Odkrywanie wiedzy klasyfikacyjnej z niezbalansowanych danych

Learning classifiers from imbalanced data

Wykład ZED dostosowany dla SE
party based on PAN PhD Summer School

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska
Poznań

Poznań, grudzien 2011 update 2019

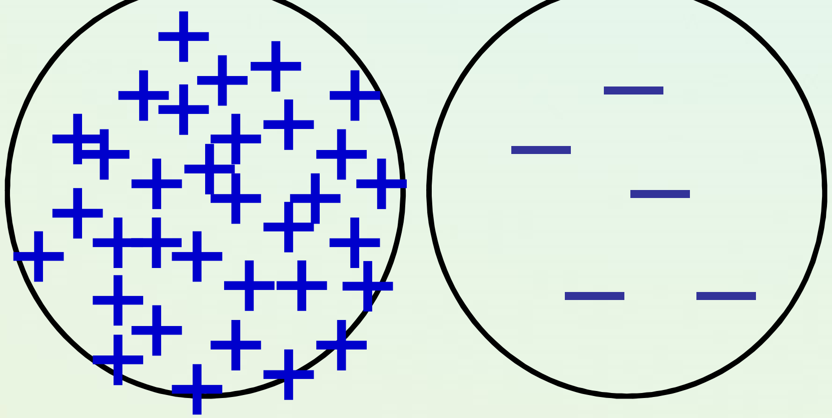
Outline of the presentation

1. Introduction
2. Class imbalance – nature of the problem
3. Types of difficult examples and their influence on learning classifiers
4. Pre-processing - SPIDER
5. Modification of SMOTE based on local neighbourhood
6. Rule-based classifiers – BRACID
7. Ensembles
8. Cost sensitive approach

Imbalanced data

- **Class imbalance** → one (minority) class includes much smaller number of examples than other (majority) classes
  - A minority class is often of primary interest
  - Diagnosing a rare disease
- Typical examples:
  - Medical problems,
  - Technical diagnostics, fault monitoring tasks, prediction of equipment failures, image recognition, fraud detection
  - Text categorization and information retrieval, …

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



More about occurrence of class imbalance

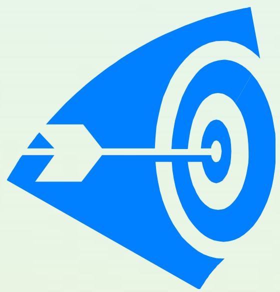
- Literature cases:
  - Medical problems – rare but dangerous illness.
  - Helicopter Gearbox Fault Monitoring
  - Discrimination between Earthquakes and Nuclear Explosions
  - Document Filtering
  - Direct Marketing
  - Detection of Oil Spills
  - Detection of Fraudulent Telephone Calls

- See some reviews:

  - 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.
  - He H, Garcia, Mining imbalanced data. IEEE Trans. Data and Knowledge 2009.

Difficulties for classifiers

- Many learning algorithms → they assume that data sets are balanced
  - there are as many positive examples of the concept (class) as for other (concepts)
- The classifiers are biased
  - Search focused on more frequent classes,...
  - Better recognition of majority classes and difficulties to classify new objects from the minority class
- An example of information retrieval system highly imbalanced (the minority class ~ 1%) → total accuracy ~100%, but fails to recognize the important class



Introduction to Imbalanced Data Sets

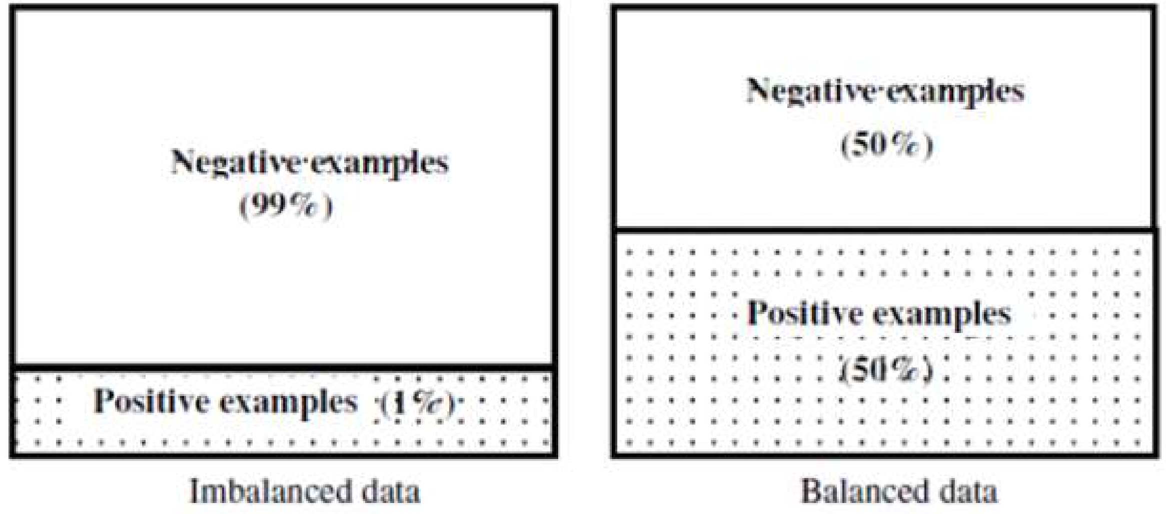
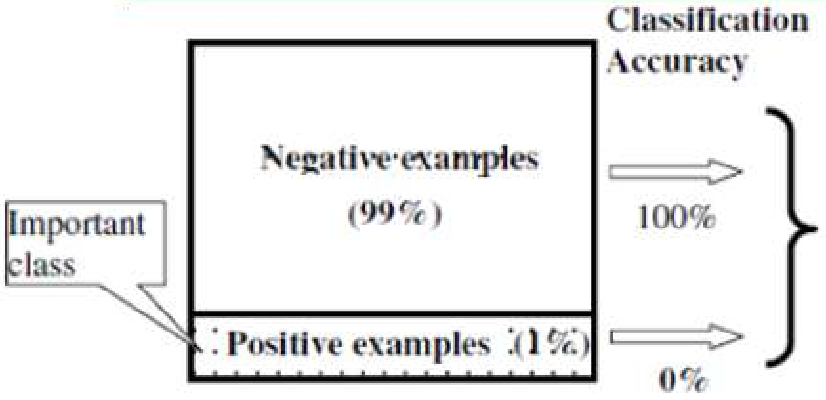


Fig. 1. Imbalanced and balanced data sets.

biased towards the majority class



We need to change the way to evaluate a model performance!

Fig. 2. The illustration of class imbalance problems.

Evaluation issues

- Evaluation of classification performance
  - The standard total accuracy is not useful!
- Performance for the minority class
  - Sensitivity and specificity,
  - ROC curve analysis + AUC

Original	Predicted	
	+	-
+	*TP*	*FN*
-	*FP*	*TN*

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

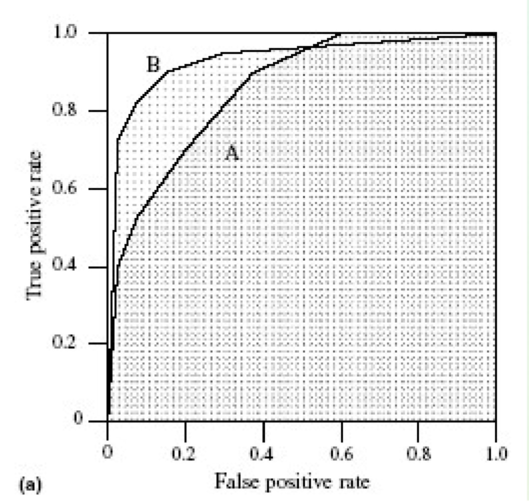
$$Specificity = \frac{TN}{TN + FP}$$

- Aggregated measures:

$$G\text{-}mean = \sqrt{Sensitivity * Specificity}$$

$$F\text{-}measure = \frac{(1+\beta)^2 * Precision * Recall}{\beta^2 * Recall + Precision}$$

$$Precision = \frac{TP}{TP + FP}$$



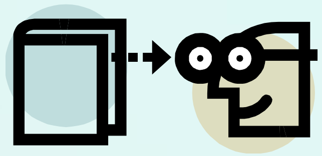
Performance of rule and tree classifiers

Sensitivity for several imbalanced data sets

Data	Modlem rules	C4.5 trees
Acl	0.805	0.855
Breast	0.319	0.387
Bupa	0.520	0.491
Cleveland	0.085	0.237
Ecoli	0.400	0.580
Haberman	0.240	0.410
Hepatitis	0.383	0.432
New-thyr.	0.812	0.922
Pima	0.485	0.601

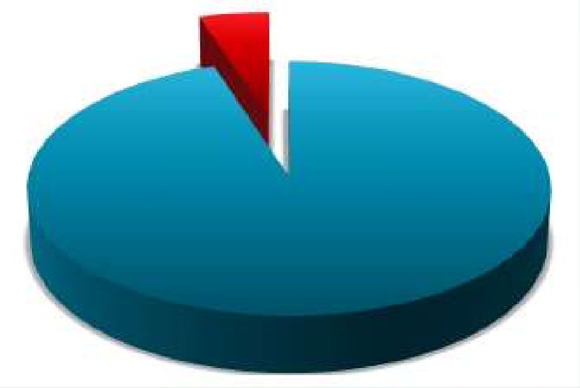
J.Stefanowski, Sz.Wilk. Selective pre-processing of imbalanced data for improving classification performance. DAWAK 2008

Several methods



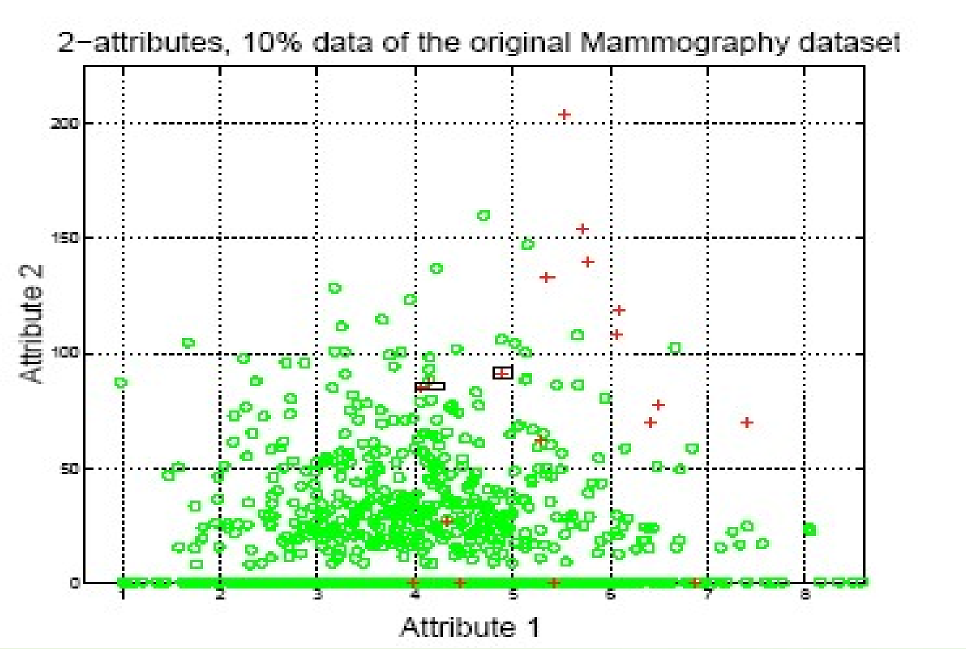
- Some reviews
  - Weiss G.M., Mining with rarity: a unifying framework. ACM Newsletter, 2004.
  - Chawla N., Data mining for imbalanced datasets: an overview. In The Data mining and knowledge discovery handbook, Springer 2005.
  - He H, Garcia, Mining imbalanced data. IEEE Trans. Data and Knowledge 2009.
- General categorization of approaches
  - Data level (preprocessing)
  - Algorithm level
- Different methods
  - Re-sampling or re-weighting,
  - Modify inductive bias, search, evaluation criteria (np. AUC)
  - New classification strategies
  - Ensemble approaches (boosting, bagging or …)
  - Hybrid algorithms
  - One-class-learning
  - Transformation to „cost-sensitive learning”
  - …

Class imbalances – what is it about?



Defining the problem

- ❑ Skewed class distribution
- ❑ Imbalance ratio

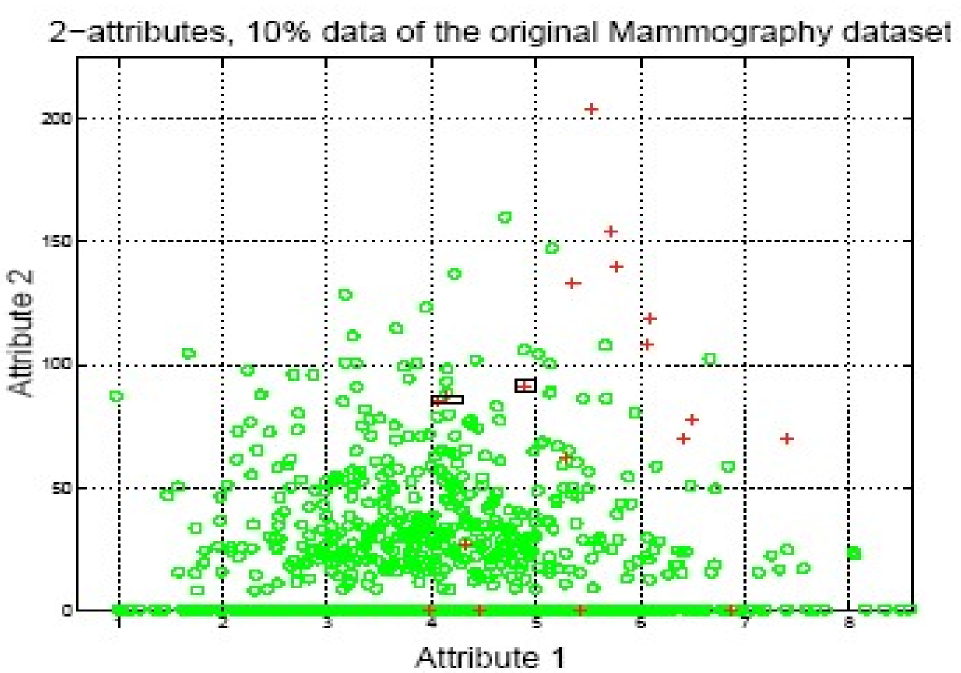


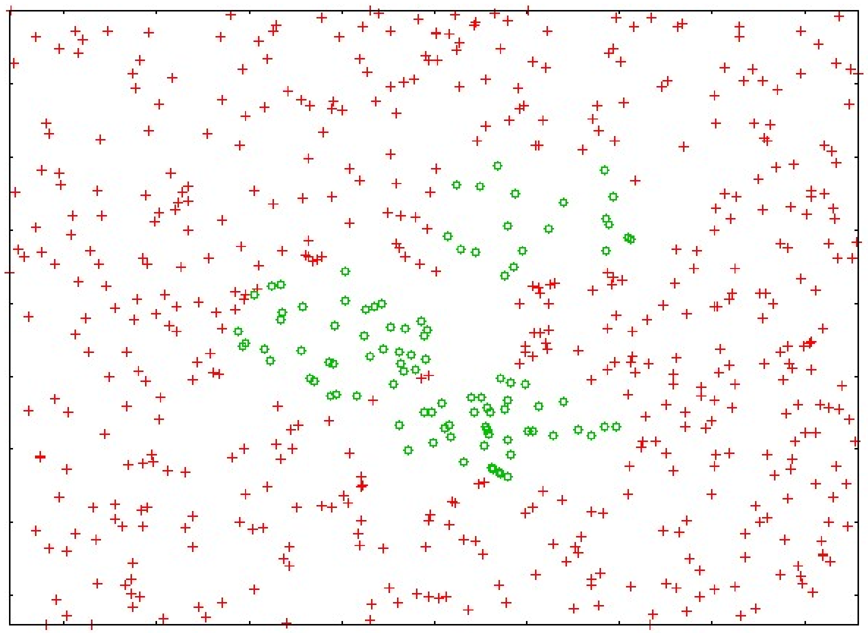
Still many questions

Another point of view

- ❑ Unsatisfactory recognition of the minority class (performance measure)

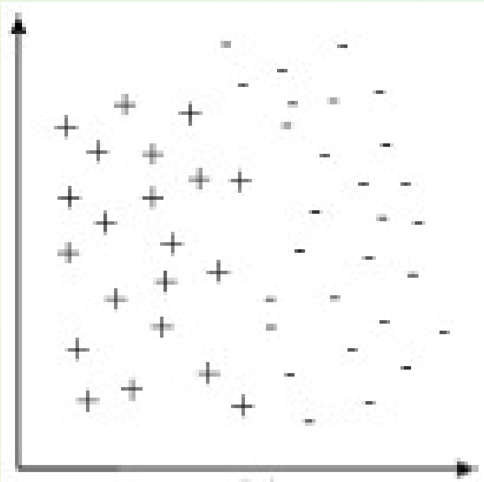
Imbalanced data distributions



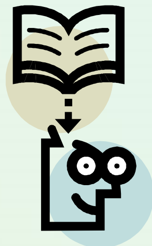


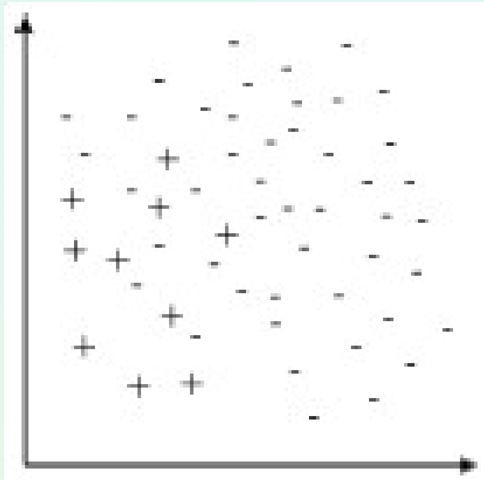
- ❑ The nature of the problem with respect to data distributions
- ❑ Sources of difficulties that deteriorate learning classifiers

Data factors for class imbalance



Imbalance – why is it difficult?





An easier problem

More difficult one

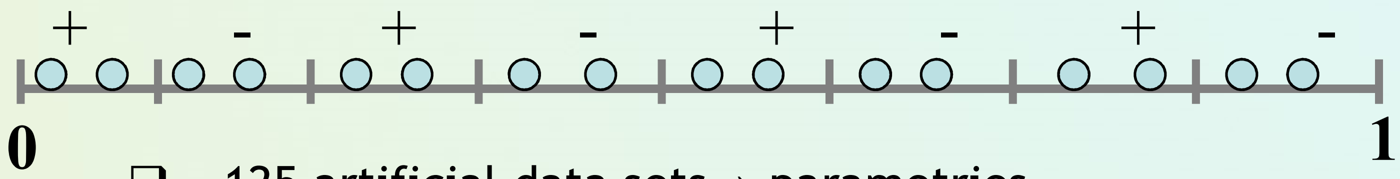
Some of sources of difficulties:

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

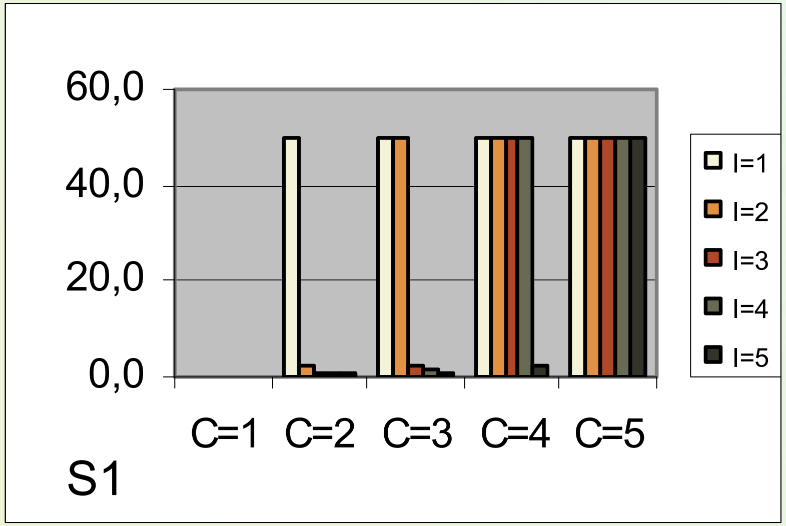
Majority classes overlaps the minority class:

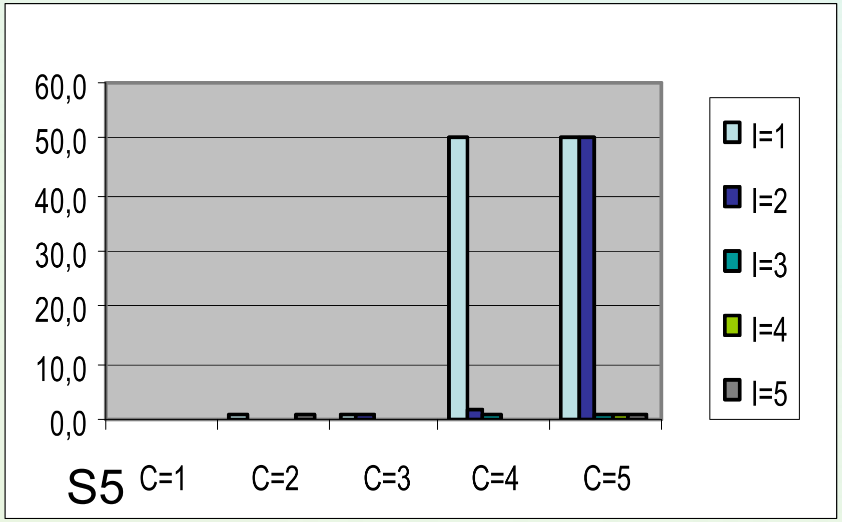
- ❑ Ambiguous boundary between classes
- ❑ Influence of noisy examples

Studies of N.Japkowicz and co-operators



- 125 artificial data sets→ parametries
  - Imbalance ratio (I)
  - Number of examples (S)
  - Concept complexity (C) – sub-concepts
- Algorithms trees C4.5, MLP BP i SVM





Large Imbal.

