The International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2013)

Racing for unbalanced methods selection

Andrea DAL POZZOLO, Olivier CAELEN, Serge WATERSCHOOT and Gianluca BONTEMPI

22/10/2013

Machine Learning Group
Université Libre de Bruxelles

Business Analytics Competence Center - Atos Worldline

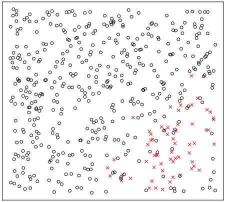
Table of contents

Introduction

- A common problem in data mining is dealing with unbalanced datasets in which one class vastly outnumbers the other in the training data.
- State-of-the-art classification algorithms suffer when the data is skewed towards one class [8].
- Several techniques have been proposed to cope with unbalanced data.
- However no technique appears to work consistently better in all conditions.
- We propose to use a racing method to select adaptively the most appropriate strategy for a given unbalanced task.

Unbalanced problem

A dataset is unbalanced when the class of interest (minority class) is much rarer than the other (majority class).



- The unbalanced nature of the data is typical of many applications such as medical diagnosis, text classification and credit card fraud detection.
- The cost of missing a minority class is typically much higher that missing a majority class.
- Proposed strategies essentially belong to the following categories: sampling, ensemble, cost-based and distance-based.

Existing methods for unbalanced data

- Sampling methods
  - Undersampling [5]
  - Oversampling [5]
  - SMOTE [3]
- Ensemble methods
  - BalanceCascade [11]
  - EasyEnsemble [11]
- Cost based methods
  - Cost proportional sampling [6]
  - Costing [19]
- Distance based methods
  - Tomek link [15]
  - Condensed Nearest Neighbor (CNN) [7]
  - One side Selection (OSS) [9]
  - Edited Nearest Neighbor (ENN) [17]
  - Neighborhood Cleaning Rule (NCL) [10]

Unbalanced strategies

- Sampling techniques up-sample or down-sample a class to rebalance the classes.
- SMOTE generates synthetic minority examples.
- Ensemble techniques combine an unbalanced method with a classifier to explore the majority and minority class distribution.
- Cost based techniques consider the misclassification cost to rebalance the dataset.
- Distance based techniques use distances between input points to undersample or to remove noisy and borderline examples.

Sampling methods

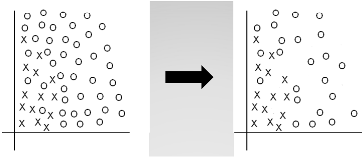


Figure: Undersampling

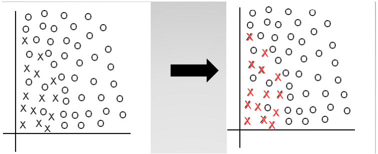


Figure: Oversampling

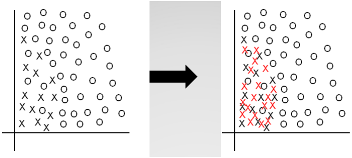


Figure: SMOTE [3]

Fraud detection problem

- Credit card fraud detection [13, 4, 14] is a highly unbalanced problem.
- Fraudulent behaviour evolves over the time changing the distribution of the frauds and a method that worked well in the past could become inaccurate afterward.

Datasets

- 1 real credit card fraud dataset provided by a payment service provider in Belgium.
- 9 datasets from UCI [1]

Dataset ID	Dataset name	Size	Input	Prop 1	Class 1
1	**fraud**	**527026**	51	**0.39%**	Fraud = 1
2	breastcancer	698	10	34.52%	class =4
3	car	1727	6	3.76%	class = Vgood
4	forest	38501	54	7.13%	class = Cottonwood/Willow
5	letter	19999	16	3.76%	letter = W
6	nursery	12959	8	2.53%	class = very_recom
7	pima	768	8	34.89%	class = 1
8	satimage	6433	36	9.73%	class = 4
9	women	1472	9	22.62%	class = long-term
10	spam	4601	57	42.14%	class =1

Some datasets are reduced to speed up computations.

