Odkrywanie wiedzy klasyfikacyjnej z niezbalansowanych danych

Learning classifiers from imbalanced data

Wykład spec. projekt eksploracji danych

JERZY STEFANOWSKI

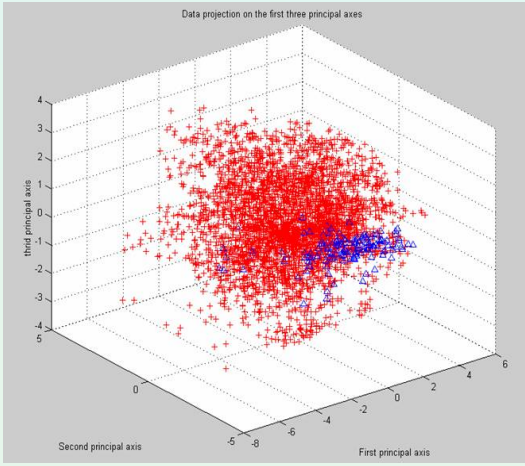
Instytut Informatyki
Politechnika Poznańska
Poznań

Poznań, 2020

Plan wykładu

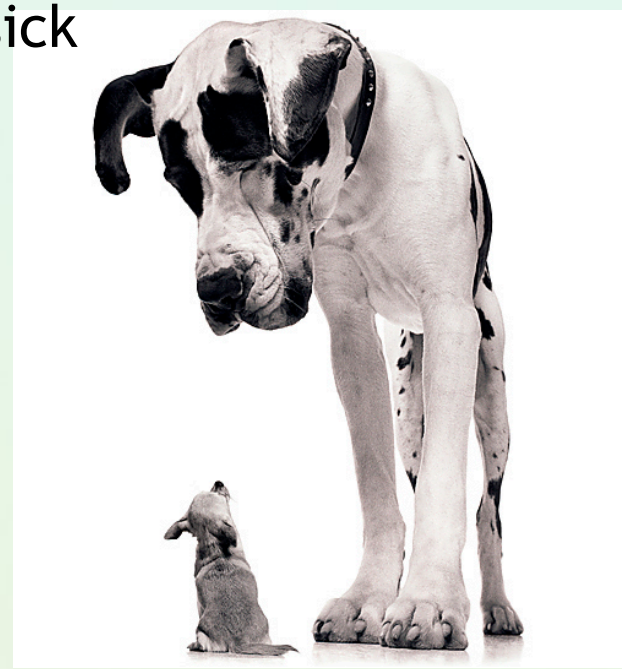
Dzisiaj cz. 1

1. Niezbalansowanie klas (przykłady; miary oceny – wstęp)
2. Czynniki trudności i charakterystyka danych
3. Kategoryzacja metod + dostęp do ich implementacji
4. Przetwarzanie wstępne
   - Under-, over- sampling, SMOTE
   - Metody hybrydowe
5. Wybrane modyfikacje algorytmów
   1. BRACID – reguły
   2. Cost sensitive learning
   3. Zespoły klasyfikatorów (RBB i inne generalizacje)
6. Ocena klasyfikatorów i metod pre-processing
7. Inne zagadnienia i wyzwania

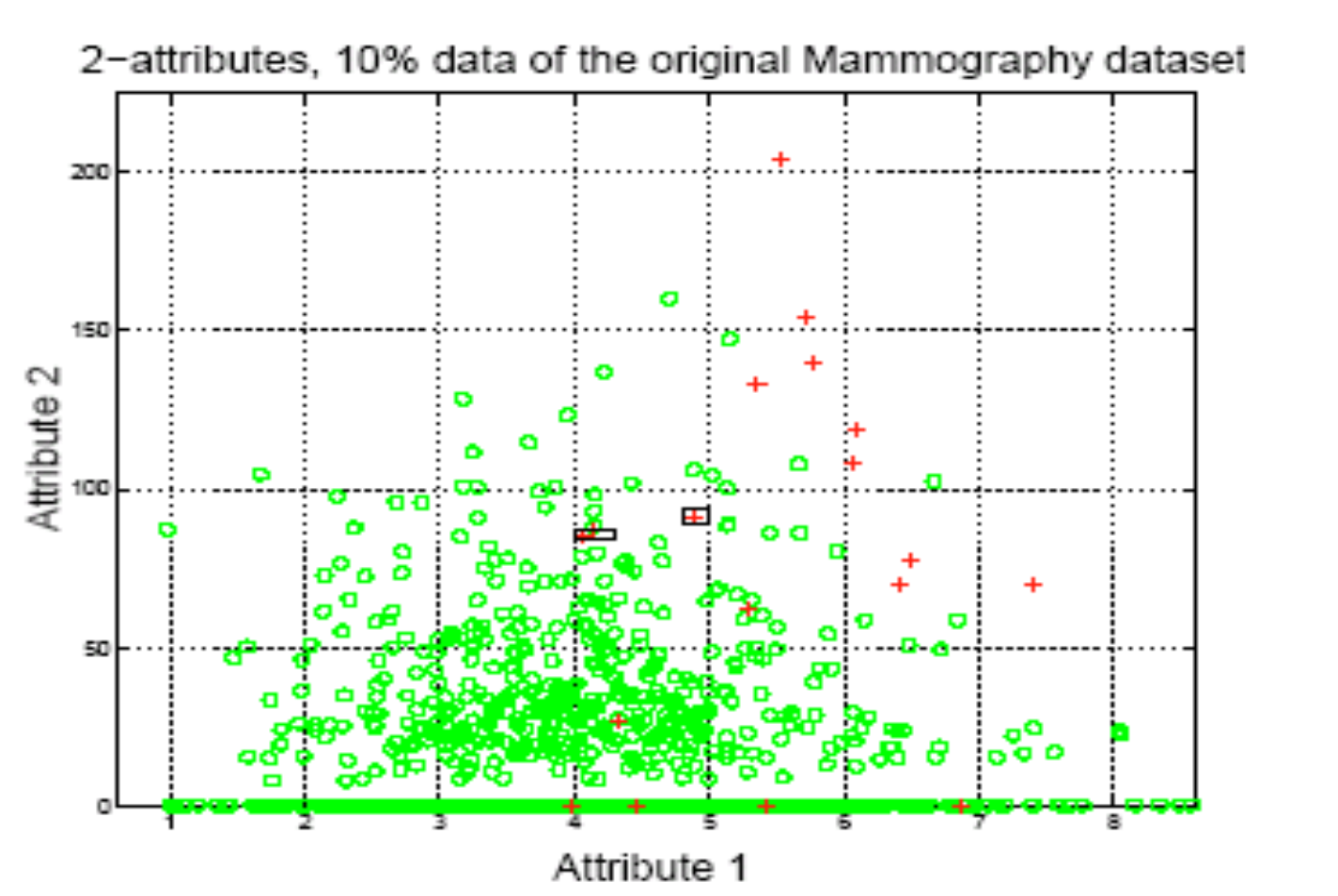


Uczenie się klasyfikatorów z niezbalansowanych danych

- Zadajmy pytanie o rozkład przykładów w klasach w zbiorze uczącym
- Standardowe założenie:
  - Dane są zrównoważone – rozkłady liczności przykładów w klasach względnie podobne
  - ***Przykład:*** „A database of sick and healthy patients contains as many examples of sick patients as it does of healthy ones.”
- Czy takie założenie jest realistyczne?



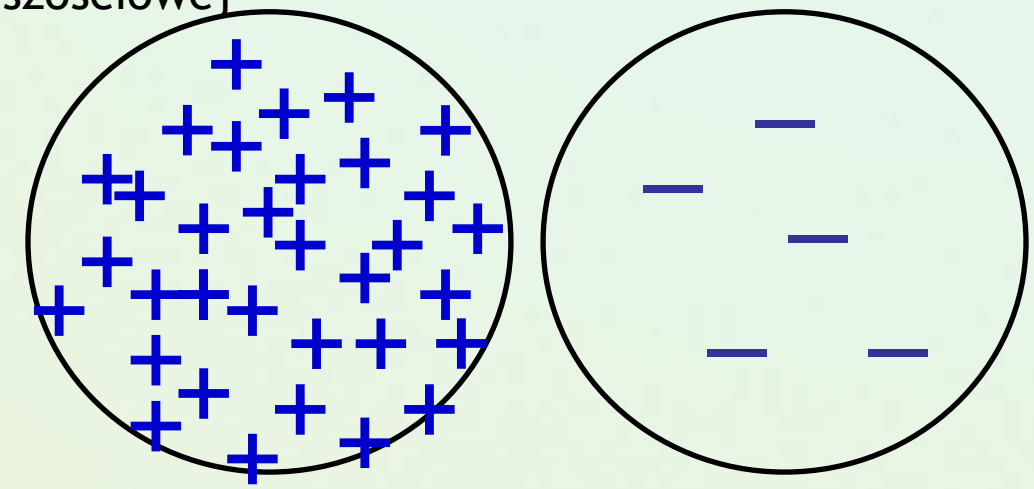
Przykład danych medycznych Chawla et al. SMOTE 2002



Dane – 11183 przykładów, 6 atrybutów, klasa mniejszościowa 2.3%

Niezbalansowanie rozkładu przykładów w klasach

- Dane są niezbalansowane (imbalanced) jeśli klasy nie są w przybliżeniu równo liczne
  - Klasa mniejszościowa (minority class) zawiera wyraźnie mniej przykładów niż inne klasy
- Przykłady z klasy mniejszościowej są często najważniejsze i ich poprawne rozpoznawanie jest głównym celem.
  - Rozpoznawanie rzadkiej, niebezpiecznej choroby
- CLASS IMBALANCE → powoduje trudności w fazie uczenia i obniża zdolność predykcyjną
  - Niektóre klasyfikatory pomimo wysokiej globalnej trafności nie rozpoznają kl. mniejszościowej



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

Przykłady rzeczywistych problemów

- Niezbalansowanie klas naturalne w :
  - 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

- Przegląd innych problemów i zastosowań
  - 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.

Globalne niezbalansowanie (Imbalance Ratio)

- Naturalnie rozważany problem binarny – klasa mniejszościowa -> specjalne znaczenie w zastosowaniu
  - Przykład: diagnoza rzadkiej, lecz niebezpiecznej choroby; błędne nierozpoznanie chorego pacjenta ważniejsze niż sytuacja odwrotna
  - Problemy wieloklasowe – wyzwania w kolejnym wykładzie cz 2.
- Prosta charakterystyka – stopień niezbalansowania
  - N_W – liczba przykładów z klasy większościowej
  - N_M – liczba przykładów z klasy mniejszościowej

  Różne definicje w literaturze
    - IR = N_W / N_M (ile razy większa klasa W)
    - IR [%] – jaki procent N_M w całości N_M+N_W
- Brak wyraźnej granicy IR, kiedy zbiór jest mniejszościowy
  - Może być 15%, 10%, 5%, 1% itd

Przykład charakterystyk benchmark data

Dataset	No of examples	Imbalance ratio [%]	No of attributes (numeric)	Minority class name
breast-w	699	34.47	9(9)	malignant
abdominal-pain	723	27.94	13 (0)	positive
acl	140	28.57	6 (4)	1
new-thyroid	215	16.28	5 (5)	hyper
vehicle	846	23.52	18 (18)	van
nursery	12960	2.53	8(0)	very-recom
satimage	4435	9.35	36(36)	4
car	1728	3.99	6 (0)	good
scrotal-pain	201	29.35	13 (0)	positive
credit-g	1000	30	20 (7)	bad
ecoli	336	10.42	7 (7)	imU
hepatitis	155	20.65	19 (6)	die
ionosphere	351	35.89	34 (34)	bad
haberman	306	26.47	3 (3)	died
cmc	1473	22.61	9 (2)	l-term
breast-cancer	286	29.72	9 (0)	rec-events
cleveland	303	11.55	13 (6)	positive
glass	214	7.94	9 (9)	v-float
hsv	122	11.48	11 (9)	4.0
abalone	4177	8.02	8 (7)	0-4 16-29
postoperative	90	26.66	8 (0)	S
seismic-bumps	2584	6.57	18(14)	1
solar-flare	1066	4.03	12 (0)	F
transfusion	748	23.8	4 (4)	yes
yeast	1484	3.44	8 (8)	ME2
balance-scale	625	7.84	4(4)	B

Za artykuł: K.Napierała, J.Stefanowski:
Types of minority class examples and their influence on learning classifiers. JIIS (2016)

Jak oceniać klasyfikatory dla niezbalansowanych danych?

- Standardowa trafność bezużyteczna
  - Wyszukiwanie informacji (klasa mniejszościowa ~ 1%) → ogólna trafność klasyfikowania ~100%, lecz źle rozpoznawana wybrana klasa
- Miary powiązane z klasą mniejszościową
  - Analiza binarnej macierzy pomyłek confusion matrix
  - Sensitivity i specificity → G-mean
  - ROC curve analysis (AUC)
  - Cost-Precision curves

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

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

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

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

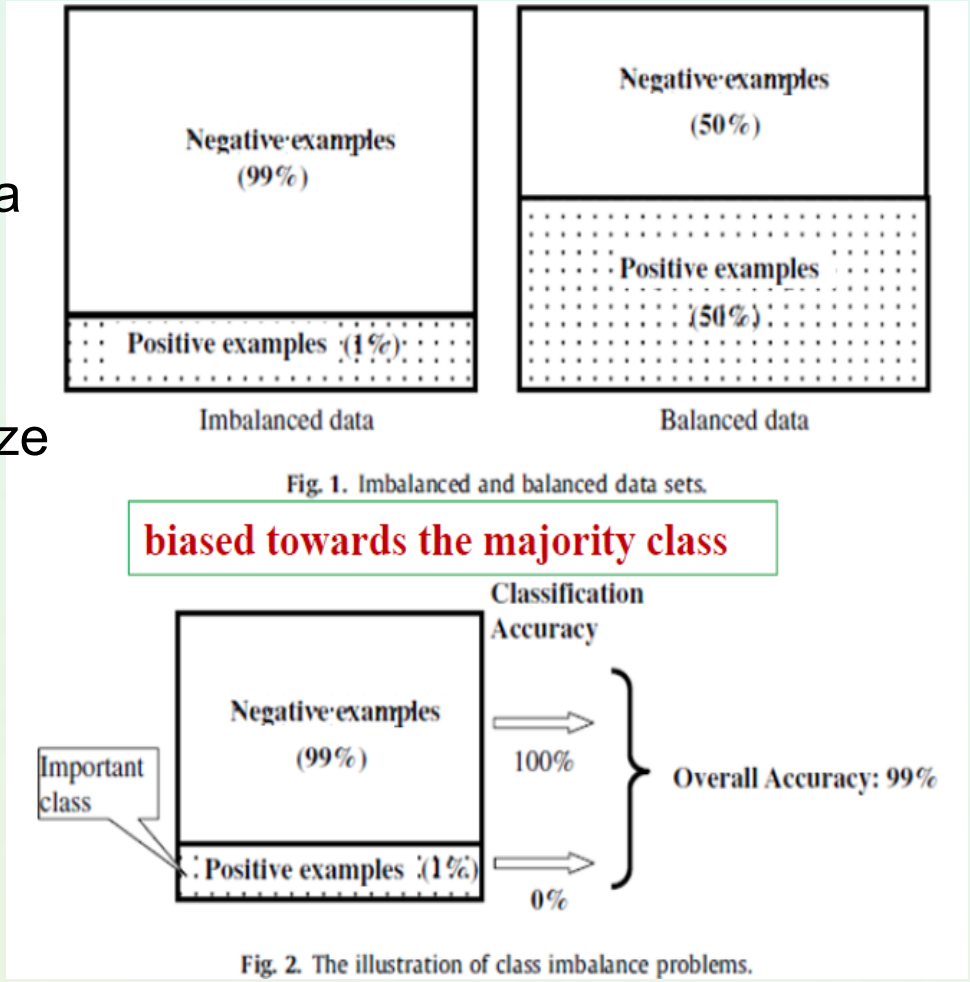
$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

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

Więcej informacji później

Standardowe klasyfikatory?

- Standardowe algorytmy uczące – zakłada się w przybliżeniu zrównoważenie klas
- Typowe strategie przeszukiwania optymalizują globalne kryteria (błąd, miary entropii, itp.)
  - Przykłady uczące są liczniej reprezentowane przy wyborze hipotez
- Metody redukcji (pruning) faworyzują przykłady większościowe
- Strategie klasyfikacyjne ukierunkowane na klasy większościowe



Słaba skuteczność klasyfikatorów

Table 1. Characteristics of evaluated data sets (N – the number of examples, N_A – the number of attributes, C – the minority class, N_C – the number of examples in the minority class, N_O – the number of examples in the majority class, $R_C = N_C/N$ – the ratio of examples in the minority class)

Data set	N	N_A	C	N_C	N_O	R_C
Acl	140	6	with knee injury	40	100	0.29
Breast cancer	286	9	recurrence-events	85	201	0.30
Bupa	345	6	sick	145	200	0.42
Cleveland	303	13	positive	35	268	0.12
Ecoli	336	7	imU	35	301	0.10
Haberman	306	3	died	81	225	0.26
Hepatitis	155	19	die	32	123	0.21
New-thyroid	215	5	hyper	35	180	0.16
Pima	768	8	positive	268	500	0.35

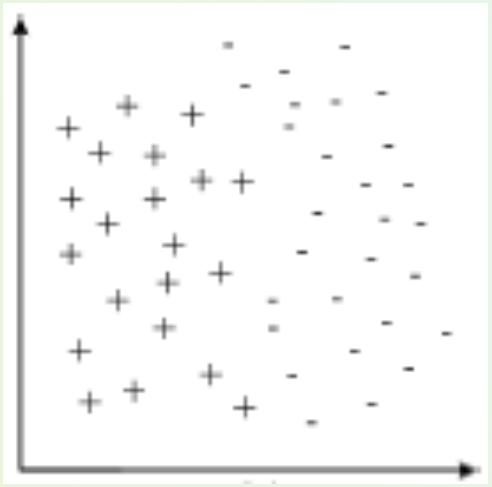
Sensitivity klasy mniejszościowej

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

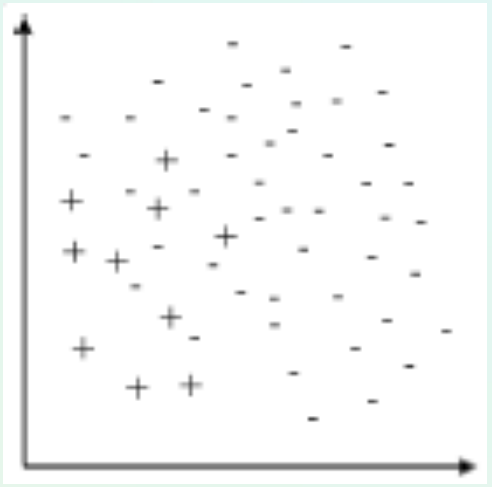
Lepsze rozpoznawania tylko Acl I New Thyroid

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

Na czym polega trudność?



