When is Undersampling Effective in Unbalanced Classification Tasks?

Conference Paper · September 2015

DOI: 10.1007/978-3-319-23528-8_13

CITATIONS
110

READS
1,980

3 authors:

Andrea Dal Pozzolo
Université Libre de Bruxelles
8 PUBLICATIONS **1,787** CITATIONS
SEE PROFILE

Olivier Caelen
Worldline
63 PUBLICATIONS **3,469** CITATIONS
SEE PROFILE

Gianluca Bontempi
Université Libre de Bruxelles
376 PUBLICATIONS **18,038** CITATIONS
SEE PROFILE

When is undersampling effective in unbalanced classification tasks?

Andrea Dal Pozzolo[1], Olivier Caelen[2], and Gianluca Bontempi[1,3]

[1] Machine Learning Group (MLG), Computer Science Department, Faculty of Sciences ULB, Université Libre de Bruxelles, Brussels, Belgium
[2] Fraud Risk Management Analytics, Worldline, Brussels, Belgium
[3] Interuniversity Institute of Bioinformatics in Brussels (IB)[2], Brussels, Belgium

Abstract. A well-known rule of thumb in unbalanced classification recommends the rebalancing (typically by resampling) of the classes before proceeding with the learning of the classifier. Though this seems to work for the majority of cases, no detailed analysis exists about the impact of undersampling on the accuracy of the final classifier. This paper aims to fill this gap by proposing an integrated analysis of the two elements which have the largest impact on the effectiveness of an undersampling strategy: the increase of the variance due to the reduction of the number of samples and the warping of the posterior distribution due to the change of priori probabilities. In particular we will propose a theoretical analysis specifying under which conditions undersampling is recommended and expected to be effective. It emerges that the impact of undersampling depends on the number of samples, the variance of the classifier, the degree of imbalance and more specifically on the value of the posterior probability. This makes difficult to predict the average effectiveness of an undersampling strategy since its benefits depend on the distribution of the testing points. Results from several synthetic and real-world unbalanced datasets support and validate our findings.

Keywords: Undersampling; Ranking; Class Overlap, Unbalanced classification

1 Introduction

In several binary classification problems, the two classes are not equally represented in the dataset. For example, in fraud detection, fraudulent transactions are normally outnumbered by genuine ones [5]. When one class is underrepresented in a dataset, the data is said to be unbalanced. In such problems, typically, the minority class is the class of interest. Having few instances of one class means that the learning algorithm is often unable to generalize the behavior of the minority class well, hence the algorithm performs poorly in terms of predictive accuracy [14].

When the data is unbalanced, standard machine learning algorithms that maximise overall accuracy tend to classify all observations as majority class instances. This translates into poor accuracy on the minority class (low recall),

which is typically the class of interest. Degradation of classification performance is not only related to a small number of examples in the minority class in comparison to the number of examples in the majority classes (expressed by the class imbalance ratio), but also to the minority class decomposition into small sub-parts [19] (also known in the literature as *small disjuncts* [15]) and to the overlap between the two classes [16] [3] [11] [10]. In these studies it emerges that performance degradation is strongly caused by the presence of both unbalanced class distributions and a high degree of class overlap. Additionally, in unbalanced classification tasks, the performance of a classifier is also affected by the presence of noisy examples [20] [2].

One possible way to deal with this issue is to adjust the algorithms themselves [14] [23] [7]. Here we will consider instead a data-level strategy known as *undersampling* [13]. Undersampling consists in down-sizing the majority class by removing observations at random until the dataset is balanced. In an unbalanced problem, it is often realistic to assume that many observations of the majority class are redundant and that by removing some of them at random the data distribution will not change significantly. However the risk of removing relevant observations from the dataset is still present, since the removal is performed in an unsupervised manner. In practice, sampling methods are often used to balance datasets with skewed class distributions because several classifiers have empirically shown better performance when trained on balanced dataset [22] [9]. However, these studies do not imply that classifiers cannot learn from unbalanced datasets. For instance, other studies have also shown that some classifiers do not improve their performances when the training dataset is balanced using sampling techniques [4] [14]. As a result, the only way to know if sampling helps the learning process is to run some simulations. Despite the popularity of undersampling, we have to remark that there is not yet a theoretical framework explaining how it can affect the accuracy of the learning process.

In this paper we aim to analyse the role of the two side-effects of undersampling on the final accuracy. The first side-effect is that, by removing majority class instances, we perturb the a priori probability of the training set and we induce a warping in the posterior distribution [18, 8]. The second is that the number of samples available for training is reduced with an evident consequence in terms of accuracy of the resulting classifier. We study the interaction between these two effects of undersampling and we analyse their impact on the final ranking of posterior probabilities. In particular we show under which conditions an under sampling strategy is recommended and expected to be effective in terms of final classification accuracy.