Full balance

- Deterioration → high complexity with rare examples
- Small disjucts problem (Holte, Porter)

Rare cases and small sub-concepts

Rarity: Rare Cases versus

Rare Classes

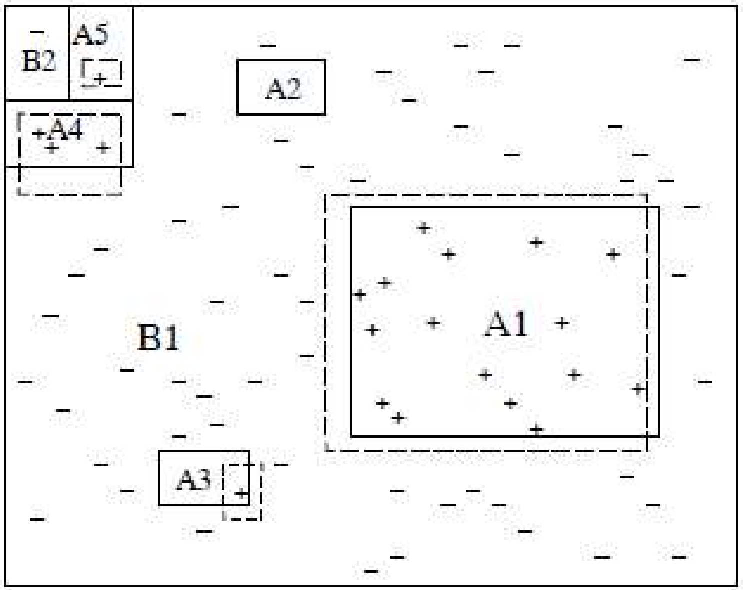


Figure 1: Graphical representation of a rare class and rare case

Class A is the rare (minority class and B is the common (majority class).

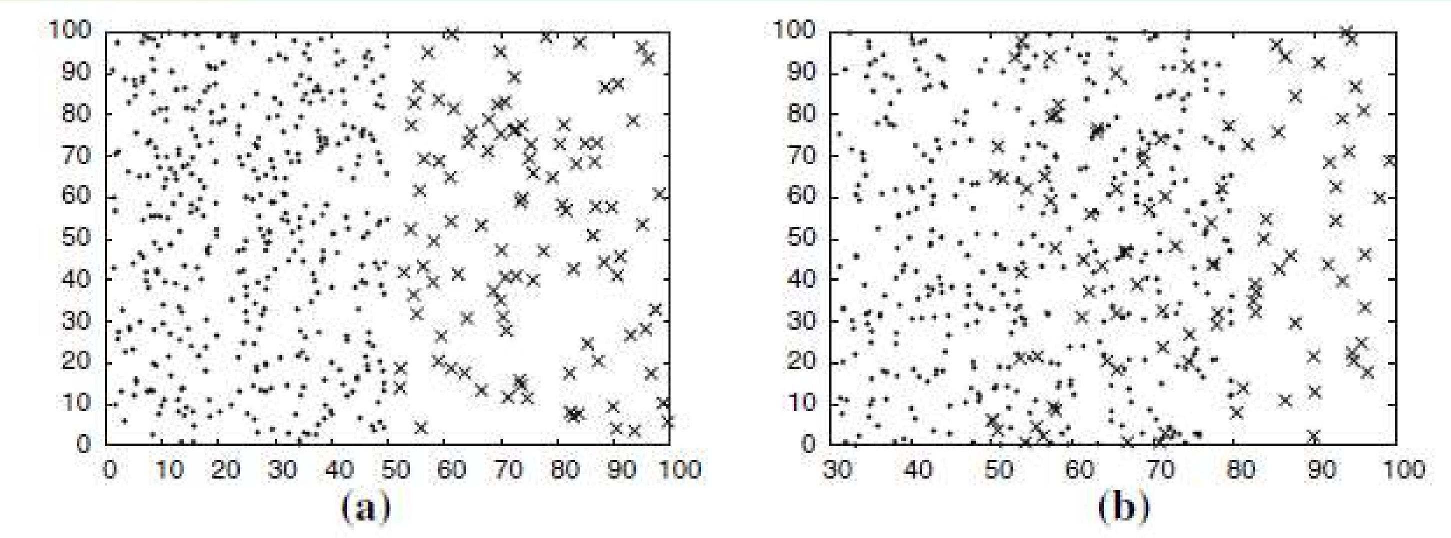
Subconcepts A2-A5 correspond to rare cases, whereas A1 corresponds to a fairly common case, covering a substantial portion of the instance space.

Subconcept B2 corresponds to a rare case, demosnstrating that common classes may contain rare cases.

G.M. Weiss. Mining with Rarity: A Unifying Framework. SIGKDD Explorations 6:1 (2004) 7-19

V. Garcia, J. Sanchez, R. Mollineda: An empirical study of the behaviour of classifiers on imbalanced and overlapped data set, PPRIAA, 2007

Overlapping



Two different levels of class overlapping: a 0% and b 60%

The positive examples are defined on the X-axis in the range [50–100], while those belonging to the majority class are generated in [0–50] for 0% of class overlap, and moves

Overlapping more important than the imbalance ratio + local density

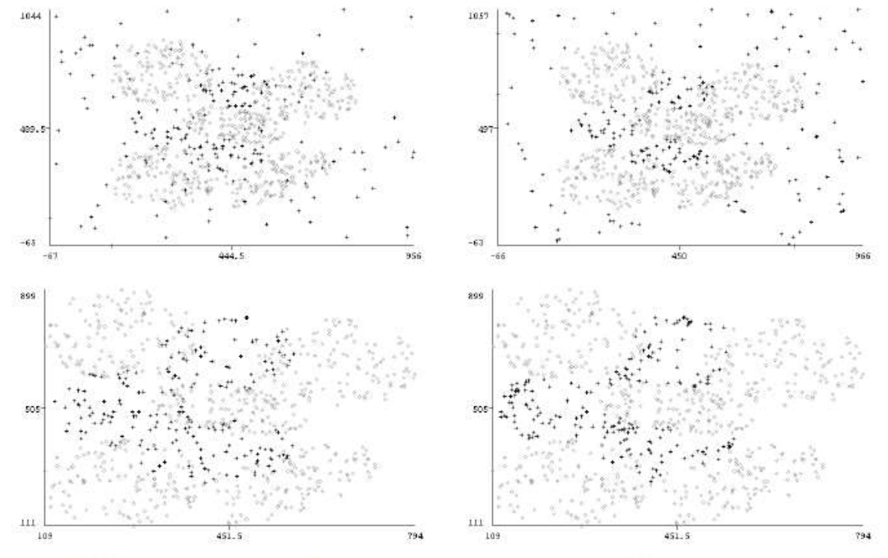
Different behaviours of classifiers

Prati, Batista, Monard 2004 other experiments

Other more complex data

J. Stefanowski, K.Kałużny 2008

- Factors (concept shape, fragmentation into sub-clusters, overlapping i rare examples / noise, imbalance ratio)
- Classifiers C4.5, Ripper and K-NN
- Fragmentation more influential than ratio $\rightarrow$for non-linear
- Overalapping and rare examples decrease the classifier performance



Rysunek 4.9 Zbiory wygenerowane dla przełącznika a (na górze) i b (na dole) parametru n_type oraz dla przełącznika i (po lewej) i o (po prawej) parametru n_transp

Number of subclusters	800				600				400			
	0%	10%	20%	30%	0%	10%	20%	30%	0%	10%	20%	30%
3	0.96	0.84	0.70	0.56	0.94	0.85	0.70	0.55	0.9	0.82	0.7	0.42
4	0.94	0.84	0.68	0.4	0.92	0.82	0.58	0.3	0.89	0.7	0.4	0.34
5	0.9	0.82	0.56	0.36	0.9	0.78	0.52	0.32	0.87	0.68	0.24	0.18
6	0.88	0.64	0.40	0.34	0.85	0.6	0.36	0.3	0.5	0.22	0.14	0.08

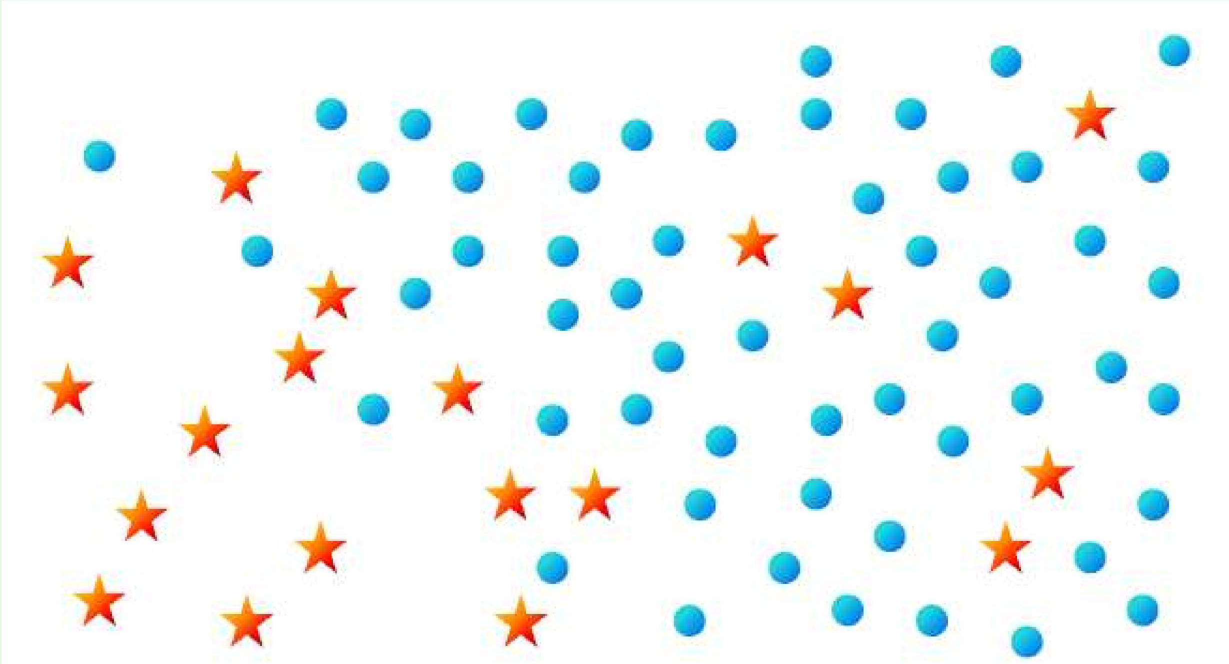
Classifier	TPR				AUC			
	0%	10%	20%	30%	0%	10%	20%	30%
Tree	0.5	0.25	0.10	0.08	0.72	0.62	0.55	0.52
Rules	0.68	0.44	0.38	0.35	0.8	0.72	0.68	0.62
KNN	0.90	0.88	0.72	0.62	0.95	0.92	0.82	0.78

More in:
JS: Overlapping, Rare Examples and Class Decomposition in Learning Classifiers from Imbalanced Data, 2013

Hypothesis on types of examples

- Safe
- Border
- Rare case
- Outliers

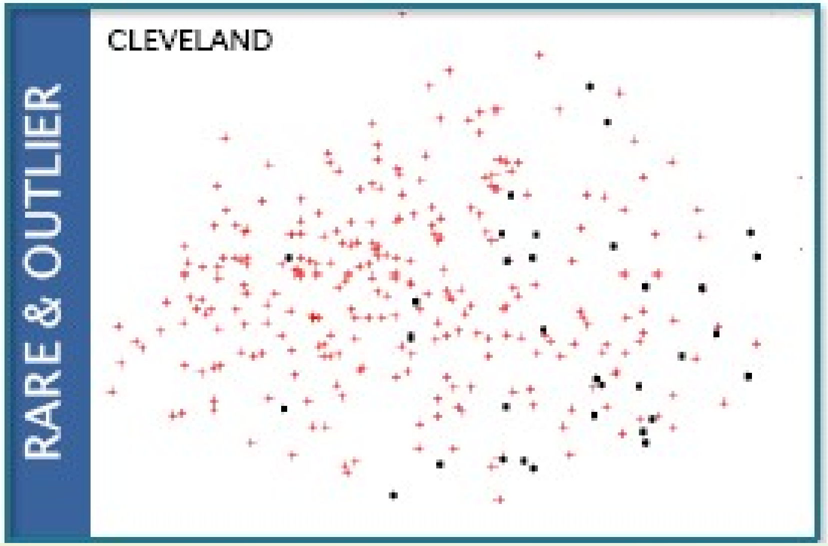
- What is noise?

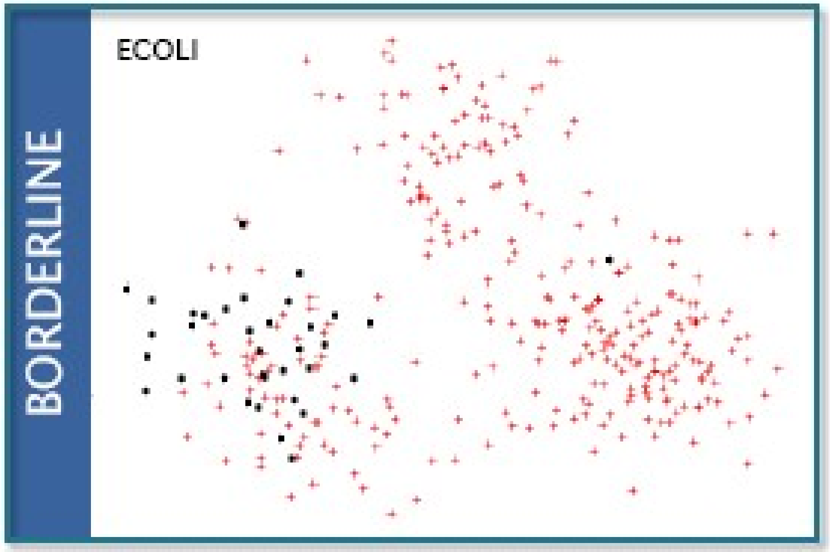


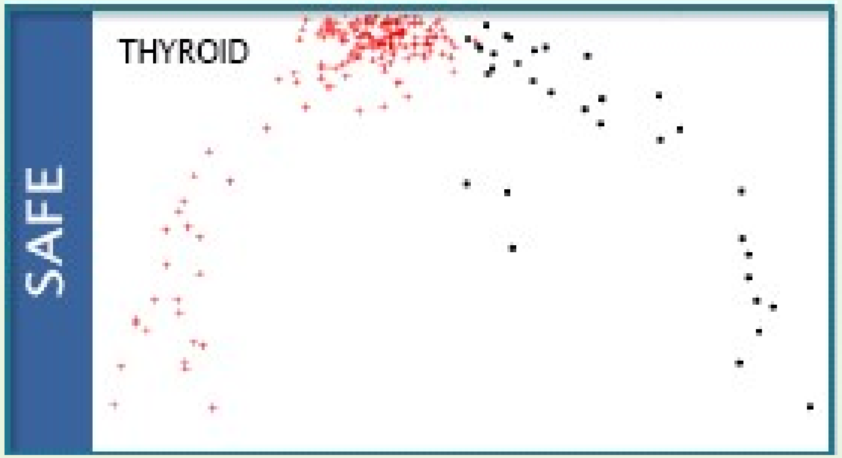
K.Napierała, J.Stefanowski: Identification of Different Types of Minority Class Examples in Imbalanced Data. Proc. HAIS 2012, Part II, LNAI vol. 7209, Springer 2012, 139–150

MDS visualisations of imbalanced data sets

- ❑ Could one notice differences?







K. Napierała, J. Stefanowski, Sz. Wilk: *Learning from imbalanced data in presence of noisy and borderline examples*. RSCTC 2010, LNAI Springer.

- Problem
  - Influence of different examples (safe, border, rare, outliers) on classifiers (rules and trees) and pre-processing methods
- Preprocessing
  - SPIDER, NCR, *cluster-oversampling* and random oversampling
- Data
  - Artificial data sets (sub-clusters, paw, clover - flower)

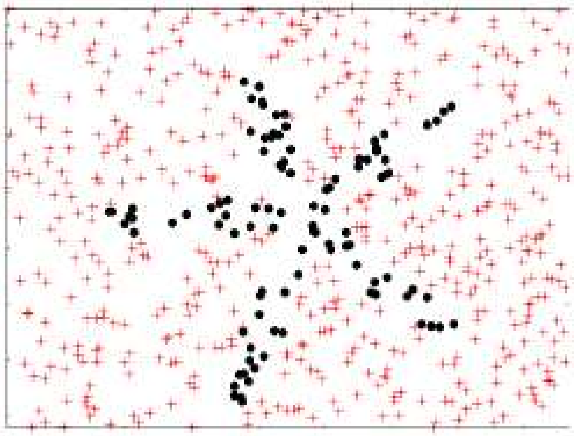


Fig. 1. Clover data set

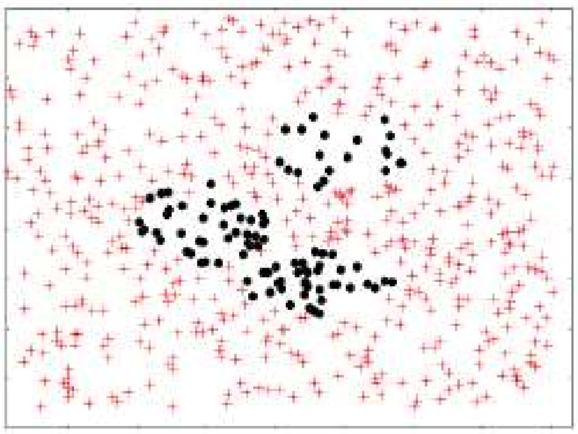


Fig. 2. Paw data set

Some results

Dataset	Base	Oversampling	Filtr Japkowicz	NCR	SPIDER
subclus-0	0.9540	0.9500	0.9500	0.9460	**0.9640**
subclus-30	0.4500	0.6840	0.6720	0.7160	**0.7720**
subclus-50	0.1740	0.6160	0.6000	0.7020	**0.7700**
subclus-70	0.0000	0.6380	0.7000	0.5700	**0.8300**
clover-0	0.4280	0.8340	**0.8700**	0.4300	0.4860
clover-30	0.1260	0.7180	0.7060	0.5820	**0.7260**
clover-50	0.0540	0.6560	0.6960	0.4460	**0.7700**
clover-70	0.0080	0.6340	0.6320	0.5460	**0.8140**
paw-0	0.5200	**0.9140**	0.9000	0.4900	0.5960
paw-30	0.2640	0.7920	0.7960	0.8540	**0.8680**
paw-50	0.1840	0.7480	0.7200	0.8040	**0.8320**
paw-70	0.0060	0.7120	0.6800	0.7460	**0.8780**

Sensitivity

C4.5

Table 3. Sensitivity for artificial data sets with different types of testing examples

Data set	MODLEM					C4.5				
	Base	RO	CO	NCR	SP2	Base	RO	CO	NCR	SP2
subcl-safe	0.5800	0.5800	0.6200	0.7800	0.6400	0.3200	0.8400	0.8600	0.9800	1.0000
subcl-B	0.8400	0.8400	0.8400	0.8600	0.8400	0.0000	0.8200	0.8400	0.3600	0.92[illegible]
subcl-C	0.1200	0.1000	0.1600	0.2400	0.2600	0.0000	0.5400	0.0000	0.0000	0.5200
subcl-BC	0.4800	0.4700	0.5000	0.5500	0.5500	0.0000	0.6800	0.4200	0.1800	0.7200
clover-safe	0.3000	0.3800	0.4400	0.7000	0.6000	0.0200	0.9600	0.9200	0.0400	0.9800
clover-B	0.8400	0.8200	0.8200	0.8400	0.8600	0.0400	0.9400	0.9200	0.0400	0.9400
clover-C	0.1400	0.0800	0.1400	0.2400	0.3600	0.0000	0.3000	0.0200	0.0000	0.4000
clover-BC	0.4900	0.4500	0.4800	0.5400	0.6100	0.0200	0.6200	0.4700	0.0200	0.6700
paw-safe	0.8400	0.9200	0.8400	0.8400	0.8000	0.4200	0.9000	0.9600	0.7400	1.0000
paw-B	0.8800	0.8800	0.8600	0.8800	0.9000	0.1400	0.9000	0.9000	0.4000	0.9200
paw-C	0.1600	0.1400	0.1200	0.2600	0.1600	0.0400	0.2000	0.0000	0.0000	0.3400
paw-BC	0.5200	0.5100	0.4900	0.5700	0.5300	0.0900	0.5500	0.4500	0.2000	0.6300

For a small number of difficult examples
- *cluter-oversampling*

Over 30% - SPIDER or SMOTE

What about assessing types of real examples?

