When is undersampling effective in unbalanced classification tasks?

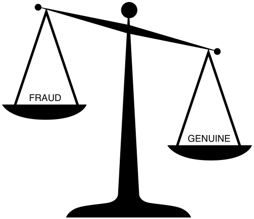
Andrea Dal Pozzolo, Olivier Caelen,
and Gianluca Bontempi

09/09/2015

ECML-PKDD 2015
Porto, Portugal

Introduction

- In several binary classification problems, the two classes are not equally represented in the dataset.
- In Fraud detection for example, fraudulent transactions are rare compared to genuine ones (less than 1% [3]).
- Many classification algorithms performs poorly in with unbalanced class distribution [7].
- A standard solution to unbalanced classification is rebalancing the classes before training a classifier.



UNDERSAMPLING

- Undersampling is a well-know technique used to balanced a dataset.
- It consists in down-sizing the majority class by removing observations at random until the dataset is balanced.
- Some works have empirically shown that classifiers perform better with balanced dataset [10] [6].
- Other show that balanced training set do not improve performances [2] [7].
- There is not yet a theoretical framework motivating undersampling.

Objective of this study

- We aim to analyse the role of the two side-effects of undersampling on the final accuracy:
  - The warping in the posterior distribution [5, 8].
  - The increase in variance due to samples removal.
- We analyse their impact on the final ranking of posterior probabilities.
- We show under which condition undersampling is expected to improve classification accuracy.

THE PROBLEM

- Let us consider a binary classification task $f : R^n \rightarrow \{+, -\}$
- $\mathbf{X} \in R^n$ is the input and $\mathbf{Y} \in \{+, -\}$ the output domain.
- $+$ is the minority and $-$ the majority class.
- Given a classifier $\mathcal{K}$ and a sample (x, y), we are interested in estimating the posterior probability $p(y = +|x)$.
- We want to study the effect of undersampling on the posterior probability.

The problem II

- Let $(X, Y) \subset (\mathcal{X}, \mathcal{Y})$ be the balanced sample of $(\mathcal{X}, \mathcal{Y})$, i.e. (X, Y) contains a subset of the negatives in $(\mathcal{X}, \mathcal{Y})$.
- Let **s** be a random variable associated to each sample $(x, y) \in (\mathcal{X}, \mathcal{Y})$, $\mathbf{s} = 1$ if the point is in $(x, y) \in (X, Y)$ and $\mathbf{s} = 0$ otherwise.
- Assume that **s** is independent of the input x given the class y (*class-dependent selection*):

$$p(s|y, x) = p(s|y) \Leftrightarrow p(x|y, s) = p(x|y)$$

.

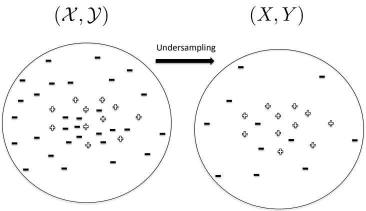


Figure : Undersampling: remove randomly majority class examples.

Posterior probabilities

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

In undersampling we have $p(s=1|+,x)=1$, so we can write:

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

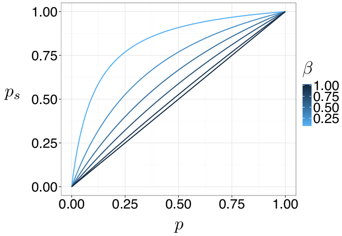
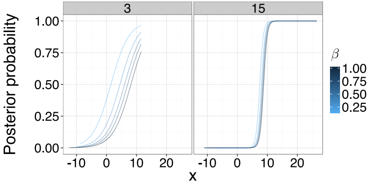
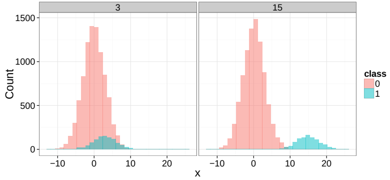


Figure : p and p_s at different β.

WARPING AND CLASS SEPARABILITY



(a) p_s as a function of β



(b) Class distribution

Figure : Class distribution and 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 $\beta < 1$.

Ranking error

- Let $\hat{p}$ (resp. $\hat{p}_s$) denote the estimation of p (resp. p_s).
- Assume $p_1 < p_2$, $\Delta p = p_2 - p_1$ with $\Delta p > 0$.
- 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.

We have a wrong ranking if $\hat{p}_1 > \hat{p}_2$ and its probability is:

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

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

$$P(\epsilon_1 - \epsilon_2 > \Delta p) = 1 - \Phi\left(\frac{\Delta p}{\sqrt{2\nu}}\right) \tag{3}$$

where Φ is the cumulative distribution function of the standard normal distribution.

