# Comparison of balancing techniques for unbalanced datasets


Andrea Dal Pozzolo, Olivier Caelen and Gianluca Bontempi

Atos Worldline Belgium

Machine Learning Group
Université Libre de Bruxelles
Belgium


## Introduction

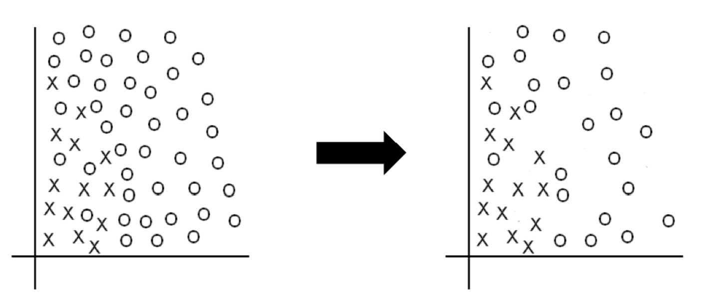
A Dataset is unbalanced when the class of interest (minority class) is much smaller or rarer than normal behaviour (majority class). Classification algorithms in general suffer when the data is skewed towards one class. In this poster we present a comparison of existing methods for dealing with unbalanced data.

## Unbalanced problem

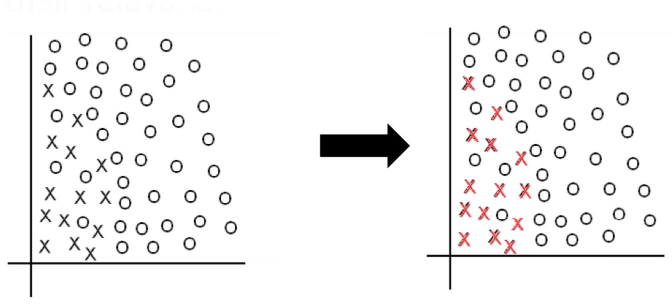
- The cost of missing a minority class is typically much higher that missing a majority class.
- Most learning systems are not prepared to cope with large difference between the number of cases belonging to each class
- Classification algorithm underperform when data is unbalanced[4].
- The unbalance problem is typical of many applications such as **fraud detection**, medical diagnosis, text classification, oil spills detection, ecc.

## Existing methods for unbalanced data

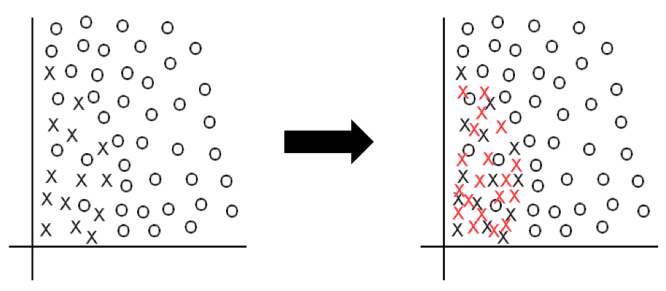
**Sampling methods** Many of the existing methods for classification with unbalanced dataset take advantage of sampling techniques to balance the dataset[4].



