Uczenie klasyfikatorów z niezbalansowanych danych

Wykład 7

Jerzy Stefanowski
Instytut Informatyki PP

2021

Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI-TECH) projekt finansowany z środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

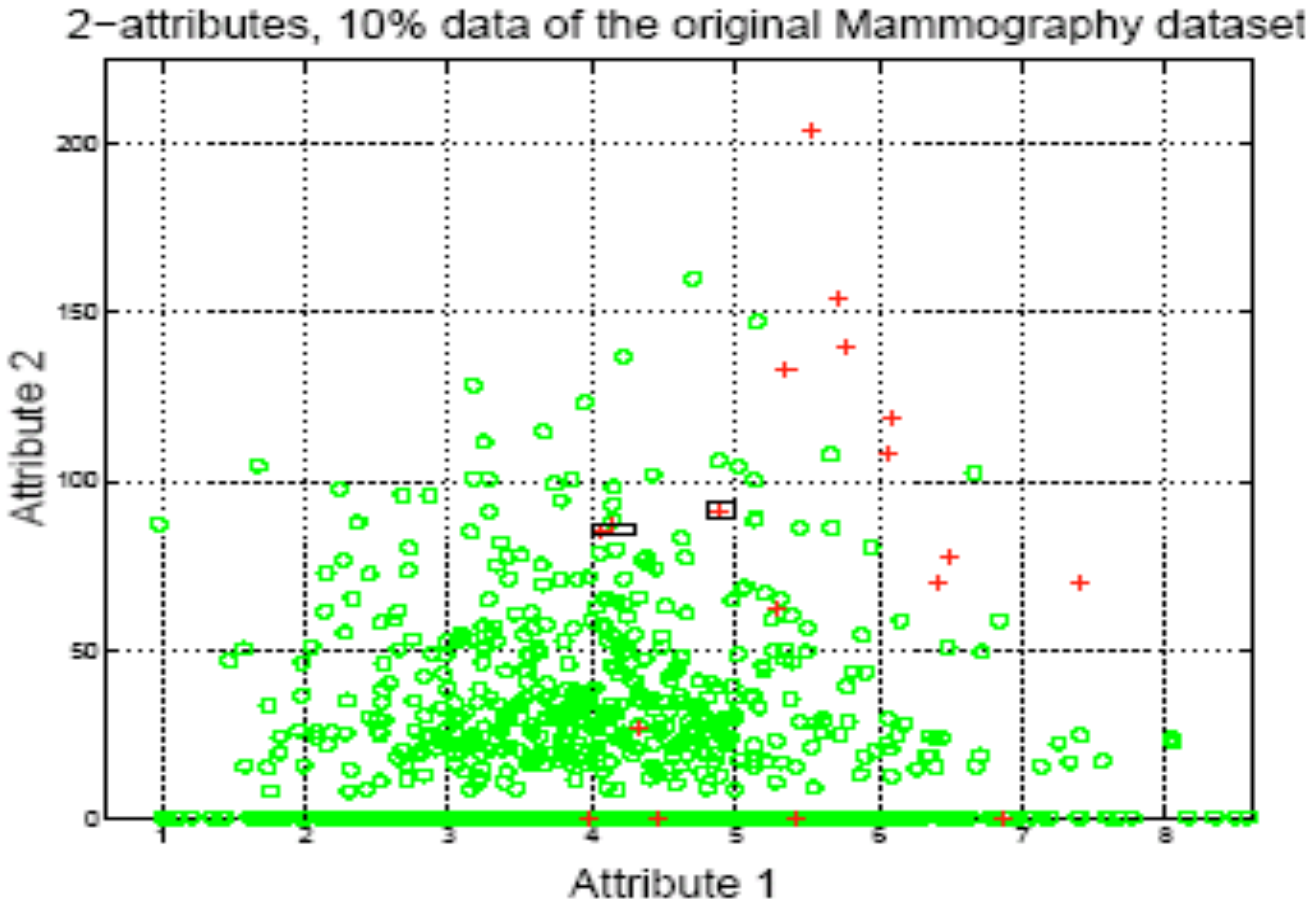
Plan wykładu

1. Niezbalansowanie liczności klas (przykłady; miary oceny – wstęp)
2. Czynniki trudności i charakterystyka danych
3. Taksonomia podstawowych metod
4. Przetwarzanie wstępne
   - Under-, over- sampling, SMOTE
   - Metody hybrydowe
5. Wybrane modyfikacje algorytmów
   1. Cost sensitive learning
   2. Zespoły klasyfikatorów (RBB i inne generalizacje) = częściowo przesunięte na późniejszy wykład
6. Ocena klasyfikatorów
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 /**zbalansowane** – rozkłady liczności przykładów w klasach względnie podobne
- Czy takie założenie jest realistyczne?
  - ***Lecz:*** „Czy medyczne dane o diagnozie rzadkich chorób są zbalansowane”
  - Rozważ, np., dane N.Chawla nt. badań mamograficznych - 11183 przykładów, 6 atrybutów, klasa mniejszościowa 2.3%
- Inne przykłady praktyczne

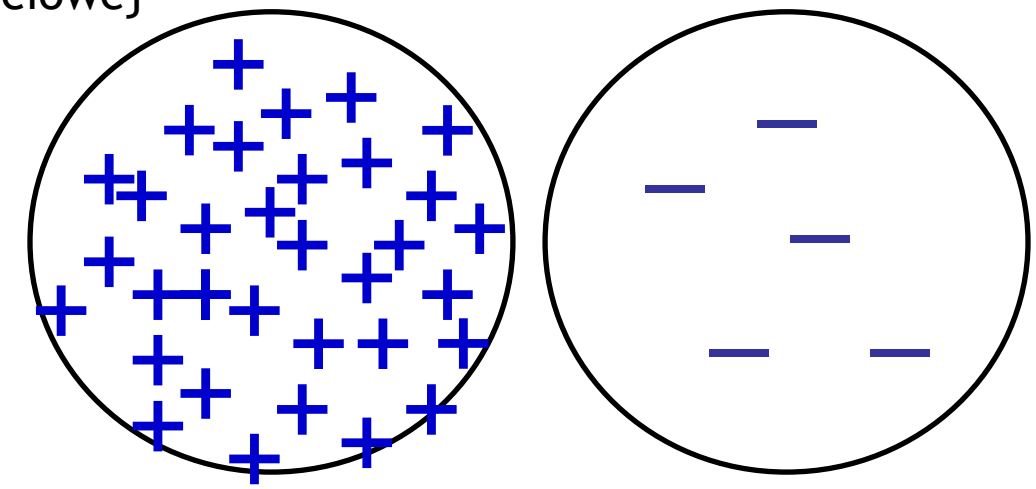
Przykład danych medycznych Chawla et al. SMOTE 2002



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, np.:
  - 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



„Niezbalansowanie to nie to samo co uczenie z kosztami”

Przykłady rzeczywistych problemów

- Niezbalansowanie klas naturalne w :
  - Analiza danych medycznych – leczenie i diagnostyka
  - Monitorowanie uszkodzeń urządzeń technicznych
  - Odróżnianie trzęsień ziemi od prób nuklearnych
  - Filtrowanie wiadomości
  - Marketing bezpośredni
  - Tzw. problem ucieczki klientów (kompanie telekomunikacyjne)
  - ……
- 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.
  - Książki = dwie monografie anglojęzyczne

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 – rzadziej badane
- 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

Złożone miary

Najpopularniejsze:

$$G - mean = \sqrt{sensitivity \cdot specificity}$$

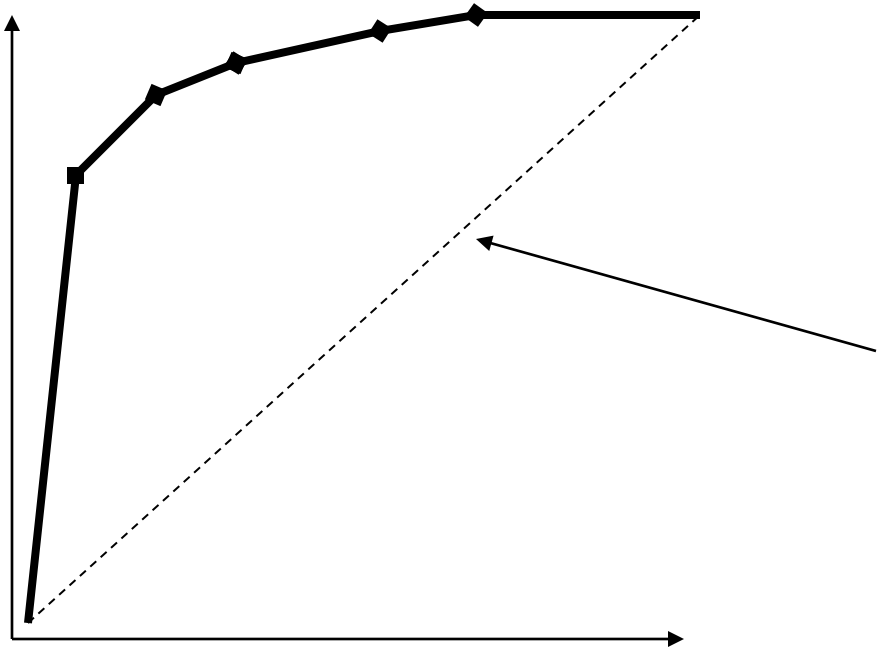
$$F_{\beta} = \frac{(1+\beta^2) \cdot \mathrm{Re}\,call \cdot \mathrm{Pr}\,ecision}{\beta^2 \cdot \mathrm{Re}\,call \cdot \mathrm{Pr}\,ecision}$$

Matthews correlation coefficient, MCC

$$MCC = \frac{TPxTN - FPx\,FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

(MCC) expresses a correlation between the actual and predicted classification and returns a value between −1 (total disagreement) and +1 (perfect agreement); 0 classifiers performs randomly

Krzywa ROC oraz AUC



Im krzywa bardziej wygięta ku górnemu lewemu narożnikowi, tym lepszy klasyfikator .