Ranking error with undersampling

- Let $\hat{p}_{s,1} = p_{s,1} + \eta_1$ and $\hat{p}_{s,2} = p_{s,2} + \eta_2$, where $\eta \sim N(b_s, \nu_s)$.
- $\nu_s > \nu$, i.e. variance is larger given the smaller number of samples.
- $p_{s,1} < p_{s,2}$ and $\Delta p_s = p_{s,2} - p_{s,1} > 0$ because (2) is monotone.

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

$$P(\eta_1 - \eta_2 > \Delta p_s) = 1 - \Phi\left(\frac{\Delta p_s}{\sqrt{2\nu_s}}\right) \tag{4}$$

Condition for a better ranking with undersampling

A classifier $\mathcal{K}$ has better ranking with undersampling when

$$P(\epsilon_1 - \epsilon_2 > \Delta p) > P(\eta_1 - \eta_2 > \Delta p_s) \tag{5}$$

or equivalently from (3) and (4) when

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

since Φ is monotone non decreasing and $\nu_s > \nu$:

$$\boxed{\frac{dp_s}{dp} > \sqrt{\frac{\nu_s}{\nu}}} \tag{6}$$

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

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

FACTORS INFLUENCING (6)

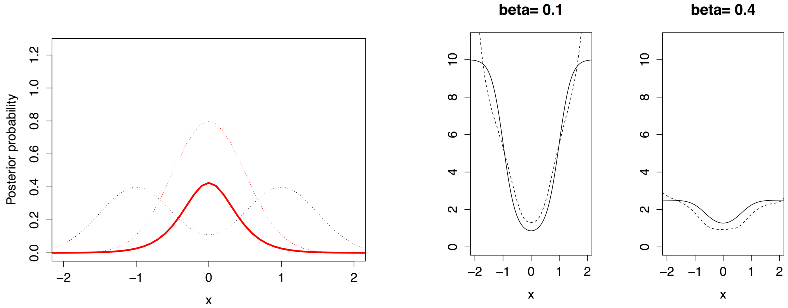
The value of inequality (6) depends on several terms:

- The rate of undersampling β impacts the terms p_s and ν_s.
- The ratio of the variances $\frac{\nu_s}{\nu}$.
- The posteriori probability p of the testing point.

The condition (6) is hard to verify: β can be controlled by the designer, but $\frac{dp_s}{dp}$ and $\frac{\nu_s}{\nu}$ vary over the input space.

This means that (6) does not necessarily hold for all the test points.

UNIVARIATE SYNTHETIC DATASET

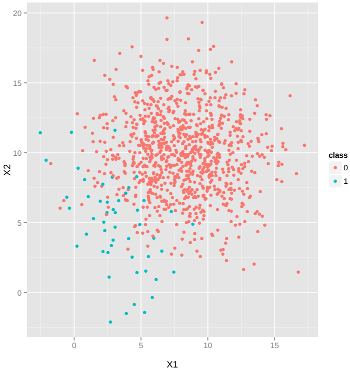


(a) Class conditional distributions (thin lines) and the posterior distribution of the minority class (thicker line).

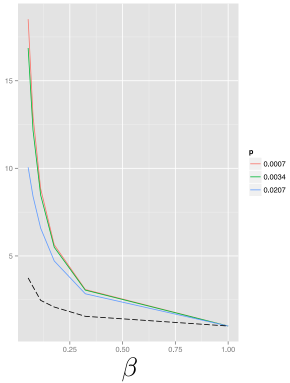
(b) $\frac{dp_s}{dp}$ (solid lines), $\sqrt{\frac{\nu_s}{\nu}}$ (dotted lines).

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

BIVARIATE SYNTHETIC DATASET



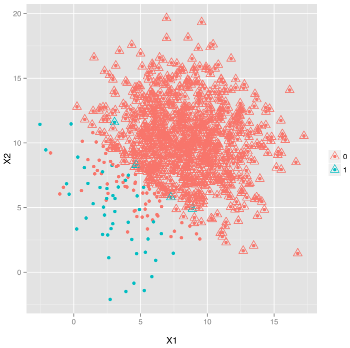
(a) Synthetic dataset 1



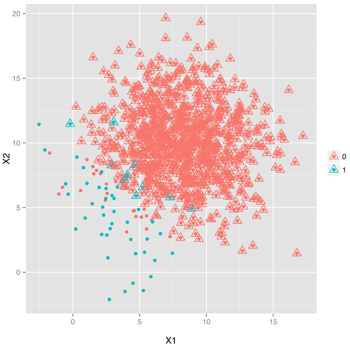
(b) $\sqrt{\frac{\nu_s}{\nu}}$ and $\frac{dp_s}{dp}$ for different β

Figure : 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).

BIVARIATE SYNTHETIC DATASET II



(a) Undersampling with $\beta = 0.053$



(b) Undersampling with $\beta = 0.323$

Figure : Regions where undersampling should work. Triangles indicate the testing samples where the condition (6) holds for the dataset in Figure 5.

BIVARIATE SYNTHETIC DATASET III

Table : Classification task in Figure 5: 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 (6) is (not) satisfied.