Undersampling

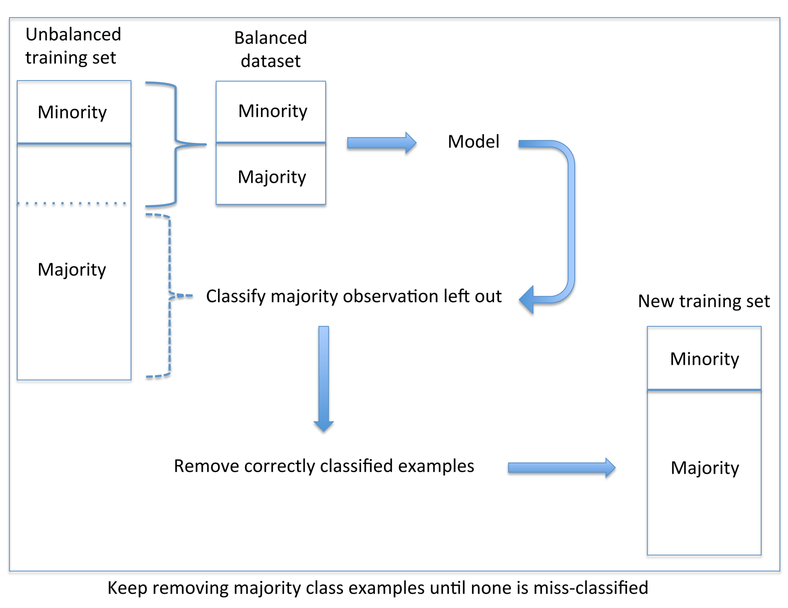


Oversampling

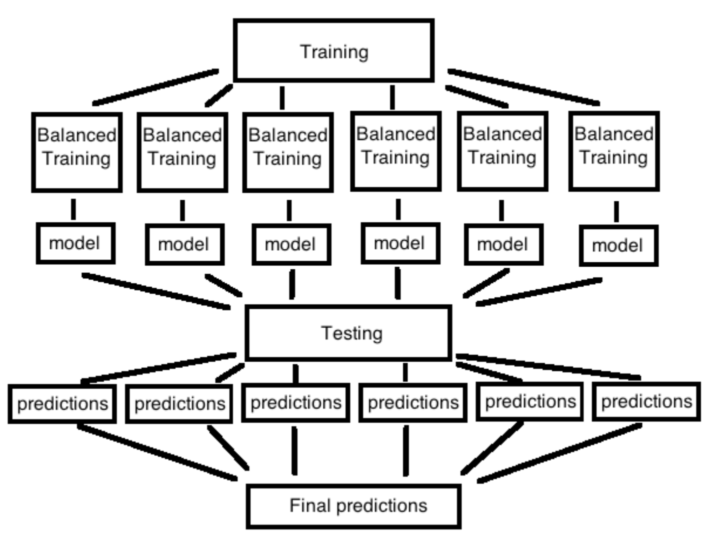


Smote[1]

**Ensemble methods** BalanceCascade, explore the majority class in a supervised manner, whereas EasyEnsemble, learns different aspects of the original majority class in an unsupervised manner.




Balance Cascade[6]




Easy Ensemble[6]

**Cost based methods** Type of learning that takes the misclassification costs into consideration [5] (Cost FN >> cost FP).

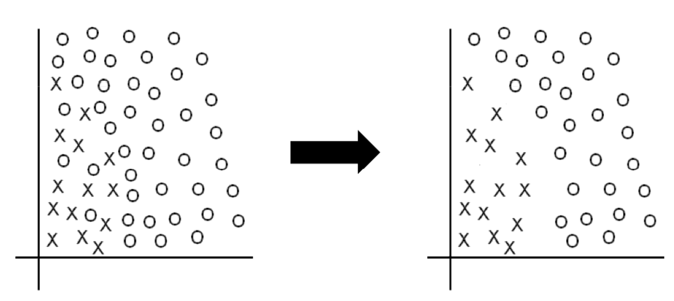
Cost-insensitive algorithm can be converted into cost-sensitive using a wrapper approach: modify the class distribution of the training data and then apply the cost-insensitive algorithm.

- Cost proportional sampling [2], positive and negative examples sample by the ratio:

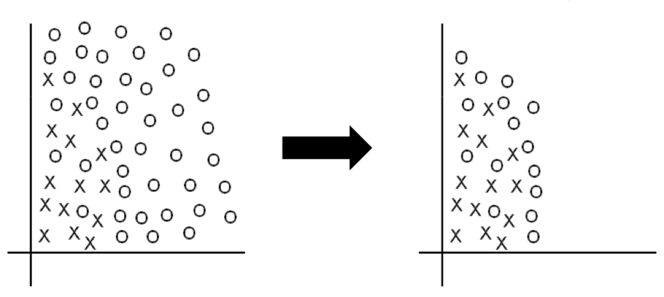
$$p(majority)FNcost : p(minority)FPcost$$

- Costing [8], accept an instance into the sample with the accepting probability C(i)/Z, where C(i) is the misclassification cost of class i, and Z is an arbitrary constant such that $Z \geq maxC(i)$

**Other methods** Goal is to remove both noise and borderline examples or instances from the majority class that are distant from the decision border, considered less relevant for learning.



