Tworzenie i ocena klasyfikatorów

Jest procesem trzyetapowym:

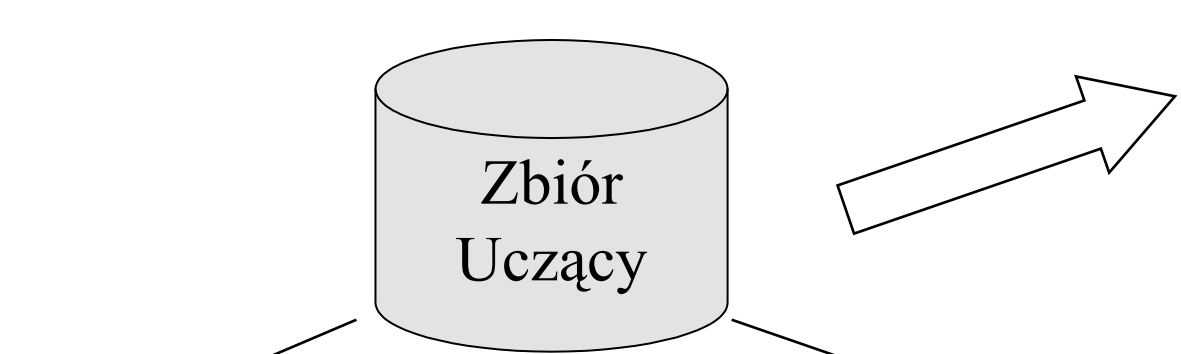
1. Konstrukcja modelu w oparciu o zbiór danych wejściowych (przykłady uczące).

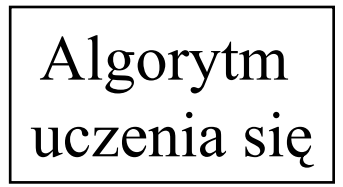
   Przykładowe modele - klasyfikatory:

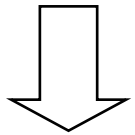
   - drzewa decyzyjne,
   - reguły (IF .. THEN ..),
   - sieci neuronowe.

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

Proces Klasyfikowania (I) – Uczenie się



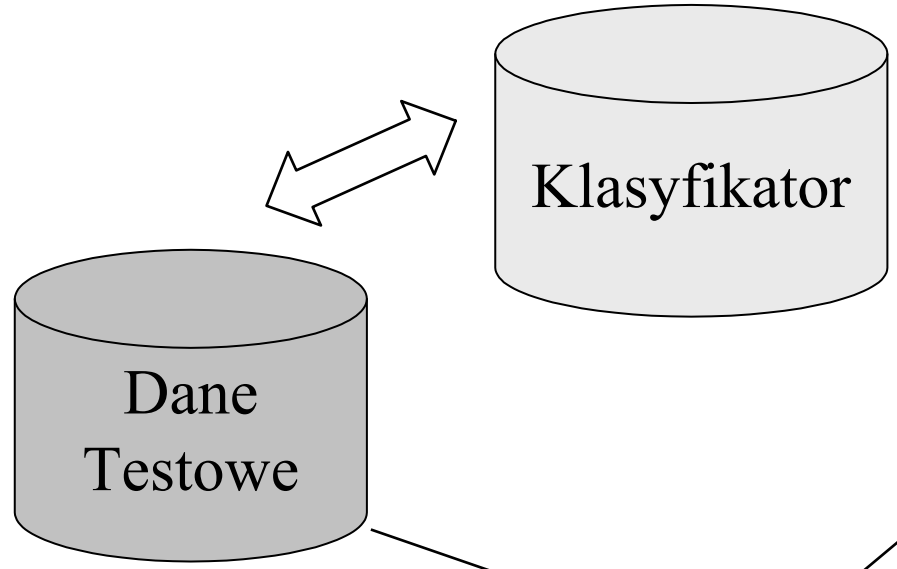


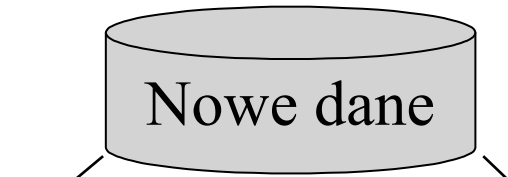


NAME	INCOME	AGE	Rating
Mike	low	30	no
Mary	high	36	yes
Bill	high	45	yes
Jim	medium	50	yes
Dave	medium	27	no
Anne	low	28	no

IF income = ‘high’
OR age > 35
THEN rating = ‘yes’

Proces Klasyfikowania obiektów

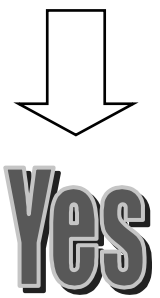




(Jeff, high, 40)

NAME	INCOME	AGE	RATING
Tom	low	24	no
Merlisa	medium	33	no
George	high	33	yes
Joseph	low	40	yes

Rating?



