Data mining for imbalanced data: Improving classifiers by selective pre-processing of examples

JERZY STEFANOWSKI
co-operation Szymon Wilk*

Institute of Computing Sciences,
Poznań University of Technology
* also with University of Ottawa

COST Doctoral School, Troina 2008

Outline of the presentation

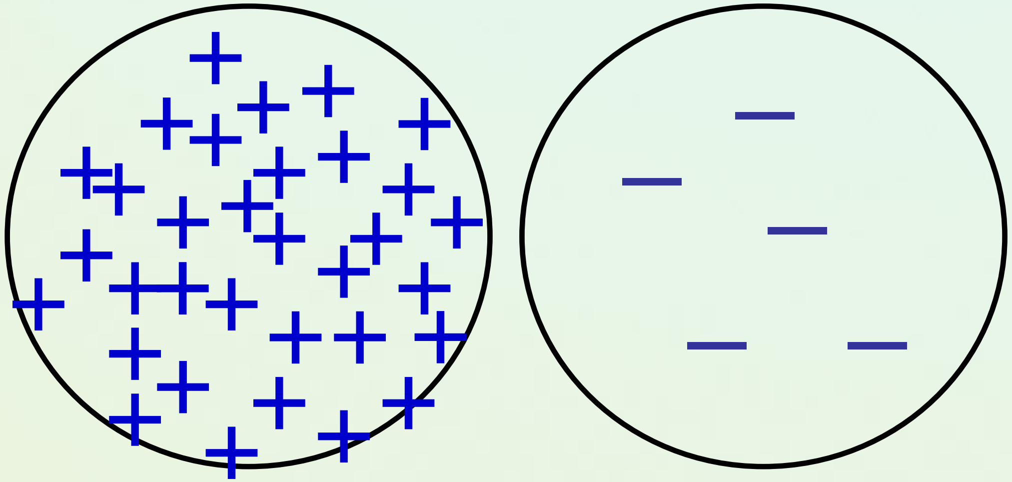
1. Introduction
2. Performance measures
3. Related works
4. Changing classifiers for the minority class
5. Re-sampling strategies
6. Experiments
7. Conclusions

Introduction

- Let us ask a question about class distribution in the input data?
- Standard assumption for discovering classification knowledge from data:
  - The data sets should be balanced: i.e., there are as many positive examples of the concept (class) as for other (concepts).
  - <u>*Example:*</u> A database of sick and healthy patients contains as many examples of sick patients as it does of healthy ones.

Introduction

- A data set is imbalanced if the classes are not approximately equally represented.
  - One class (a minority class) includes much smaller number of examples than other classes.
- Rare examples /class are often of special interest.
- Quite often we are interested in recognizing a particular class
- CLASS IMBALANCE → causes difficulties for learning and decrease the classifier performance.



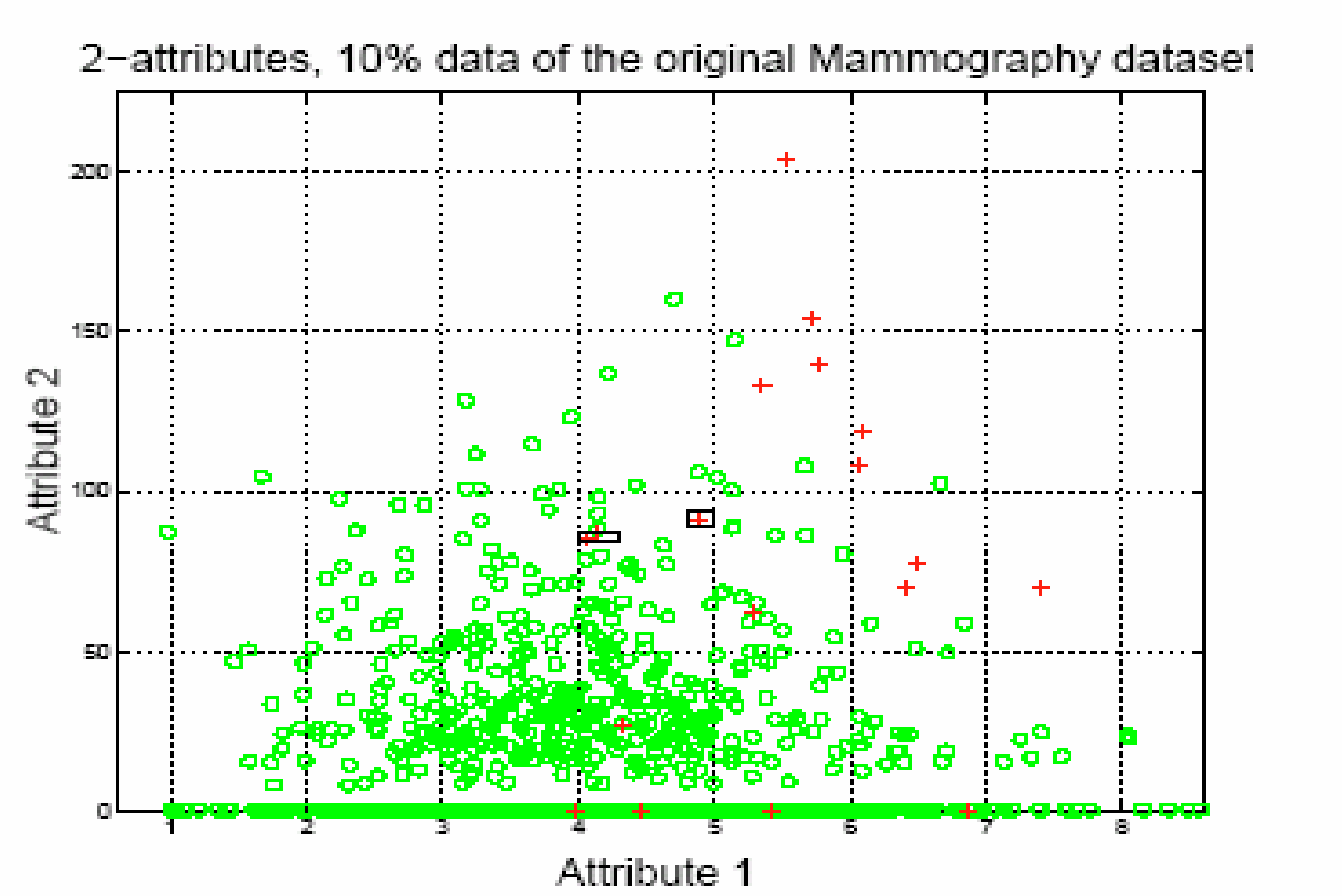
Class imbalance is not the same as COST sensitive learning.
In general cost are unknown!