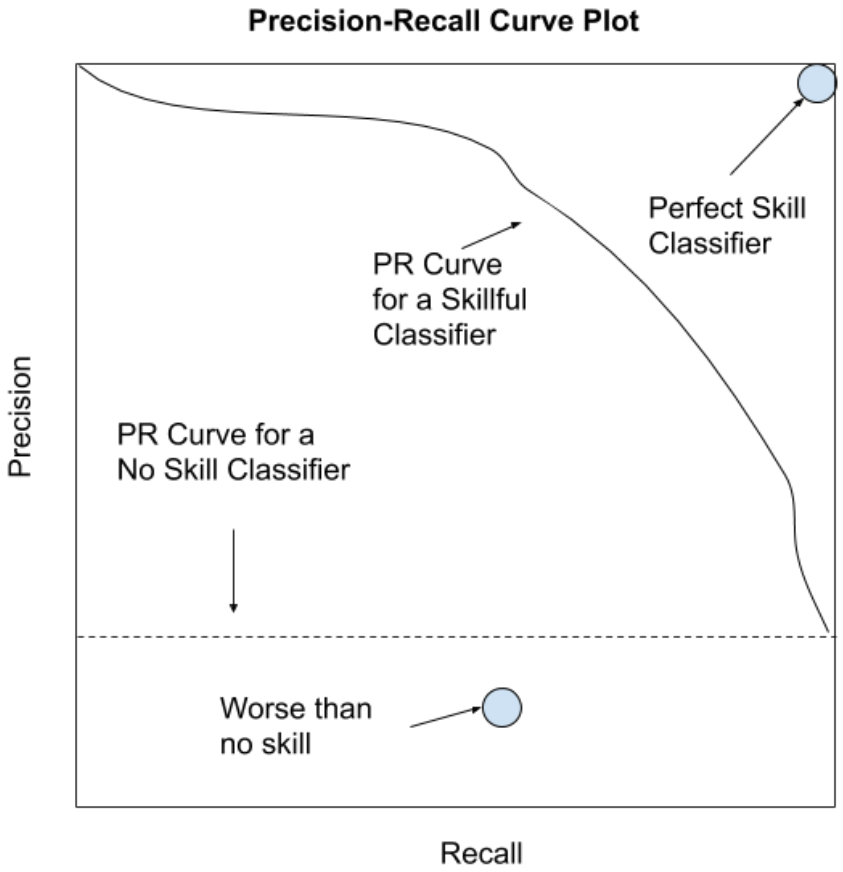
Przekątna odpowiada losowemu „zgadywaniu”. Im bliżej niej, tym gorszy klasyfikator

Można porównywać działanie kilku klasyfikatorów.
Miary oceny np. AUC – pole pod krzywą,.. Powinno być więcej niż 0.5

Precision Recall Curve

Pomimo dobrego zachowania AUC, może być zbyt optymistyczna dla silnego niezbalansowania /b. mała liczba przykładów mniejszościowych

Alternatywa – analiza krzywej precision recall – mocniej skupia się na predykcji klasyfikatora dla klasy mniejszościowej



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 (ang. pruning) faworyzują przykłady większościowe
- Strategie klasyfikacyjne ukierunkowane na klasy większościowe

Konkluzja – nie są wystarczająco dobrze przystosowane do radzenia sobie z niezbalansowaniem

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

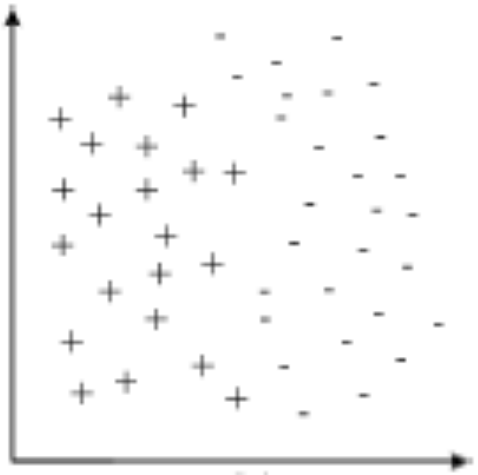
Sensitivity klasy mniejszościowej

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

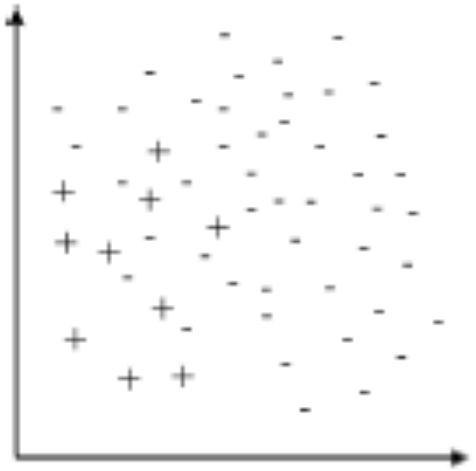
Lepsze rozpoznawania tylko Acl I New Thyroid

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:

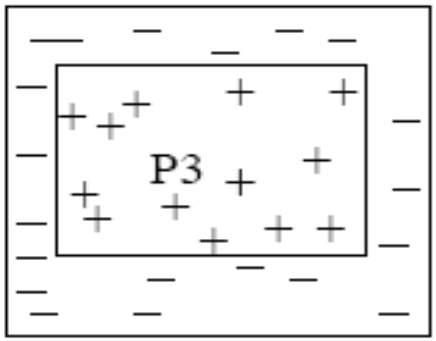
- ❑ Niejednoznaczne przykłady brzegowe
- ❑ Outliers and rare cases
- ❑ Wpływ „szumu” (noisy examples)

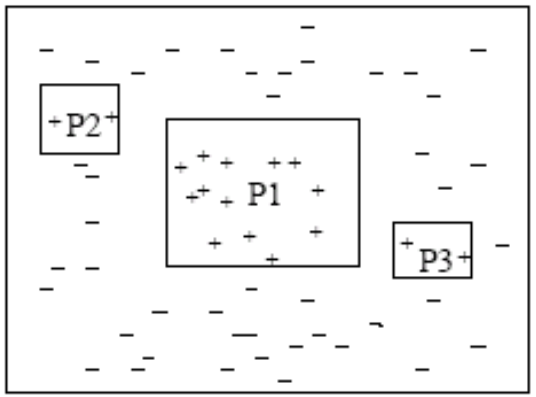
Przeglądowe prace:

- 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)



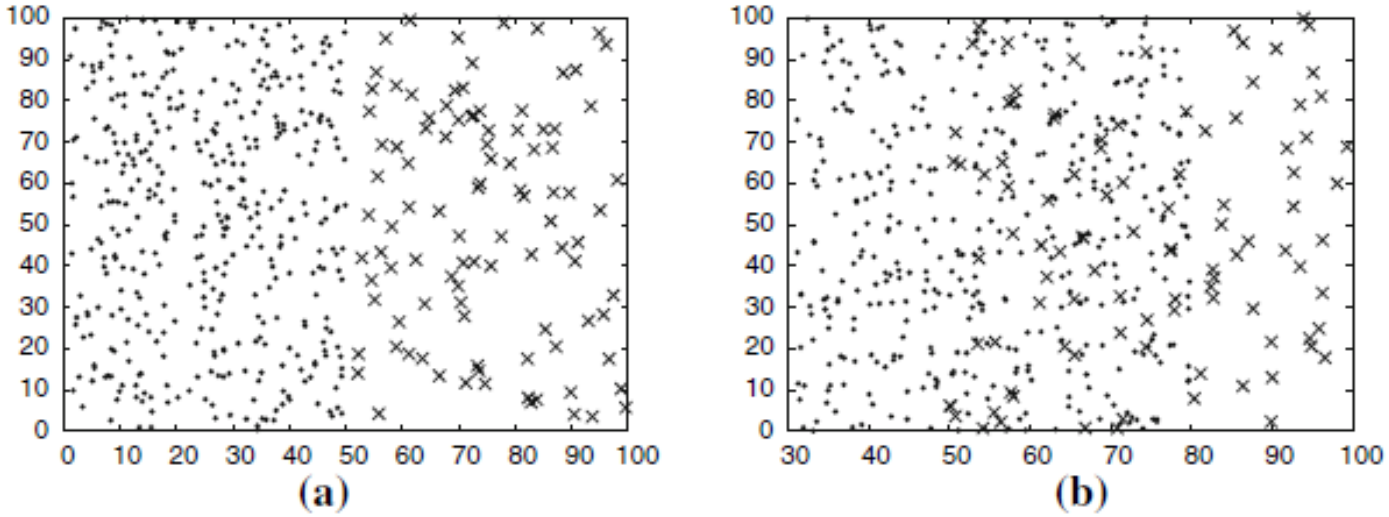


Niektóre prace eksperymentalne z dysuksją źródeł trudności, e.g:

- 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.

