Zaawansowana eksploracja danych: Metody oceny wiedzy klasyfikacyjnej odkrytej z danych

Jerzy Stefanowski
Instytut Informatyki
Politechnika Poznańska

Wykład dla spec. Mgr TPD
Poznań 2008 – popr. 2010

Ocena wiedzy klasyfikacyjnej wykład dla TPD

1. **Różne metody odkrywania wiedzy klasyfikacyjnej**
2. **Metody oceny wiedzy o klasyfikacji obiektów – przegląd miar**
3. **Eksperymentalna ocena klasyfikatorów**
4. **Porównanie wielu klasyfikatorów**
5. **Miary oceny w perspektywie opisowej na przykładzie reguł**

Uwagi do źródeł

Przygotowując wykład korzystałem m.in. z książek:

- S.Weiss, C.Kulikowski: Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems, Morgan Kaufmann 1991.
- K.Krawiec, J.Stefanowski: Uczenie maszynowe i sieci neuronowe, WPP 2004.
- J.Han, M.Kember: Data mining.
- Opis pakietu WEKA

oraz inspiracji ze slajdów wykładów nt. data mining następujących osób:

- J.Han; G.Piatetsky-Shapiro; Materiały związane z WEKA
- i prezentacji W.Kotłowski nt. oceny systemów uczących się.

Wybrane artykuły

1. Kohavi, R. (1995): A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. of the 14th International Joint Conference on Artificial Intelligence*, 1137—1143.
2. Salzberg, S. L. (1997): On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1, 317—328.
3. Dietterich, T. (1998): Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:7, 1895-–1924.
4. Bouckaert, R. R. (2003): Choosing between two learning algorithms based on calibrated tests. *ICML 2003*.
5. Bengio, Y., Grandvalet, Y. (2004): No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5, 1089—1105.
6. Demsar, J. (2006): Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.
7. Sesja spacjalna nt. oceny systemów uczących (N.Japkowicz) na ICML 2007.

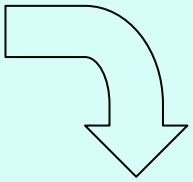
Wiedza o klasyfikacji obserwacji

- Problem określania zasad przydziału obiektów / obserwacji do znanych wstępnie klas na podstawie analizy danych o przykładach klasyfikacji.

Wiek	Zawód	dochód	…	Decyzja
21	Prac. fiz.	1220	…	Nie kupi
26	Menedżer	2900	…	Kupuje
44	Inżynier	2600	…	Kupuje
23	Student	1100	…	Kupuje
56	Nauczyciel	1700	…	Nie kupi
…	…	…	…	…
45	Lekarz	2200	…	Nie kupi
25	Student	800	…	Kupuje

Przykłady uczące

Algorytm eksploracji



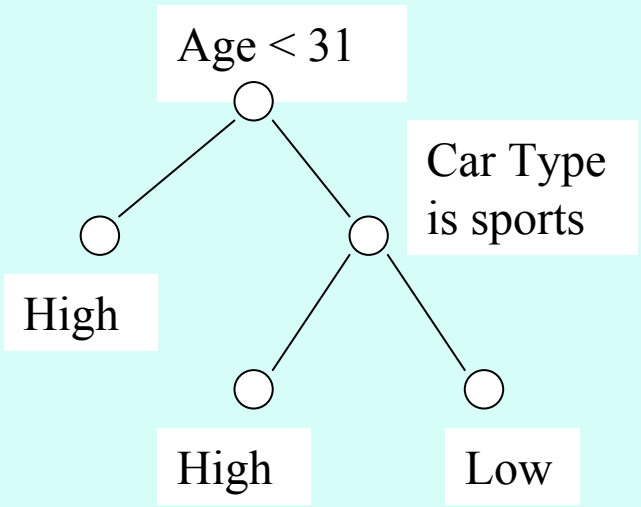
Reprezentacja wiedzy:
np. reguły
R1. Jeżeli student to kupuje komputer
R2. Jeżeli dochód > 2400 …

Podstawowe metody (klasyfikacyjne)

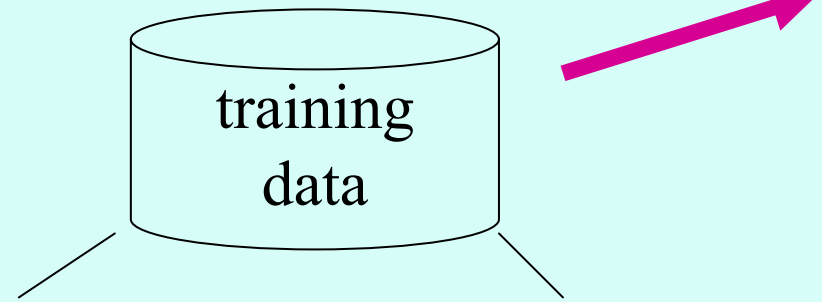
- metody symboliczne (drzewa i reguły decyzyjne),
- metody oparte na logice matematycznej (ILP),
- sztuczne sieci neuronowe,
- metody k-najbliższych sąsiadów,
- klasyfikacja bayesowska (Naive Bayes),
- analiza dyskryminacyjna (statystyczna),
- metody wektorów wspierających,
- regresja logistyczna,
- klasyfikatory genetyczne.
- **...**

Drzewa decyzyjne

Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low



Reguły decyzyjne

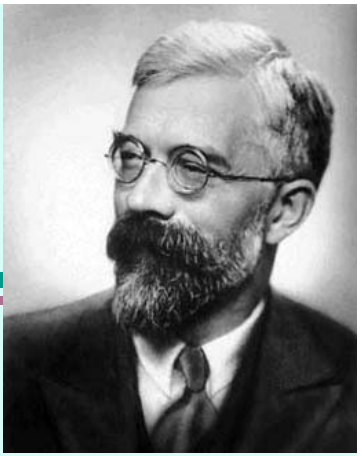


Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low

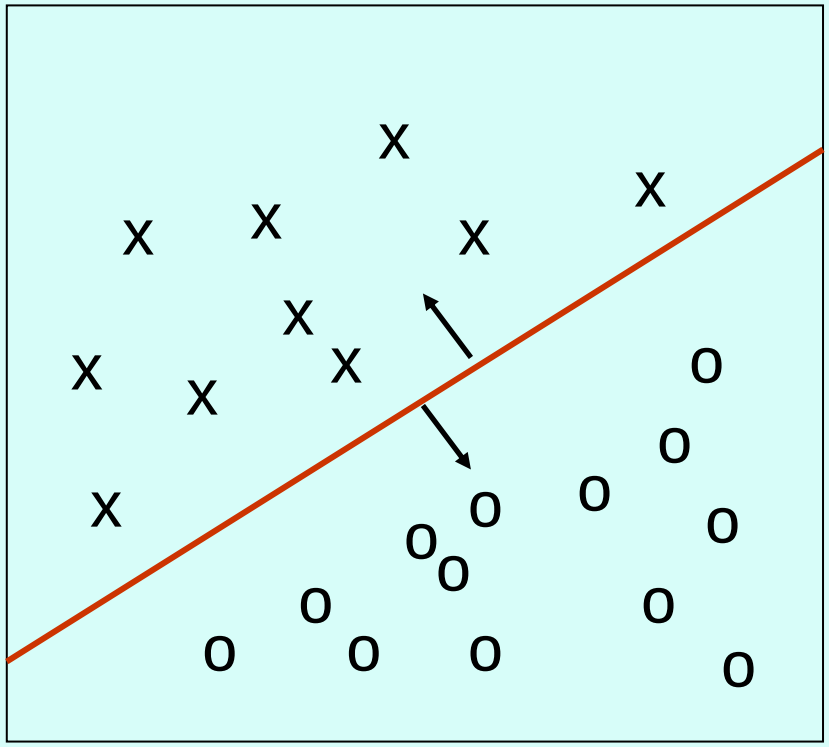
Classification / Learning algorithm

if age < 31
or Car Type =Sports
then Risk = High

Klasyfikacja – podejście statystyczne



sir Ronald Fisher



- Binarna klasyfikacja (uogólnienie na więcej klas)
- Poszukiwanie przybliżenia „granicy decyzyjnej” – ang. decision boundary
- Obserwacje ponad linią przydziel do klasy ‘x’
- Obserwacje pod linią przydziel do klasy ‘o’
- Przykłady: Fisher-owska analiza dyskryminacyjna, SVM, ANN

Algorytm k najbliższych sąsiadów

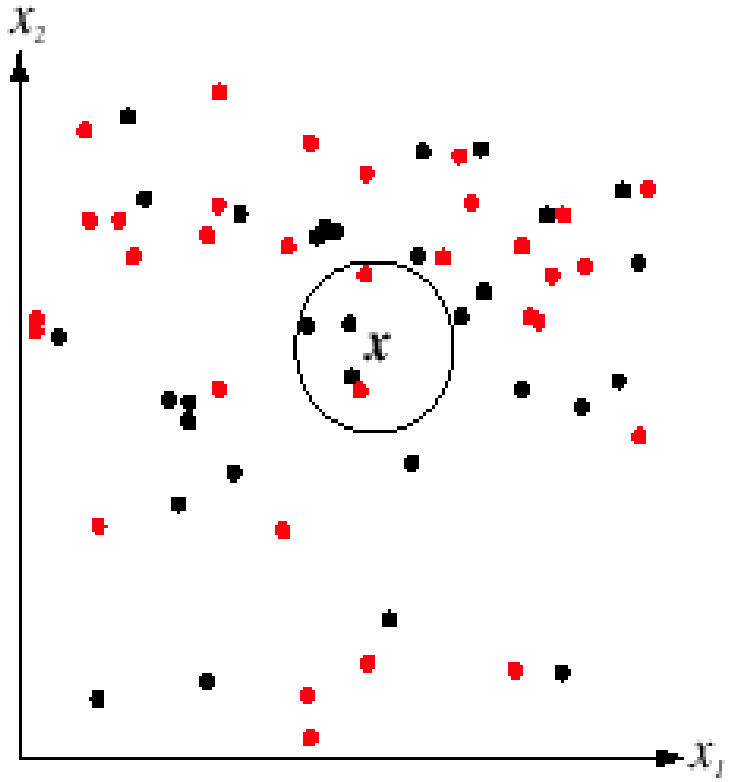


FIGURE 4.15. The k-nearest-neighbor query starts at the test point **x** and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point **x** would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Poszukiwanie i ocena wiedzy klasyfikacyjnej

- **Perspektywy odkrywania wiedzy**
  - **Predykcja klasy (-fikacji)–** przewidywanie przydziału nowych obiektów do klas / reprezentacja wiedzy wykorzystywana jako tzw. **klasyfikator** (ocena zdolności klasyfikacyjnej – na ogół jedno wybrane kryterium).
  - **Opis klasyfikacji obiektów –** wyszukiwanie wzorców charakteryzujących właściwości danych i prezentacja ich użytkownikowi w zrozumiałej formie (ocena wielokryterialna i bardziej subiektywna)**.**

Tworzenie i ocena klasyfikatorów

Jest procesem trzyetapowym:

1. Konstrukcja modelu w oparciu o zbiór danych wejściowych (przykłady uczące).

   Przykładowe modele - klasyfikatory:

   - drzewa decyzyjne,
   - reguły (IF .. THEN ..),
   - sieci neuronowe.