β	$\mathcal{K}$	$\mathcal{K}_s$	$\mathcal{K}_s - \mathcal{K}$	%points satisfying (6)
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 DATASETS

Table : Selected datasets from the UCI repository [1][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

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

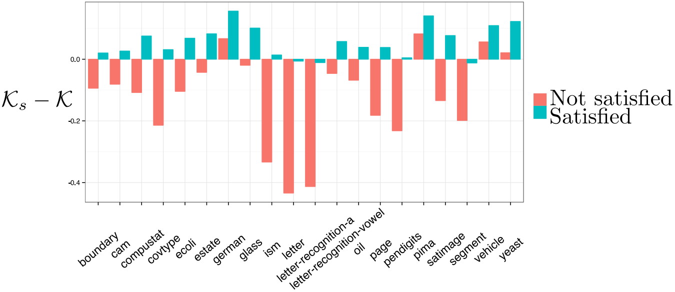


Figure : Difference between the Kendall rank correlation of $\hat{p}_s$ and $\hat{p}$ with p, namely $\mathcal{K}_s$ and $\mathcal{K}$, for points having the conditions (6) satisfied and not. $\mathcal{K}_s$ and $\mathcal{K}$ are calculated as the mean of the correlations over all βs.

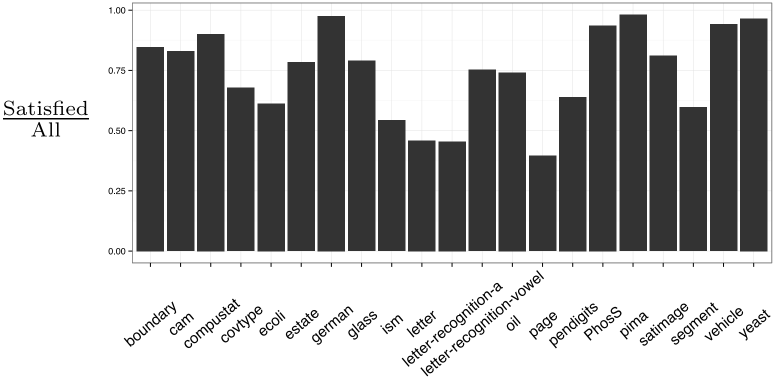


Figure : Ratio between the number of sample satisfying condition 6 and all the instances available in each dataset averaged over all the βs.

SUMMARY

Undersampling has two major effects:

- it increases the variance of the classifier
- it produces warped posterior probabilities.

Countermeasures:

- averaging strategies (e.g. UnderBagging [9])
- calibration of the probability to the new priors of the testing set [8].

Despite the popularity of undersampling, it is not clear how these two effects interact and when undersampling leads to better accuracy in the classification task.

Conclusion

- When (6) is satisfied the posterior probability obtained after sampling returns a more accurate ordering.
- Several factors influence (6) (e.g. β, variance of the classifier, class separability)
- Practical use (6) is not straightforward since it requires knowledge of p and $\frac{\nu_s}{\nu}$ (not easy to estimate).
- This result warning against a naive use of undersampling in unbalanced tasks.
- We suggest the adoption of adaptive selection techniques (e.g. racing [4]) to perform a case-by-case use of undersampling.

Code: `https://github.com/dalpozz/warping`
Website: `www.ulb.ac.be/di/map/adalpozz`
Email: adalpozz@ulb.ac.be

Thank you for the attention

Research is supported by the *Doctiris* scholarship funded by Innoviris, Brussels, Belgium.

BIBLIOGRAPHY

[1] D. N. A. Asuncion.
UCI machine learning repository, 2007.

[2] G. E. Batista, R. C. Prati, and M. C. Monard.
A study of the behavior of several methods for balancing machine learning training data.
ACM SIGKDD Explorations Newsletter, 6(1):20–29, 2004.

[3] A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi.
Learned lessons in credit card fraud detection from a practitioner perspective.
Expert Systems with Applications, 41(10):4915–4928, 2014.

[4] A. Dal Pozzolo, O. Caelen, S. Waterschoot, and G. Bontempi.
Racing for unbalanced methods selection.
In *Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning*. IDEAL, 2013.

[5] C. Elkan.
The foundations of cost-sensitive learning.
In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Citeseer, 2001.

[6] A. Estabrooks, T. Jo, and N. Japkowicz.
A multiple resampling method for learning from imbalanced data sets.
Computational Intelligence, 20(1):18–36, 2004.

[7] N. Japkowicz and S. Stephen.
The class imbalance problem: A systematic study.
Intelligent data analysis, 6(5):429–449, 2002.

[8] M. Saerens, P. Latinne, and C. Decaestecker.
Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure.
Neural computation, 14(1):21–41, 2002.

[9] S. Wang, K. Tang, and X. 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.

[10] G. M. Weiss and F. Provost.
The effect of class distribution on classifier learning: an empirical study.
Rutgers Univ, 2001.