KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

dr inż. Artur Klepaczko

Eksploracja danych
Uczenie nienadzorowane

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

Metody nienadzorowane

Modyfikacje algorytmu *k*–średnich

Grupowanie rozmyte

Fuzzy c–means

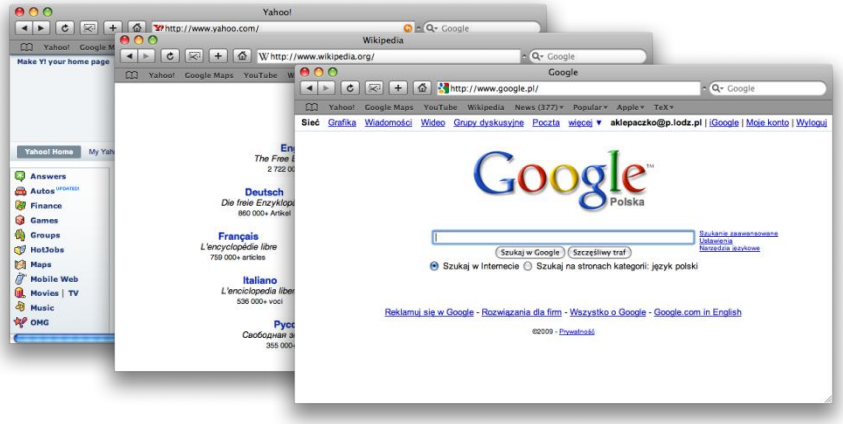
Grupowanie probabilistyczne

Nienadzorowana selekcja cech

Metody częściowo nadzorowane

Analiza skupień w klasyfikacji

Co–training



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Automatyczne wyznaczanie liczby klasterów

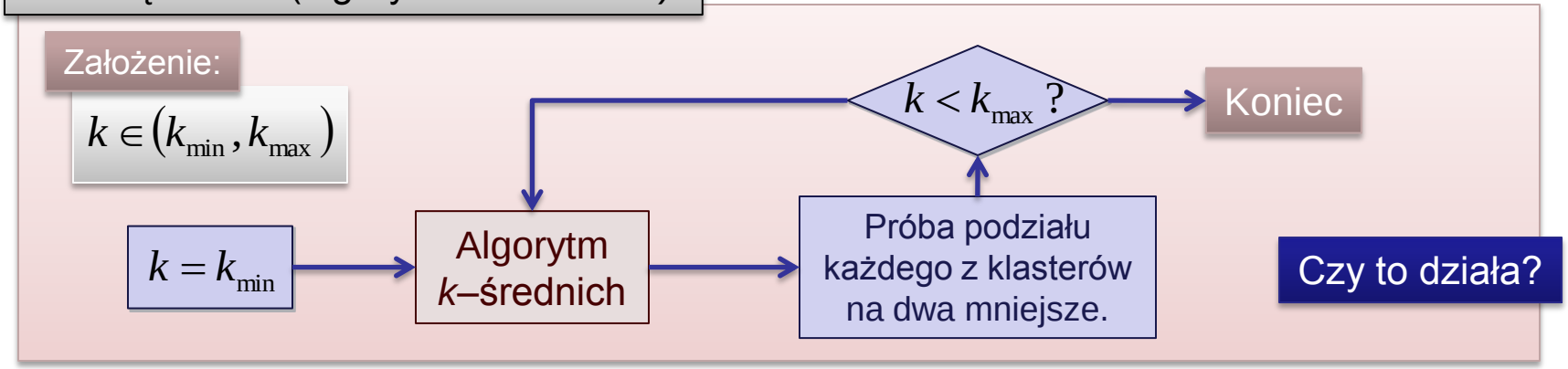
Kluczowy problem w przypadku klasycznego algorytmu *k*–średnich wiąże się z koniecznością określenia z góry właściwej liczby klasterów. W rzeczywistych zastosowaniach grupowania, informacja taka może być niedostępna.

Rozwiązanie 1:

Algorytm *k*–średnich wykonuje się kilkakrotnie, za każdym razem zakładając inną liczbę klasterów. Ostateczną decyzję o liczbie skupień podejmuje się na podstawie porównania poszczególnych podziałów pod kątem określonej miary jakości klasteryzacji.

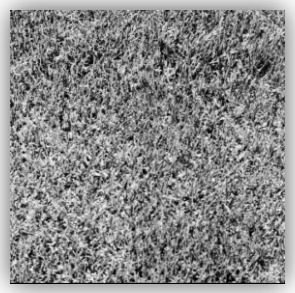
Czasochłonne ☹

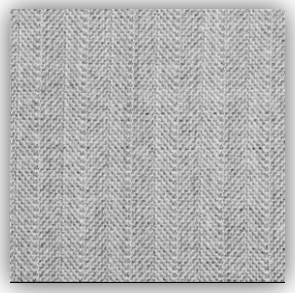
Rozwiązanie 2 (algorytm X–średnich):