2. Ocena modelu (przykłady testujące)
3. Użycie modelu (np. klasyfikowanie nowych faktów lub interpretacja regularności)

Kryteria oceny metod klasyfikacyjnych

- **Trafność klasyfikacji (Classification / Predictive accuracy)**
- Szybkość i skalowalność:
  - czas uczenia się,
  - szybkość samego klasyfikowania
- Odporność (Robustness)
  - szum (noise),
  - missing values,
- Zdolności wyjaśniania: np. drzewa decyzyjne vs. sieci neuronowe
- Złożoność struktury, np.
  - rozmiar drzew decyzyjnego,
  - miary oceny reguły

Trafność klasyfikowania

- Użyj przykładów testowych nie wykorzystanych w fazie indukcji klasyfikatora:
  - N_t – liczba przykładów testowych
  - N_c – liczba poprawnie sklasyfikowanych przykładów testowych
- Trafność klasyfikowania (ang. classification accuracy):

$$\eta = \frac{N_c}{N_t}$$

- Alternatywnie błąd klasyfikowania.

$$\varepsilon = \frac{N_t - N_c}{N_t}$$

Inne możliwości analizy:

- macierz pomyłek (ang. confusion matrix),
- koszty pomyłek i klasyfikacja binarna,
- miary Sensitivity i Specificity / krzywa ROC

Więcej o ocenie trafności /błędu

- Podejście klasyfikacji statystycznej
- Minimalizacja tzw. funkcji straty – prediction risk:

$$R(f) = E_{xy} L(y, f(x))$$

- where $L(y, \hat{y})$ is a loss or cost of predicting value $\hat{y}$ when the actual value is y and E is expected value over the joint distribution of all (x,y) for data to be predicted.
- Typowa klasyfikacja → zero-one loss function

$$L(y, \hat{y}) = \begin{cases} 0 & if\ y = f(y) \\ 1 & if\ y \neq f(y) \end{cases}$$

Macierz pomyłek

- Analiza pomyłek w przydziale do różnych klas przy pomocy tzw. macierz pomyłek (ang. *confusion matrix*)
- Macierz $r \times r$, gdzie wiersze odpowiadają poprawnym klasom decyzyjnym, a kolumny decyzjom przewidywanym przez klasyfikator; na przecięciu wiersza i oraz kolumny j - liczba przykładów $n\text{-}ij$ należących oryginalnie do klasy i-tej, a zaliczonej do klasy j-tej

Przykład:

	Przewidywane klasy decyzyjne		
Oryginalne klasy	K_1	K_2	K_3
K_1	50	0	0
K_2	0	48	2
K_3	0	4	46

Klasyfikacja binarna

- Niektóre problemy → jedna z klas posiada szczególne znaczenie, np. diagnozowanie poważnej choroby → klasyfikacja binarna.

Oryginalne klasy	Przewidywane klasy decyzyjne	
	Pozytywna	Negatywna
Pozytywna	*TP*	*FN*
Negatywna	*FP*	*TN*

- Nazewnictwo (inspirowane medycznie):
  - *TP* (ang. *true positive*) – liczba poprawnie sklasyfikowanych przykładów z wybranej klasy (ang. *hit*),
  - *FN* (ang. *false negative*) – liczba błędnie sklasyfikowanych przykładów z tej klasy, tj. decyzja negatywna podczas gdy przykład w rzeczywistości jest pozytywny (błąd pominięcia - z ang. *miss*),
  - *TN* (ang. *true negative*) – liczba przykładów poprawnie nie przydzielonych do wybranej klasy (poprawnie odrzuconych z ang. *correct rejection*),
  - *FP* (ang. *false positive*) – liczba przykładów błędnie przydzielonych do wybranej klasy, podczas gdy w rzeczywistości do niej nie należą (ang. *false alarm*).

Miary stosowane w analizie klasyfikacji binarnej

- Dodatkowe miary oceny rozpoznawania wybranej klasy:
  - **Wrażliwość / czułość (ang. *sensitivity*) = *TP* / (*TP*+*FN*),**
  - **Specyficzność (ang. *specificity*) = *TN* / (*FP*+*TN*).**
- Inne miary:
  - ***False-positive rate* = *FP* / (*FP*+*TN*), czyli 1 – specyficzność.**
- Wnikliwszą analizę działania klasyfikatorów binarnych dokonuje się w oparciu o analizę krzywej **ROC**, ang. *Receiver Operating Characteristic*).

Oryginalne klasy	Przewidywane klasy decyzyjne	
	Pozytywna	Negatywna
Pozytywna	*TP*	*FN*
Negatywna	*FP*	*TN*

Analiza macierzy… spróbuj rozwiązać…

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

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

Co przewidywano

Rzeczywista Klasa	**1**	**0**
1	60	30
0	80	20

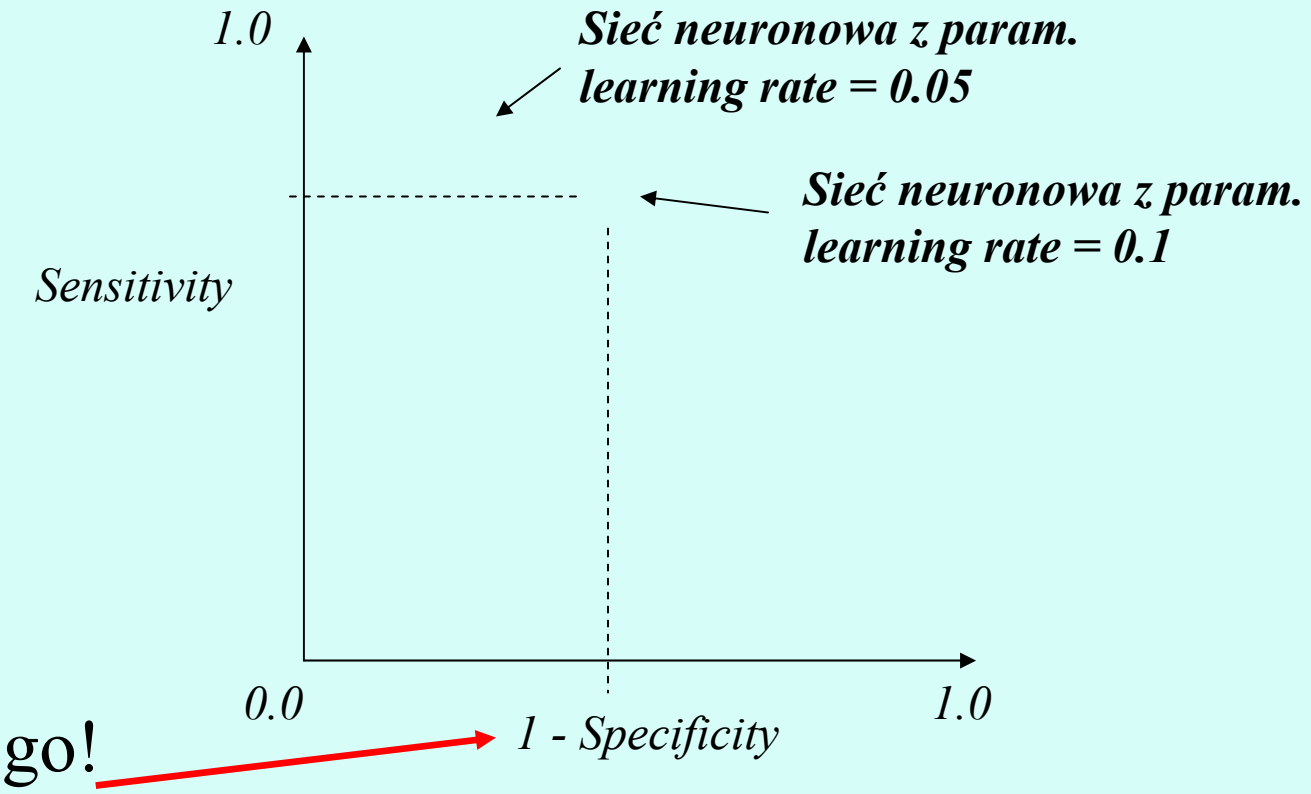
60+30 = 90 przykładów w danych należało do Klasy 1

80+20 = 100 przykładów było w Klasy 0

90+100 = 190 łączna liczba przykładów

Analiza krzywej ROC

Każda technika budowy klasyfikatora może być scharakteryzowana poprzez pewne wartości miar ‘sensitivity’ i ‘specificity’. Graficznie można je przedstawić na wykresie ‘sensitivity’ vs. 1 – ‘specificity’.



Dlaczego!

