THIELE CENTRE
FOR APPLIED MATHEMATICS IN NATURAL SCIENCE

High dimensional classifiers in the imbalanced case

Britta Anker Bak and Jens Ledet Jensen

Research Report

No. 04 I May 2015

High dimensional classifiers in the imbalanced case

Britta Anker Bak and Jens Ledet Jensen

Department of Mathematics, Aarhus University

Abstract

We consider the binary classification problem in the imbalanced case where the number of samples from the two groups differ. The classification problem is considered in the high dimensional case where the number of variables is much larger than the number of samples, and where the imbalance leads to a bias in the classification. A theoretical analysis of the independence classifier reveals the origin of the bias and based on this we suggest two new classifiers that can handle any imbalance ratio. The analytical results are supplemented by a simulation study, where the suggested classifiers in some aspects outperform multiple undersampling. For correlated data we consider the ROAD classifier and suggest a modification of this to handle the bias from imbalanced group sizes.

1 Introduction

During the last decade much research in the statistical community has been on classifiers for high dimensional data where the sample size is small, see e.g. Donoho and Jin (2009), Cai and Liu (2011) and Fan et al. (2012). Typically, this research has not focussed on the imbalance problem where the sample sizes of the groups differ. In real life experiments, on the other hand, imbalanced data sets are the norm rather than the exception. Even if scientists decide to collect a balanced data set, missing data due to for example patients dropping out of the experiment or invalid measurements commonly leads to imbalance.

Faced with imbalance most classifiers tend to classify observations from a binary classification problem to the majority group at the expense of the minority group. It appears to be overlooked or neglected that this imbalance problem becomes much more pronounced in high dimensional settings. To briefly illustrate this Table 1 gives the mean and standard deviation of the probability of correct classification for both groups in a few instances for the thresholded independence classifier. It is clearly seen that even rather small imbalances seriously harm classification, pointing to the need of correcting for all imbalances.

The imbalance problem has, however, been addressed recently in the computer science and engineering communities. Here the focus has been on reducing to the balanced case by either undersampling or oversampling. Lin et al. (2009), Yang et al.

(2014) and Liu et al. (2009) introduced Meta Imbalanced Classification Ensemble (MICE), Sample Subset Optimization (SSO) and BalanceCascade, respectively. Those are all ensemble methods, where several classifiers are build on all observations in the minority group and wisely selected subsamples of the majority group. Chawla et al. (2002) propose a technique where the minority group is extended by adding observations on the line segments between an existing minority observation and its nearest neighbours. The above classifiers are studied empirically rather than theoretically, and are all shown to handle imbalanced classification problems well. Typically, the high dimensional situation is not addressed as a problem in itself.

The aim of the present paper is to analyse the imbalance problem in relation to high dimensional binary classification and, building on this analysis, to suggest classifiers that are not based on undersampling or oversampling. Ideally, we want our classifiers to involve a small number of variables only, while maintaining a high probability of correct classification. To this end we consider a simple classification problem between two groups with independent normally distributed variables. The assumption of independent variables is a simplification in relation to most data sets, but the setting is useful for studying the imbalance problem in high dimensional settings, and the classifiers are also of practical relevance for correlated variables.

After detecting the origin of the bias problem for imbalanced data in Section 2, we suggest in Section 3 two new classifiers with, practically, no bias. We discuss the properties of the suggested classifiers both theoretically and empirically. Turning to a situation with correlated variables in Section 6, we find that the corrections introduced for the case of independent variables can be combined with the ROAD classifier of Fan et al. (2012) for the imbalanced case. This suggests that the introduced correction methods can be helpful for a range of linear classifiers in more general situations.

Table 1: Average probability of correct classification of the thresholded independence classifier for a new observation from each of two groups. There are n samples from group 1 and m samples from group 2. Each observation has 1000 variables of which only 10 have a differential expression of size 1. Values are based on 1000 simulated data sets.

		Group 1		Group 2	
n	m	Mean	Std	Mean	Std
15	15	70.5	7.0	70.3	7.1
16	14	76.8	6.2	63.2	7.7
18	12	87.4	4.5	44.9	8.7
20	10	94.9	2.2	24.7	7.2

2 The bias problem for imbalanced data

The model we consider is as follows. Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$ be p-dimensional observations from group 1 and group 2, respectively. Assume all observations and variables are independent with distributions $x_{ij} \sim N(\mu_j, \sigma_j^2)$ and $y_{ij} \sim N(\mu_j + \delta_j \sigma_j, \sigma_j^2)$. Let $\bar{x}_j$ and $\bar{y}_j$ denote the sample means of variable j for each of the two groups, and let s_{xj}^2 and s_{yj}^2 be the corresponding sample variances. Define the imbalance factor as $\rho = (n-m)/(nm)$, and let $f = n+m-2$ be the degrees of freedom for the joint sample variance. We call δ_j the (scaled) differential expression.

To describe the independence classifier with thresholding we first define for $j = 1, \ldots, p$

$$s_j^2 = \frac{(n-1)s_{xj}^2 + (m-1)s_{yj}^2}{n+m-2}, \qquad t_j = \frac{\bar{y}_j - \bar{x}_j}{\sqrt{s_j^2(1/n+1/m)}},$$

and let $w(t)$ be a weight function. Hard thresholding, which we use throughout this paper, corresponds to $w(t) = 1(|t| > \Delta)$. The independence classifier with thresholding allocates a new observation z to group 1 if $D(z) < 0$ and to group 2 if $D(z) > 0$, where

$$D(z) = \sum_{j=1}^{p} \frac{\bar{y}_j - \bar{x}_j}{s_j^2}\big[z_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j)\big]w(t_j). \tag{2.1}$$

The probability of correct classification for a new observation from either group 1 or group 2 is

$$\Phi\Big(\frac{\xi_D}{\tau_D}\Big) \qquad \text{and} \qquad \Phi\Big(\frac{\tilde{\xi}_D}{\tau_D}\Big), \tag{2.2}$$

where $\xi_D = -D(\mu) = \sum_{j=1}^p \xi_{Dj}$, $\tilde{\xi}_D = D(\mu+\delta\sigma) = \sum_{j=1}^p \tilde{\xi}_{Dj}$, $\tau_D^2 = \sum_{j=1}^p w(t_j)(\bar{y}_j - \bar{x}_j)^2\sigma_j^2/s_j^4$ and

$$\begin{aligned} \xi_{Dj} &= -\frac{\bar{y}_j - \bar{x}_j}{s_j^2}\big[\mu_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j)\big]w(t_j), \\ \tilde{\xi}_{Dj} &= \frac{\bar{y}_j - \bar{x}_j}{s_j^2}\big[\mu_j + \delta_j\sigma_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j)\big]w(t_j). \end{aligned}$$

To describe the means of these terms define

$$T_{a,b}(\delta; n, m) = E\Big[\frac{(d+\delta)^a}{v^b}w(t)\Big],$$

where $d \sim N(0, 1/n+1/m)$, $v \sim \chi^2(f)/f$ with $f = n+m-2$ and $t = (d+\delta)/\sqrt{v(1/n+1/m)}$.