2 The warping effect of undersampling on the posterior probability

Let us consider a binary classification task $f : R^n \to \{0, 1\}$, where $\mathbf{X} \in R^n$ is the input and $\mathbf{Y} \in \{0, 1\}$ the output domain. In the following we will also use the label negative (resp. positive) to denote the label 0 (resp. 1). Suppose that the

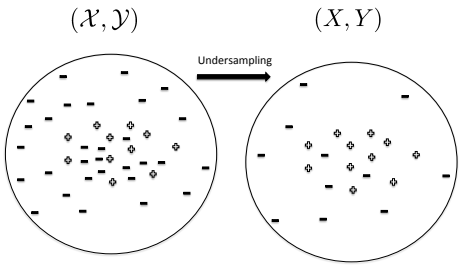


Fig. 1. Undersampling: remove majority class observations until we have the same number of instances in the two classes.

training set $(\mathcal{X}, \mathcal{Y})$ of size N is unbalanced (i.e. the number N^+ of positive cases is small compared to the number N^- of negative ones) and that rebalancing is performed by undersampling. Let $(X, Y) \subset (\mathcal{X}, \mathcal{Y})$ be the balanced sample of $(\mathcal{X}, \mathcal{Y})$ which contains a subset of the negatives in $(\mathcal{X}, \mathcal{Y})$.

Let us introduce a random binary selection variable s associated to each sample in $(\mathcal{X}, \mathcal{Y})$, which takes the value 1 if the point is in (X, Y) and 0 otherwise. We now derive how the posterior probability of a model learned on a balanced subset relates to the one learned on the original unbalanced dataset, on the basis of the reasoning presented in [17]. Let us assume that the selection variable s is independent of the input x given the class y (*class-dependent selection*):

$$p(s|y, x) = p(s|y) \tag{1}$$

where $p(s = 1|y, x)$ is the probability that a point (x, y) is included in the balanced training sample. Note that the undersampling mechanism has no impact on the class-conditional distribution but that it perturbs the prior probability (i.e. $p(y|s = 1) \neq p(y)$).

Let the sign $+$ denote $y = 1$ and $-$ denote $y = 0$, e.g. $p(+, x) = p(y = 1, x)$ and $p(-, x) = p(y = 0, x)$. From Bayes' rule we can write:

$$p(+|x, s = 1) = \frac{p(s = 1|+, x)p(+|x)}{p(s = 1|+, x)p(+|x) + p(s = 1|-, x)p(-|x)} \tag{2}$$

Using condition (1) in (2) we obtain:

$$p(+|x, s = 1) = \frac{p(s = 1|+)p(+|x)}{p(s = 1|+)p(+|x) + p(s = 1|-)p(-|x)} \tag{3}$$

With undersampling we have:

$$p(s = 1|+) = 1 \tag{4}$$

and

$$\frac{p(+)}{p(-)} \leq p(s = 1|-) < 1 \tag{5}$$

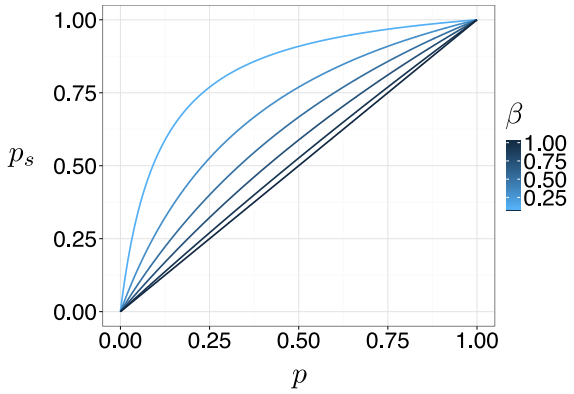


Fig. 2. p and p_s at different β. When β is low, undersampling is strong, which means it is removing a lot of negatives, while for high values the removal is less strong. Low values of β leads to a more balanced problem.

Note that if we set $p(s = 1|-) = \frac{p(+)}{p(-)}$, we obtain a balanced dataset where the number of positive and negative instances is the same. At the same time, if we set $p(s = 1|-) = 1$, no negative instances are removed and no undersampling takes place. Using (4), we can rewrite (3) as

$$p_s = p(+|x, s = 1) = \frac{p(+|x)}{p(+|x) + p(s = 1|-)p(-|x)} = \frac{p}{p + \beta(1 - p)} \tag{6}$$

where $\beta = p(s = 1|-)$ is the probability of selecting a negative instance with undersampling, $p = p(+|x)$ is the true posterior probability of class $+$ in the original dataset, and $p_s = p(+|x, s = 1)$ is the true posterior probability of class $+$ after sampling. Equation (6) quantifies the amount of warping of the posterior probability due to undersampling. From it, we can derive p as a function of p_s:

$$p = \frac{\beta p_s}{\beta p_s - p_s + 1} \tag{7}$$

The relation between p and p_s (parametric in β) is illustrated in Figure 2. The top curve of Figure 2 refers to the complete balancing which corresponds to $\beta = \frac{p(+)}{p(-)} \approx \frac{N^+}{N^-}$, assuming that $\frac{N^+}{N^-}$ provides an accurate estimation of the ratio of the prior probabilities.

Figure 3 illustrates the warping effect for two univariate ($n = 1$) classification tasks. In both tasks the two classes are normally distributed ($\mathcal{X}_- \sim N(0, \sigma)$ and $\mathcal{X}_+ \sim N(\mu, \sigma)$), $\sigma = 3$ and $p(+) = 0.1$ but the degree of separability is different (on the left large overlap for $\mu = 3$ and on the right small overlap for $\mu = 15$). It is easy to remark that the warping effect is larger in the low separable case.

