Odkrywanie wiedzy klasyfikacyjnej z niezbalansowanych danych

Learning classifiers from imbalanced data

Część 2 – modyfikacje algorytmów

Wykład spec. projekt eksploracji danych

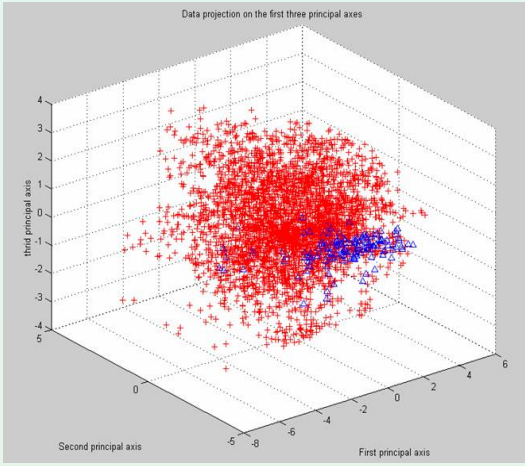
JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska
Poznań

Poznań, 2020

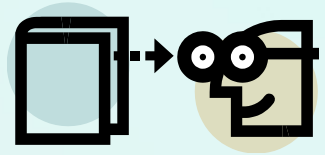
Plan wykładu

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



Dzisiaj cz. 2

Modyfikacje algorytmów



- Dwa podstawowe kierunki działania
  - Modyfikacje danych (preprocessing i re-sampling)
  - **Modyfikacje algorytmów**
- Najbardziej popularne podejścia w ramach drugiej grupy
  - Re-sampling or re-weighting,
  - Transformacje do zadania „cost-sensitive learning“
  - 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
  - …

Inne podejścia do modyfikacji algorytmów uczących

- Zmiany w indukcji drzew decyzyjnych (np. Hellinger distances lub asymetryczne entropie)
  - 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-sensitive learning“
  - 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.

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

Cost sensitive learning

Transparent → interpretacja pracy alg. / klasyfikatora

np: specific cost-sensitive algorithms, some of the weighting approaches, threshold modyfing

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

Black box → złożone, słabo interpretowalne podejścia

np.: cost-sensitive ANN, MetaCost, some boosting approaches

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

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

Cost Sensitive SVM

negative classes, SVM can be extended to the cost-sensitive setting by introducing an additional parameter that penalizes the errors asymmetrically.

Consider that we have a binary classification problem, which is represented by a data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$, where $x_i \subset \mathbb{R}^k$ represents a k-dimensional data point and $y_i \in \{+1, -1\}$ represents the class of that data point, for $i = 1, \ldots, l$. Let $I_+ = \{i : y_i = +1\}$ and $I_- = \{i : y_i = -1\}$. The support vector technique requires the solution of the quadratic programming problem as follows [20]:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C^+ \sum_{i \in I_+} \xi_i + C^- \sum_{i \in I_-} \xi_i \tag{1}$$

subject to

$$\begin{aligned} y_i(w \cdot \phi(x_i) + b) &\geq 1 - \xi_i, \\ \xi_i &\geq 0, \quad i = 1, \ldots, l, \end{aligned} \tag{2}$$

where the training vectors x_i are mapped into a higher-dimensional space by the function ϕ. Parameter C^+ represents the cost of misclassifying the positive sample, and C^- represents the cost of misclassifying the negative sample. The optimal result can be obtained when C^-/C^+ equals the minority-to-majority class ratio. The slack variables $\xi_i > 0$ hold for misclassified samples, and therefore, $\sum_{i=1}^{l} \xi_i$ can be thought of as a measure of the amount of misclassifications. This quadratic-optimization problem can be

Wprowadzić koszty C do sformułowania zadania programowania matematycznego w SVM

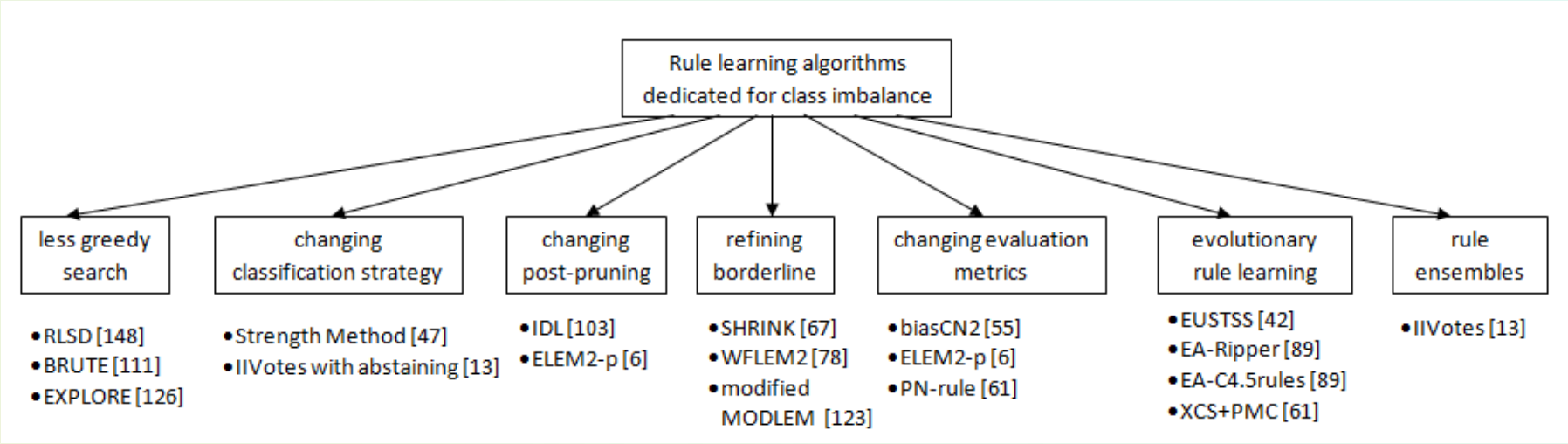
Reguły i niezblanasowanie klas

- 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

Modyfikacje klasyfikatorów regułowych



Większość uzwględnia pojedyncze ograniczenia

Review →
K.Napierała: Improving Rule Classifiers for Imbalanced Data. Ph.D. Thesis, PUT, 2013.
K.Napierała, J. Stefanowski: BRACID A comprehesive approach to rule induction from imbalanced data. Int. Journal of Intelligent Information Systems, 2012.