Eksperymenty Garcia et al. ze strefami brzegowymi

Nakładanie się rozkładów klas (ang.overlapping)

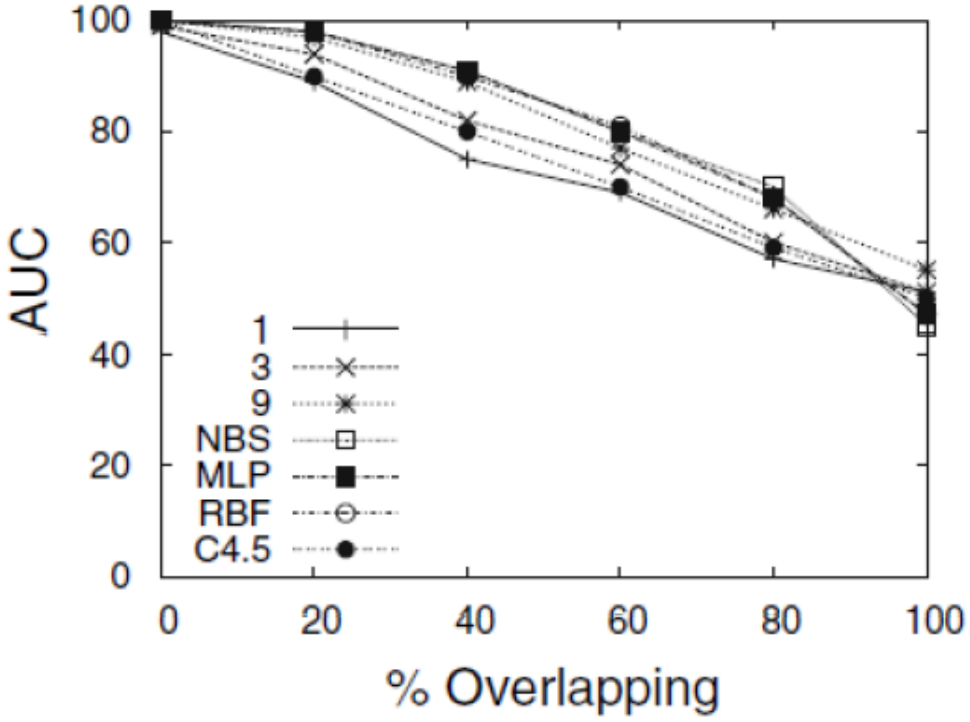


Dwa różne poziomy nakładania się 0% i 60%

Źródło: V García, J Sánchez, R Mollineda: An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. 2007.

Eksperymenty Garcia et al. ze strefami brzegowymi

Niektóre z wyników

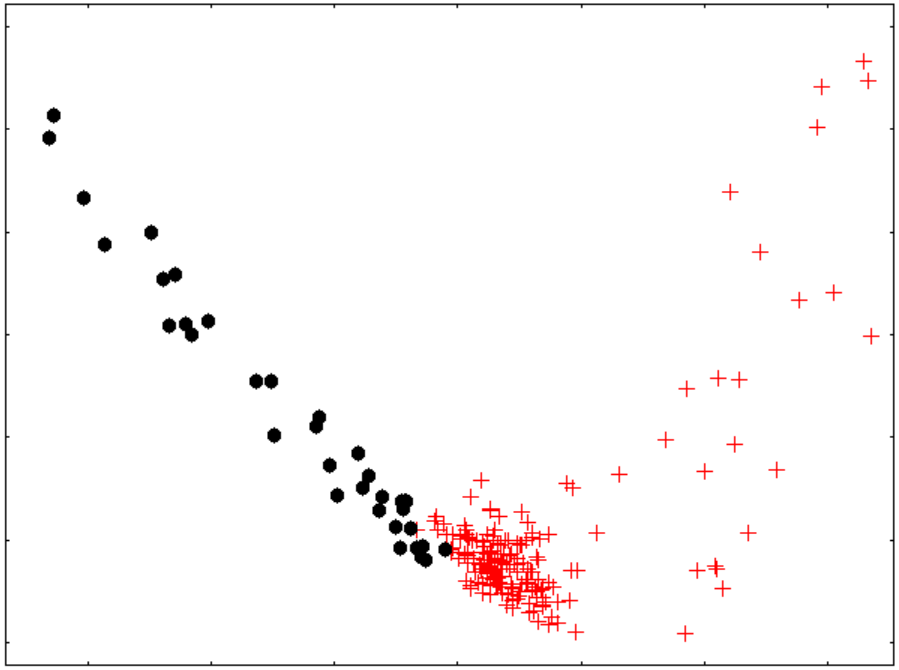


Skuteczność różnych klasyfikatorów – wzrost niejednoznaczność strefy brzegowej silniej obniża AUC niż wzrost niezrównoważenia

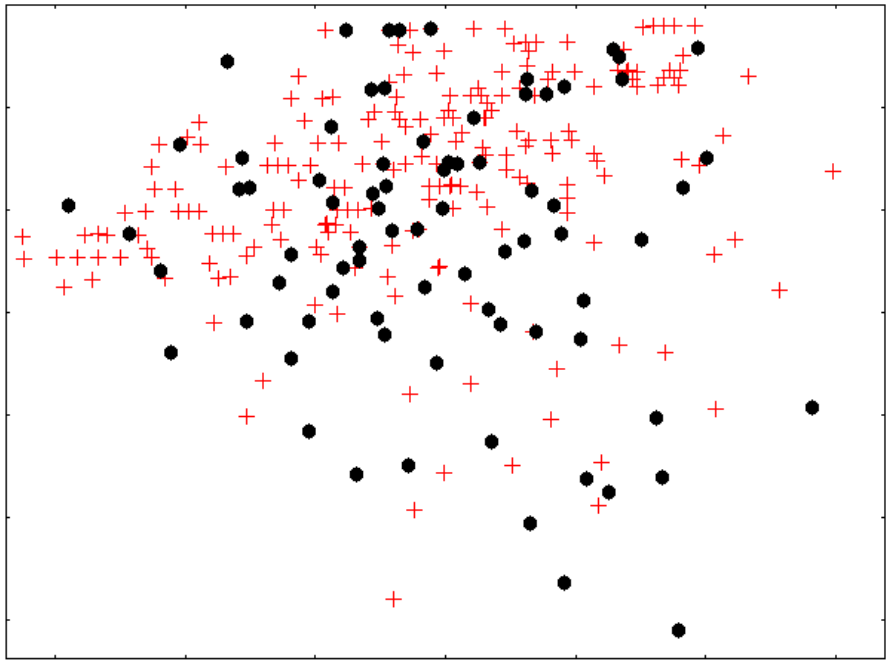
W dalszych eksperymentach zauważony wpływ lokalnej gęstości przykładów!

Co z rzeczywistymi danymi?

Wizualizacja 2 pierwszych składowych w metodzie MDS (PCA)
eksperyment własny autora i K.Napierały



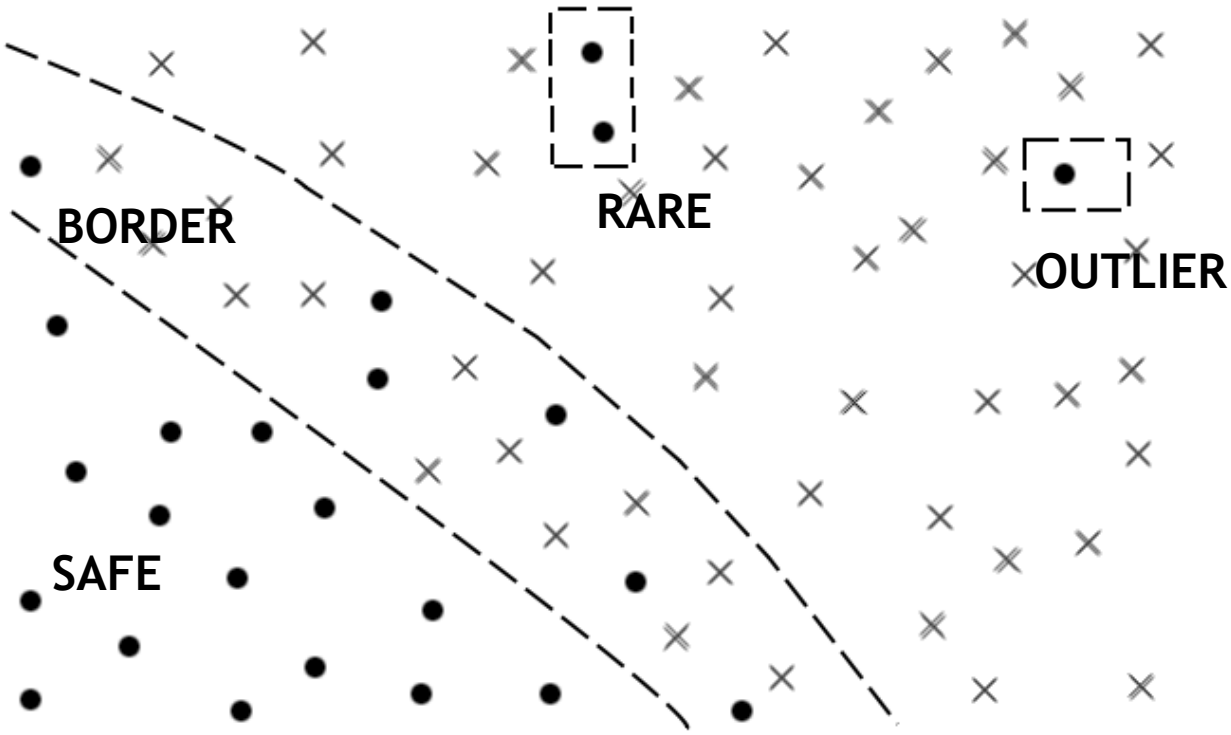
Thyroid
215 ob./ 35 mniejsz.
Prosty dla klasyfikacji