As a final remark, consider that when $\beta = \frac{N^+}{N^-}$, the warping due to undersampling maps two close and low values of p into two values p_s with a larger

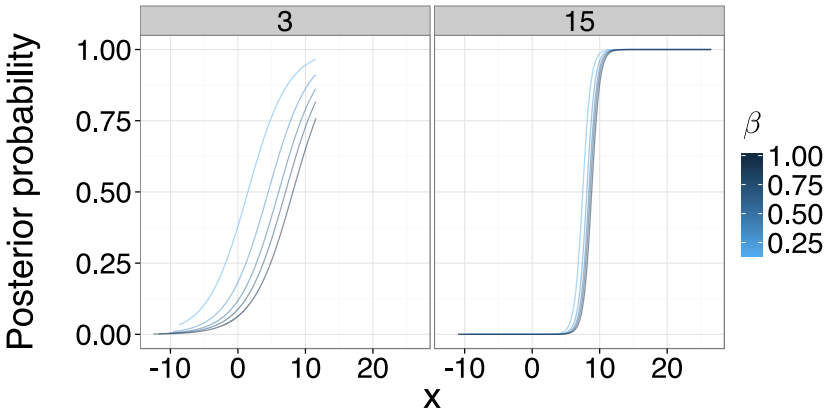


Fig. 3. Posterior probability as a function of β for two univariate binary classification tasks with norm class conditional densities $\mathcal{X}_- \sim N(0, \sigma)$ and $\mathcal{X}_+ \sim N(\mu, \sigma)$ (on the left $\mu = 3$ and on the right $\mu = 15$, in both examples $\sigma = 3$). Note that p corresponds to $\beta = 1$ and p_s to $\beta < 1$.

distance. The opposite occurs for high values of p. In Section 3 we will show how this has an impact on the ranking returned by estimations of p and p_s.

3 The interaction between warping and variance of the estimator

The previous section discussed the first consequence of under sampling, i.e. the transformation of the original conditional distribution p into a warped conditional distribution p_s according to equation (6). The second consequence of undersampling is the reduction of the training set size which inevitably leads to an increase of the variance of the classifier. This section discusses how these two effects interact and their impact on the final accuracy of the classifier, by focusing in particular on the accuracy of the ranking of the minority class (typically the class of interest).

Undersampling transforms the original classification task (i.e. estimating the conditional distribution p) into a new classification task (i.e. estimating the conditional distribution p_s). In what follows we aim to assess whether and when under sampling has a beneficial effect by changing the target of the estimation problem.

Let us denote by $\hat{p}$ (resp. $\hat{p}_s$) the estimation of the conditional probability p (resp. p_s). Assume we have two distinct test points having probabilities $p_1 < p_2$ where $\Delta p = p_2 - p_1$ with $\Delta p > 0$. A correct classification aiming to rank the most probable positive samples should rank p_2 before p_1, since the second test sample has an higher probability of belonging to the positive class. Unfortunately the values p_1 and p_2 are not known and the ranking should rely on the estimated

values $\hat{p}_1$ and $\hat{p}_2$. For the sake of simplicity we will assume here that the estimator of the conditional probability has the same bias and variance in the two test points. This implies $\hat{p}_1 = p_1 + \epsilon_1$ and $\hat{p}_2 = p_2 + \epsilon_2$, where ϵ_1 and ϵ_2 are two realizations of the random variable $\varepsilon \sim N(b, \nu)$ where b and ν are the bias and the variance of the estimator of p. Note that the estimation errors ϵ_1 and ϵ_2 may induce a wrong ranking if $\hat{p}_1 > \hat{p}_2$.

What happens if instead of estimating p we decide to estimate p_s, as in undersampling? Note that because of the monotone transformation (6), $p_1 < p_2 \Rightarrow p_{s,1} < p_{s,2}$. Is the ranking based on the estimations of $p_{s,1}$ and $p_{s,2}$ more accurate than the one based on the estimations of p_1 and p_2?

In order to answer this question let us suppose that also the estimator of p_s is biased but that its variance is larger given the smaller number of samples. Then $\hat{p}_{s,1} = p_{s,1} + \eta_1$ and $\hat{p}_{s,2} = p_{s,2} + \eta_2$, where $\eta \sim N(b_s, \nu_s)$, $\nu_s > \nu$ and $\Delta p_s = p_{s,2} - p_{s,1}$.

Let us now compute the derivative of p_s w.r.t. p. From (6) we have:

$$\frac{dp_s}{dp} = \frac{\beta}{(p + \beta(1-p))^2} \tag{8}$$

corresponding to a concave function. Let λ be the value of p for which $\frac{dp_s}{dp} = 1$:

$$\lambda = \frac{\sqrt{\beta} - \beta}{1 - \beta}$$

It follows that

$$\beta \leq \frac{dp_s}{dp} \leq \frac{1}{\beta} \tag{9}$$

and

$$1 < \frac{dp_s}{dp} < \frac{1}{\beta}, \qquad \text{when } 0 < p < \lambda$$

while

$$\beta < \frac{dp_s}{dp} < 1 \qquad \text{when } \lambda < p < 1.$$

In particular for $p = 0$ we have $dp_s = \frac{1}{\beta} dp$ while for $p = 1$ it holds $dp_s = \beta dp$.