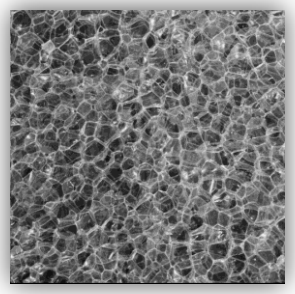


Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytm X–średnich w zastosowaniu do zbioru tekstur







Class attribute: category
Classes to Clusters:
  0 1 2 3 <-- assigned to cluster
112 126 18 0 | grass
117 139 0 0 | herringbone_weave
  0 13 142 101 | plastic_bubbles
Cluster 0 <-- grass
Cluster 1 <-- herringbone_weave
Cluster 2 <-- plastic_bubbles
Cluster 3 <-- No class

Incorrectly clustered instances : 375.0 48.8281 %

X-średnich, założony przedział dla liczby klaterów: (2,10)

Duży błąd, ale zbliżony do wyniku uzyskanego dla standardowego algorytmu *k*-średnich. Jaka może być przyczyna tego błędu?

Zbiór danych: 3 klasy, 256 wektorów danych w klasie, 10 cech tekstury (wybranych w trybie nadzorowanym – przy użyciu współczynnika Fishera)

k-średnich

Class attribute: category
Classes to Clusters:
Cluster 0 <-- grass
Cluster 1 <-- plastic_bubbles
Cluster 2 <-- herringbone_weave
Incorrectly clustered instances : 346.0 45.0521 %

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Grupowanie rozmyte

Wagi przynależności wektorów danych do poszczególnych centrów skupień

$$u_{ij} = \frac{1}{\left\| \mathbf{x}_i - \boldsymbol{\mu}_j \right\|}$$

Im dalej od centrum klastera, tym mniejszy jest skojarzony z nim stopień przynależności danego wektora.

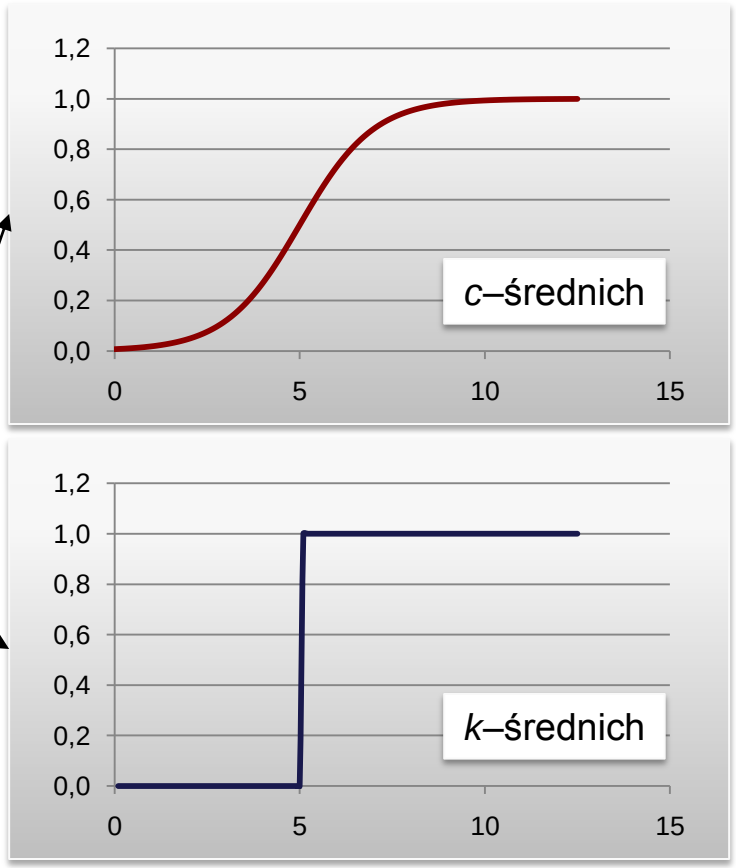
Centra klasterów

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{Q} u_{ij}^m \cdot \mathbf{x}_i}{\sum_{i=1}^{Q} u_{ij}^m}$$