- We analyse class labels in the local neighbourhood of the given example
- How to model this local neighbourhood
  - HVDM – distance measure
  - K-NN or kernel functions

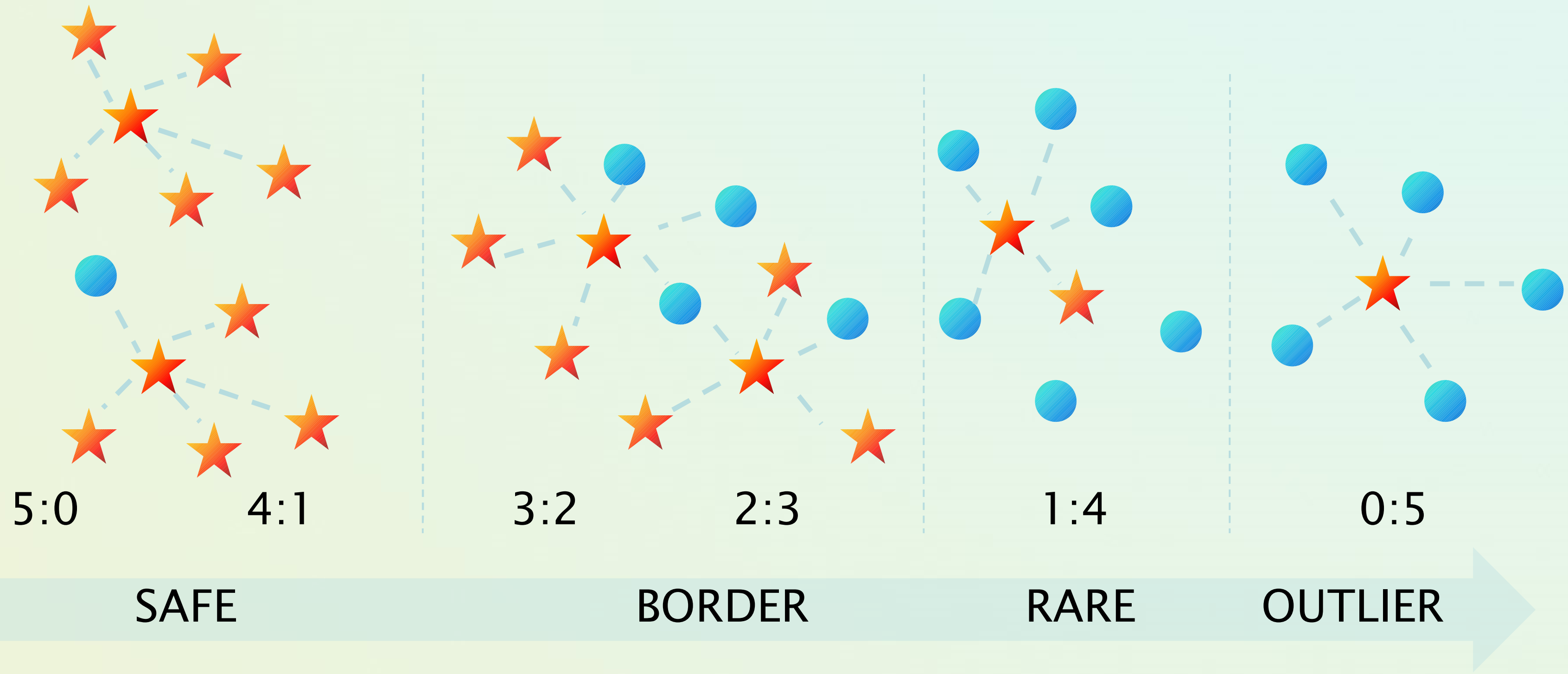
Identification of examples - local approach

- Analyse the distribution in the local neighbourhood
  - K-NN (k=5, and others)
  - Distance HVDM

$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^p$$

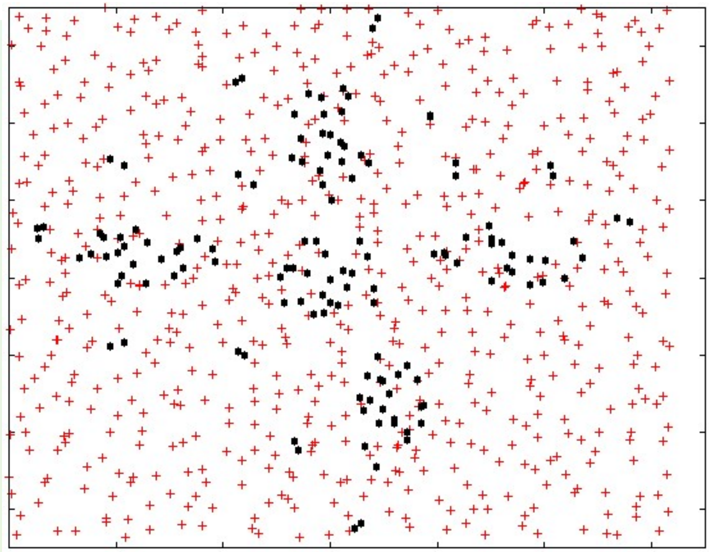
V_1, V_2 – corresponding feature values

- C_1 – total number of occurrences of V_1
- C_{1i} – total number of occurrences of V_1 for class i
- n – number of classes, p – constant (usually 1)



Re-discovery of known distributions

Dataset Description					Identified Labels			
Imbalance Ratio	Sub-concepts	Border [%]	Rare [%]	Outlier [%]	Safe [%]	Border [%]	Rare [%]	Outlier [%]
1:5	1	60	20	0	17.04	60.74	21.48	0.74
1:5	3	60	20	0	18.52	57.78	23.70	0.00
1:5	5	60	20	0	17.78	64.44	17.78	0.00
1:5	5	0	0	10	64.44	25.93	0.00	9.63
1:7	5	0	0	10	54.00	36.00	0.00	10.00
1:9	5	0	0	10	52.00	36.00	2.00	10.00

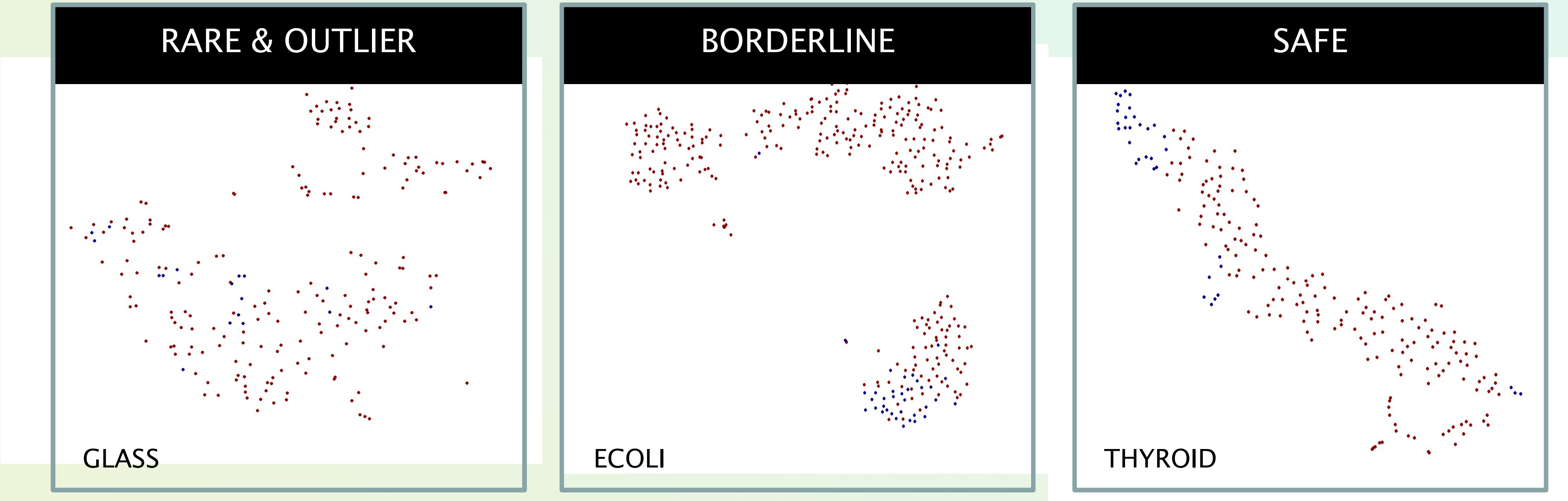


Experiments with UCI data sets

Dataset	Ab.	Size	Ratio [%]	Min.
abd-pain	AP	723	27.94	positive
acl	AC	140	28.57	1
new-thyroid	NT	215	16.28	hyper
vehicle	VE	846	23.52	van
car	CA	1728	3.99	good
scrotal-pain	SP	201	29.35	positive
credit-g	CG	1000	30	bad
ecoli	EC	336	10.42	imU
hepatitis	HE	155	20.65	die
ionosphere	IO	351	35.89	bad
haberman	HA	306	26.47	died
cmc	CM	1473	22.61	l-term
b-cancer	BC	286	29.72	rec-ev
cleveland	CL	303	11.55	positive
glass	GL	214	7.94	v-float
hsv	HS	122	11.48	4.0
abalone	AB	4177	8.02	0-4 16-29
postoperate	PO	90	26.66	S
solar-flare	SF	1066	4.03	F
transfusion	TR	748	23.8	yes
yeast	YE	1484	3.44	ME2

Different data distributions

Dataset	Safe	Border	Rare	Outlier	Category
new-thyroid	**68,57**	31,43	0,00	0,00	S
ecoli	28,57	**54,29**	2,86	14,29	B
glass	0,00	35,29	**35,29**	**29,41**	R, O



K.Napierała, J.Stefanowski: Identification of Different Types of Minority Class Examples in Imbalanced Data. HAIS 2012

Categories of data sets

Dataset	Safe [%]	Border [%]	Rare [%]	Outlier [%]	Category
abdominal pain	59,90	22,28	8,90	7,92	S
acl	67,50	30,00	0,00	2,50	S
new-thyroid	68,57	31,43	0,00	0,00	S
vehicle	74,37	24,62	0,00	1,01	S
car	47,83	39,13	8,70	4,35	B
ionosphere	44,44	30,95	11,90	12,70	B
scrotal pain	38,98	45,76	10,17	5,08	B
credit-g	9,33	63,67	10,33	16,67	B
ecoli	28,57	54,29	2,86	14,29	B
hepatitis	15,63	62,50	6,25	15,63	B
haberman	4,94	61,73	18,52	14,81	B, R
cmc	17,72	44,44	18,32	19,52	R
breast-cancer	24,71	25,88	32,94	16,47	R
cleveland	0,00	31,43	17,14	51,43	R, O
glass	0,00	35,29	35,29	29,41	R, O
hsv	0,00	0,00	28,57	71,43	R, O
abalone	8,36	20,60	20,60	50,45	R, O
postoperative	0,00	41,67	29,17	29,17	R, O
solar-flare	0,00	48,84	11,63	39,53	O
transfusion	18,54	47,19	11,24	23,03	O
yeast	5,88	47,06	7,84	39,22	O

Sensitivity of classifiers

S: 70–90%

B: 30–60%

R: 0–40%

O: 0–30%

DS	L	1NN	3NN	J48	PAR	RBF	SVM
AP	S	76.4	78.5	69.8	72.6	75.0	71.8
AC	S	72.0	78.5	85.5	80.0	84.0	82.5
NT	S	96.3	90.2	92.2	93.3	99.5	89.8
VE	S	89.1	87.9	87.0	88.3	88.0	95.2
CA	B	3.1	3.1	77.7	90.0	49.6	88.2
SP	B	58.4	58.7	55.3	63.4	62.5	65.9
CG	B	50.3	39.9	46.5	47.7	43.6	52.2
EC	B	52.2	50.8	58.0	42.0	54.7	58.5
HE	B	44.0	37.0	43.2	45.7	60.7	51.5
IO	B	69.4	65.5	82.7	84.0	94.2	89.0
HA	BR	30.1	26.9	41.0	33.4	18.3	1.3
CM	R	37.6	33.8	39.2	37.7	12.1	5.2
BC	R	40.4	27.6	38.7	41.1	40.8	45.3
PO	RO	4.3	0.0	4.7	10.3	13.7	7.0
CL	RO	20.3	12.5	23.7	25.2	9.5	9.0
GL	RO	30.0	16.0	30.0	34.0	25.0	0.0
HS	RO	0.0	0.0	0.0	2.0	1.0	0.0
AB	RO	20.5	16.5	30.4	18.8	12.3	0.2
SF	O	9.1	8.2	20.9	18.7	10.2	15.7
TR	O	31.9	34.3	41.3	42.9	32.9	2.2
YE	O	38.1	26.2	30.9	26.7	15.1	0.0

Pre-processing with respect to labels of testing examples

Friedman Tests

- 4.3 • NCR
- 3.8 • SPIDER
- 3.4 • SMOTE
- 2.0 • RO
- 1.3 • NONE

- • SPIDER • NCR
- • SMOTE
- • RO
- • NONE

- 4.3 • SMOTE
- 3.5 • SPIDER
- 3.4 • NCR
- 2.2 • RO
- 1.5 • NONE

❑ 1NN, J48: similar orders

❑ RBF: RO slightly better

B

DS	NONE	RO	NCR	SM	SP
IO	92.1	92.1	95.2	93.3	92.7
CA	91.1	69.6	92.6	89.6	86.7
SP	64.0	69.6	74.4	68.8	77.6
CG	53.3	54.1	76.9	58.8	67.9
EC	32.9	60.0	78.8	90.6	80.0
HE	65.7	80.0	82.9	80.0	80.0
HA	48.2	69.4	73.5	85.3	86.5

R

DS	NONE	RO	NCR	SM	SP
HA	20.6	49.0	48.4	62.6	64.5
CM	34.9	40.4	56.1	41.4	45.1
BC	26.7	28.7	59.3	35.3	44.7
CL	22.2	22.2	33.3	22.2	22.2
GL	25.0	25.0	45.0	37.5	35.0
HS	0.0	30.0	0.0	20.0	20.0
AB	12.4	37.1	26.5	52.1	48.8
PO	8.0	18.0	42.0	6.0	32.0
SF	32.0	58.0	66.0	52.0	60.0
TR	21.2	42.4	31.2	62.4	58.8
YE	20.0	42.0	12.0	38.0	24.0

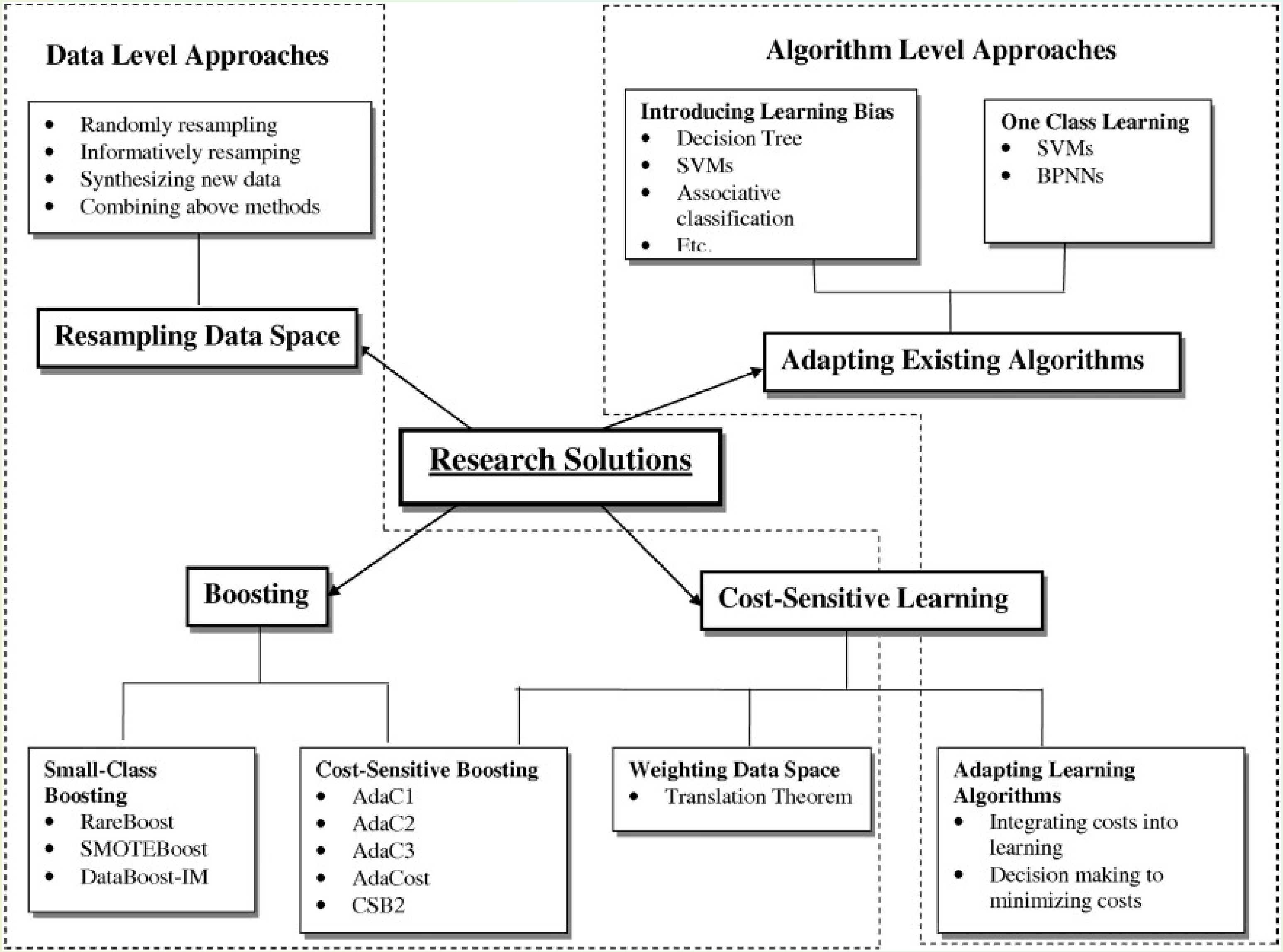
O

DS	NONE	RO	NCR	SM	SP
CM	19.1	24.0	28.0	25.5	30.2
BC	11.7	18.3	33.3	20.0	26.7
CL	16.7	11.1	37.8	21.1	10.0
GL	28.0	16.0	48.0	52.0	32.0
HS	4.0	4.0	12.0	16.0	8.0
AB	10.4	27.7	16.6	41.5	39.1
PO	5.7	5.7	28.6	22.9	14.3
SF	2.4	16.5	12.9	12.9	27.1
TR	1.6	22.9	4.9	45.3	49.4
YE	2.0	7.0	9.0	26.0	13.0

Summary of experiments (Napierała, Stefanowski 2012)

- Types of distributions / learning examples is an additional influential factor
- Quite limited number of safe data sets → easy even for simple classifiers
- Most data sets contain all types of examples
- Different performance depending on types
- Classifiers
  - S – all classifiers comparable
  - B - SVM -> trees/rules, RBF -> kNN
  - R/O – trees/rules 1NN >
- Preprocessing
  - B: undersampling (NCR)
  - R: hybrid (SPIDER) > SMOTE
  - O: SMOTE
  - S: over-sampling (partly with RBF)

Taxonomy Methods: Data level vs Algorithm Level

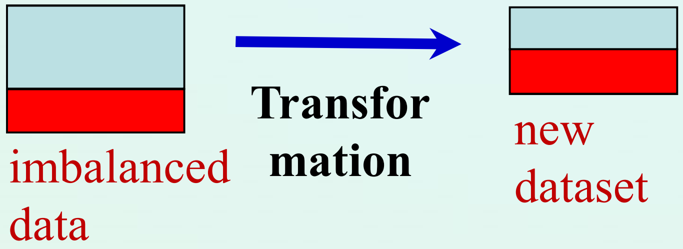


Y. Sun, A. K. C. Wong and M. S. Kamel. Classification of imbalanced data: A review. International Journal of Pattern Recognition 23:4 (2009) 687-719.

