When is undersampling effective in unbalanced classification tasks?

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

[1]Machine Learning Group, Computer Science Department, 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

Introduction

A Dataset is unbalanced when the class of interest (minority class) is much smaller or rarer than normal behaviour (majority class), e.g. in fraud detection we want to classify transactions as fraud or genuine, but fraud class is rare. Classification algorithms in general suffer when the data is skewed towards one class. A standard solution is *undersampling*, i.e. removing observations from the majority class until the datasets is balanced.

Warping effect on the posterior probability

Let p be the posterior probability of a classifier to predict an instance as belonging to the minority class. After undersampling we get p_s and we can write:

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

where β is the probability of selecting a majority instance.

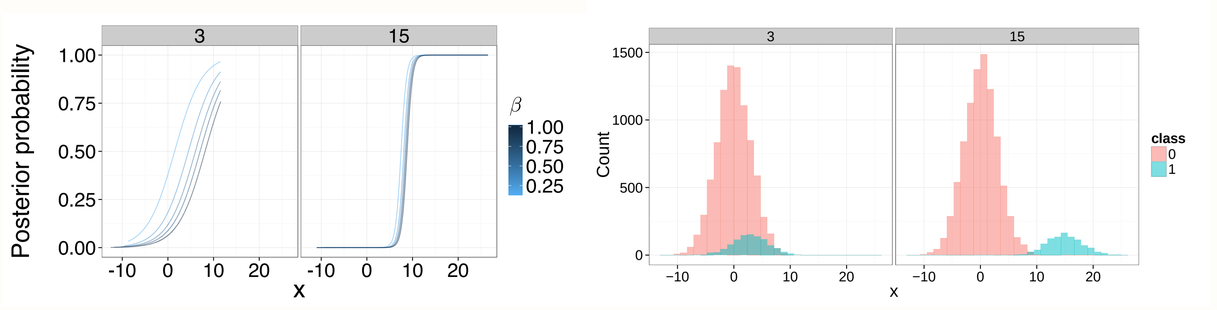

Figure 1: p_s as a function of β (left) and class distribution (right).

Posterior probability ranking

Let us denote by $\hat{p}$ (resp. $\hat{p}_s$) the estimation of p (resp. p_s). Assume we have two distinct test points with $p_1 < p_2$ where $\Delta p = p_2 - p_1 > 0$. Let $\hat{p}_1 = p_1 + \epsilon_1$ and $\hat{p}_2 = p_2 + \epsilon_2$, with $\varepsilon \sim N(b, \nu)$ where b and ν are the bias and the variance of the estimator of p. By making an hypothesis of normality we have a wrong ranking if $\hat{p}_1 > \hat{p}_2$ with probability:

$$P(\hat{p}_2 < \hat{p}_1) = P(p_2 + \epsilon_2 < p_1 + \epsilon_1) = P(\epsilon_1 - \epsilon_2 > \Delta p) = 1 - \Phi\left(\frac{\Delta p}{\sqrt{2\nu}}\right) \tag{2}$$

where $\epsilon_2 - \epsilon_1 \sim N(0, 2\nu)$ and Φ is the cumulative function of the normal distribution. Let $\hat{p}_{s,1} = p_{s,1} + \eta_1$ and $\hat{p}_{s,2} = p_{s,2} + \eta_2$, where $\eta \sim N(b_s, \nu_s)$, $\nu_s > \nu$ and $\Delta p_s = p_{s,2} - p_{s,1}$.

$$P(\hat{p}_{s,2} < \hat{p}_{s,1}) = P(\eta_1 - \eta_2 > \Delta p_s) = 1 - \Phi\left(\frac{\Delta p_s}{\sqrt{2\nu_s}}\right) \tag{3}$$

A classifier trained after undersampling has better ranking w.r.t. a classifier learned with unbalanced distribution when $P(\hat{p}_2 < \hat{p}_1) > P(\hat{p}_{s,2} < \hat{p}_{s,1})$, using (2) and (3):

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

since Φ is monotone non decreasing and we can assume that $\nu_s > \nu$.
Then it follows that undersampling is useful (better ranking) when

$$\frac{dp_s}{dp} = \frac{\beta}{(p + \beta(1-p))^2} > \sqrt{\frac{\nu_s}{\nu}} \tag{4}$$

where $\frac{dp_s}{dp}$ is the derivative of p_s w.r.t. p.