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

Predictor Error Measures (zmienna y liczbowa)

- **Measure predictor accuracy: measure how far off the predicted value is from the actual known value**
- Loss function**: measures the error betw. y_i and the predicted value y_i^**
  - **Absolute error: $| y_i - y^_i|$**
  - **Squared error: $(y_i - y^_i)^2$**
- **Test error (generalization error): the average loss over the test set**
  - **Mean absolute error:** $\frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$ **Mean squared error:** $\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$
  - **Relative absolute error:** $\frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{\sum_{i=1}^{n} | y_i - \bar{y} |}$ **Relative squared error:** $\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$
  - **The mean squared-error exaggerates the presence of outliers**
  - **Popularly use (square) root mean-square error, similarly, root relative squared error**

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 zastosowania → jedna z klas posiada szczególne znaczenie, np. diagnozowanie poważnej choroby. Problem → 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

Niezrównoważone klasy – pamiętaj o innych podejściach i miarach oceny (sensitivity, 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.

Inne miary budowane w oparciu o macierze binarne

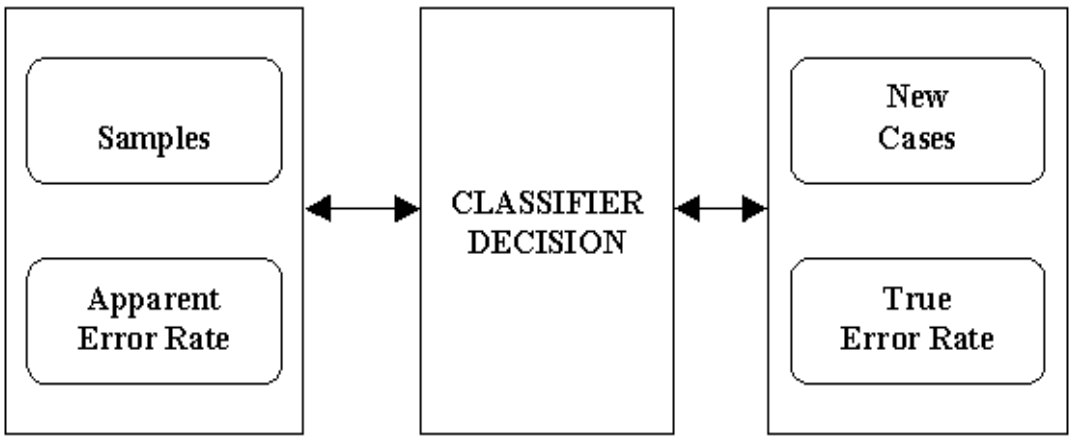
	Domain	Plot	Explanation
Lift chart	Marketing	TP Subset size	TP (TP+FP)/ (TP+FP+TN+FN)
ROC curve	Communications	TP rate FP rate	TP/(TP+FN) FP/(FP+TN)
Recall-precision curve	Information retrieval	Recall Precision	TP/(TP+FN) TP/(TP+FP)

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!
  - Choć zasada reprezentatywności próbki uczącej …
- Problem przeuczenia (ang. overfiting)
  - Nadmierne dopasowanie do specyfiki danych uczących powiązane jest najczęściej z utratą zdolności uogólniania (ang. generalization) i predykcji nowych faktów!