Pre-processing approaches

Transform original data distribution:

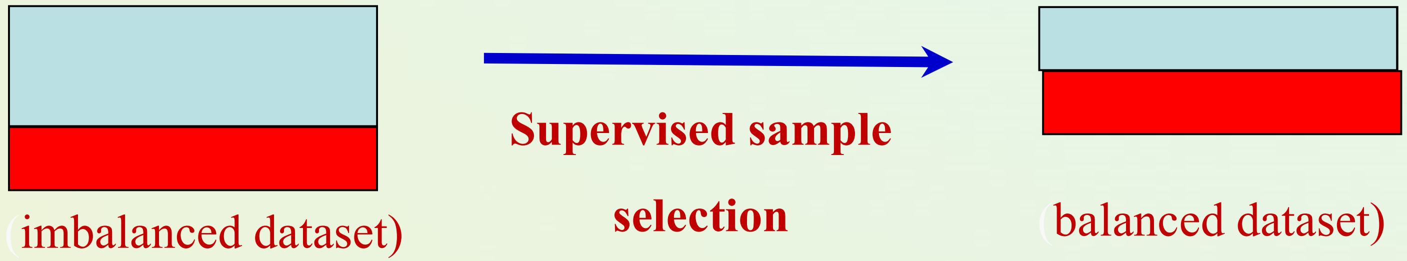
- Simple random sampling
  - „Over-sampling" – minority class
  - „Under-sampling"- majority class
- Specializaed over sampling
  - Cluster-oversampling (Japkowicz)
- Informed and focused transformation
  - More clearning majority examples
    - One-side-sampling (Kubat, Matwin) z Tomek Links
    - Laurikkala's edited nearest neighbor rule
  - Oversampling
    - SMOTE $\rightarrow$ Chawla et al.
    - Borderline SMOTE, Safe Level, Surrounding SMOTE, …
  - Hybrid ones
    - SPIDER
    - SMOTE i undersampling *ENN, ….)
  - Modifications of ensembles



Random resampling the original data sets

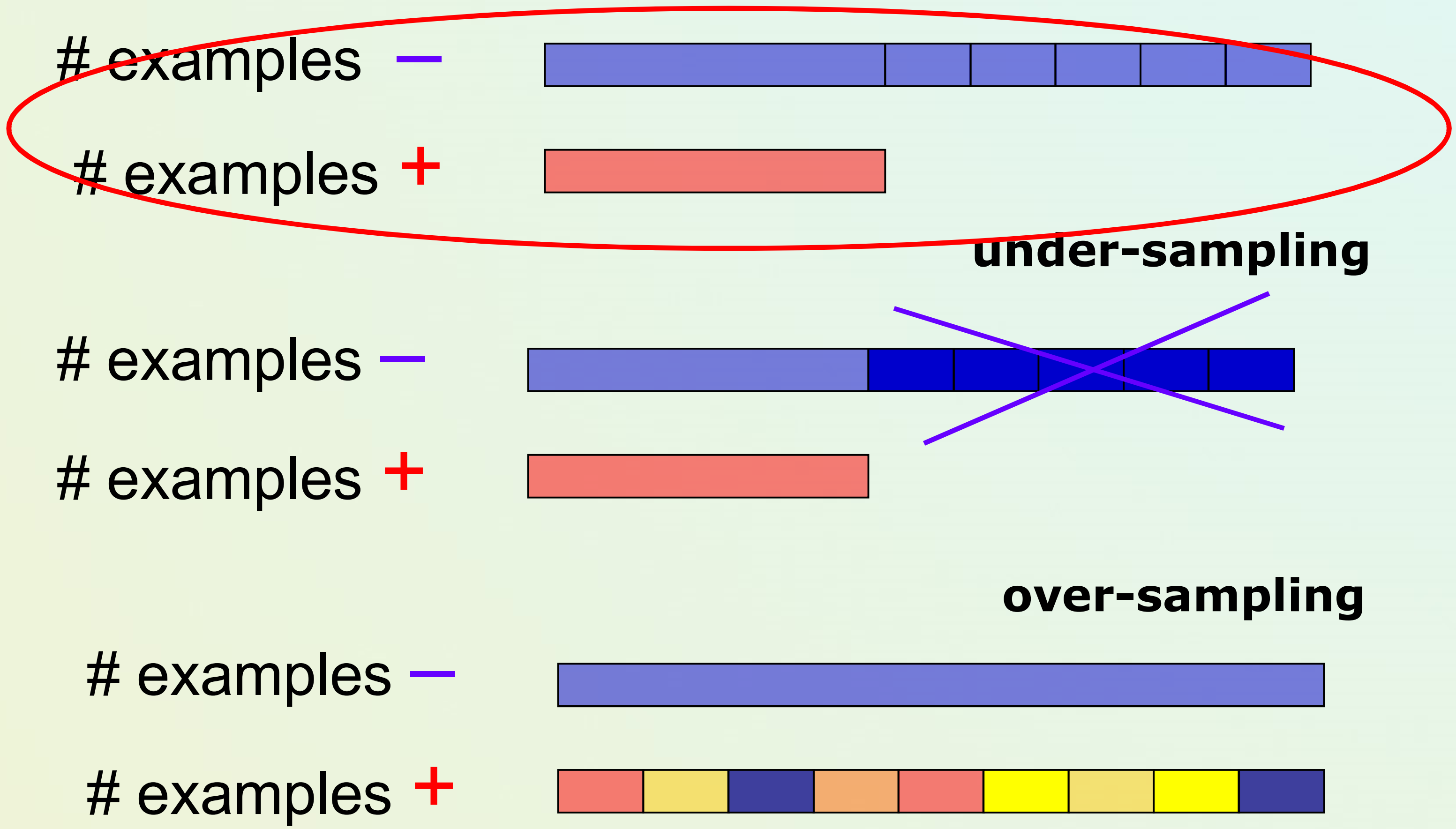
Resampling is the process of manipulating the distribution of the training examples (in a pre-processing step) in an effort to improve the performance of classifiers.

There is no guarantee that the training examples occur in their optimal distribution in practical problems, and thus, the idea of resampling is “to add or remove examples with the hope of reaching the optimal distribution of the training examples” and thus, realizing the potential ability of classifiers.



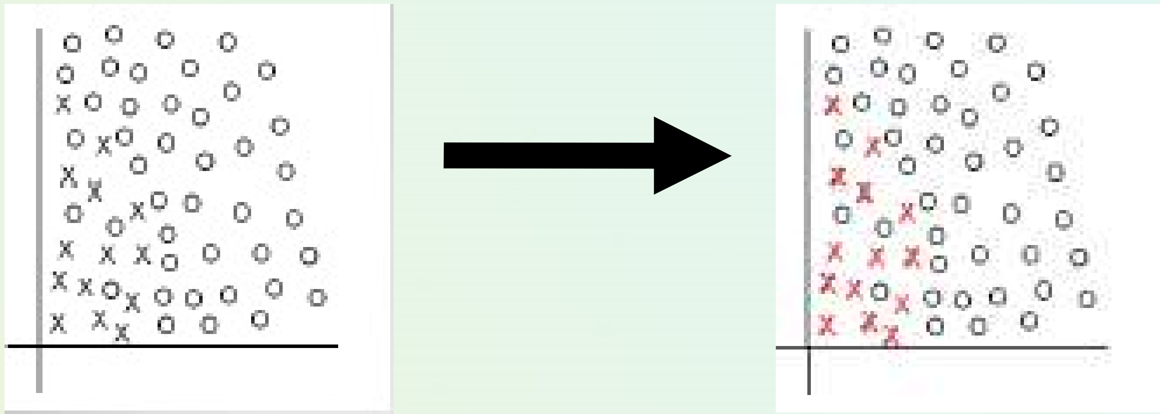
Plain random approaches

Undersampling vs oversampling

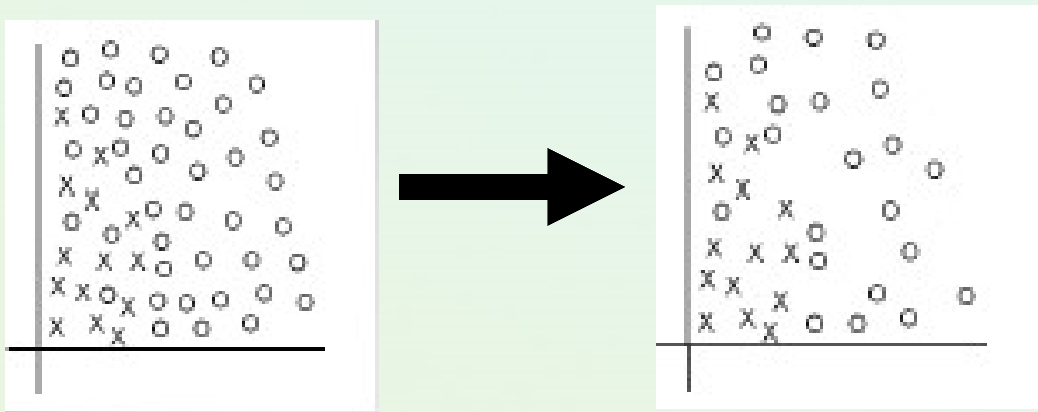


Discussion of random resampling

Random oversampling → just coping minority examples



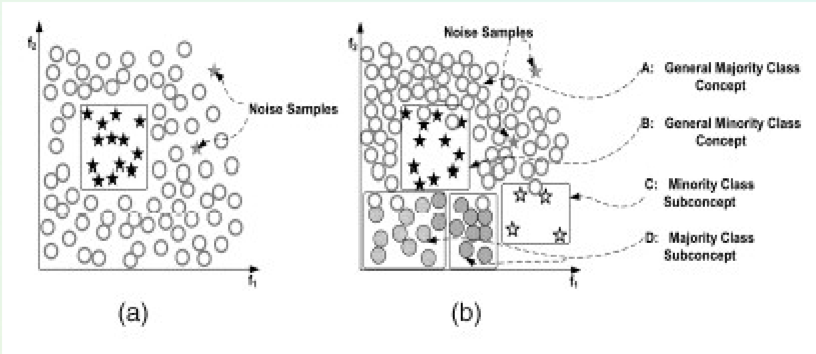
- Is balance 1:1 the best option?
- You may overfit!
- Random undersampling:
- Remove majority examples



- Loosing valuable examples

Cluster oversampling – Japkowicz decomposition

Decomposition → within and between -class imbalance



How find „small sub-concept" – is it easy?

Think about clustering and randomly oversample clusters!

Once the training examples of each class have been clustered, oversampling starts. In the majority class, all the clusters, except for the largest one, are randomly oversampled so as to get the same number of training examples as the largest cluster. Let *maxclasssize* be the overall size of the large class. In the minority class, each cluster is randomly oversampled until each cluster contains *maxclasssize/Nsmallclass* where *Nsmallclass* represents the number of subclusters in the small class.

Cluster-based resampling identifies rare regions and re-samples them individually, so as to avoid the creation of small disjuncts in the learned hypothesis.

T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49

Illustration of Cluster based Oversampling

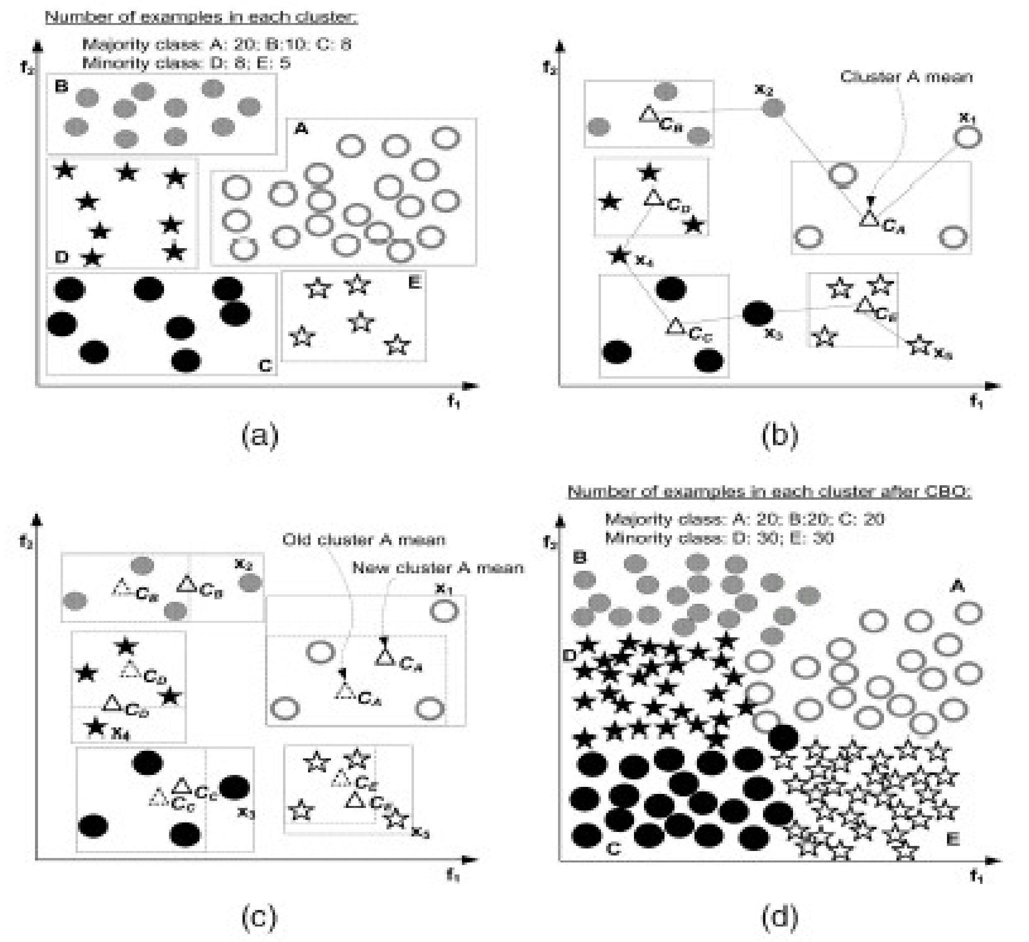


Fig. 6. (a) Original data set distribution. (b) Distance vectors of examples and cluster means. (c) Newly defined cluster means and cluster borders. (d) The data set after cluster-based oversampling method.

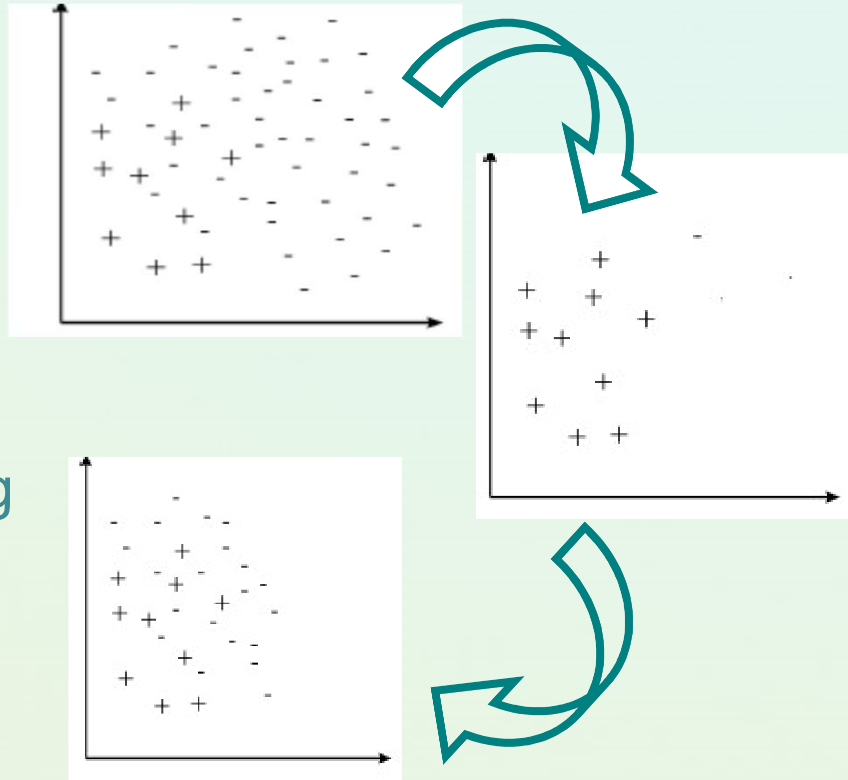
- ❑ Approach with k-means / however k=?

Under-sampling with CNN

CNN – Condensed Nearest Neighbours → Edited K-NN

The general idea (quite old one).
Find a subset *E'* of *E*, which reclassify correctly all examples from *E* with basic 1NN algorithm. Duda, Hart 1968.

- Remove difficult examples
- Schema:
  - Let E be the original training set
  - Let E' contains all positive examples from S and one randomly selected negative example
  - Classify E with the 1-NN rule using the examples in E'
  - Move all misclassified example from E to E'

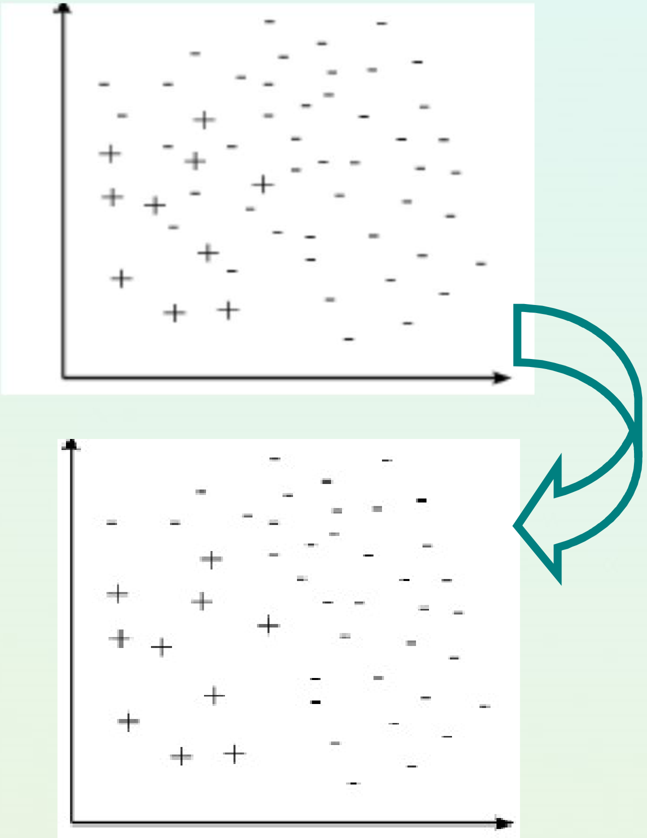


Under-sampling the original data sets with Tomek links

Tomek Links

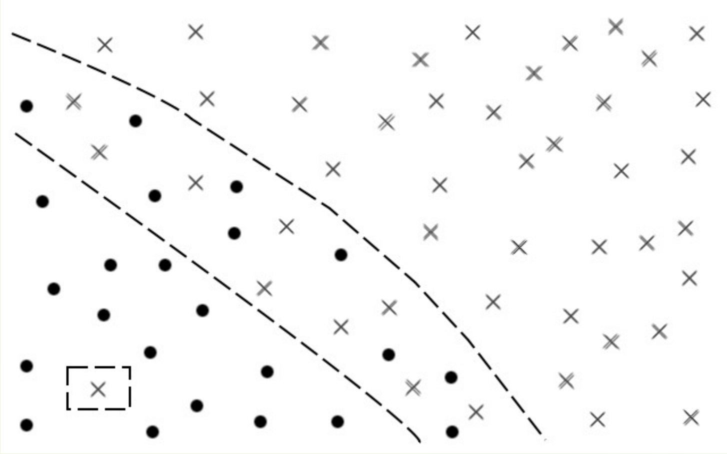
- To remove both noise and borderline examples of the majority class
- Tomek link
  - –Ei, Ej belong to different classes,

    d (Ei, Ej) is the distance between them.
  - –A (Ei, Ej) pair is called a Tomek link if there is no example El, such that d(Ei, El) < d(Ei, Ej) or d(Ej , El) < d(Ei, Ej).