BRACID

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

Założenia:

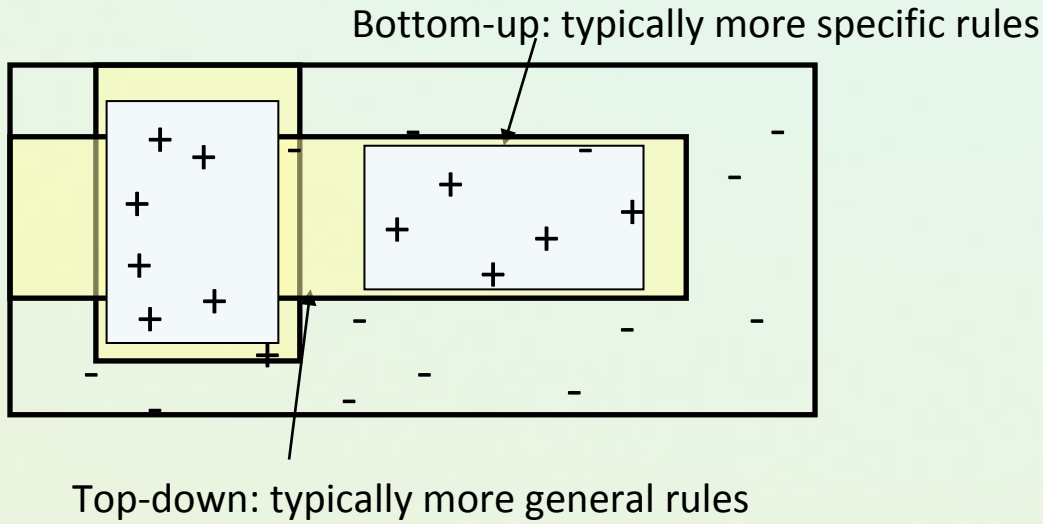
- Hybrid knowledge representation: rule and instances
- Induction rules by bottom-up strategy
- Resigning from greedy sequential covering
- Some inspirations from RISE [P.Domingos 1996]
- Considering info about types of difficult examples
- Local neighbors with HVDM
- Internal evaluation criterion (F-miara)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

wiecej →

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

Od przykładów do reguł → bottom up generalization

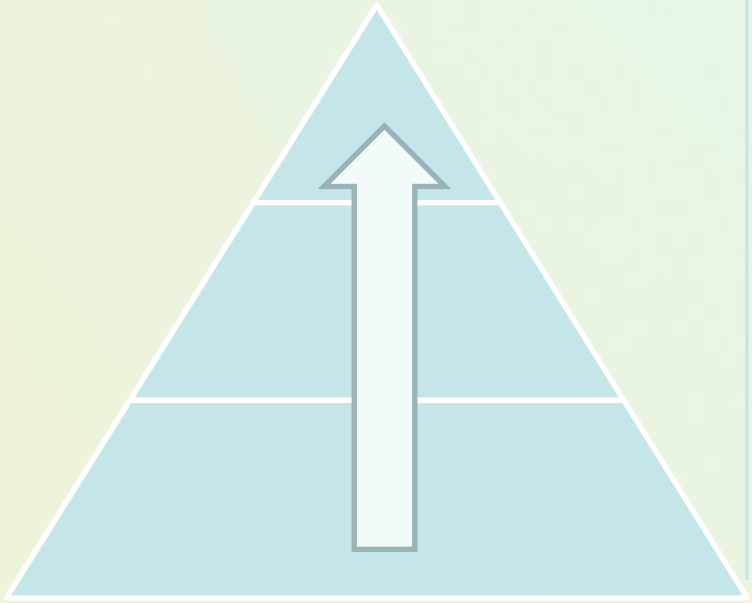
- Single example → „seed” for the most specific rule
  - **x1** → (a1=L),(a2=s),(a3=2),(a4=3) – Class A
  - **r1**: IF (a1=L) and (a2=s) and (a3=2) and (a4=3) THEN Class A
- Bottom up generalization
  - New examples and the nearest rule
  - **x2** → (a1=L),(a2=t),(a3=2.7),(a4=3) – Class A
  - **r1’**: IF (a1=L) and (a3∈[2;2.7]) and (a4=3) THEN Class A
- Rule syntax
  - For nominal attribute (ai = value_ij)
  - For numerical attribute (vi_lower ≤ a_i ≤ vi_upper)



BRACID

Bottom-up induction of Rules And Cases from Imbalanced Data

- Bottom-up
- Non-sequential covering
- Evaluation of new rules with F-measues – efficient updating classification records