Fraud Data - Fmeasure

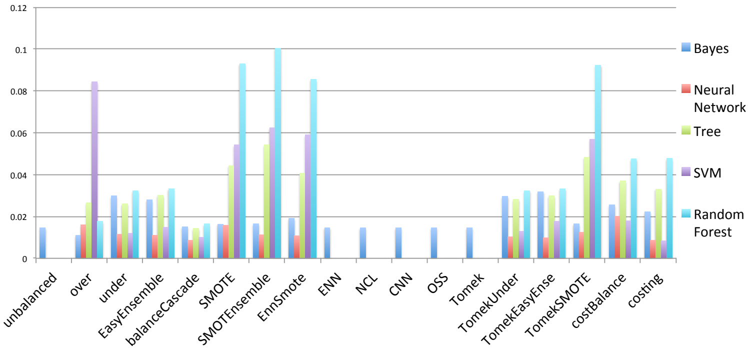


Figure: Comparison of strategies for unbalanced data in terms of F-measure for the Fraud dataset using different supervised algorithms, where F-measure $= 2 \times \frac{Precision \times Recall}{Precision + Recall}$.

Friedman test over all dataset using RF and F-measure

In the table a cell is marked as (+) if the rank difference between the method in the row and the method the column is positive, (-) otherwise.

The table shows the level of significance using *** ($\alpha = 0.001$), ** ($\alpha = 0.01$), * ($\alpha = 0.05$), . ($\alpha = 0.1$).

	Cascade	CNN	costBal	costing	EasyEns	ENN	EnnSmt	NCL	OSS	over	SMT	SMTEns	Tomek	TomekE	TomekS	TomekU	unbal	under
Cascade		(+) .														(+) .	(+) .	
CNN	(-) .						(-) .				(-) .	(-) *			(-) *			
costBal																		
costing																		
EasyEns																		
ENN												(-) .						
EnnSmt		(+) .																
NCL												(-) .			(-) .			
OSS																		
over																		
SMT		(+) .														(+) .	(+) .	
SMTEns		(+) *				(+) .		(+) .								(+) *	(+) *	(+) .
Tomek																		
TomekE																		
TomekS		(+) *						(+) .								(+) .	(+) .	
TomekU	(-) .										(-) .	(-) *			(-) .			
unbal	(-) .										(-) .	(-) *			(-) .			
under												(-) .						

Figure: Comparison of strategies using a post-hoc Friedman test in terms of F-measure for a RF classifier over multiple datasets.

Racing idea

- With no prior information about the data distribution is difficult to decide which unbalanced strategy to use.
- No single strategy is coherently superior to all others in all conditions (i.e. algorithm, dataset and performance metric)
- Under different conditions, such as fraud evolution, the best methods may change.
- 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

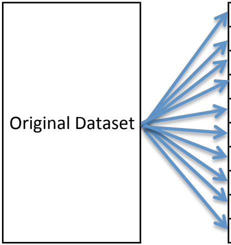
Automatically select the most adequate technique for a given dataset.

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 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 Cross Validation 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			