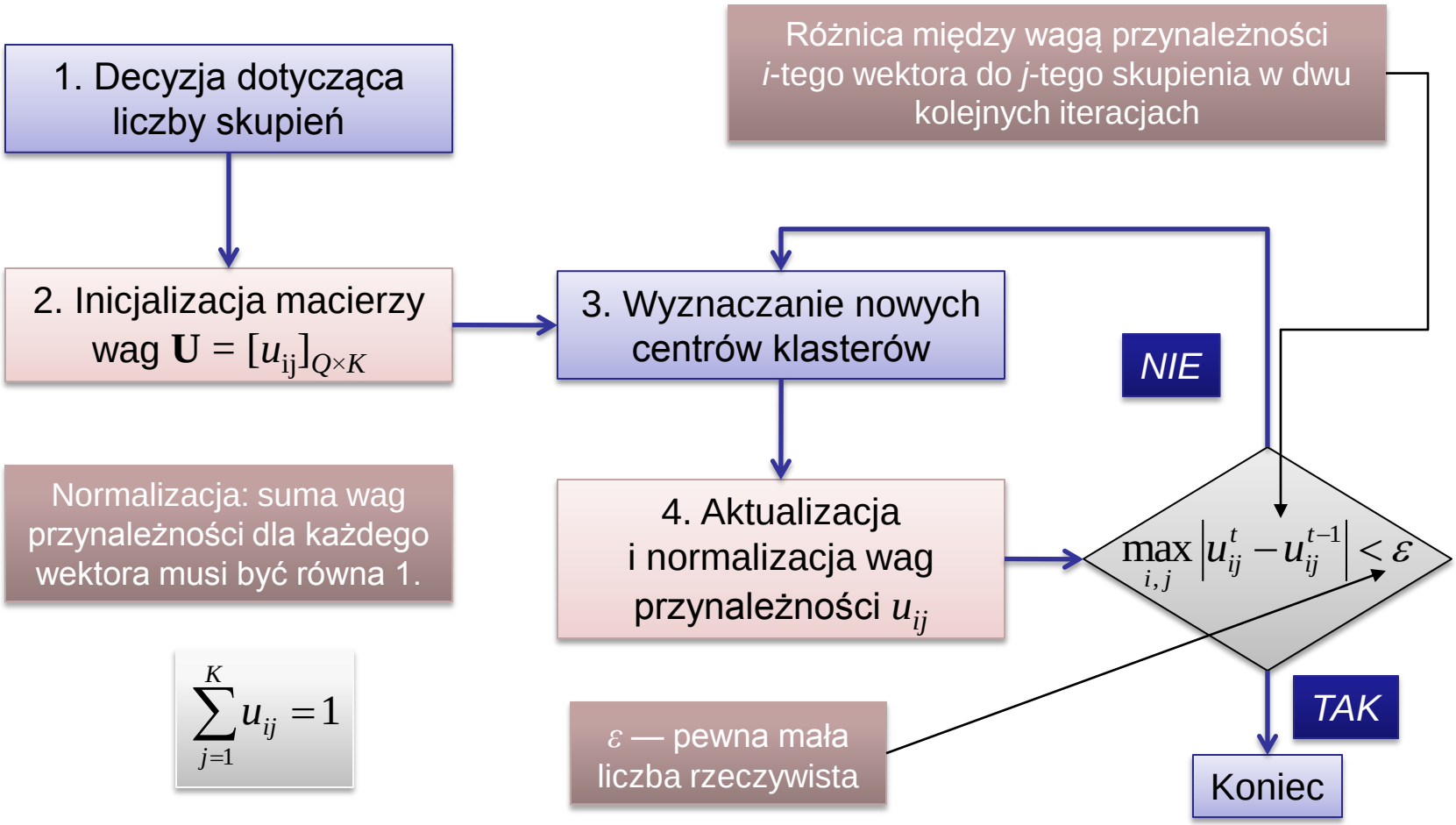
Przypadek wektorów jednowymiarowych

$m \geq 1$ — parametr fuzyfikacji (gdy m≈1, algorytm zbliża się do klasycznego *k*–średnich

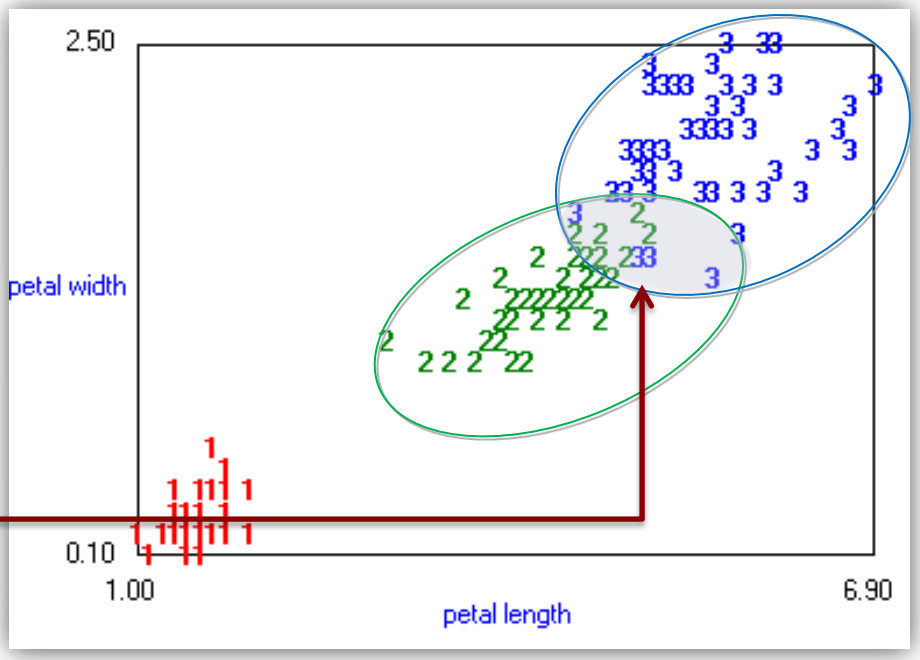
Funkcja przynależności wektorów do klasterów



Algorytm *c*–średnich



Grupowanie probabilistyczne



Zbiór *Iris*, http://archive.ics.uci.edu/ml/

Wektory leżące na granicy klasterów, z niemal równym prawdopodobieństwem należą do nich obu.