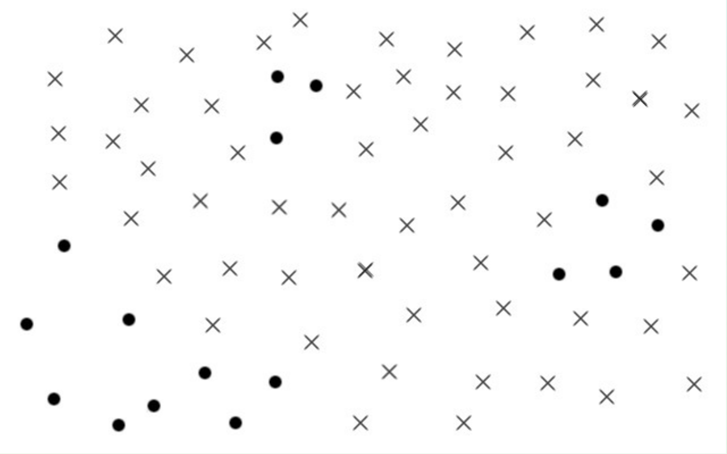


Two different aspects of data distribibition

Matwin and Kubat
→ one side sampling



Decompositon / fragmentation for over-sampling



How do they influence sampling in pre-processing methods?

Typology of examples in data distributions

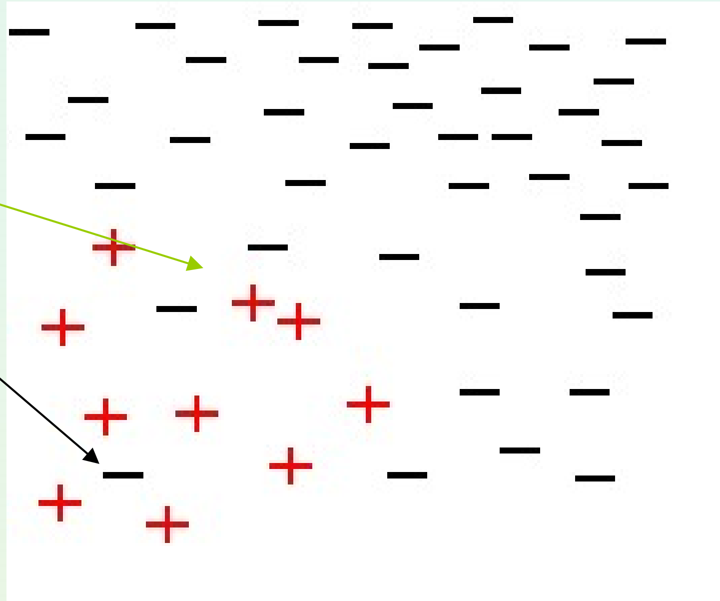
Re-sampling should be focused on some types of examples

One –side-sampling - Kubat, Matwin 1997

They distinguish different types of examples (majority ones)

- ❑ Noise examples
- ❑ Borderline examples

  Borderline examples are unsafe since a small amount of noise can make them fall on the wrong side of the decision border.
- ❑ Redundant examples

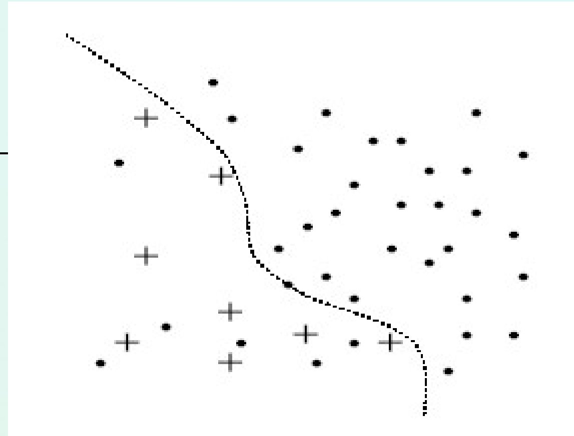


- ❑ Safe examples

Identify them and remove some of them (Tomek links, CNN)

One –side-sampling Kubat, Matwin 1997

- One-sided selection
  - Tomek links + CNN
  - may remove too many examples from the majority class
- CNN + Tomek links
  - F. Herrera
  - Finding Tomek links is computationally demanding, it would be computationally cheaper if it was performed on a reduced data set



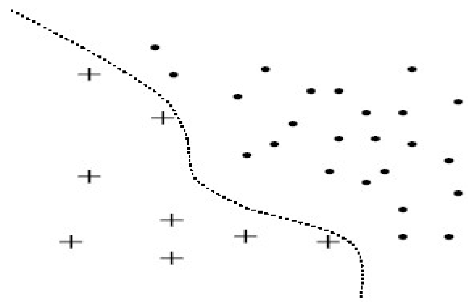


Figure 3: The training set without the borderline and noisy negative examples.

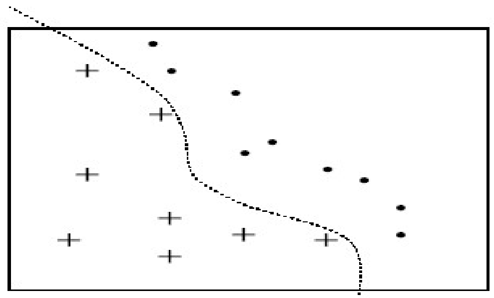


Figure 4: The training set after the removal of redundant negative examples.

Nearest Cleaning Rule

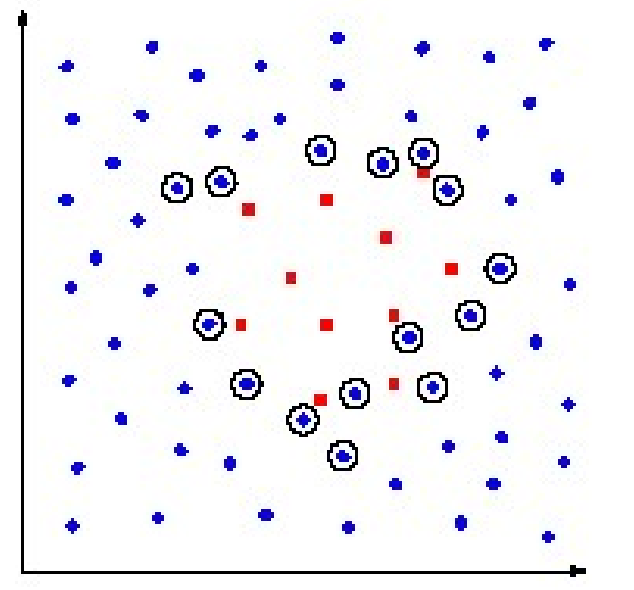
- NCL Nearest Cleaning Rule - Jorma Laurikkala 2001,

  Differnt to OSS, more „cleans” boundaries than removes so many examples

  Algorytm:

  – Find three nearest neighbors for each example Ei in the training set
  – If Ei belongs to majority class, & the three nearest neighbors classify it to be minority class, then remove Ei
  – If Ei belongs to minority class, and the three nearest neighbors classify it to be majority class, then remove the majority nearest neighbors

❑ Simple illustration



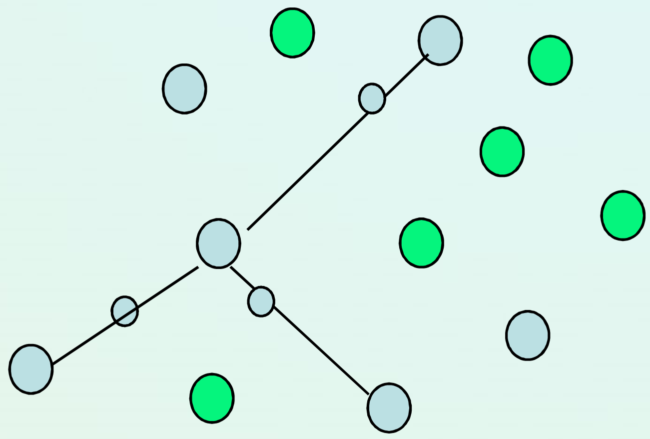
Rysunek 5.3: Neighbourhood Cleaning Rule

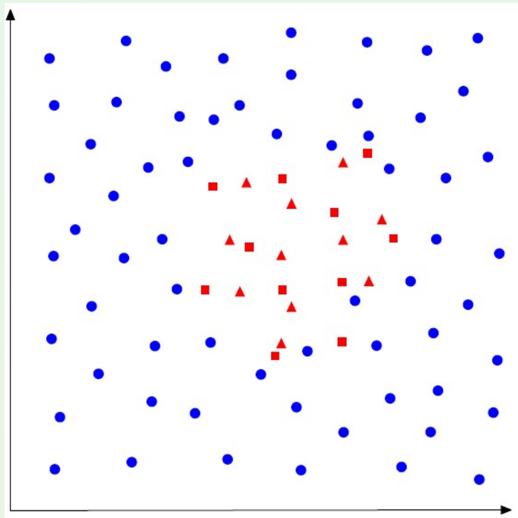
SMOTE - Synthetic Minority Oversampling Technique

- N.Chawla, Hall, Kegelmeyer 2002
- For each *p* from the minority class
  - Find its ***k***-nearest neighbours also from the minority class
    - HVDM distance
  - Randomly select ***o*** of these neighbours

    (*o* - the amount of over-sampling desired)
  - Generate a synthetic example along the line between *p* and randomly selected example *n*

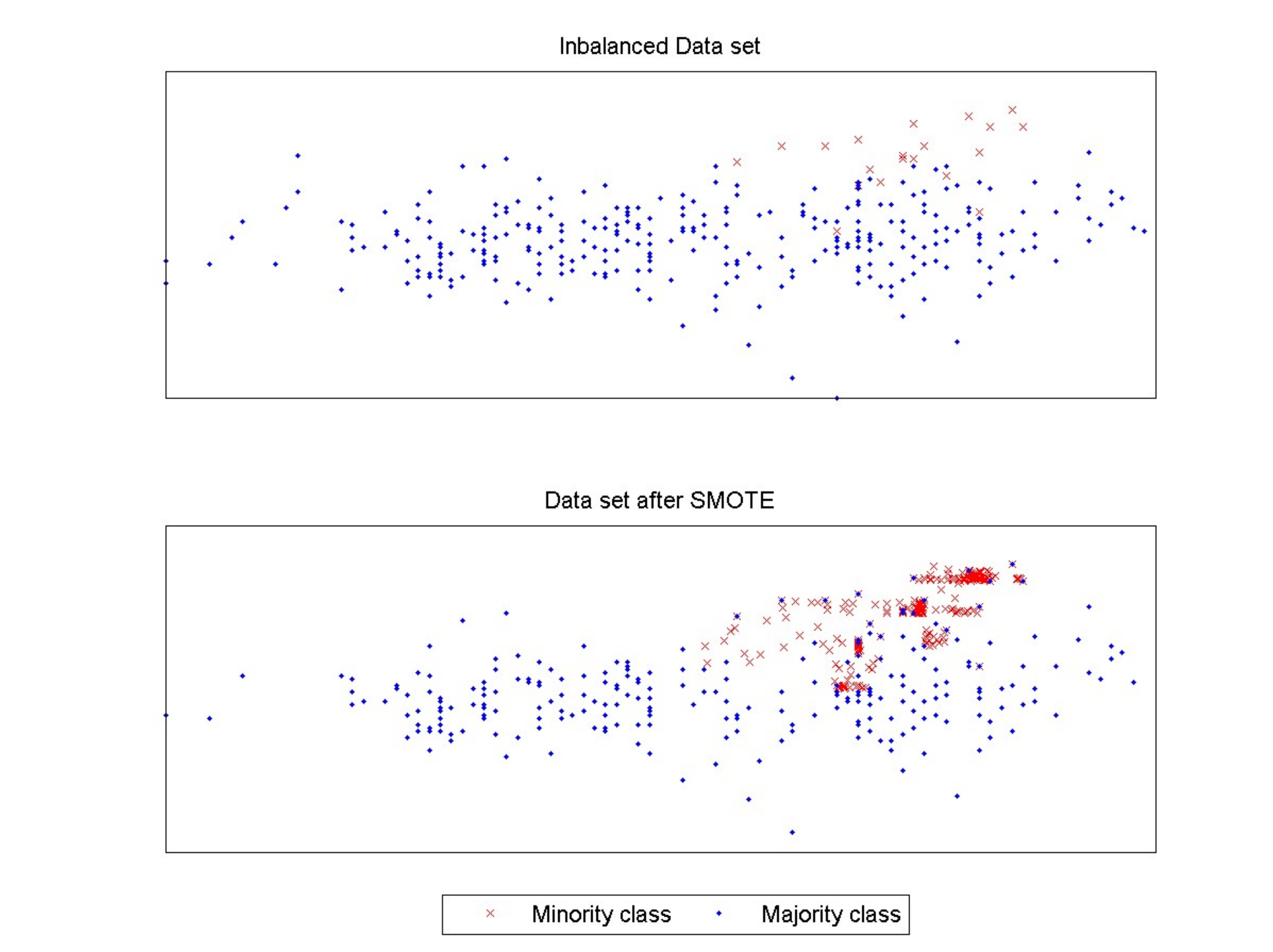
$$x_{new} = p_i + (n_i - p_i) \cdot \delta$$





- It generalizes the minority class regions without causing overfitting
- Quite efficient, also if combined with under-sampling

SMOTE : Example of a run



SMOTE – przykład oceny AUC

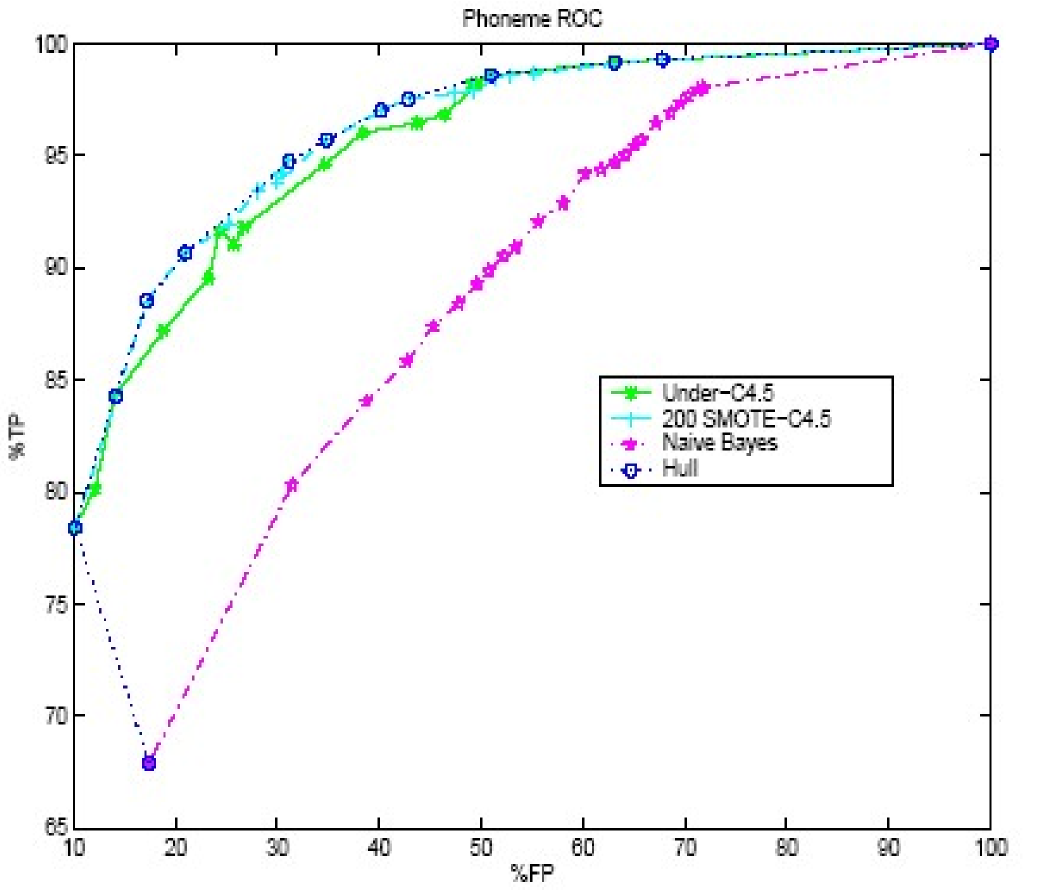


Figure 7: Phoneme. Comparison of SMOTE-C4.5, Under-C4.5, and Naive Bayes. SMOTE-C4.5 dominates over Naive Bayes and Under-C4.5 in the ROC space. SMOTE-C4.5 classifiers are potentially optimal classifiers.

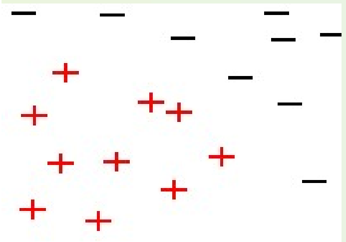
SMOTE – Chawla's results

- K=5 neighboirs, different oversampling ratio (e.g. 100% increases twice the cardinality of the minority class)

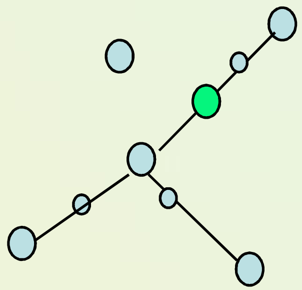
Dataset	Under	50 SMOTE	100 SMOTE	200 SMOTE	300 SMOTE	400 SMOTE	500 SMOTE
Pima	7242		**7307**				
Phoneme	8622		8644	**8661**			
Satimage	8900		8957	**8979**	8963	8975	8960
Forest Cover	9807		9832	9834	**9849**	9841	9842
Oil	8524		8523	8368	8161	8339	**8537**
Mammography	9260		9250	9265	9311	**9330**	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.

Critical remarks on related works –
J.Stefanowski, Sz.Wilk, ECML/PKDD workshop 2007

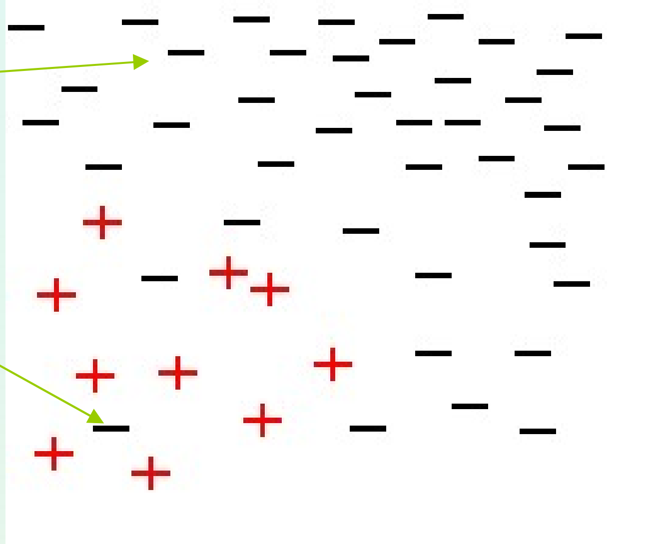


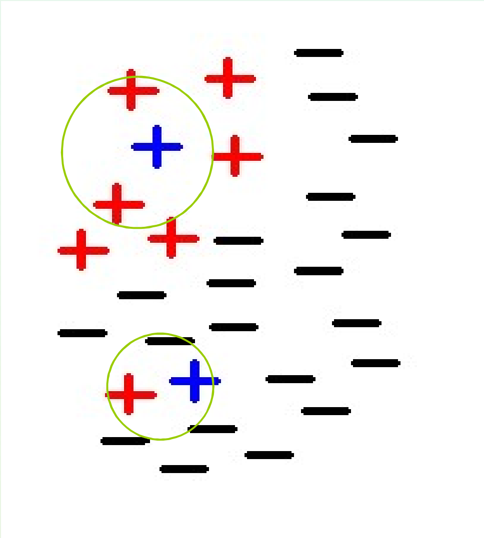
- NCR and one-side-sampling
  - Greedy removing of (too) many examples from the majority class!
  - Focused on improving sensitivity of the minority class
- SMOTE
  - Introduction of many random examples from the minority class may be difficult to interpret in some domains (medicine),
  - „Blind" over-generalization in the directions of neighbors from majority classes,
  - Number of synthetic examples - o – a global parameter requiring tuning.



