Biometrical Letters
Vol. 48 (2011), No. 1, 41-54

The performance of discriminant analysis for differentiating between genotoxic and non-genotoxic carcinogens

Małgorzata Ćwiklinska-Jurkowska[1], Tomasz Burzykowski[2], Magdalena Wietlicka-Piszcz[3]

[1]Collegium Medicum, Nicolaus Copernicus University, Dept. of Theoretical Foundations of Biomedical Sciences and Medical Informatics, Bydgoszcz, Poland, mjurkowska @ cm.umk.pl
[2]I-BioStat, Hasselt University, Diepenbeek, Belgium, tomasz.burzykowski@uhasselt.be
[3]Collegium Medicum, Nicolaus Copernicus University, Bydgoszcz, Poland, mpiszcz@cm.umk.pl

SUMMARY

The aim of the present study was to examine the use of a variety of statistical discriminant functions in the classification of genotoxic and non-genotoxic carcinogens. To this purpose, the data from an experiment conducted by van Delft *et al.* (2005) were used. The investigated methods included DQDA, DLDA, boosting trees, bagging trees, bagboosting trees, KNN, and SVM. Two gene selection methods were examined: first using tests based on a linear model (Smyth *et al.*, 2004), with a multiple-testing correction of the resulting p-values, and the second based on the tests applied to re-sampled datasets. The outcomes suggest that, when the discrimination between genotoxic and non-genotoxic carcinogens is of interest, the proper choice of discrimination method is essential. Misclassification errors may also be a confirmation of the correctness of gene selection methods.

Key words: Classification, discriminant analysis, misclassification errors, genotoxicity, carcinogenic chemical compounds, microarray data, gene expression data

1. Introduction

Chemical compounds can be genotoxic. Genotoxicity, in turn, may result in carcinogenicity. Therefore the screening of chemical compounds for their genotoxicity is an important issue for the control of risk of cancer. An important question related to the problem of screening of chemical compounds for their

genotoxicity is that of which genes are affected most after exposure to genotoxic compounds.

Because microarrays enable the simultaneous investigation of the expression levels of thousands of genes, they are a useful tool for discrimination between genotoxic and non-genotoxic compounds.

Discrimination between classes is a well-known problem in statistical methodology. There are many methods available for this purpose. However, the majority of them were developed under the assumption that the number of features (variables) which can be used to build a discrimination (classification) rule is smaller than the number of observations on whose basis the rule can be constructed. In the microarray context the situation is different: the number of features (genes) is much larger than the number of observations (arrays).

Hence when discrimination based on microarray data is considered, the choice of a suitable discrimination procedure becomes an important issue. Recently the issue has attracted considerable attention (see e.g. Dudoit *et al.*, 2002; Lee *et al.*, 2005; Statnikov *et al.*, 2005; Van Sanden *et al.* 2007, 2008). The results indicate that, although a few methods – such as random forests or support vector machines – seem to perform better than the others, there is no single method that would be suitable for all applications.

Additionally, it has been reported that the selection of genes for inclusion in the discrimination rule may be an important issue as well. Lee *et al.* (2005) mention that various methods of selection of active genes applied to the same set of microarray data can give different sets of genes and consequently lead to different discrimination results.

From these reports it is clear that the result of discrimination between genotoxic and non-genotoxic compounds based on microarray data may depend on the applied gene selection and discrimination methods.

Van Delft *et al.* (2005) applied several techniques (Pearson correlation analysis, nearest shrunken centroids analysis, K-nearest neighbour analysis, and weighted voting) to discriminate between sets of 11 genotoxic carcinogens and 9 non-genotoxic carcinogens based on microarray data. They did not observe

clear differences between the results obtained for the different methods (apart from the poorer performance of Pearson analysis). The aim of the current study is to extend the investigation conducted by van Delft *et al.* (2004, 2005) by considering alternative gene selection and discrimination methods.