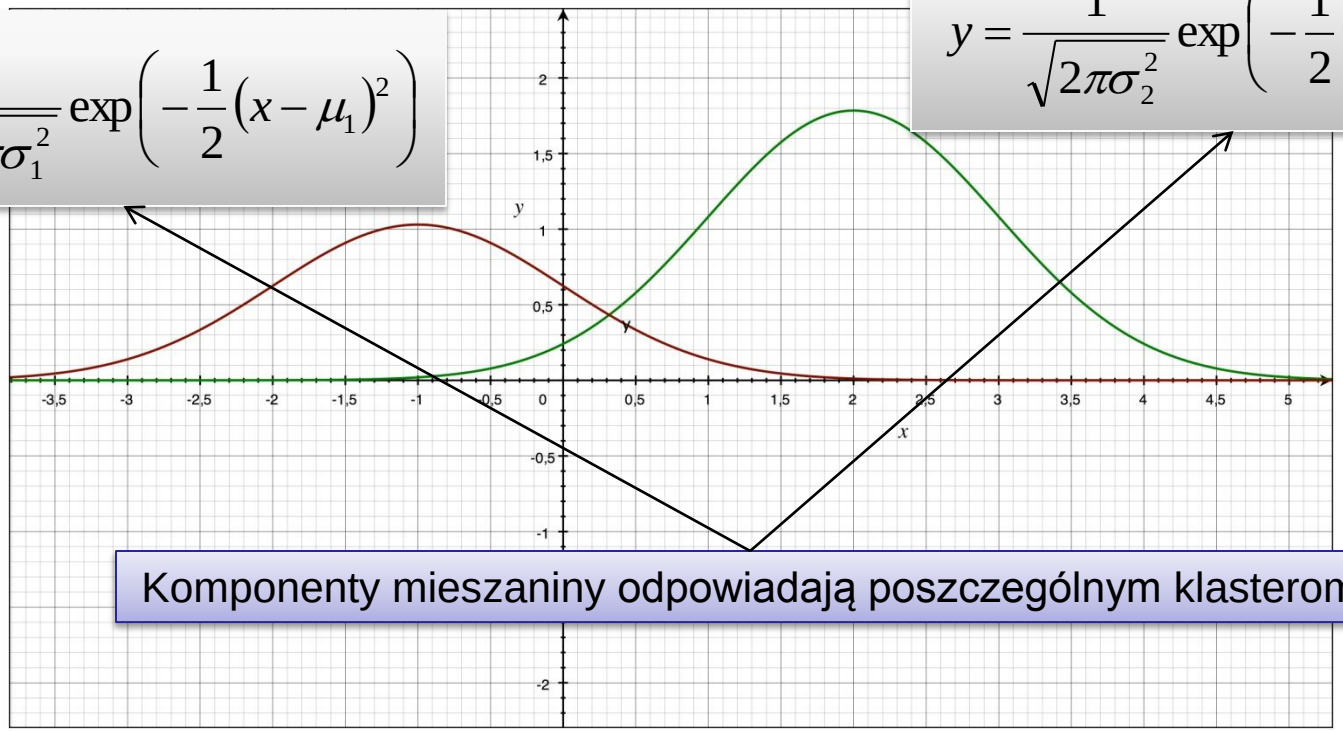
Grupowanie probabilistyczne wyznacza prowdopodobieństwo przynależności wektorów danych do poszczególnych klasterów.

Nr wektora	Klaster		
	#1	#2	#3
1	0.1	0.2	0.7
2	0.3	0.3	0.4
3	0.8	0.2	0.0
…	…	…	…
i	0.2	0.5	0.3