SPIDER assumptions

- Distinguish two types of examples:
  - **Safe** → should be classified correctly,
  - **Unsafe** → more likely to be misclassified; require special attention
  - Later on → borderline and noisy outliers
- Assumptions:
  - All examples from the minority class are preserved,
  - Unsafe majority ones may be changed
- Use Wilson's edited nearest neighbor rule:
  - Compare example's label with its neighbors,
  - Safe → correctly classified by its *k* nearest neighbors,
  - Unsafe → otherwise





Selective Preprocesing of Imbalanced Data → SPIDER
J.Stefanowski, Sz.Wilk, ECML/PKDD workshop 2007

- Increasing sensitivity without so strong decrease of specificity - could be done without artificial examples + not so extensive changes of class cardinalities?
- Hybrid approach → limited filtering i and local copying of some minority examples

Two phases

- Identifying types of examples
- For the majority class selective removing noise examples or relabeling them
- The minority class – re-sampling borderline examples and some of noisy ones.
  - weak or strong amplification
    - amplify by creating as many copies as there are *O*-safe examples in the *k*-neighborhood
  - Some *C*-noisy examples → introduce more copies (k = 3 → 5)

More about options in phase 2

Weak amplification:

1. All *C*-noisy examples → amplify by creating as many copies as there are *O*-safe examples in the *k*-neighborhood (increase their „weight”).

Relabeling and amplification:

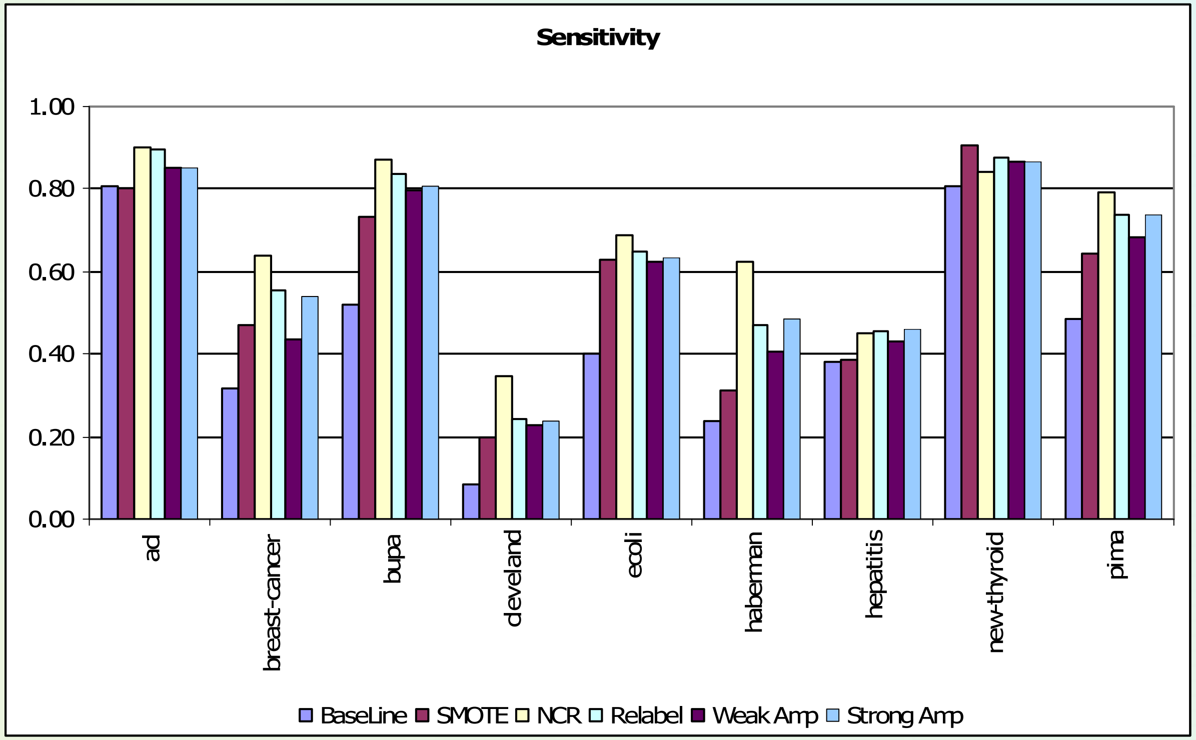
1. *O*-noisy examples from the *k*-neighborhood of *C*-noisy examples → change their class label from *O* to *C* (*extend cover*),
2. All *C*-noisy examples → amplify by creating as many copies as there are *O*-safe examples in the *k*-neighborhood.

Strong amplification

1. Some *C*-noisy examples → introduce more copies ($k = 3 \rightarrow 5$),
2. *C*-safe examples → duplicate depending on *O*-safe neighbors.

Finally all remaining *O*-noisy examples are removed.

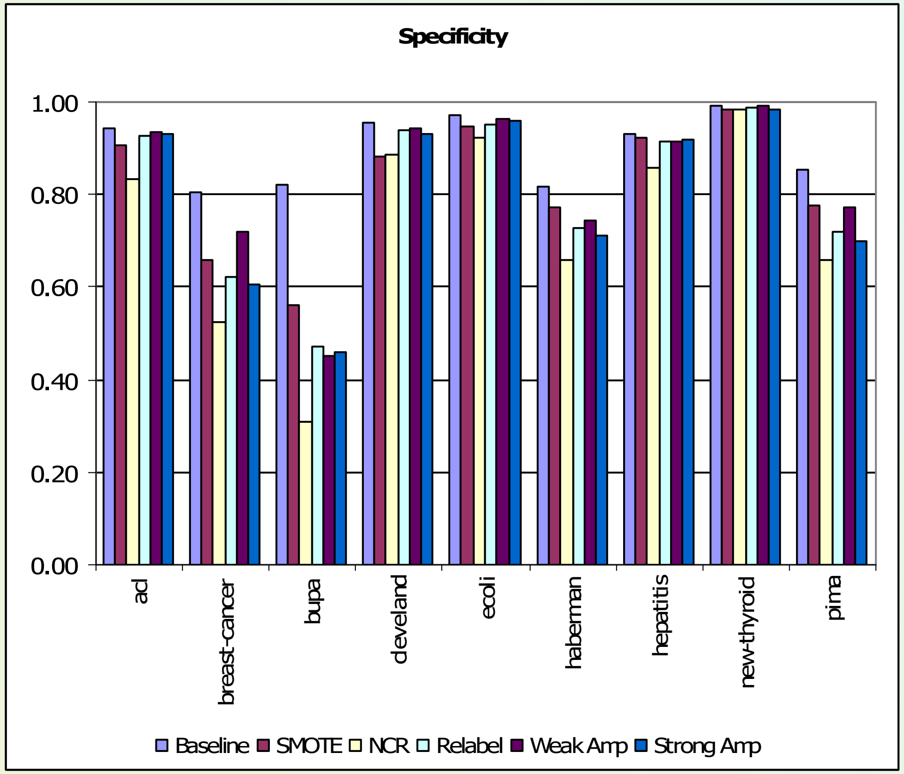
MODLEM rules → sensitivity

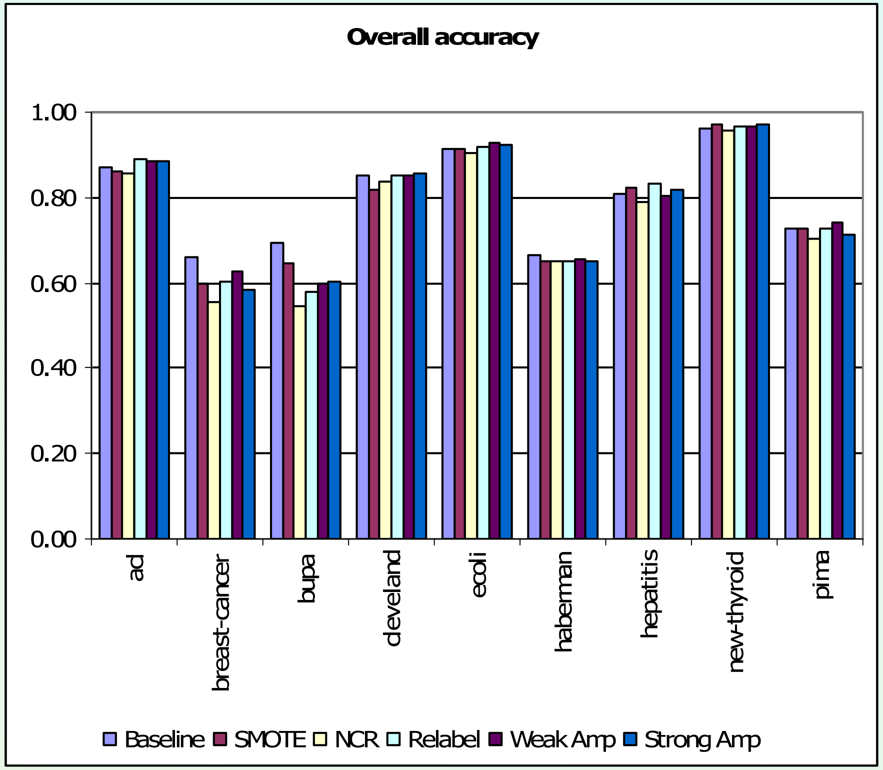


- All approaches outperform the baseline approach.
- NCR – the highest improvement (haberman 0.386, bupa 0.353).
- Relabeling or strong amplification – the second best approach (7 of 9 sets), then weak amplification or SMOTE

More: J.Stefanowski, Sz.Wilk. Selective pre-processing of imbalanced data for improving classification performance. DAWAK 2008

MODLEM rules → specificity and total accuracy





- The best specificity and accuracy for the baseline approach.
- NCR – the worst approach; for some data high decrease of specificity (bupa 0.512) and also deterioration of accuracy.
- Other approaches between baseline and NCR.
- Weak amplification is able to maintain the values.

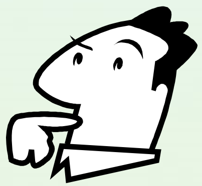
Changes in the class distribution

	SMOTE		NCR		Relabel				Weak Amp		Strong Amp	
Data set	N_C	N_O	N_C	N_O	N_C	N_O	N_R	N_A	N_C	N_O	N_C	N_O
acl	120	100	40	83	59	98	2	17	57	98	67	98
breast-cancer	255	201	85	101	197	167	24	88	173	167	253	167
bupa	290	200	145	81	271	145	35	91	236	145	309	145
cleveland	245	268	35	198	110	255	8	67	102	255	147	255
ecoli	210	301	35	266	69	288	11	23	58	288	77	288
haberman	162	225	81	121	193	182	31	81	162	182	223	182
hepatitis	64	123	32	90	68	113	7	29	61	113	88	113
new-thyroid	175	180	35	174	40	179	0	5	40	179	47	179
pima	536	500	268	280	493	409	63	162	430	409	573	409

- Larger changes led to better performance.
- NCR removed the largest number of examples from the majority classes (up to 50%).
- SMOTE increased the minority class on average by 250%.
- New approach not so greedy:
  - Only strong amplification similar to SMOTE,
  - More amplified examples than relabeled.

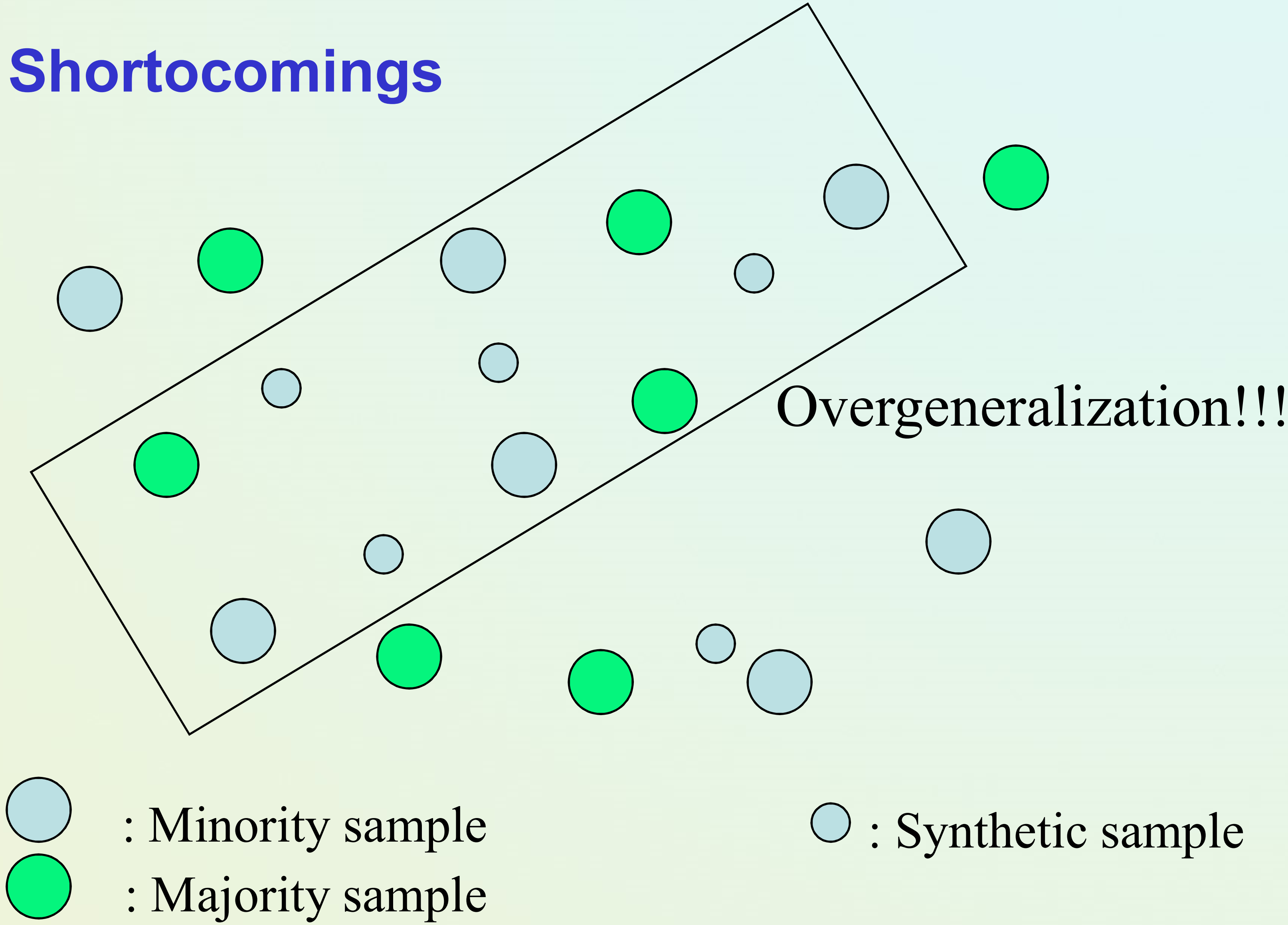
SMOTE - again

- Do not distinguish any type of examples
  - Each minority class examples $\rightarrow$ a seed for oversampling
- „Blind" over-generalization in the directions of neighbors from majority classes
  - Can address a class fragmentation into sub-concepts?
- Two directions
  - Combine with post-processing, e.g. SMOTE+ENN
  - Try to modify internal elements of SMOTE



Resampling the original data sets

SMOTE Shortocomings



SMOTE: Hybridization

- Problem with Smote: might introduce the artificial minority class examples too deeply in the majority class space.

- Tomek links: data cleaning

- Smote + Tomek links: Instead of removing only the majority class examples that form Tomek links, examples from both classes are removed

SMOTE hybridization: SMOTE + Tome links

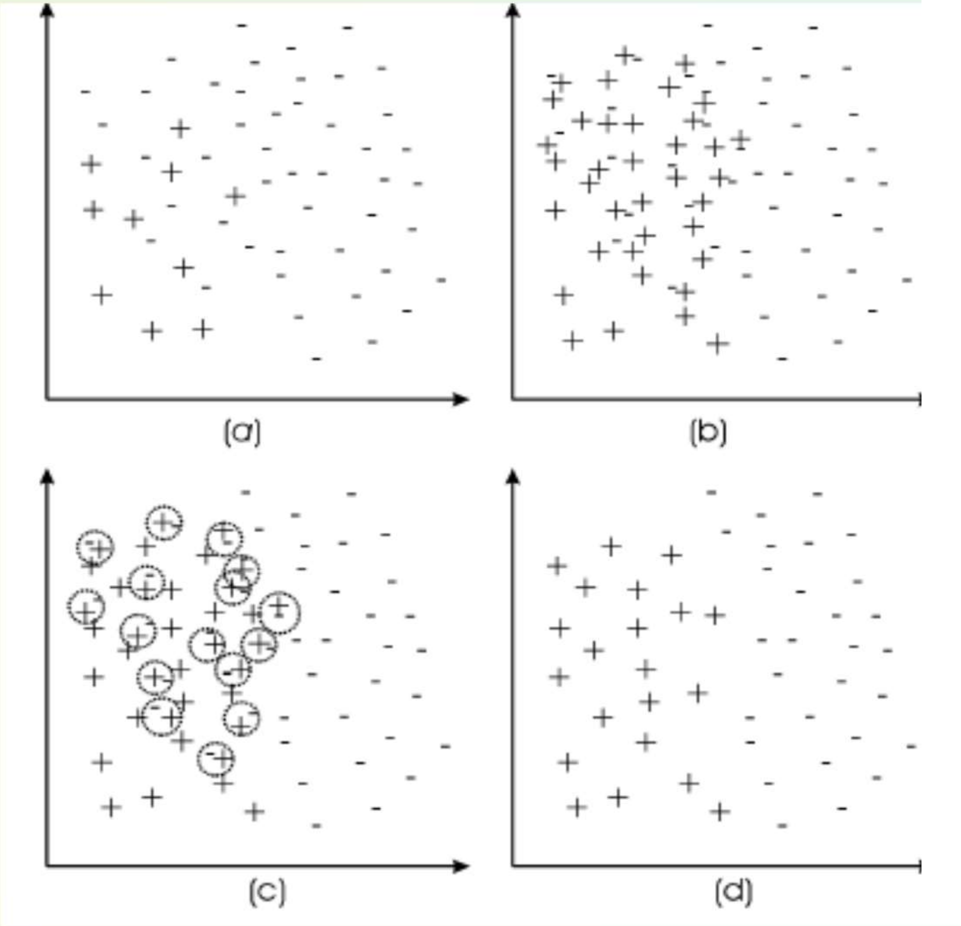
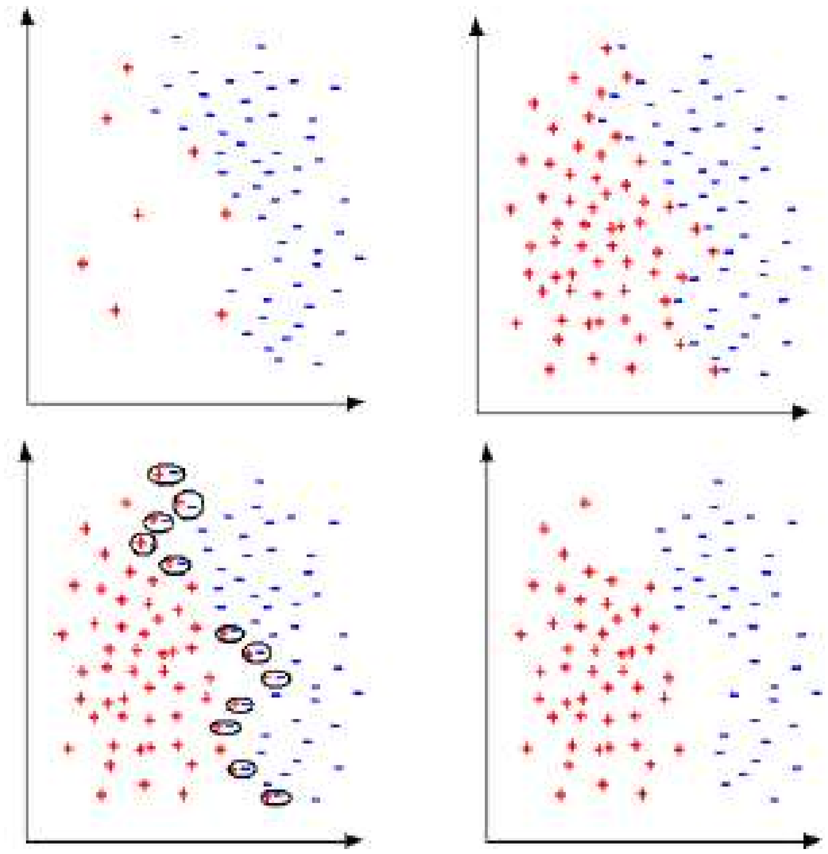


Figure: SMOTE+TomekLink



SMOTE hybridization: SMOTE + ENN

- ENN removes any example whose class label differs from the class of at least two of their neighbors
- ENN remove more examples than the Tomek links does
- ENN remove examples from both classes

SMOTE and hybridization: Analysis

Table 6: Performance ranking for original and balanced data sets for pruned decision trees.

