Metodološki zvezki, Vol. 9, No. 1, 2012, 25-45

Impact of Class-Imbalance on Multi-Class High-Dimensional Class Prediction

Rok Blagus and Lara Lusa[1]

Abstract

The goal of multi-class supervised classification is to develop a rule that accurately predicts the class membership of new samples when the number of classes is larger than two. In this paper we consider high-dimensional class-imbalanced data: the number of variables greatly exceeds the number of samples and the number of samples in each class is not equal. We focus on Friedman's one-versus-one approach for three-class problems and show how its class probabilities depend on the class probabilities from the binary classification sub-problems. We further explore its performance using diagonal linear discriminant analysis (DLDA) as a base classifier and compare its performance with multi-class DLDA, using simulated and real data.
Our results show that the class-imbalance has a significant effect on the classification results: the classification is biased towards the majority class as in the two-class problems and the problem is magnified when the number of variables is large. The amount of the bias depends also, jointly, on the magnitude of the differences between the classes and on the sample size: the bias diminishes when the difference between the classes is larger or the sample size is increased. Also variable selection plays an important role in the class-imbalance problem and the most effective strategy depends on the type of differences that exist between classes. DLDA seems to be among the least sensible classifiers to class-imbalance and its use is recommended also for multi-class problems. Whenever possible the experiments should be planned using balanced data in order to avoid the class-imbalance problem.

1 Introduction

The goal of many studies is to develop a rule (classifier) that can be used to reliably predict the class membership of new samples based on some of their observable characteristics (variables). The process used to derive a classifier is called class prediction (Bishop, 2007): the variables available on a set of samples for which the class membership is known (training set) are combined to derive the classifier, which is eventually used to predict the class of new samples (test set). Increasingly often the number of variables available for each sample is large, sometimes exceeding the number of samples included in the training set: if this is the case the data are called high-dimensional and special care

[1] Institute for Biostatistics and Medical Informatics, University of Ljubljana, Vrazov trg 2, Ljubljana, Slovenia, Rok.Blagus@mf.uni-lj.si, Lara.Lusa@mf.uni-lj.si

is needed in the development and validation of the performance of the classifier. For example, gene expression microarrays (Brown and Botstein, 1999) measure simultaneously the expression of tens of thousands of genes for each sample, while the number of samples included in a study rarely exceeds a few hundreds: many studies tried to predict the outcome of diseases using the gene-expression data derived from microarrays.

Another source of complexity that is increasingly acknowledged in the class prediction studies is the imbalance in the number of samples from each class included in the training set (class-imbalance problem, He and Garcia, 2009); the classifiers trained on class-imbalanced data tend to classify most of the new samples in the majority class. For class prediction problems involving two classes it was recently shown that the bias towards the majority class is further increased when data are high-dimensional (Blagus and Lusa, 2010). The high-dimensionality affects each classifier in a slightly different way but in most cases the additional bias arises because the sampling variability is larger in the minority class, and extreme values are likely to arise when thousands of variables are measured. Diagonal linear discriminant analysis (DLDA) was the least sensitive classifier to class-imbalance among those considered by Blagus and Lusa (2010). Its performance was further improved if used in combination with undersampling techniques, which use smaller but balanced training sets; the variability of the results was reduced by multiple undersampling, a method that combines many small balanced training sets sampled from the original training set.

Often the number of classes to predict is larger than two (multi-class classification, Tsoumakas and Katakis, 2007). Some two-class classifiers cannot be straightforwardly extended to deal with more than two classes; for example, the standard SVM classifiers are formulated for only two classes. For this reason many approaches were proposed to deal with multi-class class-prediction as a series of binary class-prediction sub-problems (Allwein et al., 2001; Dietterich and Bakiri, 1995; Friedman, 1996; Hastie and Tibshirani, 1998): the two main approaches are the one-versus-rest strategy (the binary sub-problems consist in comparing one class with all the others) and one-versus-one strategy (all pairs of classes are compared). The one-versus-one strategies have larger variability but are less affected by class imbalance compared to one-versus-rest strategies, as they use smaller but less imbalanced training sets. Moreover, the one-versus-rest strategies usually do not work because they produce regions that do not belong to any of the classes (Izenman, 2008).

The multi-class problems are common also in the high-dimensional setting where the number of classes is generally limited. For example, Hedenfalk et al. (2001) tried to distinguish three types of primary breast cancers (sporadic, with BRCA1 or with BRCA2 mutation) based on their gene-expression. Sotiriou et al. (2006) combined the data from previously published studies and analyzed the gene-expression data from breast cancer patients with histologic tumor grade 1 ($n = 167$), 2 ($n = 218$) or 3 ($n = 256$).

Some previous studies addressed the problem of multi-class classification using high-dimensional gene expression data (Berrar et al., 2003; Romualdi et al., 2003; Statnikov et al., 2005). For example, using nine multi-class gene expression data sets, Statnikov et al. (2005) compared the performance of nine multi-class classifiers (six multi-class support vector machine (MC-SVM) algorithms, K-nearest neighbors (k-NN), backpropagation neural networks and probabilistic neural networks) and showed that the MC-SVM performed better than the other three classifiers. The majority class of their most imbal-

anced multi-class data set included 68.5% of the samples; however, their work did not focus explicitly on the class-imbalance problem.

The aim of our study was to investigate how class imbalance affects the multi-class classification for high-dimensional class-imbalanced data, a problem that to our knowledge has not been systematically addressed so far. We focused mainly on DLDA because of its good behavior in the two-class problems with high-dimensional class-imbalanced data; another reason for choosing DLDA was the straightforward generalization of the two-class DLDA to the multi-class situation (multi-class DLDA, mDLDA). We compared mDLDA and the Friedman's one-versus-one approach (1996), which breaks down the multi-class problem in a series of two-class classification problems and assigns new samples to the class having most votes. Friedman's approach was chosen because of its wide applicability and simplicity, and because it was previously indicated as beneficial when the classes are imbalanced or when the number of classes is large (Speed, 2003). We chose a one-versus-one rather than a one-versus-all strategy because we expected that it would be less affected by class-imbalance.

The two approaches were evaluated both on simulated data and on four publicly available data sets from breast cancer gene expression microarray studies (Huang et al., 2003; Ivshina et al., 2006; Miller et al., 2005; Pittman et al., 2004; Sotiriou et al., 2003); we assessed both the overall and the class specific predictive accuracies. We simulated data where there was no difference between the classes (null case) and three scenarios where the classes were different (alternative case). We evaluated if random undersampling could reduce the bias arising from the class-imbalance also in the multi-class classification.

The paper is organized as follows. In the Methods section we briefly describe the classification approaches and the strategies to deal with the class-imbalance problem that we used; we also describe the simulations that were performed and the gene expression microarray data sets used in our analyses. In the Results section we outline the expected properties of the multi-class classifier in the null case and show the actual performance of the methods on simulated and real data. In the Discussion we summarize the results and outline the problems related to multi-class classification for high-dimensional data.