Gaussowski model mieszany

$$y = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2}(x-\mu_1)^2\right)$$

$$y = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2}(x-\mu_2)^2\right)$$

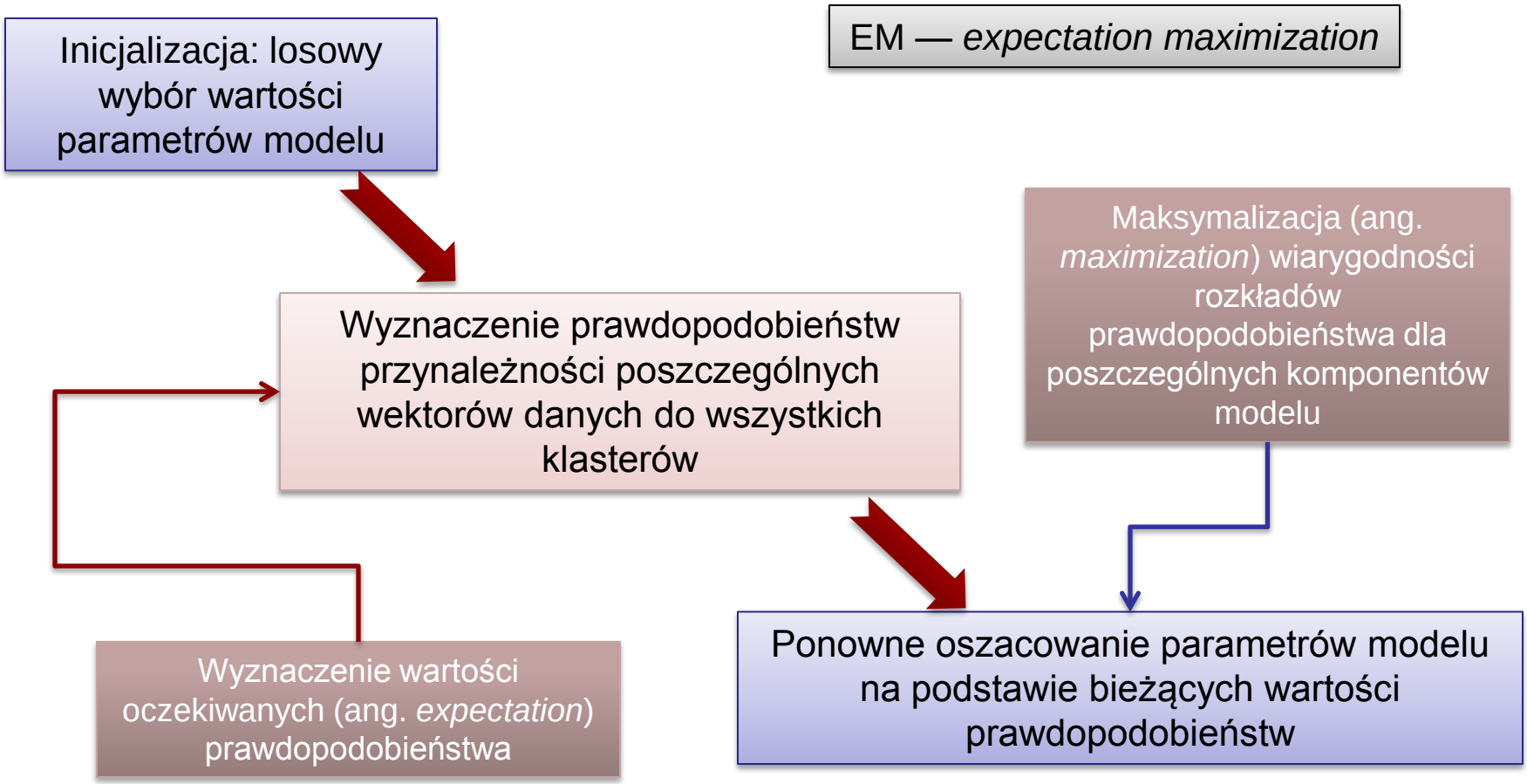


Komponenty mieszaniny odpowiadają poszczególnym klasterom

Parametry modelu: prawdopodobieństwa komponentów, ich wartości średnie oraz odchylenia standardowe.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytm EM



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Szacowanie parametrów modelu

Wartość średnia

$$\mu_A = \frac{\sum_{i=1}^{Q} \Pr[x_i \mid A] \cdot x_i}{\sum_{i=1}^{Q} \Pr[x_i \mid A]}$$

Wariancja

$$\sigma_A^2 = \frac{\sum_{i=1}^{Q} \Pr[x_i \mid A](x_i - \mu_A)^2}{\sum_{i=1}^{Q} \Pr[x_i \mid A]}$$

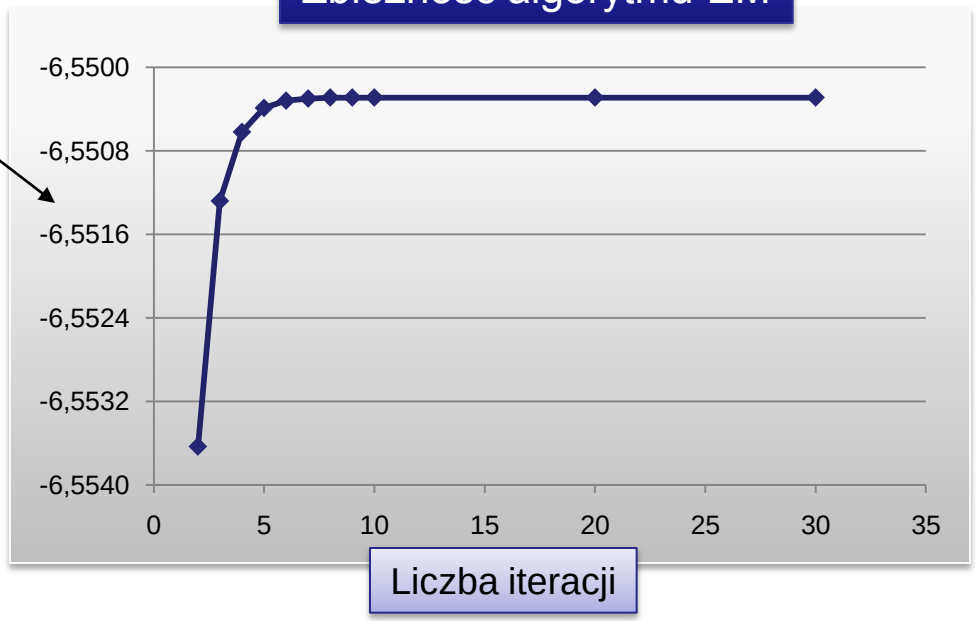
Prawdopodobieństwo przynależności *i*-tego wektora do klastera A