2. Materials and methods

2.1. Microarray experiment

The influence of carcinogenic compounds on the expression levels of a set of 596 genes obtained from HepG2 microarrays was investigated. Each gene was spotted four times on each array. Twenty carcinogens were used in the study: 11 of them were genotoxic (GTX) and 9 were non-genotoxic (NGTX). The dataset included 44 microarrays, obtained from 22 dye-swap comparisons. In each comparison, a carcinogen was compared to a control (a solvent). The microarrays were divided into two sets (see Table 1): a training dataset (32 microarrays, with 16 carcinogenic chemical compounds) and a testing dataset (12 microarrays, with 6 carcinogenic chemical compounds). BaP and DEHP were used both in the training and in the testing sets. A detailed description of the experiment can be found in van Delft *et al.* (2004).

Table 1. Chemical compounds (after van Delft et al 2004, 2005)

Chemical treatment	GTX class /NGTX	Chemical treatment	GTX class /NGTX	Chemical treatment	GTX class /NGTX
Training set		Training set		Testing set	
MMS	ALK	MMC	X-LINK	DMN	ALK
NMU	ALK	DEHP-1	NGTX	BaP-2	PAH
NNK	ALK	DIOX	NGTX	carboPt	X-LINK
BaP-1	PAH	PCP	NGTX	DEHP	NGTX
DBA	PAH	PhB	NGTX	RES	NGTX
FA	PAH	TCDD	NGTX	TCP	NGTX
cisPt	X-LINK	TCE	NGTX		
CPh	X-LINK	TPA	NGTX		

2.2. Microarray data normalization, gene selection, and discrimination methods

Normalization

The raw data in the ImaGene format were transferred to R and Bioconductor. Poor spots, i.e. spots that were flagged by ImaGene during the experiment, were not included in the further analysis.

First, the intensity values were corrected for background using a convolution of the normal and exponential distributions, where the normal part represents the background and the exponential part represents the signal (Smyth, 2005).

Next, within-array normalization of background-corrected intensities for chemical compounds vs. the control (solvents) was carried out using the print-tip loess procedure (Yang et.al., 2002). The normalized intensity values were used to obtain logarithmic transforms of the ratios of red (Cy3) and green (Cy5) intensities (M values).

Finally, for each gene, a linear model was fitted to the M values (Smyth 2004) obtained from the training dataset. The aim of this step was (1) to adjust for the dye effect and (2) to choose genes differentially expressed in GTX and NGTX groups. The model contained the indicator variables for the type of the carcinogen (GTX vs. NGTX) and for the dye (Cy5 vs. Cy3). Because each array contained 4 replicates of each gene, the between-replicate correlation was incorporated into the model (Smyth et. al., 2005). The M values were corrected for the dye effect by subtracting the estimate obtained from the model.

Gene selection

Based on the linear model, subsets of differentially expressed genes, i.e. genes with a statistically significant, at the two-sided 5% significance level, difference in expression for the GTX and NGTX carcinogens, were obtained by applying a t-test. The tests were adjusted for multiplicity using the Benjamini & Hochberg method (BH; Benjamini and Hochberg, 1995).

The choice of differentially expressed genes was also made by applying the linear model to bootstrap samples. Bootstrap samples were created by sampling with replacement from the set of 44 microarrays. The resampling procedure was repeated 1000 times. For each bootstrap sample a linear model, as described in Section 2.2, was fitted, with the use of the BH multiple-testing procedure. For each gene, the mean of the adjusted p-values was calculated and the genes were ordered according to the mean.

2.3. Correlations used for discriminating between GTX and NGTX compounds

For such selected genes, the values of ratios of gene expression logarithms from each matrix, representing the individual substance (microarray) GTX (and correspondingly NGTX) with "averaged matrices" of the group GTX (respectively NGTX), belonging to the learning group, were investigated in terms of correlation. Classification in the discriminated group of chemical compounds may be made to the group with the largest Pearson correlation coefficient.

2.4. Discriminant procedures

Two sets of genes (see Table 2) 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), bagging trees, adaptive boosting trees, and bagboosting trees (Webb 2002, Dettling 2004). For each of the two sets of genes presented in Table 2, the discrimination methods were applied to subsequently enlarged sets of genes, which included 2, 5, 10, 15, …, 100 of the highest ranked genes.

Evaluation of the misclassification error for the constructed discrimination procedures was performed by estimation of the error rate on the test set and by 8-fold cross-validation applied to the whole dataset.