Typical examples

- There exist many domains that do not have a balanced data set:
  - Medical problems – rare but dangerous illness.
  - Helicopter Gearbox Fault Monitoring
  - Discrimination between Earthquakes and Nuclear Explosions
  - Document Filtering
  - Direct Marketing.
  - Detection of Oil Spills
  - Detection of Fraudulent Telephone Calls
- For more examples see, e.g.

  - Japkowicz N., Learning from imbalanced data. AAAI Conf., 2000.
  - Weiss G.M., Mining with rarity: a unifying framework. ACM Newsletter,2004.
  - Chawla N., Data mining for imbalanced datasets: an overview. In The Data mining and knowledge discovery handbook, Springer 2005.

Example from Chawla et al. SMOTE 2002



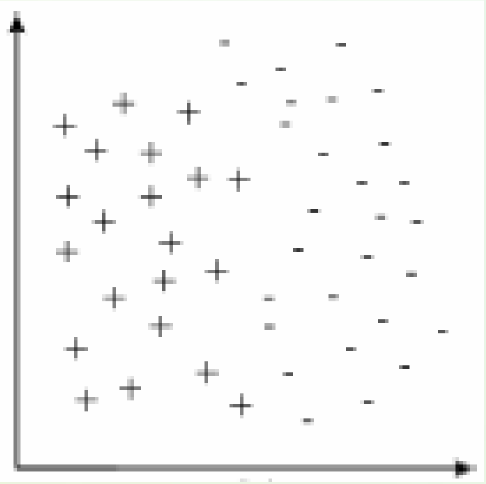
Difficulties for inducing classifiers

- Many learning algorithms $\rightarrow$ assuming that data sets are balanced.
- The standard classifiers are biased
  - Focus search no more frequent classes,…
  - Toward recognition of majority classes and have difficulties to classify new objects from minority class.
- An example of information retrieval system (Lewis and Catlett 1994)
  highly imbalanced (~ 1%)
  $\rightarrow$ total accuracy ~100%
  but fails to recognize the important (minority) class.

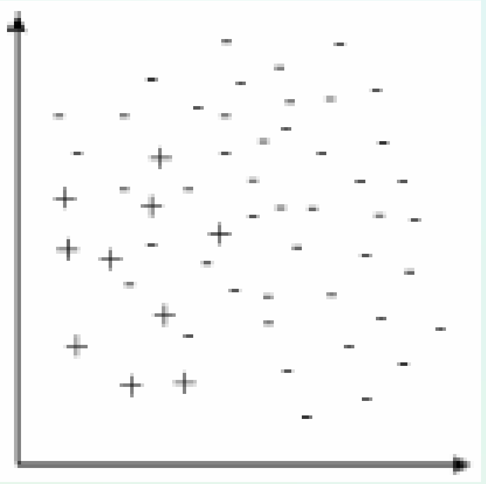
Growing research interest in mining imbalanced data

- Although the problem known in real applications, it received attention from machine learning and data mining community in the last decade.
- A number of workshops:
  - AAAI’2000 Workshop, org:R. Holte, N. Japkowicz, C. Ling, S. Matwin.
  - ICML’2000 Worshop also on cost sensitive. Dietterich T. et al.
  - ICML’2003 Workshop, org.: N. Chawla, N. Japkowicz, A. Kolcz.
  - ECAI 2004 Workshop, org.: Ferri C., Flach P., Orallo J. Lachice. N.
- Special issues:
  - ACM KDDSIGMOD Explorations Newsletter, editors: N. Chawla, N. Japkowicz, A. Kolcz.