Haberman
306 ob. / 81 mniejsz.
trudny

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

Rozróżniamy 4-y typy przykładów:



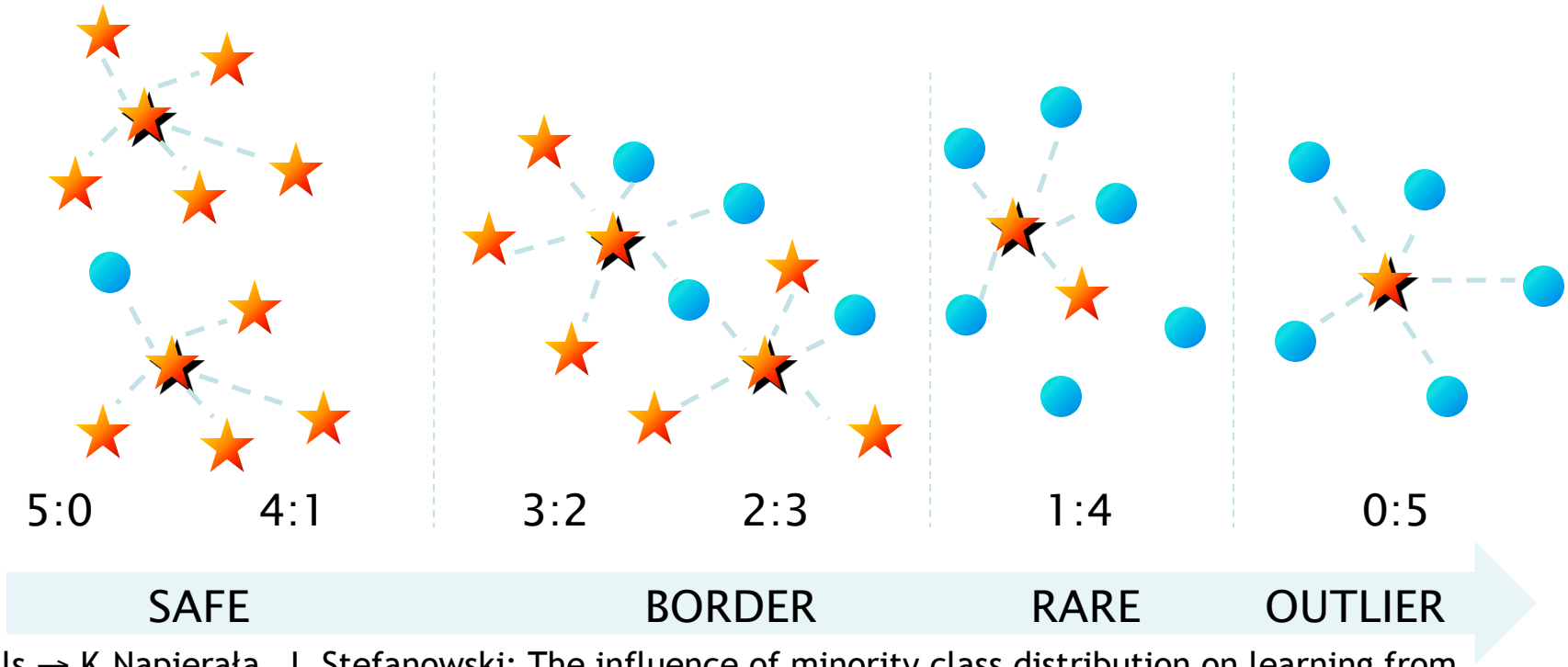
Więcej → K.Napierała, J. Stefanowski: The influence of minority class distribution on learning from imbalance data. HAIS, 2012.

Podejście do identyfikacji typów przykładów

Analizuj rozkład etykiet wśród najbliższych sąsiadów x

- K-NN (k=5, 7, ...) – HVDM distance
- Kernel functions (parametr rozproszenia)

Określ typ przykładu x

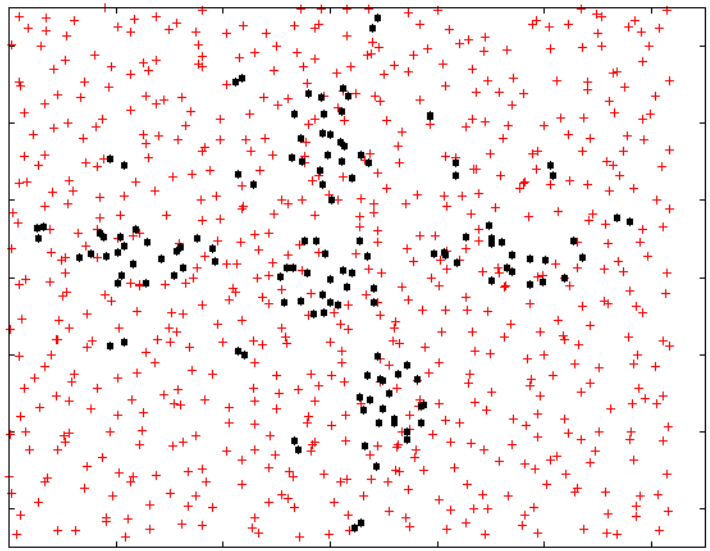


Details → K.Napierała, J. Stefanowski: The influence of minority class distribution on learning from imbalance data. HAIS, 2012.