Łatwiejszy problem



Trudniejszy

Źródła trudności:

- Zbyt mało przykładów z klasy mniejszościowej (IR),
- „Zaburzenia” brzegu klas,
- Segmentacja klasy
- ...

Klasa większ. „nakłada” się na mniejszościowe:

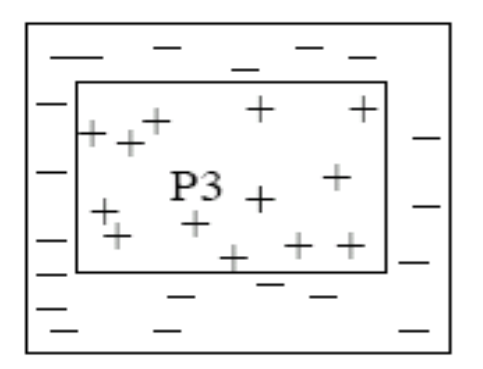
- ❑ Niejednoznaczne przykłady brzegowe
- ❑ Outliers and rare cases
- ❑ Wpływ „szumu” (noisy examples)

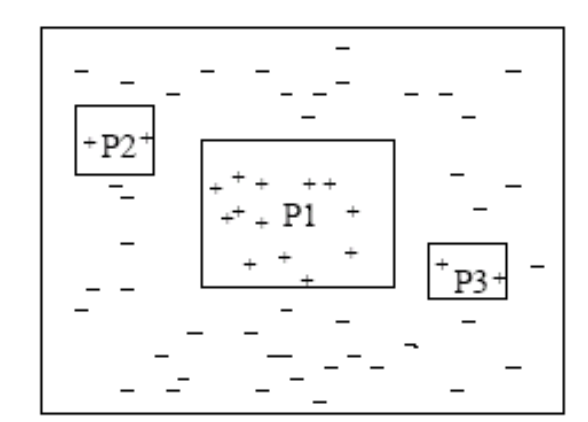
Przeglądowe prace:

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

Czy zawsze „niezbalansowanie” jest trudnością?

- Przeanalizuj studia eksperymentalne N.Japkowicz lub przeglądy G.Weiss – nie wszystkie niezbalansowane dane są trudne dla standardowych algorytmów.
- Japkowicz „The minority class contains small sub-clusters of interesting examples surrounded by other examples” (pełnią rolę tzw, small „disjuncts”, które częściej prowadzą do błędnych decyzji – Holte)





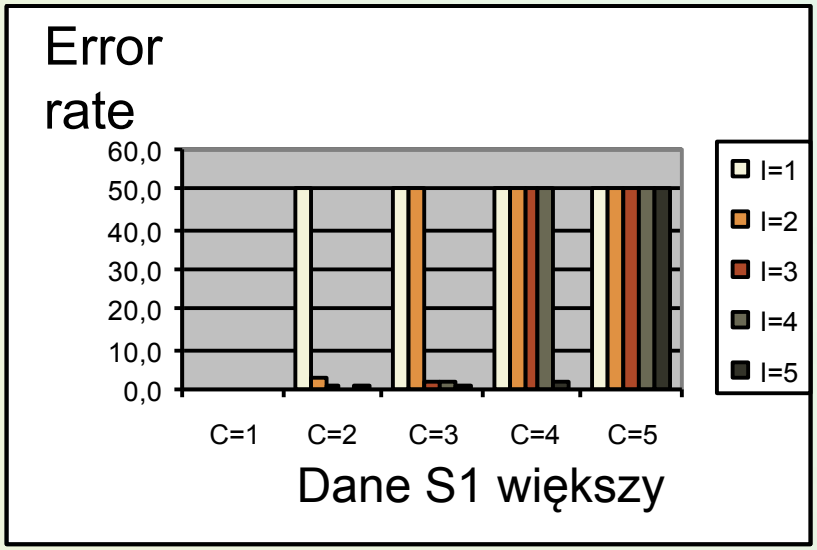
Niektóre prace eksperymentalne z dysuksją źródeł trudności, e.g:

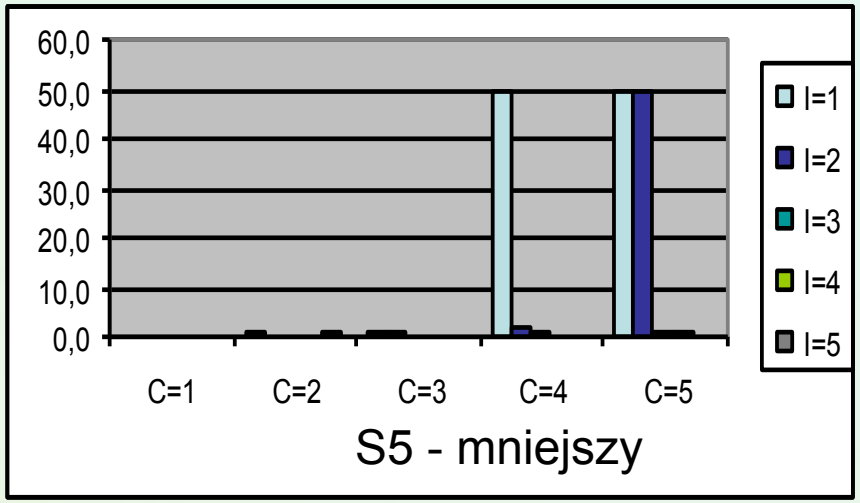
- T. Jo, N. Japkowicz. Class imbalances versus small disjuncts. SIGKDD Explorations 6:1 (2004) 40-49
- V. García, R.A. Mollineda, J.S. Sánchez. On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Anal Applic (2008) 11: 269-280
- Stefanowski J et al. Learning from imbalanced data in presence of noisy and borderline examples. RSCTC 2010.

Rezultaty eksperymentów Japkowicz i inni

Obszerne studium z 125 sztucznymi danymi, każdy zróżnicowany poprzez: the concept complexity (C), the size of the training set (S) and the degree of imbalance (I) zmieniane krokowo 1-5 (lepsze – gorsze).

- Wybrany wynik dla C4.5





Skuteczność klasyfikacji zależy od:

- the **degree** of class **imbalance**;
- the **complexity of the concept** (dekompozycja klas na subconcepts);
- the overall **size of the training set** (**rarity** połączone z powyższymi);
- the **classifier** involved (w mniejszym stopniu).

Eksperymenty Garcia et al. ze strefami brzegowymi

Overlapping

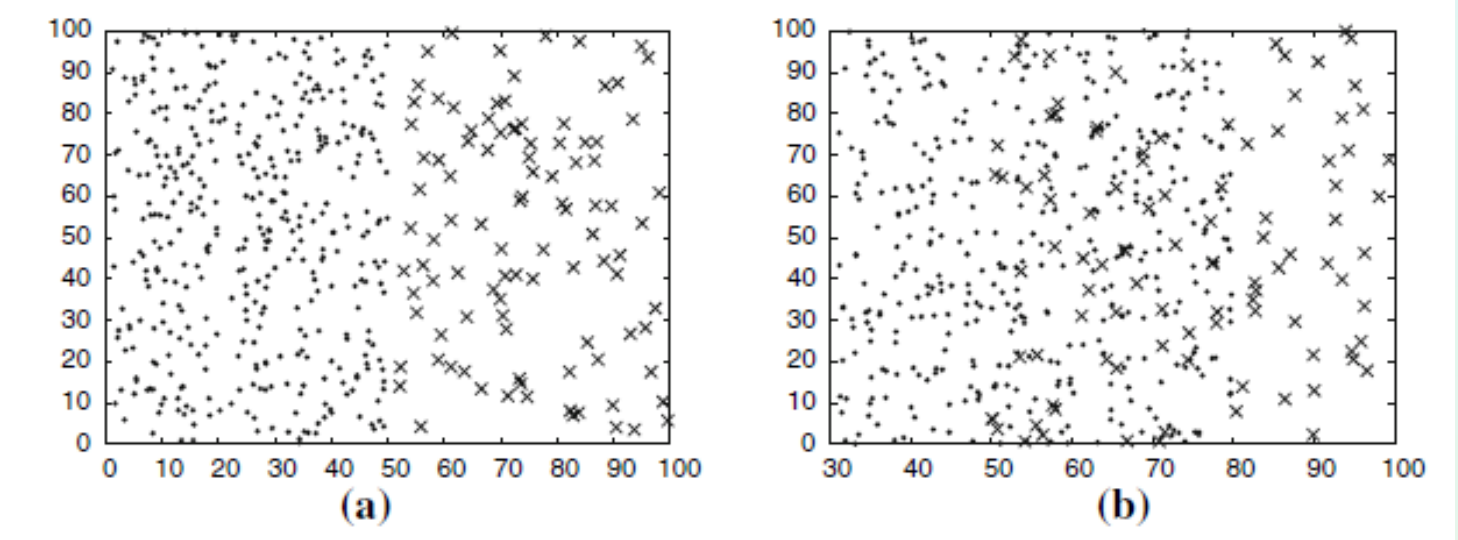


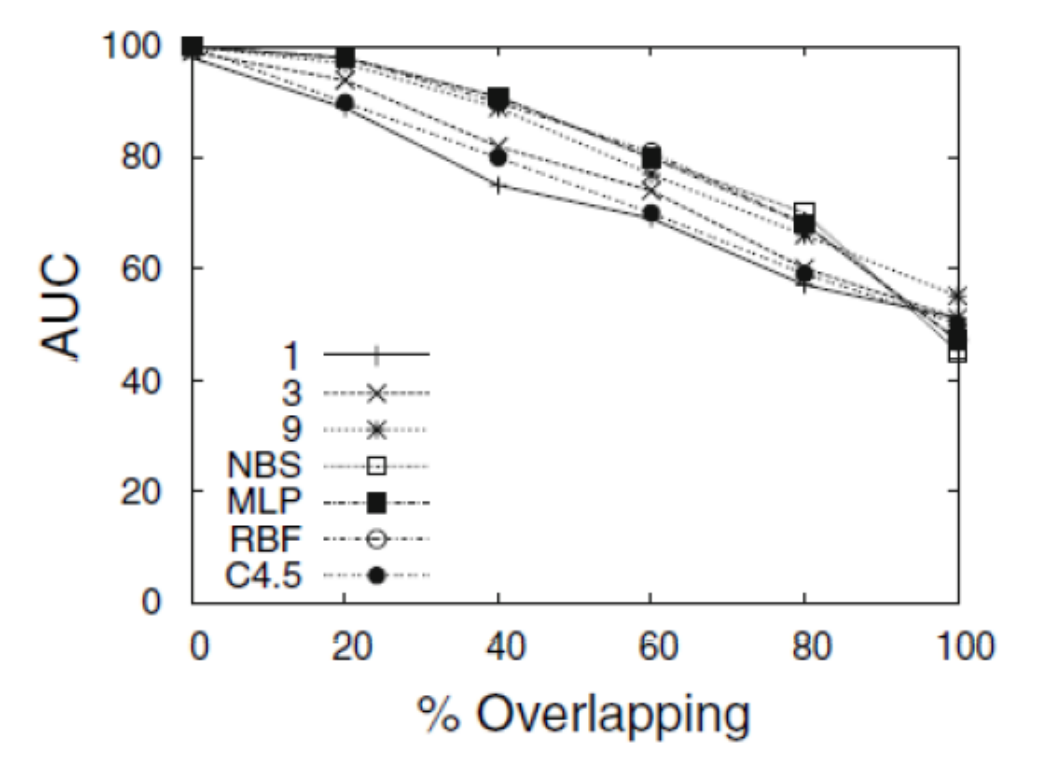
Fig. Two different levels of class overlapping: a 0% and b 60%

Experiment I: 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, [10–60] for 20%, [20–70] for 40%, [30–80] for 60%, [40–90] for 80%, and [50–100] for 100% of overlap.

The overall imbalance ratio matches the imbalance ratio corresponding to the overlap region, what could be accepted as a common case.

Eksperymenty Garcia et al. ze strefami brzegowymi

Niektóre z wyników

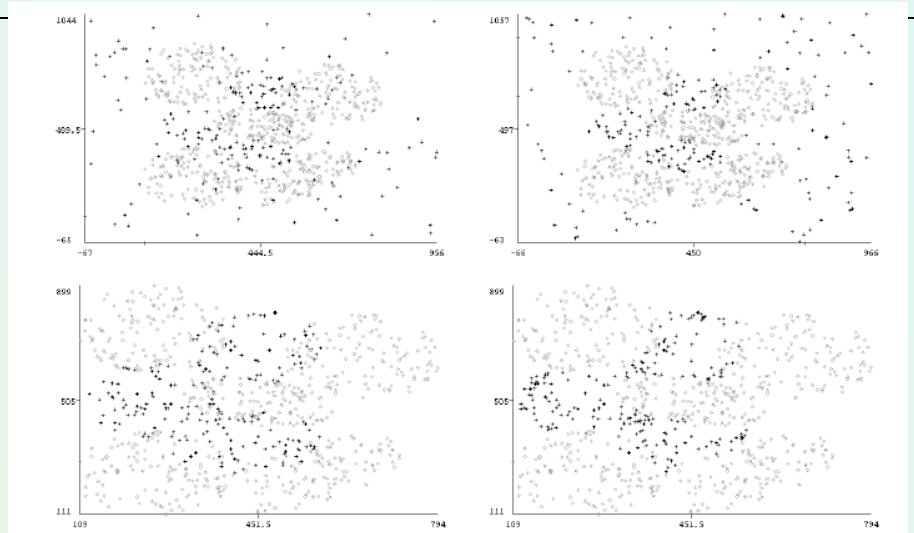


Skuteczność różnych klasyfikatorów – wzrost niejednoznaczność strefy brzegowej silniej obniża AUC niż wzrost niezrównoważenia

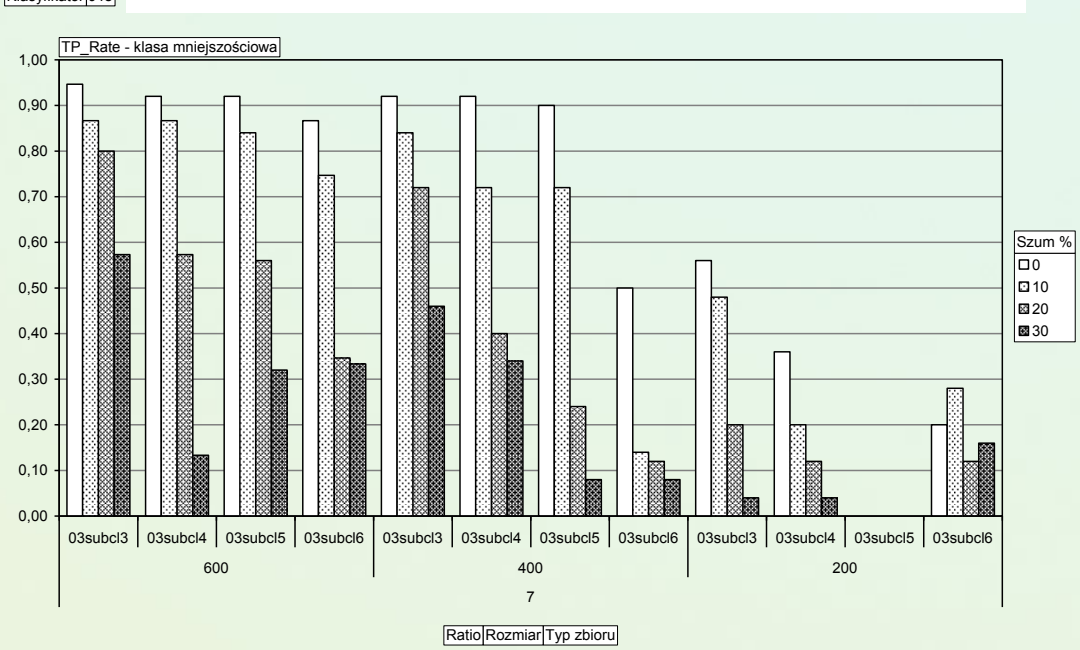
W dalszych eksperymentach zauważony wpływ lokalnej gęstości przykładów!