Imbalance – why is it difficult?



An easier problem



More difficult one

Some of sources of difficulties:

- Lack of data,
- Imbalance ratio,
- Small disjuncts,
- …

Majority classes overlaps the minority class:

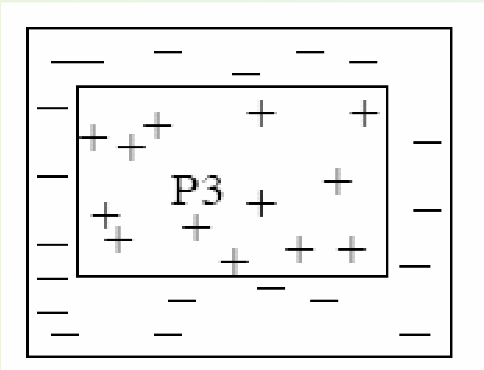
- ❑ Ambiguous boundary between classes
- ❑ Influence of noisy examples

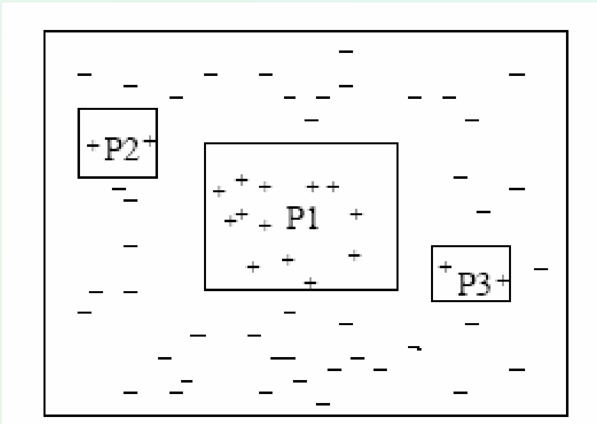
Some review studies, e.g:

- Japkowicz N., Learning from imbalanced data. AAAI Conf., 2000.
- Weiss G.M., Mining with rarity: a unifying framework. ACM Newsletter,2004.

Is always „imbalance” data difficult one?

- See some papers by N.Japkowicz or G.Weiss.
- The minority class contains small „disjuncts” – sub-clusters of interesting examples surrounded by other examples.