Probabilistyczne podstawy ROC

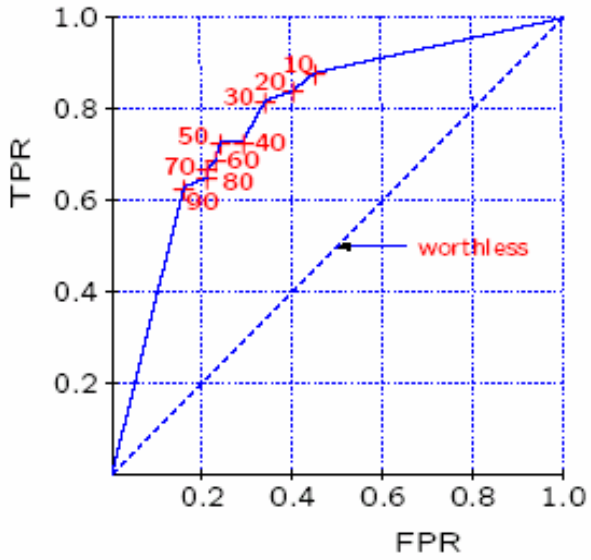
- Output of probabilistic classifier:

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

may not yield the best performance

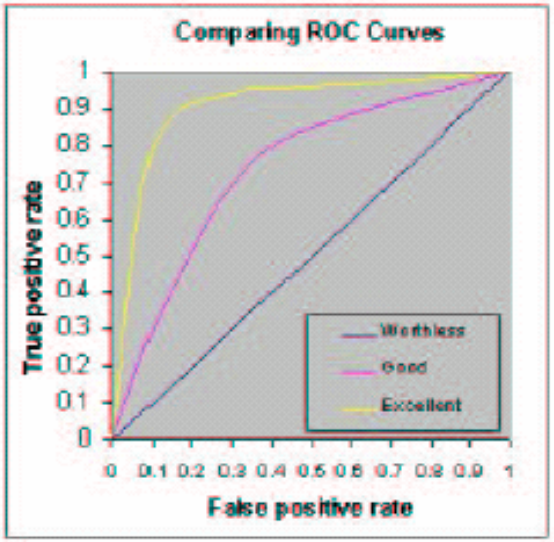
- Alternative: Receiver Operating Characteristic (ROC): determine threshold d, such that

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



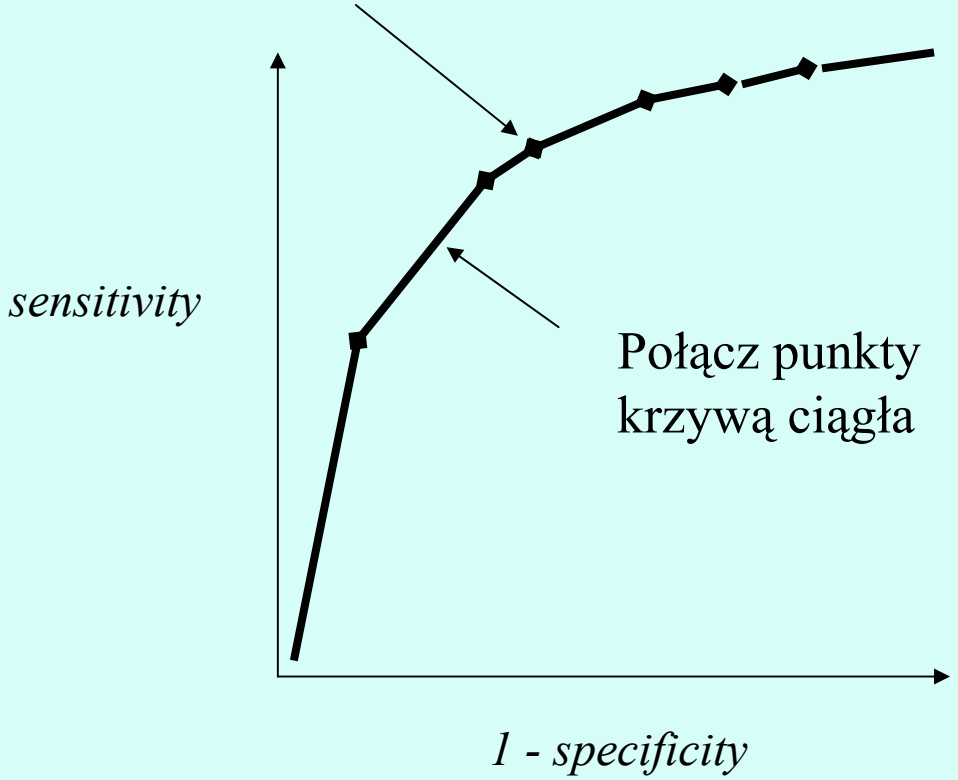
When comparing various techniques:

- 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



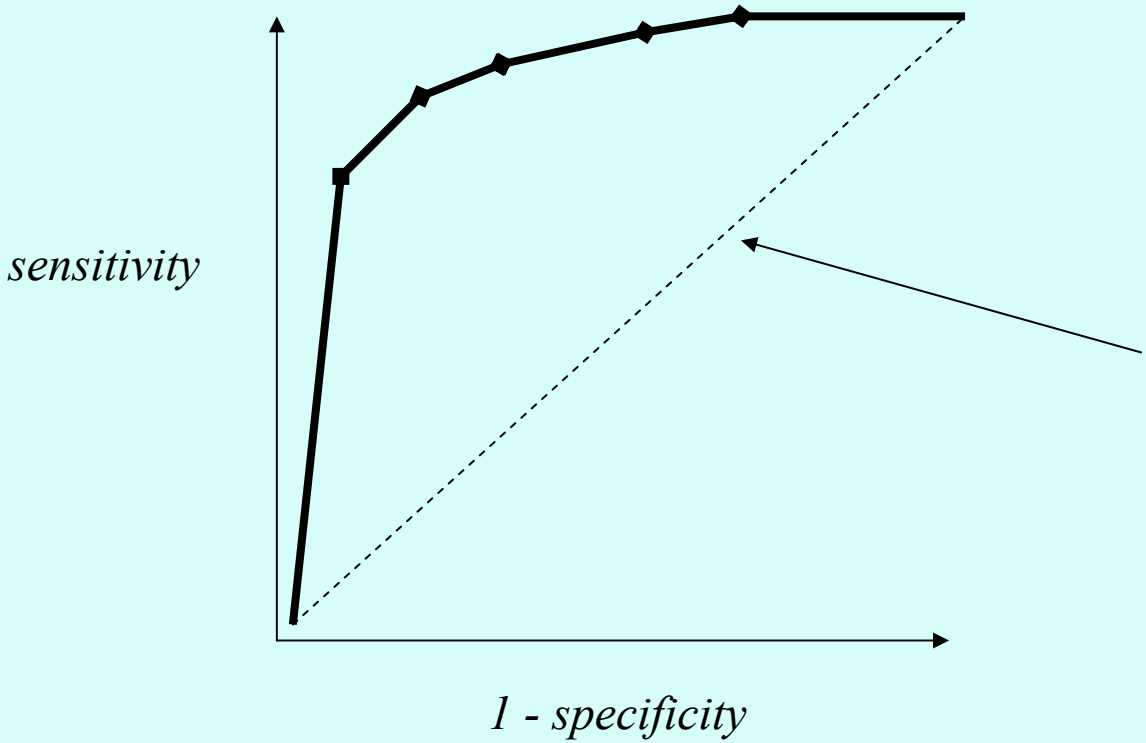
ROC - analiza

Algorytm może być parametryzowany, i w rezultacie otrzymuje się serie punktów odpowiadających doborowi parametrów



Wykres nazywany ‘krzywą’ ROC.

Krzywa ROC



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,..

Niezrównoważone klasy – pamiętaj o innych podejściach i miarach oceny (G-means, F, AUC)

- Czasami klasy mają mocno niezrównoważoną liczebność
  - Attrition prediction: 97% stay, 3% attrite (in a month)
  - medical diagnosis: 90% healthy, 10% disease
  - eCommerce: 99% don't buy, 1% buy
  - Security: >99.99% of Americans are not terrorists
- Podobna sytuacja dla problemów wieloklasowych.
- Skuteczność rozpoznawania klasy większościowej 97%, ale bezużyteczne dla klasy mniejszościowej o specjalnym znaczeniu.

Dane niezrównoważone - Sensitivity

Data set	Standard classifier	Under-sampling	Over-sampling	New filtering
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 poprawia wartość Sensitivity w porównaniu do innych metod przetwarzania wstępnego danych.

Porównywanie działania klasyfikatorów na ROC

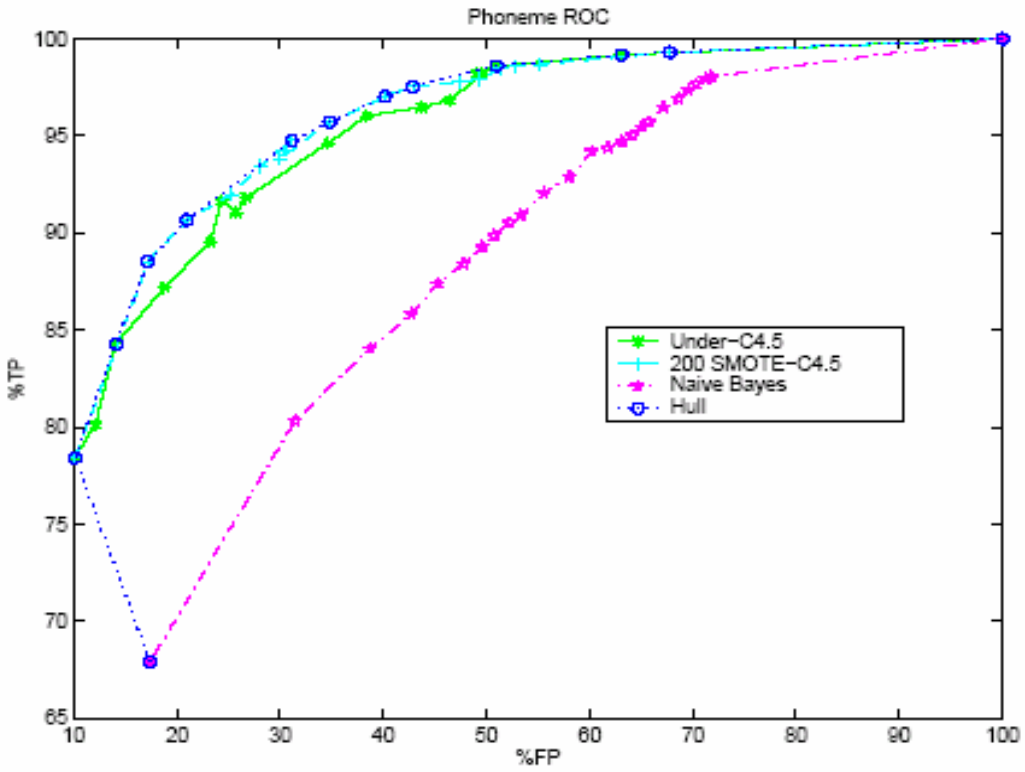


Figure 7: Phoneme. Comparison of SMOTE-C4.5, Under-C4.5, and Naive Bayes. SMOTE-C4.5 dominates over Naive Bayes and Under-C4.5 in the ROC space. SMOTE-C4.5 classifiers are potentially optimal classifiers.