2 Methods

In Section 2.1 we describe the notation used in the paper, the classifiers that were used, focusing on mDLDA, on the Friedman's approach to multi-class prediction using two-class DLDA and on the random undersampling method that was used to reduce the class imbalance problem; we also describe how we selected the variables that were used in the classification rule and the measures used to evaluate the performance of the classifiers. In Section 2.2 we describe the simulation settings for the null and the alternative case. In Section 2.3 we briefly present the microarray data sets that we used and the multi-class classification problems that we addressed.

2.1 Class prediction methods

2.1.1 Notation

Through the paper we indicate the number of samples with n, the number of variables with p and the number of variables selected and used in the classification rule with G, these variables are the most informative about class distinction; K is the number of classes while the class membership of the samples is indicated with integers from 1 to K; the classes are non-overlapping and each sample belongs to exactly one class, the number of samples in Class k is denoted by n_k.

Let x_{ij} be the expression of jth variable ($j = 1, ..., p$) on ith sample ($i = 1, ..., n$). For sample i we denote the set of G selected variables by $\mathbf{x}_i$. Let $\bar{x}_g^{(k)}$ denote the mean expression of the gth selected variable in Class k. The mean expression of the gth variable in Class k is defined as

$$\bar{x}_g^{(k)} = \frac{1}{n_k} \sum_{i \in C_k} x_{ig}, \tag{2.1}$$

and let $\mathbf{x}^*$ represent the set of selected variables for a new sample.

2.1.2 Multi-class DLDA

Discriminant analysis methods are used to find linear combination of variables that maximize the between-class variance and at the same time minimize the within-class variance (Simon et al., 2004; Speed, 2003). Diagonal linear discriminant analysis (DLDA) is a special case of discriminant analysis that assumes that the variables are independent and have the same variance in all classes.

The multi-class DLDA (mDLDA) classification rule for a new sample $\mathbf{x}^*$ is linear and is defined as

$$\mathcal{C}(\mathbf{x}) = \text{argmin}_k \sum_{g=1}^{G} \frac{\left(x_g^* - \bar{x}_g^{(k)}\right)^2}{s_g^2}, \tag{2.2}$$

where s_g^2 is the sample estimate of the pooled variance for variable g and x_g^* is the gth selected variable of the new sample. The two-class DLDA is a special case of mDLDA.

2.1.3 Friedman's approach

In Friedman's approach, also known as the win-max rule, the class-prediction problem for $K > 2$ classes is divided in $\binom{K}{2}$ binary class-prediction problem, one for all pairs of classes. Within each binary class-prediction problem we build a rule for class-prediction (train a classifier) and a new sample is classified in one of the two classes. The final class-prediction in one of the K classes is defined with majority voting, assigning the new sample to the class with most votes.

The motivation for the Friedman's approach lies in the fact that a multi-class Bayes decision rule can be obtained by separately constructing an optimal rule for discriminating between every pair of classes, ignoring the samples from the other classes (Friedman, 1996). Friedman's approach considers only the class-membership when deciding for the

final classification of the new samples. The probabilities of belonging to a certain class rather than the class membership derived from the binary sub-problems could also be considered (Hastie and Tibshirani, 1998).

It can be shown that Friedman's approach and multi-class linear discriminant analysis yield to exactly the same classification results if the within-class variances of the variables are the same in all the classes; however, when the assumption of equal within-class variances is violated the results obtained using the two methods can be slightly different, as the estimate pooled variances depend on the pairs of classes being compared and can differ substantially from the estimate obtained using all the classes, as in mDLDA. Also the variable selection method used to reduce the number of variables can be different from the method used for mDLDA. See section 2.1.5 for details.

2.1.4 Simple undersampling

Simple undersampling (down-sizing) consists of obtaining a class-balanced training set by removing a subset of randomly selected samples from the larger class (Batista et al., 2004; He and Garcia, 2009). In mDLDA undersampling consisted in using $min(n_1, n_2, n_3)$ samples from each class, randomly selecting which samples from the majority class(es) should be removed. With Friedman's approach each pairwise comparison was undersampled if the size of the classes was not equal ($n_k \neq n_j$). The classification rule was derived on the balanced training set as described for the original data, and evaluated on the test set.

In our reanalysis of the microarray data we repeated the process of random selection of samples to be included in the class-balanced training set 100 times in order to reduce the variability of the estimated predictive accuracy. In this case we report also the standard deviation of the predictive accuracy.

2.1.5 Variable selection

The $G < p$ variables that were most informative about class distinction were selected on the training set and used to define the classification rules (Eq. 2.2). Variable selection was based on two sample t-test with assumed equal variances for the Friedman's approach, or F-test for the equality of more than two means for mDLDA. A limited set of simulations and real data analysis were carried out using the F-test with Friedman's approach. We selected $G = 40$ variables in most cases; $G = 30$ variables were selected in the third alternative scenario, see below. Note that when using Friedman's approach a different set of variables could be used for the training of each binary classifier.

In the null case we also considered the situation where all the variables were used ($G = p$). We carried out also a limited set of simulations where the variables used with the Friedman's approach were selected with the F-test using all the classes.

To reduce the computational burden in the reanalysis of the breast cancer gene-expression data sets we considered only the $p = 1000$ variables with the largest variances. The variable selection consisted in selecting on each training set the $G = 40$ variables with the smallest p-values, as described for the simulated data.

2.1.6 Evaluation of the performance of the classifiers

The performance of the classifiers was evaluated on the independent test sets. It is well known that for imbalanced data the proportion of correctly classified samples can be a misleading measure of the performance of a classifier (Pepe, 2003, chapter 2). For this reason four different measures of performance were considered: (i) overall predictive accuracy (PA, the number of correctly classified subjects from the test set divided by the total number of subjects in the test set), (ii) predictive accuracy of Class 1 (PA_1, i.e., PA evaluated using only samples from Class 1), (iii) predictive accuracy of Class 2 (PA_2 i.e., PA evaluated using only samples from Class 2) and (iii) predictive accuracy of Class 3 (PA_3). Their standard deviations were also reported. When we reanalyzed the real gene expression data the predictive accuracies (overall and class-specific) were estimated using leave-one-out cross-validation (LOOCV).

2.2 Data simulation

We simulated $p = 40$ or 1000 independent variables for each of the $n_{train} = 90$ samples in the training set and $n_{test} = 300$ samples of the test set, considering three classes.

Class 1 was always assumed to be the majority class, while classes 2 and 3 were smaller and balanced ($n_1 \geq n_2 = n_3$); four different levels of class imbalance were considered (n_1 (%) $= 30$ $(33.3\%), 36$ $(40\%), 40$ (44.4%) or 60 (66.7%)) in the training sets, while the test set was always balanced. All the simulations were repeated 500 times.