Table 2. Sets of most important genes differentiating the GTX and NGTX carcinogens.

Linear model - LS set1	SAT, ACTG1, TYMS, BAX, AHR, JAG1, ANXA5, SMPD2, VMP1, RPS19, MT1X, PCNA, BHMT2, PIM1, AUTL1, RPL13 , ORM1, ABCC3, SLC2A1, CTSB, TTR, ID1, CASP8, MCL1, APOC3, HAMP, MYC, MDM2 , C4BPA, ITGB1, PSME1, AMD1, JUND, POLB, HIF1A, ODC1, VR22, CDK3 , PPP3R1, PTGS1, Calreticulin, SOD1, CEACAM6, FUT1, RPL13A, ZNF9, DNCL1, CDKN1A, CHK, MT2A, ACTB, PBP, TK1, NF1, SHB, SULT1A3, FDXR, AMACR, TFRC, YWHAZ , HPRT1 , FGA, KIAA0101, PTK2, ALB, ERCC1, PRDX1 , ADM, SERPINA3, PC326, PTPN9, RPL13 , ZFP36, PTMA, LIPC, CDH2, SEC61A2, BHMT, TIMP1, TOP1, RAD52, BIRC3, NDUFS8, SERPINB2, MMP3, P311, ABCG2 , YWHAZ , PSMD3, AMBP, ADH4, DBI, SERPINA1, CSF1R, APG-1, RAD23B, MRPL40, UGT1A1, HYOU1, CASP3
Linear model –bootstrap sample - set2	TYMS, ACTG1, AHR, BAX, PCNA, JAG1, BHMT2, SAT, MYC, CASP8, PIM1, ANXA5, AUTL1, ORM1, DNCL1, ABCC3, TTR, HAMP, POLB, ODC1, ACTB, ID1, CTSB, PTGS1, GAPD, KIAA0101, PRDX2, MDM2 , RPS19, Calreticulin-, SHB, VMP1, SMPD2, PRDX1 , PTK2, FGA, PSME1, MT1X, AMD1, SOD1, PPP3R1, CDKN1A, T-cell cyclophilin, SLC2A1, CHK, PTMA, SULT1A3, FUT1, CEACAM6, ZFP36, ITGB1, UGT1A1, TK1, ATF3, RPL13 , PBP, ABCG2 , BHMT, STMN1, FDXR, JUND, NF1, MCL1, PC326, PPARA, HSPA5, PGRMC1, TOP1, MRPL40, SERPINA3, ERCC1, C5R1, CDK3 , PSMD3, APOC3, HIF1A, SULT1C1, AMBP, VR22, SERPINB2, TPH, P311, GADD153, PRDX1, EPHX1, TUBA, CGI-45, ACAA1, HPRT1 , NDUFS8, LGALS3, CD66e, HRAS, TIMP1, PTPN9, ENC1, APG-1, C4BPA, RODH-4, SEC61A2
van Delft	AHR, SERPINB2, ACTGP3, CASP8, JAG1, PPP3R1, MYC, CALB1, ZFP36, SAT, ODC1, HIF1A, MTIE, SLC26A3, HDAC1, COL15A1, PCNA, TYMS, UGTIA3, BAX, CDKNIA, UCP2, CEACAM6, TTR, NOL5A, CASP4

3. Results

3.1. Comparison of the selected sets of genes

Table 2 presents the highest ranked 100 genes obtained by each of the gene selection methods described in Section 2.2.

The first row presents the set of the 100 most significant genes ordered by adjusted p-value obtained from the linear model, which was applied to 32 matrices from the learning set. In what follows, this set is called *set1*.

The second row of Table 2 presents the 100 highest ranked genes obtained by the use of linear model combined with resampling, as described in Section 2.2. In what follows, this set is called *set2*.

Note that the order of genes included in *set1* and *set2* differs. This affects the discriminating power of the successive subsets (see Section 3.2). There are 77 genes contained in both *set1*and *set2*.

The third row of Table 2 presents the 27 genes selected by van Delft *et al.* (2005). In what follows, this set of genes is called *set3*. There are 17 genes from *set3* which are also contained in *set1* and *set2*.