Za artykułem N.Chawla et al. nt. SMOTE

Inne miary budowane w oparciu o macierze binarne

	Domain	Plot	Explanation
Lift chart	Marketing	TP Subset size	TP (TP+FP)/ (TP+FP+TN+FN)
ROC curve	Communications	TP rate FP rate	TP/(TP+FN) FP/(FP+TN)
Recall-precision curve	Information retrieval	Recall Precision	TP/(TP+FN) TP/(TP+FP)

Jak szacować wiarygodnie ?

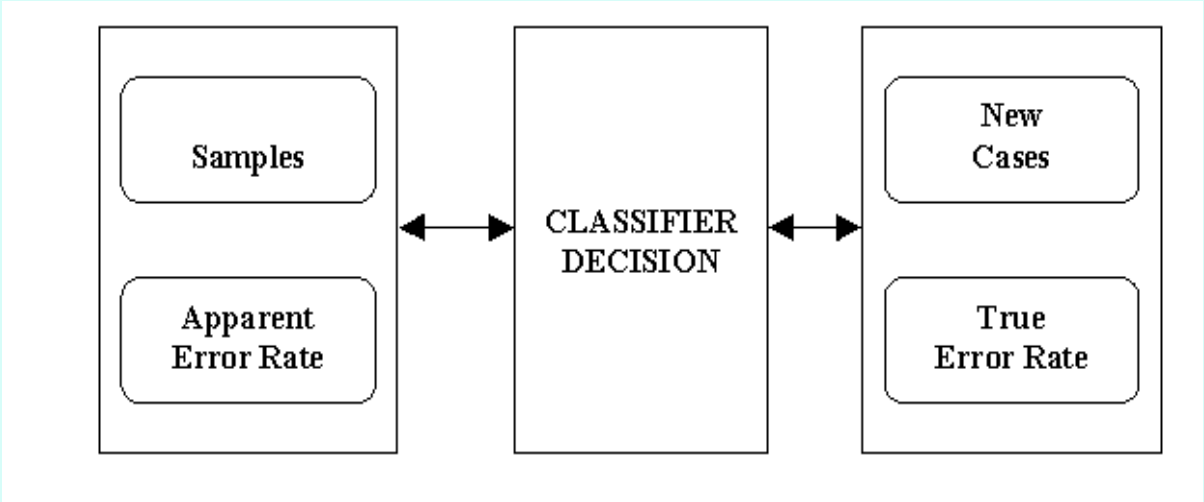
- **Zależy od perspektywy użycia wiedzy:**
  - **Predykcja klasyfikacji albo opisowa**
- **Ocena na zbiorze uczącym nie jest wiarygodna jeśli rozważamy predykcję nowych faktów!**
  - Nowe obserwacje najprawdopodobniej nie będą takie same jak dane uczące!
  - aparent vs. true error …
- Problem przeuczenia (ang. overfiting)
  - Nadmierne dopasowanie do specyfiki danych uczących powiązane jest najczęściej z utratą zdolności uogólniania (ang. generalization) i predykcji nowych faktów!

Podejścia teoretyczne

- **Obliczeniowa teoria uczenia się (COLT)**
  - PAC model (Valiant)
  - Wymiar Vapnik Chervonenkis → VC Dimension
- Pytania o ogólne prawa dotyczące procesu uczenia się klas pewnych funkcji z przykładów - rozkładów prawdopodobieństwa.
- Silne założenia i ograniczone odniesienia do problemów praktycznych (choć SVM?).

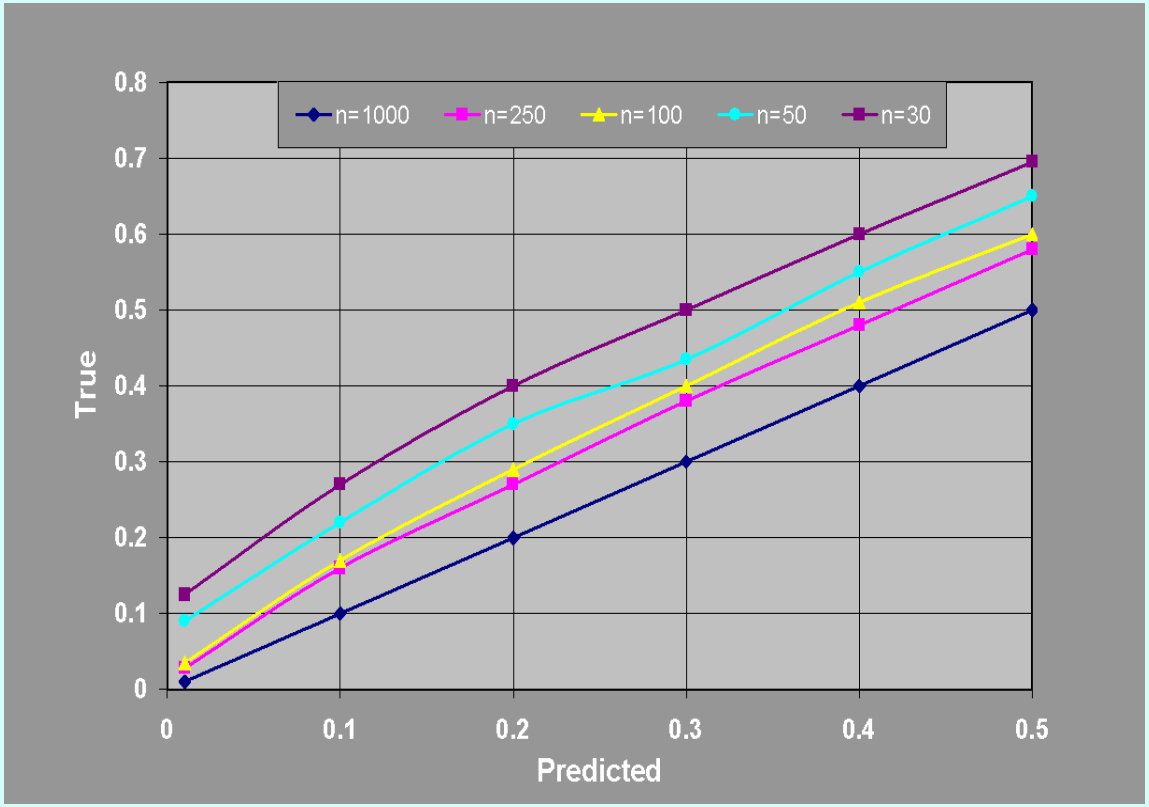
Podejście empiryczne

- Zasada „Train and test”
- Jeśli nie ma podziału zadanego „przez nauczyciela”, to wykorzystaj losowe podziały.
- Nadal pytanie jak szacować wiarygodnie?



True vs. aparent accuracy i liczba przykładów testowych

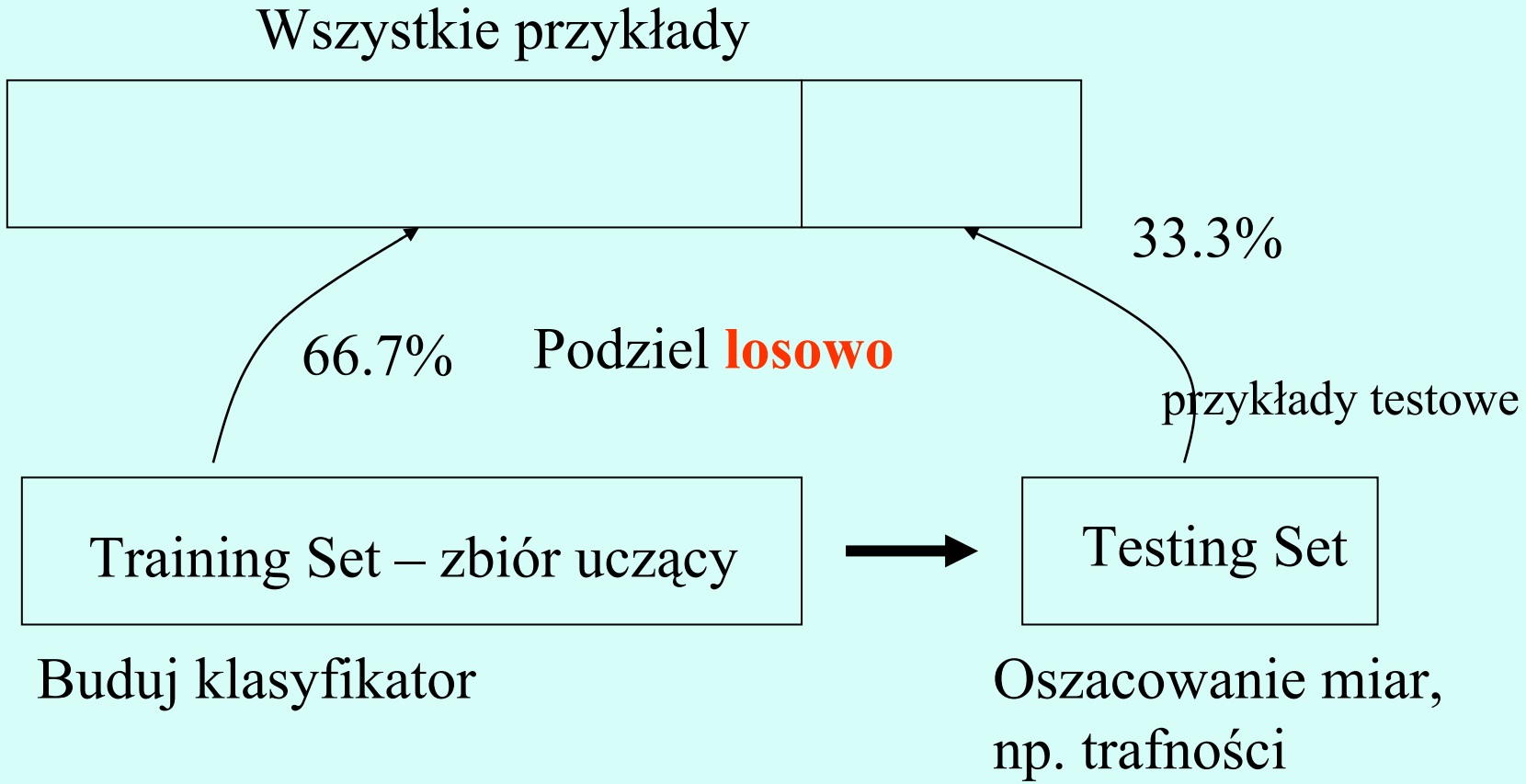
- Za S.Weiss, C.Kulikowski – podsumowanie eksperymentów ze sztucznymi danymi – ile testowych przykładów z zadanego rozkładu jest potrzebne w odpowiednim podziale losowym.