Null case. Under the null case all the variables were simulated independently from the standard normal distribution (mean $\mu = 0$ and standard deviation $\sigma = 1$, $N(0, 1)$) and the class membership of the samples was randomly assigned.

Alternative case. Under the alternative case the class membership was dependent on some of the variables; for each sample p_0 variables were generated independently from $N(0, 1)$ (null variables) for all classes, while the remaining variables (p_{DE}, non-null or alternative variables) were generated independently from a normal distribution with different means between the classes (but with equal variances). Three different scenarios were considered in the alternative case; $p_{DE} = 40$ non-null variables were used in scenario 1 and 2, and $p_{DE} = 30$ in scenario 3. In the first scenario there were no real differences between classes 2 and 3, while class 1 was different. In the second scenario the three classes were all different, and class 1 was *between* classes 2 and 3: the differentially expressed variables from class 1 were simulated using mean values that were between those from class 2 and 3. In the third scenario all three classes were different but none of the classes was placed *between* the others: the non-null variables were simulated from three different normal distributions for each class. The variance of the variables was 1 for all variables. The normal distributions from which the non-null variables were simulated in the three scenarios are reported in Table 1. A graphical illustration of the three scenarios is also given (figure 1, where larger sample sizes and differences between classes were used in order to obtain a clearer graphical display).

Table 1: Distributions from which the p_{DE} differentially expressed variables were simulated independently for the three scenarios under the alternative case.

	p_{DE}	Class 1	Class 2	Class 3
Scenario 1	40	N(0,1)	N(0.7,1)	N(0.7,1)
Scenario 2	40	N(0,1)	N(0.7,1)	N(-0.7,1)
Scenario 3	10	N(-0.5,1)	N(0,1)	N(0.5,1)
	10	N(0.5,1)	N(-0.5,1)	N(0,1)
	10	N(0,1)	N(0.5,1)	N(-0.5,1)

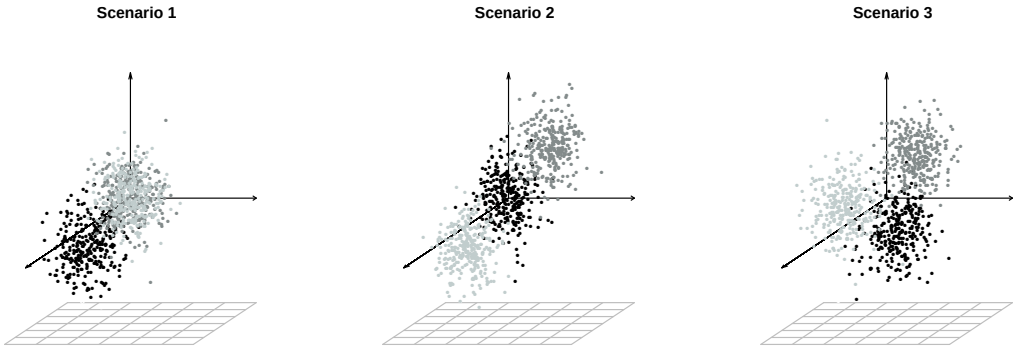

Figure 1: Illustration of different scenarios under the alternative hypothesis.

2.3 Microarray gene expression data sets

In this section we briefly describe the five breast cancer gene expression microarray data sets that we re-analyzed; data were preprocessed as described in the original publications. The number of variables measured in the original studies, the classification task(s) performed by us and the number of samples in each class are reported in table 2.

The performance of the mDLDA and Friedman's approach was evaluated using leave-one-out cross-validation (LOOCV). When undersampling was performed to balance the training set, we used 100 balanced training sets. In the results we report the average predictive accuracies (overall and class specific) and their standard deviation.

To speed up the calculations only the 1000 genes with the largest variability were considered.

Sotiriou. Sotiriou et al. (2003) analyzed cDNA gene expression profiles from 99 tumor specimens from breast cancer patients, measuring the gene expression of 7650 genes (probes). The classification problem that we considered was the prediction of the histological grade of tumors; the data set included 16 grade 1 (G1), 38 grade 2 (G2) and 45 grade 3 (G3) patients.

Table 2: Main characteristics of the five re-analyzed breast cancer data sets.

Data set	Prediction task	Variables	Samples	n_1	n_2	n_3
Sotiriou	Grade 1, 2 or 3	7650	99	16	38	45
Miller	Grade 1, 2 or 3	22283	249	67	128	54
Ivshina	Grade 1, 2 or 3	22283	289	68	166	55
Huang	ER-, ER+/++ or ER+++	12625	80	12	42	26
Huang	ER-, ER+ or ER++/+++	12625	80	12	24	44
Pittman	ER-, ER+/++ or ER+++	12625	158	48	74	36
Pittman	ER-, ER+ or ER++/+++	12625	158	48	40	70

Miller. Miller et al. (2005) derived a classifier to distinguish p53-mutant and wild-type tumors using a series of 251 breast cancer patients, for which they measured the expression of more than 20,000 genes. Also on this data set we focused on the prediction of the grade of the tumors, using 67 G1, 128 G2 and 54 G3 patients.

Ivshina. Ivshina et al. (2006) developed a classifier of histologic grade using 347 primary breast cancer samples, training the classifiers using G1 and G3 samples only; they measured the expression of more than 20,000 genes. We focused on the three class prediction of histologic grade, using 68 G1, 166 G2 and 55 G3 samples. This data set included all the samples analyzed by Miller et al. (2005) and 40 additional G2 samples.

Huang. Huang et al. (2003) used the expression of about 12,000 genes data as predictors of breast cancer outcomes. For our classification purposes we used the estrogen receptor status (ER status), which was reported in the original data set using 4 categories: ER- negative (ER-, $n = 12$), ER+ (slight intensity, $n = 24$), ER++ (moderate intensity, $n = 18$) and ER+++ (strong intensity, $n = 26$). We considered two three-class classification problems, merging the ER+ and ER++ groups for the first task, and the ER++ and ER+++ groups for the second task.

Pittman. Pittman et al. (2004) expanded the cohort of Huang et al., including in their study 78 additional patients. We performed the same two three-class prediction analyses described for the data set of Huang, using 48 ER-, 40 ER+, 34 ER++ and 36 ER+++ samples.

2.4 Analysis

Statistical analysis and simulations were carried out using R language for statistical computing (R version 2.8.1) (R Development Core Team, 2008).

3 Results

In this section we first derive the class probabilities using Friedman's approach in a three-class problem. We then present the simulation results and the results obtained re-analyzing the breast cancer microarray data sets.

3.1 Class probabilities using Friedman's approach

We derive the class probabilities using Friedman's approach as a function of the class probabilities from the binary class-prediction sub-problems. For the sake of simplicity let us consider a classification problem with $K = 3$ classes. The possible outcomes of the three possible pairwise comparisons among the three classes are listed in table 3, where the class assignments derived using Friedman's approach are also given together with the notation used to denote the probabilities of each outcome (the subscripts indicate the winning class from each binary comparison). The class is chosen at random when each of the three classes receives a vote. In practice, in a three class problem most classifiers have 0 or close to 0 probabilities of having ties (one vote for each class, verified with the simulation - data not shown).