Let us now suppose that the quantity Δp is small enough to have an accurate approximation $\frac{\Delta p_s}{\Delta p} \approx \frac{dp_s}{dp}$. We can define the probability of obtaining a wrong ranking of $\hat{p}_1$ and $\hat{p}_2$ as:

$$\begin{aligned} P(\hat{p}_2 < \hat{p}_1) &= P(p_2 + \epsilon_2 < p_1 + \epsilon_1) \\ &= P(\epsilon_2 - \epsilon_1 < p_1 - p_2) = P(\epsilon_1 - \epsilon_2 > \Delta p) \end{aligned}$$

where $\epsilon_2 - \epsilon_1 \sim N(0, 2\nu)$. By making an hypothesis of normality we have

$$P(\epsilon_1 - \epsilon_2 > \Delta p) = 1 - \Phi\left(\frac{\Delta p}{\sqrt{2\nu}}\right) \tag{10}$$

where Φ is the cumulative distribution function of the standard normal distribution. Similarly, the probability of a ranking error with undersampling is:

$$P(\hat{p}_{s,2} < \hat{p}_{s,1}) = P(\eta_1 - \eta_2 > \Delta p_s)$$

and

$$P(\eta_1 - \eta_2 > \Delta p_s) = 1 - \Phi\left(\frac{\Delta p_s}{\sqrt{2\nu_s}}\right) \tag{11}$$

We can now say that a classifier learned after undersampling has better ranking w.r.t. a classifier learned with unbalanced distribution when

$$P(\epsilon_1 - \epsilon_2 > \Delta p) > P(\eta_1 - \eta_2 > \Delta p_s) \tag{12}$$

or equivalently from (10) and (11) when

$$1 - \Phi\left(\frac{\Delta p}{\sqrt{2\nu}}\right) > 1 - \Phi\left(\frac{\Delta p_s}{\sqrt{2\nu_s}}\right) \Leftrightarrow \Phi\left(\frac{\Delta p}{\sqrt{2\nu}}\right) < \Phi\left(\frac{\Delta p_s}{\sqrt{2\nu_s}}\right)$$

which boils down to

$$\frac{\Delta p}{\sqrt{2\nu}} < \frac{\Delta p_s}{\sqrt{2\nu_s}} \Leftrightarrow \frac{\Delta p_s}{\Delta p} > \sqrt{\frac{\nu_s}{\nu}} > 1 \tag{13}$$

since Φ is monotone non decreasing and we can assume that $\nu_s > \nu$.

Then it follows that undersampling is useful in terms of more accurate ranking when

$$\frac{\beta}{(p + \beta(1-p))^2} > \sqrt{\frac{\nu_s}{\nu}} \tag{14}$$

The value of this inequality depends on several terms: the rate of undersampling β, the ratio of the variances of the two classifiers and the posteriori probability p of the testing point. Also the nonlinearity of the first left-hand term suggests a complex interaction between the involved terms. For instance if we plot the left-hand term of (14) as a function of the posteriori probability p (Figure 4(a)) and of the value β (Figure 4(b)), it appears that most favourable configurations for undersampling occur for the lowest values of the posteriori probability (e.g. non separable or badly separable configurations) and intermediate β (neither too unbalanced nor too balanced). However if we modify β, this has an impact on the size of the training set and consequently on the right-hand term (i.e. variance ratio) too. Also, though the β term can be controlled by the designer, the other two terms vary over the input space. This means that the condition (14) does not necessarily hold for all the test points.

In order to illustrate the complexity of the interaction, let us consider two univariate ($n = 1$) classification tasks where the minority class is normally distributed around zero and the majority class is distributed as a mixture of two gaussians. Figure 5 and 6 show the non separable and separable case, respectively: on the left side we plot the class conditional distributions (thin lines) and the posterior distribution of the minority class (thicker line), while on the right side we show the left and the right term of the inequality (14) (solid: left-hand

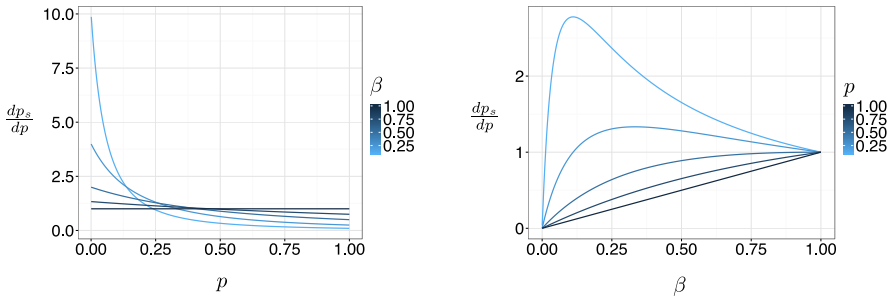
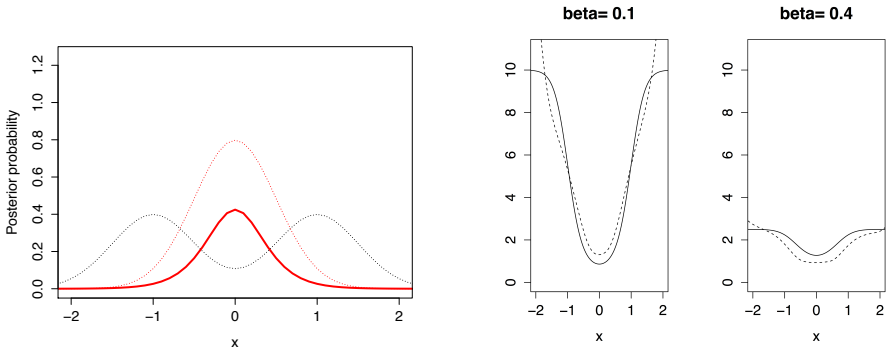


Fig. 4. Left: $\frac{dp_s}{dp}$ as a function of p. Right: $\frac{dp_s}{dp}$ as a function of β