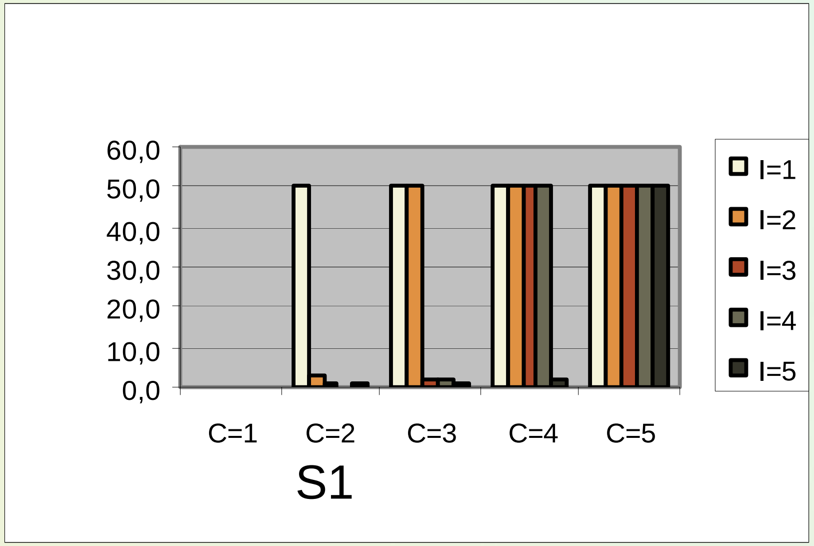


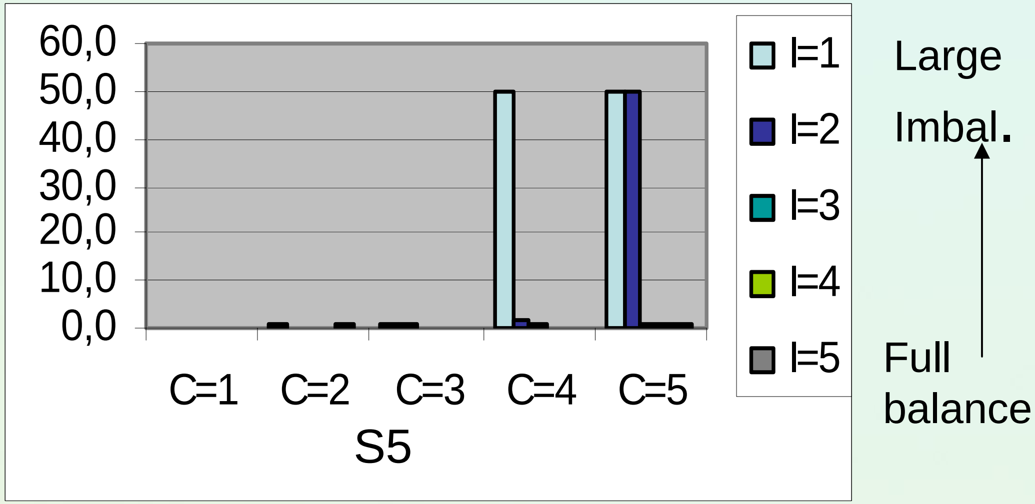
Some review studies, e.g:

- Japkowicz N., Learning from imbalanced data. AAAI Conf., 2000.
- Weiss G.M., Mining with rarity: a unifying framework. ACM Newsletter,2004.

Conclusions from N.Japkowicz study on artifical data

- Large experimental study 125 artificial, each representing a different type of class imbalance, by varying the concept complexity (C), the size of the training set (S) and the degree of imbalance (I) at different rates.
  - C5.0, MLP and SVM classifier were compared.





The *class imbalance problem* depends on

the **degree** of class **imbalance**;

the **complexity of the concept** represented by the data;

the overall **size of the training set**;

the **classifier** involved.

Imbalance – Evaluation measures

- Evaluation of classification performance
  - Standard total accuracy is not useful.
- Performance for the minority class
  - Analysis of binary confusion matrix
  - Sensitivity and specificity,
  - ROC curve analysis.

		Predicted class	
		Yes	No
Actual class	Yes	TP: True positive	FN: False negative
	No	FP: False positive	TN: True negative

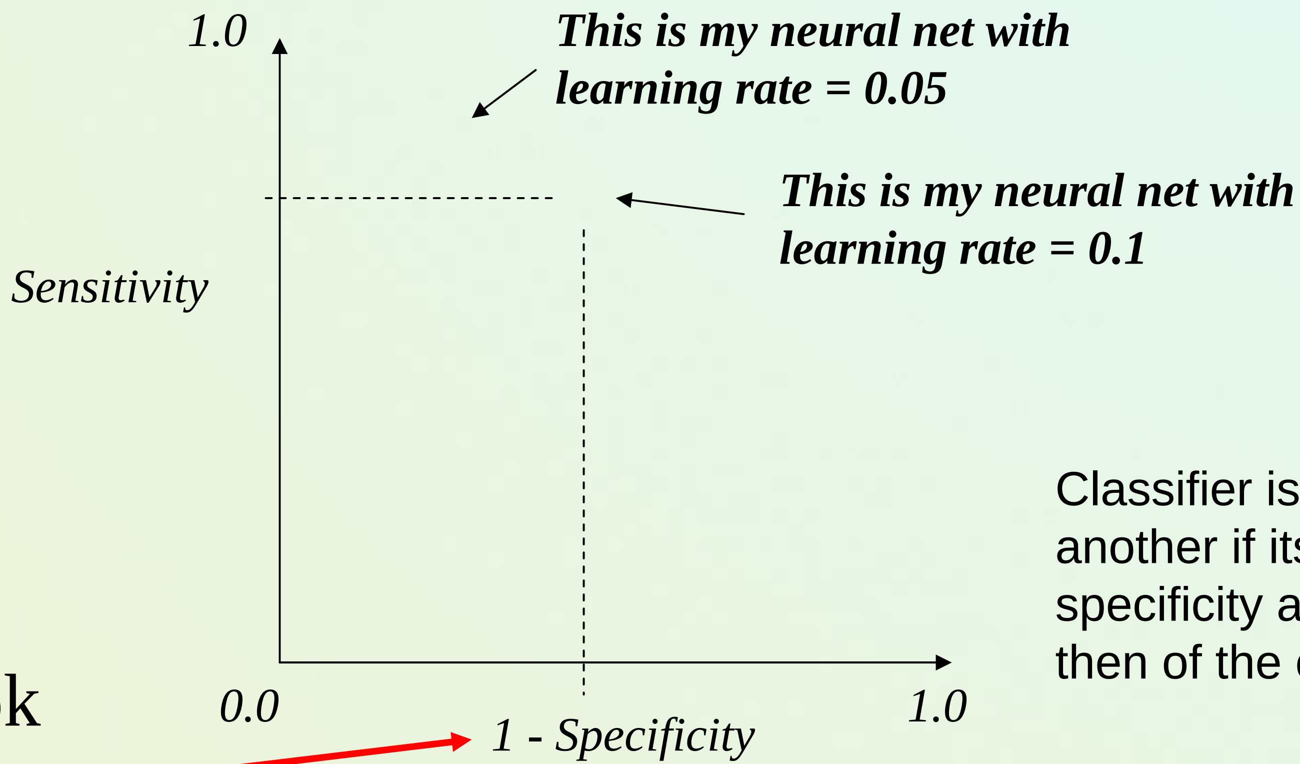
$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