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

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

25 Return RS
```

BRACID – ocena eksperymentalna

Cele:

- Zróżnicowane dane niezbalansowane
- Porównanie wielu algorytmów
  - CN2
  - MODLEM
  - C4.5 Rules
  - RIPPER
  - PART
  - MODLEM-C multiplier classification strategy
  - RISE
- Ponadto integracja z metodami przetwarzania wstępnego
  - PART + SMOTE

Ocena- G-mean

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

BRACID vers. Specjalizowane podejścia

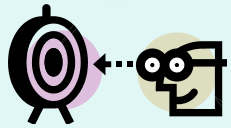


Table 8 G-mean for algorithms specialized for class imbalance

Dataset	BRACID	RISE	Modlem-C	PART SMOTE	PART SMOTE+ENN
abalone	0.650	0.345	0.513	0.643	**0,704**
abdominal-pain	0.811	0.805	0.793	0.790	**0,818**
balance-scale	**0.567**	0.000	0.000	0.346	0,462
breast-cancer	**0.559**	0.545	0.530	0.526	0,540
breast-w	**0.968**	0.963	0.947	0.959	0,962
car	0.870	0.751	0.879	**0.916**	0,842
cleveland	**0.574**	0.232	0.225	0.410	0,565
cmc	**0.637**	0.507	0.544	0.581	0,635
credit-g	0.611	0.540	0.645	0.612	**0,658**
ecoli	**0.830**	0.638	0.633	0.826	0,826
flags	**0.481**	0.025	0.000	0.224	0,224
haberman	0.576	0.375	0.532	**0.608**	0,596
hepatitis	**0.751**	0.604	0.644	0.639	0,656
ionosphere	0.912	**0.928**	0.898	0.876	0,868
new-thyroid	**0.984**	0.951	0.903	0.955	0,955
pima	**0.712**	0.666	0.704	0.681	0,660
postoperative	**0.345**	0.193	0.297	0.158	0,251
scrotal-pain	0.731	0.667	0.729	0.716	**0,732**
solar-flare	0.638	0.135	0.322	0.492	**0,651**
transfusion	**0.639**	0.507	0.579	0.601	0,621
vehicle	0.935	0.895	0.941	0.932	**0,942**
yeast	0.709	0.436	0.370	**0.749**	0,658

Friedman test:
BRACID 1.25; SMOTE+ENN+PART 2.51;
SMOTE+PART 3.16; MODLEM-C 3.86; RISE 3.98

CD = 1.3 (with Nemenyi post-hoc test)

BRACID vs. S+E+PART – Wilcoxon rejects H0

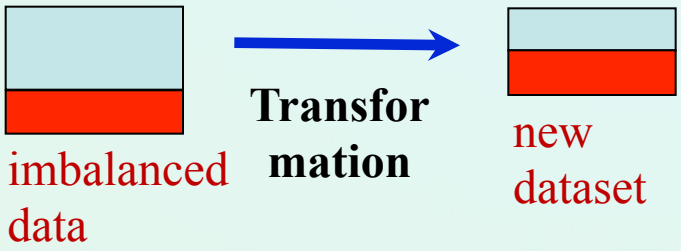