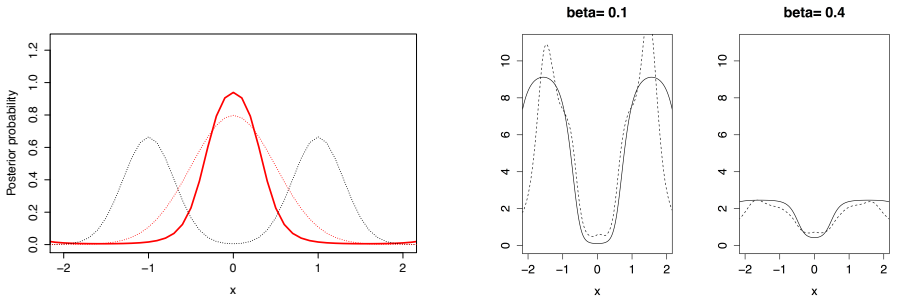
term, dotted: right-hand term). What emerges form the figures is that the least separable regions (i.e. the regions where the posteriori of the minority class is low) are also the regions where undersampling helps more. However, the impact of undersampling on the overall accuracy is difficult to be predicted since the regions where undersampling is beneficial change with the characteristics of the classification task and the rate β of undersampling.



(a) Class conditional distributions (thin lines) and the posterior distribution of the minority class (thicker line).

(b) $\frac{dp_s}{dp}$ (solid lines), $\sqrt{\frac{\nu_s}{\nu}}$ (dotted lines).

Fig. 5. Non separable case. On the right we plot both terms of inequality (14) (solid: left-hand, dotted: right-hand term) for $\beta = 0.1$ and $\beta = 0.4$



(a) Class conditional distributions (thin lines) and the posterior distribution of the minority class (thicker line).

(b) $\frac{dp_s}{dp}$ (solid lines), $\sqrt{\frac{\nu_s}{\nu}}$ (dotted lines).

Fig. 6. Separable case. On the right we plot both terms of inequality (14) (solid: left-hand, dotted: right-hand term) for $\beta = 0.1$ and $\beta = 0.4$

4 Experimental validation

In this section we assess the validity of the condition (14) by performing a number of tests on synthetic and real datasets.

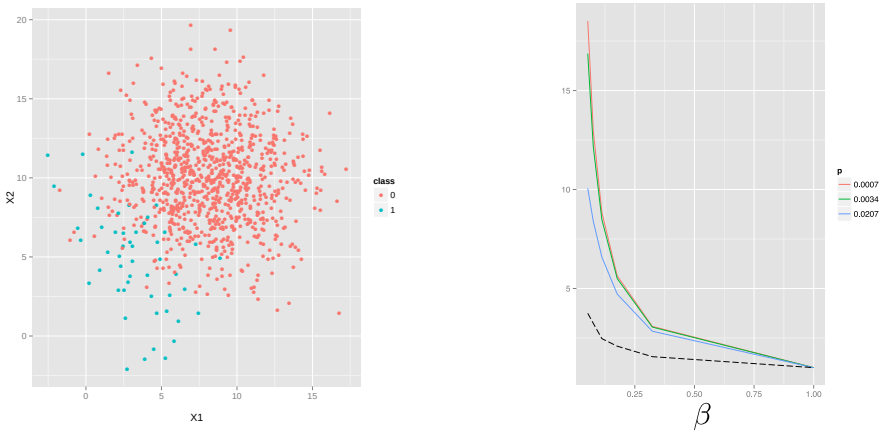
4.1 Synthetic datasets

We simulate two unbalanced tasks (5% and 25% of positive samples) with overlapping classes and generate a testing set and several training sets from the same distribution. Figures 7(a) and Figure 9(a) show the distributions of the testing sets for the two tasks.

In order to compute the variance of $\hat{p}$ and $\hat{p}_s$ in each test point, we generate 1000 times a training set ($N = 1000$) and we estimate the conditional probability on the basis of sample mean and covariance.

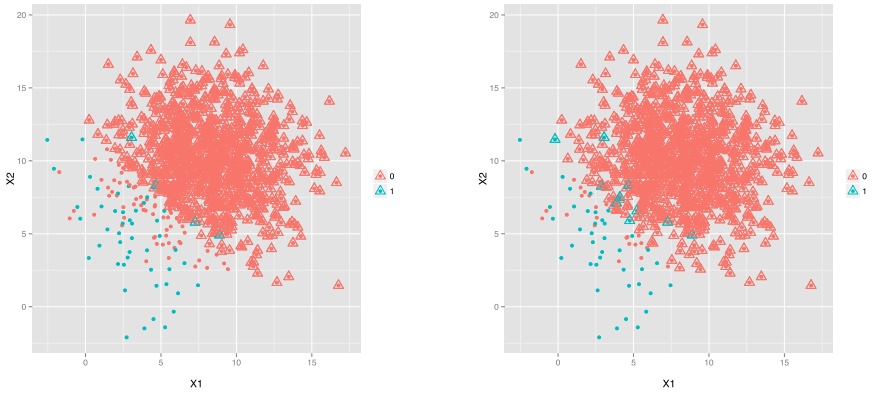
In Figure 7(b) (first task) we plot $\sqrt{\frac{\nu_s}{\nu}}$ (dotted line) and three percentiles $(0.25, 0.5, 0.75)$ of $\frac{dp_s}{dp}$ vs. the rate of undersampling β. It appears that for at least 75% of the testing points, the term $\frac{dp_s}{dp}$ is higher than $\sqrt{\frac{\nu_s}{\nu}}$. In Figure 8(a) the points surrounded with a triangle are those one for which $\frac{dp_s}{dp} > \sqrt{\frac{\nu_s}{\nu}}$ hold when $\beta = 0.053$ (balanced dataset). For such samples we expect that ranking returned by undersampling (i.e. based on $\hat{p}_s$) is better than the one based on the original data (i.e. based on $\hat{p}$). The plot shows that undersampling is beneficial in the region where the majority class is situated, which is also the area where we expect to have low values of p. Figure 8(b) shows also that this region moves towards the minority class when we do undersampling with $\beta = 0.323$ (90% negatives, 10% positives after undersampling).

