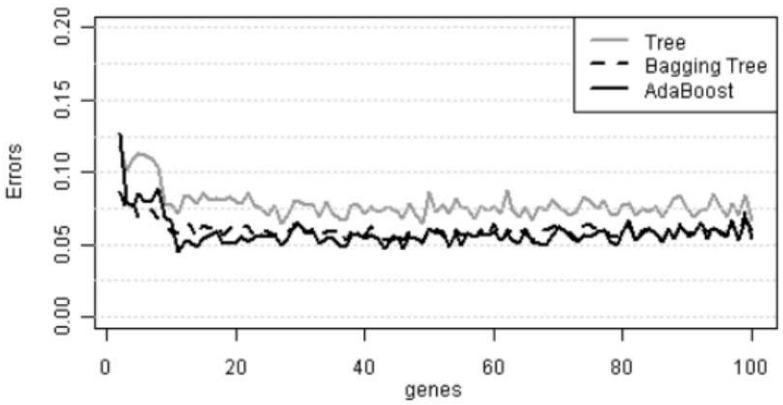

Fig. 7. Bootstrap 0.632 classification errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set1 (Limma)*, from 2 to 100 genes

Bootstrap 0.632 errors show smaller fluctuations than for CV errors across the varying number of genes and the values are between 0.04 and 0.12.

For each classification method and for each subset of set1 (LimmaBH) the standard errors were calculated. Table 1 presents the minima, maxima and mean values of standard errors for CV and bootstrap estimation of classifiers for all the subsets of set1 (LimmaBH).

Table 1

The standard errors for the two methods of error assessment

classification method	CV-10			bootstrap 0.632		
	min	max	mean	min	max	mean
kNN	0,027307	0,038233	0,030806	0,002390	0,004054	0,003246
DQDA	0,028306	0,037663	0,035850	0,002021	0,003269	0,002545
DLDA	0,028306	0,037663	0,036003	0,002094	0,003077	0,002641
SVM	0,024570	0,051926	0,028482	0,002384	0,003997	0,003332
Bagging Tree				0,002325	0,004959	0,003445
Tree				0,002708	0,005653	0,003991
AdaBoost				0,002727	0,005680	0,003543

Cross-validation errors (fig. 4–5) seem to suggest that the addition of only one or a few genes can significantly alter the performance – it can improve or worsen the performance. The CV error difference after adding only one or a few genes reaches even 0.1. The CV error can increase even two times. Such differences are not confirmed by a more stable method as bootstrap 0.632 error estimate (Fig. 6–7). On the basis of 95% confidence intervals for CV errors we can not conclude that any examined classification method significantly outperforms the other. For the bootstrap 0.632 assessments the errors have much smaller fluctuations than for CV–10 errors (Fig. 6–11). Therefore, for the comparison of the genes selection methods and classifiers in further analysis we will use bootstrap 0.632 error assessments. The ranges of standard errors for mean error rates evaluated by bootstrap 0.632, presented in Table 1, indicate much smaller diversity than CV error, so the bootstrap error assessment is more precise. Also Braga-Neto and Dougherty (2004) concluded that the cross-validated estimators show high variance and large outliers. The large outliers produced by the cross-validation estimators can cause that significantly inaccurate conclusions can be reached for a considered data set. The authors also concluded that the bootstrap estimators, in particular the bootstrap 0.632 estimator, display the best overall performance [Braga-Neto, 2004].

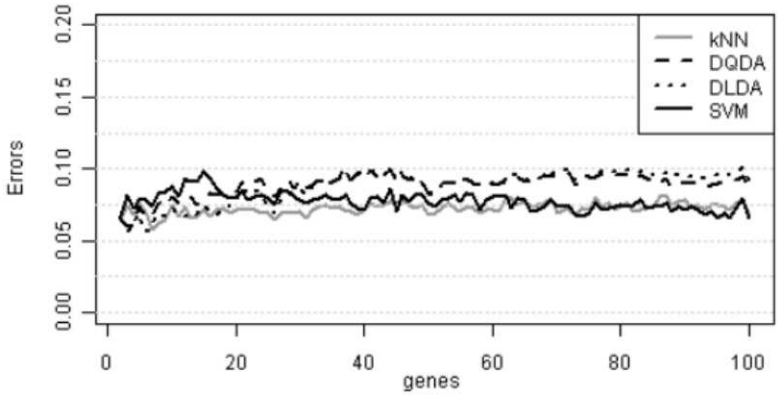

Fig. 8. Bootstrap 0.632 classification errors of discriminant methods: k-NN, DLDA, DQDA and SVM for ascending subsets of *set2* (SAM), from 2 to 100 genes