Logarytmiczne prawdopodobieństwo modelu

$$\log \prod_i (p_A \Pr[\mathbf{x}_i \mid A] + p_B \Pr[\mathbf{x}_i \mid B])$$

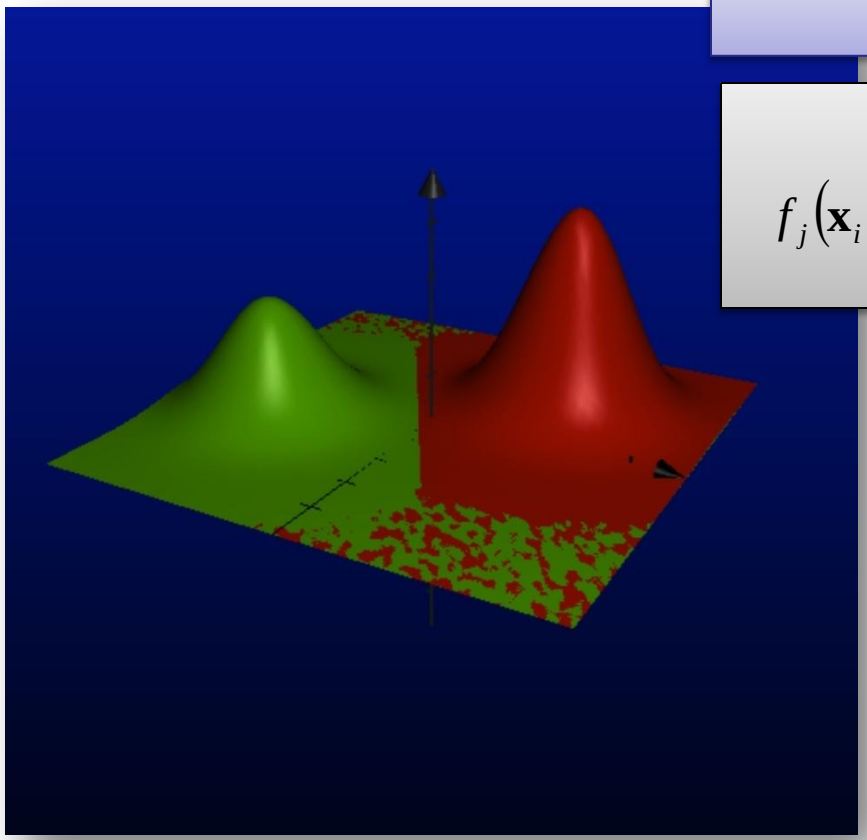
Prawdopodobieństwa klasterów A oraz B

Zbieżność algorytmu EM



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Model wielowymiarowy



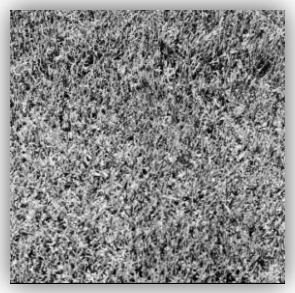
Ogólna postać N-wymiarowego rozkładu Gaussa dla pojedynczego komponentu mieszaniny

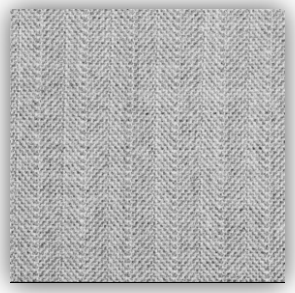
$$f_j\left(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^T \cdot \boldsymbol{\Sigma}_j^{-1} \cdot \left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\right\}}{\left(2\pi\right)^{N/2}\sqrt{\det \boldsymbol{\Sigma}_j}}$$

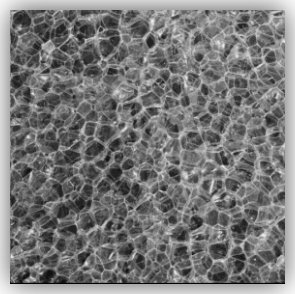
Powyższy model należy stosować wówczas, gdy poszczególne atrybuty wektorów są ze sobą skorelowane. Wówczas macierz kowariancji posiada niezerowe elementy także poza główną przekątną.

Jednakże przyjmując (naiwne) założenie o braku korelacji pomiędzy cechami, rozkład wektorów można modelować iloczynem rozkładów jednowymiarowych.