Time

F-race Vs Cross Validation

Dataset	Algo	Exploration	Method	N test	Gain	Mean	Sd	Pval
Fraud	RF	best CV	SMOTEnsemble	180	-	0.100	0.016	-
		F-race	SMOTEnsemble	44	76%			
	SVM	best CV	over	180	-	0.084	0.017	-
		F-race	over	46	74%			
Breast Cancer	RF	best CV	balanceCascade	180	-	0.963	0.035	-
		F-race	balanceCascade	180	0%			
	SVM	best CV	under	180	-	0.957	0.038	-
		F-race	under	180	0%			
Car	RF	best CV	OSS	180	-	0.970	0.039	-
		F-race	OSS	108	40%			
	SVM	best CV	over	180	-	0.944	0.052	-
		F-race	over	93	48%			
Forest	RF	best CV	balanceCascade	180	-	0.911	0.012	-
		F-race	balanceCascade	60	67%			
	SVM	best CV	ENN	180	-	0.809	0.011	-
		F-race	ENN	64	64%			
Letter	RF	best CV	balanceCascade	180	-	0.981	0.010	-
		F-race	balanceCascade	73	59%			
	SVM	best CV	over	180	-	0.953	0.022	-
		F-race	over	44	76%			
Nursery	RF	best CV	SMOTE	180	-	0.809	0.047	-
		F-race	SMOTE	76	58%			
	SVM	best CV	over	180	-	0.875	0.052	-
		F-race	over	58	68%			
Pima	RF	best CV	under	180	-	0.691	0.045	-
		F-race	under	136	24%			
	SVM	best CV	EasyEnsemble	180	-	0.675	0.071	0.107
		F-race	costBalance	110	39%	0.674	0.06	
Satimage	RF	best CV	balanceCascade	180	-	0.719	0.033	-
		F-race	balanceCascade	132	27%			
	SVM	best CV	balanceCascade	180	-	0.662	0.044	-
		F-race	balanceCascade	90	50%			
Spam	RF	best CV	SMOTE	180	-	0.942	0.015	-
		F-race	SMOTE	122	32%			
	SVM	best CV	SMOTEnsemble	180	-	0.917	0.018	0.266
		F-race	SMOTE	135	25%	0.918	0.02	
Women	RF	best CV	TomekUnder	180	-	0.488	0.051	-
		F-race	TomekUnder	150	17%			
	SVM	best CV	EnnSmote	180	-	0.492	0.073	-
		F-race	EnnSmote	102	43%			

Conclusion

- Class unbalanced problem is well known, different techniques and metrics have been proposed.
- 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.
- F-race is crucial in adapting the strategy with fraud evolution.

Future work

1. Release of an R package for unbalanced dataset
2. Adopt Racing for incremental learning / data streams

Acknowledgment
The work of Andrea Dal Pozzolo was supported by the Doctiris programe of Innoviris (Brussels Institute for Research and Innovation), Belgium.

F-race Vs Cross Validation II

- For almost all datasets F-race is able to return the best method according to the cross validation (CV) assessment.
- In Pima and Spam datasets F-race returns a sub-optimal strategy that is not significantly worse than the best (Pvalue greater than 0.05).
- The *Gain* column shows the computational gain (in percentage of the the CV tests) obtained by using F-race.
- Apart from the Breast Cancer dataset in all the other cases F-race allows a significant computational saving with no loss in performance.

UCI BreastCancer - Fmeasure

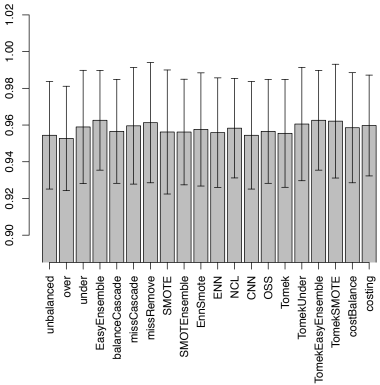
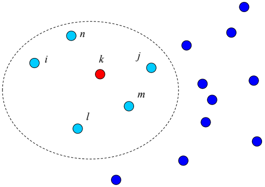
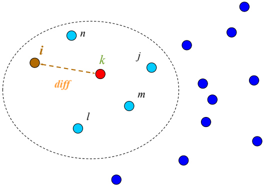


Figure: Comparison of techniques for unbalanced data with UCI Breast Cancer dataset and Random Forest classifier in terms of Fmeasure.