3.2. Investigation of correlations of gene expressions on microarray spots

Table 3 presents the results of analysis of correlations between mean expression on four repeated spots for each gene and the corresponding values of gene expression levels averaged over all GTX (or NGTX) microarrays (see Section 2.3 and Table 1). Discrimination between GTX and non-genotoxic group is done here on the basis of the higher of two considered correlations connected with the chemical compound in question.

If classification based on larger correlation coefficient is non-concordant with the learning vector (real class belonging), the appropriate microarray is denoted by a star with the description of the GTX / NGTX group. From Table 3 we can see that the two misclassified GTX microarrays belong to the PAH genotoxic group. One from the NGTX group has important, discordant difference in correlations.

The analysis points to microarrays which are atypical, difficult to classify based on correlation analysis. Those microarrays are also difficult to classify by the examined discriminant methods whose summarized results are presented in the next section (without considering in detail the microarrays which are classified incorrectly).

The most difficult to discriminate were microarrays from GTX: FA (Fluoranthene) and GTX BaP-2 (Benzopyrene). Both FA (Fluoranthene) and Ba-P (Benzopyrene) are chemical compounds from the GTX group PAX (polycyclic aromatic hydrocarbons).

One problematic microarray, Dioxane, from the NGTX group is also visible, where the difference between the larger and smaller correlation is very small (0.03), so the decision to select NGTX is doubtful. This microarray is also difficult to classify correctly in the discriminant analyses examined in the following section.

Table 3. Pearson correlation coefficients between gene expression levels for all examined microarrays with averaged gene expression on all microarrays from the GTX and NGTX groups respectively.

	Chemical compound	Sample	Correlation with averaged microarrays from group:		Non-concor-dant decision	Group
			GTX	NGTX		
1	GTX MMS	learning	0.48	-0.14		
2	GTX MNU	learning	0.72	0.01		
3	GTX NNK	learning	0.67	-0.34		
4	GTX BaP-1	learning	0.69	0.07		
5	GTX DBA	learning	0.62	-0.18		
6	GTX FA	learning	0.15	0.50	*	GTX- PAH
7	GTX cisPT	learning	0.55	0.34		
8	GTX CP	learning	0.36	0.39		
9	GTX MMC	learning	0.68	0.22		
10	GTX DMN	testing	0.61	0.50		
11	GTX BaP-2	testing	0.33	0.61	*	GTX- PAH
12	GTX carboPt	testing	0.46	0.33		
13	NGTX DEHP-1	learning	-0.01	0.91		
14	NGTX Dioxane	learning	0.57	0.60		
15	NGTX PCP	learning	0.29	0.87		
16	NGTX PhB	learning	-0.08	0.88		
17	NGTX TCDD	learning	0.05	0.85		
18	NGTX TCE	learning	0.55	0.25	*	NGTX
19	NGTX TPA	learning	-0.05	0.89		
20	NGTX DEHP-2	testing	0.30	0.88		
21	NGTX reserpine	testing	0.47	0.43	*	NGTX
22	NGTX TCP	testing	-0.11	0.76		

3.3. The performance of the discrimination methods

Misclassification errors in the test dataset

Figures 1–2 and 3–4 show the misclassification errors estimated using the test dataset for different discrimination methods applied to a sequentially enlarged set of genes from *set1* (Fig. 1–2) and *set2* (Fig. 3–4).

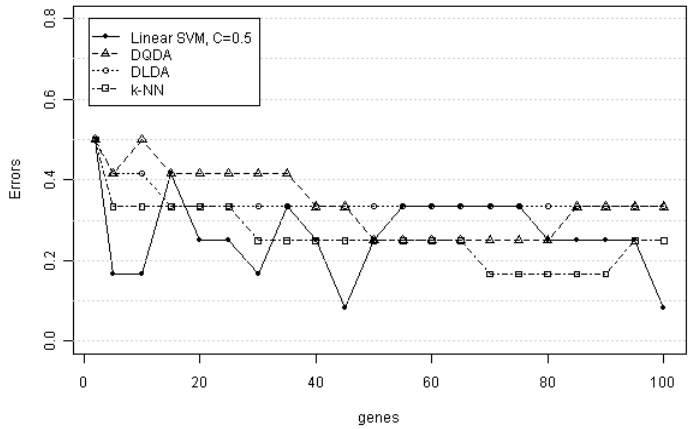

Figure. 1. Test classification errors of discriminant methods SVM, DLDA, DQDA, k-NN for ascending subsets of *set1* (from 2 to 100 genes).