ROC Analysis

“Receiver-Operator Characteristics” – used by mathematicians to analyse radar data. Applied in signal detection to show tradeoff between **hit rate** and **false alarm rate** over noisy channel.

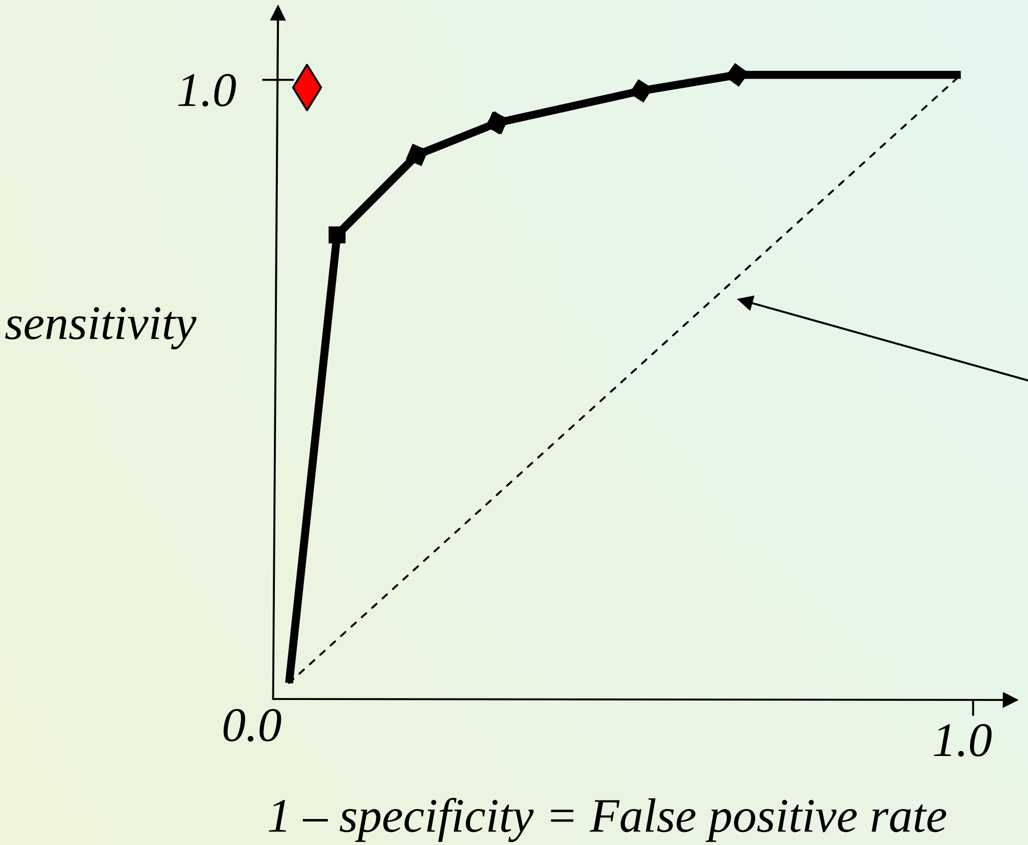
A ROC curve displays a relation between sensitivity and specificity for a given classifier (binary problems, parameterized classifier or a score classification)



Classifier is better than another if its sensitivity and specificity are both not less then of the other classifier’s.

Look close!

ROC Analysis



Ideal point .

Diagonal line – random guessing / boundary for useful classifiers

You can compare performance of several classifiers.
Quite often AUC – area under curve – is calculated.

Related works

- Review survey, e.g.,
  - Weiss G.M., Mining with rarity: a unifying framework. ACM Newsletter, 2004.
- Main approaches to deal with imbalance of data:
  - Re-sampling or re-weighting,
  - Changing search strategies in learning, use another measures,
  - Adjusting classification strategies,
  - One-class-learning
  - Using hybrid and combined approaches (boosting like re-weighing)
  - …
- Our interest in research on:
  - Modification of original data by changing the class distribution.
  - Modification of algorithms for constructing rule-based classifiers.

