Racing for unbalanced methods selection – the `unbalanced` package

Andrea Dal Pozzolo

24/2/2016

Rbelgium

About me

- I'm a Decision Analytics Consultant at VASCO Data Security.
- I hold a PhD from the Machine Learning Group (MLG) of the Université Libre de Bruxelles (ULB).
- My research focused on Machine Learning techniques for Fraud Detection in electronic transaction.

Table of Contents

INTRODUCTION

- In several binary classification problems (e.g. fraud detection), the two classes are not equally represented in the dataset.
- When one class is underrepresented in a dataset, the data is said to be unbalanced.
- In unbalanced datasets, classification algorithms perform poorly in terms of predictive accuracy [9].
- A common strategy is to balance the classes before learning a classifier [1].

Technical sections will be denoted by the symbol 

PCA projection of credit card transactions

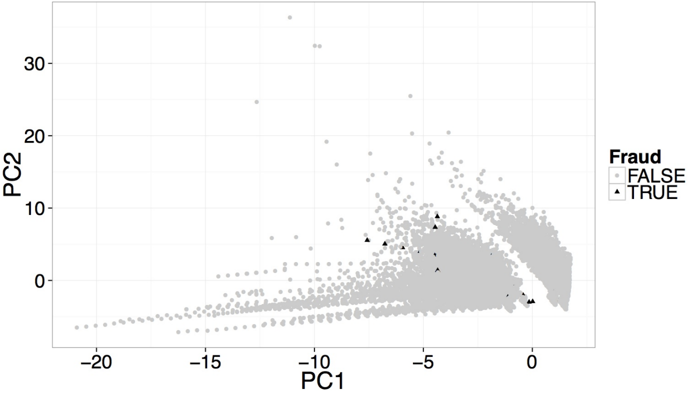


Figure : Transactions distribution over the first two Principal Components in different days.

The Fraud Detection problem

- We formalize FD as a classification task $f : \mathbb{R}^n \rightarrow \{+, -\}$.
- $X \in \mathbb{R}^n$ is the input and $Y \in \{+, -\}$ the output domain.
- $+$ is the fraud (minority) and $-$ the genuine (majority) class.
- Given a classifier $\mathcal{K}$ and a training set T_N, we are interested in estimating for a new sample (x, y) the posterior probability $\mathcal{P}(y = +|x)$.

Preprocessing methods for unbalanced data

- Sampling methods
  - Undersampling [7]
  - Oversampling [7]
  - SMOTE [3]
- Distance based methods
  - Tomek link [15]
  - Condensed Nearest Neighbor (CNN) [8]
  - One side Selection (OSS) [10]
  - Edited Nearest Neighbor (ENN) [16]
  - Neighborhood Cleaning Rule (NCL) [11]

SAMPLING METHODS

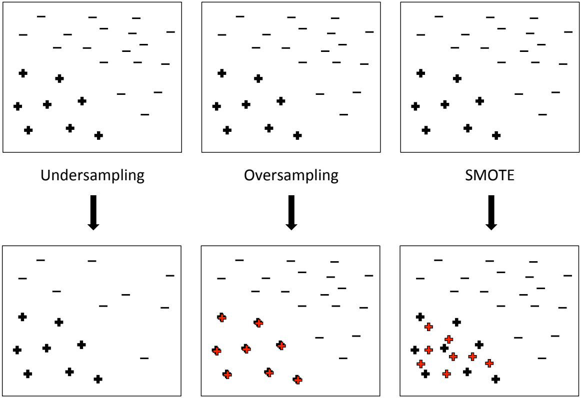


Figure : Resampling methods for unbalanced classification. The negative and positive symbols denotes majority and minority class instances. In red the new observations created with oversampling methods.

SELECTING THE BEST STRATEGY

- With no prior information about the data distribution is difficult to decide which unbalanced strategy to use.
- *No-free-lunch theorem* [17]: no single strategy is coherently superior to all others in all conditions (i.e. algorithm, dataset and performance metric)
- Testing all unbalanced techniques is not an option because of the associated computational cost.
- We proposed to use the Racing approach [12] to perform strategy selection.

Racing for strategy selection

- Racing consists in testing in parallel a set of alternatives and using a statistical test to remove an alternative if it is significantly worse than the others.
- We adopted F-Race version [2] to search efficiently for the best strategy for unbalanced data.
- The F-race combines the Friedman test with Hoeffding Races [12].

Racing for unbalanced technique selection

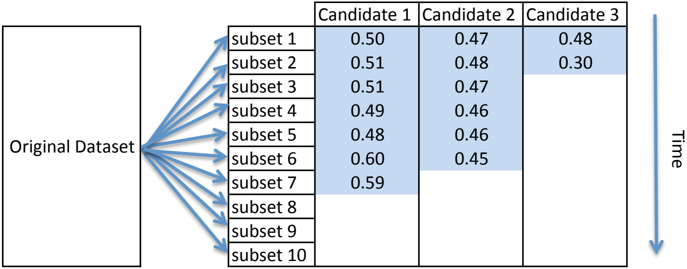
Automatically select the most adequate technique for a given dataset.

1. Test in parallel a set of alternative balancing strategies on a subset of the dataset
2. Remove progressively the alternatives which are significantly worse.
3. Iterate the testing and removal step until there is only one candidate left or not more data is available

	Candidate 1	Candidate 2	Candidate 3
subset 1	0.50	0.47	0.48
subset 2	0.51	0.48	0.30
subset 3	0.51	0.47	
subset 4	0.60	0.45	
subset 5	0.55		

F-RACE METHOD

- Use 10-fold Cross Validation (CV) to provide the data during the race.
- Every time new data is added to the race, the Friedman test is used to remove significantly bad candidates.
- We made a comparison of CV and F-race in terms of F-measure.