Proposition 1. *Let ξ_D^0 and $\tilde{\xi}_D^0$ be generic terms in the sums ξ_D and $\tilde{\xi}_D$. Then*

$$\begin{aligned} E(\xi_D^0) &= \tfrac{1}{2}\big[(1-\rho)\delta T_{1,1}(\delta, n, m) + \rho T_{2,1}(\delta, n, m)\big], \\ E(\tilde{\xi}_D^0) &= \tfrac{1}{2}\big[(1+\rho)\delta T_{1,1}(\delta, n, m) - \rho T_{2,1}(\delta, n, m)\big]. \end{aligned}$$

When $\delta = 0$ we simply get $E(\xi_D^0) = -E(\tilde{\xi}_D^0) = \frac{1}{2}\rho T_{2,1}(0, n, m)$. For the case of no thresholding, $w(t) \equiv 1$, we get in the general case

$$E(\xi_D^0) = \frac{f}{2(f-2)}[\delta^2 + \rho(1/n+1/m)], \quad E(\tilde{\xi}_D^0) = \frac{f}{2(f-2)}[\delta^2 - \rho(1/n+1/m)].$$

Proof. Letting $u = (\bar{x} + \bar{y} - 2\mu - \delta)/\sigma \sim N(0, 1/n + 1/m)$, $d = (\bar{y} - \bar{x} - \delta)/\sigma \sim N(0, 1/n + 1/m)$ and $v = s^2/\sigma^2 \sim \chi^2(f)/f$ with $f = n + m - 2$, we can write

$$\xi_D^0 = \frac{d+\delta}{2v}(u+\delta)w(t) \quad \text{and} \quad \tilde{\xi}_D^0 = \frac{d+\delta}{2v}(\delta - u)w(t),$$

with $t = (d+\delta)/\sqrt{v(1/n + 1/m)}$. Had u and d been independent, ξ_D^0 and $\tilde{\xi}_D^0$ would have the same mean and there would be no bias problem. However, in the imbalanced case we have

$$u|d \sim N\Big(\rho d, \frac{4}{n+m}\Big). \tag{2.3}$$

We then obtain

$$E(\xi_D^0) = E\Big[\frac{d+\delta}{2v}(\rho d + \delta)w(t)\Big] = \tfrac{1}{2}E\Big\{\Big[\rho\frac{(d+\delta)^2}{v} + \delta(1-\rho)\frac{d+\delta}{v}\Big]w(t)\Big\},$$

and $E(\tilde{\xi}_D^0)$ is calculated in the same way.

In the case of no thresholding, $w(t) \equiv 1$, we use that $E(1/v) = f/(f-2)$ so that

$$E\Big(\frac{(d+\delta)^2}{v}\Big) = \Big(\frac{1}{n} + \frac{1}{m} + \delta^2\Big)\frac{f}{f-2} \quad \text{and} \quad E\Big(\frac{d+\delta}{v}\Big) = \delta\frac{f}{f-2}.$$

□

The case of no differential expression ($\delta = 0$) in the proposition shows that if the expected number $pE(w(t))$ of variables with $\delta = 0$ included in the classifier is nonnegligible, then also the bias of the classifier is nonnegligible with the majority class being strongly favoured. In the general case, with $\delta \neq 0$, the formulae point to a bias in the same direction as in the $\delta = 0$ case. This is seen more directly for the case of no thresholding. Overall, the thresholding does not remove the bias problem for the imbalanced case. This can be seen more clearly from the left part of Figure 1. The two dotted curves illustrate the bias for the case of no differential expression. The figure shows the mean for a single term of ξ_D and $\tilde{\xi}_D$, conditional on this term being included in the classifier. The two dashed curves show the bias when the differential expression is one. The virtue of increasing the threshold is that we include much fewer of the $\delta = 0$ cases and keep most of the $\delta = 1$ cases. There are, however, a number of opposing effects. When the threshold is increased, the bias for each of those null cases included actually increases. Also, since the mean of τ_D^{02} is increasing with the threshold, the effect of each of the $\delta = 1$ cases in the probability (2.2) is diminished as the threshold is increased. The right part of Figure 1 relates to the classifiers proposed in the next section.

3 Bias adjusted classifiers

In this section we describe two ways of circumventing the bias problem in the imbalanced case. The origin of the bias problem is the lack of independence of $\bar{x}_j + \bar{y}_j$ and $\bar{y}_j - \bar{x}_j$ as stated in (2.3).

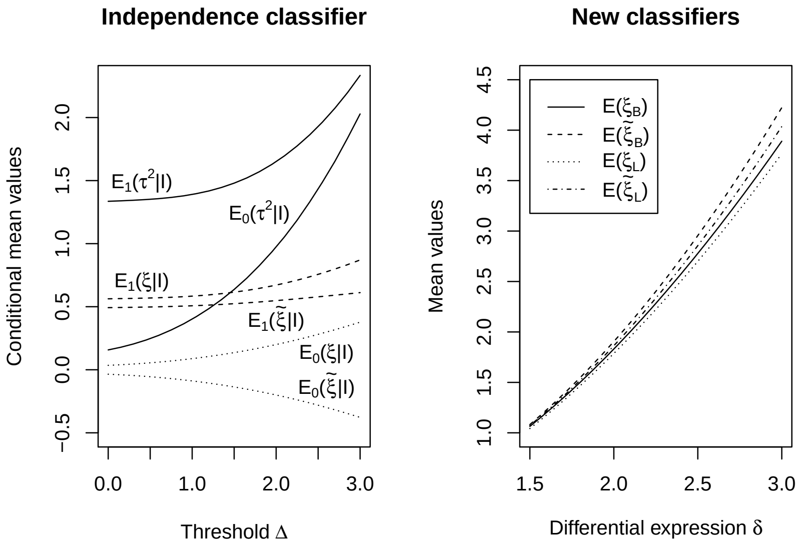

Figure 1: The left part shows the mean of a generic term ξ_D^0, $\tilde{\xi}_D^0$ and τ_D^{02} conditionally on the term being included, that is, given that $w(t) = 1$. Two cases of the differential expression are shown: $\delta = 0$ and $\delta = 1$ shown by the subscript on the mean value sign. The right part shows the mean value of ξ and $\tilde{\xi}$ for the two classifiers proposed in Section 3. The threshold here depends on the differential expression: $\Delta = \delta/\sqrt{1/n + 1/m} - 1$. In both figures $n = 30$ and $m = 10$.

The first proposal is simply to subtract the conditional mean from (2.3). Thus we consider

$$B_0(z) = \sum_{j=1}^{p} \frac{\bar{y}_j - \bar{x}_j}{s_j^2}\Big[z_j - \frac{1}{2}(\bar{x}_j + \bar{y}_j) + \frac{\rho}{2}(\bar{y}_j - \bar{x}_j)\Big] w(t_j).$$

Let $\xi_{B_0}^0$ be minus a generic term in the sum with z_j replaced by μ_j, and let $\tilde{\xi}_{B_0}^0$ be a generic term with z_j replaced by $\mu_j + \delta_j\sigma_j$. Then, with calculations as in Proposition 1, we find

$$E(\xi_{B_0}^0) = \frac{1-\rho}{2}\delta T_{1,1}(\delta, n, m), \quad E\big(\tilde{\xi}_{B_0}^0\big) = \frac{1+\rho}{2}\delta T_{1,1}(\delta, n, m).$$

Most importantly, we see here that the bias originating from those variables with $\delta = 0$ has been removed. However, there remains a bias for variables with $\delta \neq 0$, where now the minority group is favoured.