More on related works

Changing search or classification strategies

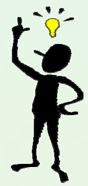
- Typical rule or tree induction:
  - Exploit a greedy search strategy and use criteria that favor the majority class.
    - The majority class rules are more general and cover more examples (strength) than minority class rules.
- Some proposals to avoid it:
  - Use another inductive bias
    - Modification of CNx to prevent small disjuncts (Holte et al.)
    - Hybrid approach with different „inductive bias” between large and small sets of examples (Ting).
  - Use less greedy search for rules
    - Exhaustive depth-bounded search for accurate conjunctions. Brute (Riddle et al..), modification of Apriori like algorithm to handle multiple levels of support (Liu at al.)
    - Specific genetic search – more powerful global search (Freitas and Lavington, Weiss et al.) …

Changing rule classification strategy

- Rules from majority classes are usually more general, stronger and shorter then these from the minority class.
- While classifying an unseen case, rules matching it and voting for the minority class are outvoted by rules voting for bigger classes.
- Grzymała proposal (2000) $\rightarrow$ leave the rule induction but change the classification strategy!
- Changing strength of rules for the minority class by an extra multiplier, while not changing the strength of rules from the secondary classes.
  - Optimization of strength multiplier by maximizing a measure gain = sensitivity + specificity –1.

Changing set of rules for the minority class

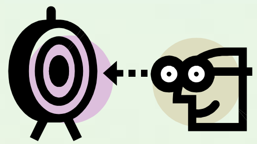
- Minority class rules have smaller chance to predict classification for new objects!
- Two stage approach (Stefanowski, Wilk):
  1. Induce minimal set of rules for all classes.
  2. Replace the set of rules for the minority class by another set $\rightarrow$ more numerous and with greater strength.
- The chance of using these rule while classifying new objects is increased.
- The use of EXPLORE (Stefanowski, Vanderpooten):
  - Induce all rules with strength greater then a threshold.
  - Modify the threshold considering gain + conditions calculated from 1 stage.



Comparison of different approaches (sensitivity)

Data set	Standard classifier	Strength multiplier	**Replace rules**
Abdominal	0.584	0.772	**0.834**
Bupa	0.324	0.365	**0.427**
Breast	0.364	**0.482**	0.471
German	0.378	0.617	**0.627**
Hepatitis	0.437	0.738	**0.753**
…	…	…	…
Pima	0.3918	0.587	**0.687**
Urology	0.1218	0.361	**0.717**

The rule replacing and strength multiplier approaches significantly outperformed the standard rule classifier, considering the sensitivity and gain measures without decreasing the total accuracy.

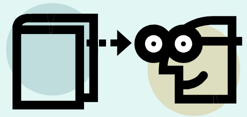


Motivations for other approach to imbalance data

- The „replace rules” approach is focused on handling „cardinality” aspects of imbalance.
  - Strengthening some sub-regions and leaving uncovered examples.
  - Some difficult examples may be uncovered depending on the procedure for tuning parameters
    - which is time consuming and sophisticated.

- However, one may focus on other characteristics of learning examples, as discussed earlier.

Related works on pre-processing of imbalanced data



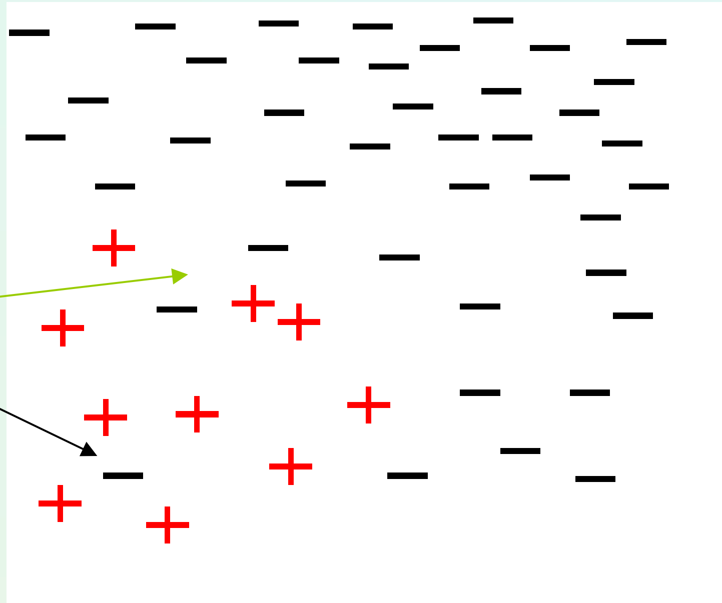
- Transforming the original class distribution into more balanced one:
  - Random sampling
    - Over-sampling
    - Under-sampling
  - Focused transformation
    - Modification of majority classes (safe, borderline, noisy, …)
      - One-side-sampling (Kubat, Matwin)
      - Laurikkala's edited nearest neighbor rule
    - Focused over-sampling
      - SMOTE → Chawla et al.