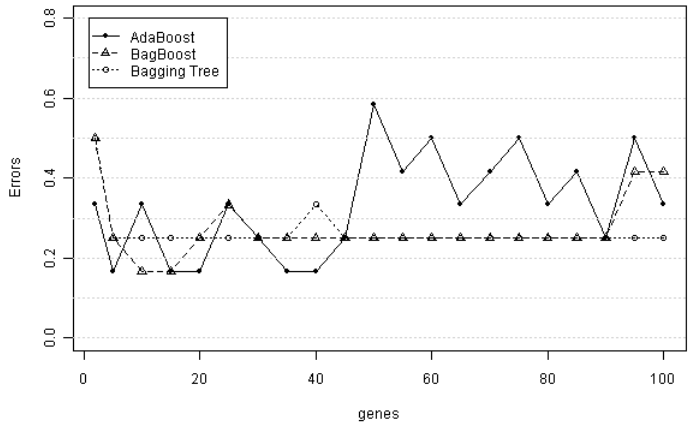

Figure 2. Test classification errors of discriminant methods: adaptive boosting trees (AdaBoost), bagboosting trees (BagBoost) and bagging trees, for ascending subsets of *set1* (from 2 to 100 genes).

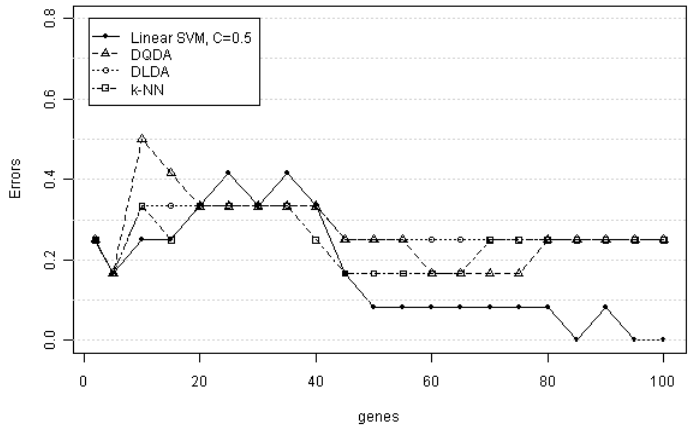

Figure 3. Test classification errors of discriminant methods: SVM, DLDA, DQDA, k-NN for ascending subsets of *set2* (from 2 to 100 genes).

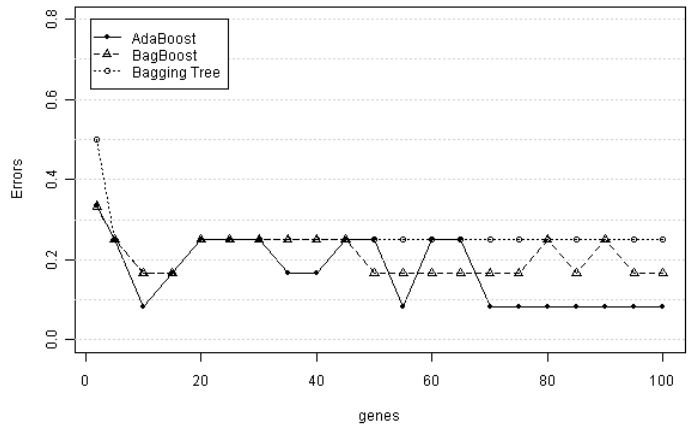

Figure 4. Test classification errors of discriminant methods: adaptive boosting trees (AdaBoost), bagboosting trees (BagBoost) and bagging trees, for ascending subsets of *set2* (from 2 to 100 genes)

The misclassification error rates estimated by an 8-fold cross-validation

Figures 5–6 and 7–8 show the misclassification error rates estimated by an 8-fold cross-validation (CV-8) for different discrimination methods applied to a sequentially enlarged set of genes from *set1* (Fig. 5–6) and *set2* (Fig. 7–8).