Table 3: Outcomes of the binary class-prediction sub-problems, class assignments and probabilities using Friedman's approach.

Vote			Probability	Class assignment
$1\ vs\ 2$	$1\ vs\ 3$	$2\ vs\ 3$		
1	1	2	p_{112}	Class 1
1	1	3	p_{113}	Class 1
1	3	2	p_{132}	Class 1, 2 or 3
1	3	3	p_{133}	Class 3
2	1	2	p_{212}	Class 2
2	1	3	p_{213}	Class 1, 2 or 3
2	3	2	p_{232}	Class 2
2	3	3	p_{233}	Class 3

The probabilities of assigning a new sample to each of the classes are (see table 3 for the definition of p_{ijk})

$$P(\mathcal{C} = 1) = p_{112} + p_{113} + \frac{1}{3}\left(p_{213} + p_{132}\right),$$

$$P(\mathcal{C} = 2) = p_{221} + p_{223} + \frac{1}{3}\left(p_{213} + p_{132}\right),$$

$$P(\mathcal{C} = 3) = p_{133} + p_{233} + \frac{1}{3}\left(p_{213} + p_{132}\right).$$

Let us assume that in each pairwise comparison the new samples are equally likely to be assigned to both classes. In this case all the outcomes listed in the table are equally

likely and it is therefore straightforward to show that the Friedman's approach would assign the new samples to each class with equal probability: $P(\mathcal{C} = 1) = P(\mathcal{C} = 2) = P(\mathcal{C} = 3) = 1/3$. This behavior would be expected when there are no true differences between classes in the training set (null case).

However, it is well known that when the classes are imbalanced the binary classification is biased towards the majority class for most classification methods, even in the null case (He and Garcia, 2009); the bias deriving from the binary classification steps is therefore carried over to the Freedman's approach.

To evaluate the extent of the bias in a practical situation let us consider a three class classification problem where there is no real difference between the classes and $p = 40$ variables are measured and used for classification; most samples belong to Class 1, while Class 2 and Class 3 are equally sized and are four times smaller (80% class-imbalance, $n_1 = 60, n_2 = n_3 = 15$, data are simulated from N(0,1) for all variables). We previously derived the probability of classifying a new sample in Class 1 in a binary problem for several classifiers in this settings (Blagus and Lusa, 2010); the probabilities for 3-NN, DLDA, random forests (RF) and SVM are reported in table 4 (binary sub-problems); using the formulae derived above we evaluated the class probabilities obtained using Friedman's approach (table 4, Friedman). The results show that the bias towards the majority class remains in the multi-class classification as well, and that the class probabilities of the minority classes become smaller if compared to those from the binary sub-problems. DLDA is the least sensitive to the class-imbalance problem. The consistency of these results was confirmed also using simulated data. Very similar results were obtained simulating correlated variables (data not shown).

Table 4: Class probabilities using binary sub-problems and Friedman's approach ($p = 40, n_1 = 60, n_2 = 15, n_3 = 15$, all variables simulated independently from N(0,1)).

Classifier	Binary sub-problems			Friedman		
	(1 vs 2) $P(\mathcal{C} = 1)$	(1 vs 3) $P(\mathcal{C} = 1)$	(2 vs 3) $P(\mathcal{C} = 2)$	$P(\mathcal{C} = 1)$	$P(\mathcal{C} = 2)$	$P(\mathcal{C} = 3)$
3-NN	0.89	0.89	0.50	0.82	0.09	0.09
DLDA	0.70	0.70	0.50	0.56	0.22	0.22
RF	1	1	0.50	1	0	0
SVM	0.96	0.96	0.50	0.93	0.03	0.03

3.2 Simulation results

In this section we present the results of a selected group of simulations where we investigated the performance of mDLDA and DLDA with the Friedman's approach, both assuming no real difference between classes (null case) or simulating differences between classes (alternative case). The focus was on the effect of dimensionality of the data and the class-imbalance; the sample size was kept fixed while the proportion of samples in the majority class varied from 1/3 (balanced situation) to 2/3 (most imbalanced situation). When the number of variables was large some additional attention was devoted to the

effect of variable selection. We present the results obtained using the original versions of the classifiers and using undersampling. The results are discussed focusing on the class-specific predictive accuracies. See the Methods section for the details on the simulation settings.

3.2.1 Null case

In the first set of simulations there were no differences between classes, as all the variables were simulated independently from the same distribution. Results are represented graphically in figure 2.

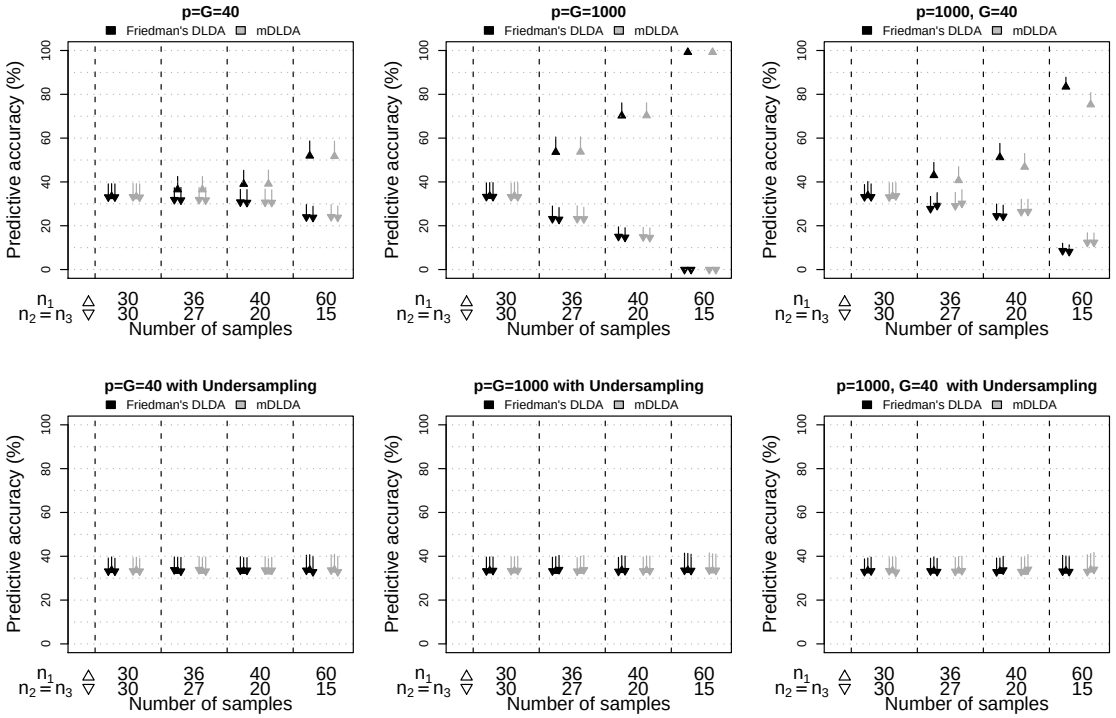

Figure 2: Null case results: class specific predictive accuracies (in %) and their standard deviations (vertical lines), using DLDA with Friedman's approach and mDLDA, varying the number of variables (simulated (p) and used (G)) and the size of the majority class (n_1); the second row shows the results obtained with random undersampling.