Empiryczne metody estymacji

- **Techniki podziału: „hold-out”**
  - Użyj dwóch niezależnych zbiorów: uczącego (2/3), testowego (1/3)
  - Jednokrotny podział losowy stosuje się dla dużych zbiorów (hold-out)
- **„Cross-validation” - Ocena krzyżowa**
  - Podziel losowo dane w *k* podzbiorów (równomierne lub warstwowe)
  - Użyj *k-1* podzbiorów jako części uczącej i pozostałej jako testującej (*k*-fold cross-validation).
  - Oblicz wynik średni.
  - Stosowane dla danych o średnich rozmiarach (najczęściej k = 10) Uwaga opcja losowania warstwowego (ang. stratified sampling).
- **leaving-one-out**
  - Dla małych rozmiarów danych.
  - „Leaving-one-out” jest szczególnym przypadkiem, dla którego liczba iteracji jest równa liczbie przykładów
- **Bootstrapping** - specyficzne losowanie ze zwracaniem

Jednokrotny podział (hold-out) – duża liczba przykładów (> tysięcy)

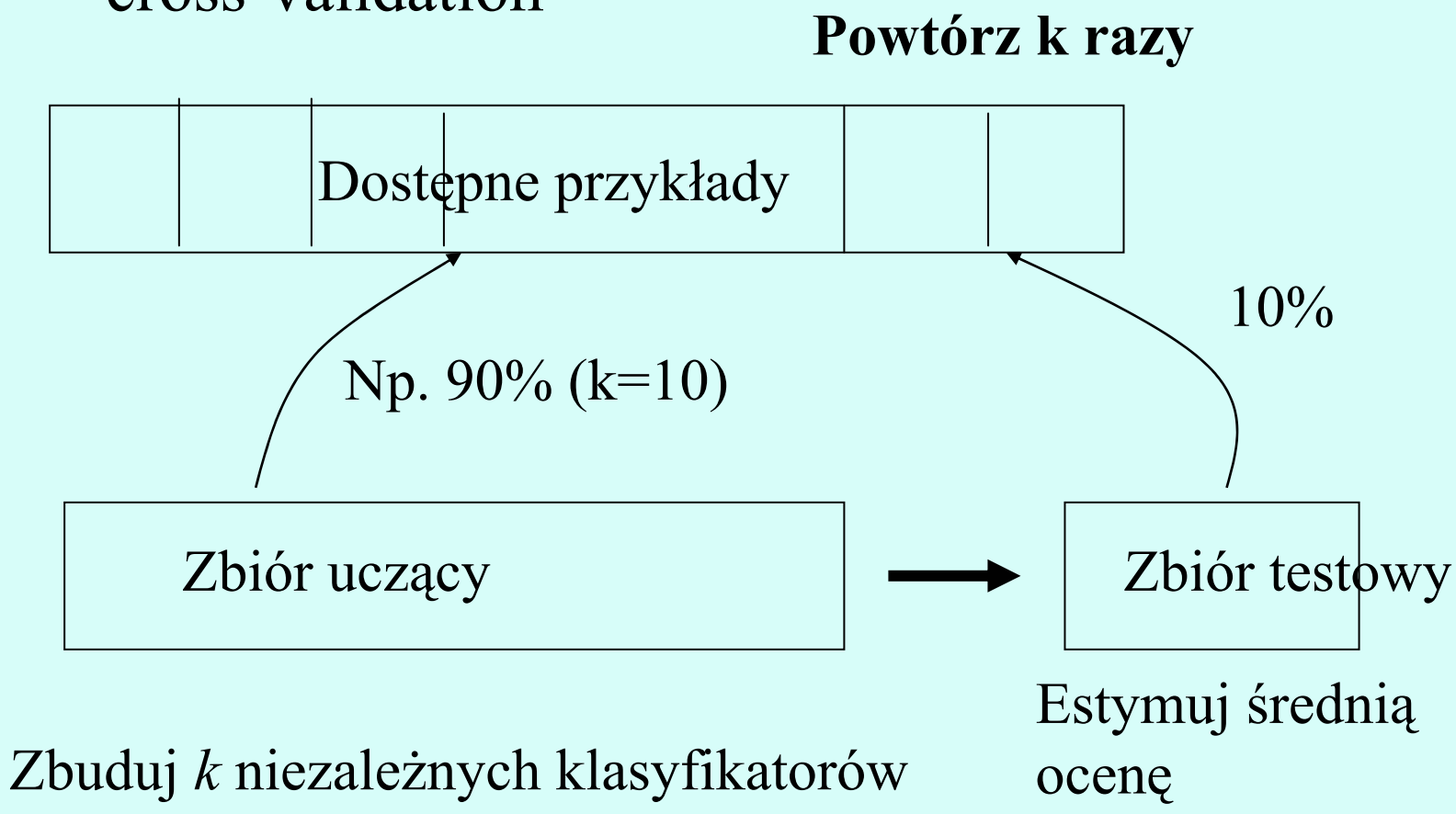


Repeated holdout method

- Estymata miar może być bardziej wiarygodna, jeśli powtarzamy proces z różnymi podzbiorami
  - W każdej iteracji pewna część jest losowo wybierana dla uczenia (najlepiej w sposób warstwowy)
    - Podziały muszą być różne
  - Estymata miary średnia z wszystkich iteracji
- Nazywamy → the *repeated holdout* method
- Trudność → możliwość nakładania się (over-lapping) zbiorów w kolejnych iteracjach

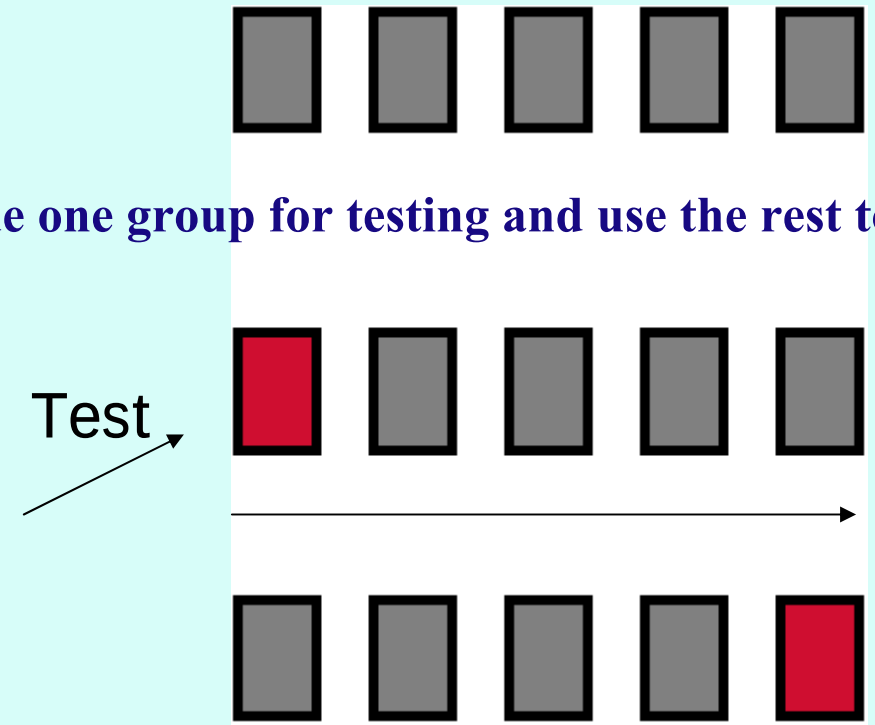
Mniejsza liczba przykładów (od 100 do kilku tysięcy)

* cross-validation



Cross-validation dokładniej:

- Randomly split data into k groups of the same size
-
-
- Hold aside one group for testing and use the rest to build model
-
- Repeat