Dalsze eksperymenty ze sztucznymi danymi

- J. Stefanowski, 2009 (przy współpracy K.Kałużny)
- Wpływ różnego typu trudności (kształt pojęc i rzadkie przykłady, nakładające się obszary brzegowe, stopnie niezbalansowania, dekompozycja klas) na działanie klasyfikatorów C4.5, Ripper i K-NN
- Dane
  - Kolekcja kilkuset sztucznych zbiorów danych o ściśle kontrolowanej charakterystyce
- Konkluzje
  - Nieliniowe kształty klas decyzyjnych oraz ich fragmentacja były zdecydowanie poważniejszym problemem niż samo niezbalansowanie klas)
  - Najpoważniejszym źródłem trudności była zaburzona granica między klasami oraz obecności szum w klasach decyzyjnych, zwłaszcza w mniejszościowej



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



Wybrane sztuczne dane i specjalizowane metody

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**

Miara oceny: sensitivity

(rozpozn. klasy mniejsz)

Algorytm: C4.5

Skuteczność metod zależy od charakterystyki danych

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

Data set	MODLEM					C4.5				
	Base	RO	CO	NCR	SP2	Base	RO	CO	NCR	
subcl-safe	0.5800	0.5800	0.6200	0.7800	0.6400	0.3200	0.8400	0.8600	0.9800	1
subcl-B	0.8400	0.8400	0.8400	0.8600	0.8400	0.0000	0.8200	0.8400	0.3600	0
subcl-C	0.1200	0.1000	0.1600	0.2400	0.2600	0.0000	0.5400	0.0000	0.0000	0
subcl-BC	0.4800	0.4700	0.5000	0.5500	0.5500	0.0000	0.6800	0.4200	0.1800	0
clover-safe	0.3000	0.3800	0.4400	0.7000	0.6000	0.0200	0.9600	0.9200	0.0400	0
clover-B	0.8400	0.8200	0.8200	0.8400	0.8600	0.0400	0.9400	0.9200	0.0400	0
clover-C	0.1400	0.0800	0.1400	0.2400	0.3600	0.0000	0.3000	0.0200	0.0000	0
clover-BC	0.4900	0.4500	0.4800	0.5400	0.6100	0.0200	0.6200	0.4700	0.0200	0
paw-safe	0.8400	0.9200	0.8400	0.8400	0.8000	0.4200	0.9000	0.9600	0.7400	1
paw-B	0.8800	0.8800	0.8600	0.8800	0.9000	0.1400	0.9000	0.9000	0.4000	0
paw-C	0.1600	0.1400	0.1200	0.2600	0.1600	0.0400	0.2000	0.0000	0.0000	0
paw-BC	0.5200	0.5100	0.4900	0.5700	0.5300	0.0900	0.5500	0.4500	0.2000	0

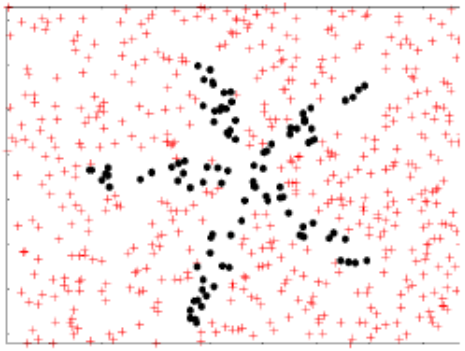


Fig. 1. Clover data set

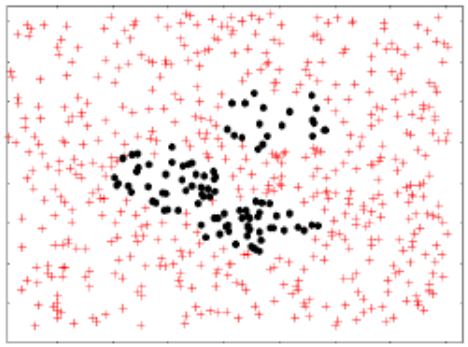
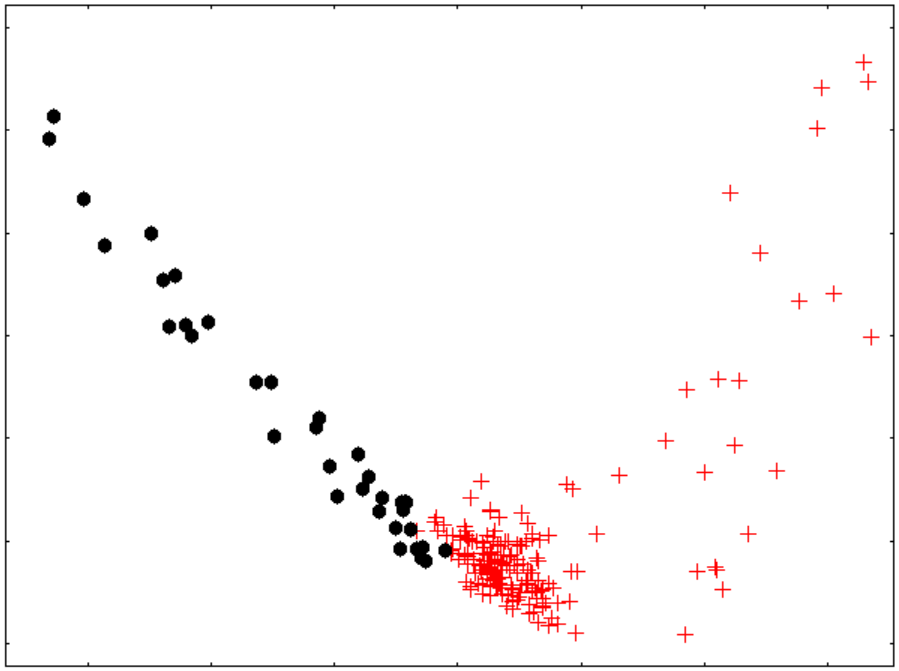


Fig. 2. Paw data set

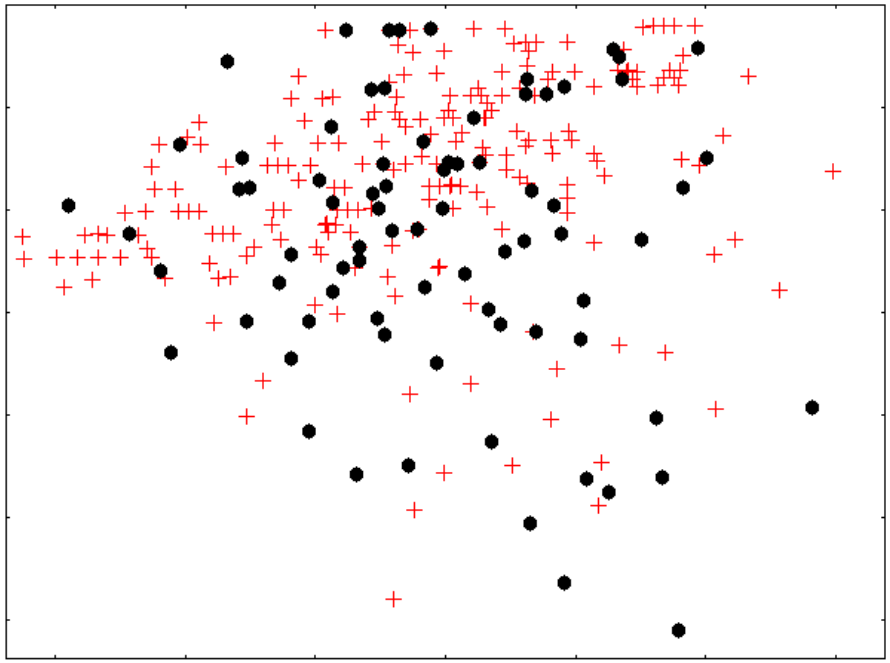
K. Napierała, J. Stefanowski, Sz. Wilk: *Learning from imbalanced data in presence of noisy and borderline examples*. RSCTC 2010, LNAI Springer.

Co z rzeczywistymi danymi?

Wizualizacja 2 pierwszych składowych w metodzie MDS (PCA)



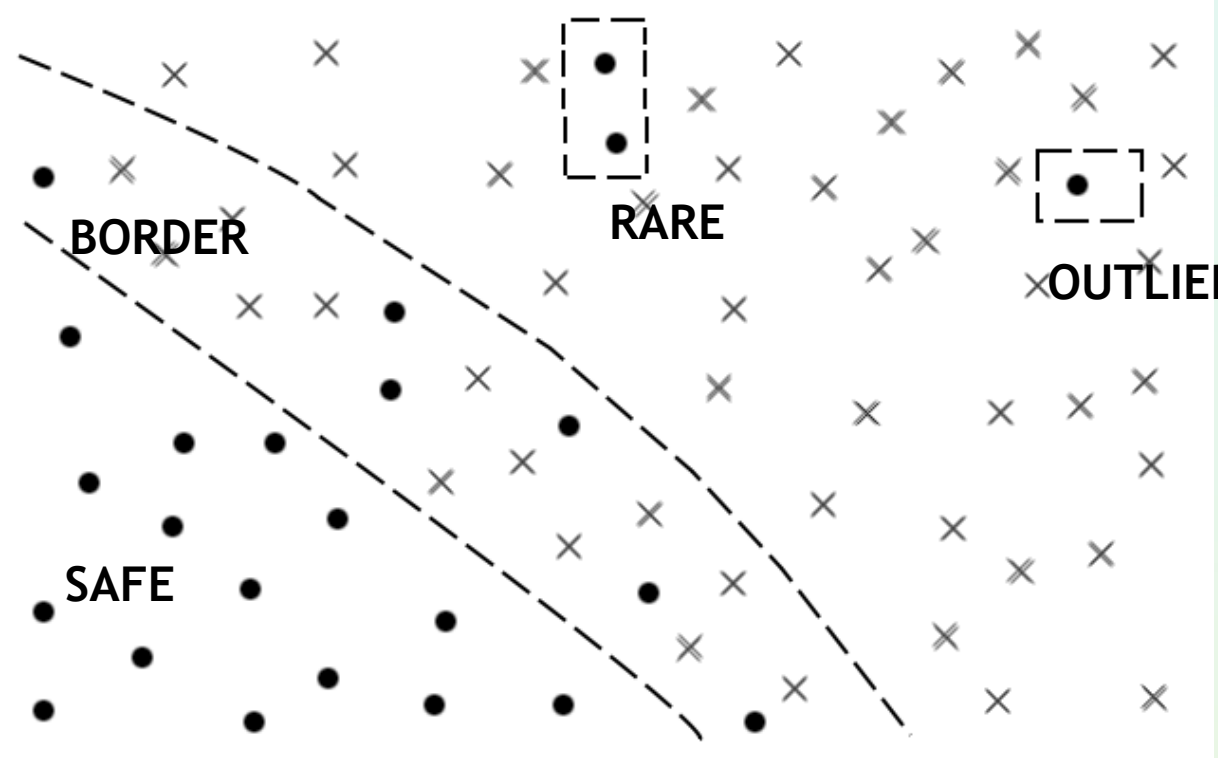
Thyroid
215 ob./ 35 mniejsz.
Prosty dla klasyfikacji



Haberman
306 ob. / 81 mniejsz.
trudny

Data Difficulty Factors → Różna lokalna charakterystyka rozkładu (typów) przykładów

Rozróżniamy 4-y typy przykładów:

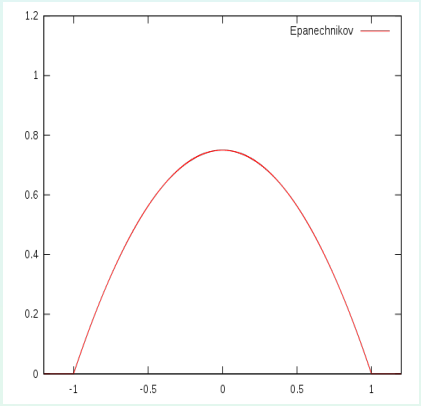


Więcej → K.Napierała, J. Stefanowski: The influence of minority class distribution on learning from imbalance data. HAIS, 2012.

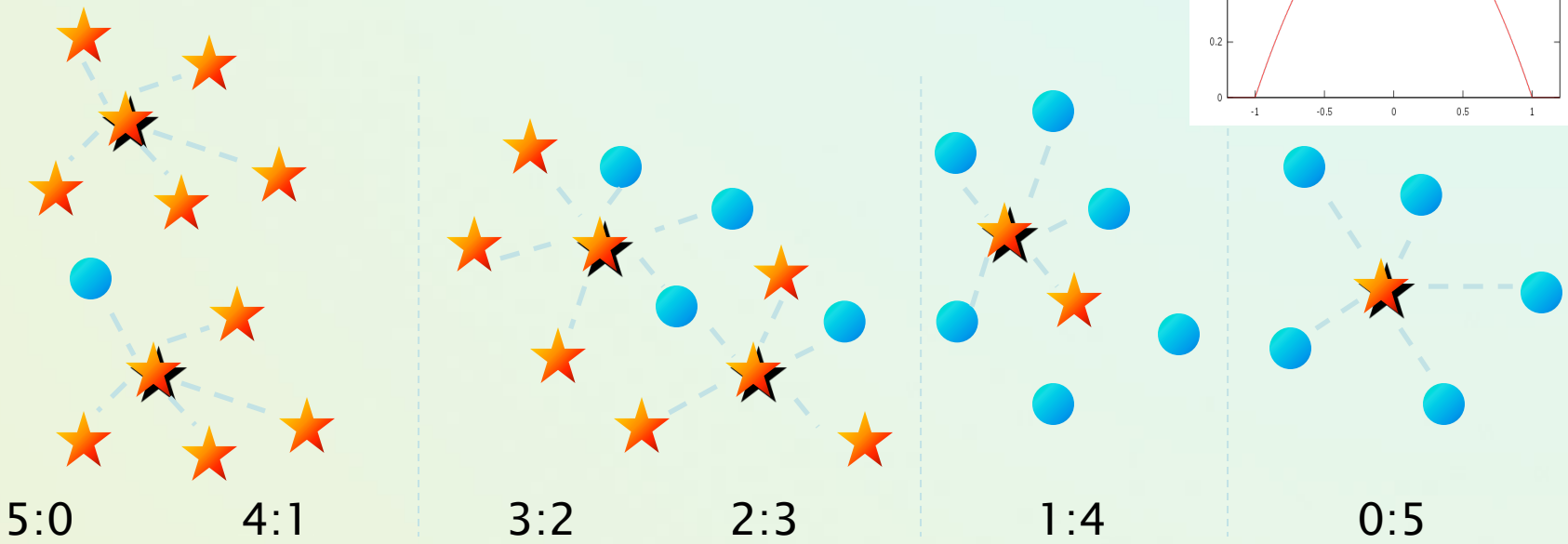
Identification of examples – local approach

Analizuj rozkład etykiet wśród najbliższych sąsiadów x

- K-NN (k=5, 7, ...) - HVDM distance
- Kernel functions (paramert Bandwidth)



Określ typ przykładu x

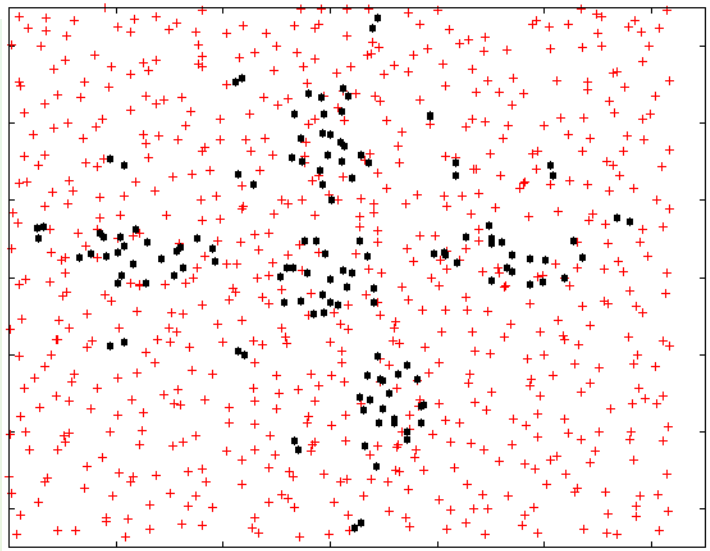


Details → K.Napierała, J. Stefanowski: The influence of minority class distribution on learning from imbalance data. HAIS, 2012.