Similar tests for Sensitivity and F-measure

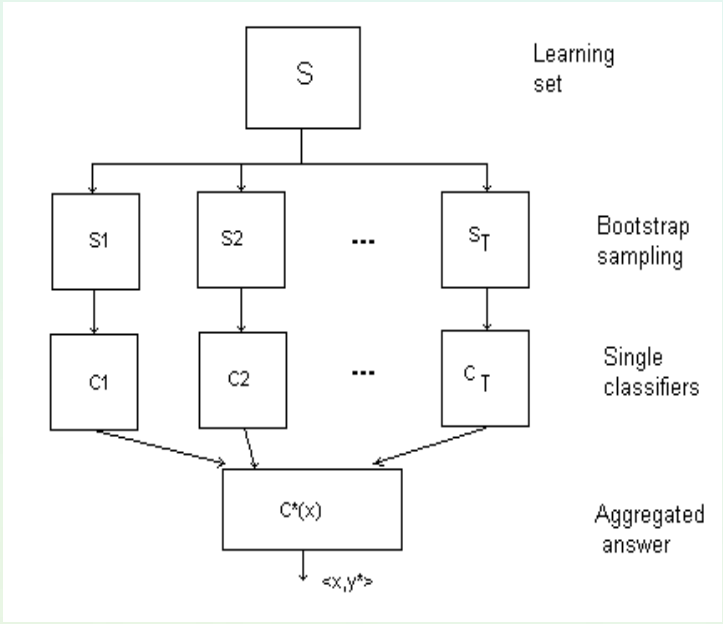
- SMOTE - Generate a synthetic example along the line between minority neighbours
- BRACID significantly better than other (Friedman test + post hoc)

More → K.Napierała, J. Stefanowski: BRACID Int. Journal of Intelligent Information Systems, 2012.

Adaptacje zespołów klasyfikatorów

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

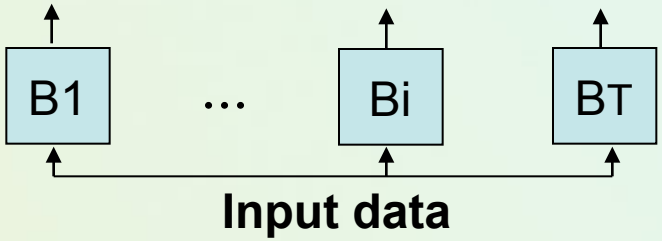


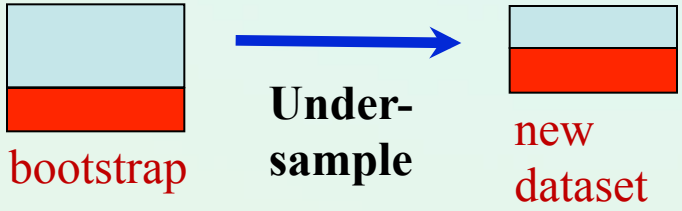


Related: Galar et. al., A Review on Ensembles for the Class Imbalance Problem. IEEE Trans. 2011

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]
  - Equal probabilities of class sampling → BS_maj
  - Sampling with replacement N_min and BS_maj

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)

Porównanie wielu zespołów klasyfikatorów

Comparative studies

Galar, Herrera et al [2011]

- Simpler generalizations better than more complex or cost based ones

Khoshgoftaar et al. [2011]

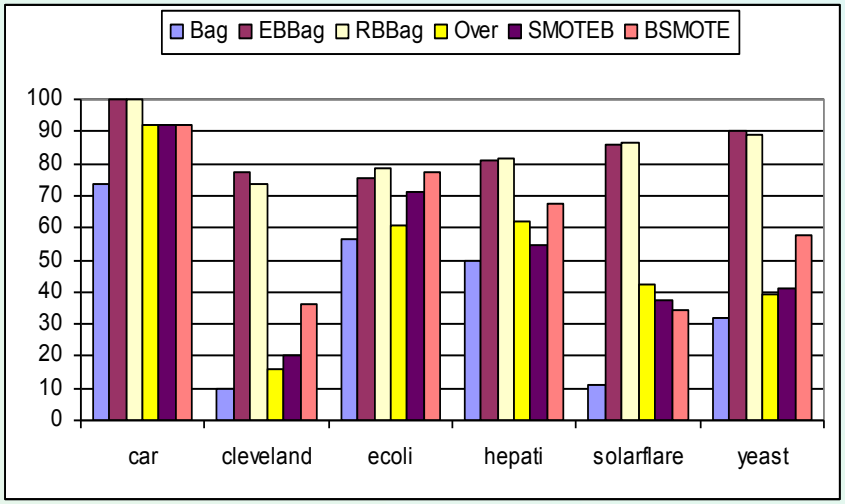
- EBBag, RBBag better then SMOTEBoost and RUBoost

Our study on bagging [2013]

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

Similar observations, e.g.

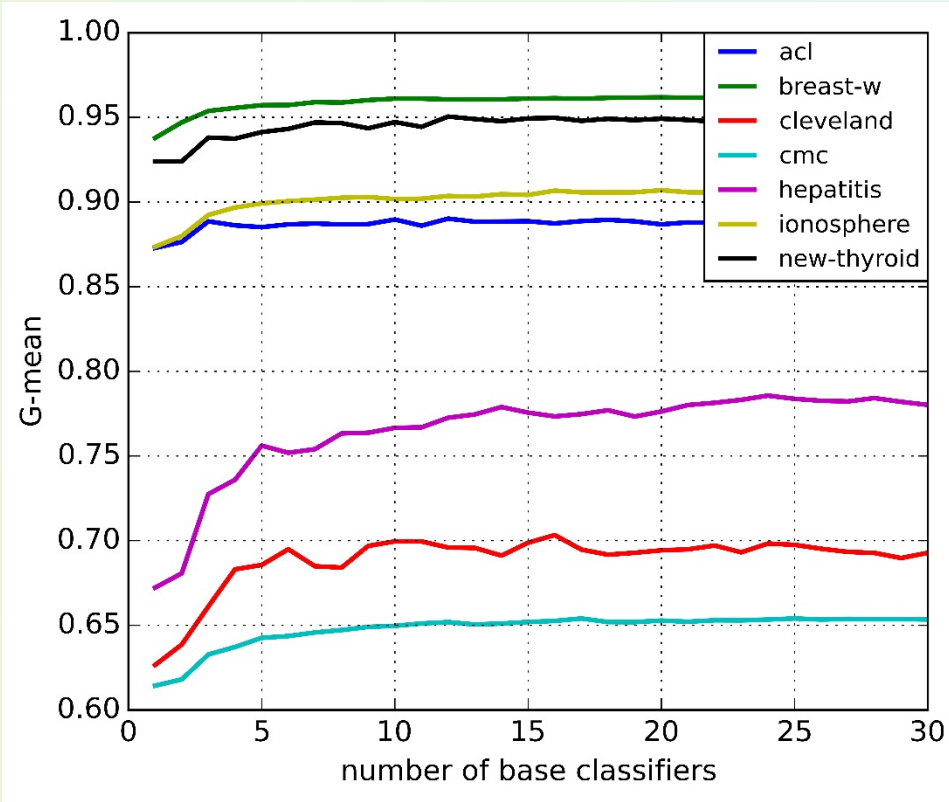
- Liu A., Zhu Zh [2013] + BalancedRF; and others,

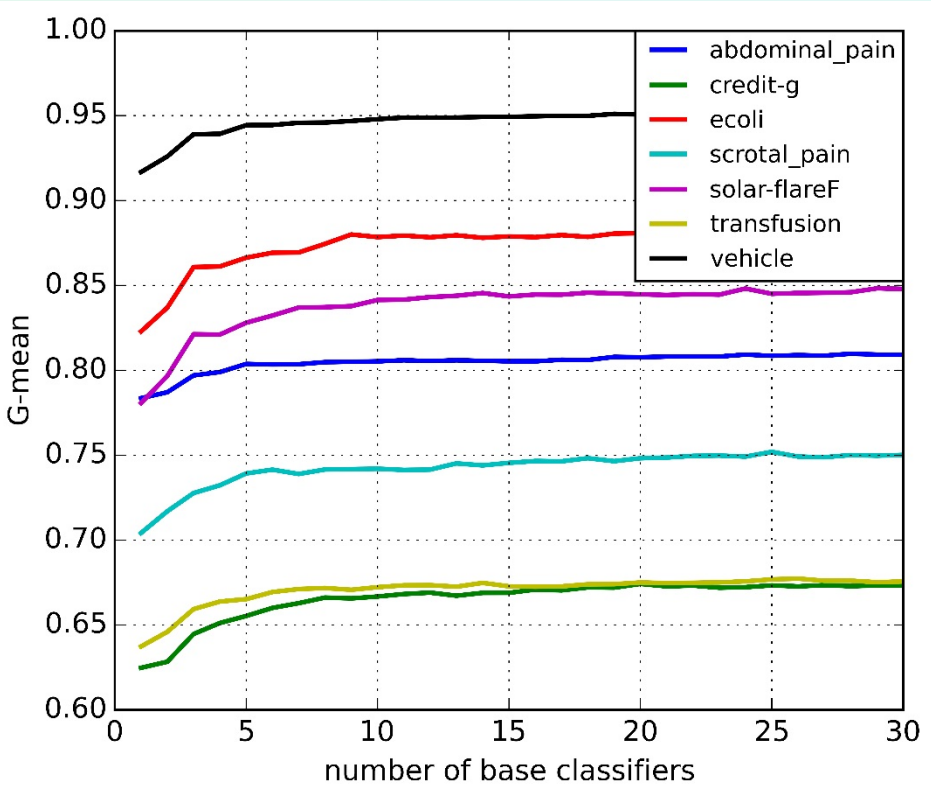


Dataset	Bag	EBBag	RBBag	OvBag	SmBag	BagSm
breast-w	95.88	96.03	96.37	96.23	95.88	96.77
abdominal-pain	78.95	80.65	80.35	79.44	80.85	79.86
acl	88.18	90.71	89.35	88.35	88.64	87.81
new-thyroid	92.41	96.91	96.58	95.36	95.18	92.89
vehicle	93.91	94.58	95.44	94.61	94.34	94.20
car	84.53	96.73	96.58	95.29	95.26	95.18
scrotal-pain	70.75	73.18	75.65	72.01	70.42	70.68
ionosphere	88.96	90.44	90.67	90.47	90.30	90.26
pima	71.54	74.22	75.64	73.54	72.33	71.38
credit-g	63.98	65.82	67.82	71.75	80.68	66.11