In order to measure the quality of the rankings based on $\hat{p}_s$ and $\hat{p}$ we compute the Kendall rank correlation of the two estimates with p, which is the true pos-



(a) Synthetic dataset 1 (b) $\sqrt{\frac{\nu_s}{\nu}}$ and $\frac{dp_s}{dp}$ for different β

Fig. 7. Left: distribution of the testing set where the positive samples account for 5% of the total. Right: plot of $\frac{dp_s}{dp}$ percentiles (25^{th}, 50^{th} and 75^{th}) and of $\sqrt{\frac{\nu_s}{\nu}}$ (black dashed).



(a) Undersampling with $\beta = 0.053$ (b) Undersampling with $\beta = 0.323$

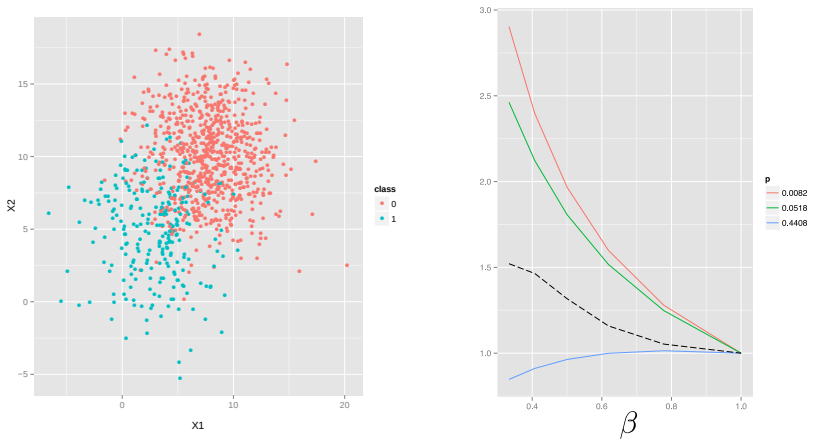
Fig. 8. Regions where undersampling should work. Triangles indicate the testing samples where the condition (14) holds for the dataset in Figure 7.

terior probability of the testing set that defines the correct ordering. In Table 1 we show the ranking correlations of $\hat{p}_s$ (and $\hat{p}$) with p for the samples where the condition (14) (first five rows) holds and where it does not (last five rows). The results indicate that points for which condition (14) is satisfied have indeed better ranking with $\hat{p}_s$ than $\hat{p}$.

Table 1. Classification task in Figure 7: Ranking correlation between the posterior probability $\hat{p}$ ($\hat{p}_s$) and p for different values of β. The value $\mathcal{K}$ ($\mathcal{K}_s$) denotes the Kendall rank correlation without (with) undersampling. The first (last) five lines refer to samples for which the condition (14) is (not) satisfied.

β	$\mathcal{K}$	$\mathcal{K}_s$	$\mathcal{K}_s - \mathcal{K}$	%points satisfying (14)
0.053	0.298	0.749	0.451	88.8
0.076	0.303	0.682	0.379	89.7
0.112	0.315	0.619	0.304	91.2
0.176	0.323	0.555	0.232	92.1
0.323	0.341	0.467	0.126	93.7
0.053	0.749	0.776	0.027	88.8
0.076	0.755	0.773	0.018	89.7
0.112	0.762	0.764	0.001	91.2
0.176	0.767	0.761	-0.007	92.1
0.323	0.768	0.748	-0.020	93.7

We repeated the experiments for the second task having a larger proportion of positives (25%) (dataset 2 in Figure 9(a)). From the Figure 9(b), plotting $\frac{dp_s}{dp}$ and $\sqrt{\frac{\nu_s}{\nu}}$ as a function of β, it appears that only the first two percentiles are over $\sqrt{\frac{\nu_s}{\nu}}$. This means that less points of the testing set satisfy the condition (14). This is confirmed from the results in Table 2 where it appears that the benefit due to undersampling is less significant than for the first task.



(a) Synthetic dataset 2 (b) $\sqrt{\frac{\nu_s}{\nu}}$ and $\frac{dp_s}{dp}$ for different β

Fig. 9. Left: distribution of the testing set where the positive samples account for 25% of the total. Right: plot of $\frac{dp_s}{dp}$ percentiles (25^{th}, 50^{th} and 75^{th}) and of $\sqrt{\frac{\nu_s}{\nu}}$ (black dashed).

Table 2. Classification task in Figure 9: Ranking correlation between the posterior probability $\hat{p}$ ($\hat{p}_s$) and p for different values of β. The value $\mathcal{K}$ ($\mathcal{K}_s$) denotes the Kendall rank correlation without (with) undersampling. The first (last) five lines refer to samples for which the condition (14) is (not) satisfied.