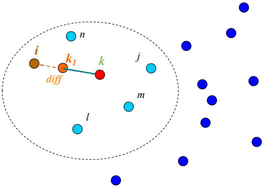
SMOTE, R package [16]



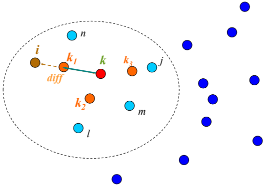
1. For each minority example k compute nearest minority class examples (i, j, l, n, m)



2. Randomly choose an example out of 5 closest points



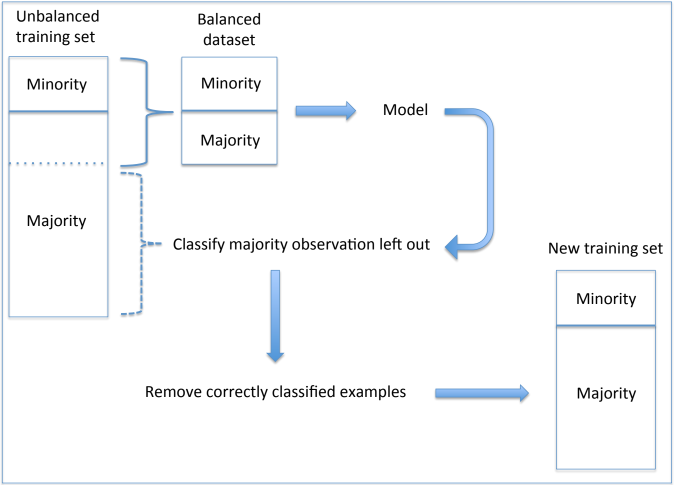
3. Synthetically generate event k_1, such that k_1 lies between k and i



4. Dataset after applying SMOTE 3 times

Balance Cascade [11]

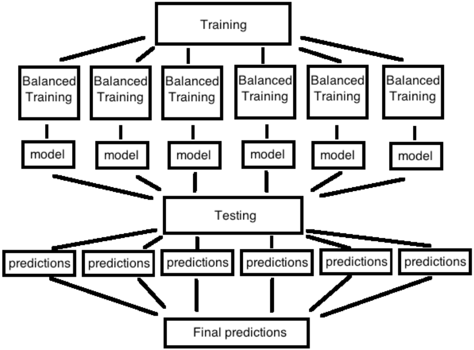
BalanceCascade, explore the majority class in a supervised manner:



Keep removing majority class examples until none is miss-classified

Easy Ensemble [11]

EasyEnsemble, learns different aspects of the original majority class in an unsupervised manner:



Cost proportional sampling [6]

- Positive and negative examples sample by the ratio:

$$p(1)FNcost : p(0)FPcost$$

  where $p(1)$ and $p(0)$ are prior class probability.
- Proportional sampling with replacement produces duplicated cases with the risk of overfitting

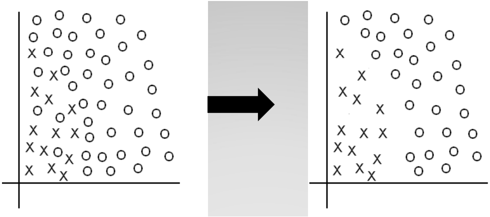
Costing [19]

Use rejection sampling to avoid duplication of instances:

1. Each instance in the original training set is drawn once
2. Accept an instance into the sample with the accepting probability C(i)/Z.
   - C(i) is the misclassification cost of class i, and Z is an arbitrary constant such that $Z \geq max C(i)$.
   - If Z = max C(i), this is equivalent to keeping all examples of the rare class, and sampling the majority class without replacement according to FPcost/ FNcost

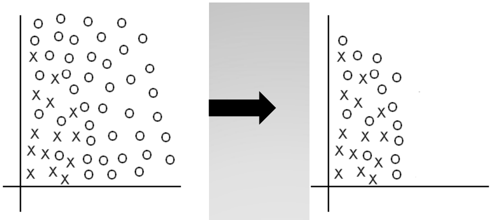
Tomek link [15]

Goal is to remove both noise and borderline examples.



Condensed Nearest Neighbor (CNN) [7]

Goal is to eliminate the instances from the majority class that are distant from the decision border, considered less relevant for learning.