ecoli	68.67	72.24	88.85	51.42	58.38	80.11
hepatitis	62.81	78.93	78.66	72.16	68.47	74.29
haberman	43.11	65.41	63.43	58.11	60.02	62.82
breast-cancer	54.30	58.82	59.37	56.17	52.57	57.25
cmc	52.76	64.61	65.27	59.95	57.74	62.77
cleveland	12.61	72.32	71.02	22.77	25.03	50.96
hsv	0.00	36.27	35.74	2.84	5.37	16.61
abalone	49.58	78.93	79.32	61.95	63.67	69.65
postoperative	1.99	24.97	34.03	15.01	1.57	11.55
solar-flare	13.70	85.39	83.21	58.07	55.04	54.40
transfusion	55.72	66.75	67.32	64.83	63.96	65.76
yeast	51.48	84.55	84.68	59.70	59.41	57.94
balance-scale	0.00	59.07	54.23	1.40	0.00	0.67
average rank	5.61	1.96	1.61	3.65	4.26	3.91

J. Blaszczynski, J., Stefanowski: Extending bagging for imbalanced data. Proc. CORES 2013.

RBBag (liczba drzew decyzyjnych)





Relatywnie mała:

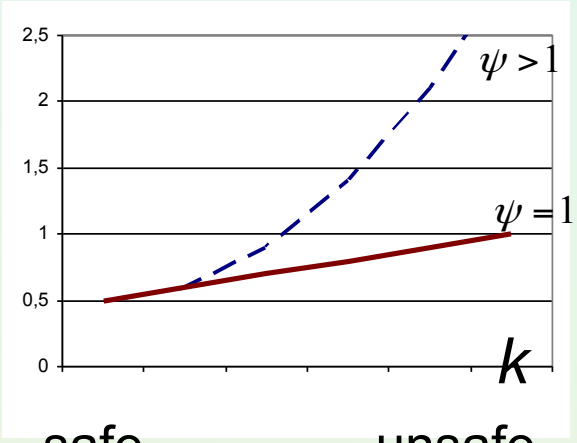
- Dla większości danych wystarczy kilkanaście

Neighbourhood Balanced Bagging

- Propozycja wykorzystujące inne zasady:
  - Zmodyfikuj prawdopodobieństwo losowania do próbki bootstrapowej z wykorzystaniem “safe level” przykładu
  - Zwiększ szanse wyboru przykładów mniejszościowych kosztem większościowych (global prob.)
- Global level
  - $p^1_{\min} = 1$ (minority class)
  - $p^1_{maj} = N_{\min} / N_{maj}$ (decrease → inverse global imbalance)
- Local level
  - Minority local neighb. $\psi \geq 1$

$$P_{global} \cdot P_{local}$$

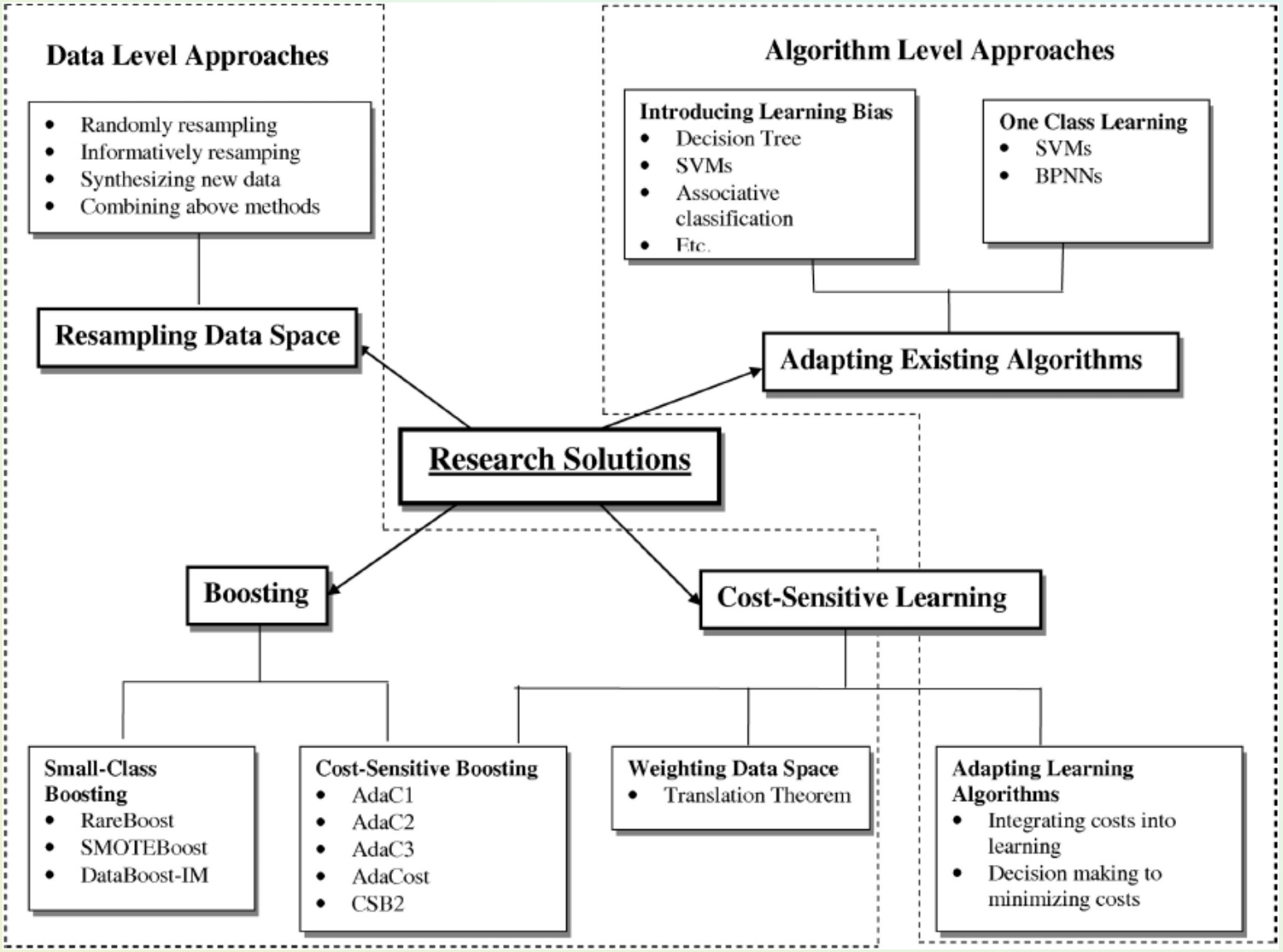
$$L = \frac{\left(N'_{maj}\right)^{\psi}}{k}$$



- Experiments – competitive to RBBag and better than other bagging variants

J.Blaszczynski, J.Stefanowski: Neighbourhood sampling in bagging for imbalanced data. Neurocomputing (2015)

Podsumowanie: 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.

Więcej o miarach oceny klasyfikatorów

Wiele propozycji (różna interpretacja i przydatność)

Typowy podział [Ferri et al, He book, Japkowicz]:

1. Point / threshold measures
2. Probabilistic measures
3. Ranking measures

Przykłady

G-mean, Fβ-measure, Kappa, MCC, IBA

ROC – AUC, Precision-Recall curves

Pierwsze ukierunkowane na analizę błędów klasyfikowania / definiowane najczęściej na podstawie zawartości macierzy pomyłek

Podstawowe miary

Binarna macierz pomyłek:

Table 1. Confusion matrix for two classes classification

	Predicted Positive	Predicted Negative
Actual positive	TP (number of True Positive)	FN (number of False Negative)
Actual Negative	FP (number of False Positive)	TP (number of True Positive)

Pojedyncze miary (dwie pierwsze nieprzydatne)

Table 2. Fundamental evaluation metrics based on confusion matrix analysis

Measure	Formula	interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	
Error rate = 1-Accuracy	$\frac{FP+FN}{TP+TN+FP+FN}$	
Sensitivity (or Recall)	$\frac{TP}{TP+FN}$	Accuracy of positive examples
Specificity	$\frac{TN}{TN+FP}$	Accuracy of Negative examples
Precision	$\frac{TP}{TP+FP}$	

Złożone miary

Najpopularniejsze:

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

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

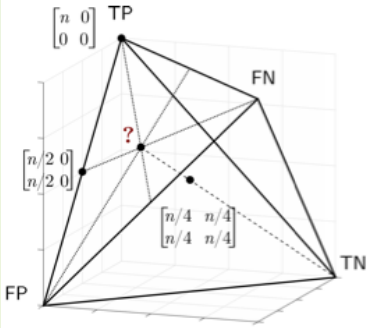
Matthews correlation coefficient, MCC

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

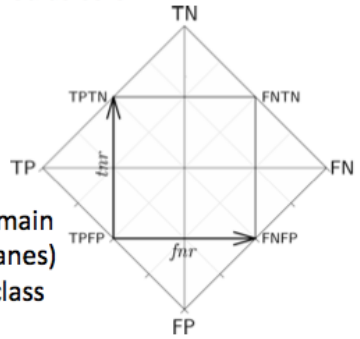
(MCC) expresses a correlation between the actual and predicted classification and returns a value between −1 (total disagreement) and +1 (perfect agreement); 0 classifiers performs randomly

Tetrahedron – analiza miar oceny klasyfikatorów

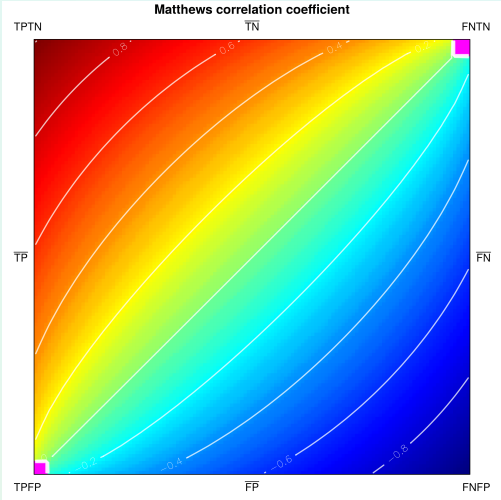
Barycentric Visualization

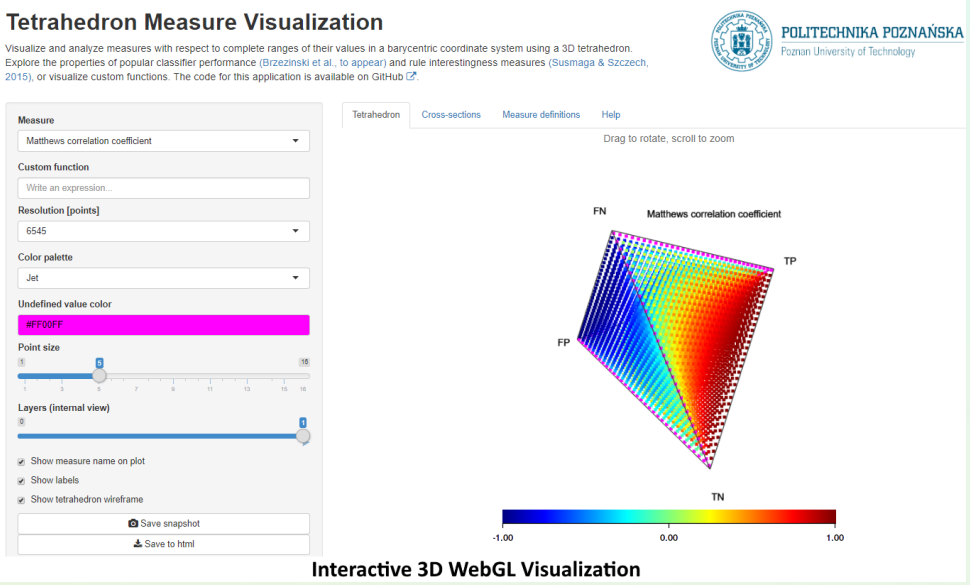


- In the barycentric coordinate system each confusion matrix is represented as a point of a 3D tetrahedron
- The value of a measure based on the depicted four values may be rendered as color