Napierala, Stefanowski., Types of minority class examples and their influence on learning classifiers. JIIS (2016)

Sprawdzenie na sztucznych rozkładach danych

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

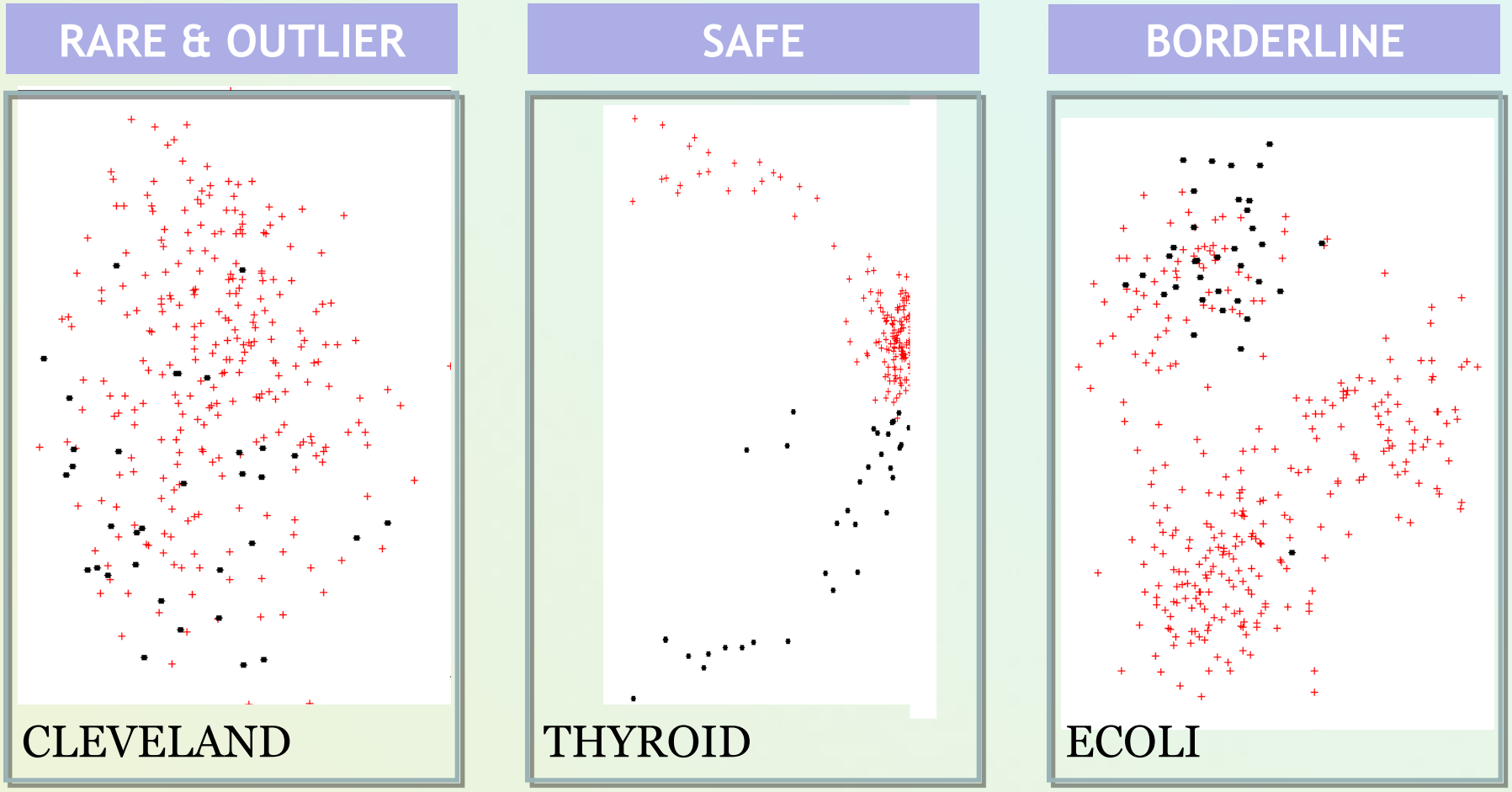


Labeling Minority Examples → UCI Data sets

Dataset	S [%]	B [%]	R [%]	O [%]
abdominal-pain	59.90	22.28	8.90	7.92
acl	67.50	30.00	0.00	2.50
new-thyroid	68.57	31.43	0.00	0.00
vehicle	74.37	24.62	0.00	1.01
car	47.83	39.13	8.70	4.35
scrotal-pain	38.98	45.76	10.17	5.08
ionosphere	44.44	30.95	11.90	12.70
credit-g	9.33	63.67	10.33	16.67
ecoli	28.57	54.29	2.86	14.29
hepatitis	15.63	62.50	6.25	15.63
haberman	4.94	61.73	18.52	14.81
breast-cancer	24.71	25.88	32.94	16.47
cmc	17.72	44.44	18.32	19.52
cleveland	0.00	31.43	17.14	51.43
glass	0.00	35.29	35.29	29.41
hsv	0.00	0.00	28.57	71.43
abalone	8.36	20.60	20.60	50.45
postoperative	0.00	41.67	29.17	29.17
solar-flare	0.00	48.84	11.63	39.53
transfusion	18.54	47.19	11.24	23.03
yeast	5.88	47.06	7.84	39.22

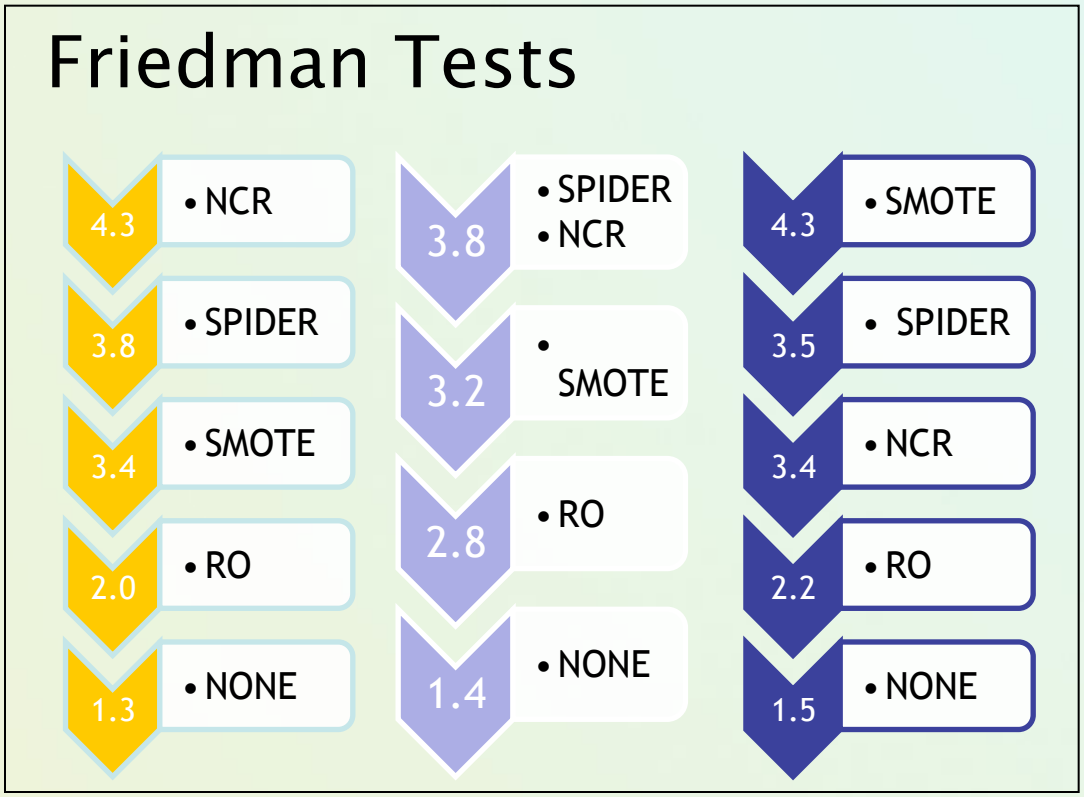
- Very unsafe distribution of the minority examples
  - cleveland → 51% outliers, no safe ones
  - solar flare, balance scale
- Majority class → quite safe
  - yeast → 98,5% S
  - ecoli → 91,7% S
- Experiences with k (7,9,..) or kernels → similar categorizations of data
- Unsafe data → deteriorate classifier performance and difficult for improvement

Dominujacy typ przykładu mniejszościowego



Napierala, Stefanowski, Types of minority class examples and their influence on learning classifiers. JIIS (2015)

Zróżnicowane działanie metod w zależności od kategorii przykładów mniejszościowych w danych



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

B

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

R

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

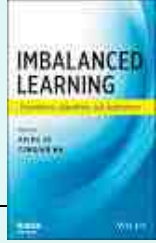
O

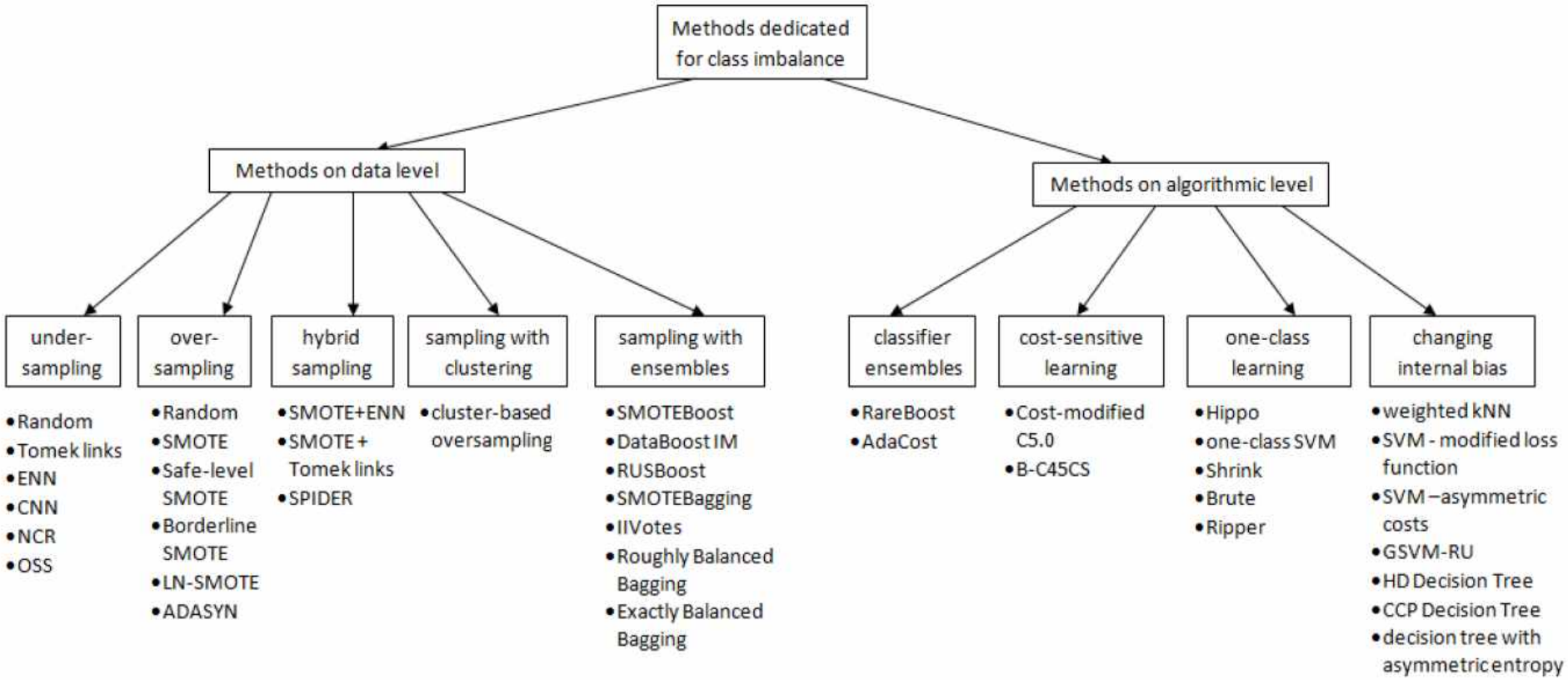
Resultaty sensitivity PART rules

1NN, J48: podobne rankingi

RBF: RO wyżej w rankingu

Quick view at methods for class imbalance





and even more, …

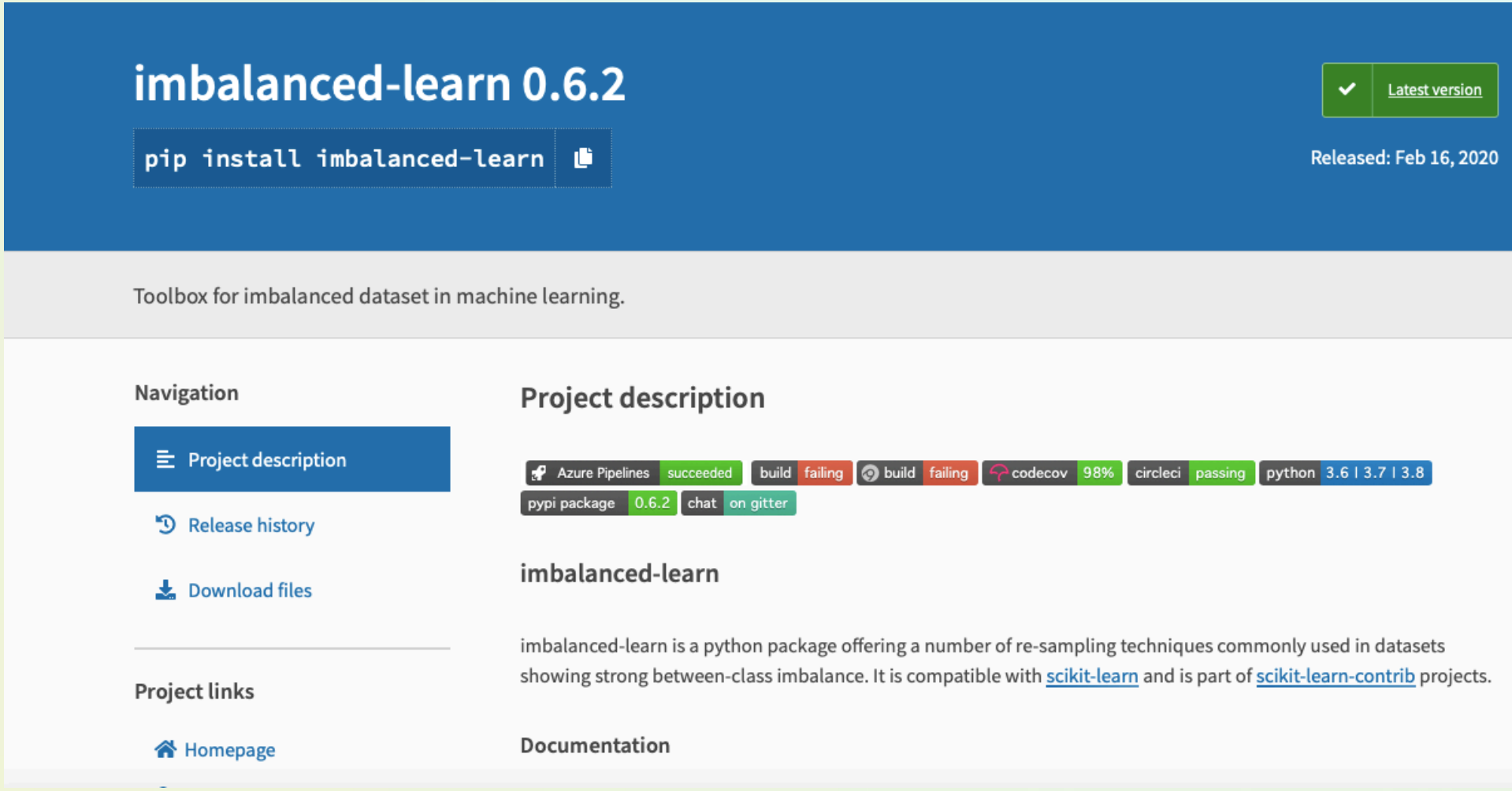
Review →

He H., Yungian, Ma (eds): Imbalanced Learning. Foundations, Algorithms and Applications. IEEE - Wiley, 2013
A.Fernandez et al.: Learning from imbalanced data sets. Springer 2018.

Python – co robić

Imbalanced-learn Toolbox (Lemaitre et al 2017)

Under- (11), over-sampling (7), some ensembles (4)



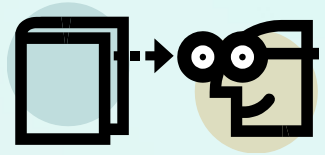
WEKA i inne

Podstawowa WEKA = resampling (SMOTE oraz random), cost-sensitive classifiers, MetaCost

KEEL – więcej algorytmów (45) over i under sampling (20) oraz rozbudowane zespoły klasyfikatorów (21)

R package ‘imbalance’ oraz IRIC: An R library for binary imbalanced classification: 29 metod re-sampling, 4 zespoły klasyfikatorów, 1 cost sensitive

Literaturowa kategoryzacja metod

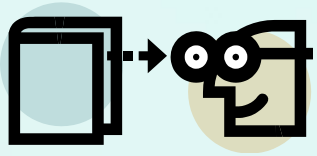


- Dwa podstawowe kierunki działanie
  - Modyfikacje danych (preprocessing)
  - Modyfikacje algorytmów
- Najbardziej popularne grupy metod
  - Re-sampling or re-weighting,
  - Zmiany w strategiach uczenia się, użycie nowych miar oceny (np. AUC)
  - Nowe strategie eksploatacji klasyfikatora (classification strategies)
  - Ensemble approaches (najczęściej adaptacyjne klasyfikatory złożone typu bagging)
  - Specjalizowane systemy hybrydowe
  - One-class-learning
  - Transformacje do zadania „cost-sensitive learning”
  - …

Inne podejścia do modyfikacji algorytmów uczących

- Zmiany w indukcji drzew decyzyjnych
  - Weiss, G.M. Provost, F. (2003) "Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction" JAIR.
- Modyfikacje w klasyfikatorach bayesowskich
  - Jason Rennie: Tackling the Poor Assumptions of Naive Bayes Text Classifiers ICML 2003.
- Wykorzystanie „cost-learning” w algorytmach uczących
  - Domingos 1999; Elkan, 2001; Ting 2002; Zadrozny et al. 2003; Zhou and Liu, 2006
- Modyfikacje zadania w SVM
  - K.Morik et al., 1999.; Amari andWu (1999)
  - Wu and Chang (2003),
  - B.Wang, N.Japkowicz: Boosting Support Vector Machines for Imbalanced Data Sets, KAIS, 2009.