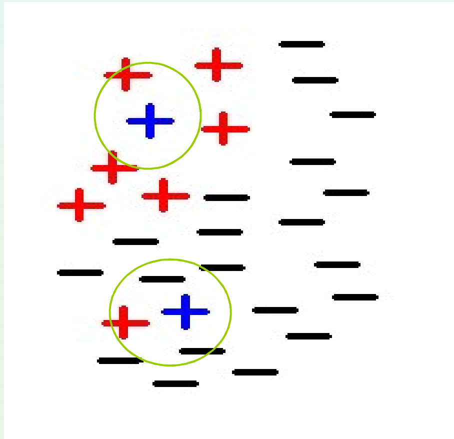
Data set	1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	11°
Pima	Smt	RdOvr	Smt+Tmk	Smt+ENN	Tmk	NCL	Original	RdUdr	CNN+Tmk	CNN*	OSS*
German	RdOvr	Smt+Tmk	Smt+ENN	Smt	RdUdr	CNN	CNN+Tmk*	OSS*	Original*	Tmk*	NCL*
Post-operative	RdOvr	Smt+ENN	Smt	Original	CNN	RdUdr	CNN+Tmk	OSS*	Tmk*	NCL*	Smt+Tmk*
Haberman	Smt+ENN	Smt+Tmk	Smt	RdOvr	NCL	RdUdr	Tmk	OSS*	CNN*	Original*	CNN+Tmk*
Splice-ie	RdOvr	Original	Tmk	Smt	CNN	NCL	Smt+Tmk	Smt+ENN*	CNN+Tmk*	RdUdr*	OSS*
Splice-ei	Smt	Smt+Tmk	Smt+ENN	CNN+Tmk	OSS	RdOvr	Tmk	CNN	NCL	Original	RdUdr
Vehicle	RdOvr	Smt	Smt+Tmk	OSS	CNN	Original	CNN+Tmk	Tmk	NCL*	Smt+ENN*	RdUdr*
Letter-vowel	Smt+ENN	Smt+Tmk	Smt	RdOvr	Tmk*	NCL*	Original*	CNN*	CNN+Tmk*	RdUdr*	OSS*
New-thyroid	Smt+ENN	Smt+Tmk	Smt	RdOvr	RdUdr	CNN	Original	Tmk	CNN+Tmk	NCL	OSS
E.Coli	Smt+Tmk	Smt	Smt+ENN	RdOvr	NCL	Tmk	RdUdr	Original	OSS	CNN+Tmk*	CNN*
Satimage	Smt+ENN	Smt	Smt+Tmk	RdOvr	NCL	Tmk	Original*	OSS*	CNN+Tmk*	RdUdr*	CNN*
Flag	RdOvr	Smt+ENN	Smt+Tmk	CNN+Tmk	Smt	RdUdr	CNN*	OSS*	Tmk*	Original*	NCL*
Glass	Smt+ENN	RdOvr	NCL	Smt	Smt+Tmk	Original	Tmk	RdUdr	CNN+Tmk*	OSS*	CNN*
Letter-a	Smt+Tmk	Smt+ENN	Smt	RdOvr	OSS	Original	Tmk	CNN+Tmk	NCL	CNN	RdUdr*
Nursery	RdOvr	Tmk	Original	NCL	CNN*	OSS*	Smt+Tmk*	Smt*	CNN+Tmk*	Smt+ENN*	RdUdr*

G.E.A.P.A. Batista, R.C. Prati, M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6:1 (2004) 20-29

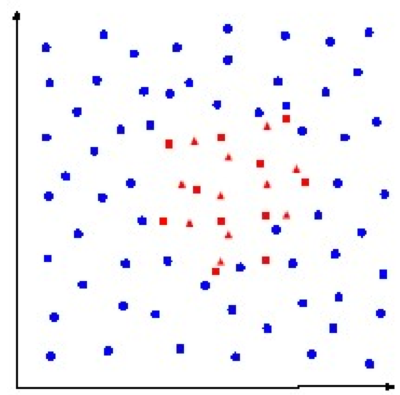
SMOTE Borderline (Han et al. 2005)

- ❑ Examples are not equally important
  Three types of minority class examples DANGER, SAFE, NOISE
- ❑ SN(p,k) – majority class among *k* neighbours of *p*
  - SAFE $\rightarrow$ SN(p,k)<k/2
  - DANGER $\rightarrow$ k/2 $\leq$SN(p,k)<k
  - NOISE $\rightarrow$ SN(p,k)=k

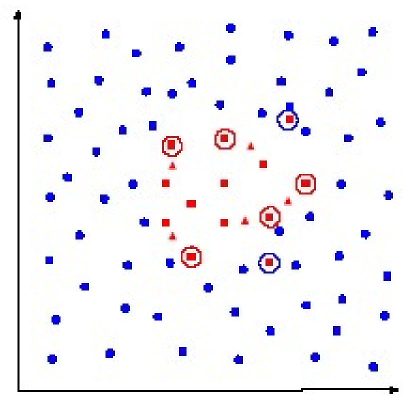
  Over-sample only DANGER
  with SMOTE procedure



- BORDERLINE 1 $\rightarrow$ neighbours from the minority class
- BORDERLINE 2 $\rightarrow$ closest neighbours from both classes



RYSUNEK 6.1: SMOTE

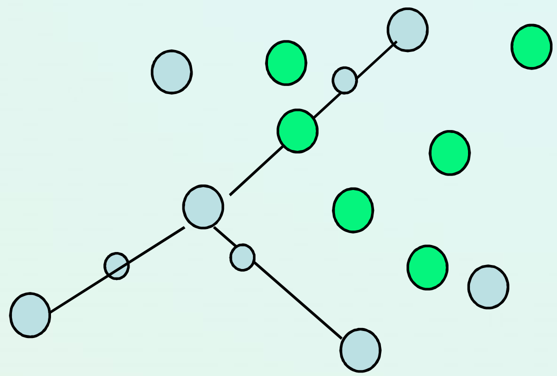


RYSUNEK 6.2: Borderline-SMOTE1

SMOTE → Safe-Level-SMOTE

SMOTE → other shortcomings

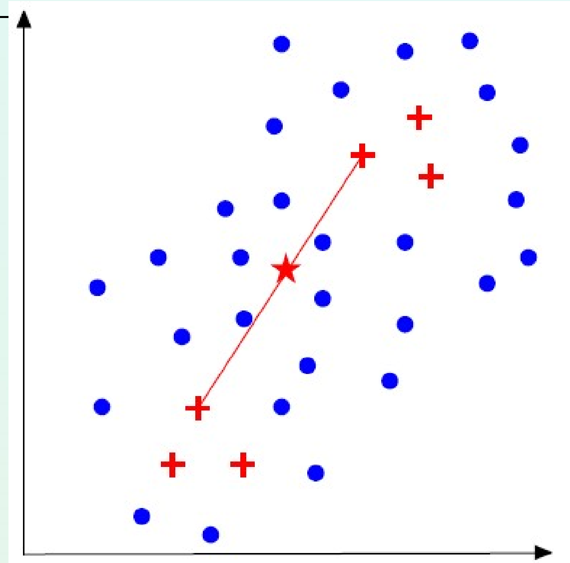
- ❑ Looking for minority class neighbours without regard to the majority class distribution
- ❑ „Blind” over-generalization in the directions of neighbors from majority classes



Safe-Level-SMOTE [2009]

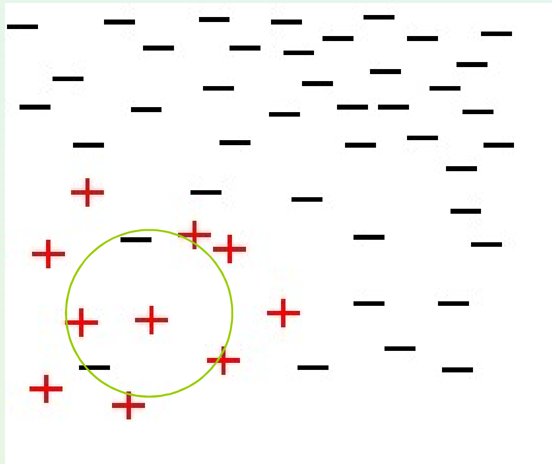
- ❑ Safe level – no. of minority examples among k neighbours of p
- ❑ For neighbour n → compare $sl(p)$ and $sl(n)$ and calculate sl ratio $sl(p)/sl(n)$
- ❑ Generation of new example x closer to the safer region
- ❑ Random gap depends on $sl\ ratio = sl(p)/sl(n)$

Analysing more local neighbourhood



- ❑ Safe level – still looking for *k* neighbours from the minority class!
- ❑ Insufficient – minority class decomposed into distant small sub-part; leads to overlapping and increasing inconsistency

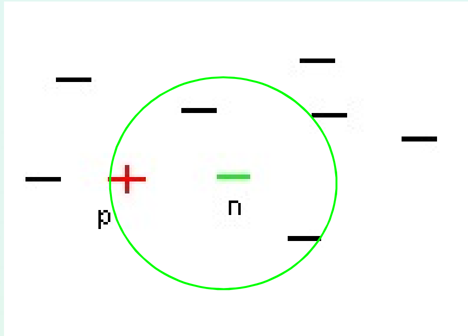
LN - SMOTE



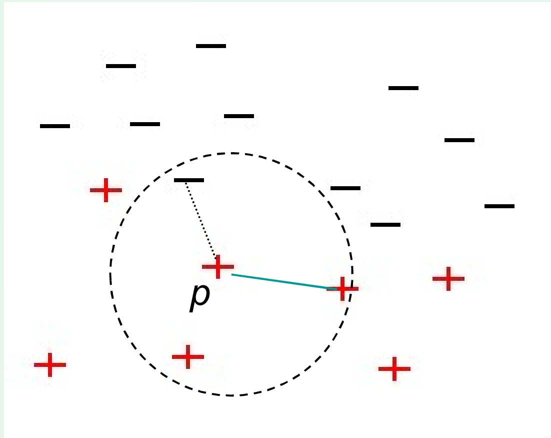
- ❑ Focus on local nearest neighbours of *p* also from the majority class
- ❑ Inspiration of safe levels and idea of generating new *x* toward safer regions
  - No simple adaptation
- ❑ Need for changes in *sl*(*p*) and *sl*(*n*) and in other points

LN – SMOTE / Maciejewski, Stefanowski IEEE CIDM 2011

- Another view on local safe levels ($p \rightarrow n$)
- If neighbour n belongs to the majority class, sl(p)=0 and sl(n)=1 (which is just p) $\rightarrow$ copy x on n
  - Change def. sl(n) $\rightarrow$ skip p and look for k+1 example



- Generation of x if n in the majority class
  - Direct x more to the minority example p by modifying random interval with τ depending on sl(n)/k



- Other changes in the algorithm

  Detailed pseudocode $\rightarrow$ see the paper!

- **LN-SMOTE 2** – combination with edited nearest rule / first remove difficult noisy examples from the majority class

C4.5 → F-measure

	None	SMO	BS1	BS2	SLS	LN1	LN2
Balance scale	0.00	9.29	8.40	11.33	8.58	16.54	16.08
Breast cancer	39.83	43.83	43.02	44.37	45.15	43.83	45.64
Cleveland	19.29	26.71	25.27	28.33	26.03	29.27	29.70
CMC	40.81	41.64	42.05	44.16	41.64	44.95	45.94
Ecoli	58.86	64.31	62.38	64.02	63.98	62.01	66.96
Flags	30.89	44.51	41.35	42.68	43.15	39.46	42.03
Germ. credit	45.51	50.30	49.98	51.01	50.02	50.91	50.46
Haberman	30.36	43.70	41.84	43.58	40.08	44.56	42.59
Hepatitis	49.20	52.10	53.94	53.00	57.10	58.57	57.86
Pima	62.05	65.51	65.68	65.61	65.02	65.13	65.06
Post-operative	5.84	22.03	22.86	19.06	20.56	20.42	19.44
Solar flare	28.79	27.84	28.85	29.93	28.68	31.60	33.08
Transfusion	47.27	48.80	50.05	51.12	48.94	49.19	50.30
Yeast	35.02	39.64	42.23	42.02	40.07	41.39	42.58

- Tuning parameters *k* and *o* →testing several combinations; for each method choose the best one with respect to F-measure
- LN SMOTE – the highest improvements (balance 7.25., solar flare 5.24)
- The best for 11 of 14 sets; LN SMOTE ver 2 > LN SMOTE ver. 1

Other results of LN-SMOTE

Decision trees → Wilcoxon test:

- ❑ F-measure → LN SMOTE outperforms the remaining methods
  - Ver. 2 the best nearly always, LN SMOTE ver 1 the second (3 the best)
  - Then, Borderline 2
- ❑ G-mean → again both LN SMOTE better then others
  - Slightly smaller difference between Ver.1 and Borderline 2
  - LN SMOTE better with respect to specificity

Decision rules → similar to trees

Naive Bayes

- ❑ Generally baseline performs better than symbolic classifiers
  - Improvements of evaluation measures smaller
  - LN SMOTE methods still wining; Superiority more visible for F-measure than G-mean
  - Differences between SMOTE, Borderline1 no significant
- ❑ Other tuning of parameters *k* and *o* → Best combination for SMOTE applied to others
  - LN SMOTE still performs better than others

Rule classifiers and class imbalance

- Data Ecoli: 336 ob. and 35 ob. (M class) ; 7 atr. numerical
- MODLEM (no pruning) 18 rules, with 7 for the minority class

r1.(a7<0.62)&(a5>=0.11) => (Dec=O); [230,76.41%, 100%]

r2.(a1<0.75)&(a6>=0.78)&(a5<0.57) => (Dec=O); [27,8.97%, 100%]

r3.(a1<0.46) => (Dec=O); [148, 148, 49.17%, 100%]

r4.(a1<0.75)&(a5<0.63)&(a2∈[0.49,0.6]) => (Dec=O); [65, 21.59%, 100%]

r5.(a1<0.75)&(a7<0.74)&(a2>=0.46) => (Dec=O); [135, 44.85%, 100%]

r6.(a2>=0.45)&(a6>=0.75)&(a1<0.69) => (Dec=O); [34, 11.3%, 100%]

...

r12.(a7>=0.62)&(a6<0.78)&(a2<0.49)&(a1 ∈[0.57,0.68]) => (Dec=M) [6, 17.14%, 100%]

r13.(a7>=0.62)&(a6<0.76)&(a5<0.65)&(a1 ∈[0.73,0.82]) => (Dec=M)[7, 20%, 100%]

r14.(a7>=0.74)&(a1>=0.47)&(a2>=0.45)&(a6<0.75)&(a5>=0.59) => (Dec=M); [3, 8.57%, 100%]

r15.(a5>=0.56)&(a1>=0.49)&(a2 ∈[0.42,0.44]) => (Dec=M); [3, 8.57%, 100%]

r16.(a7>=0.74)&(a2 ∈[0.53,0.54]) => (Dec=M); [2, 5.71%, 100%]

...

- Classification strategies:
  - Multiple matching? Voting with supports
  - No matching? – partical matching or nearest rules

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
  - Also difficulties with other strategies (m-estimate, nearest rules, etc.)
- Grzymała's proposal (2000) $\rightarrow$ leave the rule induction but change the classification strategy!
- Changing strength / support 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

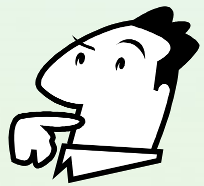
Changing set of rules for the minority class

- 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 → more numerous and with greater strength
- The chance of using these rule while classifying new objects is increased
- The use of EXPLORE (Stefanowski, Vanderpooten 94):
  - Induce all rules with strength greater then a threshold.
  - Modify the threshold considering gain + conditions calculated from 1 stage

Motivations for other approach to imbalance data

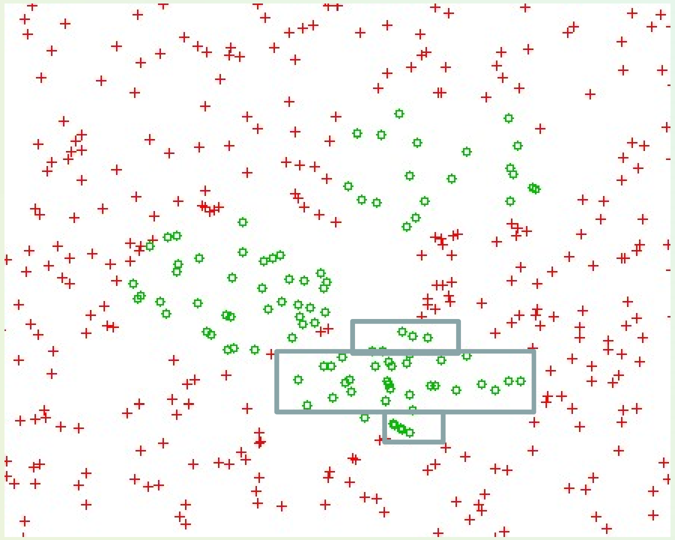
- The „replace rules" approach is focused on handling „cardinality" aspects of imbalance.
  - Strengthening some sub-regions and leaving uncovered examples.
  - Some difficult examples may be uncovered depending on the procedure for tuning parameters
    - which is time consuming and sophisticated.

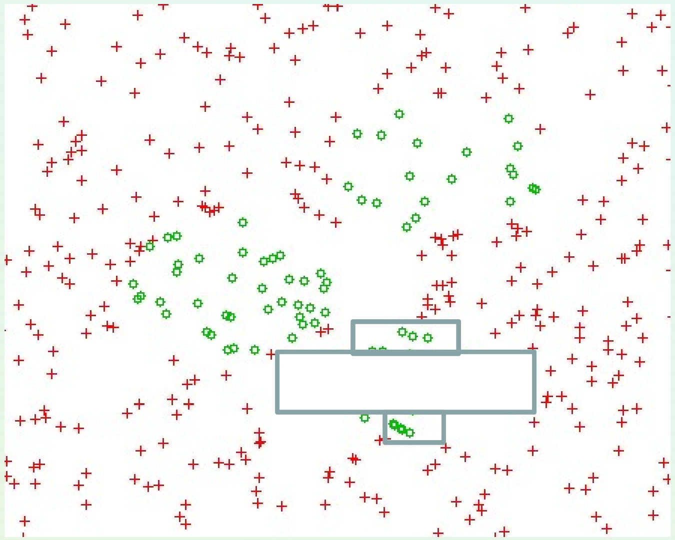
- However, one may focus on other characteristics of learning examples.