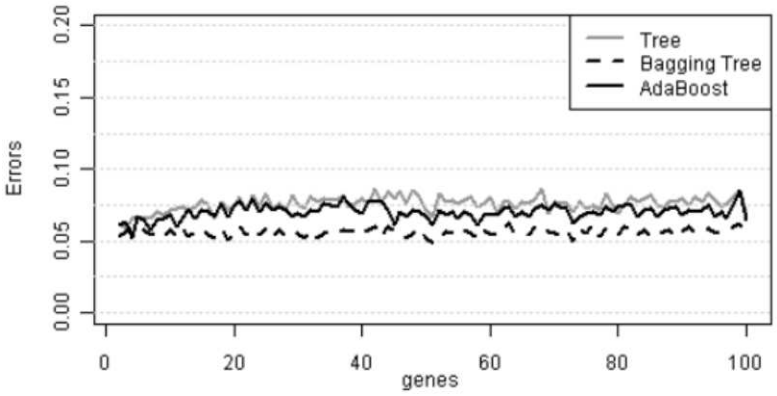

Fig. 9. Bootstrap 0.632 classification errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set2* (SAM), from 2 to 100 genes

For set1 (LimmaBH) the smallest bootstrap 0.632 assessment of error rates were reached for SVM method, for the number of genes from 18 till 45. Figures 8–11 show the misclassification errors estimated by the bootstrap 0.632 method for set2 (SAM) and for set 4 (Wilcoxon).

Bootstrap 0.632 errors for DLDA and DQDA for number of genes bigger than 40 are relatively high. From Figure 9 and 10 and using appropriate SE from Table 1 we can conclude that significantly smaller results, measured by the confidence interval of bootstrapping 0.632 error, were obtained for ensemble tree than for other examined classifiers.

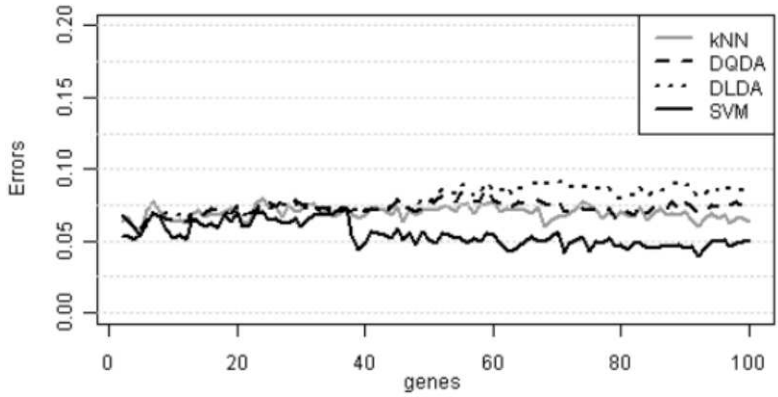

Fig. 10. Bootstrap 0.632 classification errors of discriminant methods: k-NN, DLDA, DQDA and SVM for ascending subsets of *set4* (Wilcox), from 2 to 100 genes

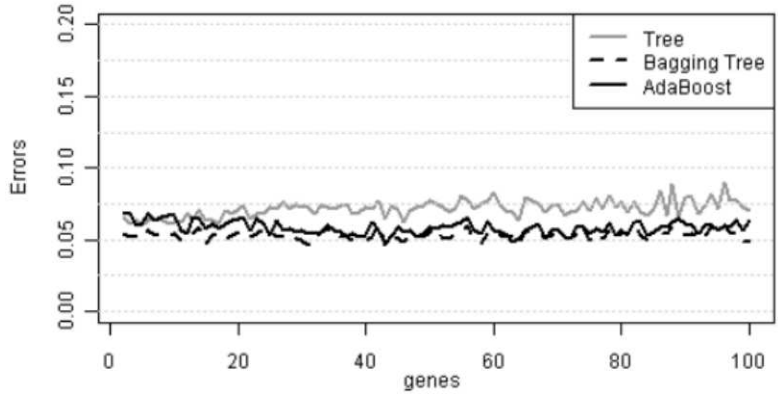

Fig. 11. Bootstrap 0.632 classification errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set4* (Wilcox), from 2 to 100 genes

The analysis of the bootstrap 0.632 error rates on Fig. 7, 9, and 11 for three sets: set1 (LimmaBH), set2 (SAM) and set3 (Wilcoxon) show that the differences between error rates are relatively small. However, for set1 and the SVM the bootstrap 0.632 obtain smallest values close to 0.05 (with small variability from 0.041–0.059) for the range between 10 and 45 probe sets.

For Wilcoxon statistic criterion selection, we obtained significantly smallest errors for SVM, however for number of genes bigger than 40 (Fig. 11).

For resampling classifiers we can not point to any optimal subset of genes for adaptive boosting and bagging while for other methods we can see the significant differences between bootstrap errors for consecutive subsamples, for example for SVM.

For set1 (LimmaBH) and set2 (Wilcoxon) the adaptive boosting and bagging tree errors oscillate around 0.05, for set3 bagging tree is also 0.05 but Adaboost is 0.07. Thus for ensemble classifiers set1 and set3 seem to have similar effectiveness. The smallest bootstrap errors – obtained for SVM and for set1 – are about 0.05 for gene sets from 10 to 40 genes, while for set3 errors are about 0.05 for 40 or more genes. For SAM and PAM all bootstrap errors are for all considered classifiers over 0.05.

The comparison of our results with results obtained by Mas et al. 2009 is not straightforward. Those authors performed the selection of genes on the whole data set (all 58 microrarys), not on the smaller subset (learning set), so the classification results based on their selection can be optimistically biased [Braga-Neto and Dougherty 2004]. This issue was raised by many researchers [e.g. Ambroise et al. 2002, Wood et al. 2007]. In our work the differentially expressed genes were selected from the learning set which consisted of 40 microarrays, so the results [*from Mas et al. and presented in current work*] are hard to compare. The authors obtained for random forest out-of-bag error equal to 0.089 for one subset containing fifteen genes selected by Gini index. They also applied logistic regression, where two variables were chosen, obtaining resubstitution error equal 0.036, though resubstitution error is known as optimistically biased.

Conclusion

For considered data set (Mas et al. 2009) concerning discrimination between cirrhotic tissue and cirrhotic tissue with concomitant hepatocellular carcinoma the SVM method, adaptive boosting and bagging gave the best classification results. However, this conclusion could not be valid for another dataset, because the differences in error rates estimates are relatively small and sometimes not significant. For microarray data classification problems the application of several classification methods coming from different sources can be useful. The comparison of considered gene selection methods seems to be a difficult task because they give similar error rates. Basing on SVM classifier results LimmaBH method for genes' selection can be recommended. However, for gene set obtained by the Wilcoxon test the same level of error was reached, although for a bigger number of genes. It is certain that further investigations in this area are necessary.

REFERENCES

[1] Alon, U. et al. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. PNAS, Vol 96, p 6745–6750.

[2] Ambroise C. Mc Lachlan G. 2002 Selection bias in gene extraction on the basis of microrray gene expression data. PNAS vol 99 No 10 pp 6562–6566.

[3] Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series* B, **57**, 289–300.

[4] Boulesteix A.L., Strobl C, Augustin T and Daumer M. (2008) Evaluating Microarray-based Classifiers: An Overview, Cancer Informatics: 6 77–97.

[5] Breiman L. (1996). "Bagging predictors". Machine Learning 24 (2): 123–140.

[6] Cortes C., Vapnik V. Support-vector network. Machine Learning, 1995, 20, 1–25.

[7] Dettling M. (2004): BagBoosting for tumor classification with gene expression data. Bioinformatics, 20 (18), 3583–3593.

[8] Duda O. R, Hart P. O., Stork D. G.: Pattern Classification. Wiley & Sons. 2001.

[9] Dudoit S., Fridlyand J., and Speed T. P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 2002, 98, 77–87.

[10] Dupuy, A. and Simon, R. 2007. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. Journal of the National Cancer Institute, 99:147–57.

[11] Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.*, 78, 316–331.

[12] Golub et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, Vol. 286: 531–537.

[13] Lee J. W., Lee J. B., Park M., and Song S. H.: Extensive comparison of recent classification tools applied to microarray data. Computational Statistics and Data Analysis. 2005, 48, 869–885.

[14] Mas VR, Maluf DG, Archer KJ, Yanek K et al. (2009) Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. Mol Me,d Mar-Apr; 15 (3–4): 85–94.

[15] McLachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley.

[16] Newton M. A., Kendziorski C. M., Richmond C. S., Blattner F. R., Tsui K. W. (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. Journal of Computational Biology, 8 (1): 37–52.

[17] Noble W. S. Support vector machine applications in computational biology In: Kernel methods in computational biology. Ed: Scholkopf et al. 2004 Massachusetts Institute of Technology.

[18] Pomeroy, S. et al. (2002) Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression. Nature, Vol 415, p 436–442.

[19] Smyth, G. K. (2004): Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, Vol. 3, No. 1, Article 3.

[20] Statnikov A., Aliferis C.F., Tsamardinos I., Hardin D., and Levy S. (2005): A comprehensive evaluation of multicateogry classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21, 631–643.

[21] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of Webb A. R. Statistical Pattern Recognition (2002), New York: Oxford University Press.

[22] Tusher, V., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America, 98, 5116–5121.

[23] Ulisses M. Braga-Neto, Edward R. Dougherty (2004), Is cross-validation valid for small-sample microarray classification?, Bioinformatics, v. 20 n. 3, p. 374–380, February 2004.

[24] Van Sanden, S., Lin, D., and Burzykowski, T. (2008): Performance of gene selection and classification methods in a microarray setting: A simulation study. Communications in Statistics – Simulation and Computation, 37, 418–433.

[25] Webb A. R. Statistical Pattern Recognition (2002), New York: Oxford University Press.

[26] Wood I Vissher P. Mengerstom K. L. (2007) Classification based upon gene expression data: bias and prediction of error rates. Bioinformatics Vol 23 No 11 pp. 1363–1370.

[27] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30** (4): e15.