Metody modyfikujące zbiór uczący



Zmiana rozkładu przykładów w klasach przed indukcją klasyfikatora (ang. pre-processing):

- Proste techniki losowe
  - „Over-sampling” – klasa mniejszościowe
  - „Under-sampling”- klasa mniejszościowa
- Specjalizowane nadlosowanie
  - Cluster-oversampling (Japkowicz)
- Ukierunkowane transformacje
  - Klasa większościowe
    - One-side-sampling (Kubat, Matwin) z Tomek Links
    - Laurikkala’s edited nearest neighbor rule
  - Klasa mniejszościowe
    - SMOTE → Chawla et al.
    - Borderline SMOTE, Safe Level, Surrounding SMOTE, …
  - Podejścia łączone (hybrydowe)
    - SPIDER
    - SMOTE i undersampling
  - Powiązanie z budową klasyfikatorów złożonych

Resampling – modyfikacja zbioru uczącego przed budową klasyfikatora

„Resampling” → pre-processing; celowa zmiana rozkładu przykładów; „balansowanie” liczności klas po to aby w kolejnej fazie móc lepiej nauczyć klasyfikator

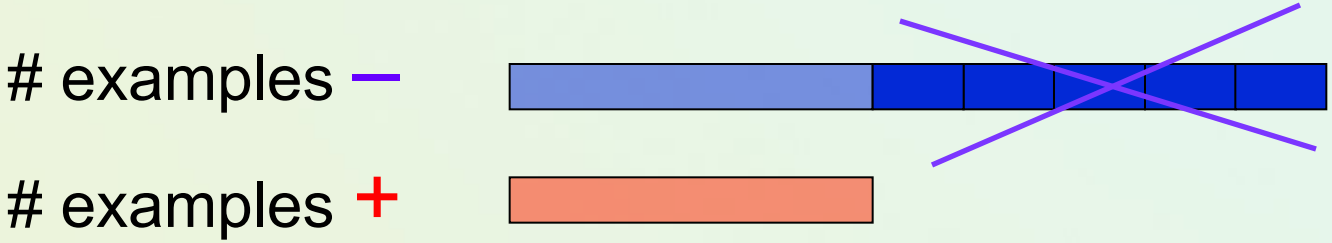
Raczej heurystyka ukierunkowana na uzyskanie lepszych rozkładów klas niż uzasadnione teoretycznie podejście [F.Herrera 2010].

Brak teoretycznej gwarancji znalezienia optymalnej postaci rozkładu!

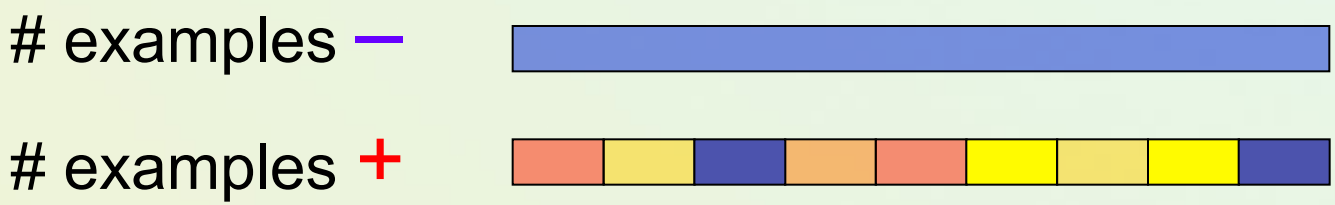
Losowe nadlosowanie lub usuwanie przykładów

Undersampling vs oversampling

under-sampling

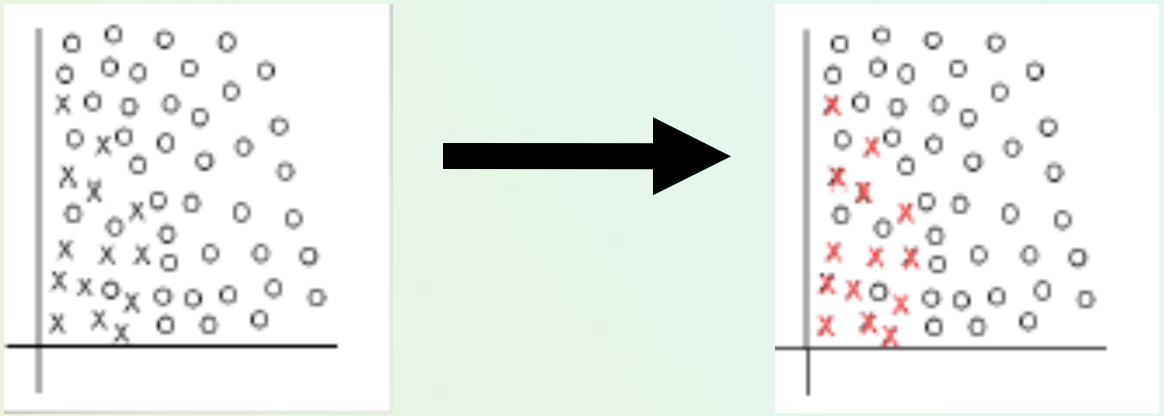


over-sampling

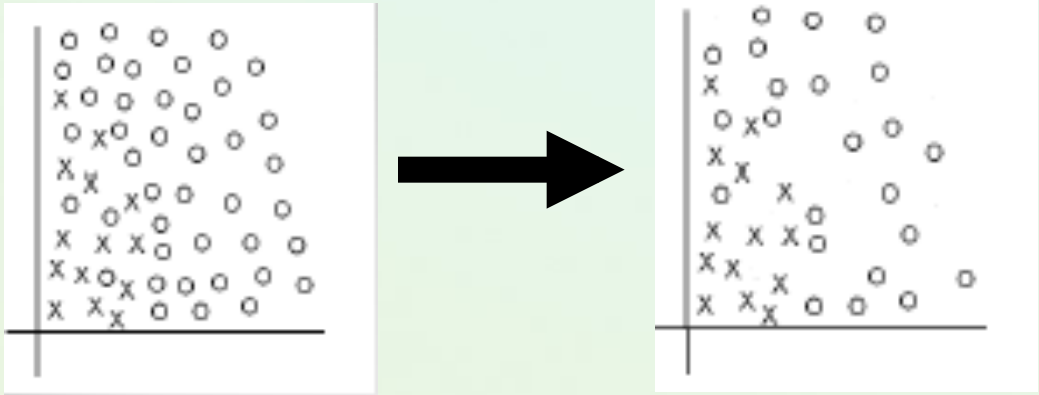


Losowe równoważenie klas - dyskusja

Random oversampling → Kopiowanie przykładów mniejszościowych



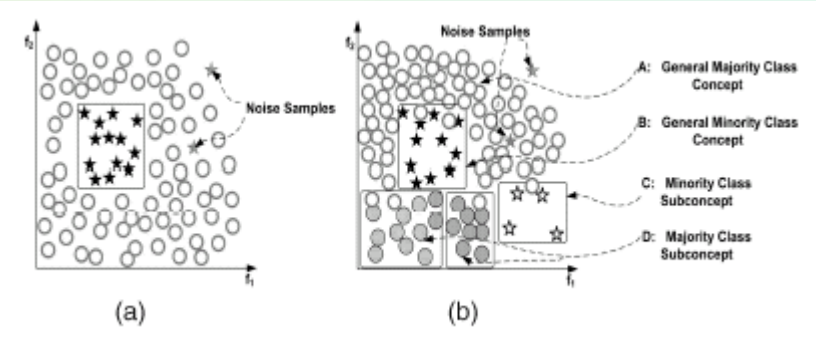
- Czy proporcja 1:1 jest optymalna? Nie zawsze.
- Ryzyko przeuczenia
- Random undersampling:
- Usuwanie przykładów większościowych



- Utrata informacji o ważnych przykładach w rozkładzie klasy

Cluster oversampling – specyficzne losowanie w przypadku dekompozycji klas i obecności „small disjuncts”

Dekompozycja klas → within and between -class imbalance



Jak zidentyfikować „small disjuncts” – rzadkie podobszary w klasach?

Użyj algorytmu analizy skupień wewnątrz każdej klasy i odpowiednio nadlosowuj!

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

Ilustracja działania Cluster based Oversampling

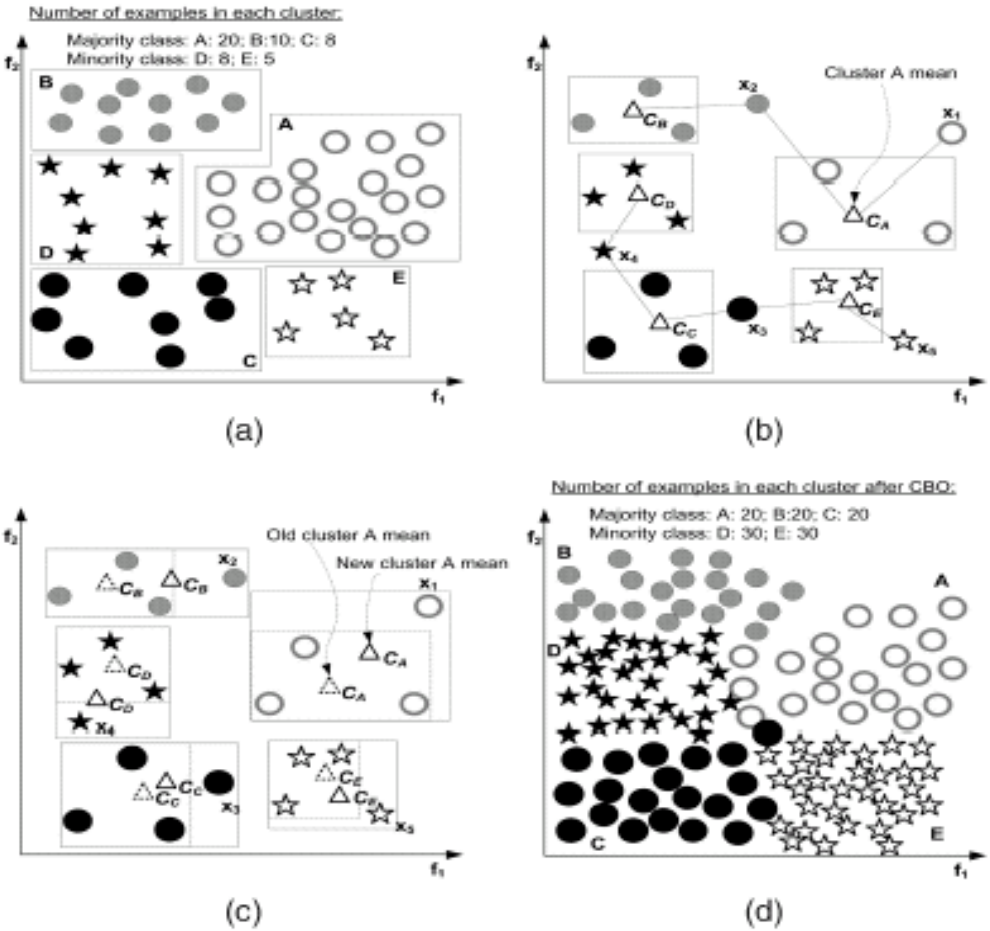


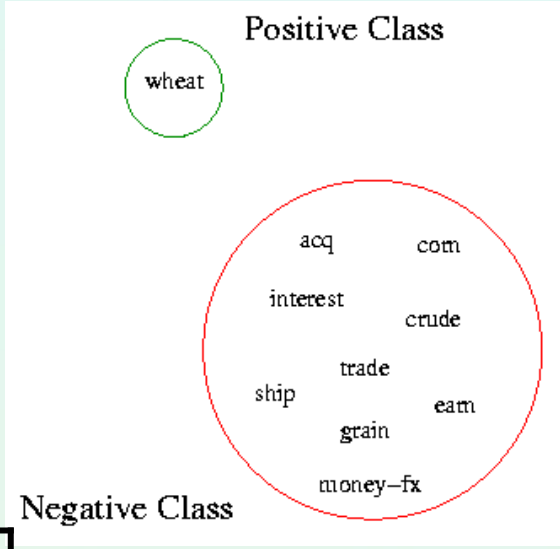
Fig. 6. (a) Original data set distribution. (b) Distance vectors of examples and cluster means. (c) Newly defined cluster means and cluster borders. (d) The data set after cluster-based oversampling method.

Jak odkryć pod-skupiska? - Wykorzystaj k-means lub algorytm gęstościowy. Ale jak je sparametryzować?

Studium Przypadku - Within-class vs Between- class Imbalances: Text Classification

Prace: Japkowicz J., Nickerson A., Milios. E

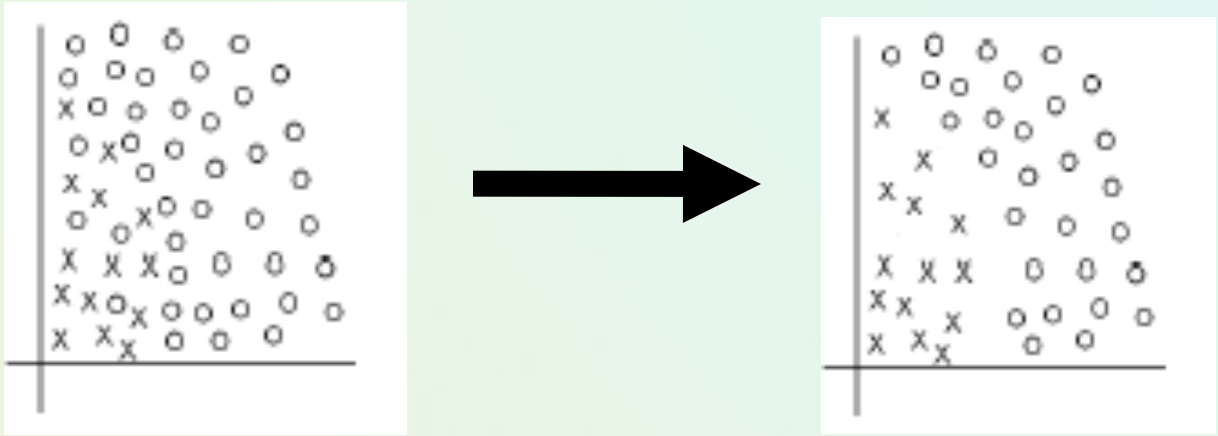
- Reuters-21578 Dataset
- Classifying a document according to its topic
- Positive class is a particular topic
- Negative class is every other topic



Method	Precision	Recall	F-Measure
Imbalanced	**0.617**	**0.394**	**0.455**
Random Oversampling	**0.580**	**0.545**	**0.560**
Guided Oversampling I (# Clusters Unknown)	**0.650**	**0.510**	**0.544**
Guided Oversampling II (Using Known Clusters)	**0.601**	**0.751**	**0.665**

Ukierunkowane modyfikacje danych

Focused resampling (Informed approaches): przetwarzaj tylko trudne obszary



- Czyszczenie borderline, redundant examples: Tomek links i one-side sampling
- Czyszczenie szumu i borderline: NCR
- Metoda SPIDER (J.Stefanowski, Sz.Wilk)
- SMOTE i jej rozszerzenia

- Czy są to typowe tricki „losowania” oraz edytowanie danych (np. rozszerzania k-NN)?

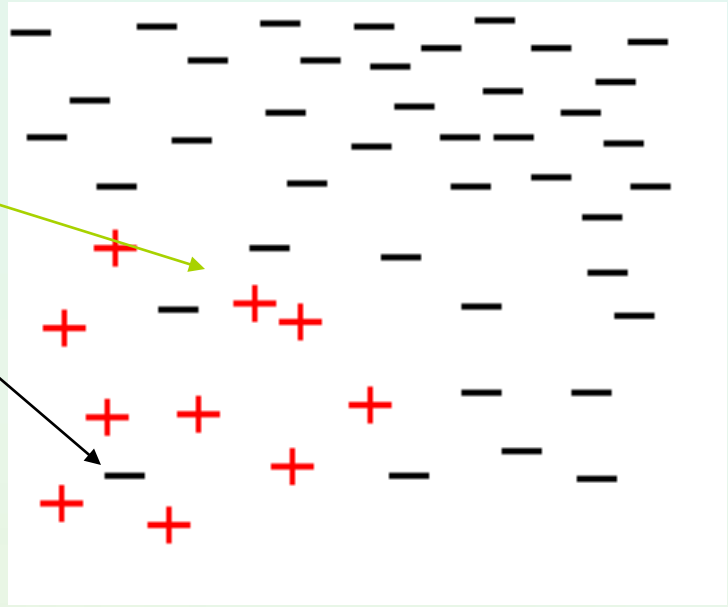
Powróćmy do charakterystyki przykładów

Typy przykładów → techniki „resampling” powinny skupić swoje działanie na niektórych z nich

Różne typy przykłady z klasy (większościowej)

- Noise examples
- Borderline examples

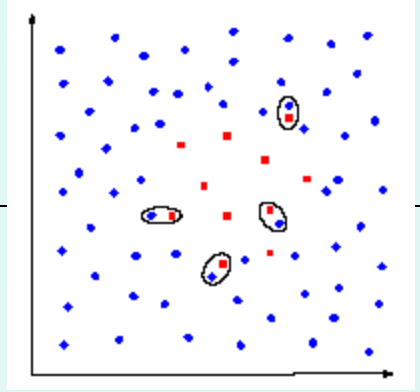
  Borderline examples are unsafe since a small amount of noise can make them fall on the wrong side of the decision border.
- Redundant examples



- Safe examples

Modyfikacje undersampling: Znajdź i usuń 2 lub 3 pierwsze typy przykładów

Under-sampling z wykorzystaniem Tomek links

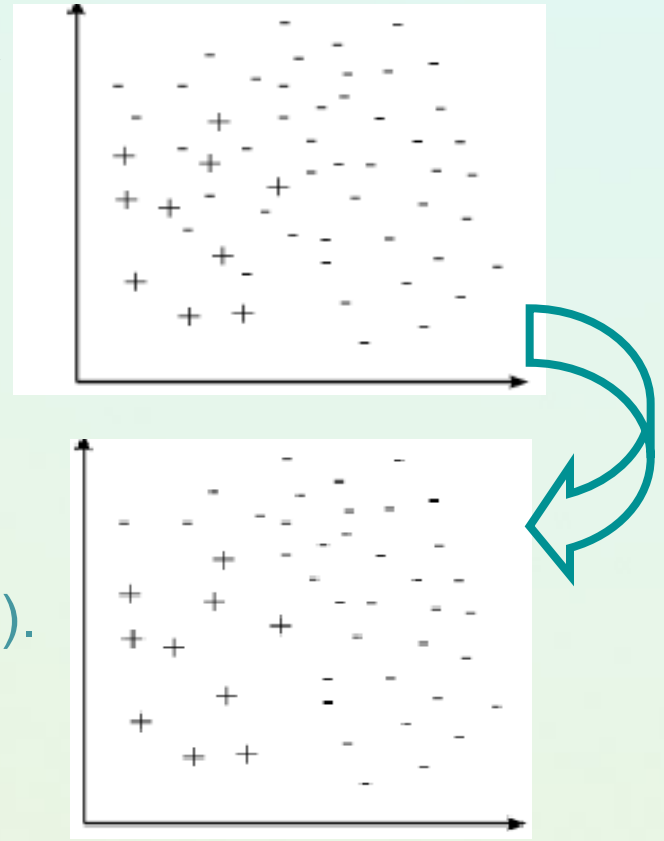


Przykład Tomek Links

- Usuwaj przykłady graniczne i szum z klasy większościowej
- „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).