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

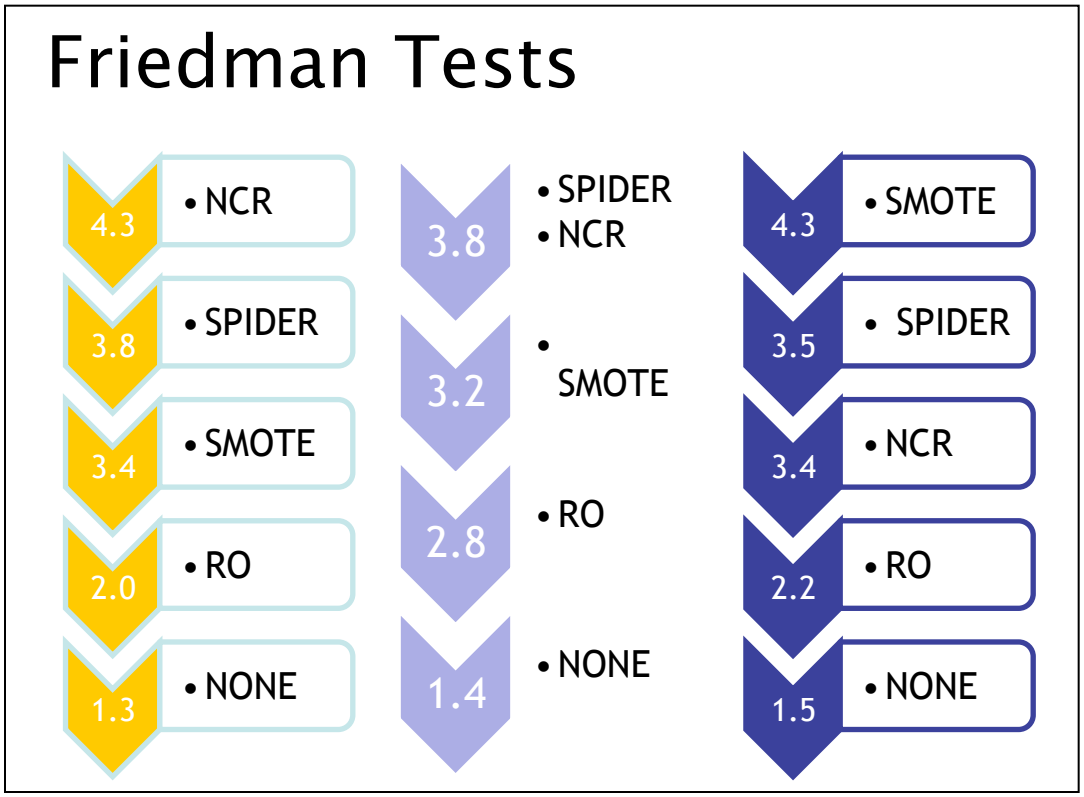


Labeling Minority Examples → UCI Data sets

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

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

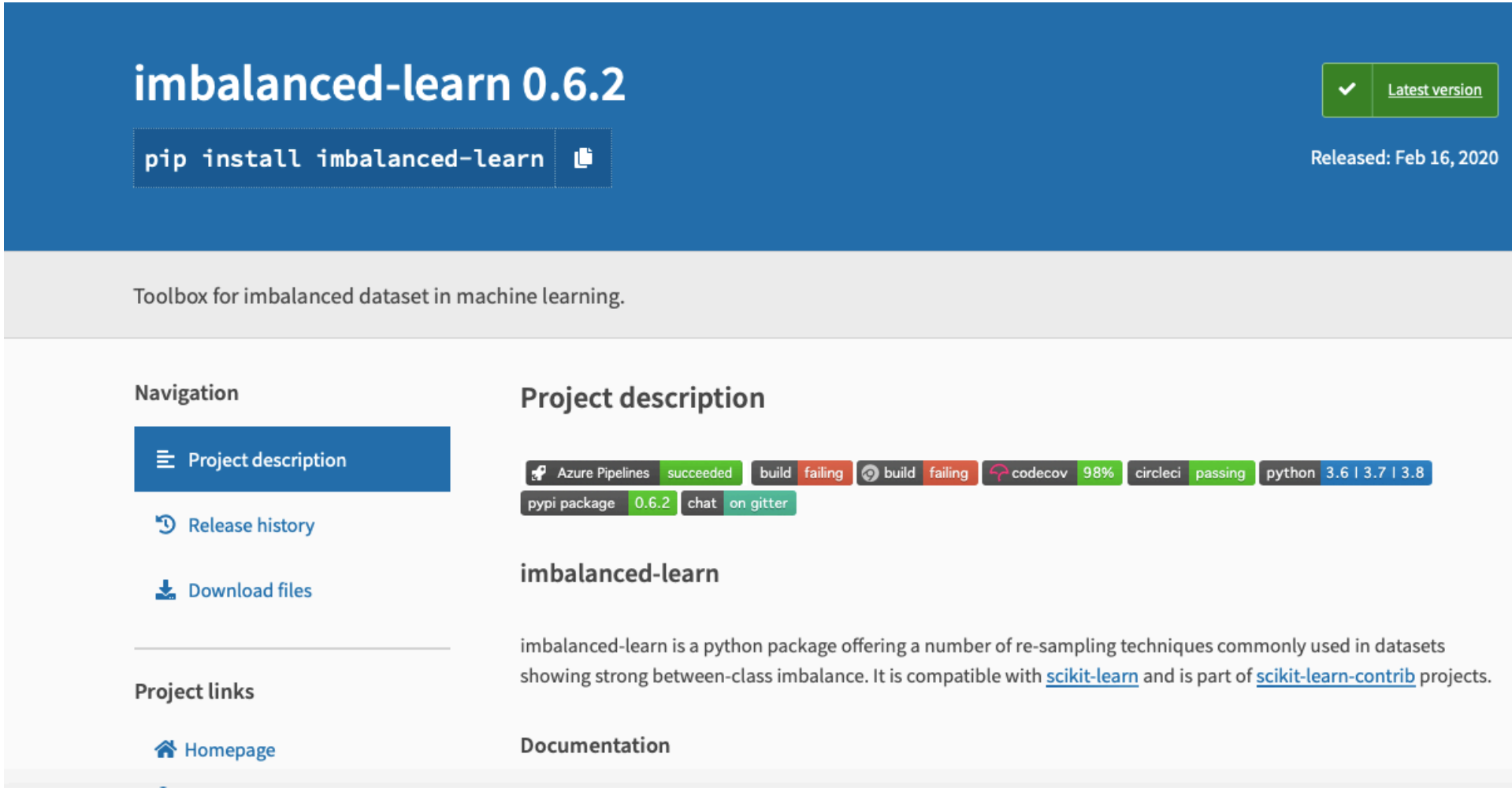
1NN, J48: podobne rankingi

RBF: RO wyżej w rankingu

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

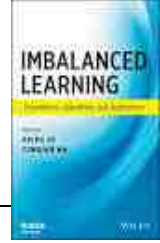
KEEL – więcej algorytmów (45) over i under sampling (20) oraz rozbudowane zespoły klasyfikatorów (21)

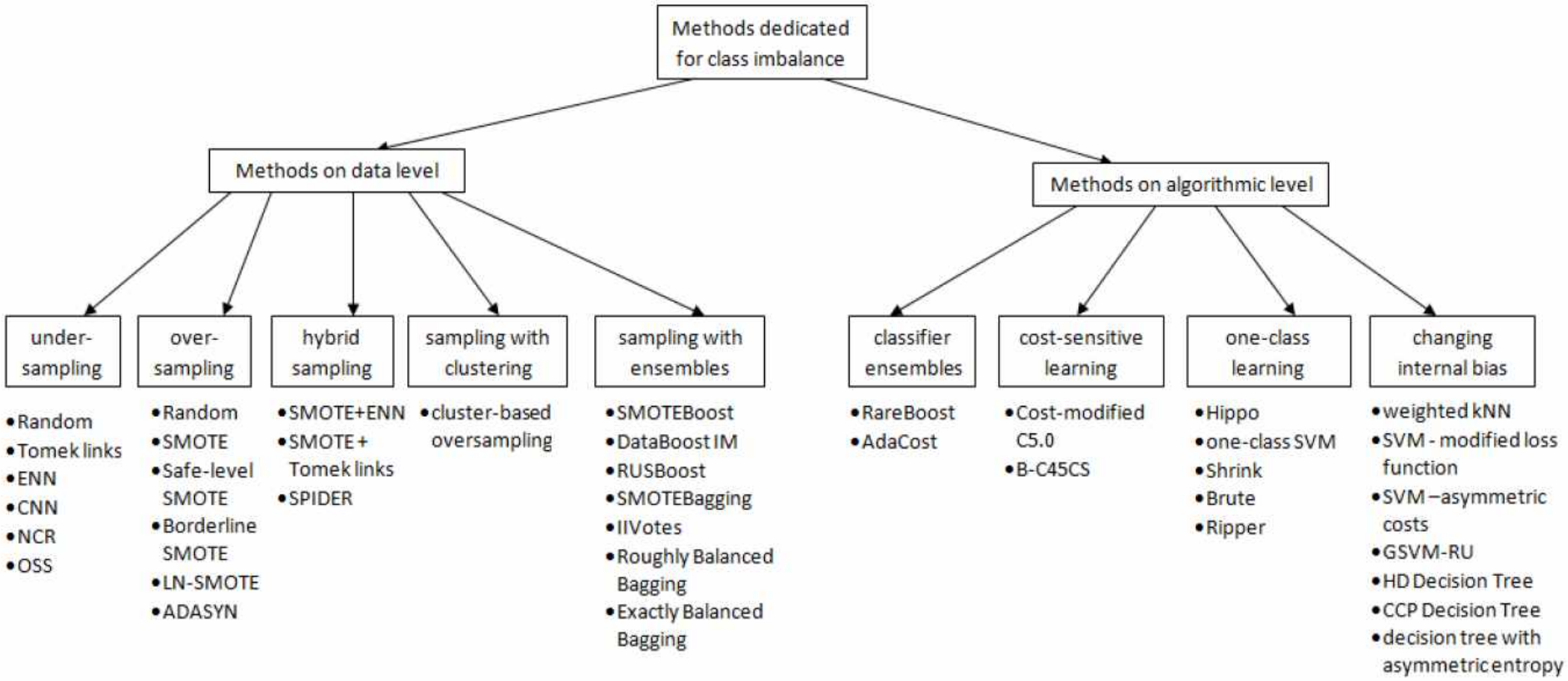
R package ‘imbalance’ oraz IRIC: An R library for binary imbalanced classification: 29 metod re-sampling, 4 zespoły klasyfikatorów, 1 cost sensitive

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”
  - …

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.

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.
  - Hellinger distances i asymetryczne entropie (Chawla et al.)
- 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):

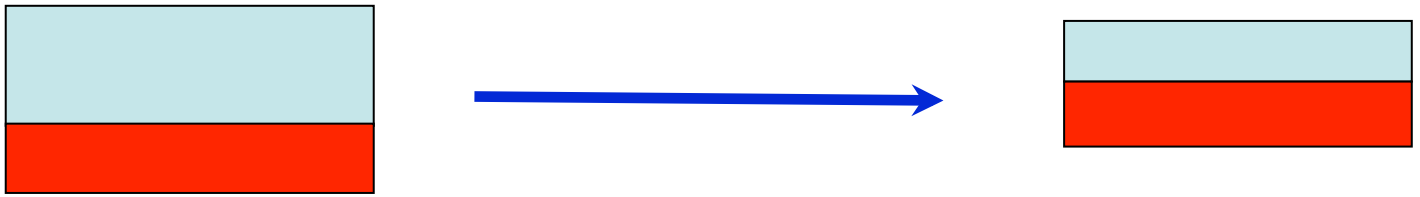
- Proste techniki losowe
  - „Over-sampling” – klasa mniejszościowe
  - „Under-sampling”- klasa mniejszościowa
- Specjalizowane nadlosowanie
  - Cluster-oversampling (Japkowicz)
- **Ukierunkowane transformacje**
  - Klasa większościowe
    - One-side-sampling (Kubat, Matwin) z Tomek Links
    - Laurikkala’s 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