When the number of simulated variables was small ($p = 40$) and the training set was balanced ($n_1 = n_2 = n_3 = 30$) both DLDA with Friedman's approach and mDLDA randomly assigned approximately one third of the new samples to each class ($PA_k \approx 1/3$). When the number of samples from the majority class (Class 1) increased, both methods assigned more new samples to the majority class; as expected, the number of samples assigned to the other two balanced minority classes was approximately equal in all the situations and there was hardly any difference between Friedman's DLDA and mDLDA. For example, in our most imbalanced setting ($n_1 = 60, n_2 = n_3 = 15$), the PA for both methods was about 52% for the majority class and about 24% for the other two classes.

The bias towards the majority class increased with the number of simulated variables: the PA of the minority classes became poorer as more samples were classified in the majority class and the bias increased with the class imbalance. Almost all new samples were assigned to the majority class when we simulated and used $p = G = 1000$ variables in our most imbalanced setting. mDLDA and Friedman's DLDA performed very similarly also when the number of variables was large, while the two methods differed when only the variables that differentiated the most between the classes were used ($p = 1000, G = 40$). Variable selection reduced the bias towards the majority class, more substantially for mDLDA: the PA_1 decreased from 99% to 75% for mDLDA and to 83% for Friedman's DLDA in the most imbalanced setting. The difference between the two methods can be ascribed to the different variable selection methods. mDLDA selects the variables that differentiate the most the three classes (with the F test), while in our implementation of Friedman's DLDA we selected the variables that differentiated the most the pairs of classes (with t-tests); therefore, different variables could be selected for each of the binary sub-problem. We ran an additional set of simulations using Friedman's DLDA, where the variables selected with F test were used in all the binary problems; the results in this case were almost identical to those obtained with mDLDA (data not shown).

Undersampling. We used simple undersampling with the aim to remove or reduce the bias towards the classification into the majority class in the null case. All the classifiers were trained on class-balanced data (see Methods for details).

In this simulation settings simple undersampling completely removed the bias caused by the class-imbalance: the new samples were randomly assigned to the three classes, regardless of the class-imbalance of the original training set. Selecting only a subset of the samples from the majority class removed the internal bias of DLDA as it ensured that the sampling variability in each class was the same. The expected drawback was the larger variability of the PA in the situations where the original class imbalance was large; the reason is the smaller size of the training set. For example, in our most imbalanced setting we used only half of the training samples, removing 45 samples from the majority class. Note however that variable selection needs to be implemented after undersampling as otherwise the results remain biased in favor of the majority class (data not shown).

3.2.2 Alternative case

We further explored the effect of class imbalance and high-dimensionality of the data in situations where some of the variables were different between the classes. Specifically, we explored three different scenarios: there was no difference between the two minority classes, while the majority class was different (scenario 1), all the classes were different, with the majority class nested between the two minority classes (scenario 2) or with no nesting between the classes (scenario 3). Results are represented graphically in figure 3.

The three scenarios produced very different results. The way in which data were simulated in scenario 3 guaranteed that the class specific PA were equal in the balanced setting, as none of the classes was more difficult to predict compared to the others. Class 1 was the easiest to predict in scenario 1 (because there was no difference between classes 2 and 3) while it was the most difficult to predict in scenario 2 (because it was nested between the two other classes, while the two minority classes were well separated).

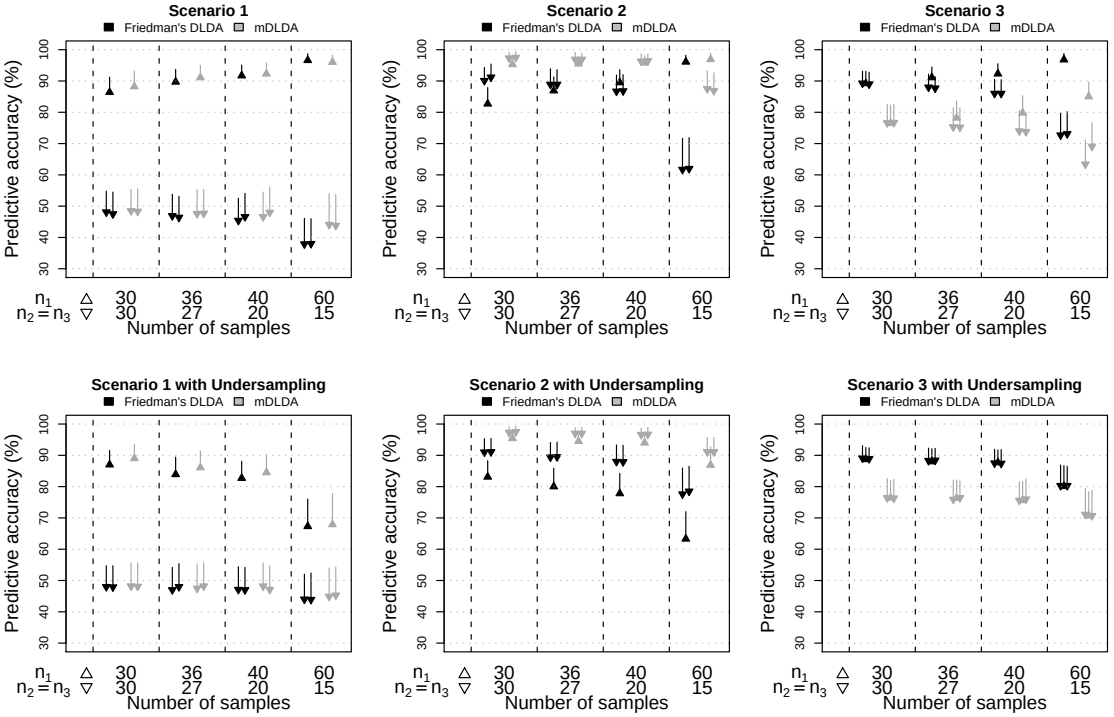

Figure 3: Alternative case results: class specific predictive accuracies (in %) and their standard deviations (vertical lines), using DLDA with Friedman's approach and mDLDA, varying the number of variables (simulated (p) and used (G)) and the size of the majority class (n_1); the second row shows the results obtained with random undersampling.