Przykład — grupowanie tekstur







Class attribute: category
Classes to Clusters:

  0   1   2   3 <-- assigned
112 126 18   0 | grass
117 139   0   0 | herringbon
  0 13 142 101 | plastic_bu

Cluster 0 <-- grass
Cluster 1 <-- herringbone_weave
Cluster 2 <-- plastic_bubbles
Cluster 3 <-- No class

Incorrectly clustered instances : 375.0 48.8281 %

Algorytm X-średnich

Class attribute: category
Classes to Clusters:

  0   1   2 <-- assigned to cluster
220 36   0 | grass
255   1   0 | herringbone_weave
  5 97 154 | plastic_bubbles

Cluster 0 <-- herringbone_weave
Cluster 1 <-- grass
Cluster 2 <-- plastic_bubbles

Incorrectly clustered instances : 323.0 42.0573 %

Algorytm EM

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Nienadzorowana redukcja wymiaru danych

W przypadku nienadzorowanej redukcji wymiaru danych, zbiór wektorów danych nie zawiera jednoznacznej sugestii w postaci etykiet kategorii wektorów danych, pozwalających powiązać informację zawartą w cechach wektorów z ich prawdziwym podziałem na klasy.

Ekstrakcja cech

– PCA
– *Random Projection*

Obiektywizm analizy — nowe cechy uwzględniają określone statystyczne właściwości całego zbioru danych.

Agregaty cech nie posiadają interpretacji fizycznej, a także niekoniecznie ułatwiają klasyfikację (są reprezentacyjne, a nie dyskryminacyjne)

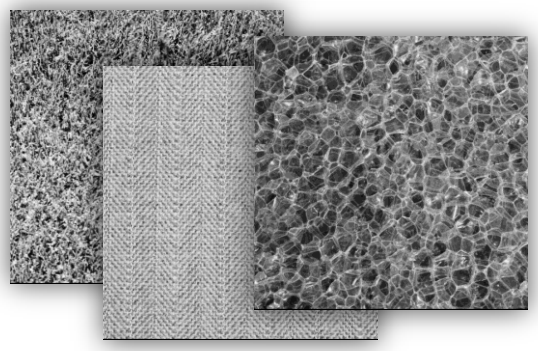
Selekcja cech

– Metody typu *filtr*
– Metody typu *powłoka* (oparte o heurystyczne miary jakości klasteryzacji)

Wyselekcjonowane cechy uwzględniają arbitralnie przyjęte miary jakości, wedle których preferowane są określone właściwości skupień.

Wybrane cechy posiadają oryginalną interpretację.

Ekstrakcja cech w zastosowaniu do zbioru tekstur



Zbiór danych: 3 klasy, 256 wektorów danych w klasie, 250 cech tekstury. ROI: 24×24 piksele.

Wyniki grupowania EM po ekstrakcji cech

Analiza PCA

Class attribute: category
Classes to Clusters:

```
 0   1   2  <-- assigned to cluster
176  80   0 | grass
143   1 112 | herringbone_weave
  3 253   0 | plastic_bubbles
```

Incorrectly clustered instances : 203.0 26.4323 %

Transformacja losowa

Class attribute: category
Classes to Clusters:

```
  0   1   2  <-- assigned to cluster
 51 204   1 | grass
  0 137 119 | herringbone_weave
242  14   0 | plastic_bubbles
```

Incorrectly clustered instances : 227.0 29.5573%

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Nienadzorowana selekcja cech — algorytm RANK

Miara podobieństwa między parą wektorów danych

$$S(\mathbf{x}_r, \mathbf{x}_s) = \exp\left(\frac{-\ln(0.5)\|\mathbf{x}_r - \mathbf{x}_s\|}{\bar{d}}\right)$$

Entropia zbioru wektorów danych

$$E = \sum_{r=1}^{Q}\sum_{s=1}^{Q}\left[S_{rs}\log(S_{rs}) + (1-S_{rs})\log(1-S_{rs})\right]$$