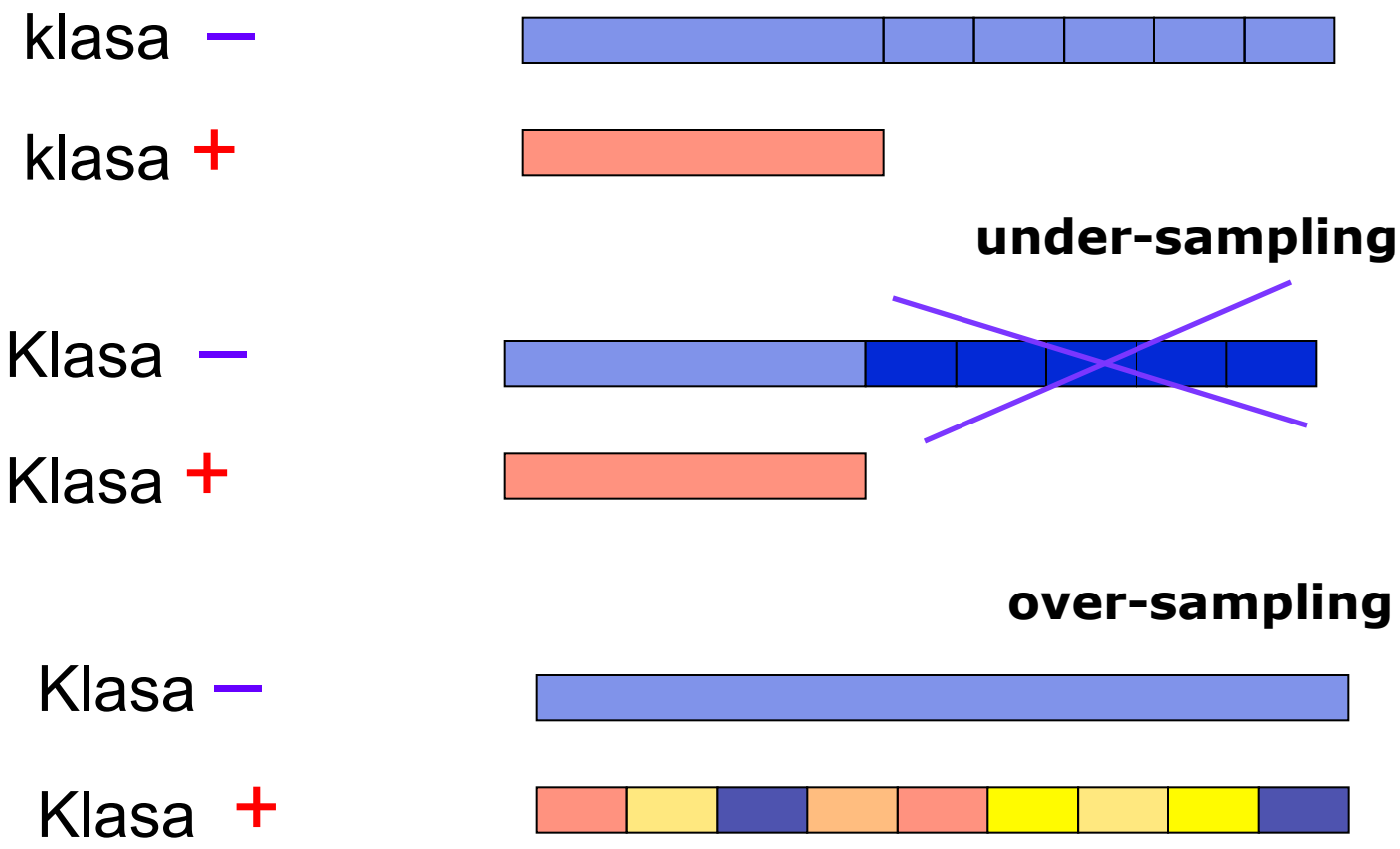
Raczej heurystyka ukierunkowana na uzyskanie lepszych rozkładów klas niż uzasadnione teoretycznie podejście [F.Herrera 2010].

Brak teoretycznej gwarancji znalezienia optymalnej postaci rozkładu!



Losowe nadlosowanie lub usuwanie przykładów

ang. undersampling vs oversampling



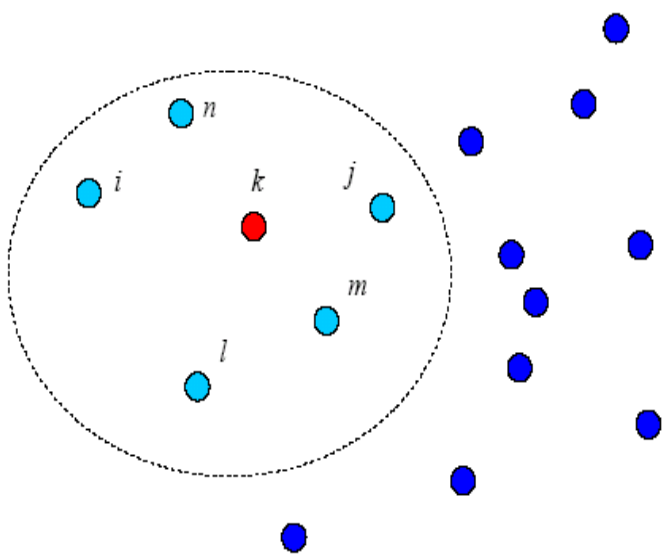
SMOTE - Synthetic Minority Oversampling Technique

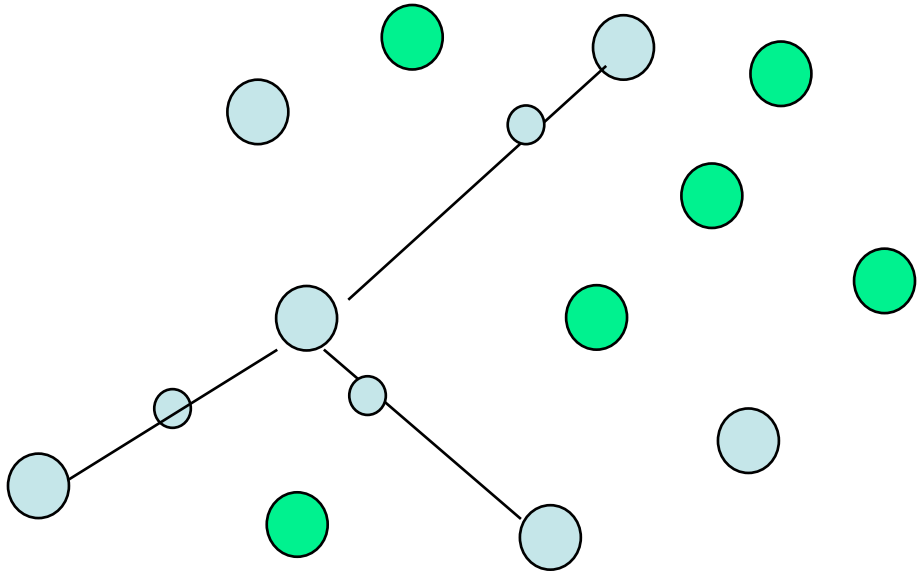
- Wprowadzona przez Chawla, Hall, Kegelmeyer 2002
- Dla każdego przykładu *p* z klasy mniejszościowej
  - Znajdź jego k-najbliższych sąsiadów (UWAGA wyłącznie z klasy mniejszościowej!)
  - Losowo wybierz j z powyższych sąsiadów
  - Losowo stwórz sztuczny przykład wzdłuż lini łączącej p z wybranym losowo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

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 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

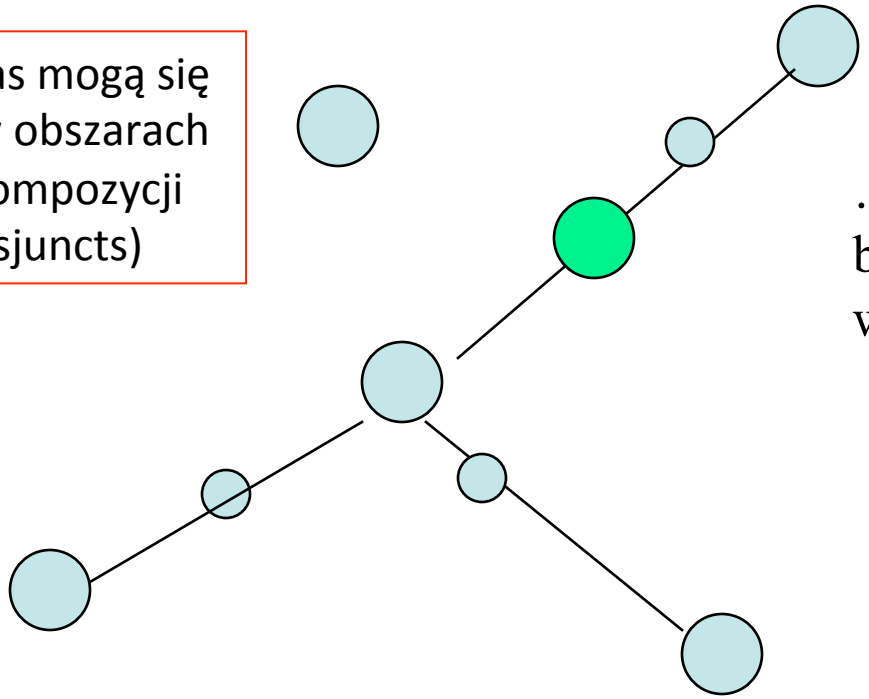
SMOTE – uwagi krytyczne

- Ślepe nadlosownie
  - 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
- Trudność strojenia
  - 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 mogą 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