In most situations the majority class had the largest class specific PA, which markedly increased with class imbalance. For example, when we used Friedman's DLDA in scenario 3 the majority class PA increased from 89% (balanced setting) to 97% (most imbalanced setting), while the PA of the minority classes decreased from 89% to 73%. In scenario 3 mDLDA performed worse than Friedman's DLDA, obtaining class specific PA that were on average about 10% points smaller than those from Friedman's DLDA.

In scenario 2 the class 1 was the most difficult to predict in the balanced setting, nevertheless it had the best class specific PA when the class imbalance was large. In this scenario mDLDA performed better than Friedman's DLDA, the differences being more marked in the most imbalanced setting: the PA of the minority classes was about 25% points better with mDLDA (87% vs 62%), while the PA of the majority class was about the same (96%) using both methods.

The performance of mDLDA and Friedman's DLDA was very similar in scenario 1. The PA of the minority classes were smaller compared to the PA of the majority class and were below 50%. This was due to the fact that in this scenario there were no real differences between the minority classes. We carried out a limited set of simulations for the class-balanced situation where we increased the difference between the majority and minority classes in this setting and we observed that as the PA of the majority class approached 100%, the PA of the other two classes approached 50% (data not shown). Also in this setting the class imbalance increased the discrepancy between the PA of the

majority and minority classes. For example, in the most imbalanced situation ($n_1 = 60$, $n_2 = n_3 = 15$) Friedman's DLDA achieved 97% PA for the majority class and 38% PA for the minority classes, while the PA were respectively 86% and 48% in the balanced situation.

Also in the alternative case the discrepancies between Friedman's DLDA and mDLDA could be attributed to the different variable selection methods that were used: the same results were obtained if both methods used the F-test to select the variables, similarly as described for the null case (data not shown). The pairwise t-tests outperformed the F-test in scenario 3, while they performed worse than F-test in scenario 2, and comparably to F-test in scenario 1. The reason for these differences can be ascribed to the statistical power of the two variable selection methods (evaluated using simulated data, data now shown). In scenario 1 the power was approximately the same for both methods, while F-test had a larger overall power in scenario 2 and 3. The best performance of the pairwise t-tests in scenario 3 can be explained focusing the subset of variables that had the largest differences between a pair of classes. For example, the first set of variables reported in table 1 for scenario 3 are those that have the biggest difference between class 1 and 3. The pairwise t-test had better power than F-test in selecting this type of variables and for this reason the pairwise classifiers included more of the variables that had (real) large differences between the pairs of classes, achieving better PA.

Undersampling. We used simple undersampling for the three alternative case scenarios. We showed in the previous section that the alternative case results varied greatly, depending on the simulated scenario and on the classification method that we used; undersampling results depended on the scenario and on the classification method as well. A general remark is that our simulations showed that by undersampling we could achieve results that were very similar to those obtainable using balanced data, in any given setting. Comparing the results obtained in the balanced setting with those of the most imbalanced (undersampled) setting in figure 3 we observe that similar patterns emerge. For example, class 1 had the best PA for in scenario 1 and the worst PA in scenario 2, while the class-specific PA were equal in scenario 3. The results obtained using undersampled training set had worse PA for the majority class and larger variability in the PA, because a smaller number of samples was used, mostly at the expense of the majority class.

Nevertheless, undersampling had the effect of balancing to some extent the class specific PA, improving the minority class PA and decreasing the majority class PA. This can be observed comparing the results that were obtained using undersampling in the most imbalanced setting with those obtained in the same setting without any correction. For example, using Friedman's DLDA in scenario 3 all three classes had 80% PA using undersampling, while the discrepancy in PA between majority and minority class without undersampling was 25% points. In scenario 2, the majority class PA decreased from 96% to 63%, while the minority class PA increased from 62% to 78%; in this scenario it is expected that the PA of the majority class is smaller than the PA of the other classes. In scenario 1 undersampling removed most of the bias towards the majority class but did not improve considerably the PA of the minority classes; this limited improvement could be expected because in scenario 1 there is no real difference between the minority classes, for which we expect at most a 50% PA.

4 Re-analysis of public microarray data sets

Five different breast cancer gene expression data sets were reanalyzed, with the aim of developing classifiers for the prediction of ER status and grade of the tumors. All the classification problems involved three classes; the sample size varied from 80 to almost 300 samples, the percentage of samples from the minority class varied from 15% to 25%, while the majority class included about a half of the samples (range: 44% to 57%, table 2). The Miller data set is a subset of the Ivshina data set, where additional 40 G2 samples where included in the analysis and the class label of two samples was changed. In the same way, the Huang data set is a subset of the Pittman data set, which included twice as many samples as the Huang data set, and had a smaller class-imbalance (30% ER- instead of 15%). See the Methods section for details on the variable selection and undersampling.

The prediction of the ER status in two classes is known to be an easy classification task, as thousands of genes distinguish the ER negative from ER positive samples. It is less clear if it is possible to develop a classifier that accurately distinguishes the intensity of ER expression; for instance, separating patients with slight or moderate ER intensity from those with strong intensity of ER. The prediction of the grade of the tumor is considered a hard classification task, as few genes distinguish G2 patients from G1 or G3 patients, while the differences between G1 and G3 are more marked: the G2 group comprises about a half of the patients and it is considered a heterogeneous group.

In all data sets the predictive accuracy for Grade 2 class was poor, even though it was the majority class in the Miller and Ivshina data sets (table 5). Friedman's DLDA and mDLDA performed very similarly in the grade prediction; Friedman's DLDA performed slightly better on the minority classes in the Miller data set and overall in the Ivshina data set. Undersampling did not worsen the performance of the classifiers, slightly improving the PA of the minority classes in some situations, and leaving the G2 PA more or less unaffected (see for example Miller data set using mDLDA). Comparison between the results on the Ivshina data set and the Miller data set showed that increasing the number of samples in the majority class (G2) somehow increased the overall performance of the classifiers, obtaining an overall better performance with Friedman's DLDA, while improving only the minority classes PA with mDLDA.

A better overall PA was obtained for the three-class ER status prediction. The Huang data set was small and had the largest class-imbalance; the classification results were poor in the first classification task (joining ER+ and ER++ samples), and slightly better in the second classification task (joining ER++ and ER+++ samples). Friedman's DLDA performed better than mDLDA; moreover, undersampling substantially improved the PA of the minority classes while the decrease of the PA of the majority class was small for the first classification task, and larger for the second one. In the Pittman data set the sample size was larger and the class-imbalance smaller. The PA of the ER- (minority) class substantially increased for both classification tasks; undersampling did not improve the performance as substantially as for the Huang data set, and seemed more helpful in improving the performance for mDLDA.

The main differences between the results obtained using Friedman's DLDA and mDLDA could be ascribed to the two different methods used to perform the variable selection. However, unlike the results obtained with the simulated data, using F-test with Friedman's DLDA did not yield exactly the same results as mDLDA (data not shown). This

difference can be explained by the fact that the within-class variances can be different in the real data, therefore the estimates of the pooled variances can differ considerably when estimated using all three classes or only the pairs of classes. Also, Friedman's approach coupled with undersampling uses more samples in the training of the classifier, compared to the undersampled mDLDA, which reduces all the classes to the size of the smallest class.