- Some reviews:
  - Weiss G.M., Mining with rarity: a unifying framework. ACM Newsletter, 2004 → A comprehensive review study.
  - Batista et al. → a study of behavior of several methods for balancing machine learning training data, 2004.

Difficult learning examples in imbalance data

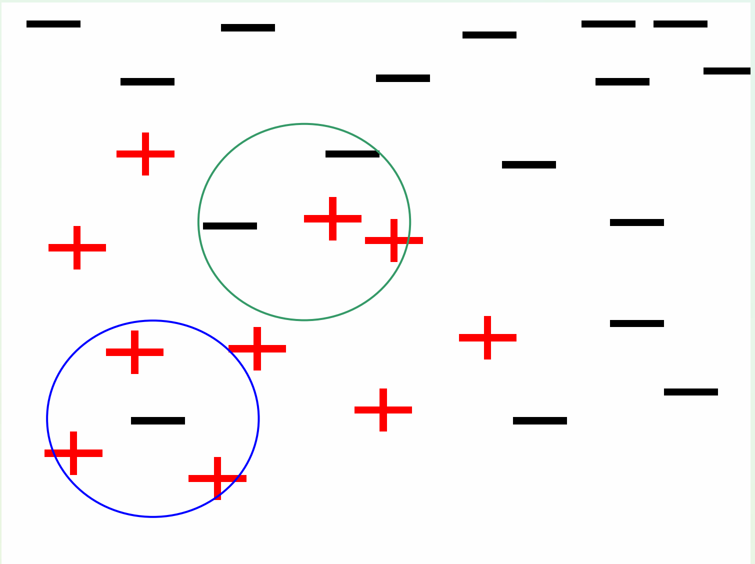
- Consider the following majority class examples:
  - Noisy examples,
  - Borderline ones.
- They may lead to misclassification of the minority ones.

- How could we handle such information?



The general idea

- Detect and remove such majority noisy and borderline examples in filtering before inducing the classifier.
- Based on the idea of Wilson's Edited Nearest Neighbor Rule $\rightarrow$ Remove these examples whose class labels differ from the class of its three nearest neighbors.
- Two phases
  - Noisy examples
  - Boundary region



Filtering Approach:

1. Split a learning set *E* into a minority class *C* and the rest of data *R*.
2. Identify noisy majority class examples from *R*:

   $\forall e_i \in R$ check if the classification given by its 3 NN contradicts its class, then add e_i to the set A_1.
3. For each $c_i \in C$: if its nearest neighbors misclassify it, then these neighbors that belong to the majority class are added to the set A_2.
4. Remove from E these majority class examples that belong to $\{A_1 \cup A_2\}$.

Use of 3-NN algorithm with a heterogeneous value distance metric:
A component distance of nominal attributes $\rightarrow$ value difference metric by Stanfill and Waltz.

Experiments

- Aims → To verify the usefulness of the new filtering approach comparing it against:
    - the standard classifier without any filtering,
    - the classifier with simple random under-sampling,
    - the classifier with simple random over-sampling.
- Conditions of experiments:
    - Rules induced by MODLEM algorithm (generalized conditions; entropy search criterion; missing attribute values),
    - Evaluation measures: sensitivity, specificity, total error,
    - 10-fold stratified cross validation.
- Data sets
    - UCI repository benchmark data and other difficult medical data.

Classification performance - Sensitivity

Data set	Standard classifier	Under-sampling	Over-sampling	**New filtering**
breast ca	0.3056	0.5971	0.4043	**0.6264**
bupa	0.7290	0.6707	0.5935	**0.8767**
ecoli	0.4167	**0.8208**	0.5150	0.7750
pima	0.4962	0.7093	0.5519	**0.8098**
Acl	0.7250	0.8485	0.7840	**0.8750**
…	…	…	…	…
Wisconsin	0.9083	0.9521	0.8326	**0.9625**
hepatitis	0.4833	**0.7372**	0.5447	0.6500

New approach improves Sensitivity for nearly all data sets, under-sampling is the second.

SMOTE - Synthetic Minority Oversampling Technique

- Technique designed by Chawla, Hall, Kegelmeyer 2002
- For each minority Sample
  - Find its k-nearest minority neighbours
  - Randomly select j of these neighbours
  - Randomly generate synthetic samples along the lines joining the minority sample and its j selected neighbours

  (j depends on the amount of oversampling desired)

- Comparing to simple random oversampling – for SMOTE larger and less specific regions are learned, thus, paying attention to minority class samples without causing overfitting.
- SMOTE currently yields the best results as far as re-sampling and combination with undersampling go (Chawla, 2003).