- Visual **comparison** of measures
- Visual detection of **properties**
- Insight into **full range** of measure's domain
- Consecutive **cross-sections** (cutting planes) depict measure behavior in changing class or prediction proportions





D.Brzeziński, J.Stefanowski, R.Susmaga, I.Szczech ECMLPKDD 2017, oraz artykuł w Information Science 2018

Score classifier – podejście rankingowe

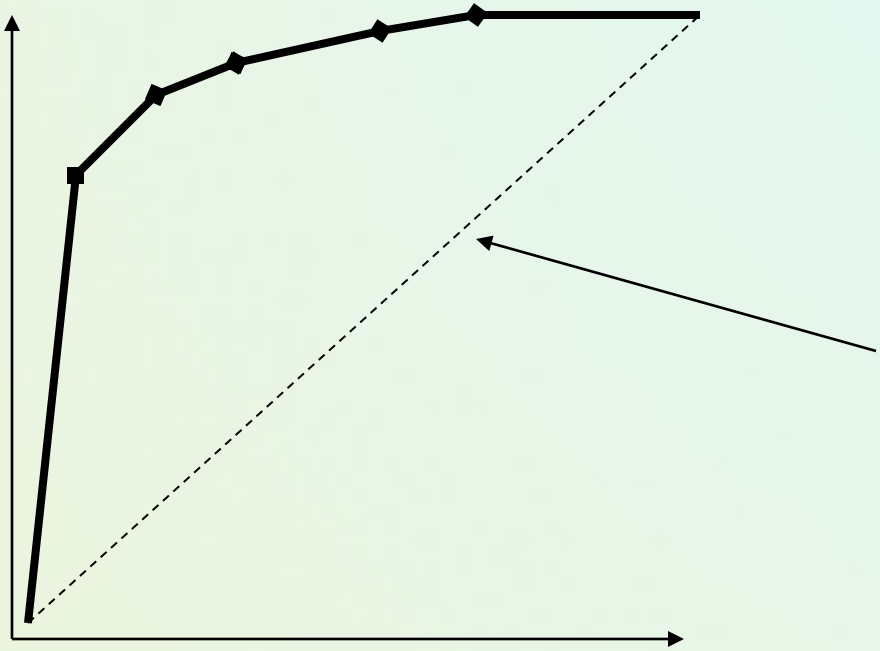
Score based models

●	Positive labelled example
○	Negative labelled example

$$\text{Prevalence} = \frac{\#\text{positives}}{\#\text{positives} + \#\text{negatives}}$$

Klasyfikator - możliwość progowania wyjścia predyktora

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

Probabilistyczne podstawy ROC

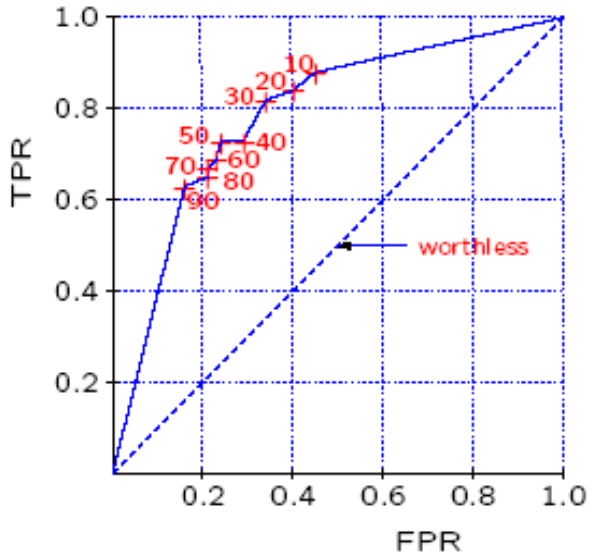
- Output of probabilistic classifier:

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

may not yield the best performance

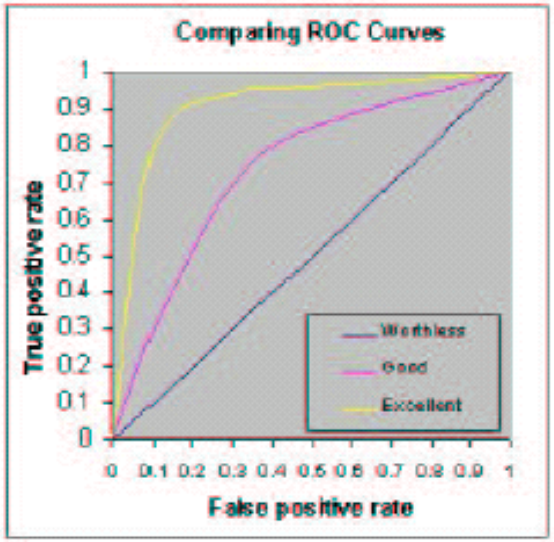
- Alternative: Receiver Operating Characteristic (ROC): determine threshold d, such that

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



When comparing various techniques:

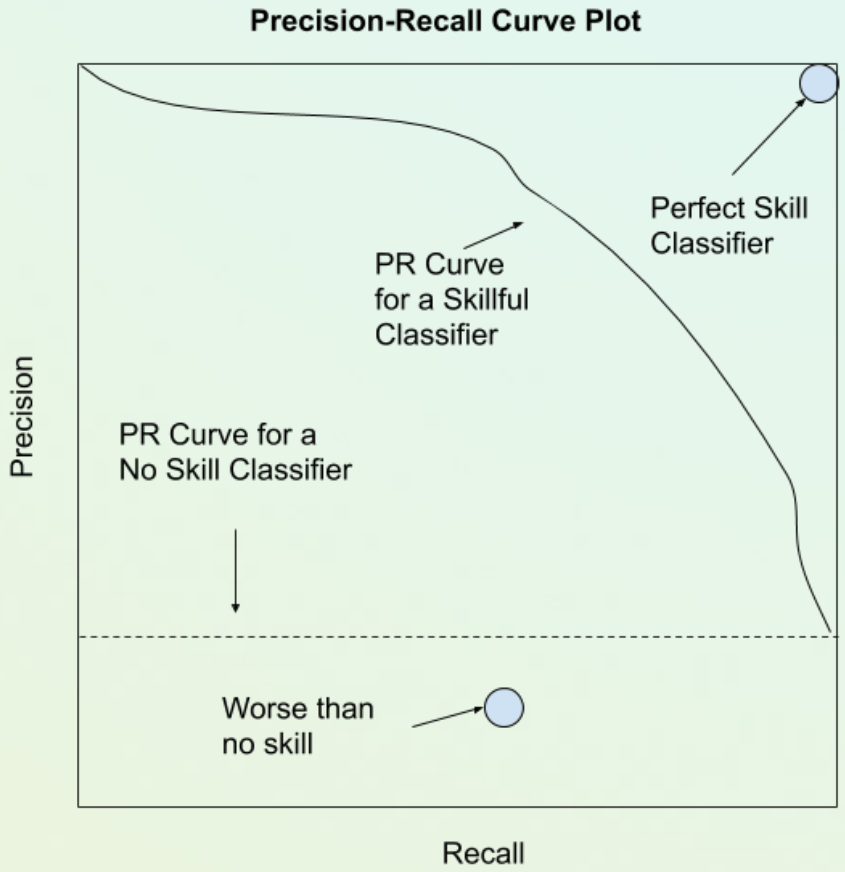
- actual performance for particular thresholds (cut-off points) may vary
- area under the ROC curve $A_f = \int_0^1 f(x)\mathrm{d}x$ offers good measure for comparison, with f relationship between FPR and TPR for classifier



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



Oceny probabilityczne klasyfikatorów

Ocena niepewności predykcji klasyfikatorów

Klasyfikator udostępnia oszacowania prawdopodobieństw

Brier Score lub LogLoss

The Brier Score is the mean square difference between the true classes and the predicted probabilities.

This function implements the original multi-class definition by Brier (1950), normalized to $[0, 1]$ as in Kruppa et al (2014). The formula is the following:

$$BS = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{K} (C_{ik} - p_{ik})^2$$

where n is the number of observations, K the number of classes, $C_{ik} = \{0, 1\}$ the indicator of class k for observation i, and p_{ik} is the predicted probability of observation i to belong to class k.

The above formulation is applicable to multi-class predictions, including the binary case. A small value of the Brier Score indicates high prediction accuracy.

The Brier Score is a strictly proper score (Gneiting and Raftery, 2007), which means that it takes its minimal value only when the predicted probabilities match the empirical probabilities.

References

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78 (1): 1-3.

Gneiting, G. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102 (477): 359-378.

Kruppa, J., Liu, Y., Diener, H.-C., Holste, T., Weimar, C., Koonig, I. R., and Ziegler, A. (2014) Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical Journal*, 56 (4): 564-583.

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 wielowymiaroych
- 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

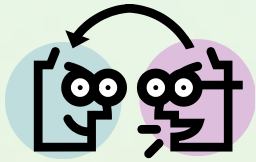
IEEE ICDM noted "Dealing with Non-static, Unbalanced and Cost-sensitive Data" among the **10 Challenging Problems in Data Mining Research**

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

Podsumowanie

- Niezbalansowany rozkład liczności klas (class imbalance)
  → źródło trudności dla konstrukcji klasyfikatorów
- Często występuje w zastosowaniach
- 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
- Ciągle ograniczona dostępność zaawansowanych lub aktualnych metod w popularnym oprogramowaniu



Dziękuję za uwagę

Pytania lub komentarze?

Więcej informacji w publikacjach!

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