β	$\mathcal{K}$	$\mathcal{K}_s$	$\mathcal{K}_s - \mathcal{K}$	% points statisfying (14)
0.333	0.586	0.789	0.202	66.4
0.407	0.588	0.761	0.172	66.6
0.500	0.605	0.738	0.133	68.1
0.619	0.628	0.715	0.087	70.3
0.778	0.653	0.693	0.040	73
0.333	0.900	0.869	-0.030	66.4
0.407	0.899	0.875	-0.024	66.6
0.500	0.894	0.874	-0.020	68.1
0.619	0.885	0.869	-0.016	70.3
0.778	0.870	0.856	-0.014	73

4.2 Real datasets

In this section we assess the validity of the condition (14) on a number of real unbalanced binary classification tasks obtained by transforming some datasets from the UCI repository [1] (Table 3)[4].

Given the unavailability of the conditional posterior probability function, we first approximate p by fitting a Random Forest over the entire dataset in order to compute the left-hand term of (14). Then we use a boostrap procedure to estimate $\hat{p}$ and apply undersampling to the original dataset to estimate $\hat{p}_s$. We repeat bootstrap and undersampling 100 times to compute the right hand term $\sqrt{\frac{\nu_s}{\nu}}$. This allows us to define the subsets of points for which the condition (14) holds.

Figure 10 reports the difference between Kendall rank correlation of $\hat{p}_s$ and $\hat{p}$, averaged over different levels of undersampling (proportions of majority vs. minority: $90/10, 80/20, 60/40, 50/50$). Higher difference means that $\hat{p}_s$ returns a better ordering than $\hat{p}$ (assuming that the ranking provided by p is correct). The plot distinguishes between samples for which condition (14) is satisfied and not. In general we see that points with a positive difference corresponds to those having the condition satisfied and the opposite for negative differences. These results seem to confirm the experiments with synthetic data, where a better ordering is given by $\hat{p}_s$ when the condition (14) holds.

In Figure 11 we show the ratio of samples in each dataset satisfying condition 14 averaged over all the (β)s. The proportion of points in which undersampling is useful changes heavily with the dataset considered. For example, in the datasets *vehicle, yeast, german* and *pima*, underdamping returns a better

[4] Transformed datasets are available at `http://www.ulb.ac.be/di/map/adalpozz/imbalanced-datasets.zip`

Table 3. Selected datasets from the UCI repository [1]

Datasets	N	N^+	N^-	N^+/N
ecoli	336	35	301	0.10
glass	214	17	197	0.08
letter-a	20000	789	19211	0.04
letter-vowel	20000	3878	16122	0.19
ism	11180	260	10920	0.02
letter	20000	789	19211	0.04
oil	937	41	896	0.04
page	5473	560	4913	0.10
pendigits	10992	1142	9850	0.10
PhosS	11411	613	10798	0.05
satimage	6430	625	5805	0.10
segment	2310	330	1980	0.14
boundary	3505	123	3382	0.04
estate	5322	636	4686	0.12
cam	18916	942	17974	0.05
compustat	13657	520	13137	0.04
covtype	38500	2747	35753	0.07

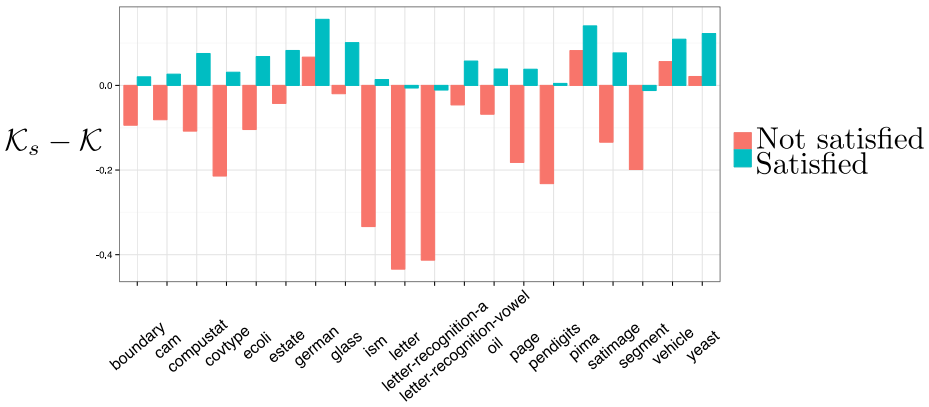


Fig. 10. Difference between the Kendall rank correlation of $\hat{p}_s$ and $\hat{p}$ with p, namely $\mathcal{K}_s$ and $\mathcal{K}$, for points having the conditions (14) satisfied and not. $\mathcal{K}_s$ and $\mathcal{K}$ are calculated as the mean of the correlations over all βs.

ordering for more than 80% of the samples, while the proportion drops to less than 50% in the *page* dataset.

This seems to confirm our intuition that the right amount of undersampling depends on the classification task (e.g. degree of non separability), the learning algorithm and the targeted test set. It follows that there is no reason to believe that undersampling until the two classes are perfectly balanced is the default strategy to adopt.

It is also worthy to remark that the check of the condition (14) is not easy to be done, since it involves the estimation of $\sqrt{\frac{\nu_s}{\nu}}$ (ratio of the the variance of the classifier before and after undersampling) and of $\frac{dp_s}{dp}$, which demands the

knowledge of the true posterior probability p. In practice since p is unknown in real datasets, we can only rely on a data driven approximation of $\frac{dp_s}{dp}$. Also the estimation of $\sqrt{\frac{\nu_s}{\nu}}$ is an hard statistical problem, as known in the statistical literature on ratio estimation [12].

