KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

dr inż. Artur Klepaczko

Eksploracja danych
Przekształcenia danych

Zadanie nr 13 – Studia podyplomowe „Przetwarzanie i analiza obrazów biomedycznych”

Prezentacja multimedialna
współfinansowana przez Unię Europejską
w ramach Europejskiego Funduszu Społecznego
w projekcie

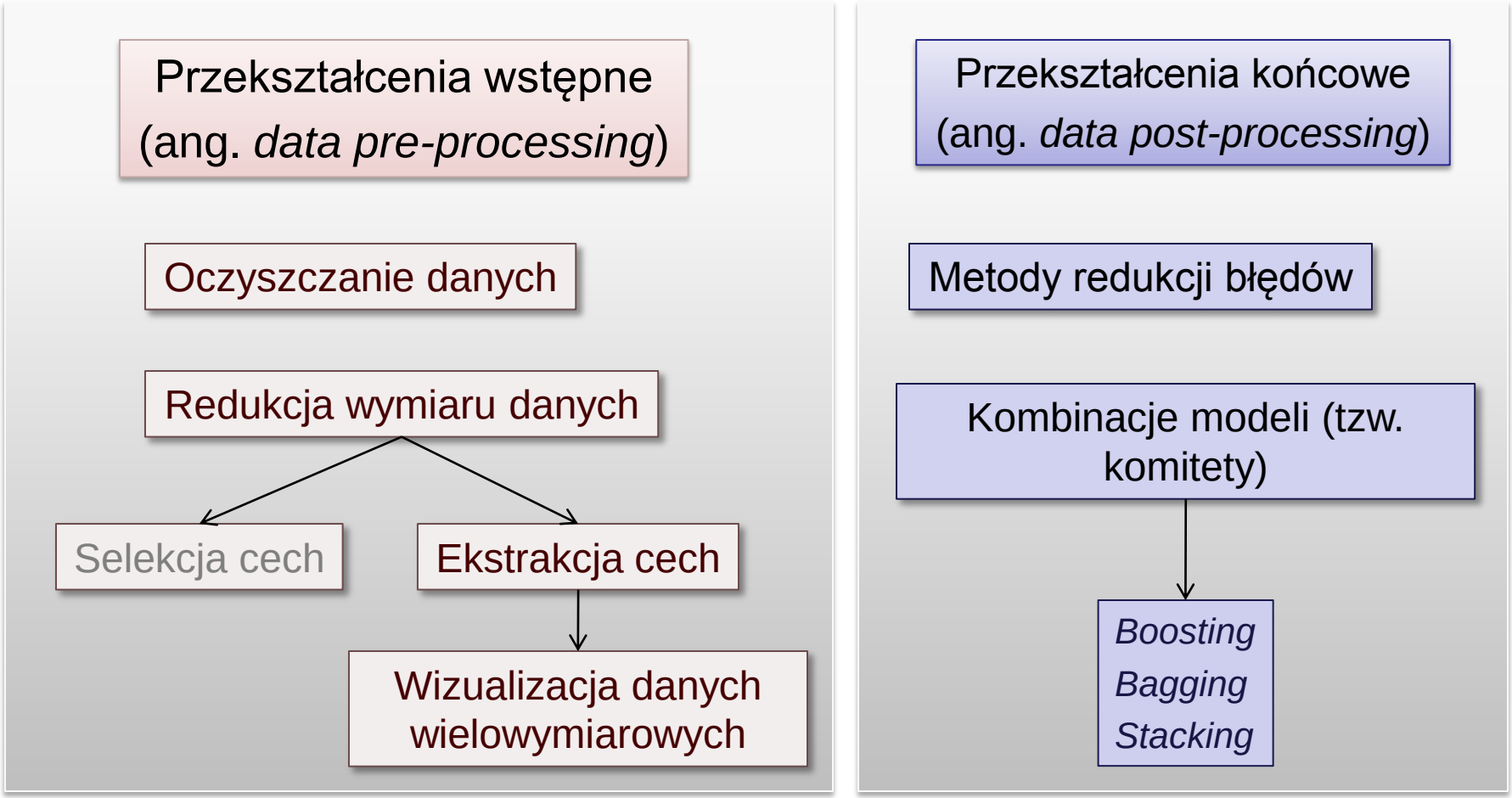
„Innowacyjna dydaktyka bez ograniczeń – zintegrowany rozwój Politechniki Łódzkiej – zarządzanie Uczelnią, nowoczesna oferta edukacyjna i wzmacniania zdolności do zatrudniania osób niepełnosprawnych”

Politechnika Łódzka
Instytut Elektroniki

90-924 Łódź, ul. Żeromskiego 116,
tel. 042 631 28 83
www.kapitalludzki.p.lodz.pl

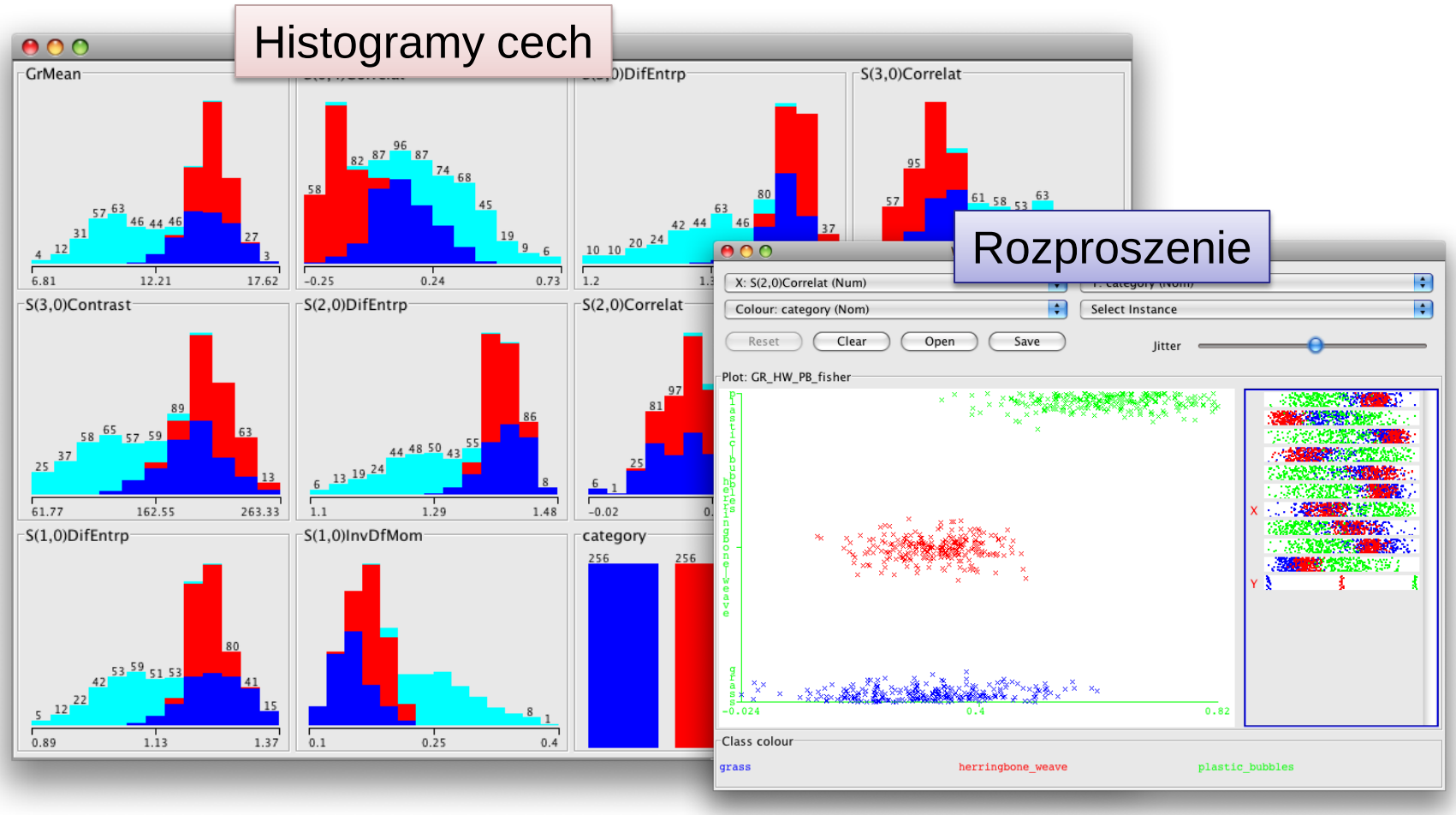
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Zagadnienia wykładu



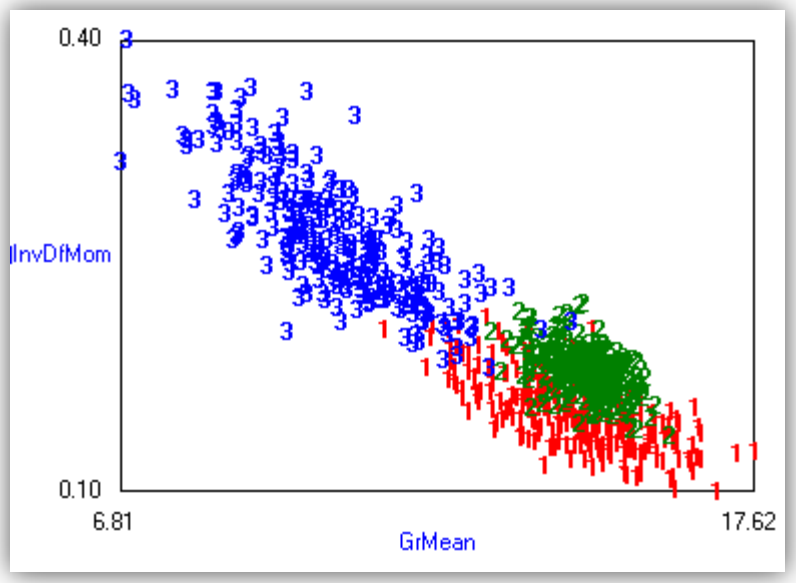
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Wizualizacja danych – wymiar 1D

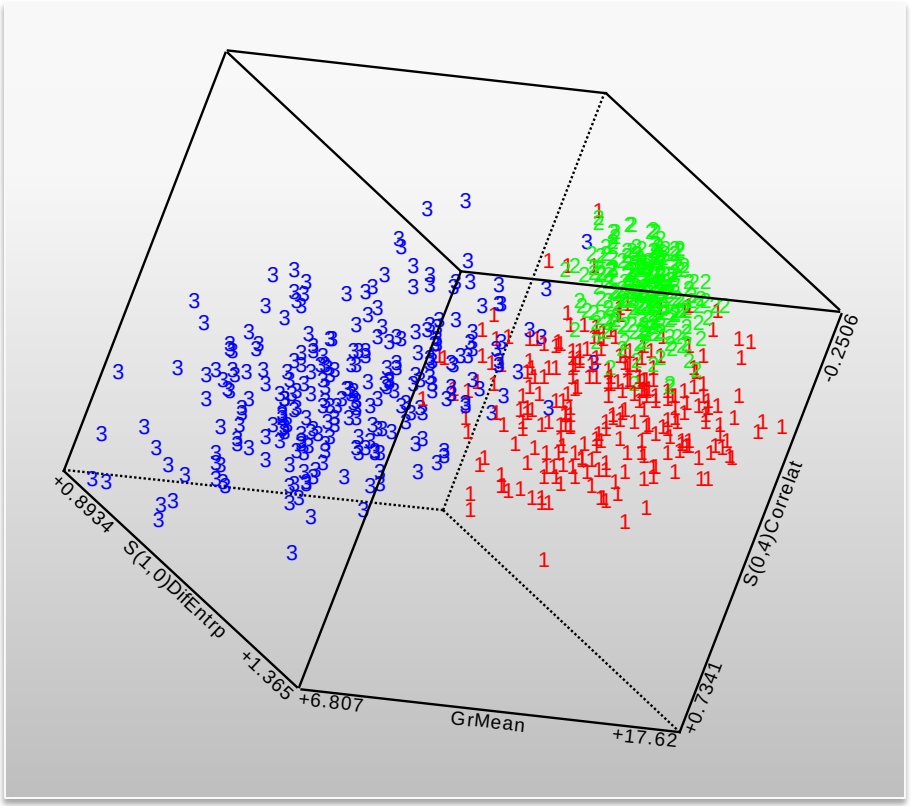


Wizualizacja danych – 2D i 3D

Opcja *Analysis→Raw data* – b11



Moduł *3D feature space* - MaZda



Jak zobaczyć dane *n*-wymiarowe w całości?

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Transformacja wektorów cech

Zbiór wektorów w oryginalnej *N*-wymiarowej przestrzeni cech

$$\mathbf{X} = [x_{ij}]_{Q \times N}$$

Q – liczba wektorów danych
N – liczba cech

Macierz transformacji

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,j} & \cdots & w_{1,N} \\ \vdots & \ddots & & & \vdots \\ w_{j,1} & & w_{j,j} & & w_{j,N} \\ \vdots & & & \ddots & \vdots \\ w_{N,1} & \cdots & w_{N,j} & \cdots & w_{N,N} \end{bmatrix}$$

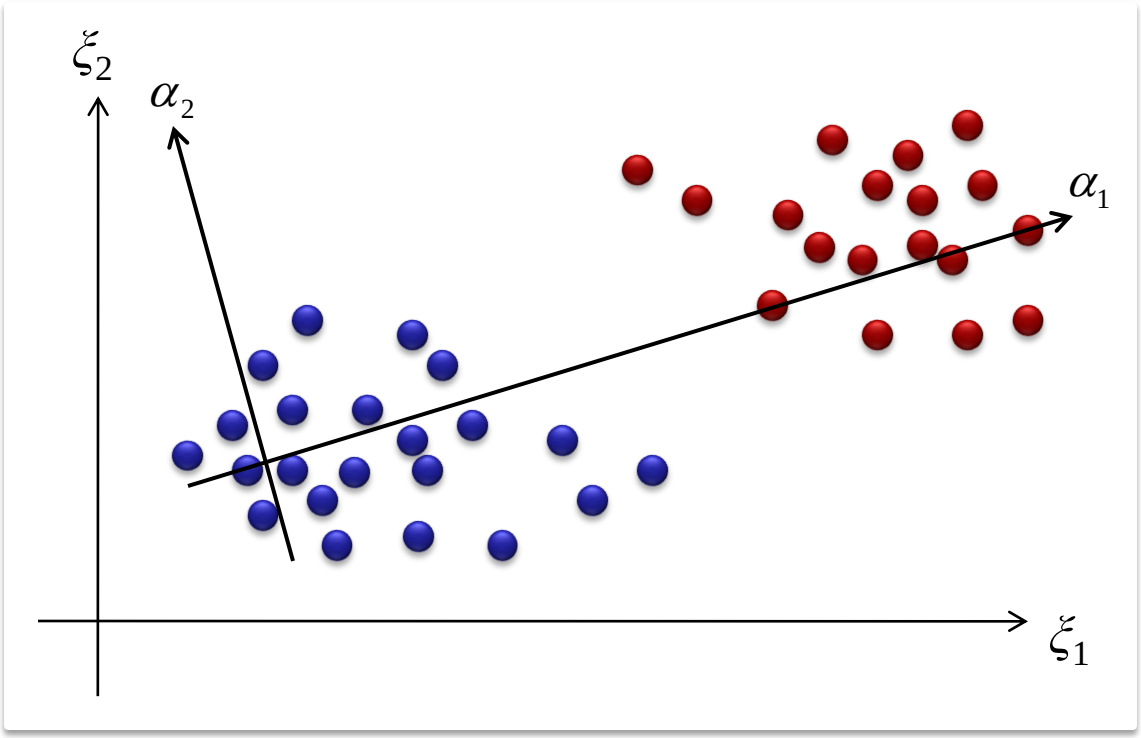
Zbiór wektorów danych w nowej, również *N*-wymiarowej przestrzeni agregatów cech.

$$\mathbf{A} = [\alpha_{i,j}]_{Q \times N}$$

Kierunki w nowej przestrzeni cech posiadają określone właściwości algebraiczne (np. są względem siebie ortogonalne) i statystyczne (przechowują pewną porcję wariancji zbioru danych). Kolejność wyznaczonych kierunków ma znaczenie.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Analiza głównych składowych (PCA)



α_1 — kierunek największej wariancji (inaczej zmienności lub odchylenia od średniej)

α_2 — kierunek prostopadły do α_1

Sumaryczna wariancja wzdłuż nowych kierunków jest równa wariancji wzdłuż kierunków oryginalnej przestrzeni cech.

PCA (ang. *principal component analysis*) — obrót układu współrzędnych w taki sposób, aby zmaksymalizować wariancję wektorów danych wzdłuż kolejnych współrzędnych.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Macierz transformacji w analizie PCA

Macierz wektorów danych

$$\mathbf{X} = [x_{ij}]_{Q \times N}$$

Macierz kowariancji

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\xi_1}^2 & \cdots & \sigma_{\xi_1}\sigma_{\xi_j} & \cdots & \sigma_{\xi_1}\sigma_{\xi_N} \\ \vdots & \ddots & & & \vdots \\ \sigma_{\xi_j}\sigma_{\xi_1} & & \sigma_{\xi_j}^2 & & \sigma_{\xi_j}\sigma_{\xi_N} \\ \vdots & & & \ddots & \vdots \\ \sigma_{\xi_N}\sigma_{\xi_1} & \cdots & \sigma_{\xi_N}\sigma_{\xi_j} & \cdots & \sigma_{\xi_N}^2 \end{bmatrix}$$

Kowariancja zmiennych ξ_1 a ξ_j — określa stopień liniowej zależności między cechami

Wariancja wektorów danych wzdłuż kierunku ξ_j

Wyznaczenie wektorów własnych macierzy kowariancji

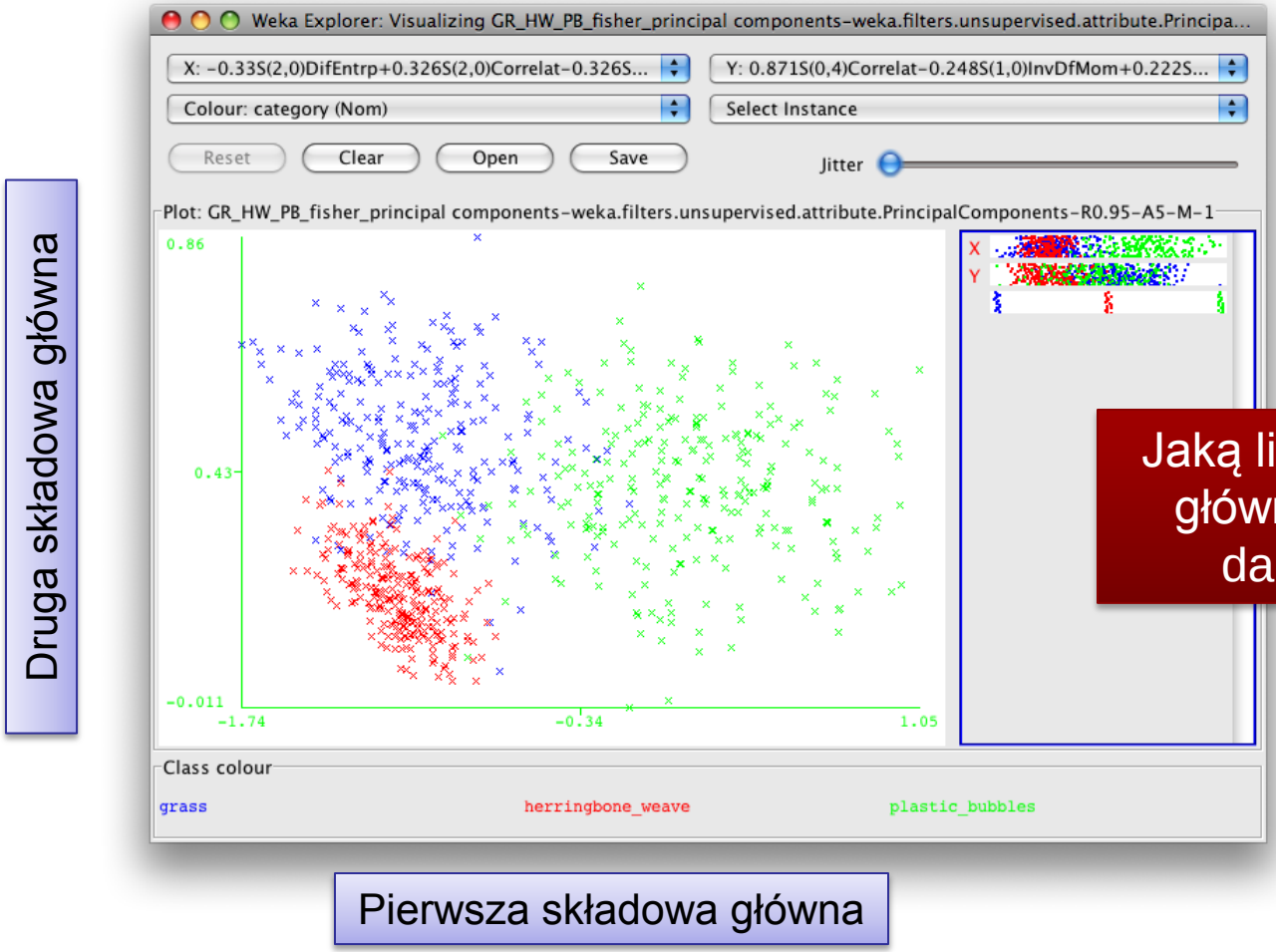
$$\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_N$$

oraz stowarzyszonych z nimi wartości własnych

$$\lambda_1, \lambda_2, \ldots, \lambda_N$$

Wektor własny skojarzony z największą wartością własną wyznacza pierwszy kierunek w nowej przestrzeni cech.

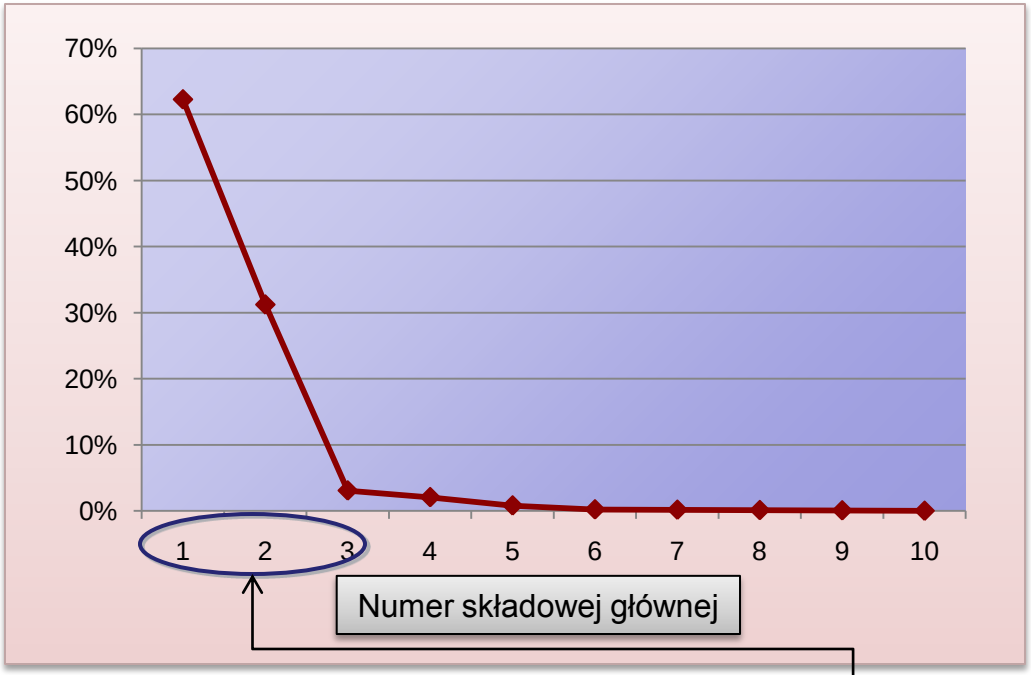
Zastosowanie PCA do wizualizacji danych



Jaką liczbę składowych głównych wybrać do dalszej analizy?

Wariancja danych w nowej przestrzeni

Porcja wariancji zbioru danych reprezentowana przez poszczególne składowe główne [%]



Wariancję składowej głównej wyznacza skojarzona z nim wartość własna macierzy kowariancji oryginalnego zbioru wektorów cech.

* Results [principal component analysis]
Eigenvalues of data covariance matrix:

5.05278E-002
2.53293E-002
2.49567E-003
1.67894E-003
6.57158E-004
1.76331E-004
1.32917E-004
8.78142E-005
5.65626E-005
3.25309E-006

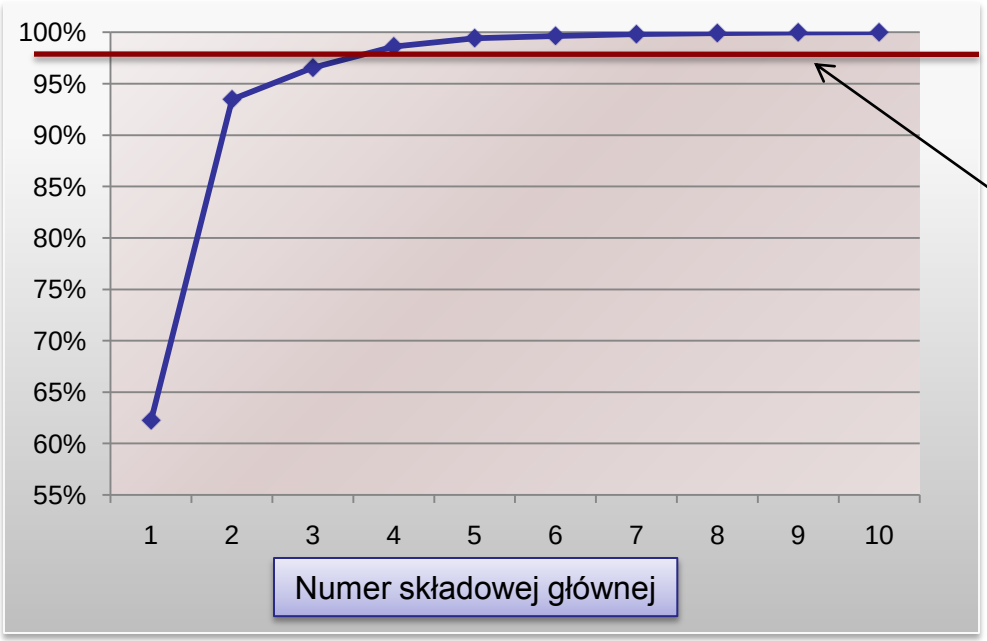
Zwykle, kilka pierwszych głównych składowych przechowuje większość wariancji zbioru danych.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Zastosowanie analizy PCA do redukcji wymiaru danych

Ekstrakcja cech: transformacja oryginalnej przestrzeni atrybutów do nowej przestrzeni o mniejszym wymiarze

Część całkowitej wariancji zbioru danych w kolejnych składowych głównych (wartości skumulowane) [%]



Arbitralnie ustalony próg (współczynnik wymiarowości liniowej):

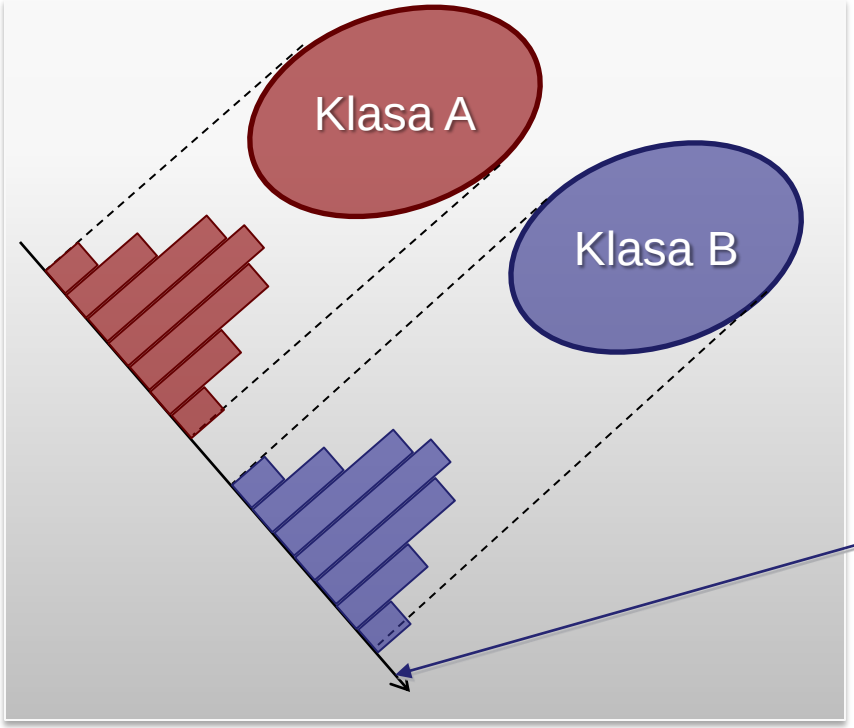
— określa, jaka część całkowitej wariancji zbioru danych ma zostać zachowana po transformacji

— wyznacza liczbę składowych głównych — wymiar nowej przestrzeni cech

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Liniowa analiza dyskryminacyjna

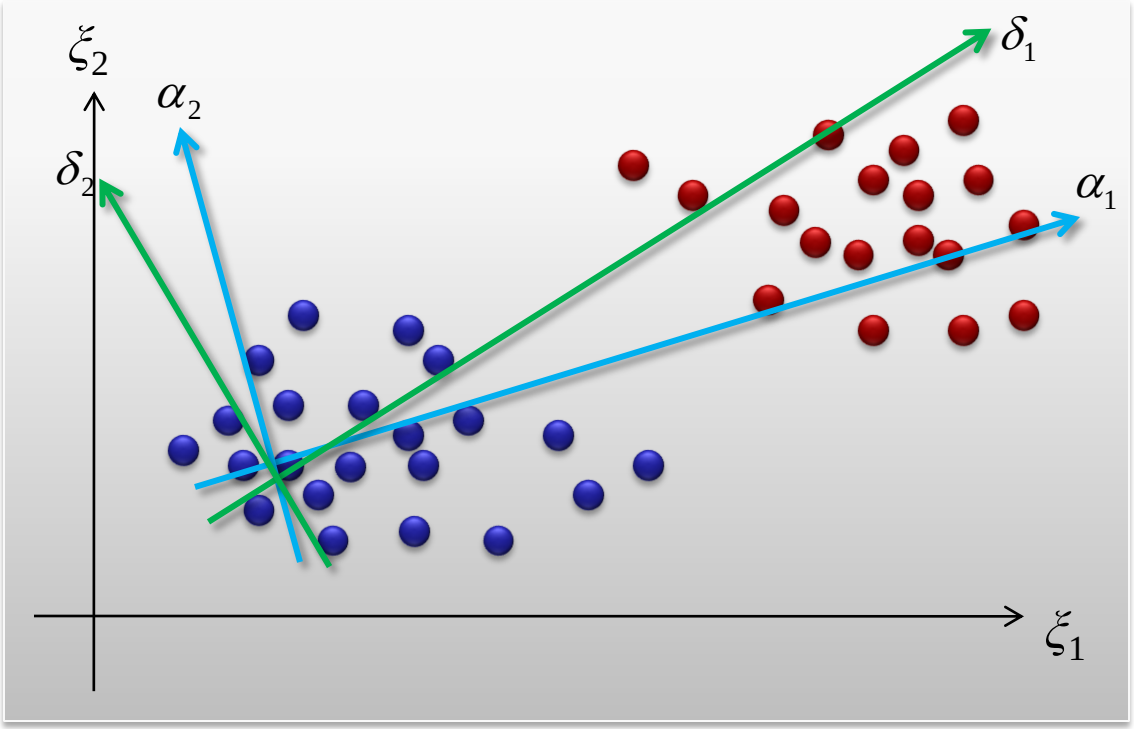
Analiza LDA — metoda klasyfikacji liniowej, polegająca na wyznaczeniu kierunków w przestrzeni cech, które najlepiej rozdzielają dane. W przypadku *K* klas, maksymalna liczba takich kierunków wynosi zawsze *K*–1.



Jeżeli wymiar oryginalnej przestrzeni cech $N > K–1$, to po zrzutowaniu wektorów danych na kierunki wyznaczone przez LDA uzyskujemy efekt redukcji wymiaru danych.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Porównanie transformacji PCA i LDA



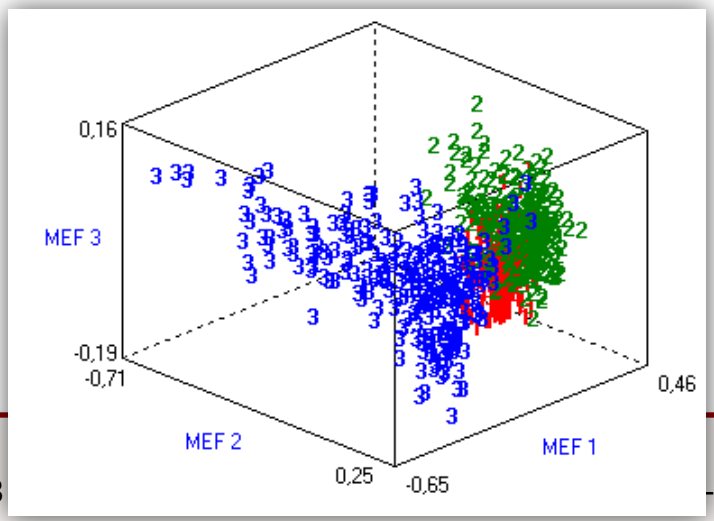
PCA wyznacza kierunki największej zmienności → cechy najbardziej reprezentatywne (ang. *Most Expressive Features*, MEF).

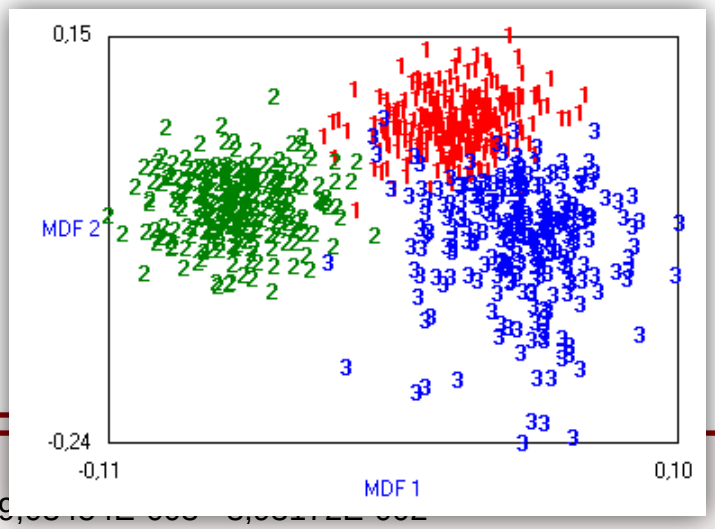
LDA wyznacza kierunki najlepszej separacji klas → cechy najbardziej dyskryminujące (ang. *Most Discriminative Features*, MDF).

W przypadku analizy PCA, przynależność wektorów danych do konkretnej klasy nie jest brana pod uwagę. Dlatego PCA stosuje się często jako metodę nienadzorowanej redukcji wymiaru danych.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Przykład — klasyfikacja wektorów danych po ekstrakcji cech





…
3 -1.29877E
3 -2.48578E-002 -1.70643E-001 3.35569E-002 2.06922E
3 -1.29480E-001 -1.85168E-001 -2.47463E-002 -5.25038E

> 1-NN classification of MEFs

Missclassified data vectors: 26/768 [or 3.39%]

Sample No: 14; Category: 1; ClassResult: 3
Sample No: 20; Category: 1; ClassResult: 3
Sample No: 23; Category: 1; ClassResult: 3
Sample No: 63; Category: 1; ClassResult: 3
…

…
3
3 3,77133E-003 -7,58219E-002

> 1-NN classification of MDFs

Missclassified data vectors: 46/768 [or 5.99%]

Sample No: 4; Category: 1; ClassResult: 3
Sample No: 9; Category: 1; ClassResult: 3
Sample No: 13; Category: 1; ClassResult: 3
Sample No: 26; Category: 1; ClassResult: 2
…

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Transformacja losowa

$$\mathbf{A}_{Q\times d} = \mathbf{X}_{Q\times N}\mathbf{R}_{N\times d}$$

d — wymiar przestrzeni docelowej

$d << N$

Kierunki w nowej przestrzeni cech są *prawie* ortogonalne (o ile jest ich dużo).

Względne odległości pomiędzy wektorami danych pozostają w przybliżeniu takie same.

$$\|\mathbf{x}_1 - \mathbf{x}_2\| = \sqrt{d/n}\,\|\mathbf{x}_1\mathbf{R} - \mathbf{x}_2\mathbf{R}\|$$

Wybór elementów macierzy transformacji jest losowy.

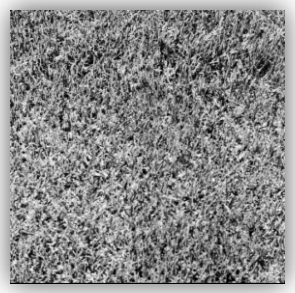
Przykład

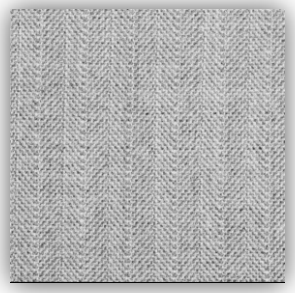
$$r_{ij} = \sqrt{3}\cdot\begin{cases} +1 & \Pr[r_{ij}=1]=1/6 \\ 0 & \Pr[r_{ij}=0]=2/3 \\ -1 & \Pr[r_{ij}=-1]=1/6 \end{cases}$$

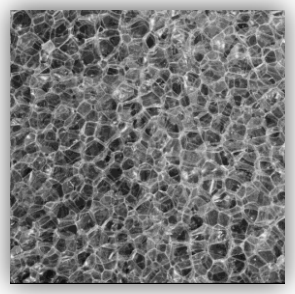
Błąd klasyfikacji wektorów w przestrzeni po transformacji jest większy niż w przypadku analizy PCA, ale złożoność obliczeniowa transformacji losowej jest mniejsza.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Transformacja losowa — przykład







Zbiór danych: 3 klasy, 256 wektorów danych w klasie, 250 cech tekstury

Transformacja losowa

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         710               92.4479 %
Incorrectly Classified Instances        58                7.5521 %
Kappa statistic                          0.8867

=== Confusion Matrix ===

   a   b   c   <-- classified as
 228  15  13 |   a = grass
  15 241   0 |   b = herringbone_w
  13   2 241 |   c = plastic_bubbles
```

Klasyfikator: SVM z liniową funkcją jądra

PCA

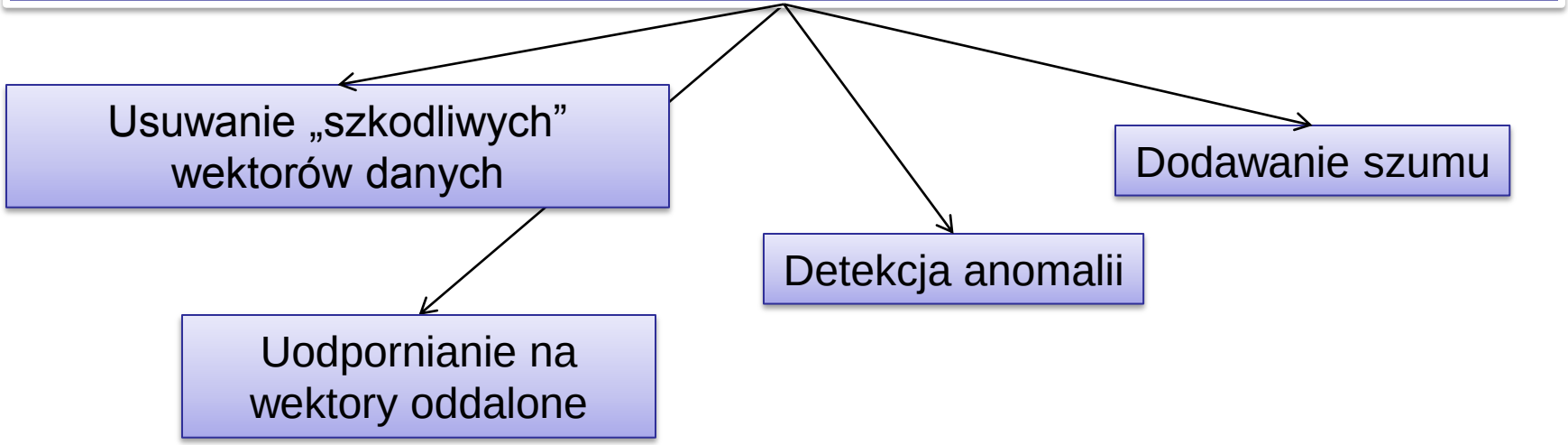
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         719               93.6198 %
Incorrectly Classified Instances        49                6.3802 %
Kappa statistic                          0.9043

=== Confusion Matrix ===

   a   b   c   <-- classified as
 243   3  10 |   a = grass
  10 246   0 |   b = herringbone_weave
  24   2 230 |   c = plastic_bubbles
```

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Oczyszczanie danych

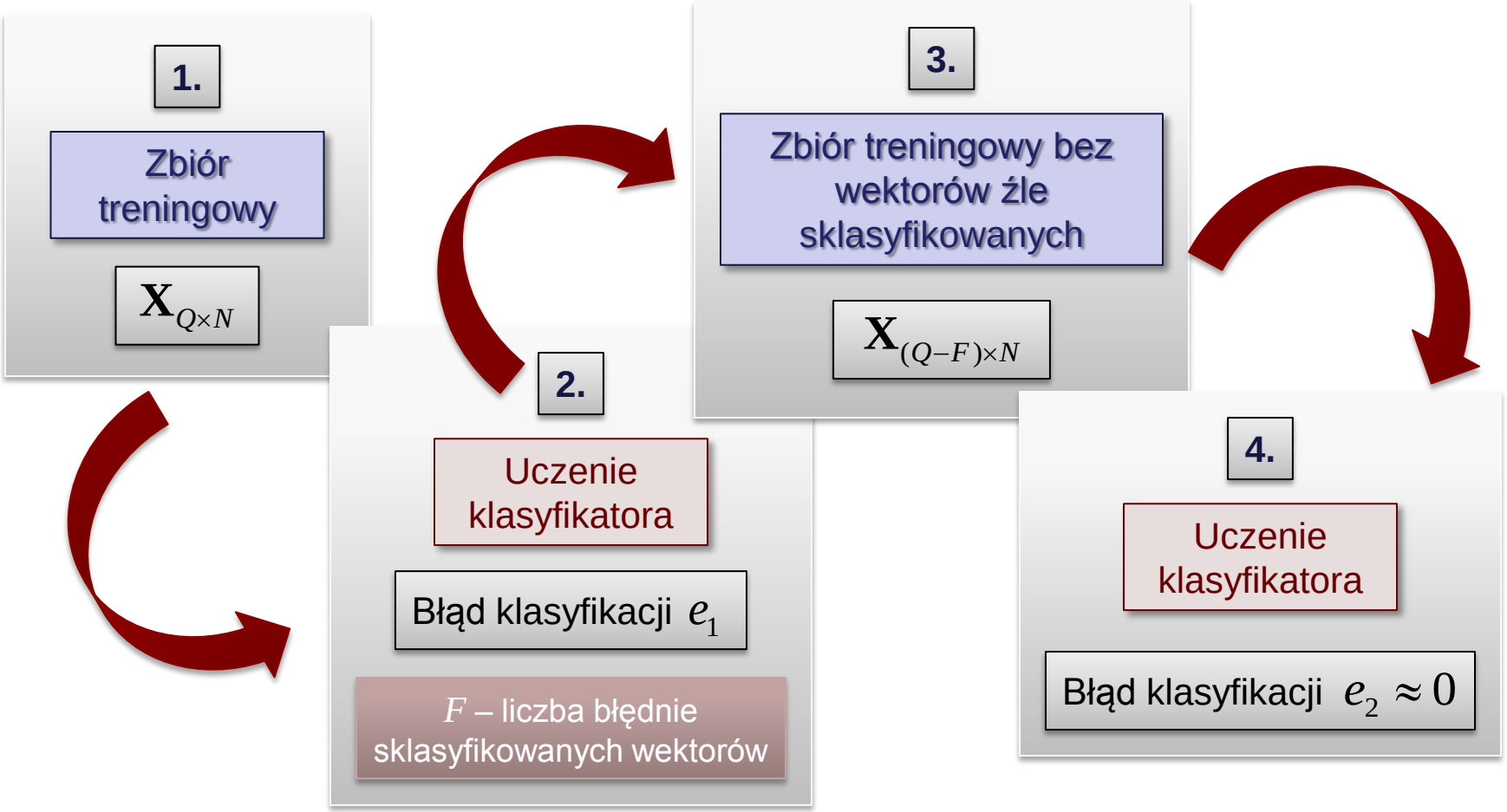
Bardzo często dane zawierają błędy, szum, wartości i wektory oddalone (ang. *outliers*), artefakty oraz inne anomalie.

Techniki oczyszczania danych mają za zadanie zwiększyć wiarygodność zbioru danych, a przez to poprawić jakość i skuteczność algorytmów uczących się.



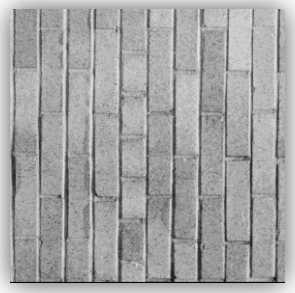
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

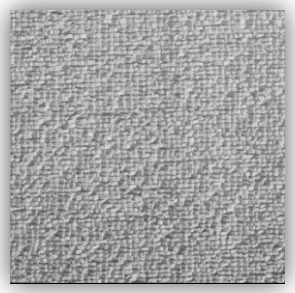
Usuwanie „szkodliwych wektorów”



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Poprawa klasyfikacji przez usunięcie „szkodliwych" wektorów





Przed…

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances	**118**	**92.1875 %**
Incorrectly Classified Instances	**10**	**7.8125 %**
Kappa statistic	0.8829	
Total Number of Instances	128	

Zbiór danych: 3 klasy, 128 wektorów danych w klasie, 10 cech tekstury

Trening: 2/3 zbioru
Test: 1//3 zbioru

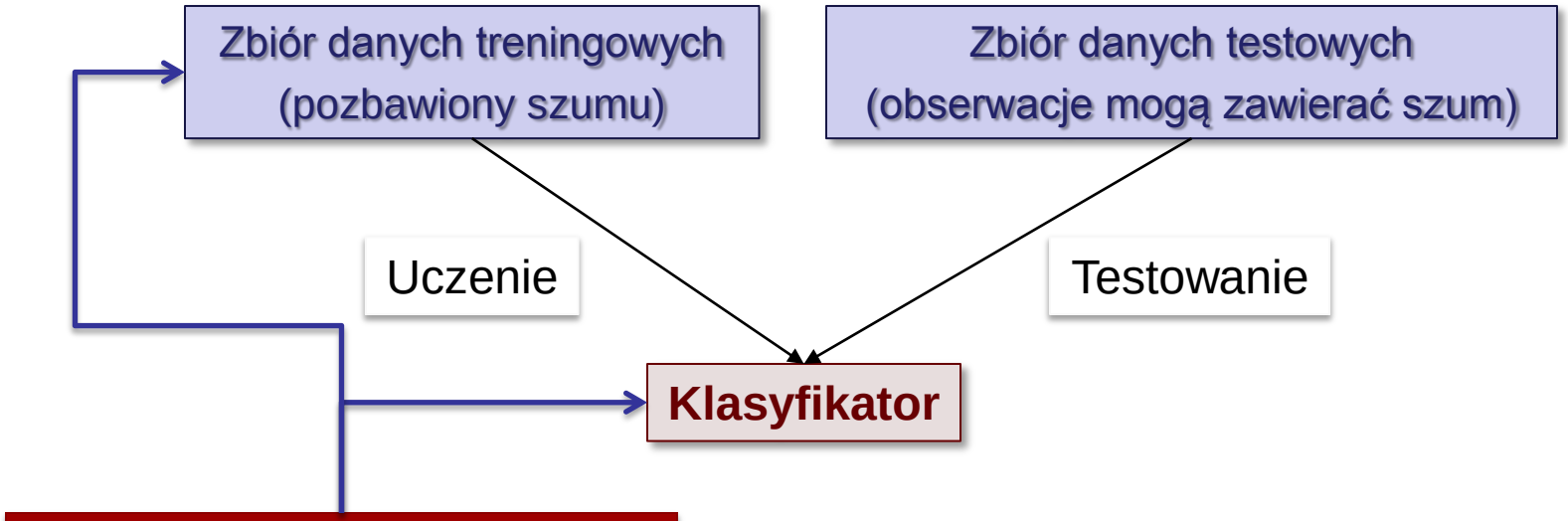
…i po usunięciu źle sklasyfikowanych wektorów

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances	**123**	**96.0938 %**
Incorrectly Classified Instances	**5**	**3.9063 %**
Kappa statistic	0.9414	
Total Number of Instances	128	

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Dodawanie szumu



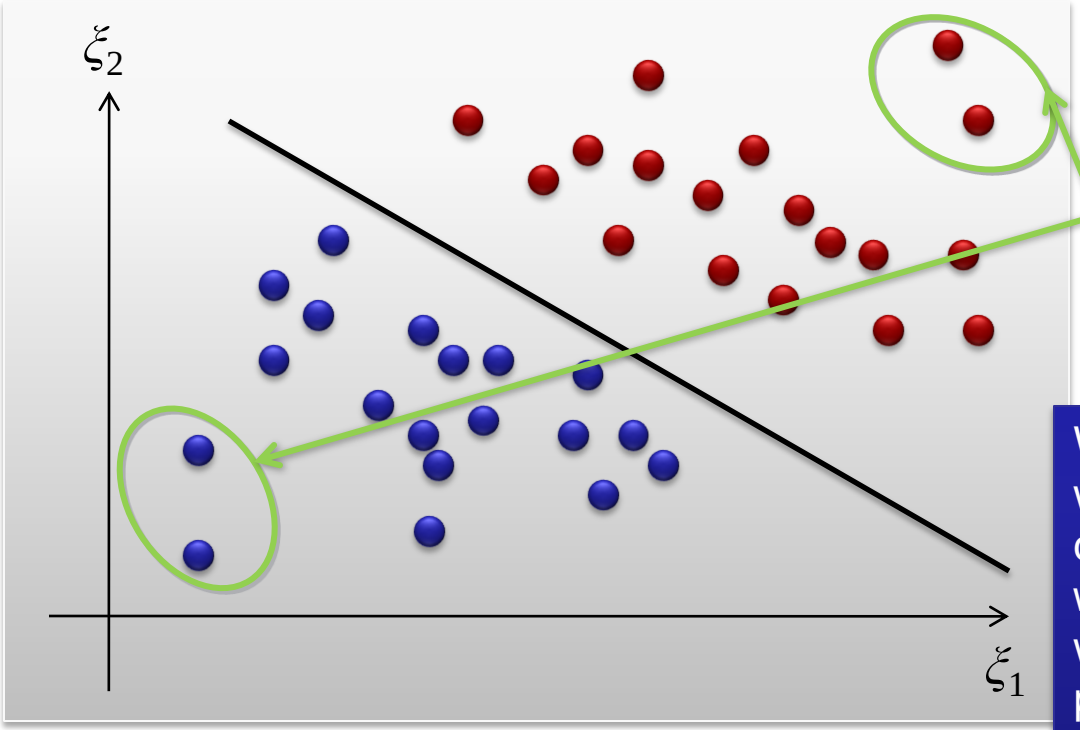
Klasyfikator w trakcie uczenia nie nabywa odporności na zakłócenia, które mogą mieć miejsce podczas akwizycji danych w rzeczywistych zastosowaniach.

Sztuczne dodawanie szumu do zbioru uczącego

Jeżeli jednak atrybut klasy zawiera szum, to trening należy przeprowadzić na „czystym” zbiorze danych.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Wykrywanie wektorów oddalonych



Wektory oddalone — statystycznie niewiarygodne elementy zbioru danych; potencjalne artefakty.

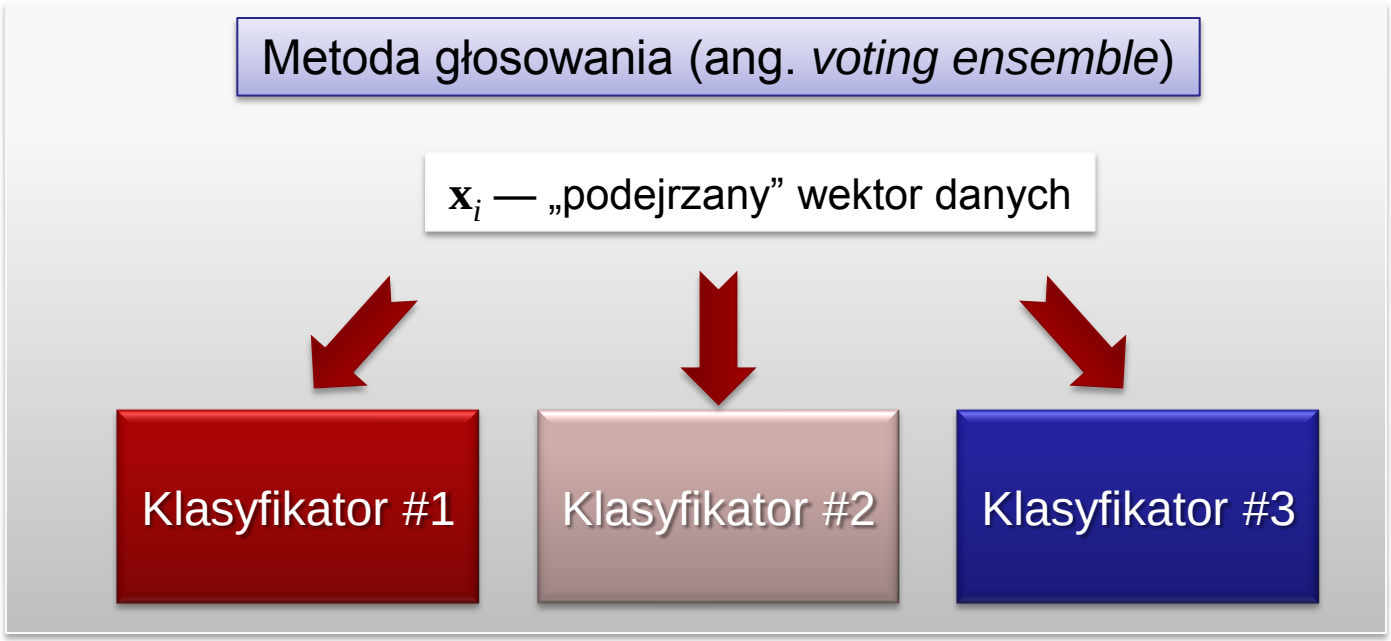
W przypadku 2 lub 3 wymiarów, wystarcza często wizualizacja danych, aby samodzielnie wykryć wektory oddalone. W przypadku większej liczby wymiarów, należy przeprowadzić wcześniej analizę PCA.

Automatyczne wykrycie wektorów oddalonych:
- obliczanie odchylenia od średniej
- obliczanie odległości od prostej aproksymującej dane

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Detekcja anomalii

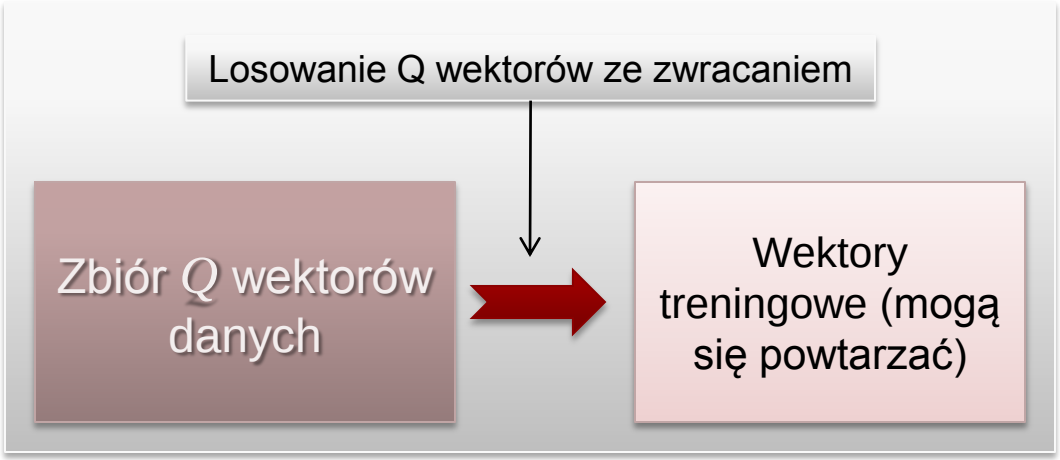
Arbitralne usunięcia wektora oddalonego lub źle klasyfikowanego — bez konsultacji z ekspertem w danej dziedzinie — może być błędem. Dane mogą być wiarygodne, tylko zastosowany model klasyfikatora do nich nie pasuje.



Jeśli każdy algorytm popełnia błąd na $\mathbf{x}_i$, to można taki wektor odrzucić.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Komitety klasyfikatorów — metoda *Bagging*



bagging = bootstrap with aggregating

Symulacja procesu losowania próby wektorów z przestrzeni danych.

Procedurę losowania wykonuje się *T* razy.

T-krotne uczenie wybranego klasyfikatora

Głosowanie: wektor testowy otrzymuje etykietę klasy, która uzyskała najwięcej głosów.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Bagging a klasyfikacja uwzględniająca koszt

W przypadku klasyfikatorów probabilistycznych, dających na wyjściu zestaw prawdopodobieństw przynależności wektora danych do poszczególnych klas, końcowa decyzja komitetu klasyfikatorów podejmowana jest na podstawie uśrednionych wartości tych prawdopodobieństw.

Zwiększona wiarygodność oszacowań kosztów predykcji, lecz trudność analizy działania klasyfikatora.

Algorytm MetaCost

Konstrukcja klasyfikatora metodą *bagging*

Uczenie nowego pojedynczego klasyfikatora na podstawie zbioru danych, uwzględniającego koszt predykcji

Ponowne zaetykietowanie wektorów danych w celu minimalizacji kosztów predykcji

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Randomizacja klasyfikatorów

Podobne efekty, które daje metoda *Bagging*, można uzyskać poprzez wykorzystanie stochastycznych elementów klasyfikatorów.

Przykład

Perceptron liniowy

Ostateczny wynik uczenia zależy od początkowego wyboru wag.

W przypadku, gdy algorytm uczenia klasyfikatora jest deterministyczny, można poddać go procedurze randomizacji.

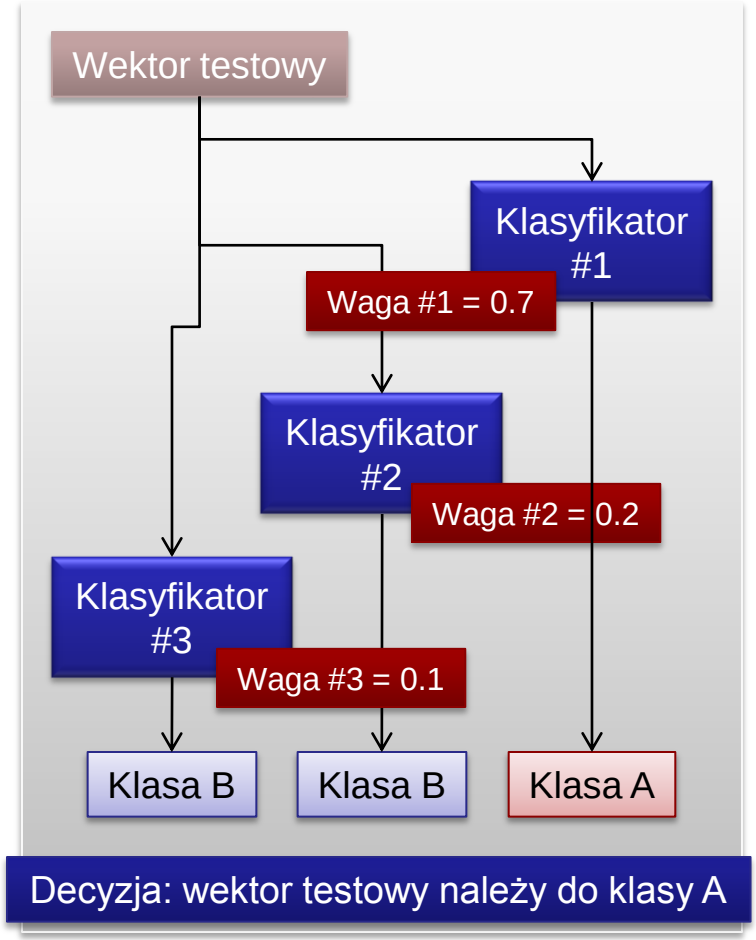
Wielokrotne wykonanie uczenia, z różnym wektorem wag początkowych pozwoli na bardziej stabilne wyniki klasyfikacji.

Np. w algorytmie SVM, wielomianowa lub radialna funkcja jądra wymaga określenia parametru wykładnika. Parametr ten najczęściej ustala się arbitralnie. Można zatem równie dobrze wylosować.

Ostateczna decyzja o klasie wektora testowego ustalana jest albo poprzez głosowanie, albo przez uśrednianie.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Wzmacnianie najlepszych w komitecie — metoda *Boosting*



Bagging	*Boosting*
Klasa jest wyznaczana motodą głosowania	
Komitet składa się z klasyfikatorów tego samego typu (np. drzewa lub reguły decyzyjne)	
Oddzielne uczenie poszczególnych klasyfikatorów	Każdy kolejny model klasyfikatora uczony jest tak, aby nie popełniał błędów poprzedników
Każdy klasyfikator jest jednakowo istotny	Najlepsze klasyfikatory w komitecie mają decydujący głos podczas decydowania o klasie wektorów

Jak uczynić kolejne klasyfikatory komplementarnymi względem siebie?

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytm *AdaBoost.M1*

1. Wektory danych otrzymują wagi o jednakowej wartości

2. Uczenie *m*–tego klasyfikatora z uwzględnieniem wag wektorów danych

3. Wyznaczenie błędu e_m. Jeśli $e_m \in (0, 0.5)$, to uczenie kończy się.

4. Wyznaczenie nowych wag poprawnie sklasyfikowanych wektorów danych:

$$w_{\mathbf{x}_i} = w_{\mathbf{x}_i} \cdot e_m / (1 - e_m)$$

5. Normalizacja wag wszystkich wektorów danych

$$w_{\mathbf{x}_i} = w_{\mathbf{x}_i} \cdot \frac{\sum_{i=1\ldots Q} w_{\mathbf{x}_i}^{stare}}{\sum_{i=1\ldots Q} w_{\mathbf{x}_i}^{nowe}}$$

6. Ustalenie waga klasyfikatorów

$$w_{C_m} = -\log \frac{e_m}{1 - e_m}$$

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Boosting — jak to działa?

2. Uczenie *m*–tego klasyfikatora z uwzględnieniem wag wektorów danych

Niektóre algorytmy klasyfikacji są ze swej istoty dostosowane do uwzględniania wag wektorów (C4.5, SVM). W pozostałych przypadkach trzeba zbiór zawierający wagi odpowiednio przekształcić do zbioru bezwagowego → np. poprzez losowanie wektorów ze zbioru treningowego, przy czym prawdopodobieństwo wylosowania wektorów o większych wagach jest odpowiednio większe.

4. Wyznaczenie nowych wag poprawnie sklasyfikowanych wektorów danych.

5. Normalizacja wag wszystkich wektorów danych

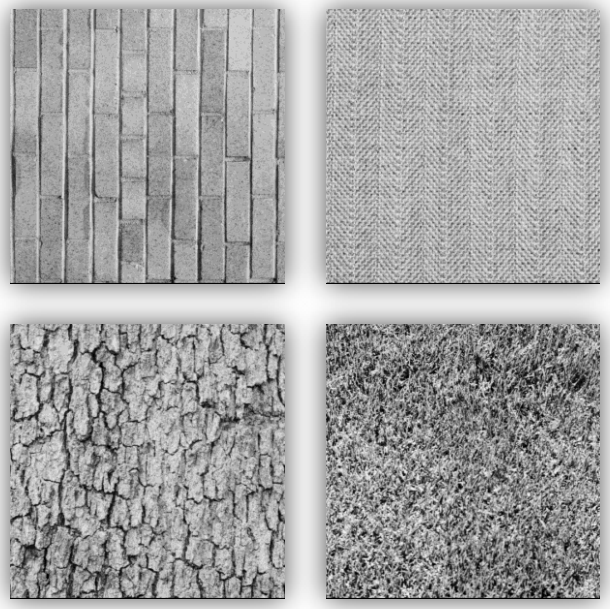
Wektory poprawnie sklasyfikowane powinny mieć mniejszy udział w uczeniu kolejnych klasyfikatorów. Dlatego ich wagi są zmniejszane. Normalizacja zapewnia z kolei zwiększenie wag wektorów źle sklasyfikowanych.

6. Ustalenie waga klasyfikatorów

Waga klasyfikatora (i jego wpływ na decyzje całego komitetu) ustalana jest na podstawie popełnianego błędu. Bezbłędne klasyfikatory nie są jednak w ogóle brane pod uwagę.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Przykład — zbiór obrazów tekstur



Wyniki klasyfikacji z zastosowaniem algorytmu 1-R

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         206               80.4688 %
Incorrectly Classified Instances        50                9.5313 %
Kappa statistic                          0.7396
```

Wyniki klasyfikacji z zastosowaniem algorytmów 1-R oraz AdaBoost.M1

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         222               86.7188 %
Incorrectly Classified Instances        34               13.2813 %
Kappa statistic                          0.8229
```

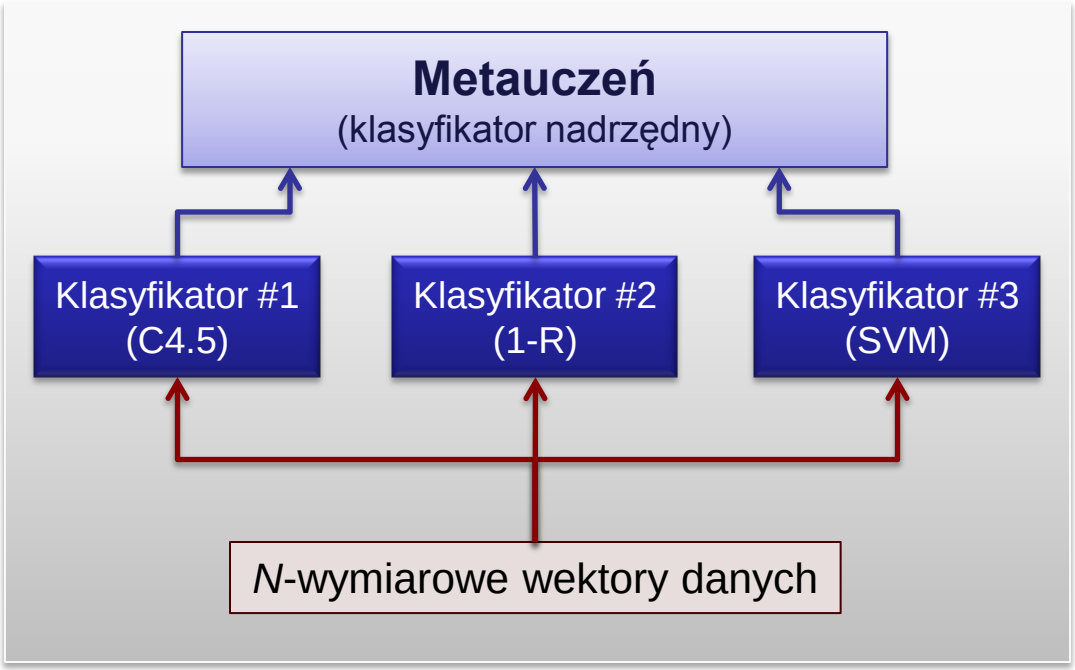
Wyniki klasyfikacji z zastosowaniem algorytmów 1-R oraz metody Bagging

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         207               80.8594 %
Incorrectly Classified Instances        49               19.1406 %
Kappa statistic                          0.7448
```

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Kombinacja różnych klasyfikatorów — metoda *Stacking*

Stacking (ang. *stacked generalization*) — uogólnianie warstwowe



W przypadku 3 klasyfikatorów dających na wyjściu jednoznaczne wskazanie na kategorię klasyfikowanego wektora, metauczeń ma na wejściu wektor o 3 składowych nominalnych.

W przypadku *m* algorytmów obliczających prawdopodobieństwo przynależności wektorów do *K* klas, wektor danych dla metaucznia posiada *K*m* składowych rzeczywistych.

Głosowanie jest zastąpione decyzją nadrzędnego klasyfikatora, który może nauczyć się rozpoznawania bardziej i mniej wiarygodnych algorytmów podrzędnych.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Uczenie metaucznia

Sposób 1 — niezalecany

Zbiór danych treningowych w całości wykorzystywany do konstrukcji klasyfikatorów podrzędnych.

Wyjścia klasyfikatorów dla poszczególnych wektorów danych wraz z odpowiadającymi im etykietami klas tworzy zbiór danych treningowych dla metaucznia.

Problem: jeden z klasyfikatorów, np. taki, który nadmiernie dopasuje się do danych, może zdominować zaufanie metaucznia.

Sposób 2 — właściwy

Zbiór danych podzielony na dwie niezależne części:

1. Część ucząca klasyfikatory podrzędne (np. 66% całego zbioru danych)
2. Część testowa (np. 33% oryginalnego zbioru) do testowania klasyfikatorów podrzędnych

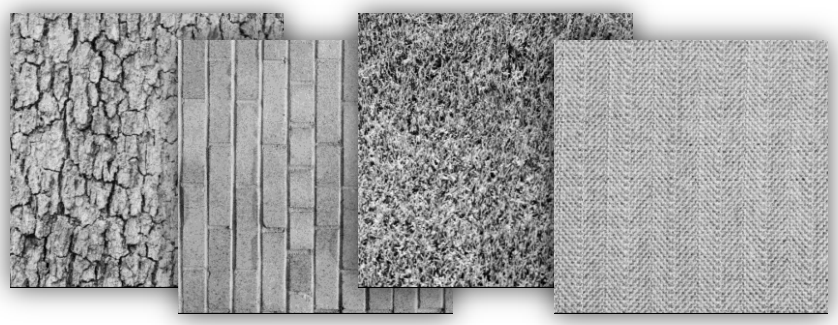
Wyjścia klasyfikatorów dla poszczególnych wektorów danych **z części testowej** wraz z odpowiadającymi im etykietami klas tworzy zbiór danych treningowych dla metaucznia.

Problem: klasyfikatory podrzędne nie wykorzystują w całości zbiorów danych treningowych

Jaki algorytm wybrać do nauki klasyfikatora nadrzędnego? Możliwie jak najprostszy.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Metoda *Stacking* w zastosowaniu do zbiorów tekstur



=== Stratified cross-validation ===
=== Summary ===

Algorytm 1-R

Correctly Classified Instances	206	80.4688 %
Incorrectly Classified Instances	50	9.5313 %
Kappa statistic	0.7396	

=== Stratified cross-validation ===
=== Summary ===

Algorytmy 1-R oraz *Bagging*

Correctly Classified Instances	207	80.8594 %
Incorrectly Classified Instances	49	19.1406 %
Kappa statistic	0.7448	

=== Stratified cross-validation ===
=== Summary ===

AdaBoost.M1

Correctly Classified Instances	222	86.7188 %
Incorrectly Classified Instances	34	13.2813 %
Kappa statistic	0.8229	

=== Stratified cross-validation ===
=== Summary ===

Stacking

Correctly Classified Instances	207	90.9091 %
Incorrectly Classified Instances	49	9.0909 %
Kappa statistic	0.8781	

Metauczeń: C4.5

Algorytmy podrzędne: naiwny klasyfikator Bayesa, SVM, 1-R