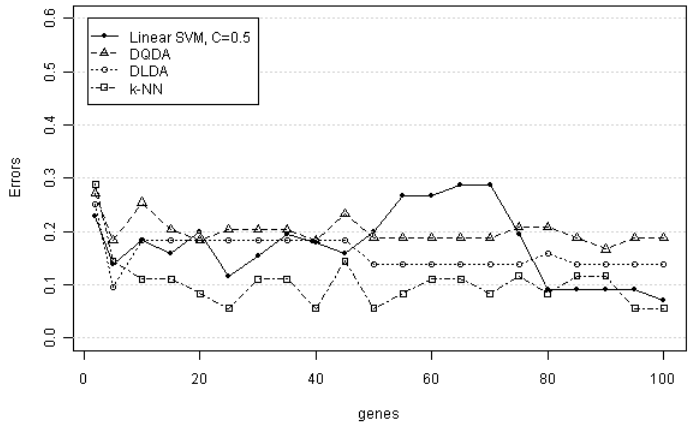

Figure 5. Classification cross-validation errors of discriminant methods SVM, DLDA, DQDA, k-NN for ascending subsets of *set1* (from 2 to 100 genes)

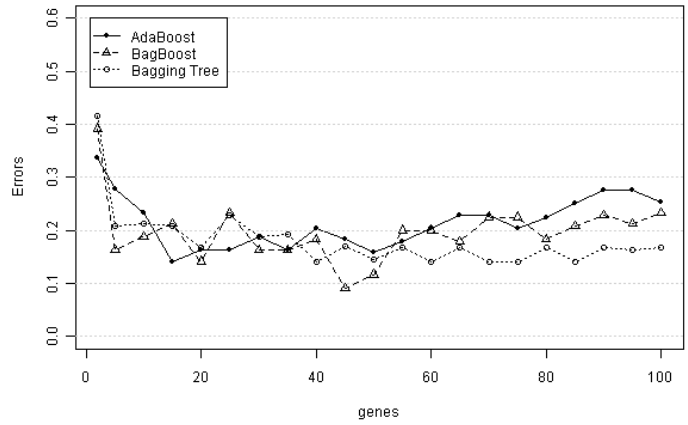

Figure 6. Cross-validation classification errors of discriminant methods: adaptive boosting trees (AdaBoost), bagboosting trees (BagBoost) and bagging trees, for successive subsets of *set1* (from 2 to 100 genes).

4. Discussion and conclusions

Figures 5–8 indicate that the misclassification error estimates obtained using the 8-fold cross-validation are, in general, smaller than the estimates obtained by applying the discrimination procedures to the test dataset (see Figures 1–4).

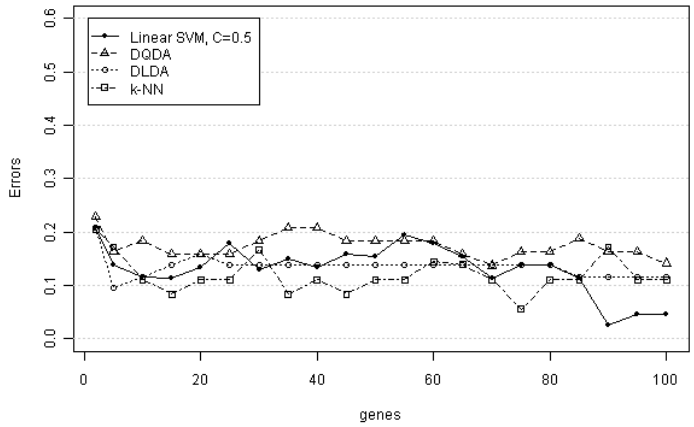

Figure 7. Cross-validation classification errors of discriminant methods: adaptive boosting trees (AdaBoost), bagboosting trees (BagBoost) and bagging trees, for successive subsets of *set2* (from 2 to 100 genes).