SMOTE - performance

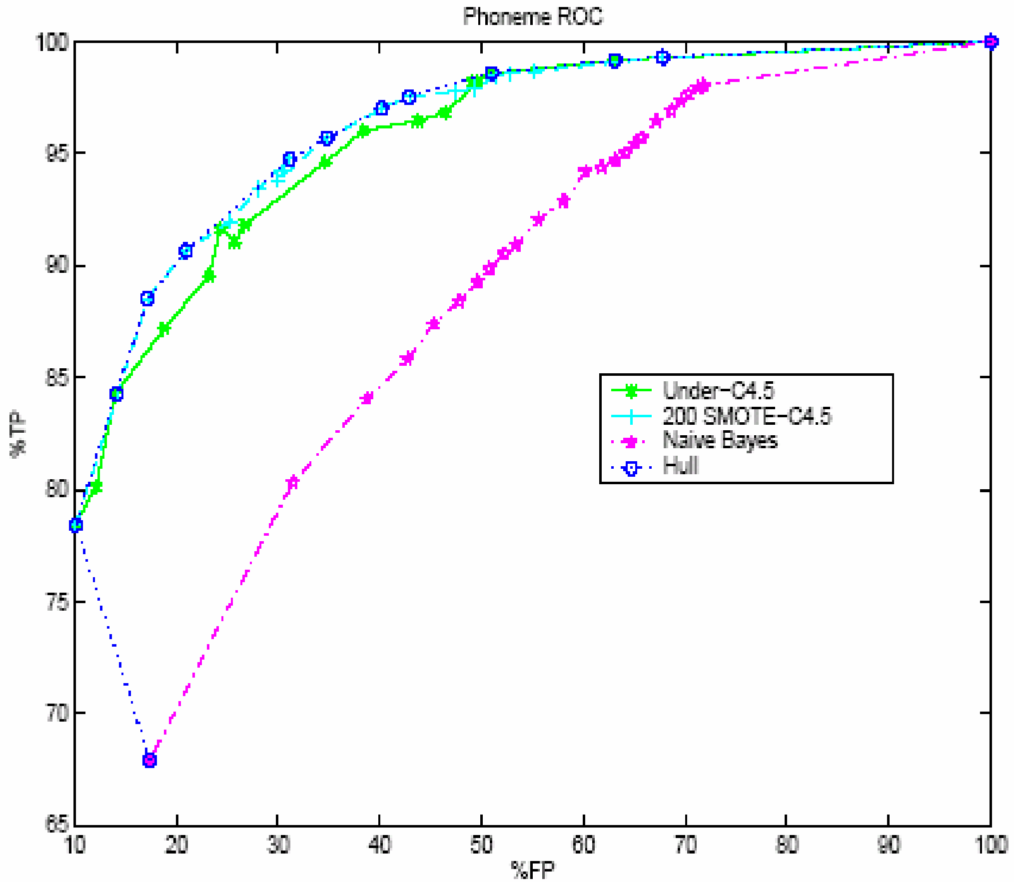


Figure 7: Phoneme. Comparison of SMOTE-C4.5, Under-C4.5, and Naive Bayes. SMOTE-C4.5 dominates over Naive Bayes and Under-C4.5 in the ROC space. SMOTE-C4.5 classifiers are potentially optimal classifiers.

SMOTE performance for different data sets

Dataset	Under	50 SMOTE	100 SMOTE	200 SMOTE	300 SMOTE	400 SMOTE	500 SMOTE
Pima	7242		**7307**				
Phoneme	8622		8644	**8661**			
Satimage	8900		8957	**8979**	8963	8975	8960
Forest Cover	9807		9832	9834	**9849**	9841	9842
Oil	8524		8523	8368	8161	8339	**8537**
Mammography	9260		9250	9265	9311	**9330**	9304
E-state	6811		6792	**6828**	6784	6788	6779
Can	9535	**9560**	9505	9505	9494	9472	9470

Table 3: AUC's [C4.5 as the base classifier] with the best highlighted in bold.

However, critical remarks on related methods

- NCR and one-side-sampling
  - Greedy removing (too) many exampled from the majority class.
  - Focused on improving sensitivity of the minority class.
  - However, it may deteriorate the recognition of examples from other (majority) classes (decreasing specificity and total accuracy).
- SMOTE
  - Introduces too many random examples from the minority class not allowing for any flexibility in the re-balancing rate.
  - SMOTE's procedure is inherently dangerous since it blindly generalizes the minority area without regard to the majority class.
  - Problematic in the case of highly skewed class distributions since, in such cases, the minority class is very sparse with respect to the majority class, thus resulting in a greater chance of class mixture.
  - Random objects may be difficult to interpret in some domains → our experience in medicine.

Aims of the study by J.Stefanowski, S.Wilk (ECML/PKDD 2007)

- To introduce a new method for selective pre-processing of imbalance data that:
  - Aims at improving sensitivity for the minority class while preserving the ability of a classifier to recognize the majority class,
  - Keeps overall accuracy at an acceptable level,
  - Does not introduce any random examples.
- This method could be combined with many classifiers;
- Its usefulness was successfully evaluated in comparative experiments.

Thank you for your attention