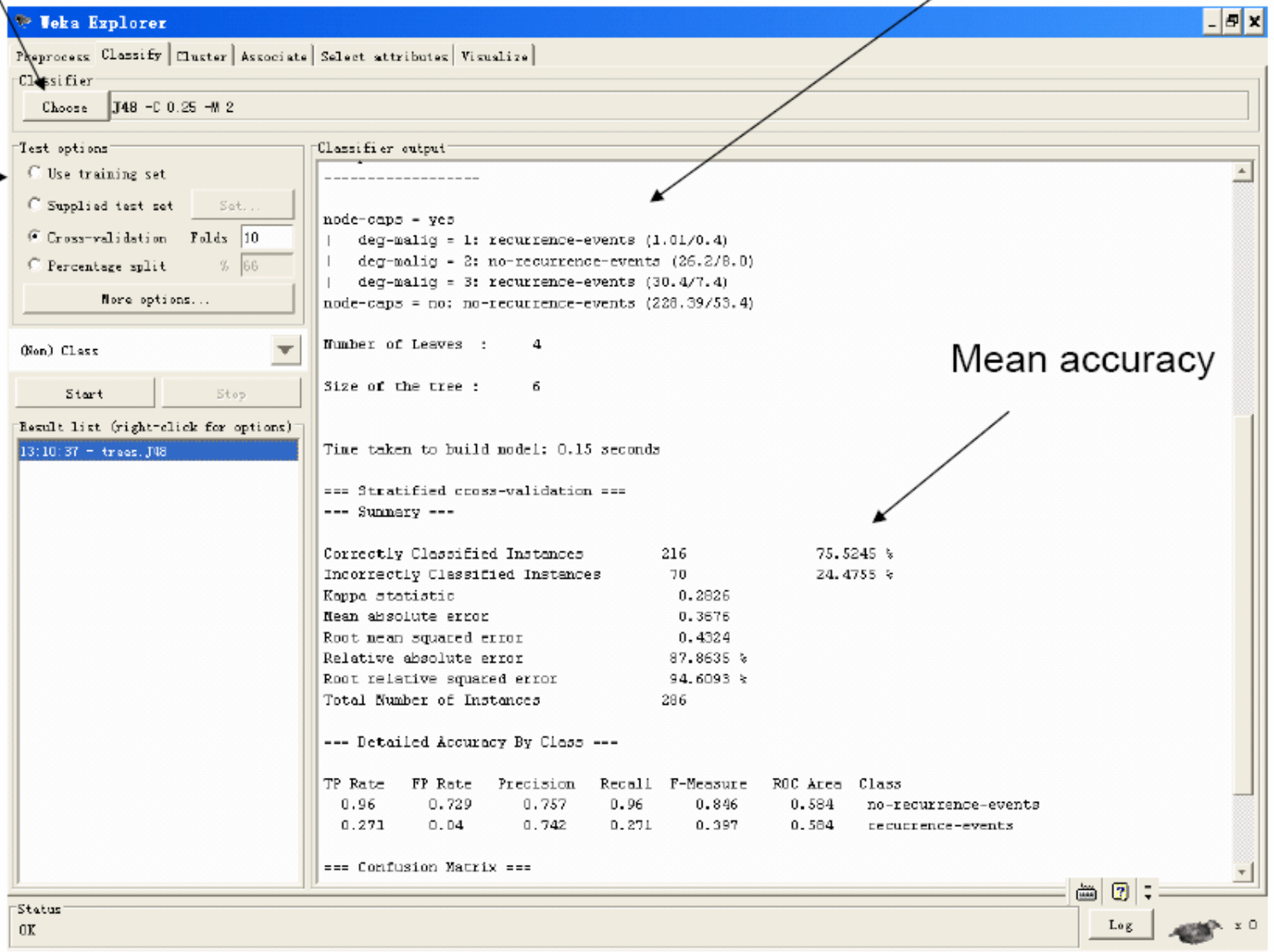
36

Classifier

WEKA Explorer

Decision Trees

Testing data

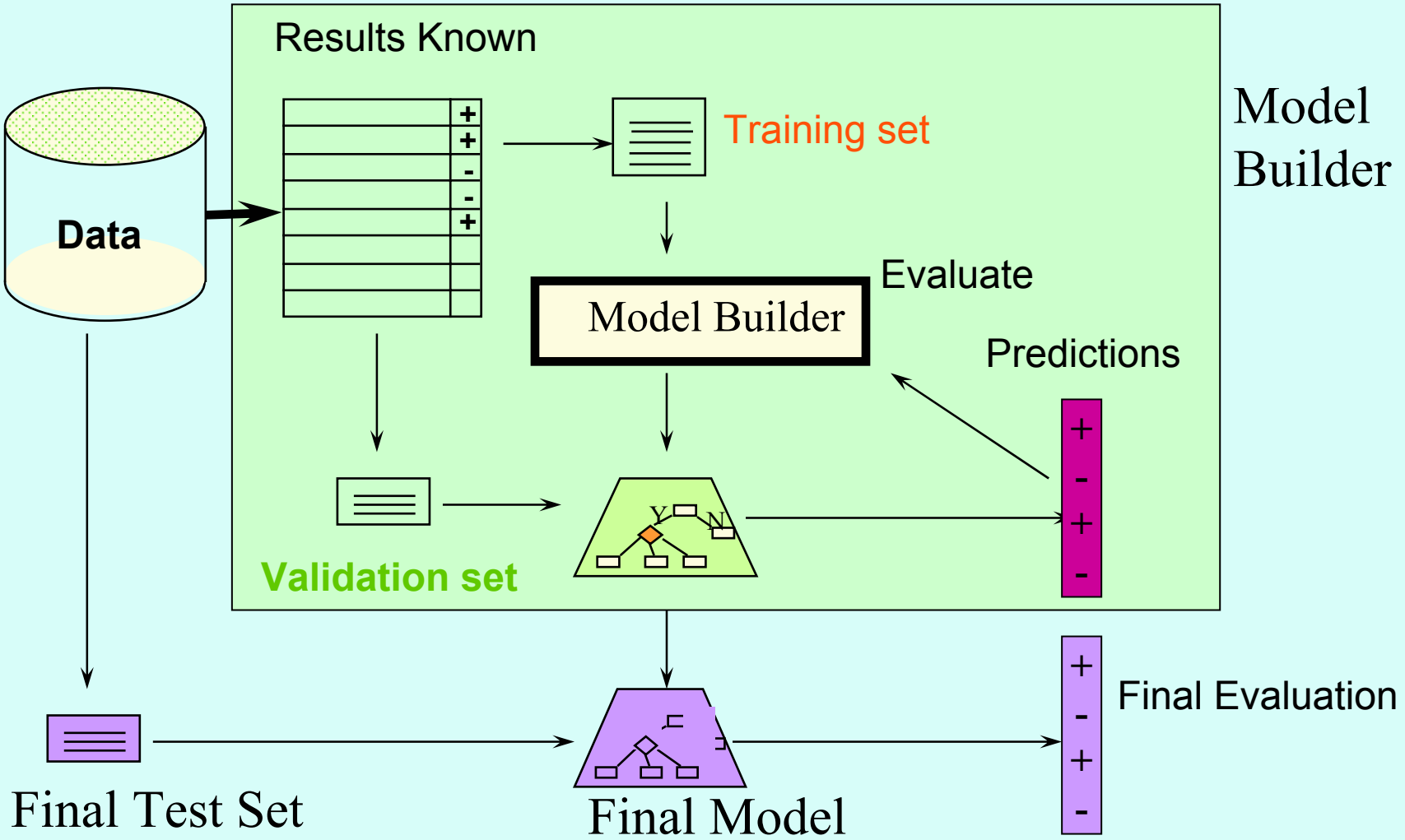


Mean accuracy

Uwagi o 10 fold cross-validation

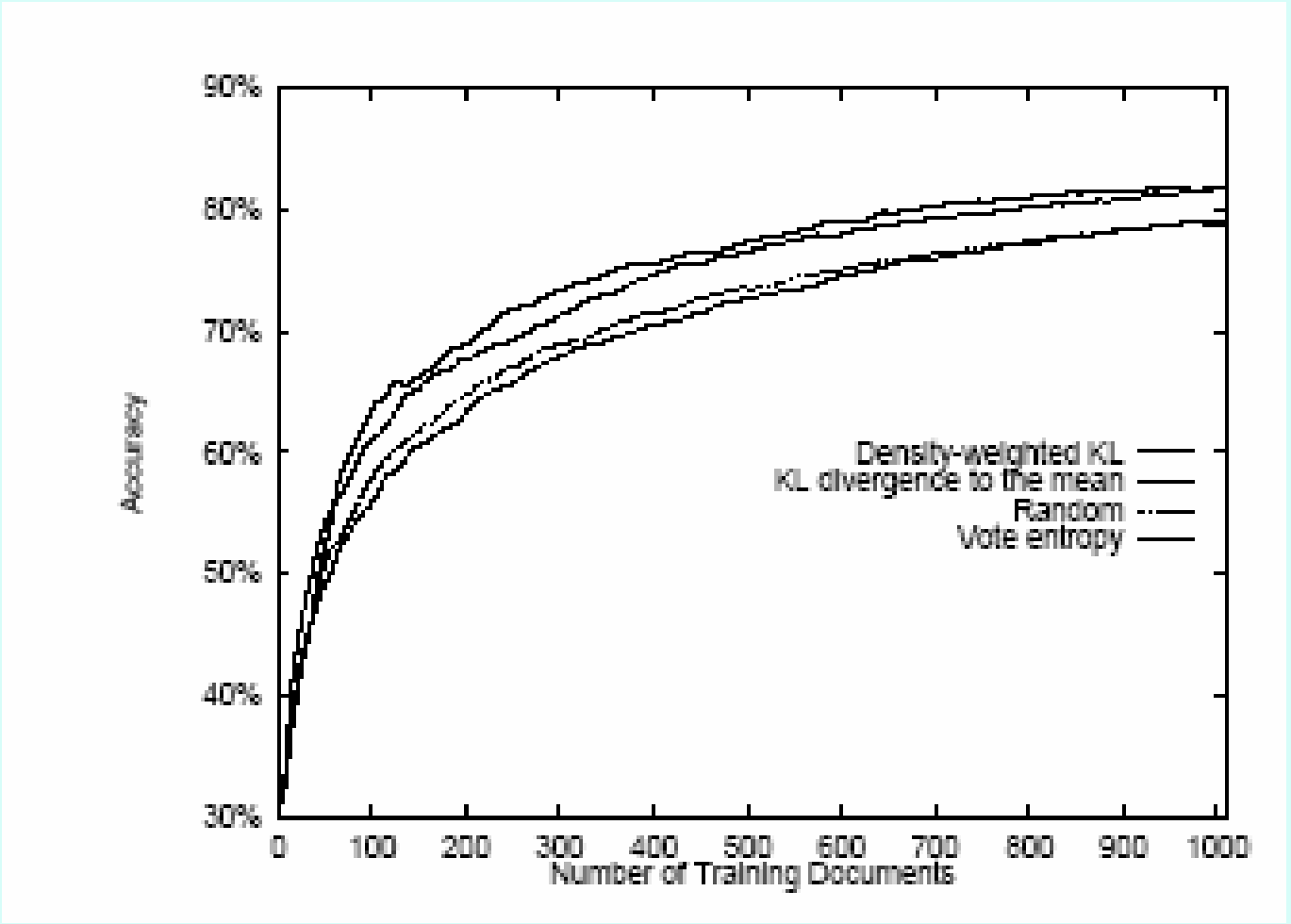
- Stosuj wersję: stratified ten-fold cross-validation
  - Warstwowe losowanie z uwagi na rozkład liczności klas
- Dlaczego 10? Doświadczenie badaczy głównie eksperymentalne (zwłaszcza związane CART)
- Stratification – warstwowość ogranicza wariancje estymaty błędu/trafności!
- Lepsza wersja: repeated stratified cross-validation"
  - np. 10-fold cross-validation is powtórzone kilka razy (z innym ziarnem rozkładu prawdopodobieństwa) i wynik średni z wielu powtórzeń.
  - Pamiętaj o obserwacji odchylenia standardowego
- Czasami 10-cv jest kilkakrotnie (3-5) powtarzany dla lepszej stabilizacji estymaty i zmniejszenia jej wariancji