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

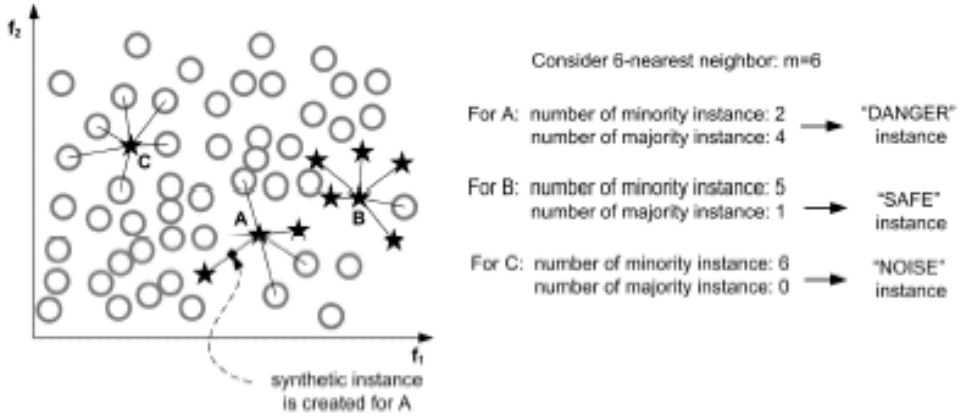
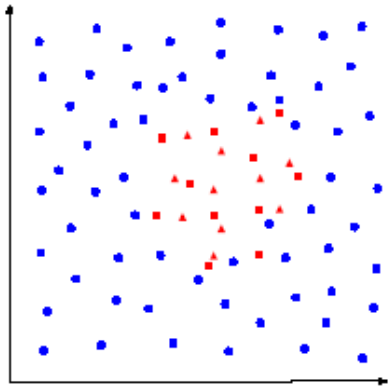
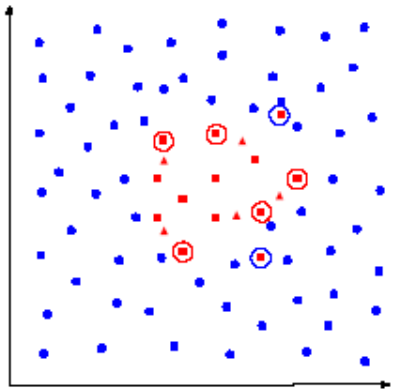


Fig. 4. Data creation based on Borderline instance.

Nadlosowyj tylko DANGER wg zasady SMOTE



RYSUNEK 6.1: SMOTE



RYSUNEK 6.2: Borderline-SMOTE1

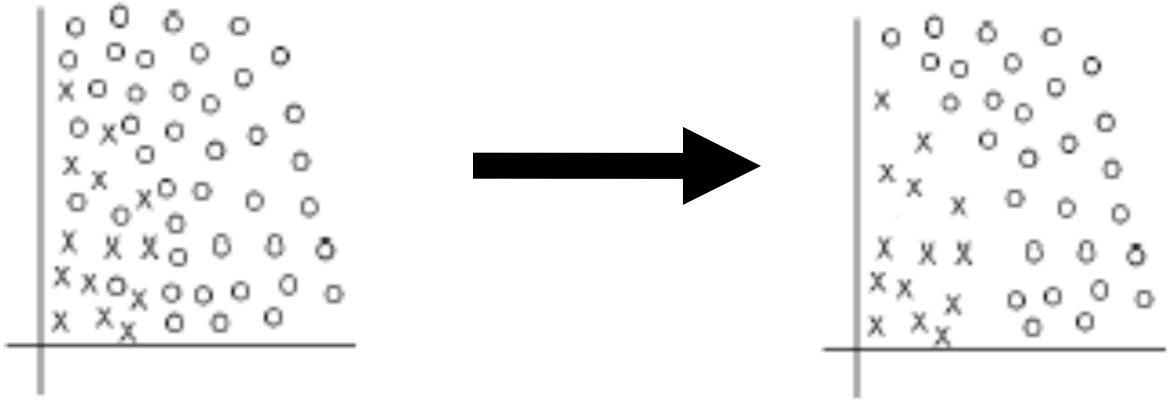
Różne rozszerzania SMOTE I C4.5 trees → F-measure

	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

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)?

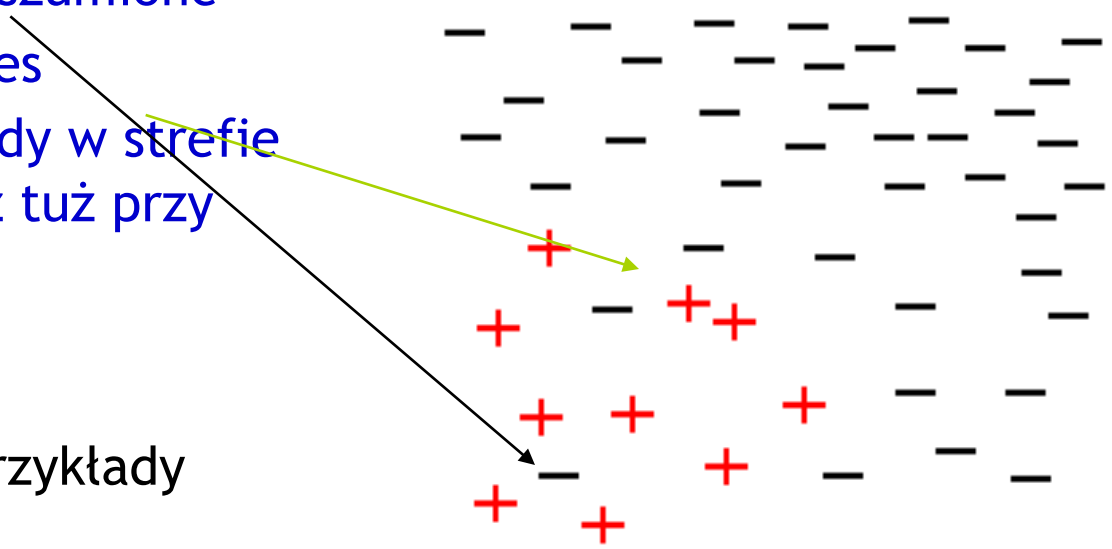
Powróćmy do charakterystyki przykładów

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

Różne typy przykłady

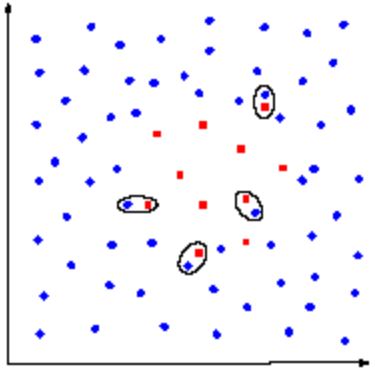
- Noise przykłady zaszumione
- Borderline examples

  Trudne przykłady w strefie brzegowej oraz tuż przy granicy.
- Rzadkie przykłady
- Safe bezpieczne przykłady



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

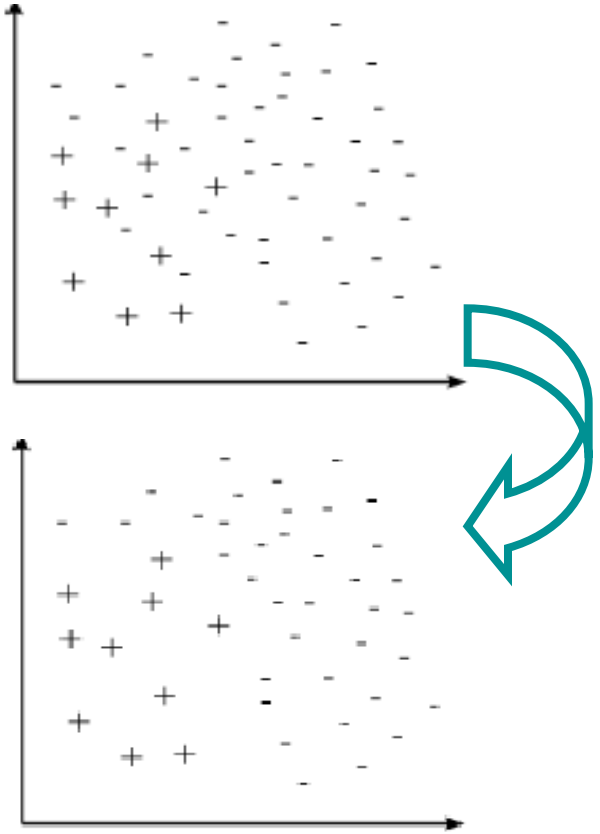
Under-sampling z wykorzystaniem Tomek links



Przykład Tomek Links

- Usuwaj przykłady graniczne i szum z klasy większościowej
- „Tomek link”
  - Ei, Ej należą do różnych klas,

    d (Ei, Ej) odległość między nimi.
  - para (Ei, Ej) jest tzw. Tomek link jeśli nie istnieje inny przykład El, spełeniający d(Ei, El) < d(Ei, Ej) or d(Ej , El) < d(Ei, Ej).