Rule induction - limitations for the minority class

- Greedy sequential covering and top down approach
  $\rightarrow$ data fragmentation + small disjuncts
- Connected with evaluation criteria in search
  (biased toward the majority classes)





Skipping covered examples

Motivations for our study $\rightarrow$
K.Napierała, J. Stefanowski: BRACID A comprehesive approach to rule induction from imbalanced data. Int. Journal of Intelligent Information Systems. 2012

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.) …

BRACID

Bottom-up induction of **R**ules **A**nd **C**ases from **I**mbalanced **D**ata

Assumptions:

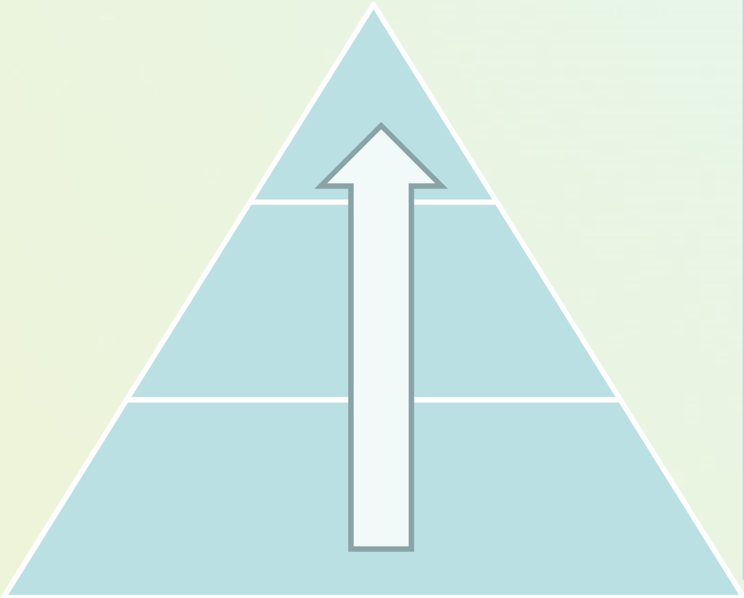
- Hybrid knowledge representation: rule and instances
- Induction rules by bottom-up strategy
- Resigning from greedy sequential covering
- Some inspirations from RISE [P.Domingos 1996]
- Considering info about types of difficult examples
- Local neighbors with HVDM
- Internal evaluation criterion (F-miara)
- Local nearest rules classification strategy

More →
K.Napierała, J. Stefanowski: BRACID A comprehesive approach to rule induction from imbalanced data. Int. Journal of Intelligent Information Systems. 2012

BRACID

Bottom-up induction of Rules And Cases from Imbalanced Data

- Bottom-up
- Non-sequential covering
- evaluation of new rules with F-measues

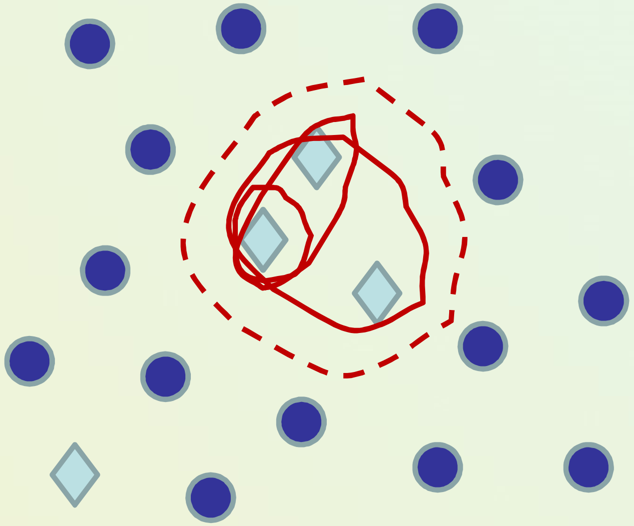


```
BRACID(Examples ES)
1 RS = ES
2 Ready_rules = empty_set
3 Labels  = Calculate labels for minority class examples
4 Iteration=0

5 Repeat
6   For each rule R in RS not belonging to Ready_rules
7      If R's class is minority class
8         Find Ek=k nearest examples to R not already covered
          by it,  and of R's class
9         If Labels[R's seed]=safe
10            Improved = AddBestRule(Ek, R,RS)
11        Else
12            Improved = AddAllGoodRules(Ek,R,RS)
13        If Improved=false and not Iteration=0
14            Extend (R)
15            Add R to Ready_rules
16     Else     #R's class is majority class
17        Find Ek=k nearest examples to R not already
          covered by it and of R's class
18        Improved = AddBestRule(Ek, R,RS, Label[R's seed])
19         If Improved=false
20             If Iteration=0    #Treat as noise
21        Remove R from RS and R's seed from ES
22             Else
23        Add R to Ready_rules
24 Until any rule improves evaluation

25 Return RS
```

BRACID

Bottom-up induction of Rules And Cases from Imbalanced Data

Generalize the rule



```
BRACID(Examples ES)
RS = ES
Ready_rules = empty_set
Labels  = Calculate labels for minority class examples
Iteration=0

Repeat
  For each rule R in RS not belonging to Ready_rules
      If R's class is minority class
Find Ek=k nearest examples to R not already covered by it,
                                      and of R's class
         If Labels[R's seed]=safe
             Improved = AddBestRule(Ek, R,RS)
         Else
             Improved = AddAllGoodRules(Ek,R,RS)
         If Improved=false and not Iteration=0
              Extend (R)
        Add R to Ready_rules
      Else    #R's class is majority class
      Find Ek=k nearest examples to R not already covered by it,
                                      and of R's class
     Improved = AddBestRule(Ek, R,RS, Label[R's seed])

        If Improved=false
             If Iteration=0    #Treat as noise
                Remove R from RS and R's seed from ES
             Else
                Add R to Ready_rules
Until any rule improves evaluation measure

Return RS
```

BRACID – experiments

Different classifiers - sensitivity

Zbiór	BRACID	RISE	kNN	C45.rules	CN2	PART	RIPPER	Modlem	Modlem-C
abalone	**0,47**	0,13	0,14	0,34	0,16	0,19	0,18	0,25	0,27
b-cancer	**0,57**	0,36	0,26	0,33	0,28	0,41	0,29	0,32	0,41
car	0,78	0,60	0,03	0,75	0,54	**0,90**	0,53	0,79	0,79
cleveland	**0,48**	0,15	0,04	0,18	0,00	0,25	0,16	0,08	0,14
cmc	**0,63**	0,29	0,31	0,40	0,10	0,38	0,07	0,26	0,36
credit-g	**0,80**	0,36	0,37	0,37	0,26	0,48	0,21	0,36	0,55
ecoli	**0,79**	0,50	0,58	0,60	0,18	0,42	0,45	0,40	0,46
haberman	**0,67**	0,22	0,18	0,24	0,18	0,33	0,18	0,24	0,41
hepatitis	**0,76**	0,49	0,47	0,36	0,05	0,46	0,42	0,38	0,55
new-thyroid	**0,98**	0,93	0,87	0,85	0,87	0,93	0,86	0,81	0,84
solar-flareF	**0,52**	0,07	0,00	0,15	0,00	0,19	0,01	0,07	0,19
transfusion	**0,74**	0,30	0,32	0,39	0,15	0,43	0,09	0,37	0,50
vehicle	**0,96**	0,83	0,87	0,87	0,33	0,88	0,87	0,86	0,92
yeast-ME2	**0,55**	0,24	0,19	0,32	0,00	0,27	0,26	0,19	0,21

Comparing classifiers - G-mean

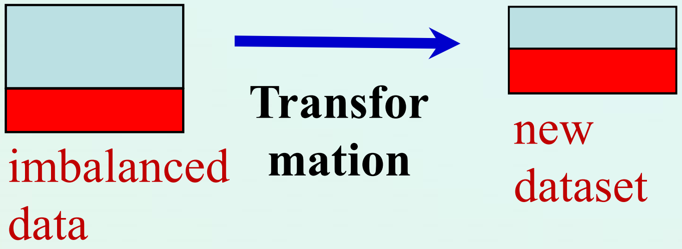
Zbiór	BRACID	RISE	kNN	C45.rules	CN2	PART	RIPPER	Modlem	Modlem-C
abalone	**0,65**	0,34	0,36	0,57	0,40	0,42	0,42	0,48	0,51
b-cancer	**0,56**	0,54	0,47	0,49	0,46	0,53	0,48	0,49	0,53
car	0,87	0,75	0,08	0,86	0,71	**0,94**	0,71	0,88	0,88
cleveland	**0,57**	0,23	0,08	0,26	0,00	0,38	0,26	0,15	0,23
cmc	**0,64**	0,51	0,52	0,59	0,26	0,54	0,25	0,47	0,54
credit-g	0,61	0,54	0,57	0,55	0,47	0,60	0,44	0,56	**0,65**
ecoli	**0,83**	0,64	0,70	0,72	0,28	0,55	0,59	0,57	0,63
haberman	**0,58**	0,38	0,33	0,43	0,35	0,47	0,36	0,40	0,53
hepatitis	**0,75**	0,60	0,62	0,51	0,05	0,55	0,50	0,50	0,64
new-thyroid	**0,98**	0,95	0,92	0,90	0,92	0,95	0,91	0,88	0,90
solar-flareF	**0,64**	0,14	0,00	0,27	0,00	0,32	0,02	0,13	0,32
transfusion	**0,64**	0,51	0,53	0,58	0,34	0,60	0,27	0,53	0,58
vehicle	**0,94**	0,90	0,91	0,91	0,51	0,92	0,92	0,92	**0,94**
yeast-ME2	**0,71**	0,44	0,34	0,51	0,00	0,42	0,45	0,34	0,37

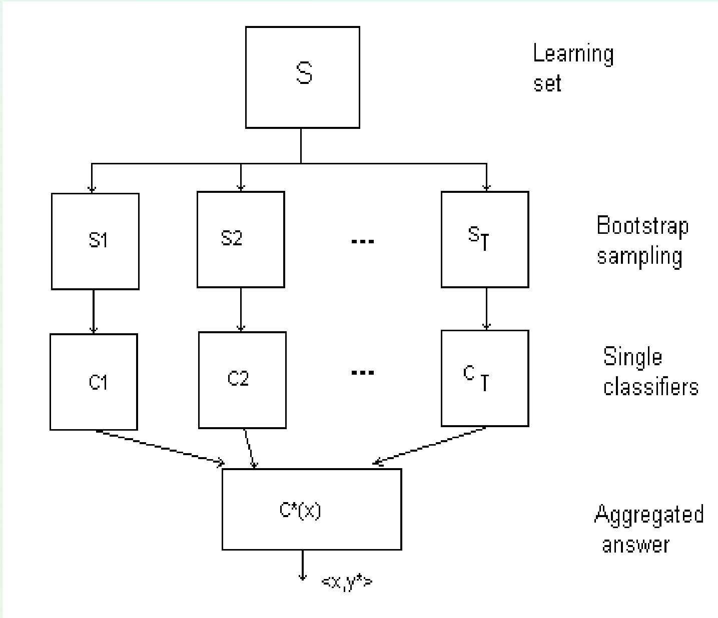
BRACID – summary

- BRACID improves recognition of the minority class
- Also G-mean, F-measure, and others
- Better than other rule classifiers
- Competitive to SMOTE/ENN used with rule classifiers
- Usually more rules but with higher supports
- Testing examples – good for border and rare ones (+ safe)
- However, still think about
  - Other classification strategies
  - Possible reducing a number of considered rules

Generalizations of Ensembles

- Data preprocessing + ensemble
  - Boosting-based
    - SMOTEBoost, DataBoost
  - Bagging-based
    - Exactly Balanced Bagging
    - Roughly Balanced Bagging
    - OverBagging
    - UnderOverBagging
    - SMOTEBagging
    - Ensemble Variation
  - IIvotes
- Others or Hybrid (EasyEnsemble)
- Cost Sensitive Boosting
  - AdaCost (C1-C3)
  - RareBoost





Related: Galar et. al., A Review on Ensembles for the Class Imbalance Problem. IEEE Trans. 2011

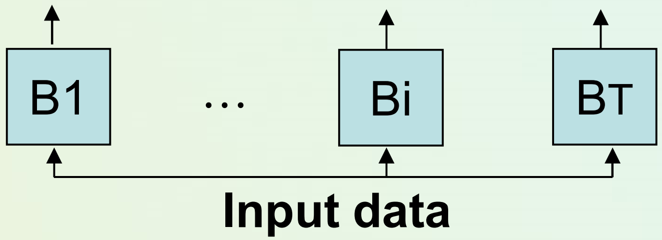
Evaluation of New Ensembles

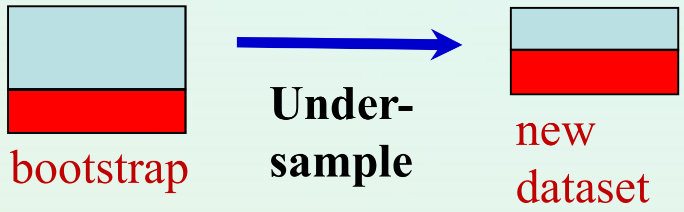
Limited comparative studies [2011]

- Galar, Herrera et al → 20 classifiers over 44 datasets
  - Simpler pre-processing generalizations better then more complex or cost based ones
  - SMOTEBagging, RUBagging, RUBoost the best ones
- Khoshgoftaar et al. → imbalanced and noisy data
  - EBBag, RBBag better then SMOTEBoost and RUBoost
- Błaszczyński, Stefanowski, Idkowiak: Extending bagging for imbalanced data. CORES2013.
  - EBBag, RBBag better then SMOTEBag and other oversampling versions of bagging

Generalizations of Bagging

- Standard Bagging → boostraps
  - sampling N examples (with replacements) equal probability





Undersampling modifications

- Exactly Balanced Bagging [Ch03]
  - bootstrap samples = copy of the minority class + randomly drawn subset of the majority class (N_maj = N_min)
- Rough Balanced Bagging [Hido 09]
  - Equal probabilities of class sampling → BS_maj
  - Sampling with replacement N_min and BS_maj

Overbagging Modifications of Bagging

Introduce more minority examples in the boostrap

- OverBag → boostrap sampling + random copying minority class until balancing classes
- SMOTEBag [WY09] → SMOTE with changing its ratio for each sample / classifier (increase its diversity)
- BagSMOTE → SMOTE with fixed ratio to balance classes

Experimental Setup [Stef 2013]

- Aim → to evaluate best versions of bagging
  - EBBag , RBBag
  - OverBag, SMOTEBag, BagSMOTE
  - Standard bagging as a baseline
- Measures → sensitivity, specificity, G-mean
- Base classifiers → decision trees J4.8 (unprunned)
  - T = 20, 50 and 100
- Design of experiments:
  - 10-fold stratified cross validation (repeated 5 x),
  - 22 UCI imbalanced data sets
  - Implementation with WEKA
- Statistical analysis – Friedman and Wilcoxon tests

Results of the Comparative Study

- RBBag and EBBag outperform all oversampling extensions
- Sensitivity

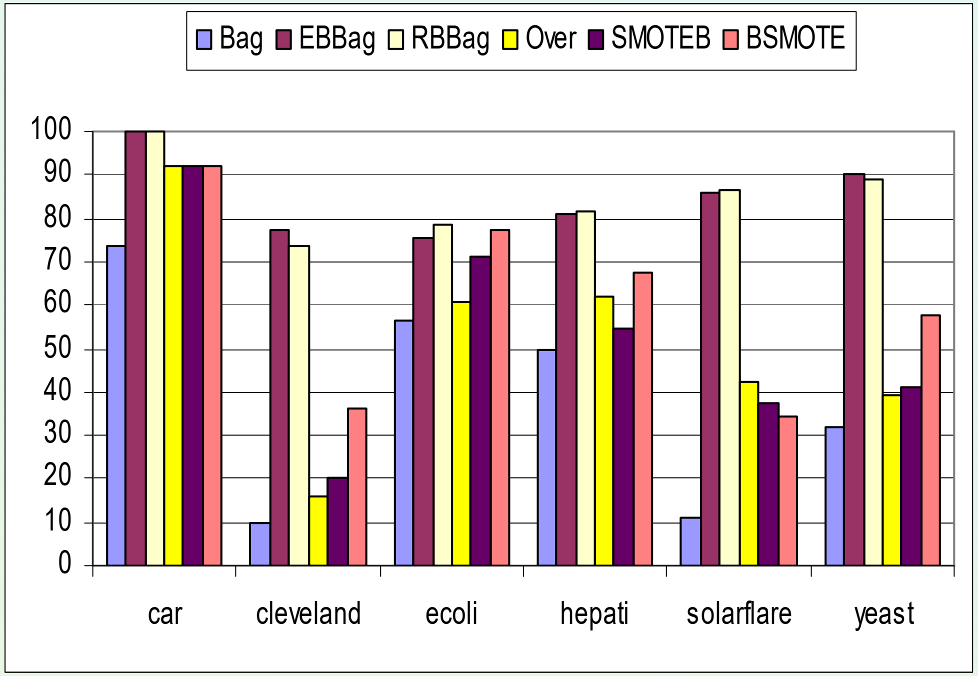
RBBag ≈ **EBBag** >BagSMOTE≈OverBag> SMOTEBag>Bagging

- G-mean

RBBag > EBBag>BagSMOTE≈OverBag> SMOTEBag>Bagging

- F-measure
  Wilcoxon : **RBBag** > EBBag
- Related works → Undersampling better; SMOTEBag does not work; use replacemnt
- Q-statistics (diversity)
  - Rather not too high
  - Best extensions less diversified than SMOTEBag

Sensitivity



Detailed tables inside the paper
Błaszczyński, Stefanowski, Idkowiak:
Extending bagging for imbalanced data. CORES2013

Cost-sensitive learning

Cost modification consists of weighting errors made on examples of the minority class higher than those made on examples of the majority class in the calculation of the training error.

This, in effect, rectifies the bias given to the majority class by standard classifiers when the training error corresponds to the simple (non-weighted) accuracy.

B. Zadrozny, J. Langford, N. Abe, Cost-sensitive learning by cost—proportionate example weighting, in: Proceedings of the 2003 IEEE International Conference on Data Mining (ICDM'03), 2003.

C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2001, pp. 973–978.

Cost-sensitive learning

Needs a cost matrix, which encodes the penalty of classifying samples from one class as another.

- Traditionally assumed a cost matrix of the form:

	True = 0	True = 1
Predict = 0	$C(0,0)$	$C(0,1)$
Predict = 1	$C(1,0)$	$C(1,1)$

- cost that depends on particular example x

	True = 0	True = 1
Predict = 0	$C(0,0,x)$	$C(0,1,x)$
Predict = 1	$C(1,0,x)$	$C(1,1,x)$

Cost-sensitive learning

Two weighting approaches

- Up-weighting, analogous to over-sampling, increases the weight of one of the classes keeping the weight of the other class at one

- Down-weighting, analogous to under-sampling, decreases the weight of one of the classes keeping the weight of the other class at one

Cost-sensitive learning

- Transparent box → need of how the algorithm works
  - Eg: specific cost-sensitive algorithms, some of the weighting approaches, threshold modyfing

Ting, K.M. An instance-weighting method to induce cost-sensitive trees (2002) *IEEE Transactions on Knowledge and Data Engineering*, 14 (3), pp. 659-665.

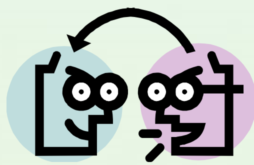
- Black box → don't need to know how the algorithm works
  - Eg: Data-level approaches, MetaCost, some boosting approaches

P. Domingos, Metacost: a general method for making classifiers cost sensitive, in: Advances in Neural Networks, International Journal of Pattern Recognition and Artificial Intelligence, San Diego, CA, 1999, pp. 155–164.

Y. Sun, M. S. Kamel, A. K. C. Wong and Y. Wang, Cost-sensitive boosting for classification of imbalanced data, *Pattern Recognition 40(12) (2007) 3358–3378*

Some questions for discussing

- Better understanding the imbalance problem
- Studies with simulated or real data
  - Noise
  - Other local analysis than k-NN
- Impact on constructing new approaches
  - Pre-processing methods
  - New ensembles
- Real model simulation of rare examples
- Evaluation issues
- Incremental learning
- Data shift and drifting concepts



Some references

- J.Błaszczyński, M.Deckert, J.Stefanowski, Sz.Wilk: Integrating Selective Pre-processing of Imbalanced Data with Ivotes Ensemble. RSCTC 2010, LNAI vol. 6086, Springer Verlag 2010, 148-157
- J.W. Grzymala-Busse, J.Stefanowski, S. Wilk: A Comparison of Two Approaches to Data Mining from Imbalanced Data, Proc. of the 8th Int. Conference KES 2004, Lecture Notes in Computer Science, vol. 3213, Springer-Verlag, 757-763
- K.Napierała, J.Stefanowski: Identification of Different Types of Minority Class Examples in Imbalanced Data. Proc. HAIS 2012, Part II, LNAI vol. 7209, Springer Verlag 2012, 139–150.
- K.Napierała, J.Stefanowski, Sz.Wilk: Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. RSCTC 2010, LNAI vol. 6086, 2010, 158-167
- K. Napierała, J. Stefanowski: BRACID Journal of Intelligent Information Systems 2013
- T. Maciejewski, J. Stefanowski: Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data. Proc. of IEEE Symposium on Computational Intelligence and Data Mining, SSCI IEEE, April 11-15, 2011, Paris, IEEE Press, 104—111
- J.Stefanowski, S. Wilk: Rough sets for handling imbalanced data: combining filtering and rule-based classifiers. Fundamenta Informaticae, vol. 72, no. (1-3) July/August 2006, 379-391.
- J.Stefanowski, Sz.Wilk: Improving Rule Based Classifiers Induced by MODLEM by Selective Pre-processing of Imbalanced Data. Proceedings of the RSKT Workshop ECML/PKDD, 2007, 54-65.
- J.Stefanowski, Sz.Wilk: Selective pre-processing of imbalanced data for improving classification performance. Proc. of 10th Int. Conf. *DaWaK 2008,LNCS* vol. 5182, Springer Verlag, 2008, 283-292.
- And more

Thanks for your attention
You are invited for „consultating"

Contact, remarks:
Jerzy.Stefanowski@cs.put.poznan.pl
http://www.cs.put.poznan.pl/jstefanowski