Uwaga na zaawansowane strojenie parametrów algorytmu – podejście „Train, Validation, Test split”



Krzywe uczenia się - learning curve

Klasyfikacja tekstów przez Naive Bayes 20 Newsgroups dataset -



[McCallum & Nigam, 1998]

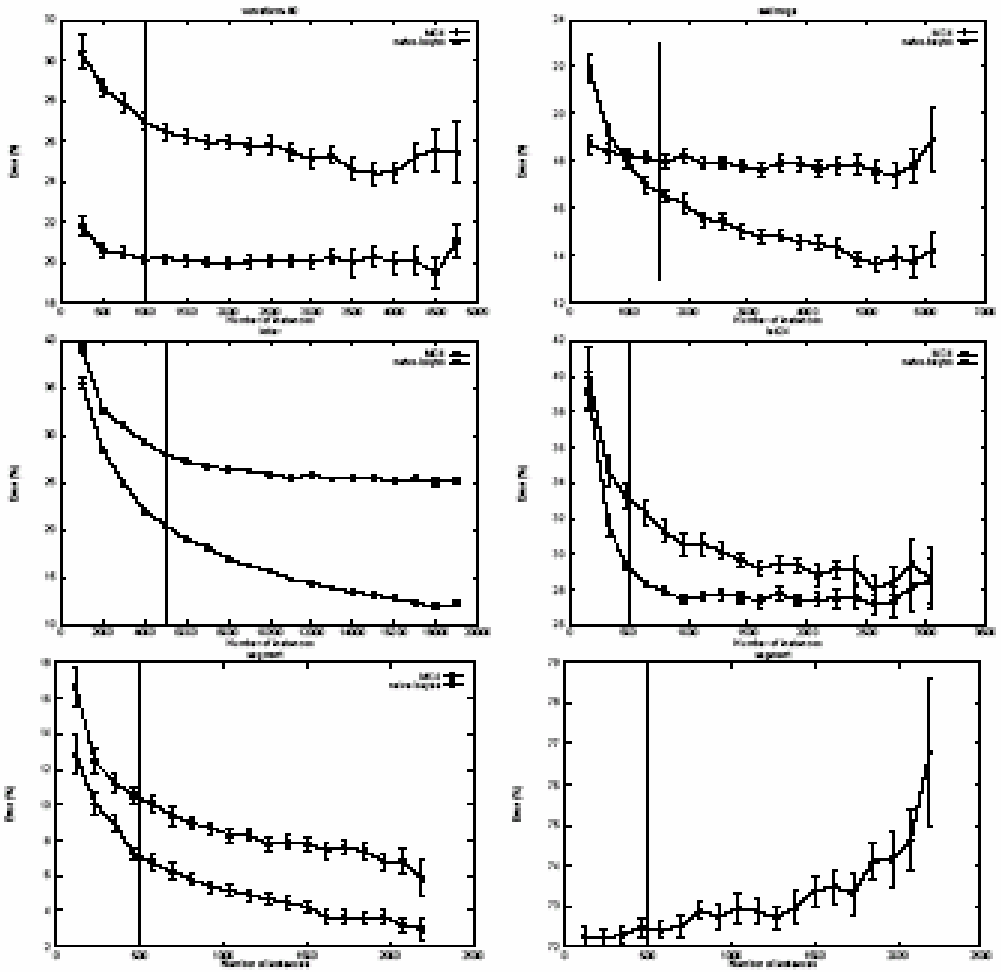
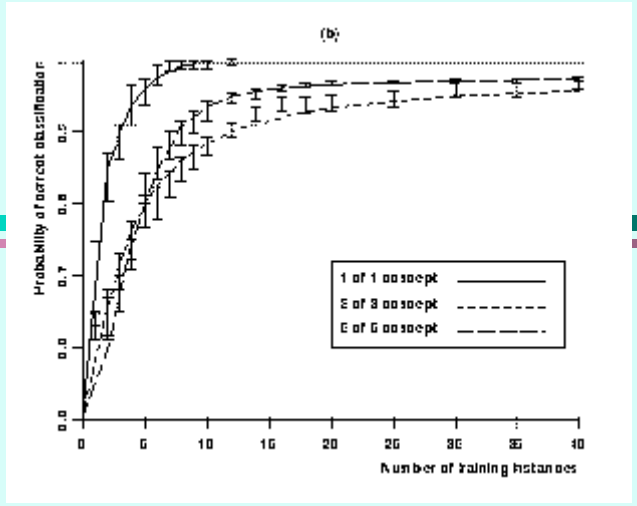
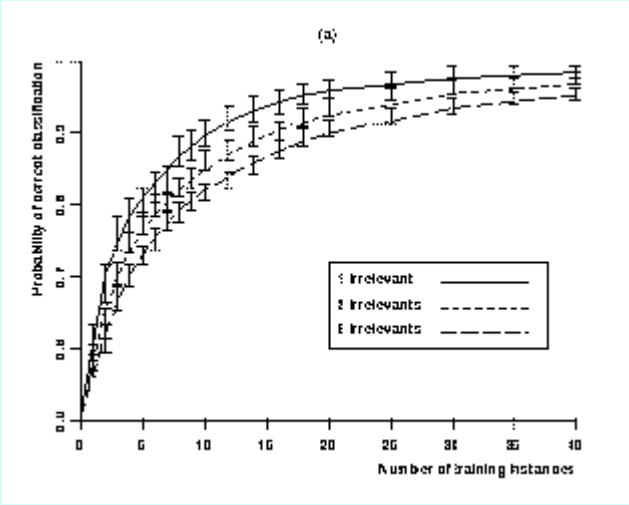


Figure 4. Learning curves for selected datasets showing different behaviors of MC4 and Naive-Bayes. Waveform represents stabilization at about 3,000 instances; satimage represents a cross-over as MC4 improves while Naive-Bayes does not; letter and segment (left) represent continuous improvements, but at different rates in letter and similar rates in segment (left); LED24 represents a case where both algorithms achieve the same error rate with large training sets; segment (right) shows MC4(1), which exhibited the surprising behavior of degrading as the training set size grew (see text). Each point represents the mean error rate for 20 runs for the given training set size as tested on the holdout sample. The error bars show one standard deviation of the estimated error. Each vertical bar shows the training set size we chose for the rest of the paper following our desiderata. Note (e.g., in waveform) how small training set sizes have high standard deviations for the estimates because the training set is small and how large training set sizes have high standard deviations because the test set is small.





- Przykład prezentacji z artykułu Bauer, Kohavi nt. porównania różnych rozwiązań w klasyfikatorach złożonych.

Porównywanie wielu klasyfikatorów

- Często należy porównać dwa klasyfikatory
- Uwaga: porównanie z niezależnością od danych?
  - Generatory losowe
  - Rzeczywiste dane (problem dependent)
- Oszacuj 10-fold CV estimates.
- Trudność: wariancja oszacowania.
- Możesz oczywiście zastosować „repeated CV”.
- Ale jak wiarygodnie ustalić konkluzję – który jest lepszy?

Porównywanie klasyfikatorów

- Jak oceniać skuteczność klasyfikacyjną dwóch różnych klasyfikatorów na tych samych danych?
- Ograniczamy zainteresowanie wyłącznie do trafności klasyfikacyjnej – oszacowanie techniką 10-krotnej oceny krzyżowej (ang. *k-fold cross validation*).
- Zastosowano dwa różne algorytmy uczące *AL*1 i *AL*2 do tego samego zbioru przykładów, otrzymując dwa różne klasyfikatory *KL*1 i *KL*2. Oszacowanie ich trafności klasyfikacyjnej (10-fcv):
  - klasyfikator *KL*1 $\rightarrow$ 86,98%
  - klasyfikator *KL*2 $\rightarrow$ 87,43%.
- Czy uzasadnione jest stwierdzenie, że klasyfikator *KL*2 jest skuteczniejszy niż klasyfikator *KL*1?

Analiza wyniku oszacowania trafności klasyfikowania

Podział	Kl_1	Kl_2
1	87,45	88,4
2	86,5	88,1
3	86,4	87,2
4	86,8	86
5	87,8	87,6
6	86,6	86,4
7	87,3	87
8	87,2	87,4
9	88	89
10	85,8	87,2
Srednia	**86,98**	**87,43**
Odchylenie	**0,65**	**0,85**

- Test statystyczny (t-Studenta dla par zmiennych/zależnych)
- H0 : ?
- *t*emp = 1,733 ($p = 0{,}117$) ???
- ALE !!! W art. naukowych zastosuj odpowiednie poprawki

przy wykonaniu testu (kwestia naruszenia założeń co do rozkładu *t*).

- Nadeau i Bengio 2003 proponują modyfikacje testu!

Przykład zastosowania „paired t-test” $\alpha = 0{,}05$

Table 1. Comparison of classification accuracies [%] obtained by the single MODLEM based classifier and the bagging approach

Name of dataset	Single MODLEM	Bagging - with different T 3	5	7	10
bank	93.81 ± 0.94	95.05 ± 0.91	94.95 ± 0.84	95.22 ± 1.02	93.95* ± 0.94
buses	97.20 ± 0.94	98.05* ± 0.97	99.54 ± 1.09	97.02* ± 1.15	97.45* ± 1.13
zoo	94.64 ± 0.67	93.82* ± 0.68	93.89* ± 0.71	93.47 ± 0.73	93.68 ± 0.70
hepatitis	78.62 ± 0.93	82.00 ± 1.14	84.05 ± 1.1	81.05 ± 0.97	84.0 ± 0.49
iris	94.93 ± 0.5	95.13* ± 0.46	94.86* ± 0.54	95.06* ± 0.53	94.33* ± 0.59
automobile	85.23 ± 1.1	82. 98 ± 0.86	83.0 ±0.99	82.74 ±0.9	81.39 ± 0.84
segmentation	85.71 ± 0.71	86.19* ± 0.82	87.62 ± 0.55	87.61 ± 0.46	87.14 ± 0.9
glass	72.41 ± 1.23	68.5 ± 1.15	74.81 ± 0.94	74.25 ± 0.89	76.09 ± 0.68
bricks	90.32* ± 0.82	90.3 * ± 0.54	89.84* ± 0.65	91.21* ± 0.48	90.77* ± 0.72
vote	92.67 ± 0.38	93.33* ± 0.5	94.34 ± 0.34	95.01 ± 0.44	96.01 ± 0.29
bupa	65.77 ± 0.6	64.98* ± 0.76	76.28 ± 0.44	70.74 ± 0.96	75.69 ± 0.7
election	88.96 ± 0.54	90.3 ± 0.36	91.2 ± 0.47	91.66 ± 0.34	90.75 ± 0.55
urology	63.80 ± 0.73	64.8 ± 0.83	65.0 ± 0.43	67.40 ± 0.46	67.0 ± 0.67
german	72.16 ± 0.27	73.07* ± 0.39	76.2 ± 0.34	75.62 ± 0.34	75.75 ± 0.35
crx	84.64 ± 0.35	84.74* ± 0.38	86.24 ± 0.39	87.1 ± 0.46	89.42 ± 0.44
pima	73.57 ± 0.67	75.78* ± 0.6	74.35* ± 0.64	74.88 ± 0.44	77.87 ± 0.39