Table 5: Re-analysis of breast cancer gene expression data: overall and class specific predictive accuracies (**PA**, in % (SD)) for different approaches (Friedman's DLDA, multi-class DLDA, with our without undersampling (US) the training set) on five different data sets. The PA are estimated using leave-one-out cross validation.

Grade	Sotiriou				Miller				Ivshina			
	PA	$\mathbf{PA}_1$	$\mathbf{PA}_2$	$\mathbf{PA}_3$	**PA**	$\mathbf{PA}_1$	$\mathbf{PA}_2$	$\mathbf{PA}_3$	**PA**	$\mathbf{PA}_1$	$\mathbf{PA}_2$	$\mathbf{PA}_3$
(n_k)	(99)	(16)	(38)	(45)	(249)	(67)	(128)	(54)	(286)	(68)	(166)	(55)
Friedman's DLDA	46.5	43.8	26.3	64.4	55.8	70.1	40.6	74.1	59.5	76.5	45.8	80.0
mDLDA	49.5	62.5	23.7	66.7	56.2	68.7	44.5	68.5	55.4	76.5	38.6	80.0
Friedman's DLDA with US	44.0	60.8	22.1	56.7	56.6	71.7	39.4	78.9	59.0	78.0	43.6	81.8
(SD)	(3.1)	(9.1)	(5.2)	(4.0)	(1.4)	(3.0)	(2.1)	(2.0)	(1.4)	(3.0)	(2.2)	(2.2)
mDLDA with US	45.8	58.1	31.3	53.7	56.7	72.5	39.0	79.3	57.0	76.1	41.3	80.9
(SD)	(3.3)	(7.5)	(5.9)	(4.9)	(1.0)	(1.9)	(1.6)	(1.7)	(1.2)	(2.3)	(1.9)	(1.8)

ER-, ER+/++, ER+++	Huang				Pittman			
	PA	$\mathbf{PA}_1$	$\mathbf{PA}_2$	$\mathbf{PA}_3$	**PA**	$\mathbf{PA}_1$	$\mathbf{PA}_2$	$\mathbf{PA}_3$
(n_k)	(80)	(12)	(42)	(26)	(158)	(48)	(74)	(36)
Friedman's DLDA	46.3	50.0	42.9	50.0	67.1	83.3	40.0	71.4
mDLDA	40.0	33.3	38.1	46.2	62.7	72.9	45.0	65.7
Friedman's DLDA with US	49.4	79.2	35.8	57.6	68.4	84.1	45.9	70.4
(SD)	(3.9)	(6.2)	(5.2)	(6.8)	(1.3)	(1.3)	(3.7)	(2.5)
mDLDA with US	48.5	76.6	40.9	48.0	68.0	81.6	42.7	73.0
(SD)	(3.9)	(6.1)	(6.1)	(8.0)	(1.3)	(1.4)	(3.4)	(1.8)

ER-, ER+, ER++/+++	Huang				Pittman			
	PA	$\mathbf{PA}_1$	$\mathbf{PA}_2$	$\mathbf{PA}_3$	**PA**	$\mathbf{PA}_1$	$\mathbf{PA}_2$	$\mathbf{PA}_3$
(n_k)	(80)	(12)	(24)	(44)	(158)	(48)	(40)	(70)
Friedman's DLDA	56.3	50.0	33.3	70.5	63.9	83.3	59.5	47.2
mDLDA	52.5	41.7	29.2	68.2	58.2	79.2	48.6	50.0
Friedman's DLDA with US	54.1	65.7	33.2	62.3	55.9	84.2	39.4	52.2
(SD)	(3.5)	(9.3)	(5.6)	(4.5)	(2.4)	(1.3)	(4.1)	(5.8)
mDLDA with US	53.7	66.0	37.7	59.1	55.4	84.2	35.9	57.4
(SD)	(4.2)	(8.8)	(8.6)	(5.6)	(1.9)	(1.2)	(3.0)	(4.9)

5 Discussion

It is recognized that multi-class classification tasks are generally significantly harder than binary classification tasks (Mukherjee, 2003). In practice, the complexity of multi-class classification problems is often further increased by the fact that some of the classes can be considerably smaller than the others, i.e., by the class imbalance problem. The lack of minority class data can be one source of additional difficulty in the class-imbalanced problems. However, most of the commonly used classifiers have a built-in bias towards the classification into the majority class, which can contribute substantially to the poor

performance in the class-imbalanced setting. It was previously shown that for most classifiers this bias is further increased when the data are high-dimensional and/or if some type of variable selection is performed (Blagus and Lusa, 2010).

In this paper we focused on class-imbalanced multi-class high-dimensional data, and on the Friedman's method, a one-versus-one approach that reduces the multi-class problem in a series of binary classification problems and assigns the samples to the class with most votes. For the three-class classification problems we showed how the class probabilities deriving from the Friedman's approach are related to the class probabilities of the binary class-prediction sub-problems: the classification bias towards the majority class observed in the binary problems is carried over to the multi-class problem. Moreover, the classifiers that are most sensitive to class-imbalance in the two-class problems are deemed to be also the most sensitive in the multi-class setting when Friedman's approach is used. Another consequence is that also for multi-class problems the high-dimensionality of the data increases the bias towards the classification in the majority class. We did not consider any one-versus-all approaches because we expected that they would be even more sensitive to class-imbalance because the binary classifiers would be trained on even more imbalanced data.

A simple comparison of four classification methods (3-NN, DLDA, RF and SVM) showed that DLDA was the least sensitive to class-imbalance also in the multi-class setting when using the Friedman's approach. We focused on DLDA and compared the results obtained with Friedman's approach with those derived from the straightforward generalization of DLDA to multi-class problems (mDLDA). In practice the two methods differ in only two aspects: (i) the distance between the new data and the class centroids is standardized using a pooled variance of the variables, which is estimated using the data from pairs of classes for the Friedman's DLDA, while it is based on all the data for mDLDA and (ii) the variable selection method can be different.

It is common practice to select the variables to include in a multi-class classifier based on some statistic that compares all the classes at the same time; for example, we used F-test to rank the variables for mDLDA. We decided to investigate Friedman's approach with variable selection based on pairwise class comparisons (pairwise t-tests), obtaining a different ranking of the variables for each binary sub-problem. This approach mimics more closely the procedure that would be used in a two-class problem. The other reason for this choice was that we observed that mDLDA and Friedman's DLDA obtained very similar results if the same variable selection method was used; the similarity was more striking in the simulations, while some differences were observed analyzing the real data. The reason was that the estimates of the pooled variances sometimes differed substantially if they were based on all classes rather than on pairs of classes. This problem did not arise for the simulated data because the within-class variability was the same for all the classes.

The simulation results under the null case confirmed the general considerations outlined above. When the classes were imbalanced most new samples were classified in the majority class and increasing the number of variables increased the bias towards the majority class; variable selection reduced the bias but did not remove it. Friedman's DLDA was more sensitive to class-imbalance when the variables were selected, as the use of the pairwise t-tests was less effective in removing the bias compared to the F-test used with mDLDA. The reason is that by using the pairwise t-tests we further increased the high-dimensionality of the data, as the number of performed statistical tests was three times

larger compared to mDLDA. It was therefore more likely that variables exhibiting spurious and large differences between the pairs of classes would be used. We expect that the differences between the F-test and pairwise t-test results would be even more pronounced if the number of classes was larger.

In multi-class problems the classes can differ in many ways, and this is the reason why the alternative case results are more are complex to interpret compared to two-class setting. For example, while it is possible that all the classes are different, it is also possible that some of the classes cannot be discriminated using the measured variables, mixing null and alternative case situations in the same classification problem. Furthermore, the majority class can be either the easiest or the most difficult to predict, depending on whether it exhibits the least or the most differences compared to the other classes. For example, in our re-analysis of breast cancer gene-expression data the patients with grade 2 were the most abundant in two of the three data sets and grade 2 was the most difficult class to predict. We tried to take some of these possibilities into account in our simulations considering different scenarios, which by no means can be considered exhaustive of all the possible situations that can arise in practice. It is therefore very difficult to make some general conclusions about the effect of class-imbalance in multi-class problems when the classes are different.

Our findings suggest that the class-imbalance reduced the predictive accuracy of the minority classes and increased the probability of classifying new samples in the majority class, similarly as it was observed in the null case. The variable selection method played an important role. The statistical power for identifying the truly differentially expressed variables was better for the F-test compared to the pairwise t-tests; the pairwise t-tests had better power for the variables that distinguished the most the pairs of classes. However, there was no clear winner between mDLDA and Friedman's DLDA. In the three simulation scenarios that we considered mDLDA performed similarly to Friedman's DLDA in scenario 1 (no difference between the minority classes), better in scenario 2 (majority class nested between the minority classes) and worse in scenario 3 (no nesting of the classes); in the real data analysis Friedman's DLDA seemed to perform slightly better than mDLDA.

In binary classification the bias in favor of the majority class can be attenuated with simple random undersampling (Blagus and Lusa, 2010). Undersampling consists in obtaining a class-balanced training set by removing a subset of samples from the majority class. Our results suggest that undersampling can attenuate the bias also in the multi-class classification. In the alternative case the results obtained using undersampling were similar to those obtainable using balanced data. Therefore, undersampling did not necessarily balance the predictive accuracy of all the classes but it improved the minority class predictive accuracy. In the analysis of real data undersampling did not worsen the performance of the classifiers, and in some situations it improved the predictive accuracy of the minority class. This is to some extent surprising, as a much smaller sample sizes were used with undersampling. This result seems to support the view that it is important to plan experiments using balanced data, whenever possible.

In our analysis we considered only DLDA as a base classifier. It is therefore possible that other classification methods will be less sensitive to class imbalance. However, our theoretical results and limited simulation results suggest that bias caused by the class-imbalance will be even larger for other classifiers. Nevertheless it would be beneficial to

compare different classifiers using simulated and real high-dimensional class-imbalanced data, simulating also correlated variables. A further limitation of our study was that we considered only random undersampling as an approach to reduce the bias. More efficient methods exist in the literature; for example, we previously showed that multiple undersampling outperforms random undersampling (Blagus and Lusa, 2010) and it would be interesting to compare different undersampling approaches also for multi-class classification problems. Note however that on our analysis of the real data sets we repeated the subset selection many times, which can be seen as a simplified form of multiple downsizing.

6 Conclusions

Our results show that the class-imbalance has a significant effect on the classification results also in multi-class problems and that its influence is magnified when the number of variables is large. The amount of bias depends also jointly on the magnitude of the differences between classes and on the sample size, i.e. the bias diminishes when the differences between the classes are larger or the sample size is increased. Also variable selection plays an important role in the class-imbalance problem and the most effective strategy depends on the type of differences that exist between classes. DLDA seems to be among the least sensible classifiers to class-imbalance and its use is recommended also for multi-class problems. Whenever possible the experiments should be planned using balanced data in order to avoid the further complications arising from the class-imbalance.

References

[1] Allwein, E. L., Schapire, R. E., and Singer, Y. (2001): Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, **1**, 113–141.

[2] Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004): A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, **6**, 20–29.

[3] Berrar, D. P., Downes, C. S., and Dubitzky, W. (2003): Multiclass cancer classification using gene expression profiling and probabilistic neural networks. *In Proceedings of the Pacific Symposium on Biocomputing*, 5–16.

[4] Bishop, C. M. (2007): *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed.

[5] Blagus, R. and Lusa, L. (2010): Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, **11**, 523.

[6] Brown, P. and Botstein, D. (1999): Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, **21**, 33–37.

[7] Dietterich, T. G. and Bakiri, G. (1995): Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, **2**, 263–286.

[8] Friedman, J. H. (1996): Another approach to polychotomous classification. Technical report.

[9] Hastie, T. and Tibshirani, R. (1998): Classification by pairwise coupling. *The Annals of Statistics*, **26**, 451–471.

[10] He, H. and Garcia, E. A. (2009): Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263–1284.

[11] Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, k., and Trent, J. (2001): Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, **344**, 539–548.

[12] Huang, E., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R., and Huang, A. T. (2003): Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–1596.

[13] Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., and et al. (2006): Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Research*, **66**, 10292–10301.

[14] Izenman, A. J. (2008): *Modern Multivariate Statistical Techniques : Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. New York: Springer.

[15] Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., and et al. (2005): An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13550–13555.

[16] Mukherjee, S. (2003): Classifying microarray data using support vector machines machines,understanding and using microarray analysis techniques: A practical guide. Boston: Kluwer Academic Publishers.

[17] Pepe, M. S. (2003): *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.

[18] Pittman, J., Huang, E., Dressman, H., Horng, C.-F., Cheng, S. H., Tsou, M.-H., Chen, C.-M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R., West, M., and

Berger, J. O. (2004): Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 8431–8436.

[19] R Development Core Team (2008): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[20] Romualdi, C., Campanaro, S., Campagna, D., Celegato, B., Cannata, N., Toppo, S., Valle, G., and Lanfranchi, G. (2003): Pattern recognition in gene expression profiling using dna array: a comparative study of different statistical methods applied to cancer classification. *Human Molecular Genetics*, **12**, 823–836.

[21] Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., and Zhao, Y. (2004): *Design and Analysis of DNA Microarray Investigations*. New York: Springer.

[22] Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003): Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Science USA*, **100**, 10393–10398.

[22] Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M., and Delorenzi, M. (2006): Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**, 262–272.

[23] Speed, T. P. (2003): *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC.

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

[25] Tsoumakas, G. and Katakis, I. (2007): Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, **2007**, 1–13.