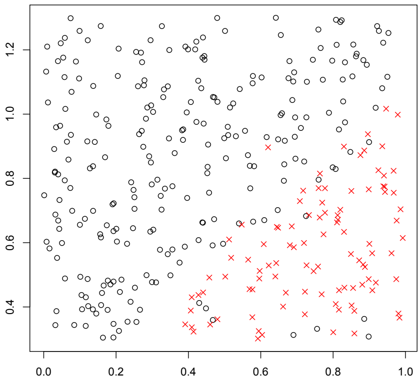
One Side Selection [9]

Hybrid method obtained from Tomek link and CNN:

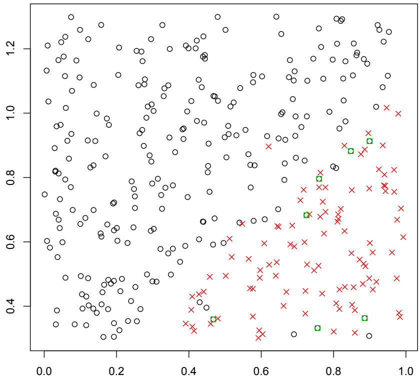
- Apply first Tomek link and then CNN
- Major drawback is the use of CNN which is sensitive to noise[18], since noisy examples are likely to be misclassified. Many of them will be added to the training.

Edited Nearest Neighbor [17]

If an instance belongs to the majority class and the classification given by its three nearest neighbours contradicts the original class, then it is removed



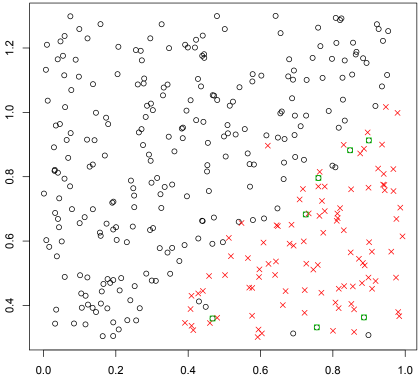
1. Before ENN



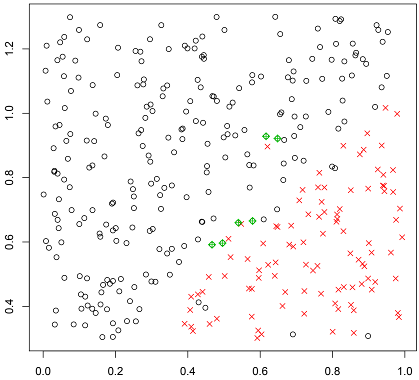
2. Majority class removed with ENN

Neighborhood Cleaning Rule [10]

Apply ENN first and If an instance belongs to the minority class and its three nearest neighbours misclassify it, then the nearest neighbours that belong to the majority class are removed.



1. Apply ENN



2. Majority class removed after ENN

D.J. Newman A. Asuncion.
UCI machine learning repository, 2007.

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.

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.

P. Clark and T. Niblett.
The cn2 induction algorithm.
Machine learning, 3(4):261–283, 1989.

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

C. Elkan.
The foundations of cost-sensitive learning.

In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Citeseer, 2001.

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

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

M. Kubat, S. Matwin, et al.
Addressing the curse of imbalanced training sets: one-sided selection.
In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 179–186. MORGAN KAUFMANN PUBLISHERS, INC., 1997.

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

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.

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

L.B.J.H.F.R.A. Olshen and C.J. Stone.
Classification and regression trees.
Wadsworth International Group, 1984.

J.R. Quinlan.
C4. 5: programs for machine learning, volume 1.
Morgan kaufmann, 1993.

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

L. Torgo.
Data Mining with R, learning with case studies.

Chapman and Hall/CRC, 2010.

D.L. Wilson.
Asymptotic properties of nearest neighbor rules using edited data.
Systems, Man and Cybernetics, IEEE Transactions on, (3):408–421, 1972.

D.R. Wilson and T.R. Martinez.
Reduction techniques for instance-based learning algorithms.
Machine learning, 38(3):257–286, 2000.

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.