Jeden z klasyfikatorów złożonych vs. pojedynczy klasyfikator – z pracy J.Stefanowski

- z pracy T.Diettricha o ensembles

Table 2: Results on Five Domains (best error rate in **boldface**)

Task	Test set size	C4.5	200-fold bootstrap C4.5	200-fold random C4.5
Vowel	462	0.5758	0.5152	**0.4870***
Soybean	376	0.1090	**0.0984**	0.1090
Part-of-Speech	3060	0.0827	**0.0765**	0.0788[a]
NETtalk	7242	0.3000	0.2670***	**0.2500*****
Letter Recognition	4000	0.2010	0.0038***	**0.0000*****

Difference from C4.5 significant at $p < 0.05^*$, 0.001^{***}. [a]256-fold random.

Problem Statement

Comparison of two algorithms on m datasets:

datasets	classifier 1	classifier 2	difference (ℓ_j)
dataset 1	0.09	0.06	$+0.03$
dataset 2	0.25	0.36	-0.11
dataset 3	0.019	0.012	$+0.007$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

We treat the difference in errors on each dataset ℓ_j, $j = 1, \ldots, m$, as a separate observation.

The differences cannot be simply compared, because the errors on each dataset may have different scales (see example above).

Globalna ocena (2 alg. wiele zb. danych)

Wilcoxon test (sparowany test rangowy)

H0: nie ma różnicy oceny klasyfikatorów

1. Różnice oceny klasyfikatorów uporządkuj wg. wartości bezwzględnych i przypisz im rangi.
2. R+ suma rang dla sytuacji gdy klasyfikator 1 jest lepszy niż klasyfikator 2 // R- sytuacja odwrotna
3. Oblicz statystykę T = min{R+;R-}

   Rozkład T jest stabelaryzowany / prosta reguła decyzyjna
4. Dla odpowiednio dużej liczby *m* zbiorów danych można stosować przybliżenie z

$$z = \frac{\min\{R+;R-\} - \frac{1}{4}m(m-1)}{\sqrt{\frac{1}{24}m(m+1)(2m+1)}}$$

Example of AUC comparison [Demšar, 2006]

	C4.5	C4.5+m	difference	rank
adult (sample)	0.763	0.768	+0.005	3.5
breast cancer	0.599	0.591	−0.008	7
breast cancer wisconsin	0.954	0.971	+0.017	9
cmc	0.628	0.661	+0.033	12
ionosphere	0.882	0.888	+0.006	5
iris	0.936	0.931	−0.005	3.5
liver disorders	0.661	0.668	+0.007	6
lung cancer	0.583	0.583	0.000	1.5
lymphography	0.775	0.838	+0.063	14
mushroom	1.000	1.000	0.000	1.5
primary tumor	0.940	0.962	+0.022	11
rheum	0.619	0.666	+0.047	13
voting	0.972	0.981	+0.009	8
wine	0.957	0.978	+0.021	10

$$R_+ = 3.5 + 9 + 12 + 5 + 6 + 14 + 11 + 13 + 8 + 10 + 1.5 = 83$$

$$R_- = 7 + 3.5 + 1.5 = 12$$

$$z = -2.54 < -1.96 \quad \text{(Null hypothesis rejected.)}$$

Recommended in [Demšar, 2006].

Porównywanie wielu klasyfikatorów na wielu zbiorach danych

Most Popular Approach: "win-loss-tie" Matrix

Example [Dietterich, 1999]

	C4.5	AdaBoost C4.5	Bagged C4.5
Randomized C4.5	14 – 0 – 19	1 – 7 – 25	6 – 3 – 24
AdaBoost C4.5	11 – 0 – 22	1 – 8 – 24	
Bagged C4.5	17 – 0 – 6		

Often only *significant* win/losses counted. Significance usually checked with t-test.

Not wrong but does not lead to a statistical hypothesis testing.

Friedman Test [Demšar, 2006]

Procedure (N classifiers, m files)

- H_0 : All classifiers perform equally well.
  H_1 : Some of the classifiers perform better then the others.
- For each dataset, rank all classifiers according to their results; calculate average ranks $\bar{r}_j$, $j = 1, \ldots, m$:

datasets	classifier 1	classifier 2	classifier 3
dataset 1	1	3	2
dataset 2	1.5	1.5	3
dataset 3	1	2	3
dataset 4	2	3	1
dataset 5	2.5	2.5	1
average	1.6	2.4	2.0

Friedman Test [Demšar, 2006]

Procedure (N classifiers, m files) – continued

- Using average ranks calculate Friedman's statistics:

$$\chi_F^2 = \frac{12m}{N(N+1)} \left(\sum_{j=1}^{m} \bar{r}_j^2 - \frac{N(N+1)^2}{4} \right)$$

  which is distributed approx. as χ^2 with $N-1$ df.
- Reject null hypothesis if $\chi_F^2 > \chi_{crit}^2$ and proceed to a post-hoc analysis.
- Post-hoc analysis (Nemenyi test): calculate critical difference:

$$CD = q_\alpha \sqrt{\frac{N(N+1)}{6m}}$$

  where q_α is the critical value (based on α and N).
- All algorithms with the differences in average ranks greater than CD are significantly different.

Podsumowanie sposobów analizy

- **Wybierz właściwe miary oceny**
- **Unikaj przeuczenia** (niezależny zbiór testowy)
- Stosuj właściwą metodę szacowania miar
  - np. warstwową ocenę krzyżową
- Bądź czujny przy porównywaniu kilku klasyfikatorów
- Przypomnij sobie metody statystycznej analizy znaczenia wyników
- Nie oczekuj, że jeden algorytm będzie najlepszy we wszystkich typach danych.
  - „Each has its own area of superiority”

Perspektywa opisowa oceny wiedzy

- Trudniejsza niż ocena zdolności klasyfikacyjnych.
- Rozważmy przykład reguł:
  - Klasyfikacyjne (decyzyjne).
  - Asocjacyjne.
- Pojedyncza reguła oceniana jako potencjalny reprezentant „interesującego” wzorca z danych
  - W literaturze propozycje tzw. ilościowych miar oceny reguł oraz sposoby definiowania „interesujących” reguł, także na podstawie wymagań podawanych przez użytkownika.

Opisowe miary oceny reguł

- Miary dla reguły *r* (*jeżeli P to Q*) definiowane na podstawie zbioru przykładów *U*, z którego została wygenerowana.
- Tablica kontyngencji dla reguły *jeżeli P to Q* :

	Q	$\neg Q$	
P	n_{PQ}	$n_{P\neg Q}$	n_P
$\neg P$	$n_{\neg PQ}$	$n_{\neg P\neg Q}$	$n_{\neg P}$
	n_Q	$n_{\neg Q}$	n

- Przegląd różnych miar, np.: Yao Y.Y, Zhong N.: An analysis of quantitative measures associated with rules, w: Proc. of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNAI 1574, Springer, 1999, s. 479-488.
- Także rozprawa habilitacyjna J.Stefanowski: Algorytmy indukcji reguł w odkrywaniu wiedzy.

Popularne miary oceny reguł

- **Wsparcie reguły** (ang. *support*) zdefiniowane jako:

$$G(P \wedge Q) = \frac{n_{PQ}}{n}$$

- **Dokładność** (ang. *accuracy*) / wiarygodność (ang. *confidence*) reguły (bezwzględne wsparcie konkluzji Q przez przesłankę P):

$$AS(Q \mid P) = \frac{n_{PQ}}{n_P}$$

- **Względne pokrycie** (ang. *coverage*) reguły zdefiniowane jako:

$$AS(P \mid Q) = \frac{n_{PQ}}{n_Q}$$

Zaawansowane miary oceny reguł

Change of support – rodzaj konfirmacji wsparcia hipotezy Q przez wystąpienie przesłanki P (propozyjca Piatetsky-Shapiro)

$$CS(Q \mid P) = AS(Q \mid P) - G(Q)$$

gdzie $$G(Q) = \frac{n_Q}{n}$$

Zakres wartości od -1 do +1 ; Interpretacja: różnica między prawdopodobieństwami a prior i a posterior; dodatnie wartości wystąpienie przesłanki P powoduje konkluzję Q; ujemna wartość wskazuje że nie ma wpływu.

Degree of independence:

$$IND(Q,P) = \frac{G(P \wedge Q)}{G(P) \cdot G(Q)}$$

Złożone miary oceny reguł

Połączenie miar podstawowych

Significance of a rule (propozycja Yao i Liu)

$$S(Q \mid P) = AS(Q \mid P) \cdot IND(Q, P)$$

Klosgen's measure of interest

$$K(Q \mid P) = G(P)^{\alpha} \cdot (AS(Q \mid P) - G(Q))$$

Michalski's weighted sum

$$WSC(Q \mid P) = w_1 \cdot AS(Q \mid P) + w_2 \cdot AS(P \mid Q)$$

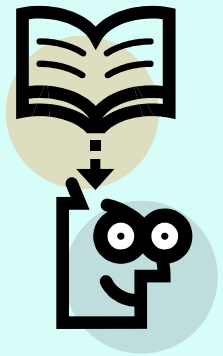
The relative risk (Ali, Srikant):

$$r(Q \mid P) = \frac{AS(Q \mid P)}{AS(Q \mid \neg P)}$$

No i na razie wystarczy

- **O innych zagadnieniach procesu odkrywania wiedzy jeszcze porozmawiamy!**

I to by było na tyle …



Nie zadawalaj się tym
co usłyszałeś – poszukuj więcej!
Czytaj książki oraz samodzielnie
badaj problemy ML!