Tomek link [7]



Condensed Nearest Neighbor [3]

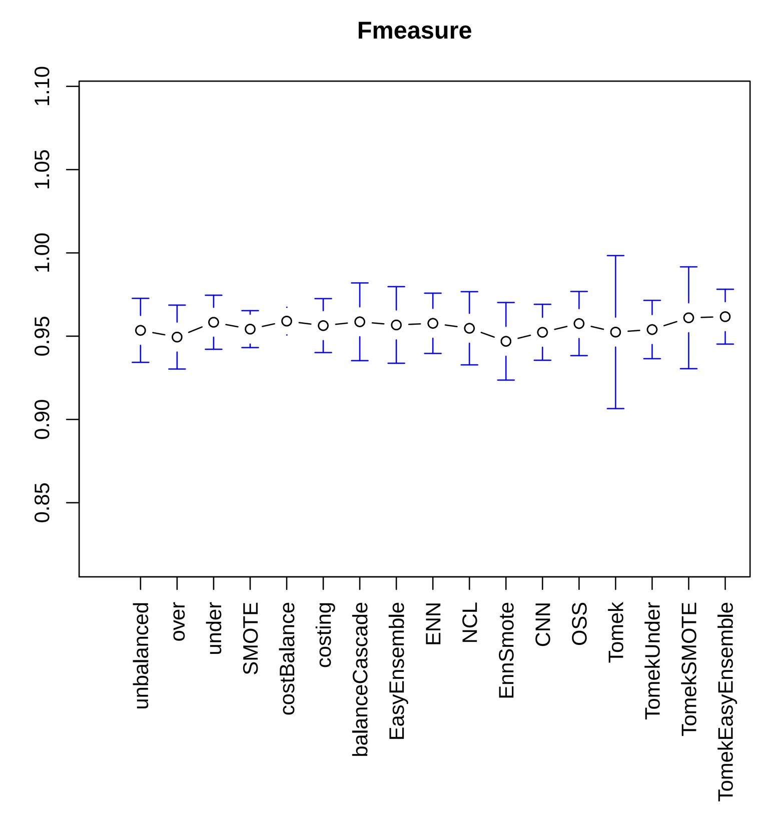
## Experimental Results

### Data

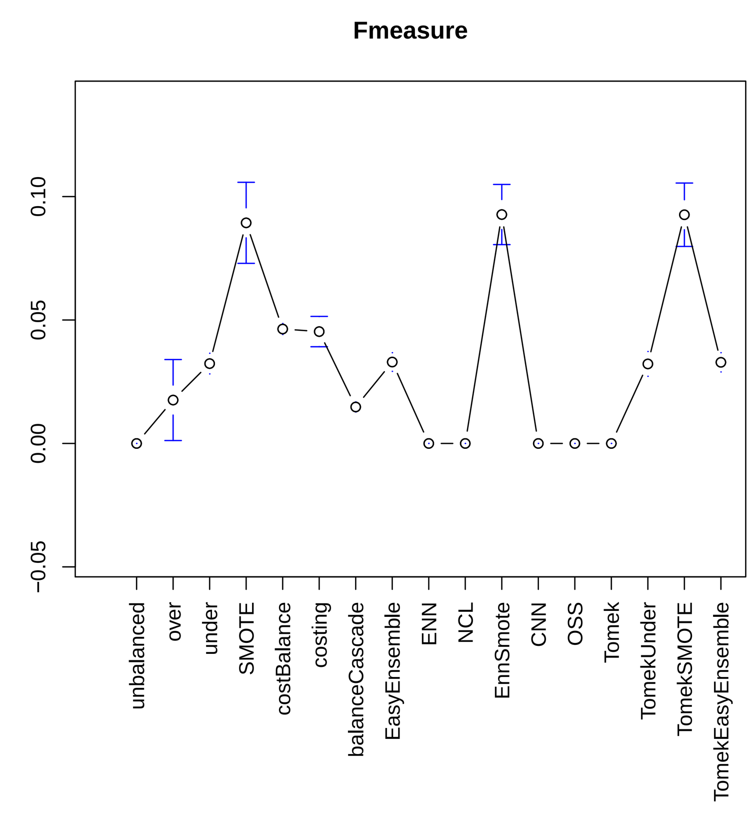
| Dataset | Size | Input | Prop 1 | Class 1 |
|---|---|---|---|---|
| breastcancer | 698 | 10 | 34.52% | class =4 |
| car | 1727 | 6 | 3.76% | class = Vgood |
| forest | 38501 | 54 | 7.13% | class = Cottonwood/Willow |
| letter | 19999 | 16 | 3.76% | letter = W |
| nursery | 12959 | 8 | 2.53% | class = very_recom |
| pima | 768 | 8 | 34.89% | class = 1 |
| satimage | 6433 | 36 | 9.73% | class = 4 |
| women | 1472 | 9 | 22.62% | class = long-term |
| spam | 4601 | 57 | 41.14% | class = 1 |
| credit | 150000 | 10 | 6.68% | SeriousDlqin2yrs = 1 |
| claim | 800000 | 34 | 0.71% | claim>1 |
| ford | 304544 | 30 | 16.41% | alert = 1 |
| kicked | 72983 | 31 | 12.29% | IsBadBuy = 1 |
| kdd99 | 398965 | 41 | 19.45% | class != normal |
| fraud | 527026 | 51 | 0.39% | Fraud = 1 |