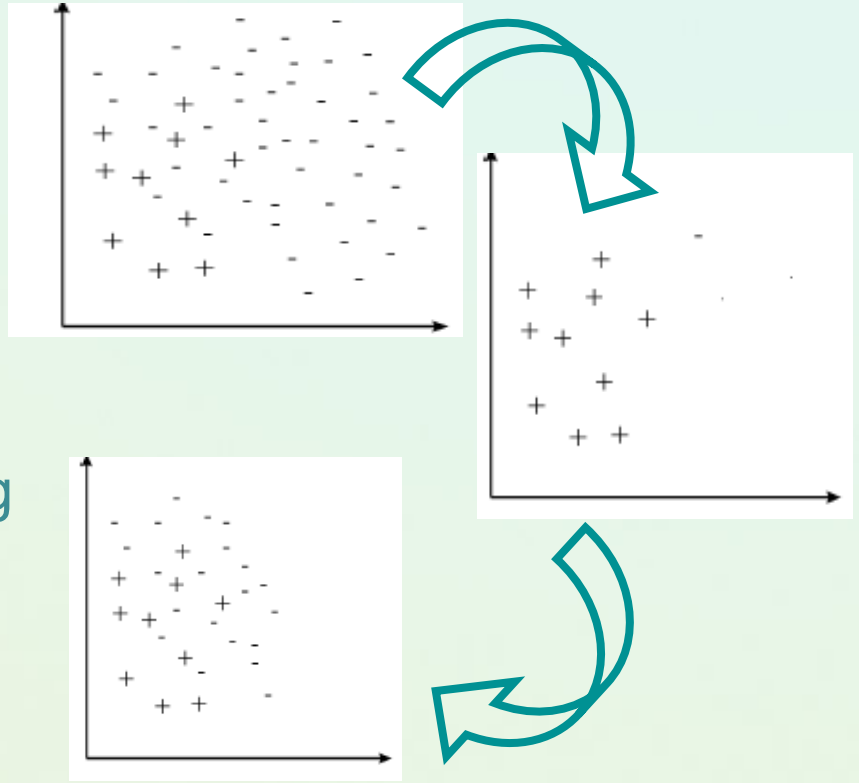


Under-sampling z wykorzystaniem zasady CNN

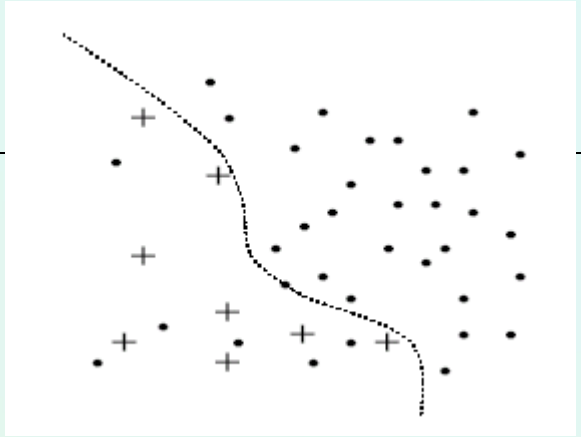
CNN – Condensed Nearest Neighbours specjalizowana odmiana edytowanego K-NN

Jedna z najstarszych metod redukcji przykładów (powstało wiele modyfikacji).
Idea znaleźć taki podzbiór *E'* zbioru *E*, który gwarantowałby poprawną klasyfikację wszystkich przykładów ze zbioru *E* za pomocą algorytmu 1NN.
Duda, Hart 1968.

- Usuwanie trudnych przykładów
- Schemat algorytmu:
  - 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'



Informed Under-sampling



Najbardziej znany One –side-sampling

Kubat, Matwin 1997

- One-sided selection
  - – Tomek links + CNN
  - – może usuwać zbyt dużo przykładów
- CNN + Tomek links
  - – wprowadził F. Herrera
  - – Poszukiwanie Tomek links duże koszty obliczeniowe → lepiej wykonuje na zredukowanym zbiorze

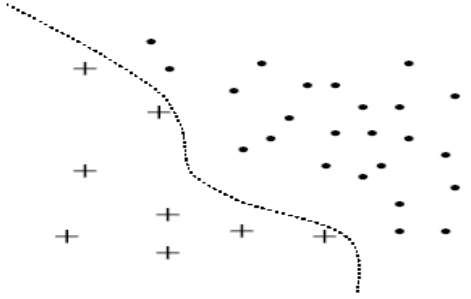


Figure 3: The training set without the borderline and noisy negative examples.

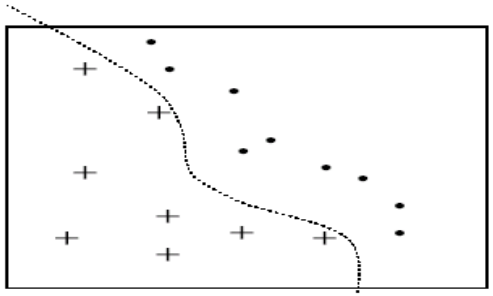


Figure 4: The training set after the removal of redundant negative examples.

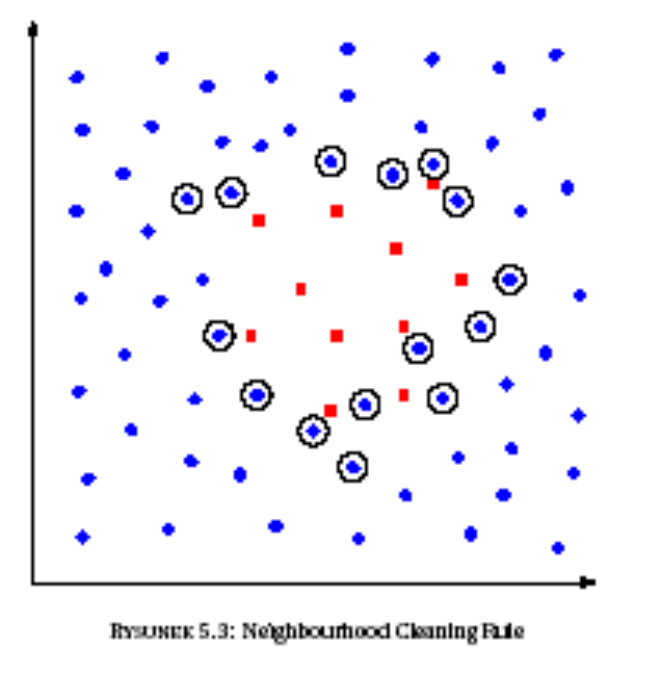
Nearest Cleaning Rule

- **NCL** Nearest Cleaning Rule - Jorma Laurikkala 2001,

  Inne od OSS, bardziej „czyści” obszary brzegowe klas niż redukuje przykłady

  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 three nearest neighbors

Ilustracja – które przykłady większościowe usuwamy



SMOTE - Synthetic Minority Oversampling Technique

- Wprowadzona przez Chawla, Hall, Kegelmeyer 2002
- Dla każdego przykładu *p* z klasy mniejszościowej
  - Znajdź jego k-najbliższych sąsiadów (UWAGA wyłącznie z klasy mniejszościowej!)
  - Losowo wybierz j z powyższych sąsiadów
  - Losowo stwórz sztuczny przykład wzdłuż lini łączącej p z wybranym losowo jego sąsiadem

  (**parametr j** - the amount of oversampling desired)
- Porównując z simple random oversampling – SMOTE rozszerza regiony klasy mniejszościowej starając się robić je mniej specyficzne, „paying attention to minority class samples without causing overfitting”.
- SMOTE – uznawana za bardzo skuteczną zwłaszcza w połączeniu z odpowiednim undersampling (wyniki Chawla, 2003).

Detale generacji sztucznego przykładu w SMOTE

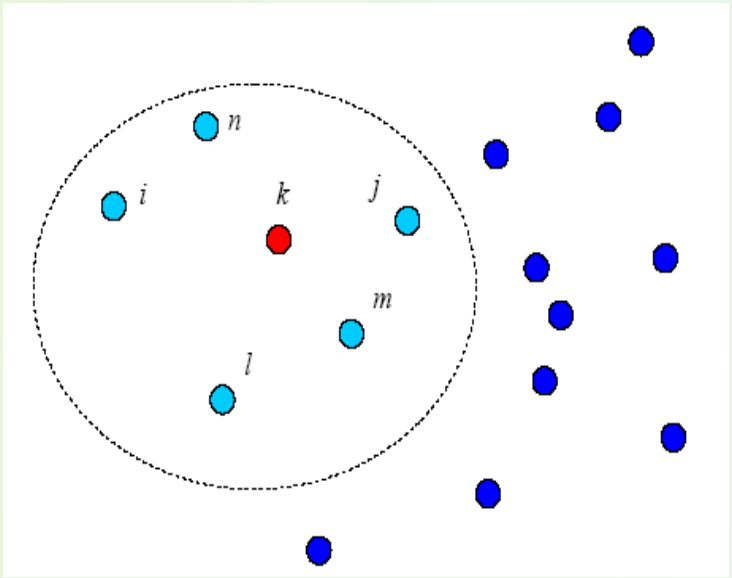
- Sąsiedzi przykładu *p* identyfikowani poprzez K-NN z odległością euklidesową lub HVDM
- Ogólny schemat dla atr. liczbowych
  - Take the difference between the feature vector (sample) under consideration and its nearest neighbor.
  - Multiply this difference by a random number between 0 and 1
  - Add it to the feature vector under consideration.
- Rozszerzenie dla atrybutów nominalnych
  - wartość najczęściej występująca w zbiorze przykładów składającym się z *p* oraz jego *k* najbliższych sąsiadów z tej samej klasy.

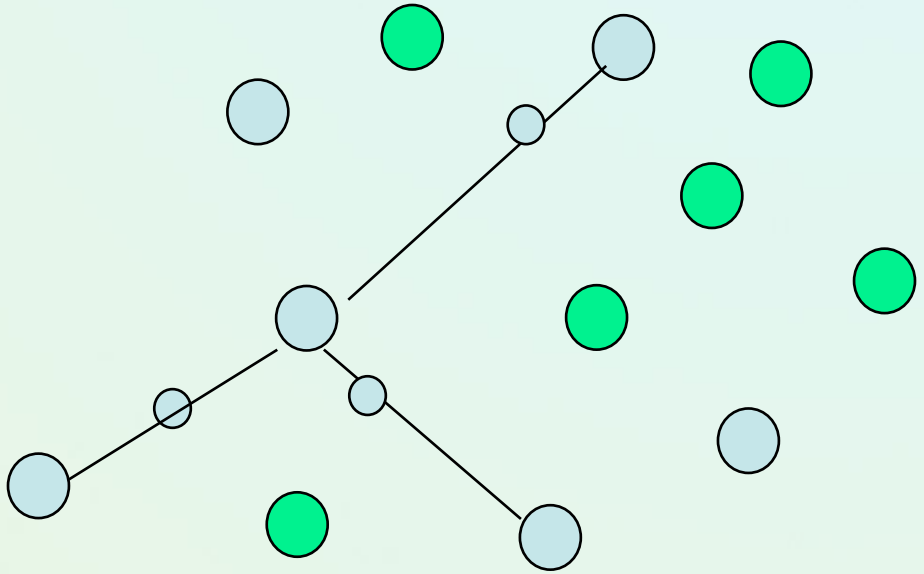
Consider a sample (6,4) and let (4,3) be its nearest neighbor.
(6,4) is the sample for which k-nearest neighbors are being identified
(4,3) is one of its k-nearest neighbors.
Let:
f1_1 = 6 f2_1 = 4 f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3 f2_2 - f1_2 = -1
The new samples will be generated as
(f1',f2') = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a random number between 0 and 1.

* N. Chawla, K. Bowyer, L. Hall, P. Kegelmeyer, SMOTE: Synthetic Minority Over-Sampling Technique, JAIR, vol. 16, 321-357, 2002.

Oversampling klasy mniejszościowej w SMOTE

SMOTE – analiza WYŁĄCZNIE klasy mniejszościowej!





: Przykład kl. mniejszościowej

: syntetyczny przykład

: Przykład kl. większościowej

Dobre rozkłady klas

SMOTE – może wstawić sztuczne przykłady w regionach klasy większościowej / wprowadza zakłócenia, szum

SMOTE – przykład oceny AUC

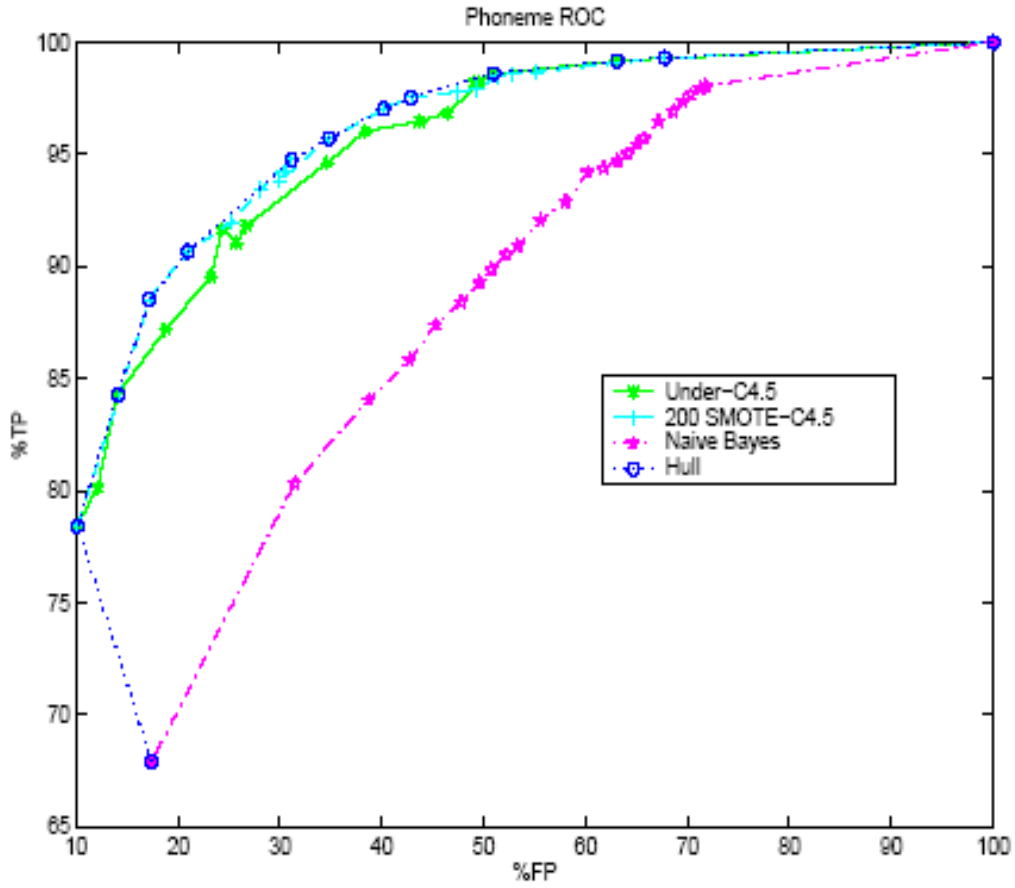


Figure 7: Phoneme. Comparison of SMOTE-C4.5, Under-C4.5, and Naive Bayes. SMOTE-C4.5 dominates over Naive Bayes and Under-C4.5 in the ROC space. SMOTE-C4.5 classifiers are potentially optimal classifiers.

SMOTE zbiorcza ocena

k = 5 sąsiadów, różny stopień nadlosowania (np. 100% to dwukrotne zwiększenie liczności klasy mniejszościowej)

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.

SMOTE

: Comparison of % Minority correct for replicated over-sampling and SMOTE fo
the Mammography dataset

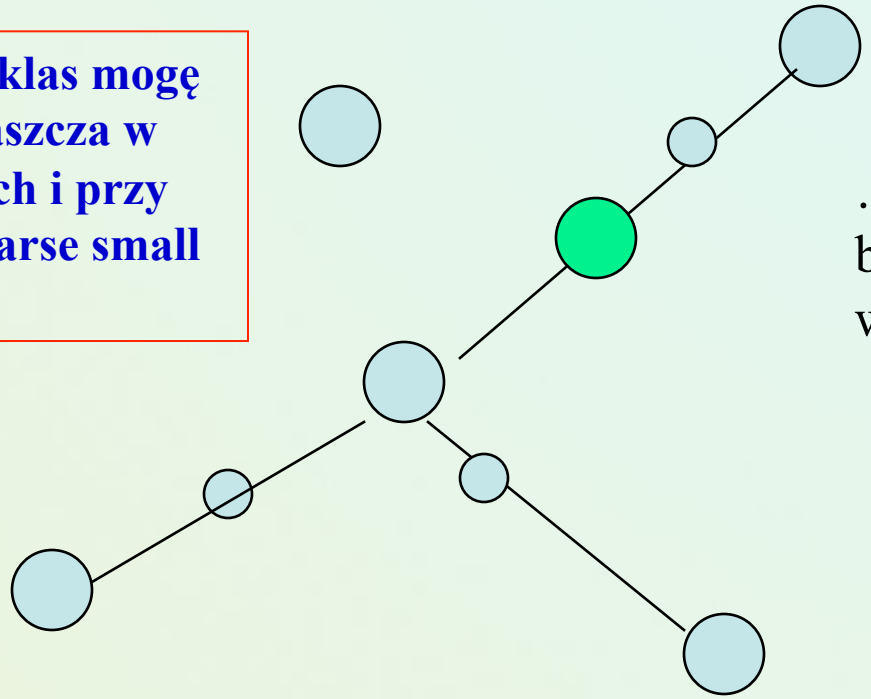
SMOTE – uwagi krytyczne

- Overgeneralization
  - SMOTE jest „naturalnie” niebezpieczna, gdyż ślepo uogólnia mniejszościowe przykłady bez rozważania rozkładów klasy większościowej
  - Szczególnie problematyczne dla mocno rozproszonych klas z tzw. small disjuncts → zwiększa szanse na nakładanie się rozkładów klas
- Lack of Flexibility
  - Liczba przykładów do nadlosowania klasy mniejszościowej musi być znana przed uruchomieniem procedury.
  - Właściwe dostrojenie parametrów silnie zależne od zadania

Oversampling klasy mniejszościowej w SMOTE

Oversampling – nie rozważa rozkładów klasy większościowej

Pamiętaj, że rozkłady klas mogę się „przenikać” zwłaszcza w obszarach brzegowych i przy dekompozycji klas (sparse small disjuncts)



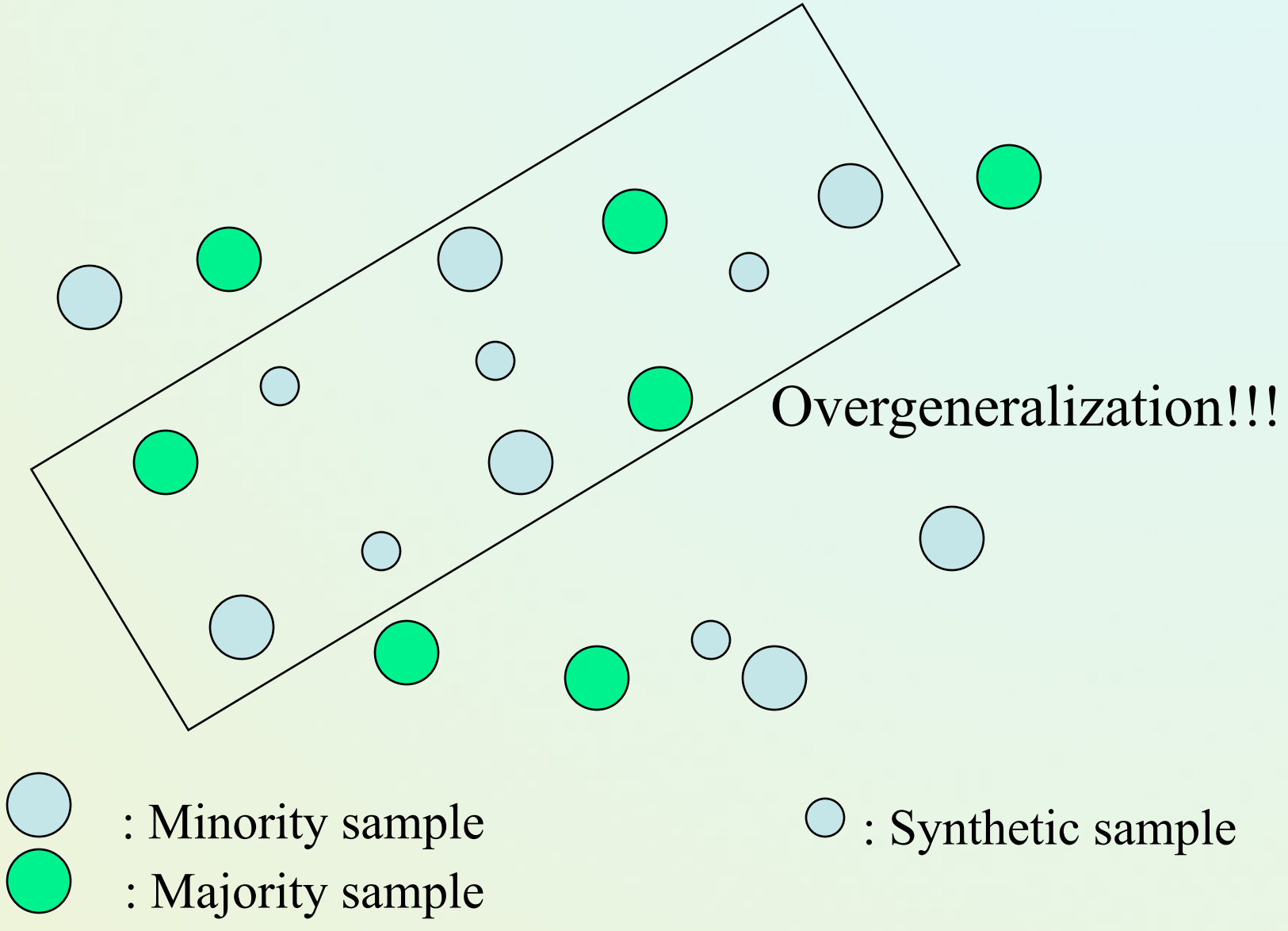
… Co się stanie gdy blisko jest przykład większościowy?

: Minority sample

: Synthetic sample

: Majority sample

Ograniczenia nadlosowania w SMOTE



SMOTE hybridization: SMOTE + Tome links (albo ENN)

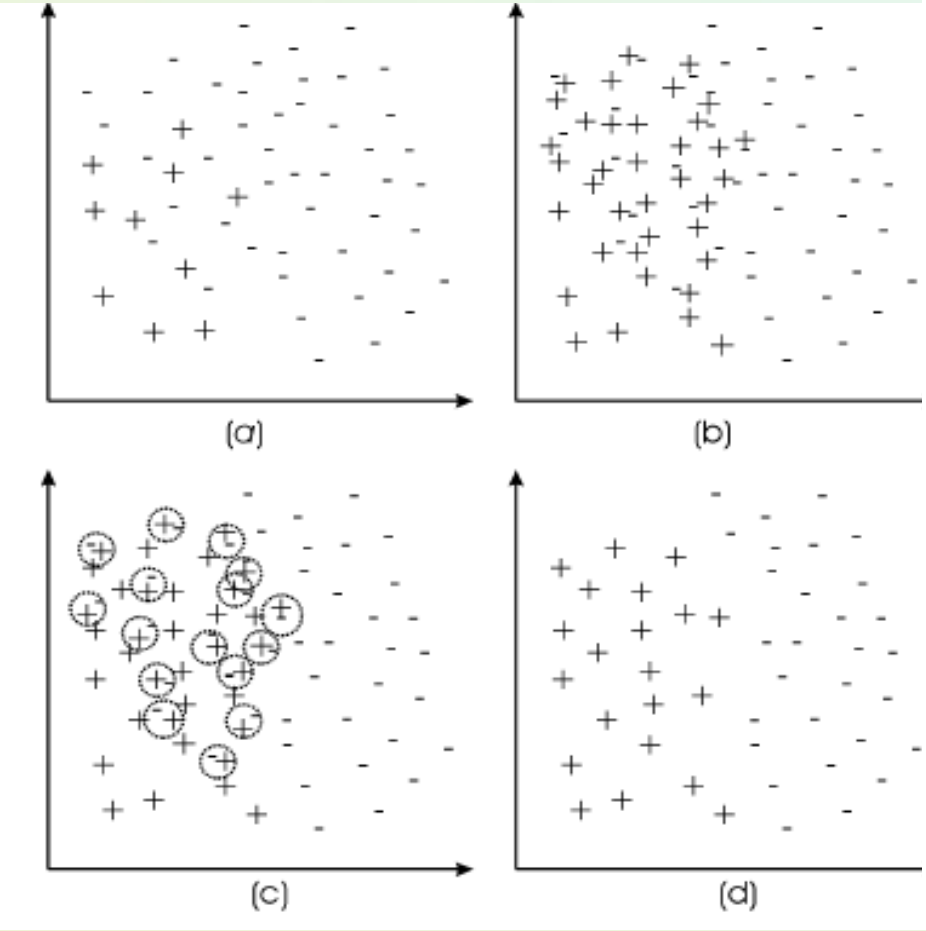
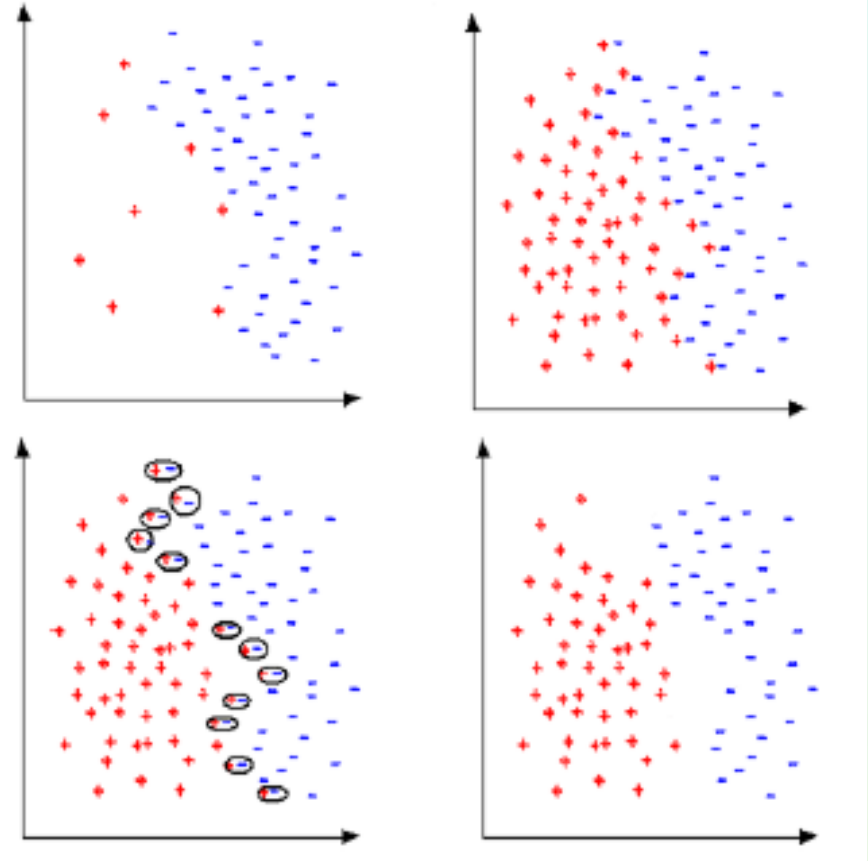


Figure: SMOTE+TomekLink



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

Najnowsze rozszerzenia SMOTE

Borderline_SMOTE: H. Han, W.Y. Wang, B.H. Mao. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International Conference on Intelligent Computing (ICIC'05). Lecture Notes in Computer Science 3644, Springer-Verlag 2005, Hefei (China, 2005) 878-887

Safe_Level_SMOTE: C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09). LNAI 5476, Springer-Verlag 2005, Bangkok (Thailand, 2009) 475-482

SMOTE_LLE: J. Wang, M. Xu, H. Wang, J. Zhang. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. IEEE 8th International Conference on Signal Processing.

LN-SMOTE: J. Stefanowski, T.Maciejewski: Local Neighbourhood in SMOTE for Mining Imbalanced Data. IEEE CIDM, 2010

SMOTE Borderline

Przykład ilustracyjny Borderline

Trzy typy przykładów mniejszościowych DANGER, SAFE, NOISE

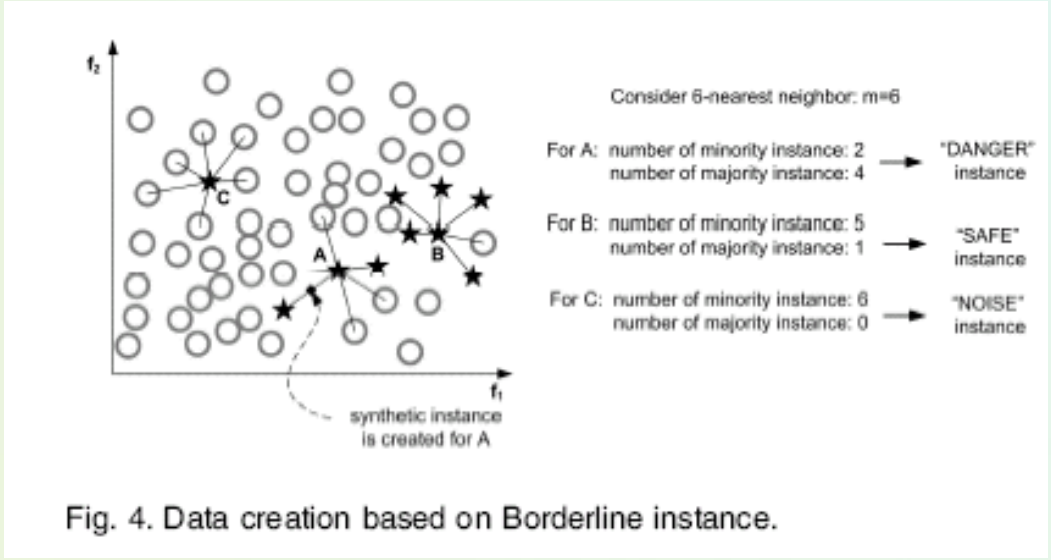
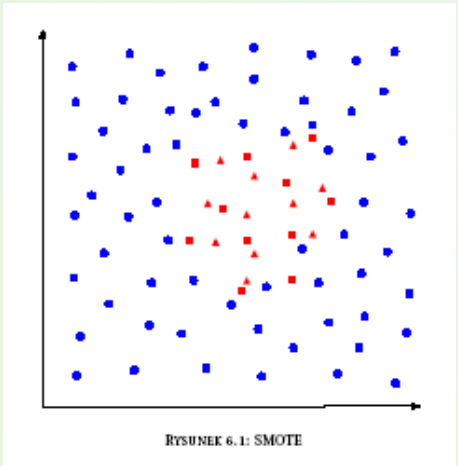
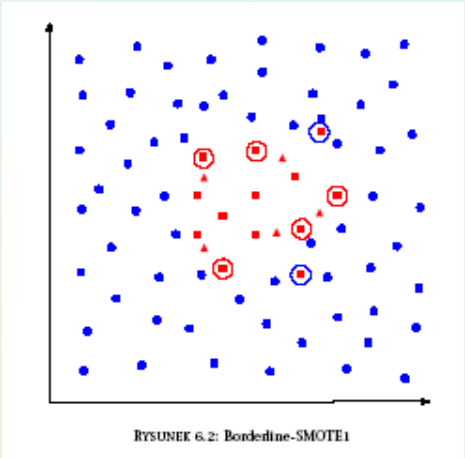


Fig. 4. Data creation based on Borderline instance.

Nadlosowyj tylko DANGER wg zasady SMOTE



RYSUNEK 6.1: SMOTE

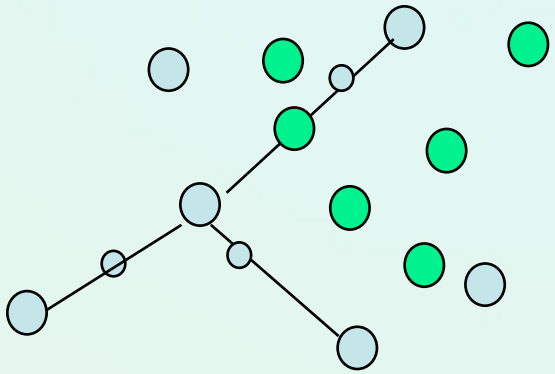


RYSUNEK 6.2: Borderline-SMOTE1

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*)
- No new synthetic examples if case of noise
- Generation of new example *x* closer to the safer region
- Random gap depends on *sl ratio*= *sl*(*p*)/*sl*(*n*)

Różne rozszerzania SMORE I C4.5 trees → 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

- LN SMOTE – największa poprawa (balance 7.25., solar flare 5.24)
- Najlepszy dla 11 z 14 danych; LN SMOTE ver 2 > LN SMOTE ver. 1
- Podobne trendy dla G-means + PART rules, k-NN

Krytyczna dyskusja → podejście hybrydowe

- NCR and one-side-sampling
  - Zachłannie usuwają (zbyt) wiele przykładów z klasy większościowej
  - Ukierunkowane głównie na poprawie wrażliwości (sensitivity) – działa lepiej dla kategorii borderline datasets
  - Mogą prowadzić do zbyt gwałtownego pogorszenia rozpoznawania klas większościowych (specificity, i inne miary) dla innych kategorii danych
- SMOTE
  - Wprowadzają b. wiele przykładów sztucznych z klasy mniejszościowej → tworzy odwrotność niezrównoważenia
  - SMOTE „ślepo” poszukuje kierunku losowania bez obserwacji położenia przykładów większościowych
  - Problematyczne w przypadku dekompozycji klasy większościowej
  - Może prowadzić do dodatkowego wymieszania klas
  - Trudność parametryzacji (k-liczba sąsiadów (5,7), lecz głównie j – stopień nadlosowania)

Selective Preprocesing of Imbalanced Data → SPIDER

- Ukierunkowane na wzrost czułości (ang. sensitivity) dla klasy mniejszościowej przy możliwie jak najmniejszym spadku specyficzności
- Rozróżnienie rodzaju przykładów: bezpieczne safe (certain lub possible); unsafe (brzegowe, noise, outliers)
- Metoda hybrydowa → ograniczony undersampling i lokalizowany over-sampling

Dwie fazy

- W przypadku klasy większościowej selektywne usunięcie noise certain i części z noise possible
  - Możliwość przetykietowania przykładów noise certain
- W przypadku klasy mniejszościowej – modyfikacje przykładów brzegowych i noise (nadlosowania)
  - weak or strong amplification / SPIDER 1 kopiowanie wybranych przykładów lub relabel (zmień etykietę większościowego)
  - Stopień wzmocnienia zależny od analizy sąsiedztwa (ENN)

Selektywny wybór przykładów – detale
J.Stefanowski, Sz.Wilk, ECML/PKDD 2007

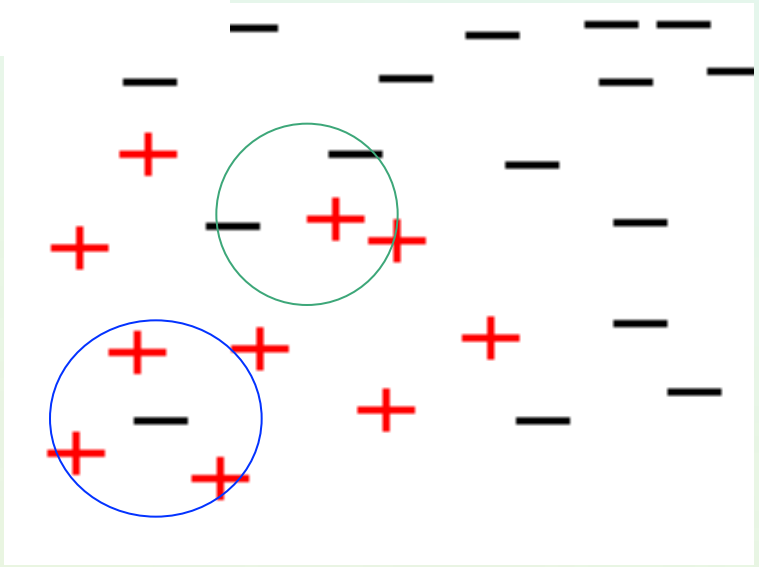
- Wykorzystanie modyfikacji zasady k- najbliższych sąsiadów (Wilson's Edited Nearest Neighbor Rule)
  → Usuń te przykłady, których klasa różni się od trzech najbliższych przykładów.
- Odległość HDVM z wykorzystaniem miary VDM Cost, Salzberg

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

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)

- Dwa etapy selekcji w klasie większościowej:
  - Identyfikacja „szumu"
  - Lokalizacja przykładów brzegowych
  - Niektóre przykłady rare i outlier → relabel

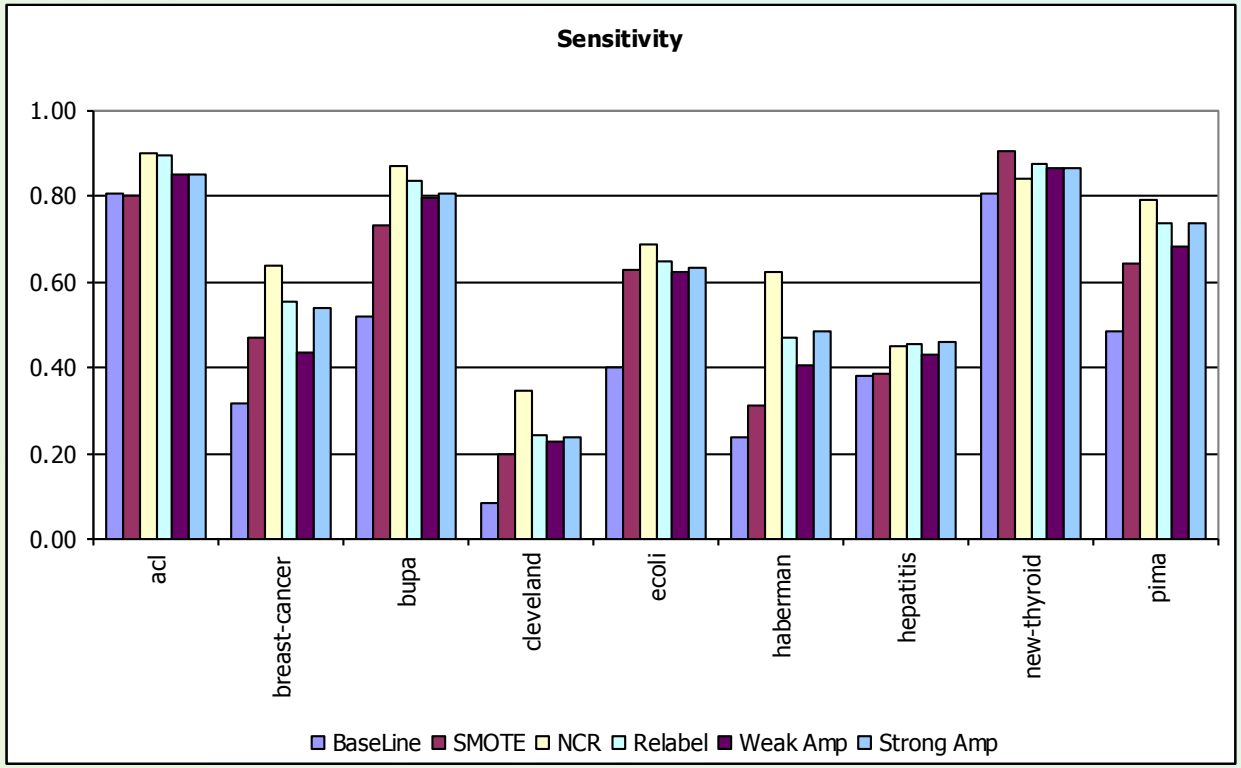


Miara czułości klasy mniejszościowej

Dane	Pojed. Klasyfik.	Under-sampling	Over-sampling	SPIDER
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

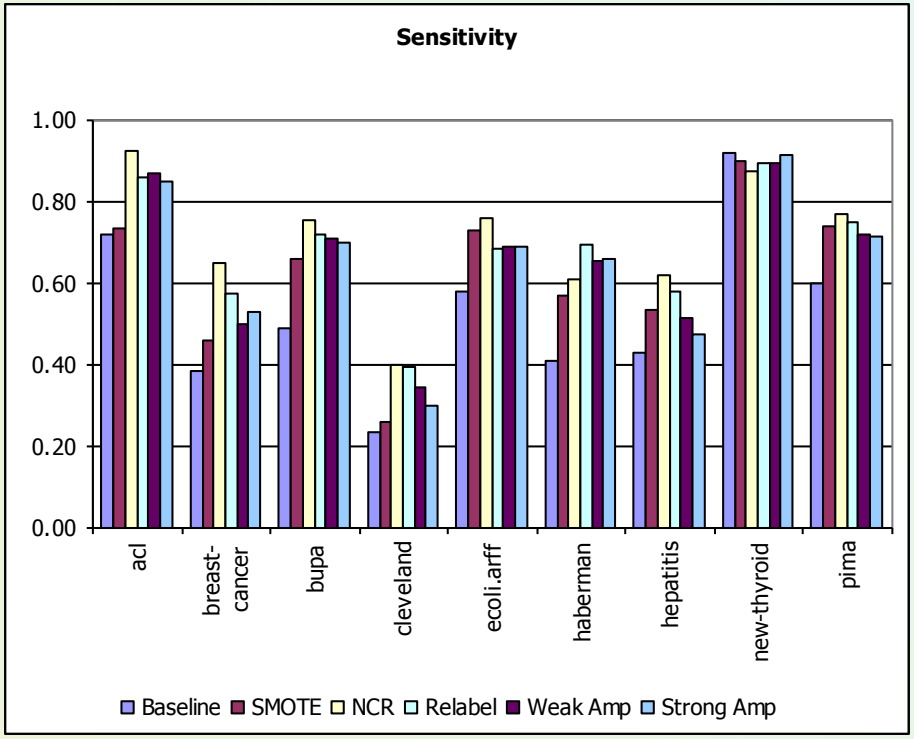
Nowe podejście zwiększa znacząco wartość miary Sensitivity

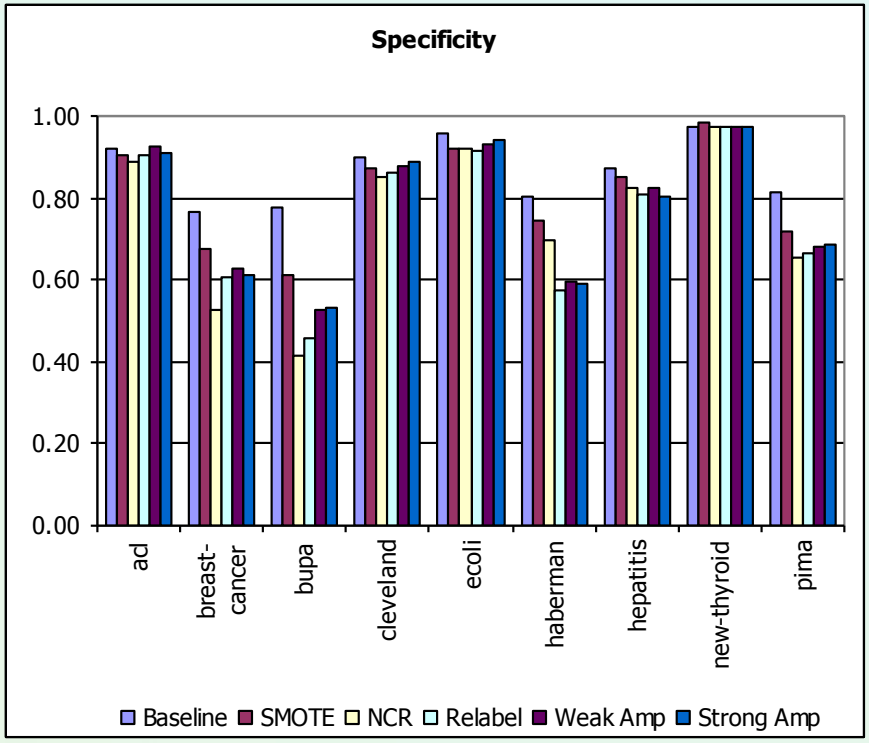
MODLEM rules → sensitivity



- 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.

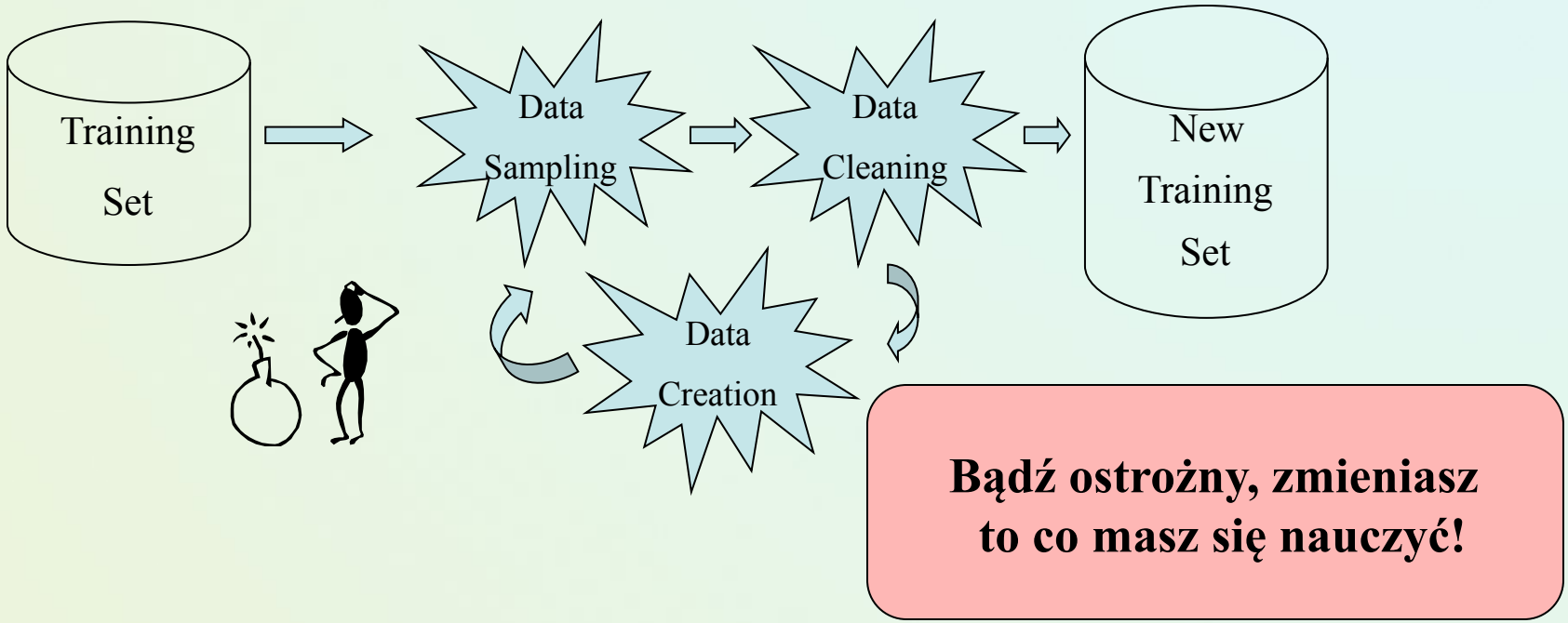
Eksperymenty dla C4.5 trees





- Similar observations (but slightly smaller changes between baseline and other approaches).
- NCR → highest increase of sensitivity at the cost of too high decrease of specificity and accuracy.
- Relabel and strong amplification → the second with respect of sensitivity.
- Weak amplification and SMOTE – quite similar performance for specificity; SMOTE - slightly better for accuracy.
- NCR → the worst overall accuracy among all approaches.

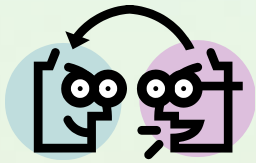
Kilka uwag podsumowujących metody informed re-sampling



- Brak jednoznacznych wskazań, które metoda jest najlepsza w danych zastosowaniu
- Problem możliwego znacznej zmiany rozkładów → klasa mniejszościowego może stać się silnie dominująca
- Trudności w doborze parametryzacji (np. SMOTE)

Podsumowanie

- Niezbalansowany rozkład liczności klas (class imbalance)
  → źródło trudności dla konstrukcji klasyfikatorów
- Typowe metody uczenia ukierunkowane są na lepsze rozpoznawanie klasy większościowej → słabe rozpoznanie mniejsz.
- Dyskusja źródeł trudności
  - Nie tylko sama niska liczność klasy mniejszościowej!
  - Rozkład przykładów i jego zaburzenia (inne czynniki)
- Rozwiązania:
  - Na poziomie danych (focused re-sampling)
  - Modyfikacje algorytmów
- Brak dobrej analizy teoretycznej metod re-sampling!!!
  - Większość niestety mocno heurystyczne
- **Nie ma jednej najskuteczniej metody** (zależne od danych)
- Dalszy wykład:
  - Modyfikacje algorytmów uczenia klasyfikatorów
  - Bardziej złożone rozkłady danych i wyzwania

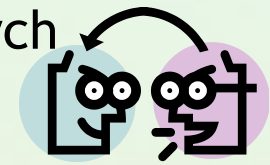


Inne odnośniki literaturowe

- 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.
- I wiele inne

Podsumowanie

- Niezbalansowany rozkład liczności klas (class imbalance) → źródło trudności dla konstrukcji klasyfikatorów
- Typowe metody uczenia ukierunkowane są na lepsze rozpoznawanie klasy większościowej → potrzeba nowych rozwiązań
- Dyskusja źródeł trudności
  - Nie tylko sama niska liczność klasy mniejszościowej!
  - Rozkład przykładów i jego zaburzenia
- Rozwiązania:
  - Na poziomie danych (focused re-sampling)
  - Modyfikacje algorytmów
- Własna metoda **selektywnego wyboru przykładów** w konstrukcji klasyfikatorów z niezrównoważonych danych
- **Podejście zmiany struktury klasyfikatora regułowego**
- Warto zapoznać się z podejściami klasyfikatorów złożonych



Dziękuję za uwagę

Koniec części pierwszej

Pytania lub komentarze?

Więcej informacji w publikacjach!

Kontakt:
Jerzy.Stefanowski@cs.put.poznan.pl