Experimental Results

Synthetic data

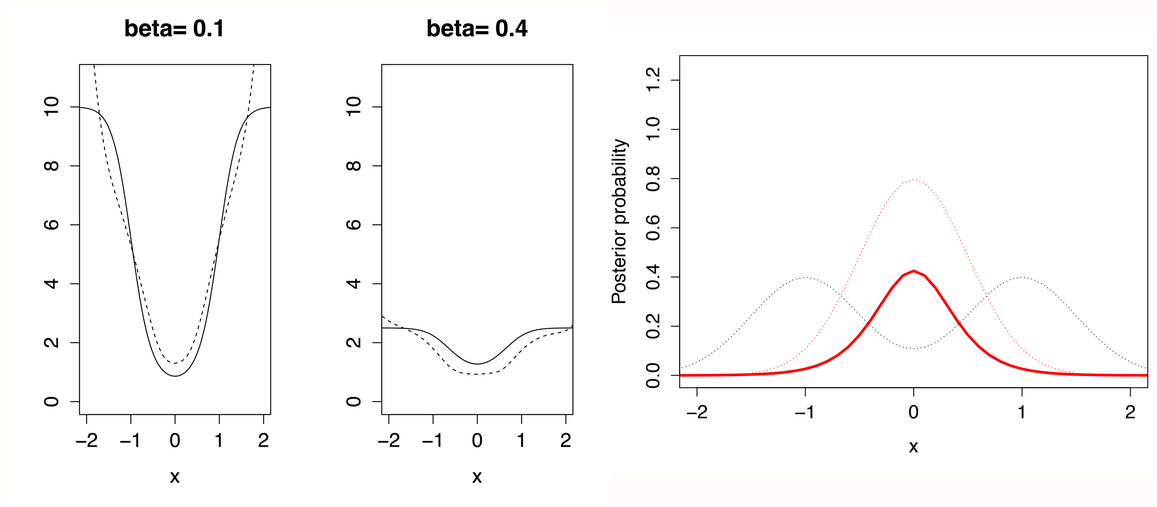

Figure 2: Left: $\frac{dp_s}{dp}$ (solid lines), $\sqrt{\frac{\nu_s}{\nu}}$ (dotted lines). Right: class conditional distributions (thin lines) and the posterior distribution of the minority class (thicker line).

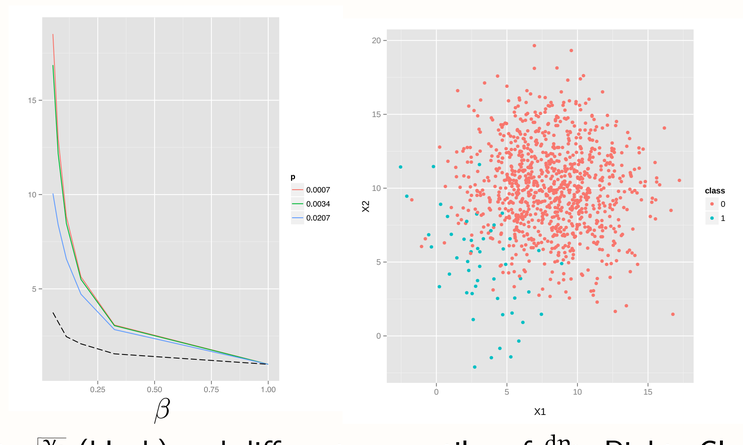

Figure 3: Left: $\sqrt{\frac{\nu_s}{\nu}}$ (black) and different percentiles of $\frac{dp_s}{dp}$. Right: Class distribution.

Table 1: 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 (4) is (not) satisfied.

β	$\mathcal{K}$	$\mathcal{K}_s$	$\mathcal{K}_s - \mathcal{K}$	%points satisfying (4)
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

Real data

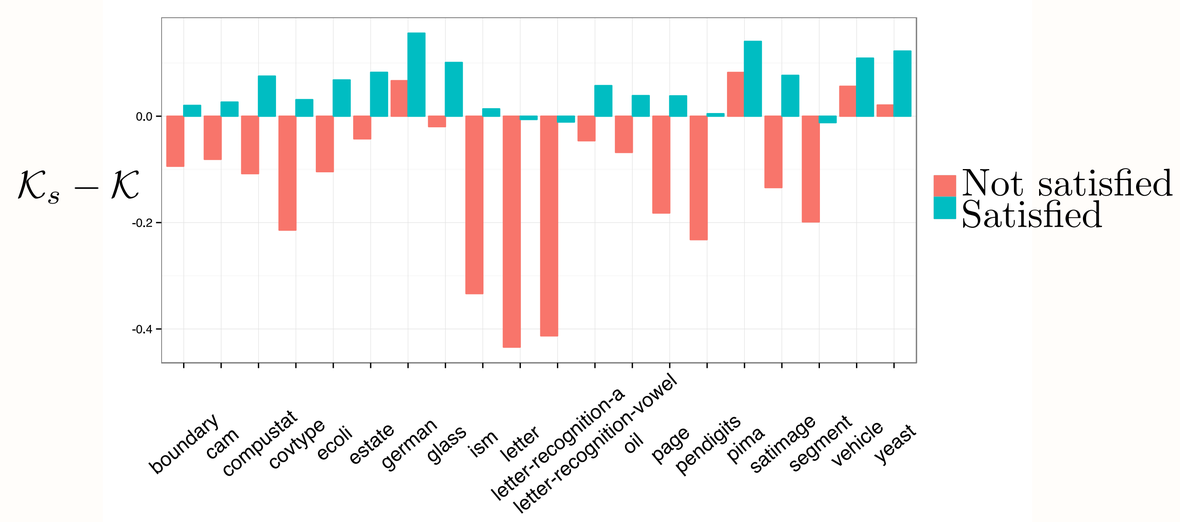

Figure 4: Difference between the Kendall rank correlation of $\hat{p}_s$ and $\hat{p}$ with p, namely $\mathcal{K}_s$ and $\mathcal{K}$, for points having the conditions (4) satisfied and not on selected datasets from the UCI repository. $\mathcal{K}_s$ and $\mathcal{K}$ are calculated as the mean of the correlations over all βs.

Summary and conclusions

Undersampling has two major effects: i) it increases the variance of the classifier and ii) it produces warped posterior probabilities. Countermeasures: i) averaging strategies (e.g. Bagging) and calibration of the probability to the new priors of the testing set [2].
When (4) is satisfied the posterior probability obtained after sampling returns a more accurate ordering. Practical use (4) requires knowledge of p and $\frac{\nu_s}{\nu}$ (not easy to estimate). Also (4) may not hold for all testing points and depends on β. This result warns against a naive use of undersampling in unbalanced tasks and suggest the adoption of adaptive selection techniques (e.g. racing [1]).

Research is funded by the *Doctiris* programme of Innoviris (Brussels capital region).

References

[1] 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.

[2] 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.

ECML-PKDD, 7 - 11 September 2015, Porto, Portugal