### Results

- Some dataset present easy problem where there are not significant differences between the methods.




UCI breast cancer dataset




Atos fraud dataset

- We did a multiple comparison of all methods over all datasets using the Friedman test with the F-measure. In the following table a cell is marked as (+) if the rank difference between the method in the row and the method the column is positive, (-) othervise.
  The table shows the level of significance using *** ($\alpha = 0.001$), ** ($\alpha = 0.01$), * ($\alpha = 0.05$), . ($\alpha = 0.1$).

| | balanceCas | CNN | costBalance | costing | EasyEnsemble | ENN | EnnSmote | NCL | OSS | over | SMOTE | Tomek | TomekEasyEns | TomekSMOTE | TomekUnder | unbalanced | under |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| balanceCas | | | | | | | | | | | | | | | | | |
| CNN | | | | | | | (-) * | | (-) . | (-) * | (-) * | | | (-) ** | | | |
| costBalance | | | | | | | | | | | | | | | | | |
| costing | | | | | | | (-) . | | | (-) * | (-) * | | | (-) * | | | |
| EasyEnsemble | | | | | | | | | | | | | | | | | |
| ENN | | | | | | | | | | (-) . | (-) . | | | (-) * | | | |
| EnnSmote | | (+) * | | (+) . | | | | | | | | | | | (+) * | (+) . | |
| NCL | | | | | | | | | | | | | | | | | |
| OSS | | (+) . | | | | | | | | | | | | | (+) . | | |
| over | | (+) * | | (+) * | | (+) . | | | | | | | | | (+) * | (+) * | (+) . |
| SMOTE | | (+) * | | (+) * | | (+) . | | | | | | | | | (+) * | (+) * | (+) . |
| Tomek | | | | | | | | | | | | | | | | | |
| TomekEasyEns | | | | | | | | | | | | | | | | | |
| TomekSMOTE | | (+) ** | | (+) * | | (+) * | | | | | | | | | (+) * | (+) * | (+) * |
| TomekUnder | | | | | | | (-) * | | (-) . | (-) * | (-) * | | | (-) * | | | |
| unbalanced | | | | | | | (-) . | | | (-) * | (-) * | | | (-) * | | | |
| under | | | | | | | | | | (-) . | (-) . | | | (-) * | | | |

## Conclusion

- Using F-measure as metric, SMOTE and its combinations with Tomek link and ENN appear to be the best methods.
- Future work: release a R package for unbalanced data.


The project is funded by the Doctiris programme of Brussels capital region.


### References


[1] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Arxiv preprint arXiv:1106.1813*, 2011.

[2] C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Citeseer, 2001.

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

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

[5] C.X. Ling and V.S. Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning*, 2008.

[6] X.Y. Liu, J. Wu, and Z.H. Zhou. Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(2):539–550, 2009.

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

[8] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 435–442. IEEE, 2003.