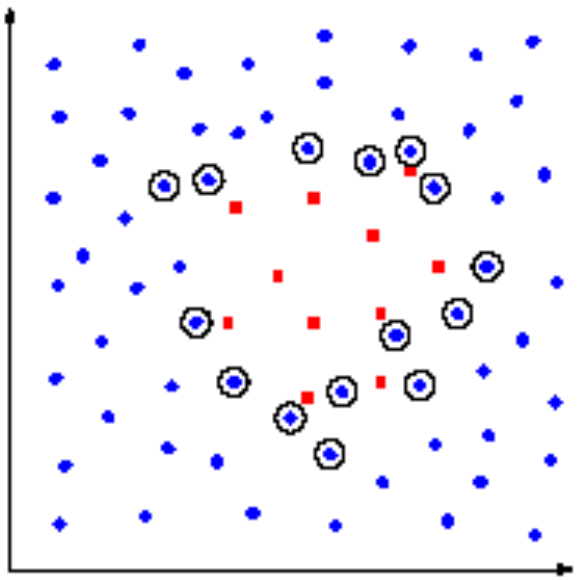
Nearest Cleaning Rule

- **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



RYSUNEK 5.3: Neighbourhood Cleaning Rule

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)

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

Cost learning

Potrzeba zdefiniowania macierzy kosztów pomyłek

	Actual = negative	Actual = positive
Predict = negative	***TN***	***FN***
Predict = positive	***FP***	***TP***

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

Positive – Minority class

Imbalanced FN is more dangerous than FP !

Zwykle C(0,1) większe niż C(1,0)

Cost learning

The cost of labeling an example incorrectly should always be greater than the cost of labeling it correctly.[C.Elklan]

C(0,1) >> C(1,0) i ….

	True = 0	True = 1
Predict = 0	***0***	***80***
Predict = 1	***5***	***0***

Jak zdefiniować precyzyjne wartości kosztów?

Jak je wykorzystać w klasyfikacji niezbalansowanych danych?

“In cost-sensitive learning instead of each instance being either correctly or incorrectly classified, each class (or instance) is given a misclassification cost. Thus, instead of trying to optimize the accuracy, the problem is then to **minimize the total misclassification cost**.”

Definiowanie kosztów (globalne dla klasy)

Wiedząc, że koszt nierozpoznanie klasy mniejszościowej jest większy C(0,1) >> C(1,0)

Prosto – ustal koszty proporcjonalnie do stopnia niezbalansowania, np.

	True = 0	True = 1
Predict = 0	***0***	***1*IR***
Predict = 1	***1***	***0***

Nguyen, Gantner, Schmidt-Thieme: Cost-sensitive learning methods for imbalanced data

Potraktuj to jako hiper-parametr o lokalnej optymalizacji (wewnętrzna ocena krzyżowa)

Koszty pomyłek mogą być zdefiniowane dla poszczególnych przykładów z klasy = trudniejsze podejście

Cost sensitive learning

Cost-Sensitive Learning is a type of learning that takes the misclassification costs (and possibly other types of cost) into consideration. The goal of this type of learning is to minimize the total cost [Ling,Sheng]

Dla danej macierzy kosztów, przykład klasyfikuje się do klasy z minimalnym oczekiwanym kosztem

$$R(i \mid x) = \sum_j P(j \mid x) \cdot C(i, j)$$

gdzie P(j|x) jest estymatą prawdopodobieństwa przydziału x do j-tej klasy.

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

Przydziel x do klasy pozytywnej / mniejszościowej, gdy

$P(0|x)C(1,0)+P(1|x)C(1,1) \leq P(0|x)C(0,0)+P(1|x)C(0,1)$

można przekształcić do

$$P(0|x)(C(1,0)-C(0,0)) \leq P(1|x)(C(0,1)-C(1,1))$$

wiedząc, że $C(0,0)=C(1,1)=0$ otrzymujemy

$P(0|x)C(1,0) \leq P(1|x)C(0,1)$ oraz $P(0|x)=1-P(1|x)$

Otrzymujemy próg p* pozwalający na klasyfikacje przykładu x do klasy pozytywnej, gdy

$$p^* = \frac{C(1,0)}{C(1,0)+C(0,1)}$$

Kalibracja – dane zbalansowane p*=0.5

Niezbalansowanie mniejszościowa p* < 0.5

Reguły i niezrównoważenie klas

- zbiór uczący Ecoli: 336 ob. i 35 ob. w klasie M ; 7 atr. liczbowych
- MODLEM (noprune) 18 reguł, w tym 7 dla 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%]
...

- A strategia klasyfikacyjna:
  - Niejednoznaczne wielokrotne dopasowanie? Głosowanie większościowe
  - Brak dopasowania? – reguły najbliższe

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

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

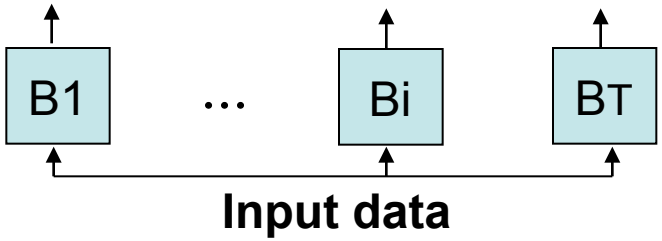
Adaptacje zespołów klasyfikatorów

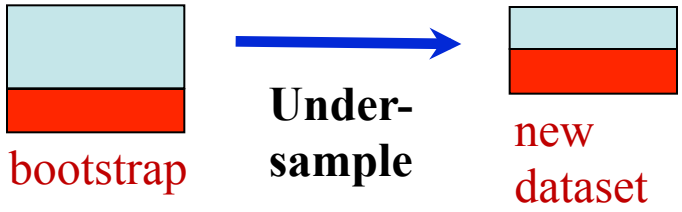
- Data preprocessing + ensemble
  - Boosting-based
    - SMOTEBoost, DataBoost
  - Bagging-based
    - Exactly Balanced Bagging
    - Roughly Balanced Bagging
    - OverBagging
    - UnderOverBagging
    - SMOTEBagging
    - NBBag
- Inne or Hybrid (EasyEnsemble)
- Cost Sensitive Boosting
  - AdaCost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

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

Under- Bagging – popularne rozszerzania

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



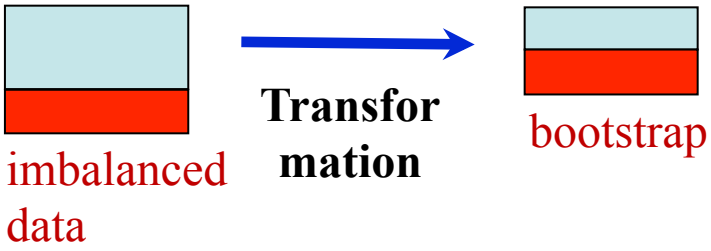


Propozycje z Undersampling

- 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]
  - Inaczej – wyrównuje prawdopodobieństwa klas w losowaniu

Roughly Balanced Bagging

Hido S., Kashima H.: Roughly balanced bagging for imbalance data (2008)



Data preprocessing + ensemble

- Under-sampling modification of Exactly Balanced Bagging
- Instead of fixing the constant sample size, it equalizes the sampling probability of each class
- For each of T iterations the size of the majority class in the bootstrap BS_{maj} is determined probabilistically according to the negative binominal distribution

For each bootstrap

- Random size BS_{maj}
- Sample with replacement N_{min} and BS_{maj}

Prediction with majority voting

- Przykładowe rozszerzenia:
  - Attribute Selection with RBBag for highly dimensional data
  - Multi-class generalization (changing sampling idea)

Lango M., Stefanowski J.: The Usefulness of Roughly Balanced Bagging for Complex and High-dimensional Imbalanced Data (2016)

Wybrane otwarte problemy

- Lepsze zrozumienie problemu
  - Analiza sztucznych I rzeczywistych danych
  - Lepsze wykrywania dekompozycji na pod-pojęcia
  - Teoretyczna analiza wybranych metod
- Multi-class imbalanced data
- Nowe miary oceny
- Rozważanie danych wielowymiarowych
- Uczenie przyrostowe
- Niezbalansowane strumienie danych i zmiany podjęć
- Niezbalansowanie regresji, alg. skupień, its.
- Large scale imbalanced learning i Big Data

Spójrz do B.Krawczyk Learning from imbalanced data: open challenges and future directions (2016)

Literatura przeglądowa

1. G. M. Weiss. Mining with Rarity: A Unifying Framework. SIGKDD Explorations, 6(1):7-19, June 2004
2. Chawla N., Data mining for imbalanced datasets: an overview. In The Data mining and knowledge discovery handbook, Springer 2005.
3. Garcia V., Sánchez J.S., Mollineda R.A., Alejo R., Sotoca J.M. The class imbalance problem in pattern classification and learning. pp. 283-291, 2007
4. Visa, S. and Ralescu, A. Issues in mining imbalanced data sets - a review paper. Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference, Dayton, pp.67-73, 2005
5. 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.
6. He, H. and Garcia, E. A. Learning from Imbalanced Data. IEEE Trans. on Knowl. and Data Eng. 21, 9 (Sep. 2009), pp. 1263-1284, 2009

IEEE ICDM noted "Dealing with Non-static, Unbalanced and Cost-sensitive Data" among the **10 Challenging Problems in Data Mining Research**

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

Pytanie i komentarze?

Dalszy kontakt:

jerzy.stefanowski@cs.put.poznan.pl

http://www.cs.put.poznan.pl/jstefanowski/

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego