Odkrywanie wiedzy klasyfikacyjnej z niezrównoważonych danych

Learning classifiers from imbalanced data
Wpływ niezrównoważenia klas na klasyfikator

Wykład ZED dla specjal. TPD

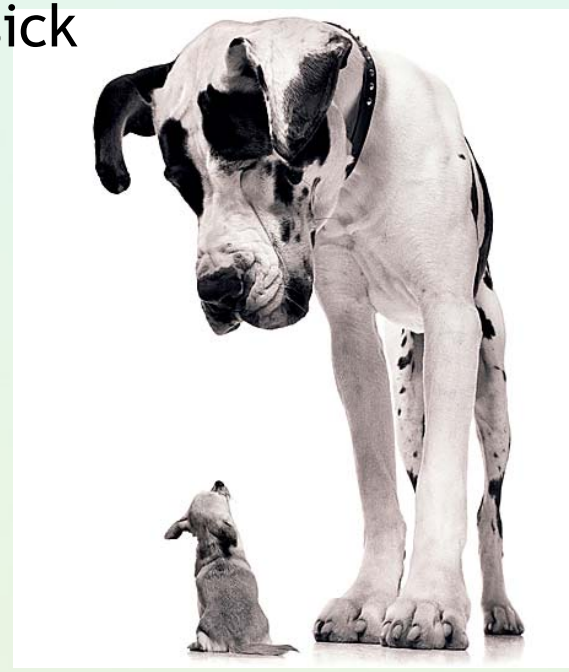
JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska
Poznań

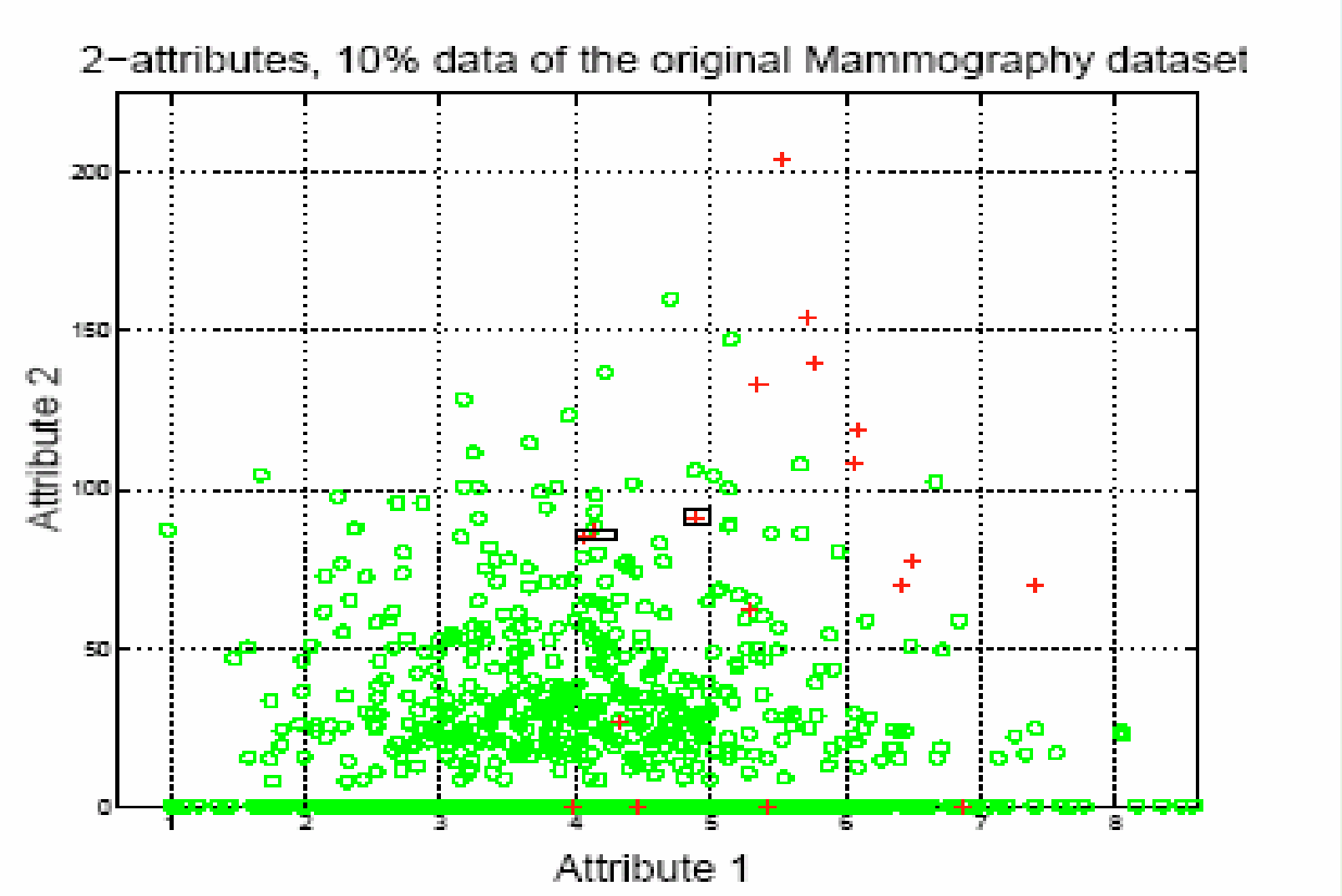
Poznań, grudzień 2010

Uczenie się klasyfikatorów z niezrównoważonych 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?

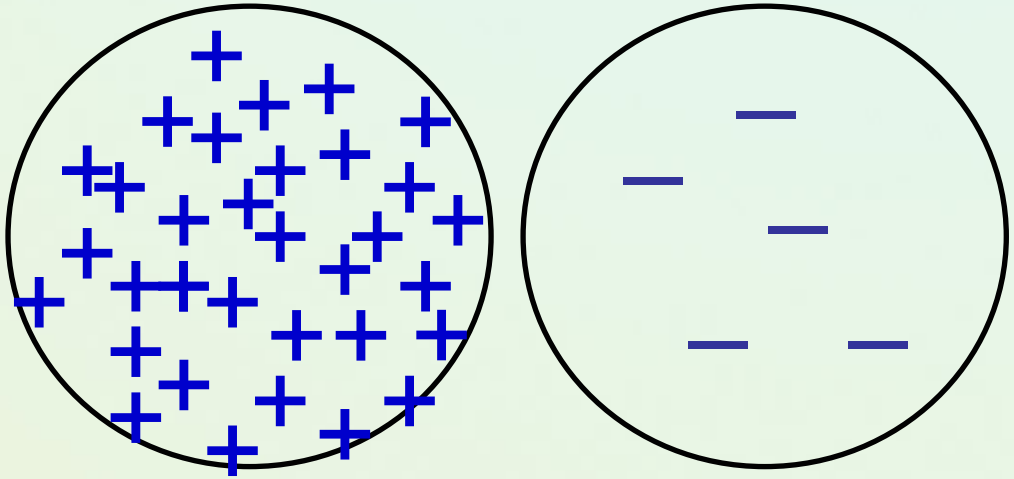


Przykład danych medycznych Chawla et al. SMOTE 2002



Niezrównoważenie rozkładu w klasach

- Dane są niezrównoważone 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



Class imbalance is not the same
s COST sensitive learning.
n general cost are unknown!"

Przykłady niezrównoważonych problemów

- Niezrównoważenie jest 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.

W czym tkwi trudność ?

- 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

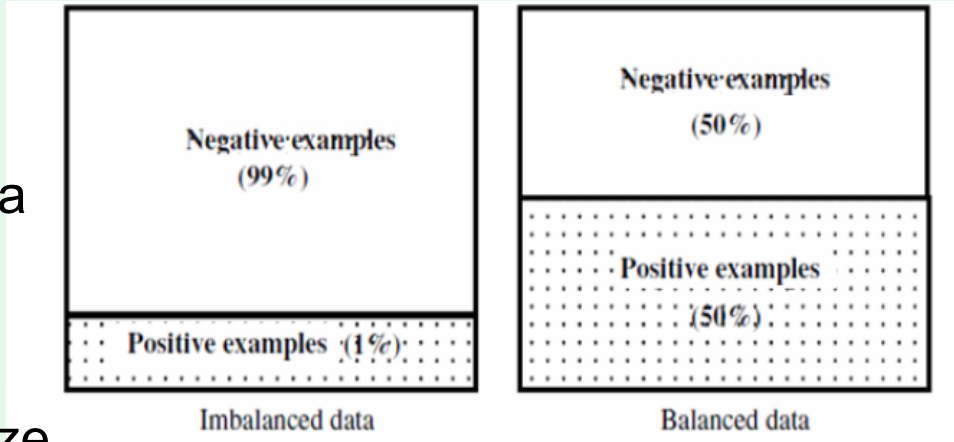


Fig. 1. Imbalanced and balanced data sets.

biased towards the majority class

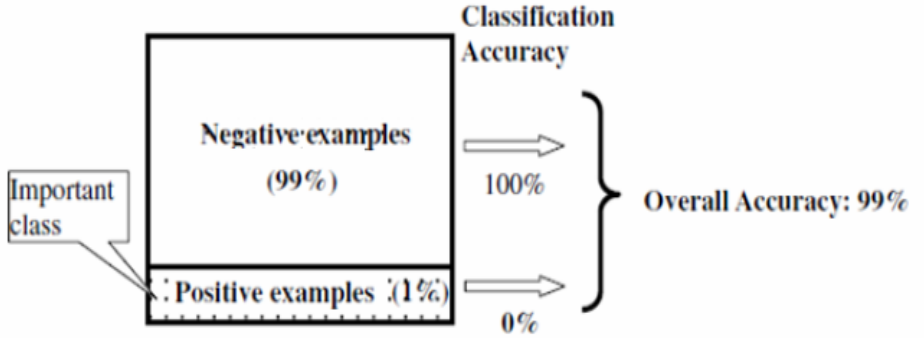


Fig. 2. The illustration of class imbalance problems.

Przykład algorytmu indukcji reguł

- MODLEM [Stefanowski 98] → minimalny zbiór reguł.
- Ocena warunków elementarnych → entropia.
- Niespójne dane → „pruning” reguł albo przybliżenia klas decyzyjnych → indukcja pewnych i możliwych reguł.
- Przetwarzanie atrybutów nominalnych i liczbowych.
- W połączeniu ze strategiami klasyfikacyjnymi → skuteczny klasyfikator
  - Dopasowanie opisu obiektu do części warunkowych.
  - Niejednoznaczności → decyzja większościowa

obj.	*a*1	*a*2	*a*3	*a*4	*D*
*x*1	m	2.0	1	a	C1
*x*2	f	2.5	1	b	C2
*x*3	m	1.5	3	c	C1
*x*4	f	2.3	2	c	C1
*x*5	f	1.4	2	a	C2
*x*6	m	3.2	2	c	C2
*x*7	m	1.9	2	b	C1
*x*8	f	2.0	3	a	C2

if (*a*1 = m) ***and*** (*a*2 ≤ 2.6) ***then*** (*D* = C1) {*x*1,*x*3,*x*7}

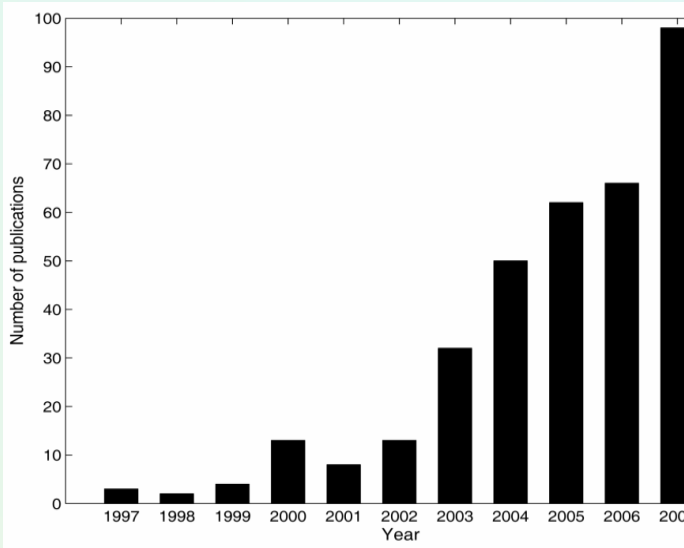
if (*a*2 ∈ [1.45, 2.4]) ***and*** (*a*3 ≤ 2) ***then*** (*D* = C1) {*x*1,*x*4,*x*7}

if (*a*2 ≥ 2.4) ***then*** (*D* = C2) {*x*2,*x*6}

if (*a*1 = f) ***and*** (*a*2 ≤ 2.15) ***then*** (*D* = C2) {*x*5,*x*8}

Zainteresowanie środowiska ML i DM

- Początkowo problem znany w zastowaniach
- Od 10lat wzrost zaintersowania badawczego
  - AAAI'2000 Workshop, org:R. Holte, N. Japkowicz, C. Ling, S. Matwin.
  - ICML'2000 Worshop also on cost sensitive. Dietterich T. et al.
  - ICML'2003 Workshop, org.: N. Chawla, N. Japkowicz, A. Kolcz.
  - ECAI 2004 Workshop, org.: Ferri C., Flach P., Orallo J. Lachice. N.
  - ICMLA 2007 – Session on Learning from Imbalanced Data
  - PAKD 2009 – Workshop on Imbalanced Data
  - RSCTC 2010 – Special Sessions CDMC
- Special issues:
  - ACM KDDSIGMOD Explorations Newsletter, editors: N. Chawla, N. Japkowicz, A. Kolcz.



Miary oceny

- Jak oceniać klasyfikatory
  - Standardowa trafność bezużyteczna
    - Wyszukiwanie informacji (klasa mniejszościowa ~ 1%) → ogólna trafność klasyfikowania ~100%, lecz źle rozpoznawana wybrana klasa
- Miary powinny być z klasą mniejszościową
  - Analiza binarnej macierzy pomyłek confusion matrix
  - Sensitivity i specificity,
  - ROC curve analysis.

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

Klasyczne miary:

Error Rate: (FP + FN)/N

Accuracy Rate: (TP + TN) /

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

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

Miary oceny wynikające z macierzy pomyłek

- G-mean

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

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

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

		True class: p	True class: n
Hypothesis output	Y	TP (True Positives)	FP (False Positives)
	N	FN (False Negatives)	TN (True Negatives)
Column counts:		P_C	N_C

- F-miara

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

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

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

Probabilistyczne podstawy ROC

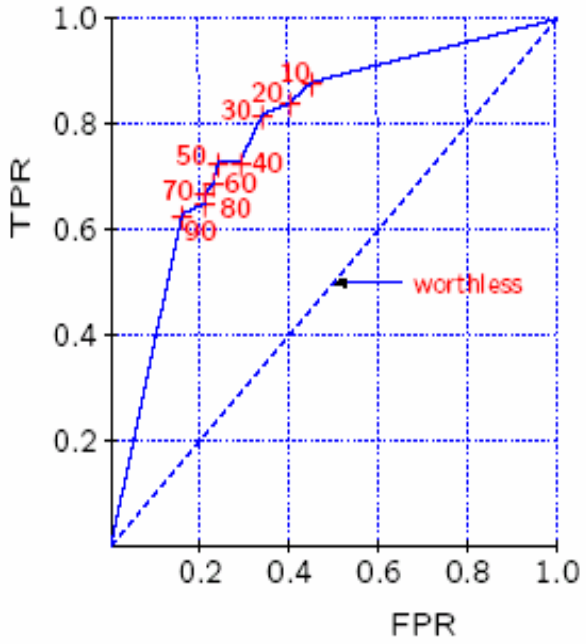
- Output of probabilistic classifier:

$$c_{max} = \arg\max_{C} P(C \mid \mathcal{E})$$

may not yield the best performance

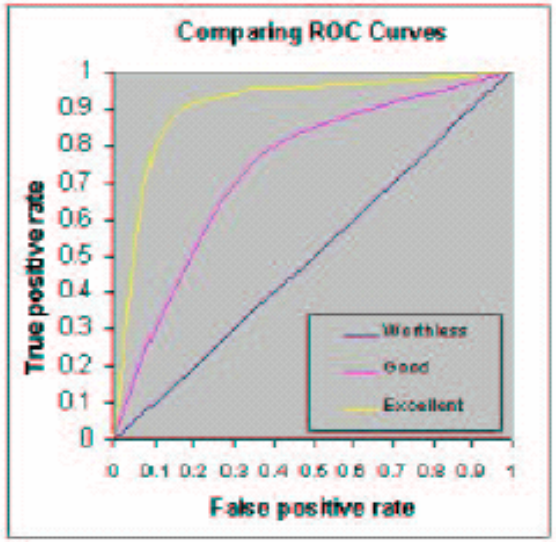
- Alternative: Receiver Operating Characteristic (ROC): determine threshold d, such that

$$C = \begin{cases} c & \text{if } P(c \mid \mathcal{E}) \geq d \\ \neg c & \text{otherwise} \end{cases}$$

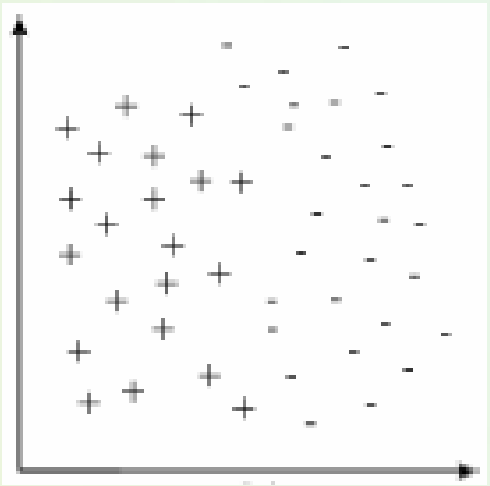


When comparing various techniques:

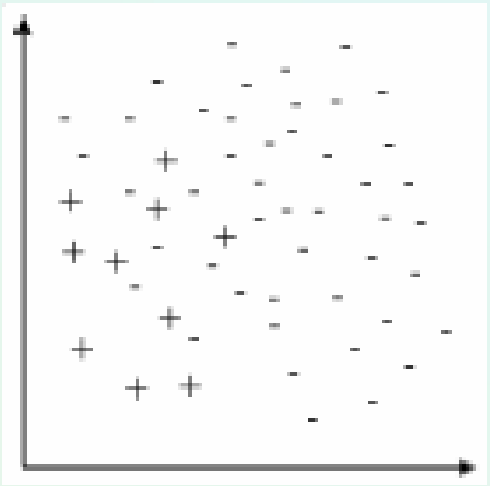
- actual performance for particular thresholds (cut-off points) may vary
- area under the ROC curve $A_f = \int_0^1 f(x)\mathrm{d}x$ offers good measure for comparison, with f relationship between FPR and TPR for classifier



Na czym polega trudność?



Łatwiejszy problem



Trudniejszy

Źródła trudności:

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

Klasa większ. „nakłada” się na mniejszościowe

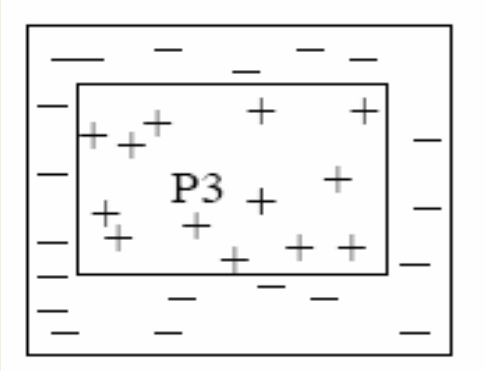
- ❑ Niejednoznaczne przykłady brzegowe
- ❑ Wpływ „szumu” (noisy examples)

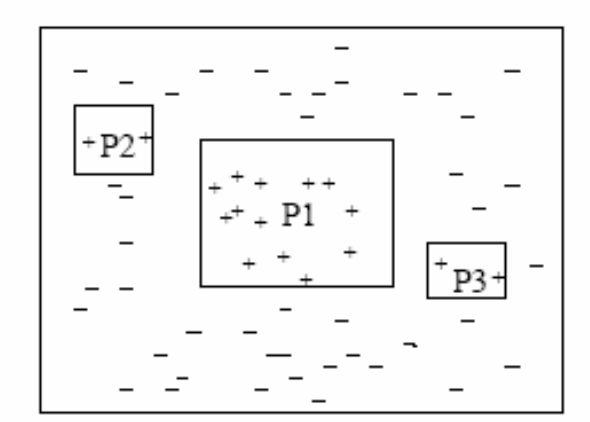
Przeglądowe prace:

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

Czy zawsze „niezrównoważenie” jest trudnością?

- Przeanalizuj studia eksperymentalne N.Japkowicz lub przeglądy G.Weiss.
- Japkowicz „The minority class contains small „disjuncts” – sub-clusters of interesting examples surrounded by other examples”





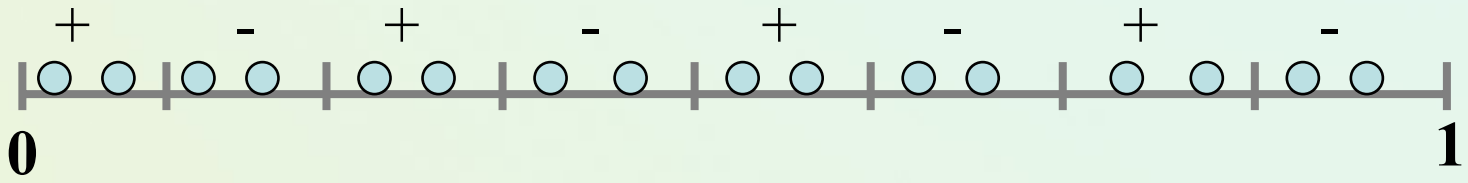
Niektóre prace eksperymentalne z dysuksją źródeł trudności, e.g:

- 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 p. Japkowicz i współpracowników

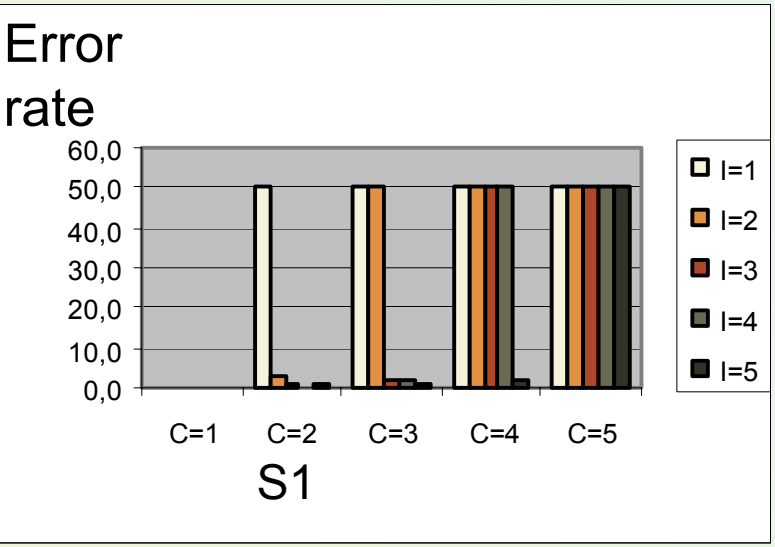
T. Jo, N. Japkowicz. Class imbalances versus small disjuncts.

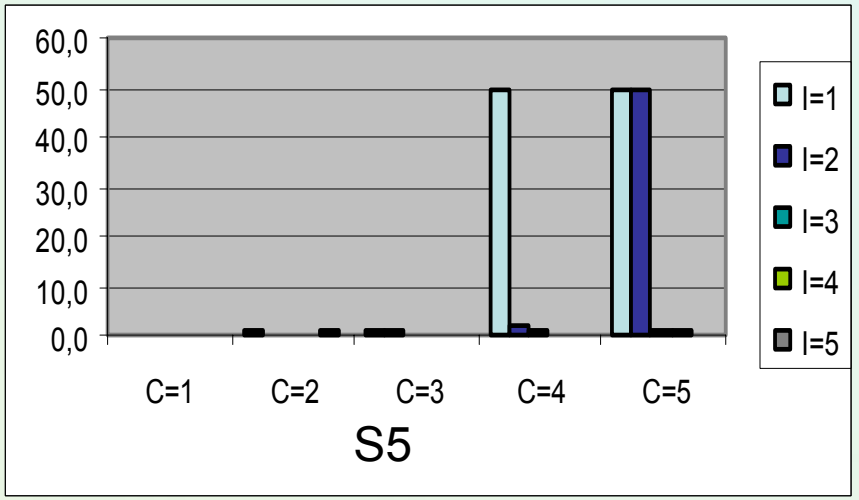
- Przygotowanie sztucznych zbiorów w których zmienia się:
  - Poziom niezrównoważenia (I)
  - Liczność zbioru uczącego (S)
  - Concept complexity (C) – dekompozycje klas na podobszary
- Algorytmy drzewa C4.5, sztuczne sieci neuronowe BP i SWM



Rezultaty eksperymentów

- 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.
  - Wybrany wynik dla C4.5





Large
Imba
Full
balan

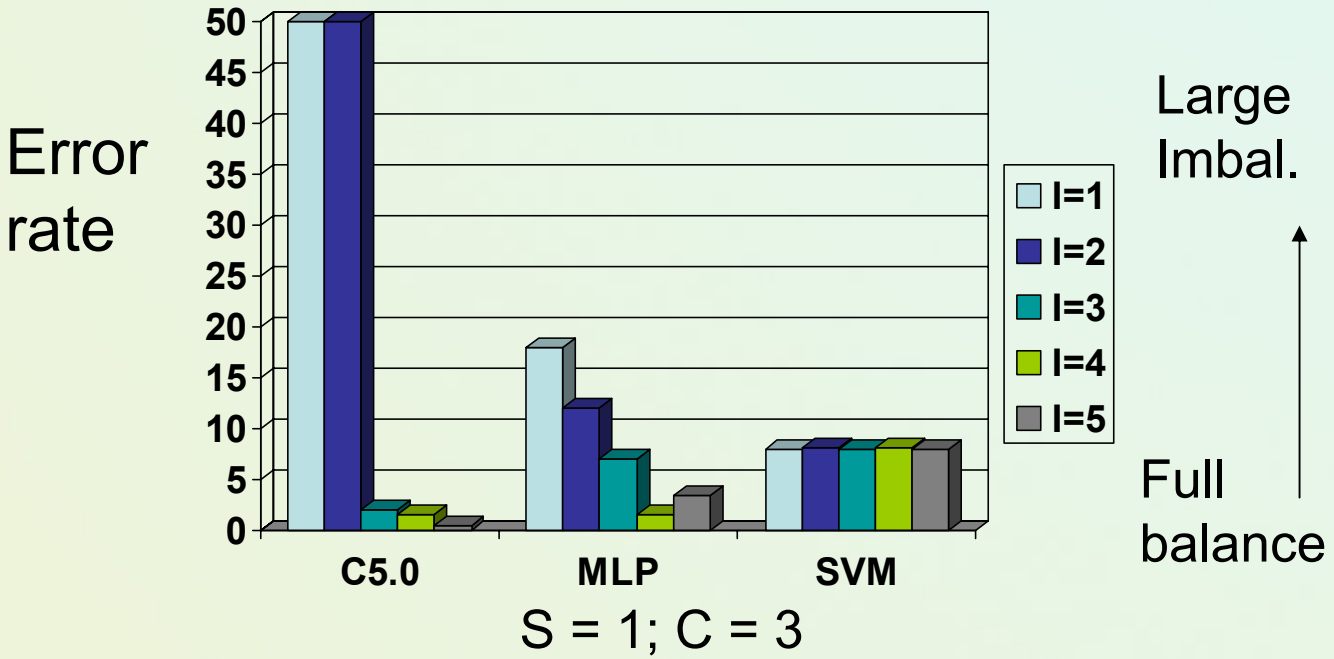
Skuteczność klasyfikacji zależy od:

the **degree** of class **imbalance**;
the **complexity of the concept** represented by the data;
the overall **size of the training set**;
the **classifier** involved.

Konkluzje co do wyboru algorytmów

Niektóre algorytmy mniej podatne na dane niezrównoważone:

- Naive Bayes, Support Vector Machines
- vs. drzewa decyzyjne, reguły decyzyjne, sieci neuronowe, k-NN



Eksperymenty dokładniej opisane przez Japkowicz, Jo

ksperymenty Garcia et al. ze strefami brzegowymi

Overlapping

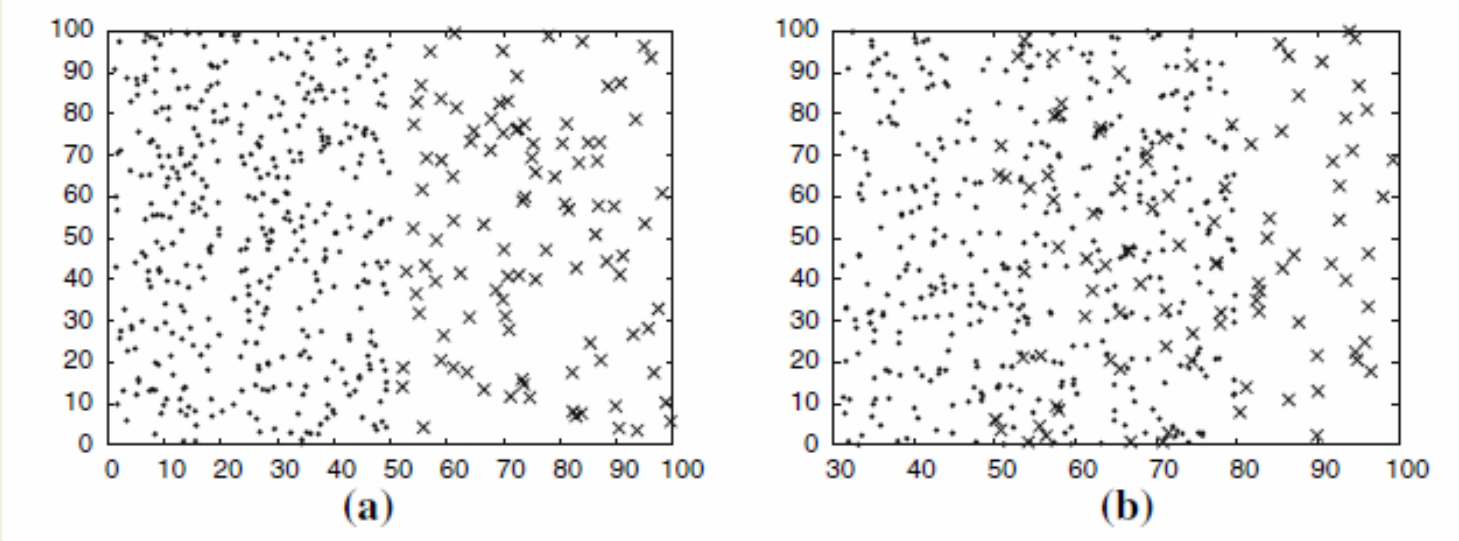


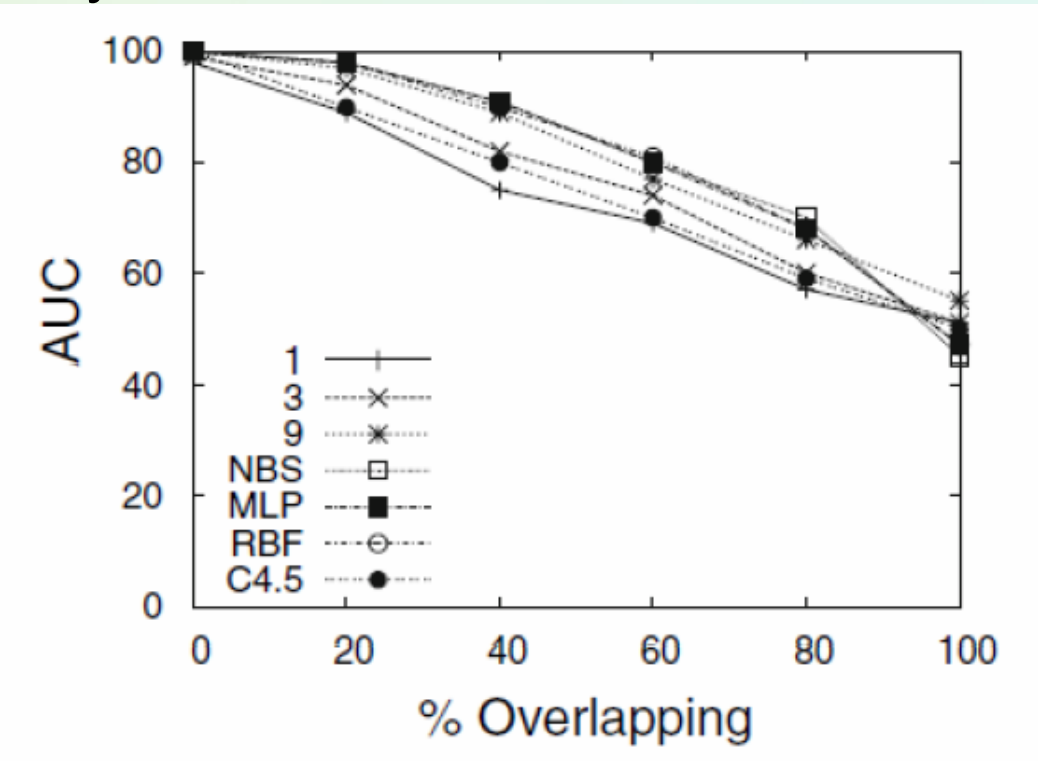
Fig. Two different levels of class overlapping: a 0% and b 60%

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.

ksperymenty Garcia et al. ze strefami brzegowymi

Niektóre z wyników

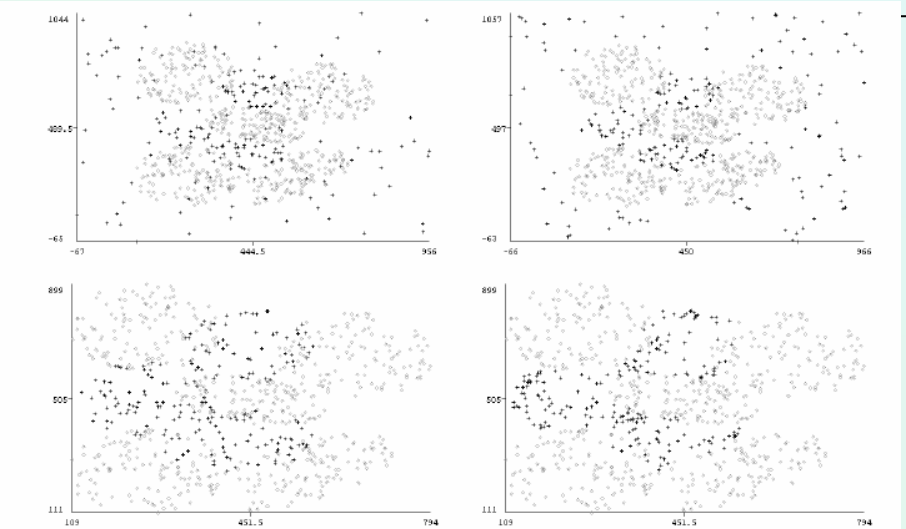


Skuteczność różnych klasyfikatorów – wzrost niejednoznaczność strefy brzegowe
silniej obniża AUC niż wzrost niezrównoważenia

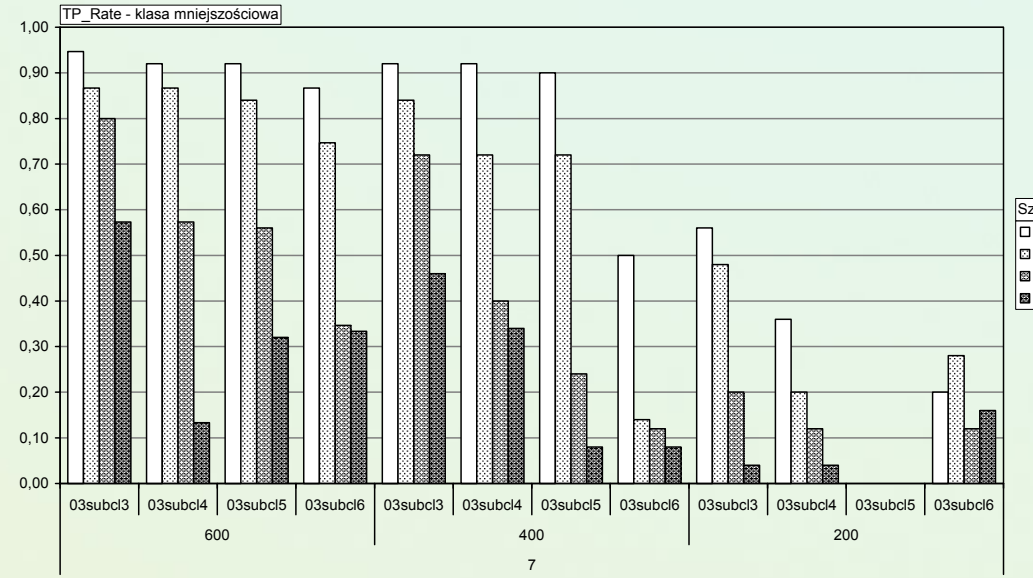
W dalszych eksperymentach zauważony wpływ lokalnej gęstości przykładów!

Dalsze eksperymenty ze sztucznymi danymi

- J. Stefanowski, 2009 (przy współpracy K.Kałużny)
- Wpływ różnego typu trudności (kształt pojęc, „szum” i rzadkie przykłady, nakładające się obszary brzegowe, stopnie niezrównoważenia, 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 niezrównoważenie klas (rozszerzanie rezultatów znanych z literatury)
  - 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



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

- Problem badawczy
  - Dokładniejsze zbadanie wpływu różnego typu zaburzeń (przypadki brzegowych i zaszumione, dekompozycja klasy mniejszościowej) na efektywność metod wstępnego przetwarzania danych niezrównoważonych
- Proponowane podejście
  - Eksperymentalne porównanie algorytmów SPIDER, NCR, *cluster-oversampling* oraz prostego nadlosowywania klasy mniejszościowej
- Dane
  - Sztuczne zbiory danych o określonej specyfice (niedoskonałości różnego typu)
  - Także niezrównoważone zbiory danych pochodzące z repozytorium UCI
- Konkluzje
  - Proste nadlosowywanie okazało się najgorszą z rozważanych metod
  - W przypadku małej ilości trudnych przypadków *cluter-oversampling* okazał się najskuteczniejszy
  - W przypadku silnie zaburzonych danych (np. szum powyżej 30%) lepsze okazały się metody SPIDER oraz częściowo NCR

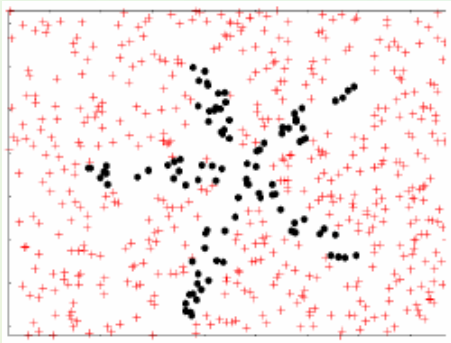


Fig. 1. Clover data set

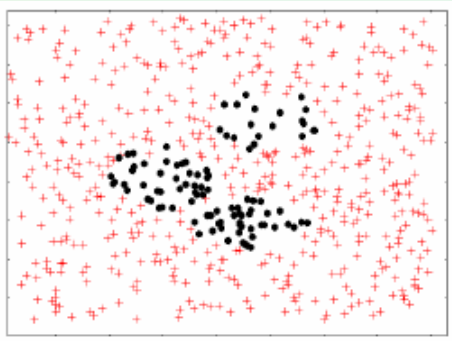


Fig. 2. Paw data set

Wybrane sztuczne dane i wyniki

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

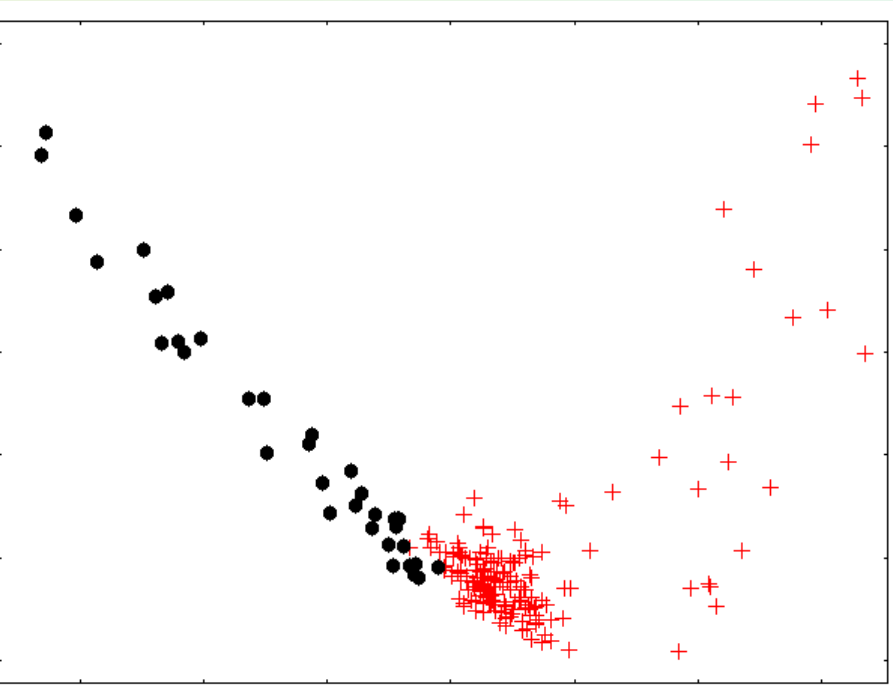
Algorytm: C4.5

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

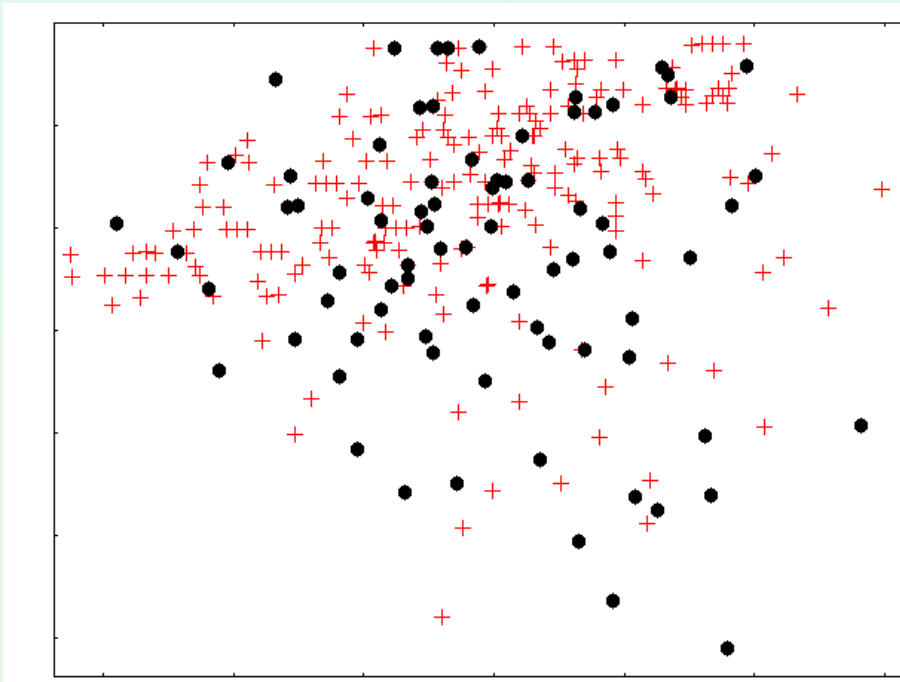
Data set	MODLEM					C4.5				
	Base	RO	CO	NCR	SP2	Base	RO	CO	NCR	SP2
subcl-safe	0.5800	0.5800	0.6200	0.7800	0.6400	0.3200	0.8400	0.8600	0.9800	1.0000
subcl-B	0.8400	0.8400	0.8400	0.8600	0.8400	0.0000	0.8200	0.8400	0.3600	0.9200
subcl-C	0.1200	0.1000	0.1600	0.2400	0.2600	0.0000	0.5400	0.0000	0.0000	0.5200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

Co z rzeczywistymi danymi?

- Wizualizacja 2 składowych w metodzie PCA

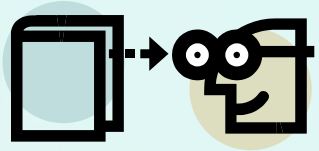


Thyroid



Haberman

Podstawowe metody



- Prace przeglądowe
  - 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.
- 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 boosting)
  - Specjalizowane systemy hybrydowe
  - One-class-learning
  - Transformacje do zadania „cost-sensitive learning”
  - …

Modyfikacje algorytmów regułowych

Zmiany w fazie poszukiwania reguł albo w strategiach klasyfikacyjnych

- Schemat tradycyjny:
  - Exploit a greedy search strategy and use criteria that favor the majority class.
    - The majority class rules are more general and cover more examples (strength) than minority class rules.
- Różne propozycje zmian:
  - Use another inductive bias
    - Modification of CN2 to prevent small disjuncts (Holte et al.)
    - Hybrid approach with different „inductive bias” between large and small sets of examples (Ting).
  - Use less greedy search for rules
    - Exhaustive depth-bounded search for accurate conjunctions. Brute (Riddle et al..), modification of Apriori like algorithm to handle multiple levels of support (Liu at al.)
    - Specific genetic search – more powerful global search (Freitas and Lavington, Weiss et al.) …

Przykład modyfikacji

Nowy Algorytm uczący *J.Stefanowski, S.Wilk: KES 2004*

- Połączenie dwóch zbiorów reguł – tzw. satysfakcjonującego (dla klasy mniejszościowej) oraz minimalnego (dla klasy większościowej)
- Klasa większościowa – algorytm LEM2:
  - Przeszukiwanie zachłanne
  - Minimalny zbiór reguł
- Klasa mniejszościowa – algorytm Explore (Stefanowski, Vanderpooten 1994,2001):
  - Wszystkie reguły o wsparciu > minsup
  - Więcej reguł o sumarycznie większym wsparciu
- Przedstawione podejście nie wymagało strojenia parametrów – progi minimalnego wsparcia dla reguł satysfakcjonujących były dobierane automatycznie

Alternatywne podejście $\rightarrow$ zmiana strategii klasyfikacyjnej

(J.W. Grzymała-Busse et al. 2000)

- Współczynnik zwiększający siłę reguł mniejszościowych
- Optymalizacja *gain* = *sensitivity* + *specificity*

3 Replacing rules for the minority class

procedure replace_rules (**input** K_{min}: the minority class;
R : initial minimal set of rules;
L : learning examples; T : validation examples;
output R^{final}: resulting set of rules)
begin
    $min\_sup \leftarrow$ minimum coverage in R for K_{min}
    $max\_sup \leftarrow$ maximum coverage in R for K_{min}
    $R_{maj} \leftarrow$ rules from R pointing at the majority classes
    $R_{min}^{min\_sup} \leftarrow$ use EXPLORE to induce rules from L for K_{min}
        with minimum required coverage set to min_sup
  for $sup = min\_sup$ **to** max_sup **do**
  begin
    $R_{min}^{sup} \leftarrow$ select these rules from $R_{min}^{min\_sup}$ for which coverage $\geq sup$
    $R^{sup} \leftarrow R_{min}^{sup} \cup R_{maj}$
    $gain \leftarrow$ evaluate R^{sup} on T
    memorize $gain$ and R^{sup}
  end
  $R^{final} \leftarrow R^{sup}$ corresponding to the best observed $gain$
end

Table 2 Characteristics of data sets used for experiments (N – number of examples, N_{Pos} – number of examples in the minority class, N_{Orh} – number of examples in the majority classes, $R_{Pos} = N_{Pos}/N$ – ratio of examples in the minority class)

Data set	N	N_{Pos}	N_{Orh}	R_{Pos}
Abdominal Pain	723	202	521	27.9%
Breast Slovenia	294	89	205	30.3%
Breast Wisconsin	625	112	513	17.9%
Bupa	345	145	200	42.0%
German	666	209	457	31.4%
Hepatitis	155	32	123	20.6%
Pima	768	268	500	34.9%
Scrotal Pain	201	59	142	29.4%
Urology	498	155	343	31.1%

Table 4 Best results of increasing rule support by multipliers (*mult* – support multiplier, *sens* – sensitivity, *spec* – specificity, *GM* – G-mean, *acc* – overall accuracy)

Data set	*Mult*	*Sens*	*Spec*	*GM*	*Acc*
Abdominal Pain	5	80.69	84.84	82.74	83.68
Breast Slovenia	1	36.47	88.56	56.83	73,08
Breast Wisconsin	5	57.14	86.74	70.41	81.44
Bupa	3	55.86	58.50	57.17	57.39
German	4	57.89	64.11	60.92	62.16
Hepatitis	18	84.38	77.24	80.73	78.71
Pima	3.5	59.33	76.40	67.32	70.44
Scrotal Pain	3	67.80	80.99	74.10	77.11
Urology	14	51.92	49.42	50.65	50.52

Table 5 Results for the *Replacing Rules* approach (*SC* – coverage threshold, *sens* – sensitivity, *spec* – specificity, *GM* – G-mean, *acc* – overall accuracy, N_R – numer of rules)

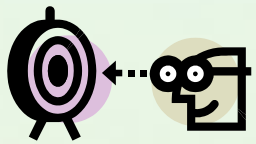
Data set	*SC*	*Sens*	*Spec*	*GM*	*Acc*	N_R
Abdominal Pain	8.0	83.14	83.68	83.41	83.54	88.0
Breast Slovenia	3.0	47.09	84.11	62.93	73.08	37.0
Breast Wisconsin	2.0	63.85	81.60	72.18	78.57	158.5
Bupa	2.0	42.75	63.00	51.90	54.50	61.5
German	5.0	62.71	72.65	67.50	69.50	73.5
Hepatitis	4.0	75.30	81.56	78.37	80.02	76.5
Pima	2.0	68.78	67.89	68.33	68.10	341.5
Scrotal Pain	4.0	68.87	87.24	77.51	81.56	12.5
Urology	4.0	71.73	43.20	55.67	51.61	691.5

Niektóre z wyników eksperymentów (sensitivity)

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

Zdecydowana poprawa miary wrażliwości; także lepsze wartości G-means

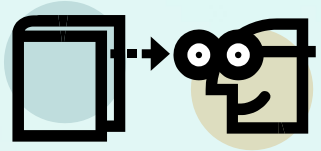
Więcej → J. Stefanowski, Sz. Wilk: *Extending rule-based classifiers to improve recognition of imbalanced classes, w Z. Ras (eds) Advanced in Data Managment, Springer 2009*



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:

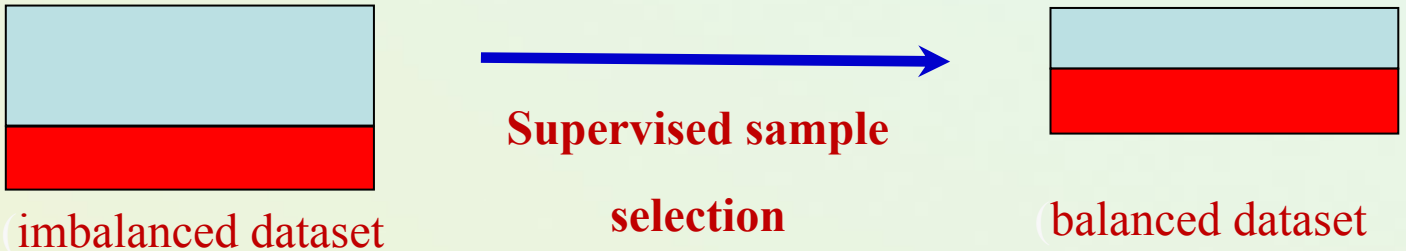
- Proste techniki losowe
  - „Over-sampling” – klasa mniejszościowe
  - „Under-sampling”- klasa mniejszościowa
- Specjalizowane nadlosowanie
  - Cluster-oversampling (Japkowicz)
- Ukierunkowane transformacje
  - Klasa większościowe
    - One-side-sampling (Kubat, Matwin) z Tomek Links
    - Laurikkala’s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

esampling - modyfikacja zbioru uczącego przed budową lasyfikatora

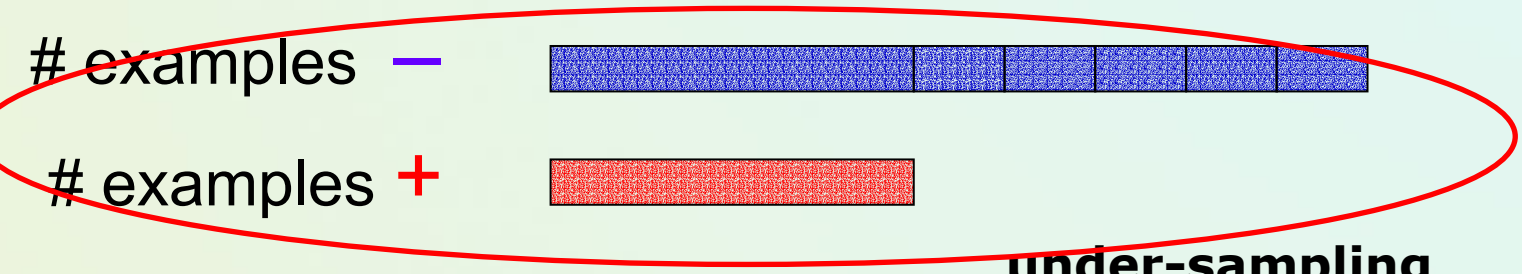
Resampling → pre-processing; celowa zmiana rozkładu przykładów; „balansowanie” liczności klas po to aby w kolejnej fazie móc lepiej nauczyć klasyfikator

Brak teoretycznej gwarancji znalezienia optymalnej postaci rozkładu

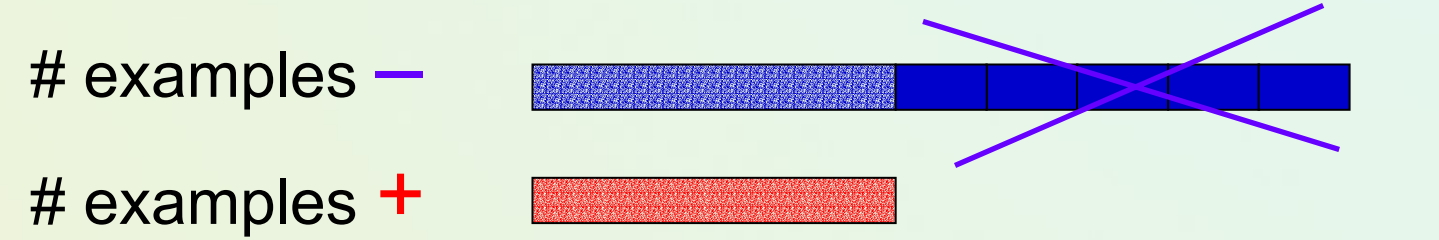
Raczej heurystyka ukierunkowana na “to add or remove examples with the hope of reaching better distribution of the training examples and thus, realizing the potential ability of classifiers” [F.Herrera 2010].



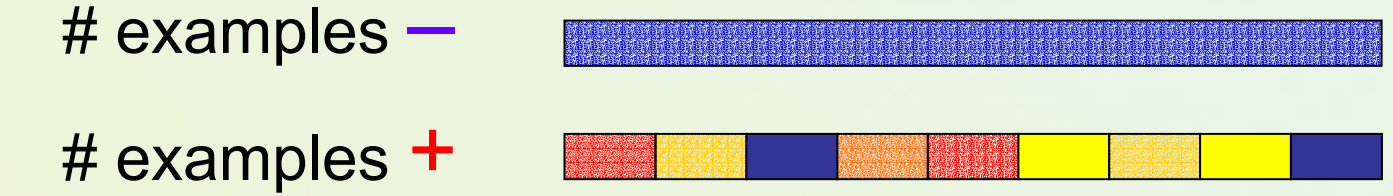
Undersampling vs oversampling



under-sampling

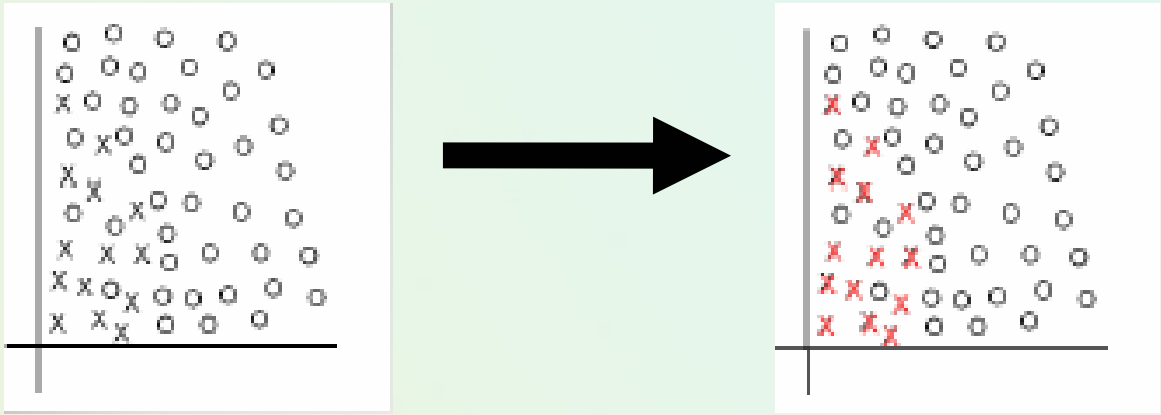


over-sampling

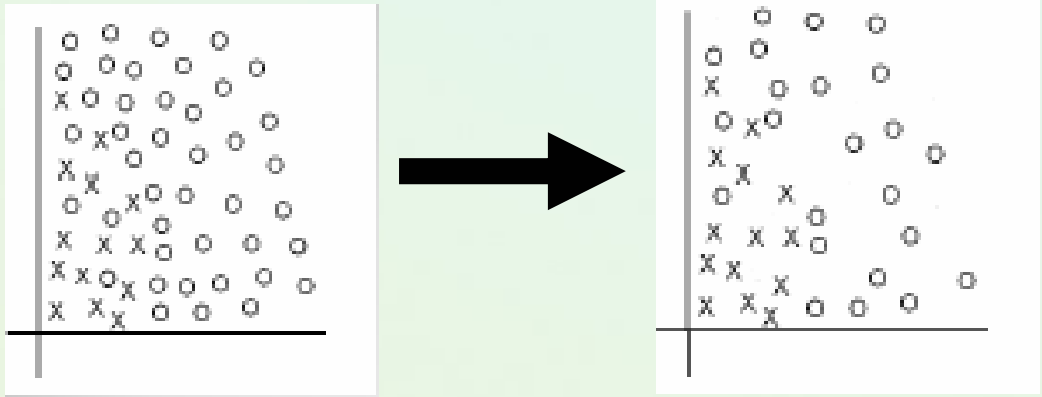


Losowe równoważenie klas - dyskusja

Random oversampling → Kopiowanie przykładów mniejszościowych



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



- Utrata informacji

luster oversampling – specyficzne losowanie dla „small isjuncts”

Dekompozycja klas → within and between -class imbalance

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

Użyj algorytmu analizy skupień wewnątrz każdej klasy i odpowiednio adlosowuj!

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.

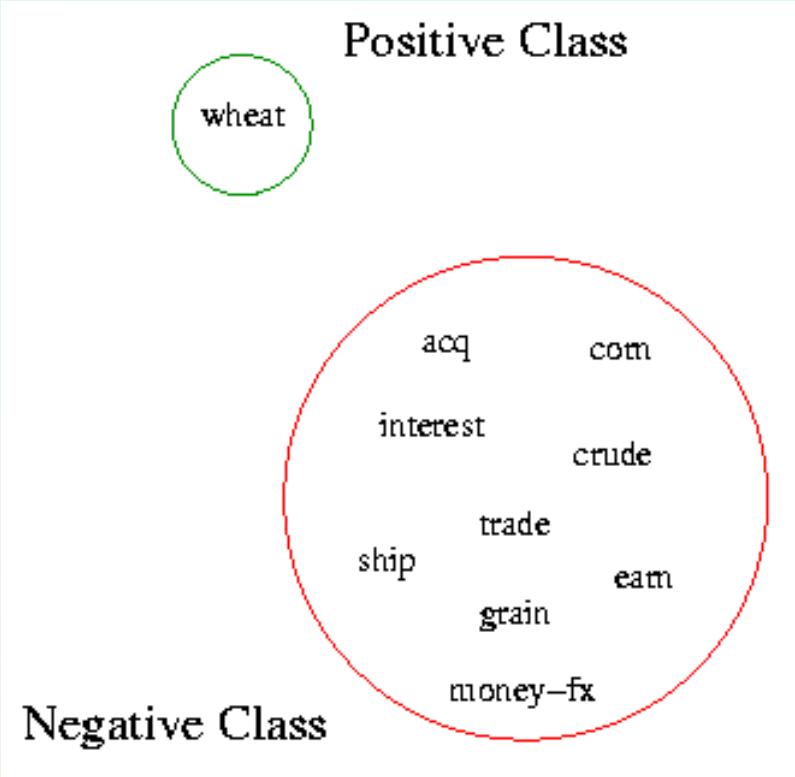
Cluster-based resampling identifies rare regions and re-samples them individually, so as to avoid the creation of small disjuncts in the learned hypothesis.

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

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

Prace Japkowicz i współpracownicy 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

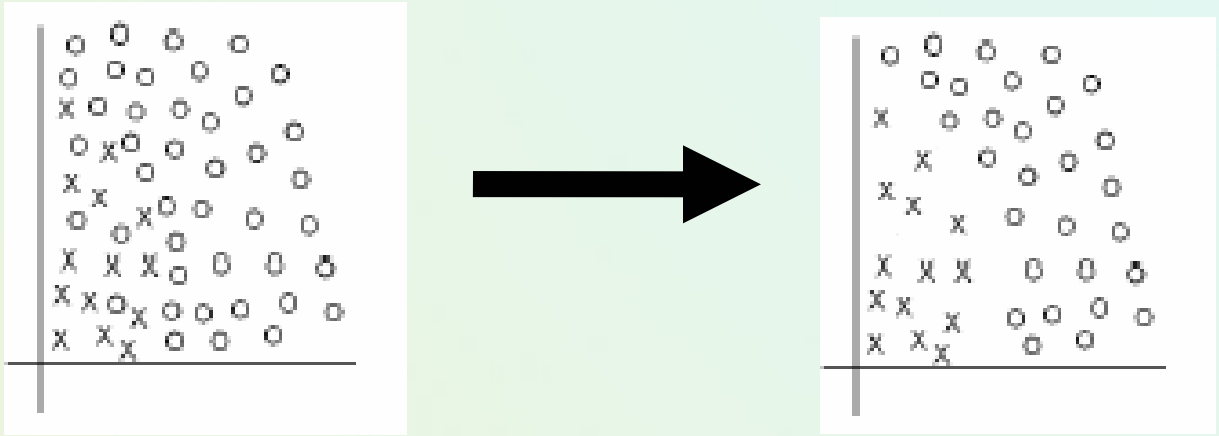


II: Within-class vs Between- class Imbalances: Text Classification

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 OversamplingII (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”?

Powróćmy do charakterystyki przykładów

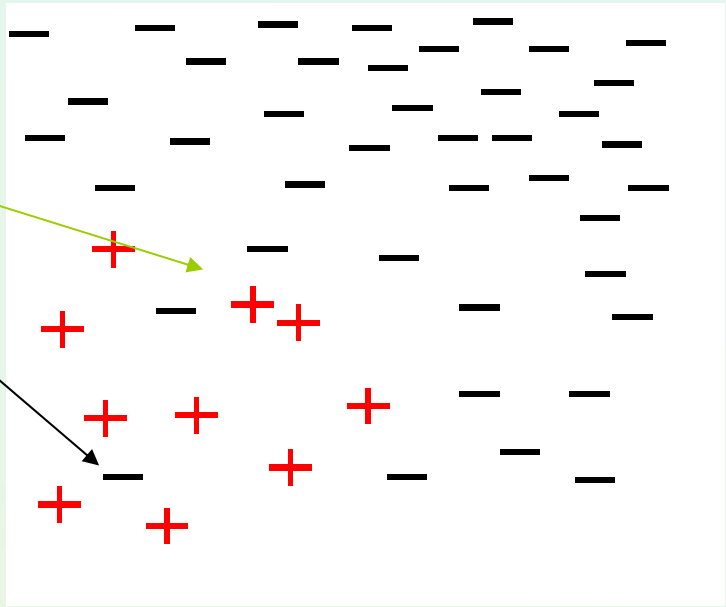
Typy przykładów → techniki „resampling” powinny skupić swoje działanie na niektórych z nich

ztery typy przykłady z klasy (większościowej)

- Noise examples
- Borderline examples

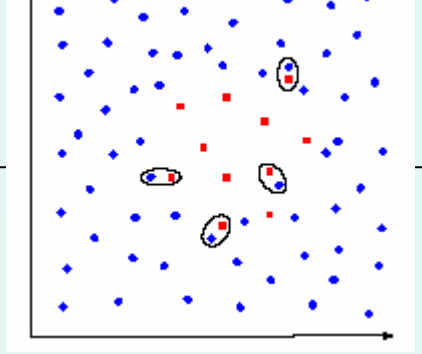
  Borderline examples are unsafe since a small amount of noise can make them fall on the wrong side of the decision border.
- Redundant examples

- Safe examples



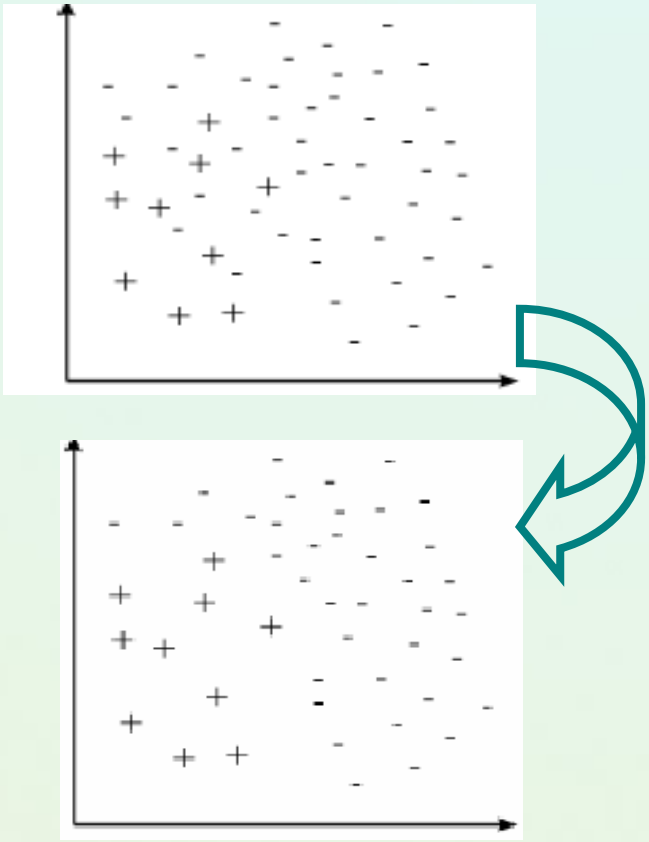
Modyfikacje undersampling: Znajdź i usuń 2 lub 3 pierwsze typy przykładów

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

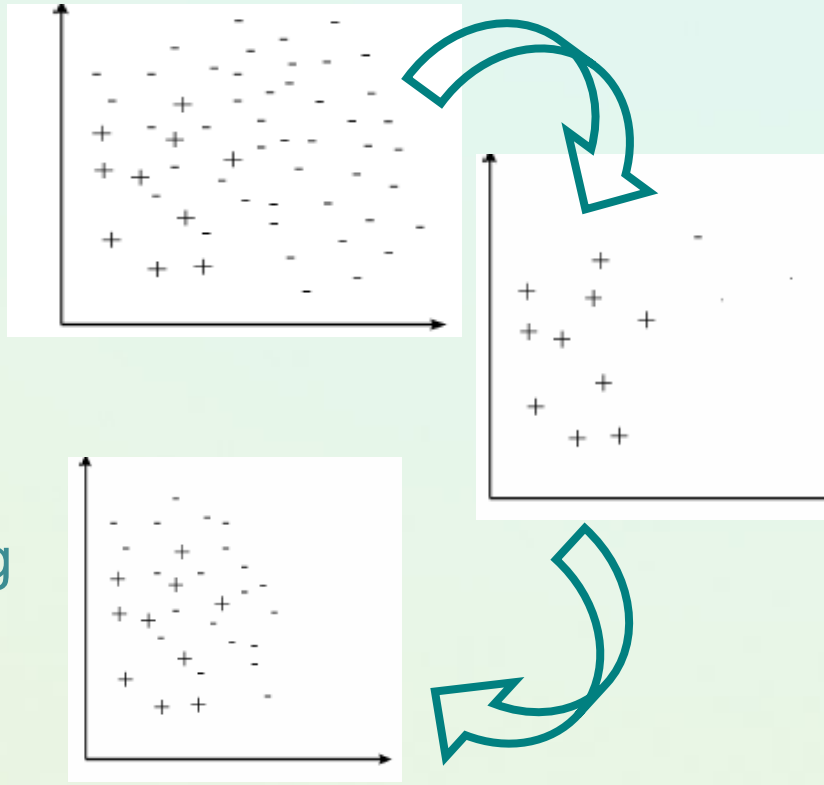
CNN – Condensed Nearest Neighbours specjalizowana odmiana edytowanego K-NN

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 przykładów granicznych i szum

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'



nformed Under-sampling

Najbardziej znany OSS → 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

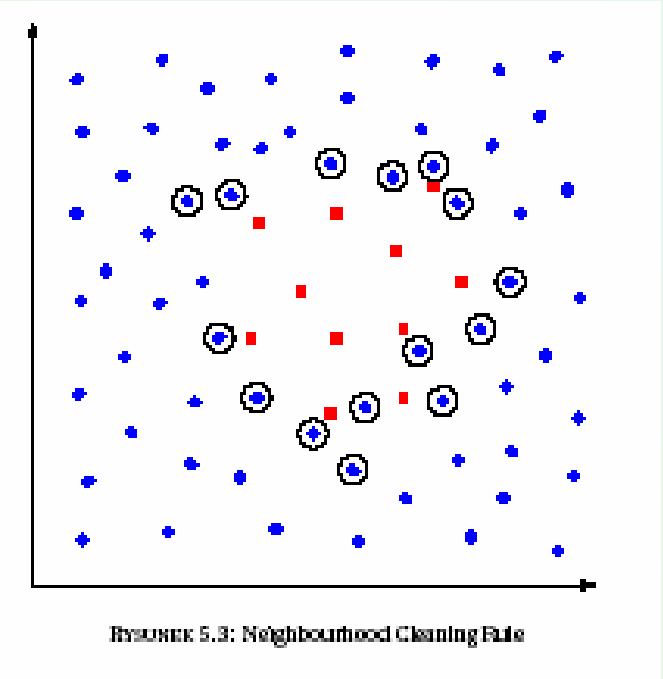
- **NCL** Nearest Cleaning Rule - Jorma Laurikkala 2001,

  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

Nearest Cleaning Rule

- Przykład ilustracyjny



Rysunek 5.3: Neighbourhood Cleaning Rule

Krytyczna dyskusja → podejście hybrydowe

- NCR and one-side-sampling
  - Zachłannie usuwają (zbyt) wiele przykładów z klasy większościowej
  - Ukierunkowane głównie na poprawie wrażliwości (sensitivity)
  - Mogą prowadzić do zbyt gwałtownego pogorszenia rozpoznawania klas większościowych (specificity, i inne miary)
- 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, r – 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
  - Stopień wzmocnienia zależny od analizy sąsiedztwa (ENN)

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

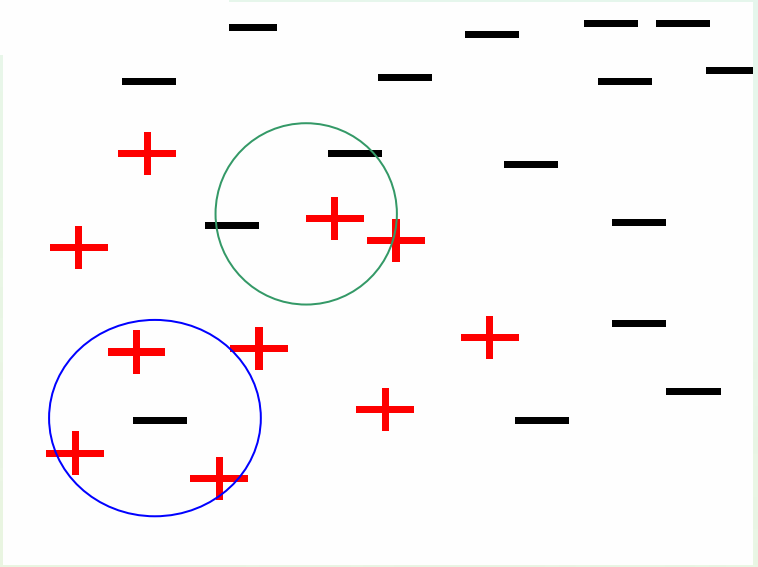
- Wykorzystanie modyfikacji zasady k- najbliższych sąsiadów (Wilson's Edited Nearest Neighbor Rule)
  → Usuń te przykłady, których klasa różni się od trzech najbliższych przykładów.
- Odległość HDVM z wykorzystaniem miary VDM Cost, Salzberg

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

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.



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

- Studium eksperymentalne
  - Porównanie z prostym losowym balansowaniem, NCR i podstawową wersją SMOTE
- Dane
  - Niezrównoważone rzeczywiste zbiory danych pochodzące z repozytorium UCI oraz własnych projektów medycznych
- Konkluzje
  - SPIDER poprawił wrażliwość klasyfikatorów opartych na drzewach i regułach decyzyjnych (np. lepszy niż SMOTE), spadki specyficzności były mniejsze niż w przypadku algorytmu NCR i losowego balansowania klas
  - SPIDER w znacznie mniejszym stopniu zmodyfikował rozkład przetwarzanych zbiorów danych, niż algorytm SMOTE

Table 2. Sensitivity

	MODLEM						C4.5					
Data set	Base	SMOTE	NCR	Weak	Relabel	Strong	Base	SMOTE	NCR	Weak	Relabel	Strong
Acl	0.805	*0.850*	*0.900*	0.830	0.835	0.826	*0.855*	0.840	*0.920*	0.835	0.835	0.860
Breast can.	0.319	0.468	*0.638*	0.437	*0.554*	0.539	0.387	0.463	*0.648*	0.500	*0.576*	0.531
Bupa	0.520	0.737	*0.873*	0.799	*0.838*	0.806	0.491	0.662	*0.755*	0.710	*0.720*	0.700
Cleveland	0.085	*0.245*	*0.343*	0.233	*0.245*	0.236	0.237	0.260	*0.398*	0.343	*0.395*	0.302
Ecoli	0.400	0.632	*0.883*	0.605	*0.643*	0.637	0.580	*0.730*	*0.758*	0.688	0.687	0.690
Haberman	0.240	0.301	*0.626*	0.404	0.468	*0.483*	0.410	0.572	0.608	0.657	*0.694*	*0.660*
Hepatitis	0.383	0.382	*0.456*	0.385	*0.438*	0.437	0.432	0.537	*0.622*	0.513	*0.580*	0.475
New-thyr.	0.812	*0.917*	0.842	0.860	*0.877*	0.865	*0.922*	0.896	0.873	0.897	0.897	*0.913*
Pima	0.485	0.640	*0.793*	0.685	*0.738*	*0.738*	0.601	0.739	*0.788*	0.718	*0.751*	0.715

Table 3. Specificity

	MODLEM						C4.5					
Data set	Base	SMOTE	NCR	Weak	Relabel	Strong	Base	SMOTE	NCR	Weak	Relabel	Strong
Acl	*0.942*	0.914	0.890	*0.934*	0.922	0.930	*0.940*	0.922	0.898	*0.924*	0.908	0.918
Breast can.	*0.804*	0.657	0.523	*0.710*	0.621	0.606	*0.767*	*0.676*	0.525	0.630	0.609	0.614
Bupa	*0.820*	*0.668*	0.308	0.453	0.473	0.459	*0.775*	*0.611*	0.415	0.534	0.459	0.532
Cleveland	*0.957*	0.887	0.884	*0.934*	0.919	0.927	*0.899*	0.870	0.849	0.877	0.864	*0.887*
Ecoli	*0.989*	0.951	0.924	0.958	0.953	*0.962*	*0.959*	0.921	0.920	0.931	0.916	*0.941*
Haberman	*0.816*	*0.782*	0.658	0.746	0.720	0.713	*0.805*	*0.747*	0.698	0.597	0.565	0.591
Hepatitis	*0.933*	*0.927*	0.894	0.918	0.907	0.908	*0.873*	*0.851*	0.823	0.822	0.807	0.803
New-thyr.	*0.987*	0.986	0.984	*0.990*	*0.990*	0.984	0.973	*0.984*	0.974	0.971	0.972	*0.976*
Pima	*0.856*	*0.778*	0.658	0.774	0.720	0.698	*0.814*	*0.716*	0.656	0.681	0.667	0.687

Table 5. Changes in the class distribution (N_C – the number of examples in the minority class, N_O – the number of examples in the majority class N_O, N_R – the number of relabeled examples, N_A – the number of amplified examples)

	SMOTE		NCR		Weak		Relabel				Strong	
Data set	N_C	N_O	N_C	N_O	N_C	N_O	N_C	N_O	N_R	N_A	N_C	N_O
Acl	120	100	40	83	57	98	59	98	2	17	67	98
Breast cancer	255	201	85	101	173	167	197	167	24	88	253	167
Bupa	290	200	145	81	236	145	271	145	35	91	309	145
Cleveland	245	268	35	198	102	255	110	255	8	67	147	255
Ecoli	210	301	35	266	58	288	69	288	11	23	77	288
Haberman	162	225	81	121	162	182	193	182	31	81	223	182
Hepatitis	64	123	32	90	61	113	68	113	7	29	88	113
New-thyroid	175	180	35	174	40	179	40	179	0	5	47	179

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

SMOTE - Synthetic Minority Oversampling Technique

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

  (j depends on the amount of oversampling desired)

- 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

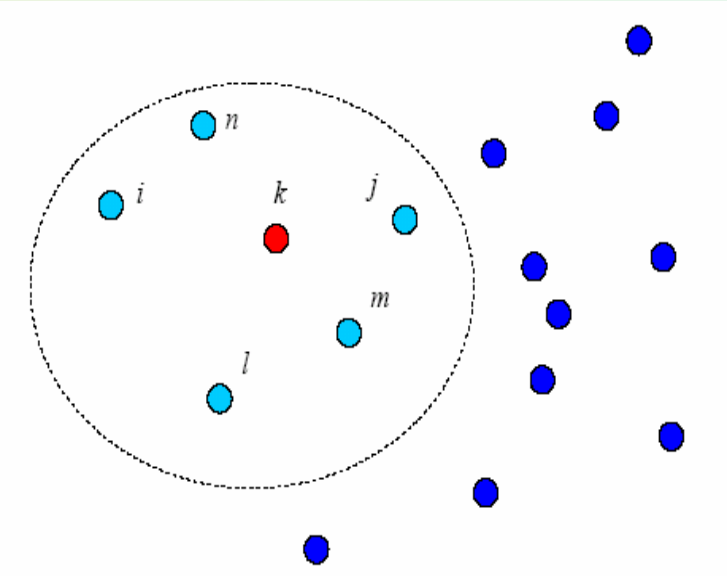
- Sąsiedzi przykładu *p* identyfikowani poprzez K-NN z odległością HVDM
- Ogólny schemat dla atr. liczbowych
  - Take the difference between the feature vector (sample) under consideration and its nearest neighbor.
  - Multiply this difference by a random number between 0 and 1
  - Add it to the feature 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.

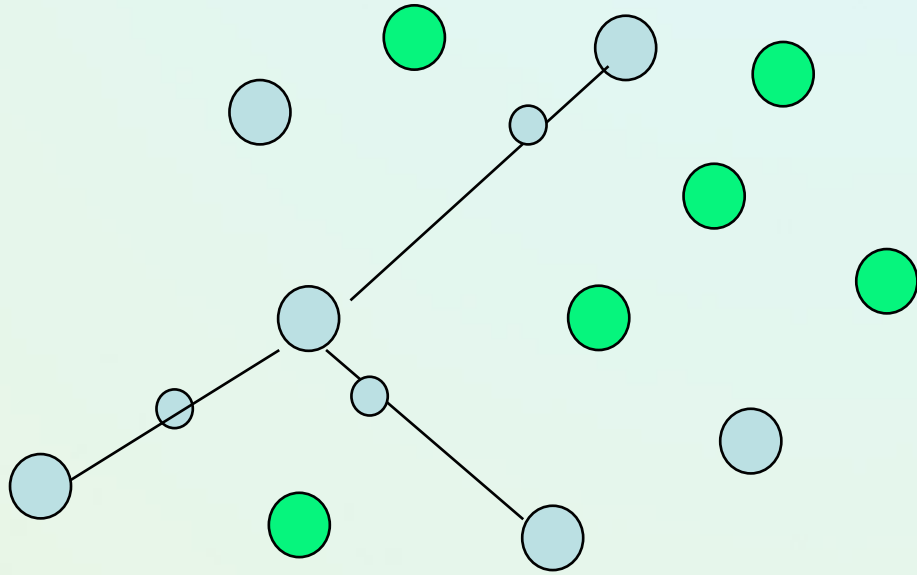
Consider a sample (6,4) and let (4,3) be its nearest neighbor.
(6,4) is the sample for which k-nearest neighbors are bei identified
(4,3) is one of its k-nearest neighbors.
Let:
f1_1 = 6 f2_1 = 4 f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3 f2_2 - f1_2 = -1
The new samples will be generated as
(f1',f2') = (6,4)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

: Przykład kl. większościow

: syntetyczny przykład

SMOTE – przykład oceny AUC

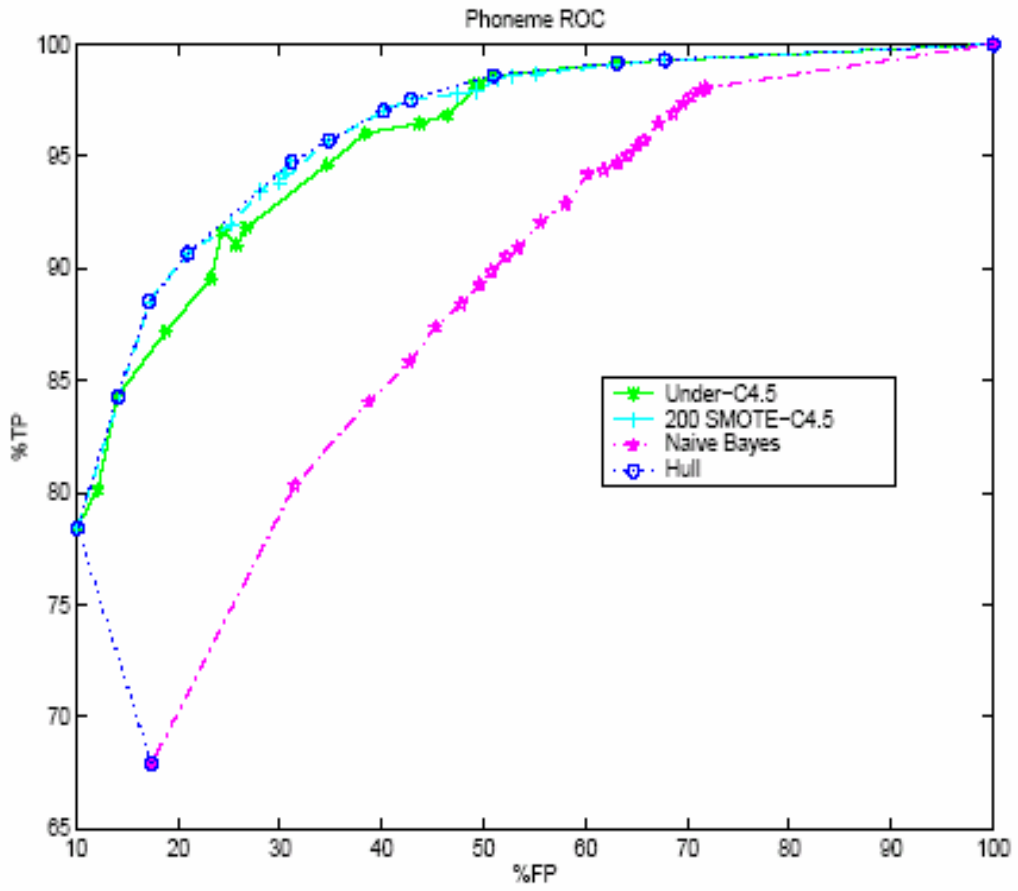


Figure 7: Phoneme. Comparison of SMOTE-C4.5, Under-C4.5, and Naive Bayes. SMOTE-C4.5 dominates over Naive Bayes and Under-C4.5 in the ROC space. SMOTE-C4.5 classifiers are potentially optimal classifiers.

SMOTE zbiorcza ocena

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

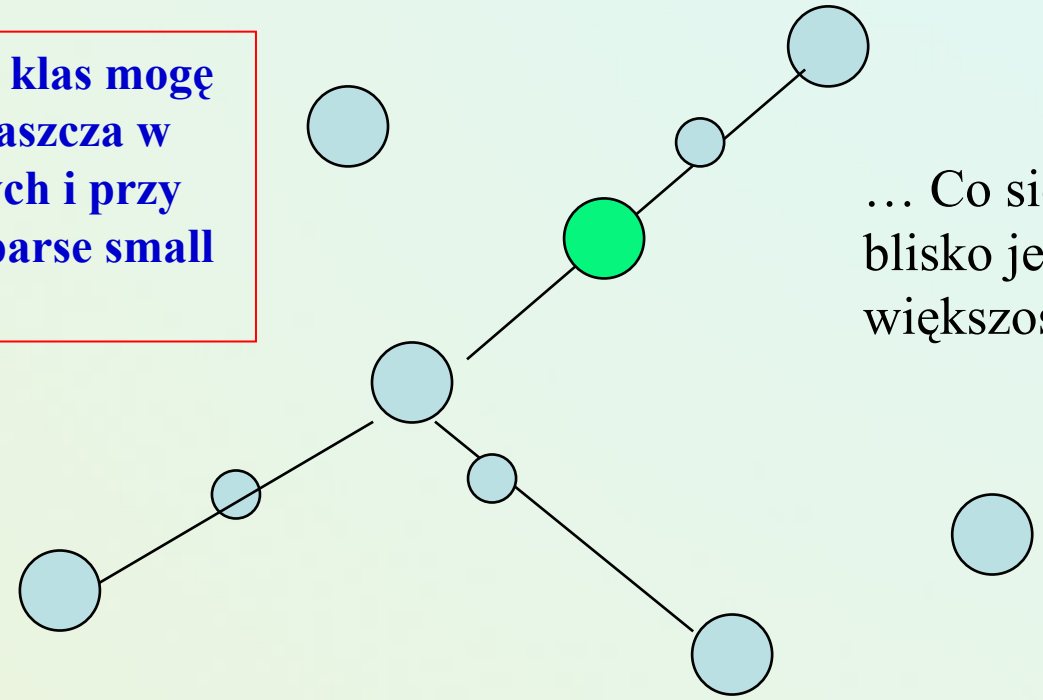
Dataset	Under	50 SMOTE	100 SMOTE	200 SMOTE	300 SMOTE	400 SMOTE	500 SMOTE
Pima	7242		**7307**				
Phoneme	8622		8644	**8661**			
Satimage	8900		8957	**8979**	8963	8975	8960
Forest Cover	9807		9832	9834	**9849**	9841	9842
Oil	8524		8523	8368	8161	8339	**8537**
Mammography	9260		9250	9265	9311	**9330**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

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

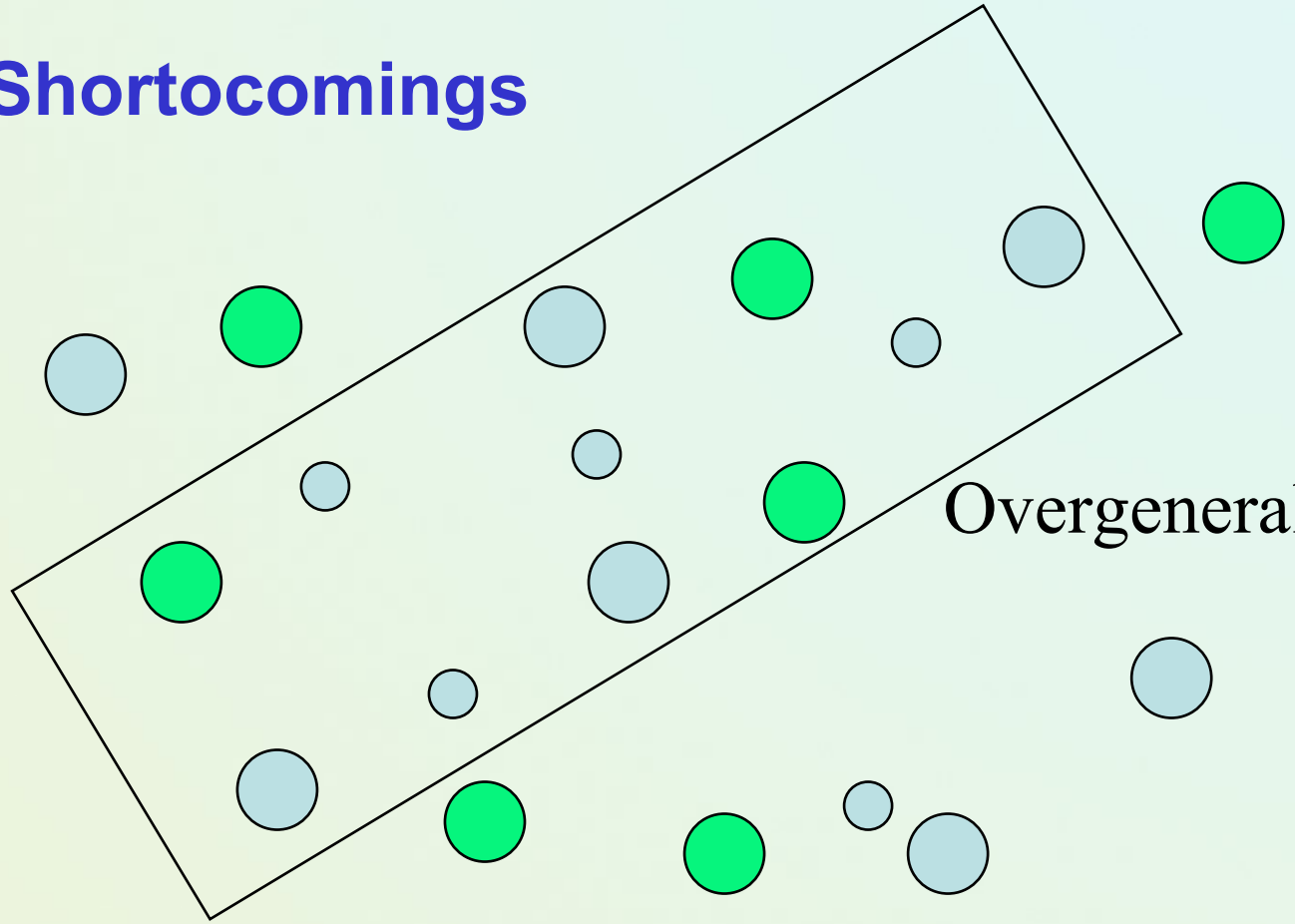
: Synthetic sample

: Majority sample

MOTE – uwagi krytyczne

- Overgeneralization
  - SMOTE's procedure is inherently dangerous since it blindly generalizes the minority area without regard to the majority class.
  - This strategy is particularly problematic in the case of highly skewed class distributions since, in such cases, the minority class is very sparse with respect to the majority class, thus resulting in a greater chance of class mixture.
- Lack of Flexibility
  - The number of synthetic samples generated by SMOTE is fixed in advance, thus not allowing for any flexibility in the re-balancing rate.

SMOTE Shortocomings



: Minority sample

: Majority sample

: Synthetic sample

lajnowsze 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

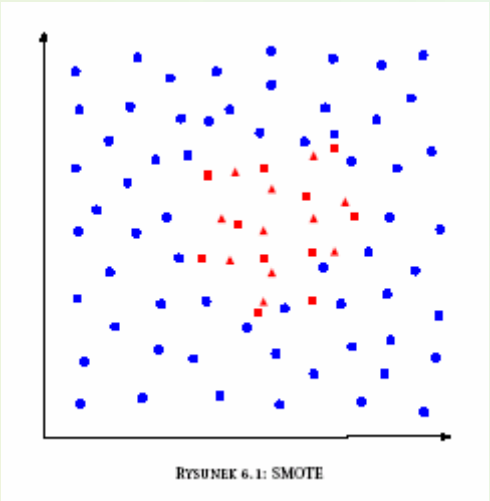
Safe_Level_SMOTE: C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09). LNAI 5476, Springer-Verlag 2005, Bangkok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

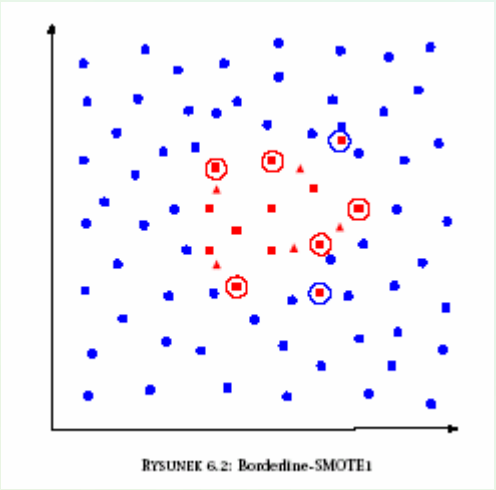
LN-SMOTE: J. Stefanowski, T.Maciejewski.

SMOTE vs. inne rozszerzenia

- Przykład ilustracyjny

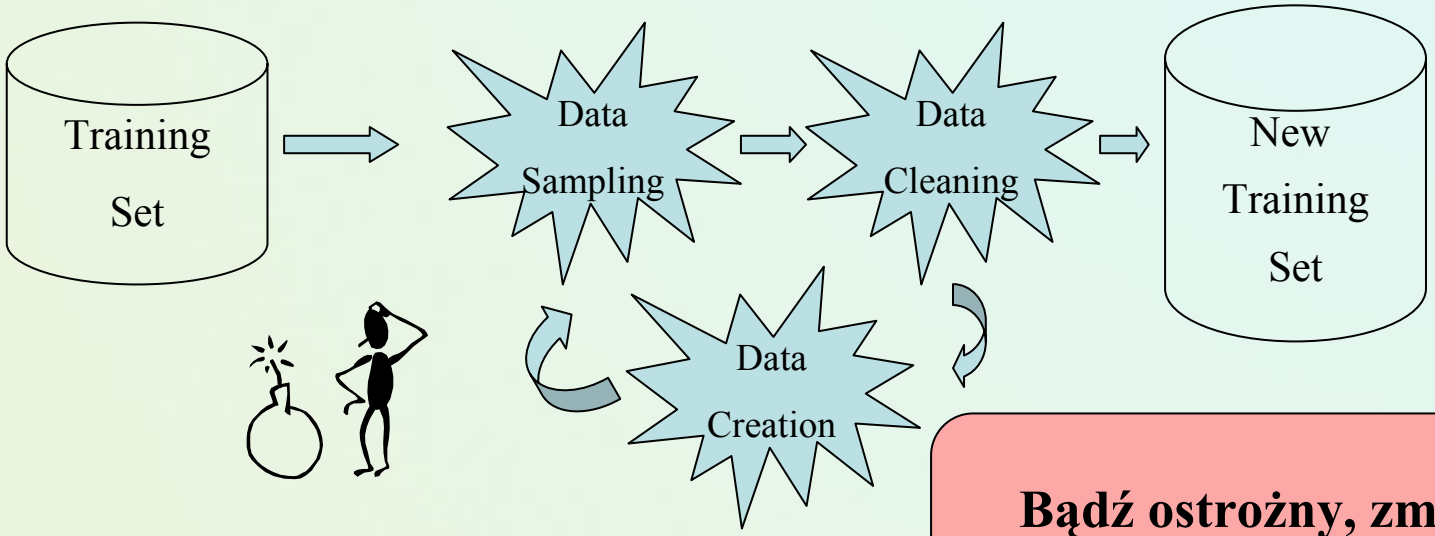


RYSUNEK 6.1: SMOTE



RYSUNEK 6.2: Borderline-SMOTE1

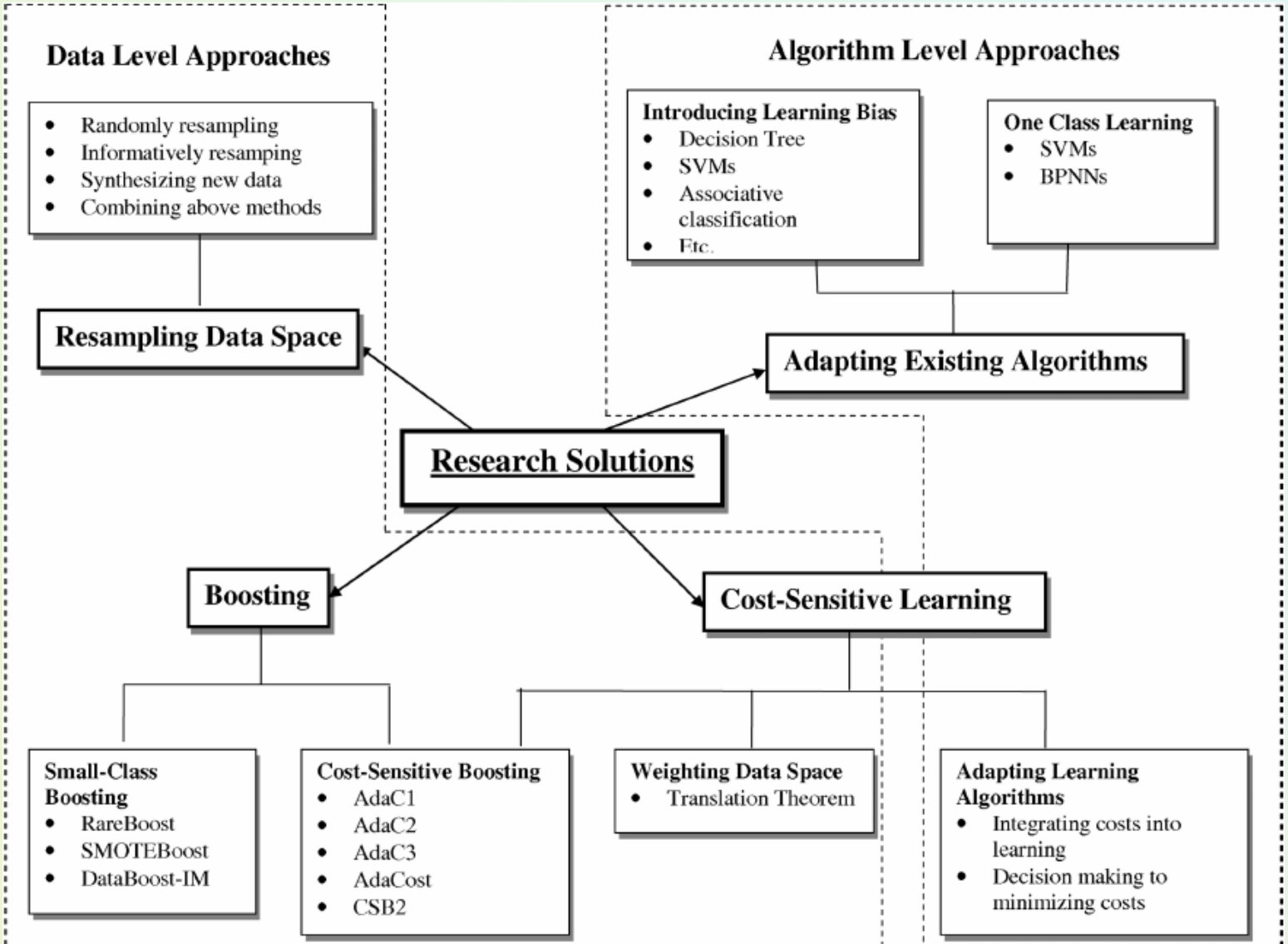
ilka uwag podsumowujących metody informed re-sampling



Bądź ostrożny, zmieniasz to co masz się nauczyć!

- 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)

ummary: Data level vs Algorithm Level



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

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**

Kilka uwag na koniec

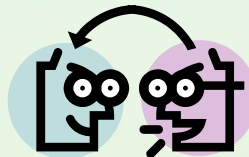
- Odkrywanie wiedzy z danych
- Potrzeby informacyjne
  - „We are drowning in the deluge of data that are being collected world-wide, while starving for knowledge at the same time”*
- Rozwój KDD i pośrednio ML dostarczył wiele metod
- Badania (tak) choć nadal wiele otwartych problemów, …
- Jak wyglądają zastosowania biznesowe?
- „Rzadkie” pojęcia słabo reprezentowane (<10% ?) lecz mogą znacząco i negatywnie wpływać na dane zadanie

“Mining needle in a haystack. So much hay and so little time”

* J. Naisbitt, Megatrends: Ten New Directions Transforming Our Lives.

Podsumowanie

- Niezrównoważony 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**
- Dalsze kierunki badań:
  - Klasyfikatory złożone (IIvotes)
  - Hybrydyzacja (elementy LN-SMOTE i SPIDER2)



Dziękuję za uwagę

Pytania lub komentarze?

Więcej informacji w moich artykułach!

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