Computational Statistics and Data Analysis 98 (2016) 46–59

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

High dimensional classifiers in the imbalanced case

Britta Anker Bak, Jens Ledet Jensen *

Aarhus University, Department of Mathematics, Ny Munkegade 118, 8000 Aarhus C, Denmark

ARTICLE INFO

Article history:
Received 22 June 2015
Received in revised form 14 December 2015
Accepted 17 December 2015
Available online 28 December 2015

Keywords:
High dimension
Imbalance
Classification

ABSTRACT

A binary classification problem is imbalanced when the number of samples from the two groups differs. For the high dimensional case, where the number of variables is much larger than the number of samples, imbalance leads to a bias in the classification. The independence classifier is studied theoretically and based on the analysis two new classifiers are suggested 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 the ROAD classifier is considered and a suggestion is given for how to modify the classifier 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, in this line of research the imbalance problem appearing when the sample sizes of the groups differ has not been of major concern. In real life experiments, on the other hand, imbalanced data sets are the norm rather than the exception, see e.g. Shipp et al. (2002) and Ramaswamy et al. (2002), where in the former the sample sizes are 64 and 12 for the two groups and in the latter the sample sizes are 58 and 19. 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 where the number of variables can be in the order of tens of thousands, while the sample sizes are in the order of tens. To briefly illustrate this, Table 1 gives the mean and standard deviation of the probability of correct classification in a few instances for the thresholded independence classifier. The classifier is described in detail in Section 2 and details for the numbers in the table are given in the caption of the table. It is clearly seen that even small imbalances can lead to rather large biases in the probability of correct classification, showing that the imbalance problem should not be ignored.

The imbalance problem has 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) proposed 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

* Corresponding author. Tel.: +45 87155797.
E-mail address: jlj@math.au.dk (J.L. Jensen).

http://dx.doi.org/10.1016/j.csda.2015.12.009
0167-9473/

Table 1
Average percentage of correct classification and the standard deviation based on 1000 simulated data sets for the thresholded independence classifier. There are two groups and the group entry of the table corresponds to the group label of a new observation. Each observation has 1000 variables which are independent and standard normally distributed. Ten of the variables have a mean difference of 1 between the two groups and for the remaining variables the means are the same in the two groups. The threshold parameter has been chosen such that the power of the t-test is 0.8 for detecting a mean difference of 1. The sample sizes are n for group 1 and m for group 2.

n	m	Group 1		Group 2	
		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

imbalanced classification problems well, although the oversampling methods do have limitations as we describe in the Appendix. 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 in a non-asymptotic setting (see Section 4 for further discussion of the setting). Building on the analysis, we 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. We study in particular the thresholded independence classifier. Without thresholding the classifier has been studied with correlated data in Bickel and Levina (2004) and with thresholding in Fan and Fan (2008) and Bak et al. (2015).

After detecting the origin of the bias problem for the independence classifier 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.

We end this introduction with a presentation of a data set and the result from using the thresholded independence classifier. We return to the example in Section 5.1 to see the benefits of using a bias corrected version of the classifier.

Example 1. Sotiriou et al. (2003) described a breast cancer study with 99 women divided into two groups according to their oestrogen receptor status. The ER+ group (65 women) are those women where the cancer has receptors for oestrogen, and the ER− group (34 women) are those without receptors. Out of 7650 variables there are 4404 measured in all 99 samples, and we use these variables for our analysis. A random split of the data into a training set and test set is made where the training set has 45 women from the ER+ group and 14 from the ER− group, and the test set has 20 women from each group. The threshold of the classifier is found using leave-one-out cross-validation searching over the range corresponding to having 1–30 expected false positives out of 4404 variables. The percentage correctly classified samples from the test set is found, and the whole procedure starting with a random split is repeated one hundred times.

For the ER+ group the percentage correctly classified is 90.5 with a standard deviation of 5.5, and for the ER− group 75.4% is correctly classified with a standard deviation of 8.0. As in Table 1 the majority group is favoured and the classifier has a marked bias. We return to this example in Section 5.1.

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 $x_{ij} \sim N(\mu_{1j}, \sigma_j^2)$ and $y_{ij} \sim N(\mu_{2j}, \sigma_j^2)$. 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 $\rho_{n,m} = (n-m)/(n+m)$, and let $f = n+m-2$ be the degrees of freedom for the joint sample variance. An important parameter is the (scaled) differential expression for variable j defined as $\delta_j = (\mu_{2j} - \mu_{1j})/\sigma_j$.

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

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