Podejście empiryczne

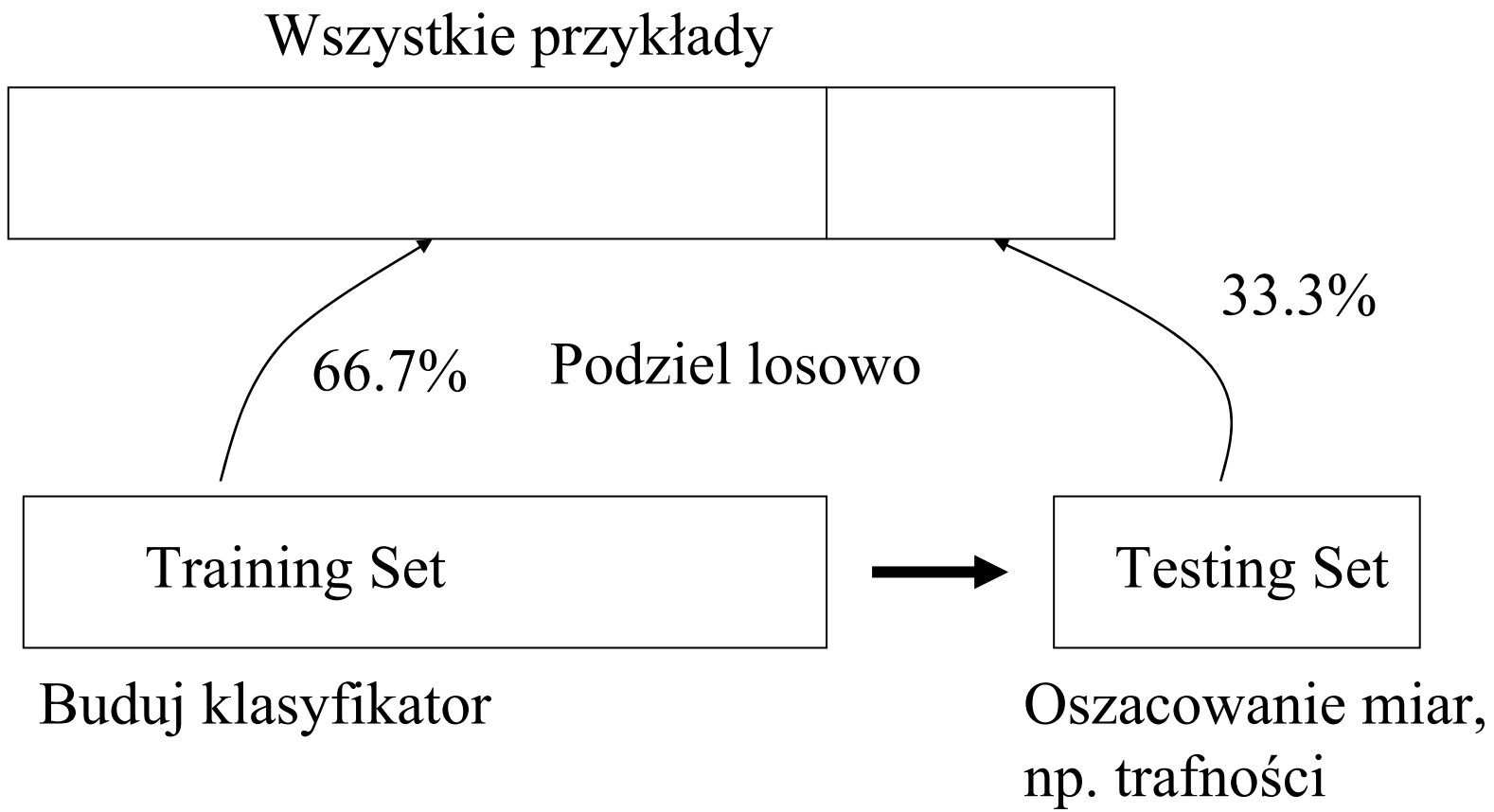
- Zasada „Train and test” (ucz i testuj)
- Gdy nie ma podziału zadanego przez nauczyciela, to wykorzystaj losowe podziały.
- Nadal pytanie jak szacować wiarygodnie?



Empiryczne metody estymacji

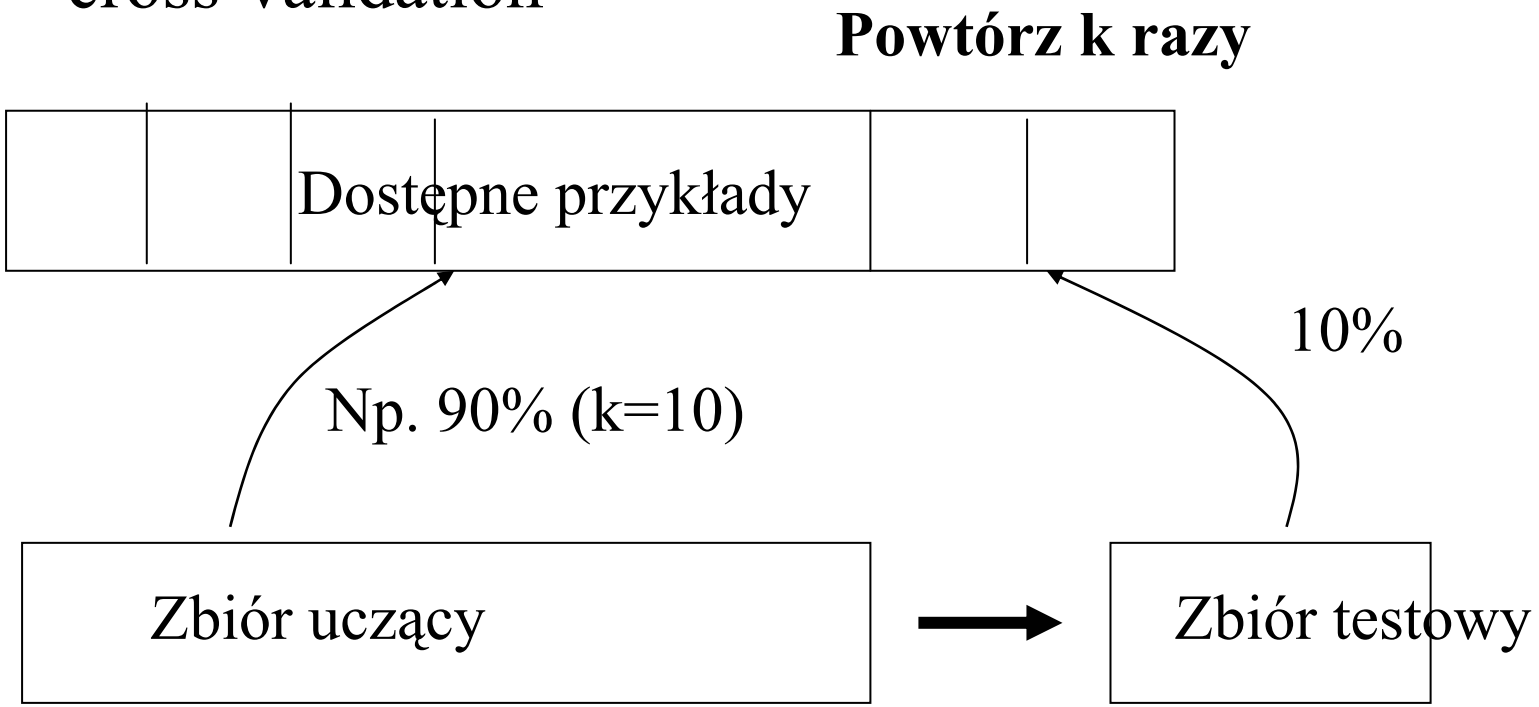
- **Techniki podziału: „hold-out”**
  - Użyj dwóch niezależnych zbiorów: uczącego (2/3), testowego (1/3)
  - Jednokrotny podział losowy stosuje się dla dużych zbiorów (hold-out)
- **„Cross-validation” - Ocena krzyżowa**
  - Podziel losowo dane w *k* podzbiorów (równomierne lub warstwowe)
  - Użyj *k-1* 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).
- **Bootstrapping i leaving-one-out**
  - Dla małych rozmiarów danych.
  - „Leaving-one-out” jest szczególnym przypadkiem, dla którego liczba iteracji jest równa liczbie przykładów

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



Mniejsza liczba przykładów (od 100 do kilku tysięcy)

* cross-validation



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

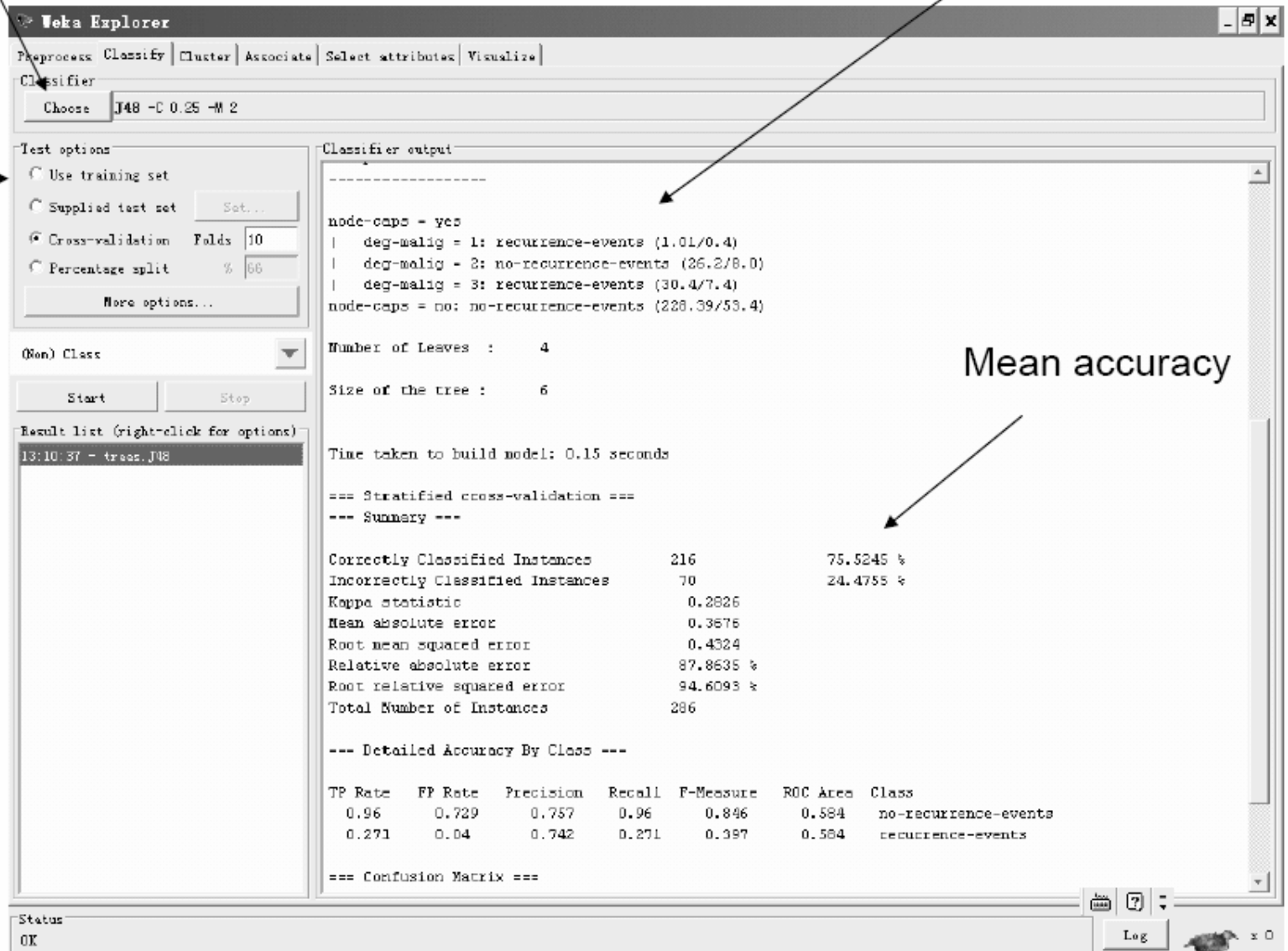
Podział	Kl_1	Kl_2
1	87,45	88,4
2	86,5	88,1
3	86,4	87,2
4	86,8	86
5	87,8	87,6
6	86,6	86,4
7	87,3	87
8	87,2	87,4
9	88	89
10	85,8	87,2
Srednia	**86,98**	**87,43**
Odchylenie	**0,65**	**0,85**

- Test statystyczny (t-Studenta dla par zmiennych/zależnych)
- H0 : $\alpha 1 = \alpha 2$ H1 : $\alpha 1 < \alpha 2$
- *t*emp = 1,733 (p = 0,117) ???

Classifier

WEKA Explorer

Decision Trees

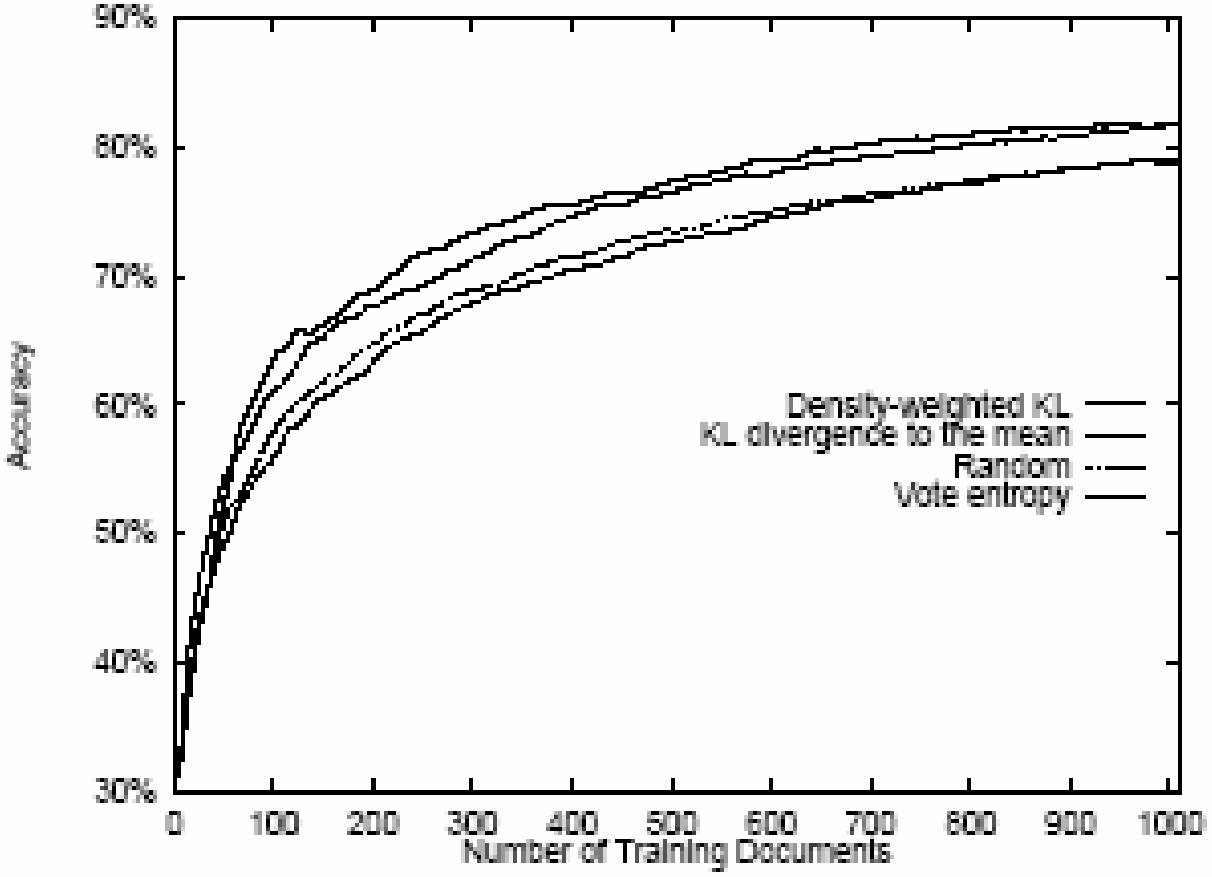


Testing data

Mean accuracy

Krzywe uczenia się - learning curve

Klasyfikacja tekstów przez Naive Bayes 20 Newsgroups dataset -



[McCallum & Nigam, 199

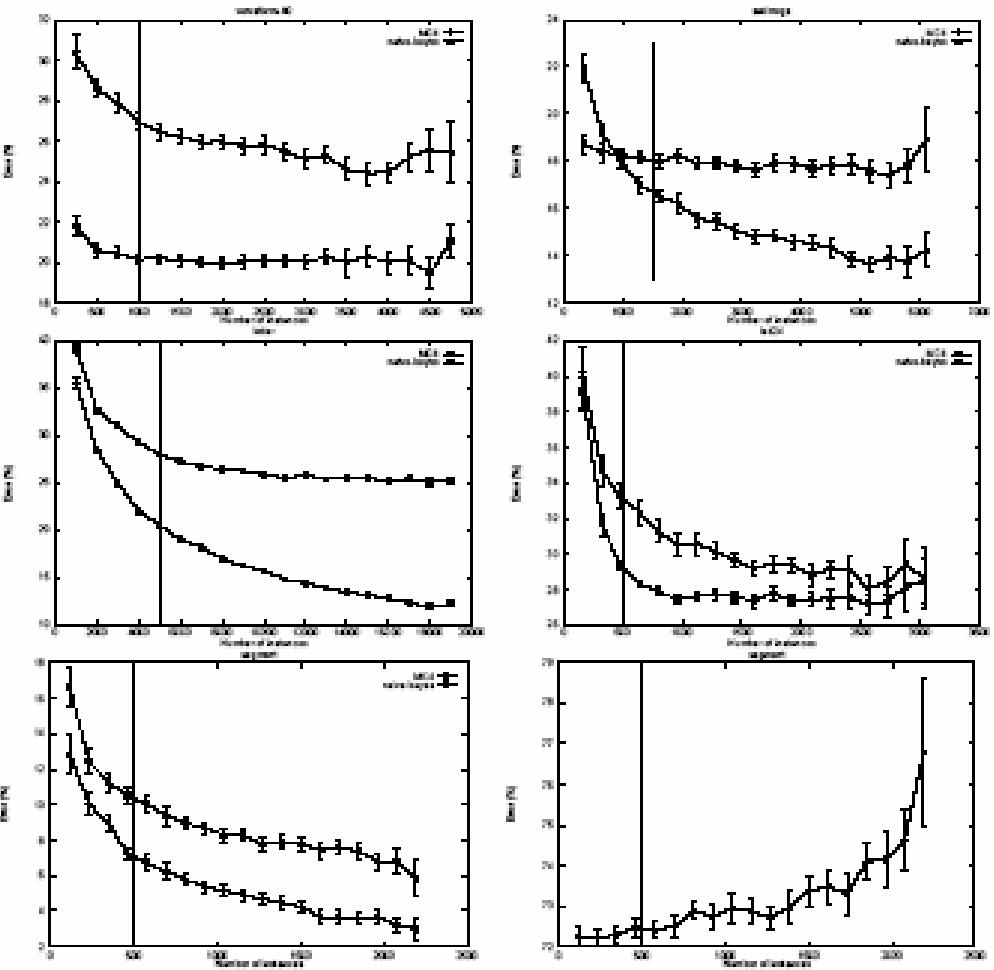
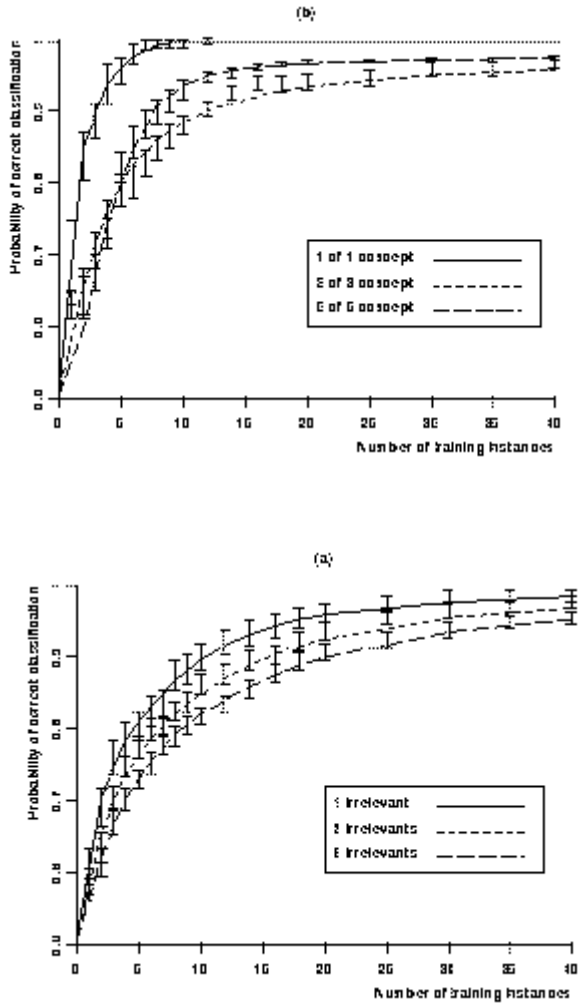


Figure 4. Learning curves for selected datasets showing different behaviors of MC4 and Naive-Bayes. Waveform represents stabilization at about 3,000 instances; satimage represents a cross-over as MC4 improves while Naive-Bayes does not; letter and segment (left) represent continuous improvements, but 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.

I to by było na tyle …

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