We therefore consider a classifier on the form $B_0(z) - \epsilon$ for some constant ϵ. Optimally, we want $\xi_{B_0} + \epsilon = \tilde{\xi}_{B_0} - \epsilon$ or $\epsilon = (\tilde{\xi}_{B_0} - \xi_{B_0})/2$. We estimate ξ_{B_0} and $\tilde{\xi}_{B_0}$ by a leave-one-out cross-validation and use these to correct the classifier. To this end we define $B_0(z; x_i)$ to be the classifier based on the reduced sample with x_i excluded and, similarly, $B_0(z; y_i)$ is based on the reduced sample with y_i excluded. Define

$$\bar{\epsilon} = \frac{1}{2}\Big[\frac{1}{n}\sum_{i=1}^{n} B_0(x_i; x_i) + \frac{1}{m}\sum_{i=1}^{n} B_0(y_i; y_i)\Big].$$

Since B_0 is a sum over all p variables, we can also write $\bar{\epsilon}$ as a sum $\bar{\epsilon} = \sum_{j=1}^{p} \bar{\epsilon}_j$, where $\bar{\epsilon}_j$ depends on the j'th coordinate of the data only. The *bias adjusted independence* classifier (BAI classifier) is now defined as

$$B(z) = B_0(z) - \bar{\epsilon} = \sum_{j=1}^{p} \Big\{ \frac{\bar{y}_j - \bar{x}_j}{s_j^2} \Big[z_j - \frac{1}{2}(\bar{x}_j + \bar{y}_j) + \frac{\rho}{2}(\bar{y}_j - \bar{x}_j) \Big] w(t_j) - \bar{\epsilon}_j \Big\}.$$

Defining ξ_B^0 and $\tilde{\xi}_B^0$ as the generic terms of $-B(\mu)$ and $B(\mu + \delta\sigma)$, we see that

$$\begin{aligned} E(\xi_B^0) &= \frac{\delta}{2}\Big\{(1-\rho)T_{1,1}(\delta, n, m) - \frac{1-\rho_1}{2} T_{1,1}(\delta, n-1, m) + \frac{1+\rho_2}{2} T_{1,1}(\delta, n, m-1)\Big\}, \\ E(\tilde{\xi}_B^0) &= \frac{\delta}{2}\Big\{(1+\rho)T_{1,1}(\delta, n, m) + \frac{1-\rho_1}{2} T_{1,1}(\delta, n-1, m) - \frac{1+\rho_2}{2} T_{1,1}(\delta, n, m-1)\Big\}, \end{aligned} \tag{3.1}$$

where $\rho = (n-m)/(n+1)$, $\rho_1 = (n-m-1)/(n+m-1)$ and $\rho_2 = (n-m+1)/(n+m-1)$. Since $\bar{\epsilon}$ is based on one less observation than B_0, the BAI classifier is not exactly unbiased, but the remaining bias is of no practical concern. The bias of the BAI classifier is illustrated in the right part of Figure 1 for the case $n = 30$ and $m = 10$. When the differential expression δ is less than 1.5, the bias is very small.

When calculating the probability of correct classification as in (2.2), the denominator is $\tau_B^2 = \sum_{j=1}^{p} w(t_j)(\bar{y}_j - \bar{x}_j)^2 \sigma_j^2 / s_j^4$, that is, the same expression as τ_D^2.

We next consider a different approach for removing the bias of the independence classifier in the imbalanced case. First, we rewrite the independence classifier as

$$D(z) = \frac{1}{2} \sum_{j=1}^{p} \Big[\frac{1}{n} \sum_{i=1}^{n} \frac{\bar{y}_j - \bar{x}_j}{s_j^2}(z_j - x_{ij}) w(t_j) + \frac{1}{m} \sum_{i=1}^{m} \frac{\bar{y}_j - \bar{x}_j}{s_j^2}(z_j - y_{ij}) w(t_j) \Big].$$

The origin of the bias problem, as given in (2.3), is here seen as the lack of independence of x_{ij} (or y_{ij}) and $\bar{y}_j - \bar{x}_j$. We suggest to solve this by removing x_{ij} (or y_{ij}) when calculating the difference $\bar{y}_j - \bar{x}_j$. Thus let $\bar{x}_j(i)$ and $\bar{y}_j(i)$ be the group averages when the i'th observation is left out, and let $s_j^2(x_i)$ and $s_j^2(y_i)$ be the within group variance when either x_i or y_i is left out. The corresponding t-value is denoted either $t_j(x_i)$ or $t_j(y_i)$. The *leave one out* independence classifier (LOUI classifier, originally suggested in Jensen (2006)) is defined as

$$\begin{aligned} L(z) = \frac{1}{2} \sum_{j=1}^{p} \Big[&\frac{1}{n} \sum_{i=1}^{n} \frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)}(z_j - x_{ij}) w(t_j(x_i)) \\ &+ \frac{1}{m} \sum_{i=1}^{m} \frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)}(z_j - y_{ij}) w(t_j(y_i)) \Big]. \end{aligned}$$

Defining ξ_L^0 and $\tilde{\xi}_L^0$ as a generic term in $-L(\mu)$ and $L(\mu + \delta\sigma)$, we see that

$$E(\xi_L^0) = \tfrac{1}{2}\delta T_{1,1}(\delta, n, m-1) \quad \text{and} \quad E\big(\tilde{\xi}_L^0\big) = \tfrac{1}{2}\delta T_{1,1}(\delta, n-1, m).$$

The difference between these two terms is very small so that the LOUI classifier is almost unbiased. An example is shown in the right part of Figure 1 for the case $n = 30$ and $m = 10$.

When calculating the probability of correct classification, as in (2.2), the denominator is now

$$\tau_L^2 = \frac{1}{4}\sum_{j=1}^{p}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)/\sigma_j^2}w(t_j(x_i)) + \frac{1}{m}\sum_{i=1}^{m}\frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)/\sigma_j^2}w(t_j(y_i))\Big]^2,$$

which is somewhat more complicated than for the independence classifier and the BAI classifier.

A comparison of the two proposed classifiers BAI and LOUI is given in Section 5.

4 Distribution approximation of the error probability

We are mostly interested in situations where the number of variables with a nonzero differential expression is quite small, and the sample sizes n and m are not sufficiently large for a complete separation between the variables with a nonzero differential expression and those with no differential expression. The classifier therefore typically includes a limited number of variables and a part of these are null variables. The probability of correct classification given through ξ/τ and $\tilde{\xi}/\tau$ in (2.2) therefore has a fairly large variance, and part of this variance stems from the variance of the denominator τ. Actually, both ξ and τ turn out to have fairly large variances and a strong correlation.

We want to be able to look at the mean and variance of ξ/τ and $\tilde{\xi}/\tau$ for various combinations of the differential expressions δ_j in an easy computable way for the case of independent variables. This means that we want to use only moment values of generic terms ξ^0, $\tilde{\xi}^0$ and τ^{02}. For this purpose we use the following rough approximation

$$\xi|\tau^2 \approx N(\alpha + \beta\tau^2, \omega^2), \quad \tilde{\xi}|\tau^2 \approx N\big(\tilde{\alpha} + \tilde{\beta}\tau^2, \tilde{\omega}^2\big), \quad \tau^2 \approx \Gamma(\lambda, \kappa). \tag{4.1}$$

The approximation is illustrated in Figure 2. The left subfigure shows the approximate linear relationship $E(\xi_B|\tau_B^2) \approx \alpha + \beta\tau_B^2$, the center figure shows approximate normality of ξ_B given τ_B^2 and the right subfigure shows the Gamma approximation to the distribution of τ_B^2. Plots for the thresholded independence classifier and the LOUI classifier show that the approximation also works well in these cases.

Lemma 2. *Under the above approximation* (4.1) *we have*

$$E\Big(\frac{\xi}{\tau}\Big) \approx \alpha\sqrt{\kappa}\frac{\Gamma(\lambda - \frac{1}{2})}{\Gamma(\lambda)} + \beta\frac{\Gamma(\lambda + \frac{1}{2})}{\sqrt{\kappa}\Gamma(\lambda)},$$

$$\mathrm{Var}\Big(\frac{\xi}{\tau}\Big) \approx (\omega^2 + \alpha^2)\frac{\kappa}{\lambda - 1} + \beta^2\frac{\lambda}{\kappa} + 2\alpha\beta - \Big\{E\Big(\frac{\xi_N}{\tau_N}\Big)\Big\}^2,$$

with similar expressions for $\tilde{\xi}$ with (α, β) replaced by $(\tilde{\alpha}, \tilde{\beta})$.

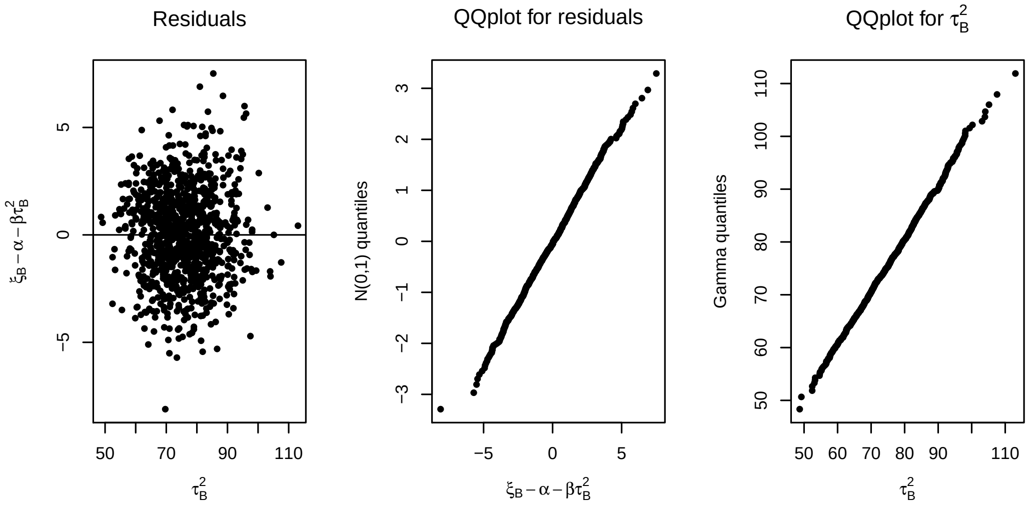

Figure 2: Illustration of the approximation (4.1) for the BAI classifier. The 1000 simulated values of ξ_B and τ_B^2 are for the case $n = 30$, $m = 10$, $\delta = 1$ and $\Delta = 2$. There are $p = 1000$ variables of which $k = 20$ are differentiably expressed. The left subfigure shows the approximate linearity of the conditional mean of ξ_B given τ_B^2, the center figure shows the conditional normality and the right subfigure illustrates the Gamma approximation to the distribution of τ_B^2.

Proof. We have $E(\xi/\tau) = \alpha E(1/\tau) + \beta E(\tau)$ and the first result follows from the properties of a gamma distribution. Next,

$$\begin{aligned}\mathrm{Var}(\xi/\tau) &= \mathrm{Var}(\alpha/\tau + \beta\tau) + E(\omega^2/\tau^2) \\ &= (\omega^2 + \alpha^2)E(1/\tau^2) + \beta^2 E(\tau^2) + 2\alpha\beta - [E(\xi/\tau)]^2.\end{aligned}$$

and the result for the variance again follows from properties of the gamma distribution. □

To use this in practice we choose the parameters in (4.1) from moment relations:

$$\begin{gathered}\frac{\lambda}{\kappa} = E(\tau^2), \quad \frac{\lambda}{\kappa^2} = \mathrm{Var}(\tau^2), \quad \mathrm{Cov}(\xi, \tau^2) = \beta\,\mathrm{Var}(\tau^2), \\ E(\xi) = \alpha + \beta E(\tau^2), \quad \mathrm{Var}(\xi) = \beta^2\,\mathrm{Var}(\tau^2) + \omega^2.\end{gathered}$$

We consider now in detail the moments for the BAI classifier. We write a generic term of the sums ξ_B and τ_B^2 as

$$\xi_B^0 = \xi_{B_0}^0 + \frac{1}{n}\sum_{i=1}^n B_0^0(x_i) + \frac{1}{m}\sum_{i=1}^n B_0^0(y_i), \quad \tau_B^{02} = \frac{(\bar{y} - \bar{x})^2}{s^4},$$

where $B_0^0(x_i)$ is a generic term in the sum $B_0(x_i; x_i)$ and $B_0^0(y_i)$ is a generic term in the sum $B_0(y_i; y_i)$. The first two moments can be simulated directly from standard normal variables $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, and calculating all the terms in ξ_B^0. However, the computational complexity can be reduced on writing variances and covariances as sums involving at most two terms from ξ_B^0. In that case we only need to simulate x_1, x_2, $\bar{x}(3) = \sum_{i=3}^n x_i/(n-3)$, $\sum_{i=3}^n (x_i - \bar{x}(3))^2$ (and similar y-terms), and calculate $\xi_{B_0}^0$, $B_0^0(x_1)$ and $B_0^0(x_2)$ from these. To this end, and supplementing the mean values in (3.1), we note the following simplifications.

Proposition 3. *For the case of hard thresholding we have the following moment relations:*

$$E\big(\tau_B^{02}\big) = T_{2,2}(\delta; n, m), \quad E\big(\tau_B^{04}\big) = T_{4,4}(\delta; n, m),$$
$$E\big(\xi_{B_0}^{02}\big) = \Big[\frac{1}{n+m} + \frac{(1-\rho)^2}{4}\delta^2\Big] T_{2,2}(\delta; n, m), \quad E\big(\xi_{B_0}^{0}\tau_B^{02}\big) = \frac{1-\rho}{2}\delta T_{3,3}(\delta; n, m),$$
$$E[B_0^0(x_1)] = -\frac{1-\rho_1}{2}\delta T_{1,1}(\delta; n-1, m),$$
$$E[B_0^0(x_1)^2] = \Big[1 + \frac{1}{n-1+m} + \frac{(1-\rho_1)^2}{4}\delta^2\Big] T_{2,2}(\delta; n-1, m).$$

Proof. The proof follows the same lines as the proof of Proposition 1. The only extra element used is that $E[(u-\rho d)^2|d] = 4/(n+m)$ from (2.3). The requirement of hard thresholding is used for the simplification $w(t)^2 = w(t)$. □

4.1 Mean and variance investigations

In Figures 3 and 4 we compare the independence classifier D, the BAI-classifier B and the LOUI-classifier L. There are k differentiable expressed variables all with the same differential expression $\delta = 1$. We consider the two cases $k = 20$ and $k = 80$. In all cases we have $n = 30$ and $m = 10$. To calculate the mean and variance of ξ/τ we use the approximation in Lemma 2. To this end, we must calculate moments of generic terms ξ^0, $\tilde{\xi}^0$ and τ^{02} for the chosen value of δ for the expressed variables, as well as the case $\delta = 0$ for the nonexpressed variables. These moments cannot be calculated analytically, and we use 10^6 simulated values to estimate the moments. Note that the mean values μ_j and variances σ_j^2 do not enter the distribution of ξ^0, $\tilde{\xi}^0$ and τ^{02} so that we can fix these at zero and one, respectively.

In Figure 3 the threshold is fixed at $\Delta = 2$, and we consider the dependency on the differential expression δ for the k expressed variables in the range $0 < \delta < 1.5$. We consider the two cases $k = 20$ and $k = 80$, and either $p = 1000$ or $p = 10\,000$ variables. It is clearly seen that the independence classifier D performs much better on the majority group than on the minority group. For both BAI and LOUI there is practically no difference between the two groups, and also practically no difference between BAI and LOUI for the considered range of δ. For this reason only the BAI classifier is shown in Figure 3. Taking into account the random variation, and looking at the case $p = 10\,000$, we will indeed encounter simulations where the classifier is worse than a random guess unless the differential expression δ is large. For $k = 20$ and $\delta = 1$ this will happen in approximately 7% of the simulations. For $p = 1000$ variables the classifier is much more useful, although there is a considerable variation in ξ/τ giving a considerable variation in the probability of correct classification.

In Figure 4 the differential expression is fixed at $\delta = 1$, and we consider the dependency on the threshold Δ. As in Figure 3 the curves for the two classifiers BAI and LOUI as well as the curves for the two groups for each classifier are indistinguisable, and only one curve is shown. Clearly, a high threshold reduces the strong bias of the independence classifier D. Still, in most cases the median probability of correct classification for the minority group is below 0.5. Looking for the value of the threshold Δ, where the mean value of ξ_B/τ_B is maximized, no clear optimal

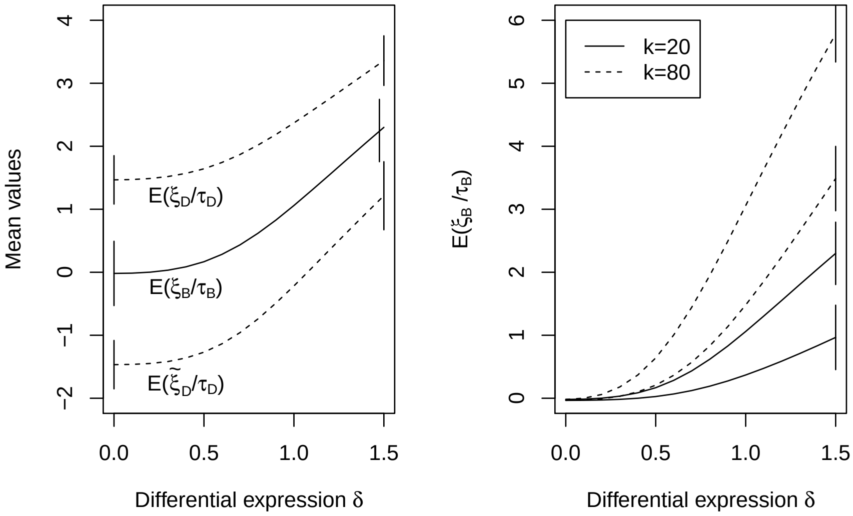

Figure 3: Performance of different classifiers for the case $n = 30$, $m = 10$ and with the threshold $\Delta = 2$. In the left part the means of the classification indices ξ_D/τ_D and $\tilde{\xi}_D/\tau_D$ are compared to the mean of ξ_B/τ_B for the case of $p = 1000$ variables with $k = 20$ having the differential expression δ, the remaining variables having no differential expression. In the right part the mean of the classification index ξ_B/τ_B is shown for different values of p and k. The value of k is shown in the legend, and the lower and upper curves of a specific line type correspond to $p = 10000$ and $p = 1000$, respectively. For the chosen settings of the parameters, the means of $\tilde{\xi}_B/\tau_B$, ξ_L/τ_L and $\tilde{\xi}_L/\tau_L$ are indistinguisable from the mean of ξ_B/τ_B. The vertical lines show plus and minus two times the standard deviation.

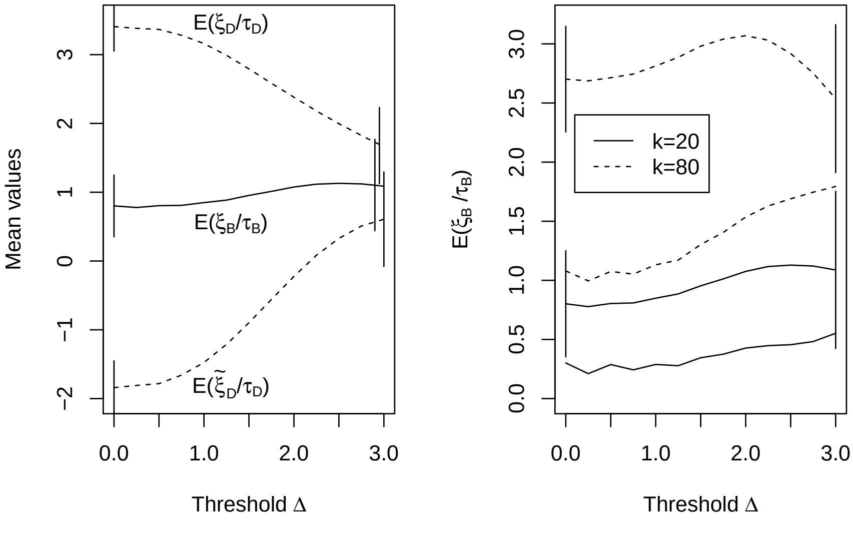

Figure 4: Performance of different classifiers for the case $n = 30$, $m = 10$ and with the differential expression $\delta = 1$. In the left part the means of the classification indices ξ_D/τ_D and $\tilde{\xi}_D/\tau_D$ are compared to the mean of ξ_B/τ_B for the case of $p = 1000$ variables with $k = 20$ having the differential expression $\delta = 1$, the remaining variables having no differential expression. In the right part the mean of the classification index ξ_B/τ_B is shown for different values of p and k. The value of k is shown in the legend, and the lower and upper curves of a specific line type correspond to $p = 10000$ and $p = 1000$, respectively. For the chosen settings of the parameters, the means of $\tilde{\xi}_B/\tau_B$, ξ_L/τ_L and $\tilde{\xi}_L/\tau_L$ are indistinguisable from the mean of ξ_B/τ_B. The vertical lines show plus and minus two times the standard deviation.

choice is seen for the case of $p = 10\,000$ variables. For $p = 1000$ the optimal value is between 2 and 2.5. However, the gain in mean value is partly reduced by having a large spread of ξ_B/τ_B when the threshold is increased.

5 Simulations

In this section we report on simulations to compare the suggested classifiers BAI and LOUI for the case of imbalanced data. We include also in the comparison a commonly used undersamling classifier, namely EasyEnsemble from Liu et al. (2009) built on top of the thresholded independence classifier. To write this explicitly, assume $n > m$ and let $D(z; A)$ be the independence classifier from (2.1) based on a subset A of the observations $x_1, \ldots, x_n$ from group 1 and all the observations from group 2 and with $|A| = m$. The undersampling classifier is based on

$$Q(z) = \frac{1}{q} \sum_{i=1}^{q} D(z; A_i), \tag{5.1}$$

where $A_1, \ldots, A_q$ are independent random subsets. In the results in Table 2 below we use a value of q such that the probability of using all the samples in the training of the classifier is at least 0.95.

We include the case of a fixed threshold in the comparisons, but we are mostly interested in the situation where the threshold Δ is chosen suitably for each simulated data set. In the simulations we have searched for a value of Δ in the range where a t-test will give between 1 and 30 false positives among p independent tests. For any classifier $H(z)$ we have used a leave-one-out cross-validation to choose Δ. Instead of using the number of correctly classified samples we use a measure that depends continuously on the threshold Δ. Define

$$\hat{\xi} = -\frac{1}{n} \sum_{i=1}^{n} H(x_i; x_i) \quad \text{and} \quad \hat{\tilde{\xi}} = \frac{1}{n} \sum_{i=1}^{m} H(y_i; y_i),$$

where $H(z; x_i)$ is the classifier constructed from the reduced sample with x_i left out and $H(z; y_i)$ defined similarly. Also let $\hat{\tau}^2$ be the empirical variance of the terms that enters $\hat{\xi}$ and $\hat{\tilde{\xi}}$. We then use $\Phi(\hat{\xi}/\hat{\tau})$ and $\Phi(\hat{\tilde{\xi}}/\hat{\tau})$ to choose Δ. Since we often see strong negative correlation between ξ and $\tilde{\xi}$, we have opted against using the average of the two terms for selecting Δ. Instead we use

$$\arg\max_{\Delta} \min\{\Phi(\hat{\xi}/\hat{\tau}), \Phi(\hat{\tilde{\xi}}/\hat{\tau})\}.$$

For the LOUI classifier it is easy to see that $\hat{\xi} = \hat{\tilde{\xi}}$ so that it is immaterial how the two terms are combined to choose Δ. We compare the above cross-validation choice with an optimal oracle selected threshold based on the true mean values, where we maximize $\min\{\Phi(\xi/\tau), \Phi(\tilde{\xi}/\tau)\}$ in the same range of Δ values as in the cross-validation approach.

The numbers in Table 2 are based on 1000 simulated data sets for each setting. It is clear from the table that the independence classifier D has an unacceptable large bias, even for the case of the optimal threshold. The bias for each of the LOUI, BAI and EasyEnsemble classifiers is very small, favouring the minority group in the fixed threshold and optimal threshold cases, and favouring the majority group in the cross-validation case. The EasyEnsemble classifier has the smallest bias, but at the same time also the smallest probability of correct classification for both groups, making it less optimal than the BAI and LOUI classifiers. The LOUI classifier typically has a slightly larger probability of correct classification as compared to the BAI classifier. However, this comes at the cost of including many more variables in the classfier. The EasyEnsemble classifier includes even more variables than the LOUI classifier.

Generally, the fixed threshold and the cross-validation threshold gives approximately the same probability of correct classification, but with the use of fewer variables for the cross-validation approach. Also, the cross-validation approach typically lowers the negative correlation between ξ/τ and $\tilde{\xi}/\tau$. For the BAI and LOUI classifiers the optimal threshold gives rise to a fairly large positive correlation. The reason for this is that in many instances the threshold will be chosen close to where the two curves for ξ/τ and $\tilde{\xi}/\tau$, as a function of Δ, intersects, so that the two values are almost identical.

In general, the BAI classifier is our preferred method since the bias is small, it has a comparable good probability of correct classification, and it uses only a small number of the variables for constructing the classifier.

5.1 Breast Cancer Data

We illustrate the imbalance bias problem with the breast cancer data from Sotiriou et al. (2003). There are 99 women in the study divided into two groups according to their estrogen receptor status. The ER+ group (65 women) are those women where the cancer has receptors for estrogen, and the ER- group (34 women) are those without receptors. In the original data there are 7650 variables, but we use here only the subset with $p = 4327$ variables measured in all 99 samples. One hundred times we split the data into a training set and a test set, the latter consisting of 20 randomly chosen observations from each group. The training set thus has 45 women in the ER+ group and 14 in the ER- goup, an imbalance ratio around 3. The threshold in the different classifiers is chosen through leave-one-out cross-validation, where the range considered corresponds to an expected number of false positives out of 4327 variables to be between 1 and 30.

In Table 3 we compare BAI, LOUI and EasyEnsemble to the thresholded independence classifier. The table gives the percentage of correctly classified samples, both when evaluated on the training set and on the test set. As expected, the independence classifier shows no bias on the training set, but has a considerable bias when evaluated on the test set. This bias is removed for all three alternatives BAI, LOUI and EasyEnsemble. The bias correction has the consequence that on the training set BAI, LOUI and EasyEnsemble perform best on the minority group. BAI obtains the same performance as LOUI and EasyEnsemble using much less variables, roughly one half of the variables used in LOUI and one third of the variables used in EasyEnsemble. It seems slightly astonishing for this data set, that although a large number of variables seem to be true positives, the classification error is still around 16%.

Table 2: Comparison of the classifiers D, LOUI, BAI and EasyEnsemble for various values of p, n and m based on 1000 simulated data sets. There are $k = 20$ differential expressed variables with $\delta = 1$. The *Fixed* columns have the threshold fixed at $\Delta = 2.5$ when $p = 1000$ and $\Delta = 3$ when $p = 10000$. In the CV columns the threshold is selected by leave-one-out cross-validation for each data set, while *Opt* denotes the optimal threshold calculated from the true parameters.

	D			LOUI			BAI			EasyEnsemble		
p n m	fixed	CV	opt	Fixed	CV	Opt	Fixed	CV	Opt	Fixed	CV	Opt
10^3 30 10												
$E(\xi/\tau)$	2.00	1.55	1.56	1.19	1.20	1.21	1.14	1.14	1.19	1.13	1.10	1.12
$E(\tilde{\xi}/\tau)$	0.32	0.58	0.81	1.25	1.14	1.27	1.18	1.14	1.25	1.14	1.08	1.16
$\mathrm{Std}(\xi/\tau)$	0.23	0.38	0.32	0.29	0.28	0.21	0.29	0.28	0.22	0.26	0.27	0.23
$\mathrm{Cor}(\xi/\tau, \tilde{\xi}/\tau)$	-0.09	-0.10	0.08	-0.28	-0.15	0.42	0.06	-0.13	0.44	-0.17	-0.12	0.06
$E(N)$	28.5	12.3	10.6	68.0	51.7	55.1	28.5	22.6	22.2	165.8	121.7	144.6
$\mathrm{Std}(N)$	4.7	9.1	5.2	7.3	34.0	27.0	4.7	14.1	11.6	11.4	74.3	65.1
10^3 50 10												
$E(\xi/\tau)$	2.28	1.76	1.77	1.30	1.28	1.34	1.24	1.22	1.31	1.20	1.17	1.21
$E(\tilde{\xi}/\tau)$	0.31	0.65	0.87	1.41	1.28	1.41	1.35	1.30	1.39	1.24	1.20	1.26
$\mathrm{Std}(\xi/\tau)$	0.22	0.36	0.30	0.29	0.24	0.21	0.29	0.25	0.21	0.26	0.25	0.23
$\mathrm{Cor}(\xi/\tau, \tilde{\xi}/\tau)$	-0.12	-0.08	0.11	-0.33	-0.11	0.43	0.10	-0.13	0.48	-0.23	-0.12	0.03
$E(N)$	27.9	11.7	10.9	62.3	38.0	50.3	27.9	19.3	22.1	254.0	175.4	226.5
$\mathrm{Std}(N)$	4.4	7.1	4.3	7.0	28.8	24.1	4.4	12.4	10.6	13.6	110.0	93.1
10^4 30 10												
$E(\xi/\tau)$	2.35	1.03	0.99	0.60	0.68	0.64	0.55	0.58	0.63	0.53	0.50	0.52
$E(\tilde{\xi}/\tau)$	-1.22	-0.02	0.21	0.63	0.53	0.68	0.58	0.54	0.68	0.52	0.46	0.52
$\mathrm{Std}(\xi/\tau)$	0.23	0.41	0.35	0.32	0.34	0.24	0.31	0.31	0.24	0.26	0.29	0.23
$\mathrm{Cor}(\xi/\tau,\tilde{\xi}/\tau)$	-0.35	-0.21	0.15	-0.56	-0.21	0.58	-0.15	-0.09	0.58	-0.49	-0.14	0.17
$E(N)$	55.7	6.3	4.5	193.0	63.8	60.1	55.7	17.6	15.5	639.1	177.3	252.7
$\mathrm{Std}(N)$	7.2	7.1	3.2	13.5	50.3	40.8	7.2	14.3	11.5	25.1	146.1	148.6
10^4 50 10												
$E(\xi/\tau)$	2.77	1.29	1.27	0.70	0.84	0.80	0.64	0.72	0.77	0.59	0.59	0.61
$E(\tilde{\xi}/\tau)$	-1.41	0.05	0.26	0.79	0.70	0.86	0.72	0.73	0.85	0.60	0.59	0.64
$\mathrm{Std}(\xi/\tau)$	0.24	0.42	0.35	0.32	0.31	0.25	0.30	0.30	0.25	0.28	0.30	0.25
$\mathrm{Cor}(\xi/\tau, \tilde{\xi}/\tau)$	-0.32	-0.26	0.11	-0.44	-0.18	0.52	-0.02	-0.14	0.52	-0.44	-0.20	0.05
$E(N)$	48.8	6.1	5.0	157.9	54.2	54.4	48.8	18.2	16.0	1066.5	335.2	480.3
$\mathrm{Std}(N)$	6.7	5.5	3.0	12.0	44.3	36.4	6.7	14.3	11.2	30.7	250.7	277.4

Table 3: Comparison of the thresholded independence classifier, BAI, LOUI and EasyEnsemble on the Breast Cancer data from Sotiriou et al. (2003). The data are randomly divided into a training set with $n = 45$ and $m = 14$ observations in the two groups ER+ and ER-, and a test set with 20 observations in each group. Numbers in the table are based on 100 random splits. The row N gives the number of variables included in the classifier and the remaining entries are percentage correctly classified samples.

	D		LOUI		BAI		EasyEnsemble	
Variable	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Training ER+	94.8	2.2	89.7	3.7	89.3	3.9	89.6	3.3
Training ER-	92.1	6.2	94.4	5.1	94.4	5.0	94.7	4.6
Test ER+	90.5	5.5	84.1	7.2	83.9	7.1	84.6	6.9
Test ER-	75.5	8.2	83.3	6.2	83.8	5.9	83.0	6.0
N	229	127	409	198	232	121	602	270

6 Correlated data: BA-ROAD and LOU-ROAD

In many high dimensional settings the variables will be correlated, and classifiers build on the independence classifier will be suboptimal. The Fisher classifier based on an estimate of the inverse covariance matrix is not directly applicable when $p \gg n$. As an alternative Fan et al. (2012) suggested the *Regularized Optimal Affine Discriminant* (ROAD) classifier based on

$$R(z) = \sum_{j=1}^{p} r_j \big[z_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j)\big],$$

where

$$r = \underset{(\bar{y}-\bar{x})^{\mathsf{T}} r = 1, |r|_1 \le c}{\arg\min} \quad r^{\mathsf{T}} \hat{\Sigma} r, \tag{6.1}$$

with $\hat{\Sigma}$ the $p \times p$ estimated covariance matrix, and with the tuning parameter c chosen by cross-validation. Fan et al. (2012) introduced an efficient algorithm for calculating r, and simulations with $n = m$ show that ROAD performs better for correlated data as compared to a number of alternative classifiers including the independence classifier. However, as seen from the first two columns of table 4, in the inbalanced case the ROAD classifier can have an appreciable bias. Inspired by the BAI and the LOUI corrections to the independence classifier, we propose the following adjustments to the ROAD classifier. First define

$$B_{0,R}(z) = \sum_{j=1}^{p} r_j \Big[z_j - \frac{1}{2}(\bar{x}_j + \bar{y}_j) + \frac{\rho}{2}(\bar{y}_j - \bar{x}_j)\Big],$$

$$\bar{\epsilon}_R = \frac{1}{2}\Big[\frac{1}{n}\sum_{i=1}^{n} B_{0,R}(x_i; x_i) + \frac{1}{m}\sum_{i=1}^{m} B_{0,R}(y_i; y_i)\Big],$$

where $B_{0,R}(x_i; x_i)$ and $B_{0,R}(y_i; y_i)$ are defined from $B_{0,R}$ in the same way as $B_0(x_i; x_i)$ and $B_0(y_i; y_i)$ are defined from B_0, that is, $B_{0,R}$ is constructed from a reduced sample with one observation left out and then evaluated on the excluded observation. The BA-ROAD classifier is next defined as

$$B_R(z) = B_{0R}(z) - \bar{\epsilon}_R.$$

In a similar spirit we define the LOU-ROAD classifier as

$$L_R(z) = \frac{1}{2}\sum_{j=1}^{p} \Big[\frac{1}{n}\sum_{i=1}^{n} r_j(x_i)(z_j - x_{ij}) + \frac{1}{m}\sum_{i=1}^{m} r_j(y_i)(z_j - y_{ij})\Big],$$

where $r(x_i)$ and $r(y_i)$ are calculated as in (6.1) based on the reduced sample with either x_i or y_i left out.

For each of the above classifiers the probability of correct classification is evaluated through ξ, $\tilde{\xi}$ and τ^2 as in (2.2). Here ξ is minus the value of the classifier evaluated at μ, and $\tilde{\xi}$ is the value at $\mu + \delta\sigma$. For both of R and B_R we have $\tau^2 = \sum_{j=1}^{p} \sigma_j^2 r_j^2$, and for L_R the formula becomes

$$\tau^2 = \frac{1}{4}\sum_{j=1}^{p} \sigma_j^2 \Big[\frac{1}{n}\sum_{i=1}^{n} r_j(x_i) + \frac{1}{m}\sum_{i=1}^{m} r_j(y_i)\Big]^2.$$

We evaluate BA-ROAD and LOU-ROAD via a set of simulations. For comparison we include the EasyEnsemble undersampling classifier built on top of ROAD, that is, the classifier (5.1) with D replaced by R. In each simulation the value of c in (6.1) is determined by five-fold cross-validation for each of the classifiers. Also, we include the BAI independence classifier where the threshold Δ is chosen by five-fold cross-validation searching over a region with 5 to 30 expected false positives. We consider the setting with $n = 30$, $m = 10$ and $p = 1000$ variables of which the first 20 variables have differential expression 1, the remaining variables having no differential expression. The numbers in Table 4 are based on 100 simulated values. We consider three models for the covariance matrix Σ:

$$\begin{aligned} &\text{Model 1:} \quad \Sigma_{ii} = 1, \;\; \Sigma_{ij} = 0.2, \; i \neq j, \\ &\text{Model 2:} \quad \Sigma_{ij} = 0.8^{|i-j|}, \\ &\text{Model 3:} \quad \Sigma = \mathrm{Cor}\Big(\hat{\Sigma}_p + \sqrt{\frac{\log(p)}{n+m}} I_p\Big), \end{aligned}$$

where $\hat{\Sigma}_p$ is the empirical variance based on the data in Golub et al. (1999), I_p is the identity matrix and Cor is the function that transforms a variance matrix to a correlation matrix ($\hat{\Sigma}_p$ has been obtained by choosing p consequtive variables where the distribution of the correlations resembles the distribution for all variables).

First of all, Table 4 shows that ROAD itself has a considerable bias in the imbalanced case. The bias is almost eliminated with the use of BA-ROAD, LOU-ROAD or the EasyEnsemble-ROAD classifier. Generally, the performance of BA-ROAD is comparable to that of ROAD in terms of the number of variables included in the classifier. LOU-ROAD and EasyEnsemble-ROAD perform slightly better on average, but at the cost of including many more variables than BA-ROAD. In terms

Table 4: Comparison of ROAD, BA-ROAD, LOU-ROAD, EasyEnsemble-ROAD (EE-ROAD) and the BAI independence classifier for the case $n = 30$, $m = 10$ and $p = 1000$ variables of which the first $k = 20$ have differential expression $\delta = 1$. Values are based on 100 simulated data sets. The variable N is the number of variables included in the classifier and *Cor* is the correlation between ξ/τ and $\tilde{\xi}/\tau$.

		ROAD		LOU-ROAD		BA-ROAD		EE-ROAD		BAI	
Model	Variable	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	ξ/τ	1.28	0.33	1.06	0.26	0.92	0.27	1.16	0.28	1.06	0.52
	$\tilde{\xi}/\tau$	0.40	0.25	1.05	0.28	0.84	0.29	1.12	0.31	1.34	0.55
	N	24	21	138	65	35	20	147	63	15	12
	Cor	−0.22		−0.10		−0.22		−0.01		−0.00	
2	ξ/τ	0.97	0.27	0.63	0.32	0.57	0.28	0.77	0.37	1.04	0.73
	$\tilde{\xi}/\tau$	0.22	0.31	0.66	0.35	0.58	0.32	0.74	0.35	1.13	0.52
	N	10	11	49	57	16	18	69	71	15	9
	Cor	−0.26		−0.33		−0.22		−0.11		0.10	
3	ξ/τ	1.45	0.29	1.22	0.21	1.13	0.25	1.23	0.23	1.13	0.48
	$\tilde{\xi}/\tau$	0.74	0.25	1.28	0.29	1.12	0.28	1.21	0.33	1.21	0.61
	N	28	18	116	64	34	17	105	64	16	18
	Cor	0.19		0.04		0.06		−0.05		−0.49	

of mean values the BAI independence classifier performs as good as the ROAD based classifiers. However, it has a somewhat larger spread. A clear message from this small simulation study is that the bias of the ROAD classifier can be handled by using the classifiers we propose in this paper.

7 Conclusion

In this paper we have analyzed the independence classifier in order to study the bias originating from imbalanced data sets. It has been found that a correction for bias is needed also for minor imbalances when considering classification in the high dimensional case. The thresholded independence classifier favours the majority group, and in the high dimensional case this can lead to classifying practically all observations to the majority group. The two suggested classifiers virtually remove the bias and have almost the same error rate.

The BAI classifier performs better in the sense that it obtains the same error rate as the LOUI classifier using much fewer variables. This can be of some practical value when implementing a classifier as a diagnostic tool in a medical setting. Simulations reveal that both classifiers have a slightly lower error rate than a variant of multiple undersampling, which is currently considered among the best methods for correcting imbalance (Blagus and Lusa, 2013). Multiple undersampling uses a high number of variables which also makes it less attractive.

For the case of correlated variables the ROAD classifier turns out to have a bias in the imbalanced case. We have suggested a modification of the ROAD classifier that removes the bias, and simulations show a good performance of this classifier. Overall, our way of correcting for bias seems of value for a broad range of linear classifiers.

References

Blagus, R. and L. Lusa (2013). Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics 14*(106).

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association 106*(496), 1566–1577.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: Synthetic miniority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Donoho, D. and J. Jin (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical transactions of the royal society A 367*, 4449–4470.

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society Series B: Statistical Methodology 74*(4), 745–771.

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

Jensen, J. L. (2006). Maximum likelihood classifiers in microarray studies. Research Report 474, University of Aarhus.

Lin, S.-C., Y.-c. I. Chang, and W.-N. Yang (2009). Meta-learning for imbalanced data and classification ensemble in binary classification. *Neurocomput. 73*(1-3), 484–494.

Liu, X.-Y., J. Wu, and Z.-H. Zhou (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transations on Sysetems, Man, and Cybernetics Part B: Cybernetics 39*(3), 539–550.

Sotiriou, C., S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences 100*(18), 10393–10398.

Yang, P., P. D. Yoo, J. Fernando, B. B. Zhou, Z. Zhang, and A. Y. Zomaya (2014). Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. *IEEE Transations on Cybernetics 44*(3), 445–455.