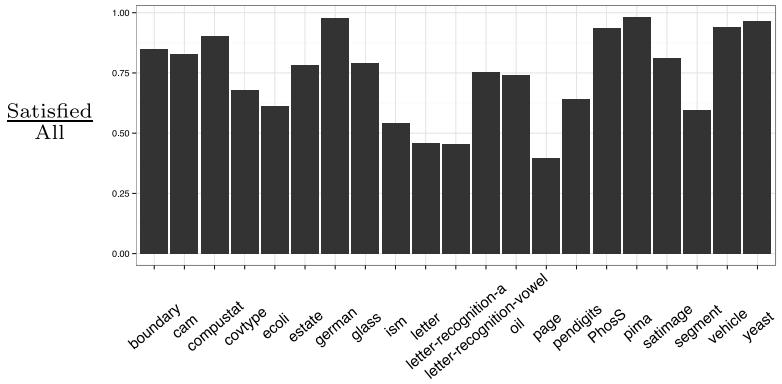


Fig. 11. Ratio between the number of sample satisfying condition 14 and all the instances available in each dataset averaged over all the βs.

5 Conclusion

Undersampling has become the de facto strategy to deal with skewed distributions, but, though easy to be justified, it conceals two major effects: i) it increases the variance of the classifier and ii) it produces warped posterior probabilities. The first effect is typically addressed by the use of averaging strategies (e.g. UnderBagging [21]) to reduce the variability while the second requires the calibration of the probability to the new priors of the testing set [18]. Despite the popularity of undersampling for unbalanced classification tasks, it is not clear how these two effects interact and when undersampling leads to better accuracy in the classification task.

In this paper, we aimed to analyse the interaction between undersampling and the ranking error of the posterior probability. We derive the condition (14) under which undersampling can improve the ranking and we show that when it is satisfied, the posterior probability obtained after sampling returns a more accurate ordering of testing instances. To validate our claim we used first synthetic and then real datasets, and in both cases we registered a better ranking with undersampling when condition (14) was met. It is important to remark how this condition shows that the beneficial impact of undersampling is strongly dependent on the nature of the classification task (degree of unbalancedness and non separability), on the variance of the classifier and as a consequence is extremely

dependent on the specific test point. We think that this result sheds light on the reason why several discordant results have been obtained in the literature about the effectiveness of undersampling in unbalanced tasks.

However, the practical use of this condition is not straightforward since it requires the knowledge of the posteriori probability and of the ratio of variances before and after undersampling. It follows that this result should be used mainly as a warning against a naive use of undersampling in unbalanced tasks and should suggest instead the adoption of specific adaptive selection techniques (e.g. racing [6]) to perform a case-by-case use (and calibration) of undersampling.

Acknowledgments

A. Dal Pozzolo is supported by the Doctiris scholarship *Adaptive real-time machine learning for credit card fraud detection* funded by Innoviris, Belgium. G. Bontempi is supported by the project *BridgeIRIS* funded by Innoviris, Belgium.

References

1. D.J. Newman A. Asuncion. UCI machine learning repository, 2007.
2. D Anyfantis, M Karagiannopoulos, S Kotsiantis, and P Pintelas. Robustness of learning techniques in handling class noise in imbalanced datasets. In *Artificial intelligence and innovations 2007: From theory to applications*, pages 21–28. Springer, 2007.
3. Gustavo EAPA Batista, Ronaldo C Prati, and Maria C Monard. Balancing strategies and class overlapping. In *Advances in Intelligent Data Analysis VI*, pages 24–35. Springer, 2005.
4. Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004.
5. Andrea Dal Pozzolo, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10):4915–4928, 2014.
6. Andrea Dal Pozzolo, Olivier Caelen, Serge Waterschoot, and Gianluca Bontempi. Racing for unbalanced methods selection. In *Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning*. IDEAL, 2013.
7. P. Domingos. Metacost: a general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 155–164. ACM, 1999.
8. Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Citeseer, 2001.
9. Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36, 2004.
10. Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez. On the k-nn performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, 11(3-4):269–280, 2008.

11. Vicente García, Jose Sánchez, and Ramon Mollineda. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 397–406. Springer, 2007.
12. HO Hartley and A Ross. Unbiased ratio estimators. 1954.
13. Haibo He and Edwardo A Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.
14. N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.
15. Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, 6(1):40–49, 2004.
16. Ronaldo Prati, Gustavo Batista, and Maria Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. *MICAI 2004: Advances in Artificial Intelligence*, pages 312–321, 2004.
17. Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.
18. Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.
19. Jerzy Stefanowski. Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In *Emerging Paradigms in Machine Learning*, pages 277–306. Springer, 2013.
20. Jason Van Hulse and Taghi Khoshgoftaar. Knowledge discovery from imbalanced and noisy data. *Data & Knowledge Engineering*, 68(12):1513–1542, 2009.
21. Shuo Wang, Ke Tang, and Xin Yao. Diversity exploration and negative correlation learning on imbalanced data sets. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 3259–3266. IEEE, 2009.
22. Gary M Weiss and Foster Provost. The effect of class distribution on classifier learning: an empirical study. *Rutgers Univ*, 2001.
23. B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, ICDM*, pages 435–442. IEEE, 2003.