	Candidate 1	Candidate 2	Candidate 3
subset 1	0.50	0.47	0.48
subset 2	0.51	0.48	0.30
subset 3	0.51	0.47	
subset 4	0.49	0.46	
subset 5	0.48	0.46	
subset 6	0.60	0.45	
subset 7	0.59		
subset 8			
subset 9			
subset 10			

F-RACE VS CROSS VALIDATION

Dataset	Exploration	Method	Ntest	% Gain	Mean	Sd
ecoli	Race	Under	46	49	0.836	0.04
	CV	SMOTE	90	-	0.754	0.112
letter-a	Race	Under	34	62	0.952	0.008
	CV	SMOTE	90	-	0.949	0.01
letter-vowel	Race	Under	34	62	0.884	0.011
	CV	Under	90	-	0.887	0.009
letter	Race	SMOTE	37	59	0.951	0.009
	CV	Under	90	-	0.951	0.01
oil	Race	Under	41	54	0.629	0.074
	CV	SMOTE	90	-	0.597	0.076
page	Race	SMOTE	45	50	0.919	0.01
	CV	SMOTE	90	-	0.92	0.008
pendigits	Race	Under	39	57	0.978	0.011
	CV	Under	90	-	0.981	0.006
PhosS	Race	Under	19	79	0.598	0.01
	CV	Under	90	-	0.608	0.016
satimage	Race	Under	34	62	0.843	0.008
	CV	Under	90	-	0.841	0.011
segment	Race	SMOTE	90	0	0.978	0.01
	CV	SMOTE	90	-	0.978	0.01
estate	Race	Under	27	70	0.553	0.023
	CV	Under	90	-	0.563	0.021
covtype	Race	Under	42	53	0.924	0.007
	CV	SMOTE	90	-	0.921	0.008
cam	Race	Under	34	62	0.68	0.007
	CV	Under	90	-	0.674	0.015
compustat	Race	Under	37	59	0.738	0.021
	CV	Under	90	-	0.745	0.017
creditcard	Race	Under	43	52	0.927	0.008
	CV	SMOTE	90	-	0.924	0.006

Table : Results in terms of G-mean for RF classifier.

DISCUSSION

- The best strategy is extremely dependent on the data nature, algorithm adopted and performance measure.
- F-race is able to automatise the selection of the best unbalanced strategy for a given unbalanced problem without exploring the whole dataset.
- For the fraud dataset the unbalanced strategy chosen had a big impact on the accuracy of the results.

CONCLUSIONS

- With racing [6] we can rapidly select the best strategy for a given unbalanced task.
- However, we see that undersampling and SMOTE are often the best strategy to adopt.
- In R [14] we have a software package called `unbalanced` [4] that implements all these algorithms.

PACKAGE DEMO

Let's see in practise how to do it ...

The code of this demo is available at

`https://github.com/dalpozz/utility/blob/master/ubDEMO.R`

Vignette: `www.ulb.ac.be/di/map/adalpozz/pdf/unbalanced.pdf`
CRAN: `http://CRAN.R-project.org/package=unbalanced`
Github: `https://github.com/dalpozz/unbalanced`
Email: dalpozz@gmail.com

Questions?

BIBLIOGRAPHY I

[1] R. Akbani, S. Kwek, and N. Japkowicz.
Applying support vector machines to imbalanced datasets.
In *Machine Learning: ECML 2004*, pages 39–50. Springer, 2004.

[2] M. Birattari, T. Stützle, L. Paquete, and K. Varrentrapp.
A racing algorithm for configuring metaheuristics.
In *Proceedings of the genetic and evolutionary computation conference*, pages 11–18, 2002.

[3] N. Chawla, K. Bowyer, L. O. Hall, and W. P. Kegelmeyer.
Smote: synthetic minority over-sampling technique.
Journal of Artificial Intelligence Research (JAIR), 16:321–357, 2002.

[4] A. Dal Pozzolo, O. Caelen, and G. Bontempi.
unbalanced: Racing For Unbalanced Methods Selection., 2015.
R package version 2.0.

[5] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi.
Calibrating probability with undersampling for unbalanced classification.
In *2015 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, 2015.

[6] 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.

[7] C. Drummond and R. Holte.
C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling.
In *Workshop on Learning from Imbalanced Datasets II*, 2003.

Bibliography II

[8] P. E. Hart.
The condensed nearest neighbor rule.
IEEE Transactions on Information Theory, 1968.

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

[10] M. Kubat, S. Matwin, et al.
Addressing the curse of imbalanced training sets: one-sided selection.
In *ICML*, volume 97, pages 179–186. Nashville, USA, 1997.

[11] J. Laurikkala.
Improving identification of difficult small classes by balancing class distribution.
Artificial Intelligence in Medicine, pages 63–66, 2001.

[12] O. Maron and A. Moore.
Hoeffding races: Accelerating model selection search for classification and function approximation.
Robotics Institute, page 263, 1993.

[13] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence.
Dataset shift in machine learning.
The MIT Press, 2009.

[14] R Development Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2011.
ISBN 3-900051-07-0.

BIBLIOGRAPHY III

[15] I. Tomek.
Two modifications of cnn.
IEEE Trans. Syst. Man Cybern., 6:769–772, 1976.

[16] D. Wilson.
Asymptotic properties of nearest neighbor rules using edited data.
Systems, Man and Cybernetics, (3):408–421, 1972.

[17] D. H. Wolpert.
The lack of a priori distinctions between learning algorithms.
Neural computation, 8(7):1341–1390, 1996.

Sampling selection bias

Sampling selection bias [13] occurs when the training and testing set comes from a different distribution. In this case we need to calibrate the posterior probability of the classifier [5].