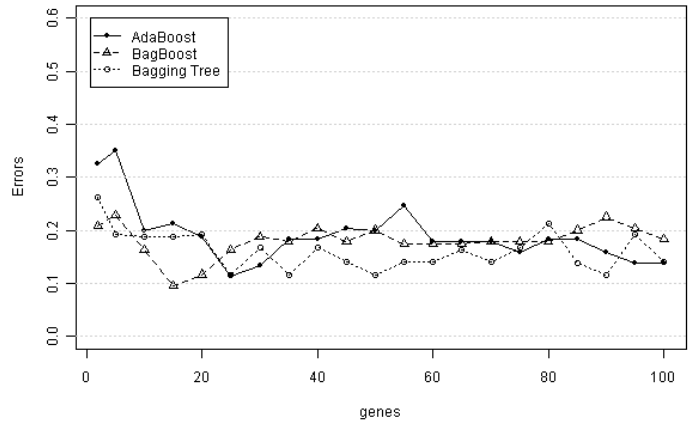

Figure 8. Cross-validation classification errors of discriminant methods: adaptive boosting trees (AdaBoost), bagboosting trees (BagBoost) and bagging trees, for succeeding subsets of *set2* (from 2 to 100 genes).

This may be due to the fact that the test dataset contains six microarrays that are "atypical" in the sense that they contain at most six differentially expressed genes, as reported by van Delft *et al.* (2004). After exclusion of these arrays, van Delft *et al.* (2004) noted an improvement in the estimates of the misclassification errors. In our analysis we retained all the 12 microarrays in

the test dataset. Note that, in the cross-validation, the "atypical" arrays are distributed in different folds. Thus their influence may be discounted.

The CV-based error estimates do not suggest any method that would consistently outperform all the others.

All the figures suggest that, in general, the use of only a few genes (2 or 5) results in a larger misclassification error. The reason may be that a limited set of genes may also have a limited discriminative power, irrespective of whether the genes themselves are important for discrimination. A similar observation was made, based on a simulation study, by Van Sanden *et al.* (2007, 2008).

In the analysis of the same dataset, van Delft *et al.* (2005) did not observe clear differences (apart from the poorer performance of Pearson analysis) between the results obtained for different discrimination methods, which included the Pearson correlation analysis, nearest shrunken centroids analysis, K-nearest neighbour analysis, and weighted voting. This is consistent with our finding, and may be due to the limited number of microarrays that are available for the analysis. However, the misclassification errors for the discrimination methods considered in the current paper, as estimated by cross-validation, are about two times smaller than those reported by van Delft *et al.* (2005) for the same dataset. This might be seen as a suggestion that the more advanced discrimination techniques might yield better results. However, it might also be due to the different pre-processing steps used by van Delft *et al.* (2005) to prepare the data for the analysis.

It is also worth noting that 17 of the genes indicated by van Delft *et al.* (2005) as important for discrimination purposes were included in the set of genes most often selected for building discrimination rules considered in our paper. The small misclassification errors may be also a confirmation of the correctness of gene selection methods. To conclude, our results suggest that, when discrimination between genotoxic and non-genotoxic carcinogens is of interest, the choice of the discrimination method may be important. However, further evaluation on more extensive data is warranted. In particular, the possibility of using ensembles of classifiers may be worth investigating.

Acknowledgments

The authors are indebted to Professor Joost Van Delft, from the Dept. of Toxicogenomics of Maastricht University, who kindly shared the microarray data used in this manuscript.

REFERENCES

Benjamini Y., 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.

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

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

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

Smyth G.K. (2004): Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3(1)

Smyth G.K., Speed T.P. (2003): Normalization of cDNA microarray data. Methods 31, 265–273.

Smyth G.K., Michaud J., Scott H. (2005): The use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics 21(9), 2067–2075.

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

van Delft J.H., van Agen E., van Breda S.G., Herwijnen M.H., Staal Y.C., Kleinjans J.C. (2005): Comparison of supervised clustering methods to discriminate genotoxic from non-genotoxic carcinogens by gene expression profiling. Mutat Res. 575(1-2):17-33.

Van Sanden S., Lin D., Burzykowski T. (2007): Performance of classification methods in a microarray setting: a simulation study. Biocybernetics and Biomedical Engineering 27(3), 15–28.

Van Sanden S., Lin D., 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.

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

Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., 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.