1. Wyznaczenie entropii E dla całego zbioru danych.

2. Wyznaczenie entropii E_j po usunięciu j-tego atrybutu z przestrzeni cech (j=1..N)

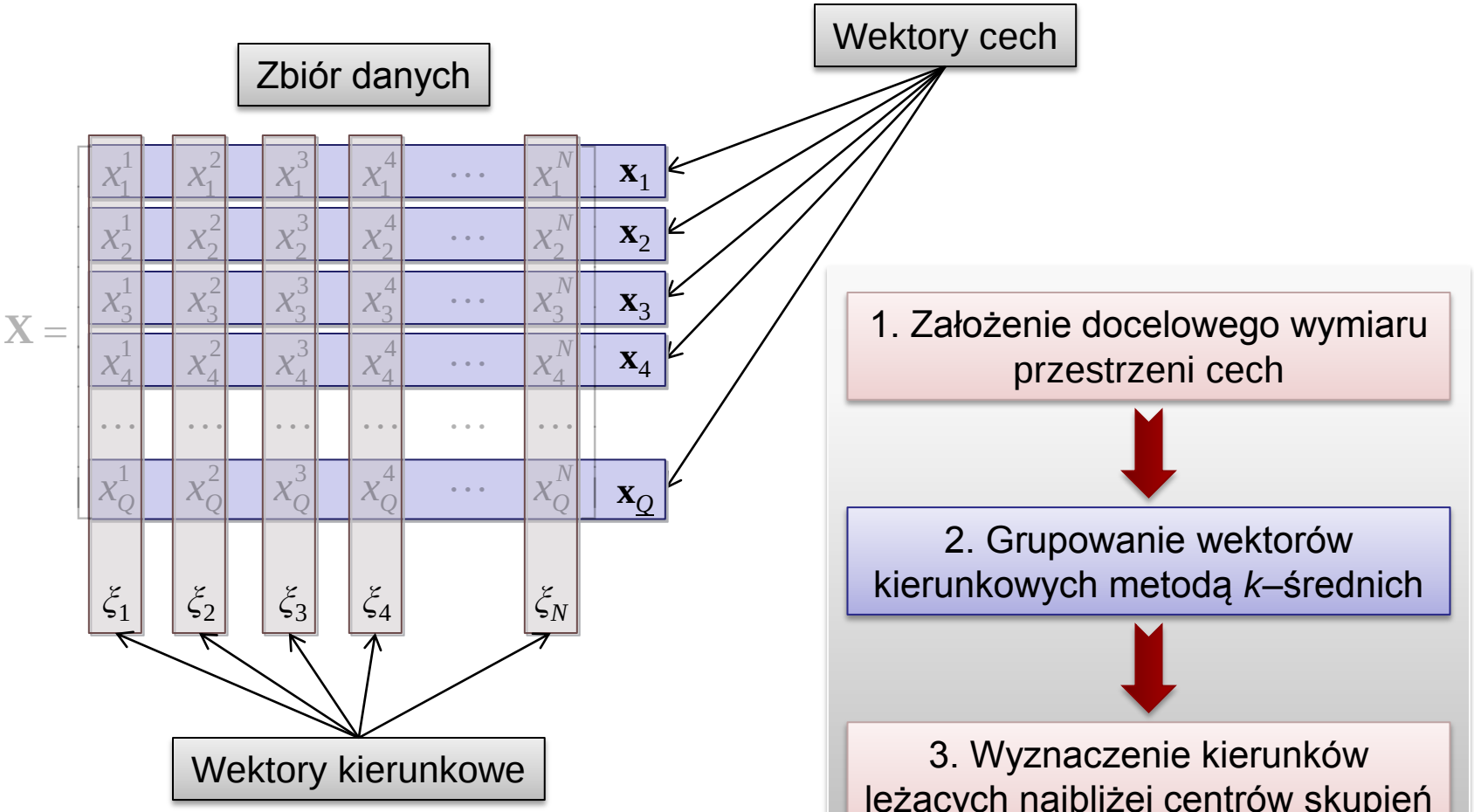
3. Uszeregowanie cech według ich wpływu na entropię zbioru danych. Pierwsza cecha w rankingu w największym stopniu obniża (lub w najmniejszym stopniu podwyższa) entropię zbioru.

Zaleta: prostota i niezależność od algorytmu klasteryzacji.

Wada: najlepsze cechy w rankingu niekoniecznie odpowiadają kierunkom najlepszej dyskryminacji

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Grupowanie cech



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Metody typu *powłoka*

Przeszukiwanie przestrzeni cech

- Sekwencyjne (SFS, SFFS)
- Ewolucyjne (algorytmy genetyczne)

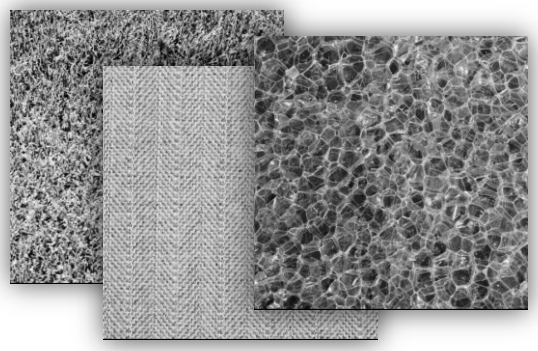
Miary oceny podzbiorów cech

- Miary jakości klasterów utworzonych w danej podprzestrzeni atrybutów
  - Współczynnik kształtu i podobne miary maksymalizujące rozproszenie zewnętrzne przy jednoczesnej minimalizacji rozproszenia wewnętrznego
  - Kryterium MDL
- Entropia zbioru wektorów danych uwzględniająca podział na klastery

Wektory danych mogą w różnych podprzestrzeniach cech tworzyć różne skupienia dobrej jakości → nienadzorowane miary oceny podzbiorów cech mają wiele optimów lokalnych → przeszukiwanie sekwencyjne jest nieefektywne.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Selekcja cech dla zbioru tekstur



Losowy charakter operacji w algorytmie genetycznym uniezależnia wybory dokonywane na danym etapie przeszukiwania od wcześniej uzyskanych wyników. Dzięki temu możliwe jest do pewnego stopnia unikanie optimów lokalnych reprezentujących nieprawidłowe rozwiązania.

Wyniki grupowania *k*–średnich po selekcji cech

Przeszukiwanie sekwencyjne

Classes to Clusters:

```
  0   1   2  <-- assigned to cluster
111   0  17 | grass
 33   0  95 | herringbone_weave
 10  39  79 | plastic_bubbles
```

Incorrectly clustered instances : 139.0 36.1979 %

Algorytm genetyczny

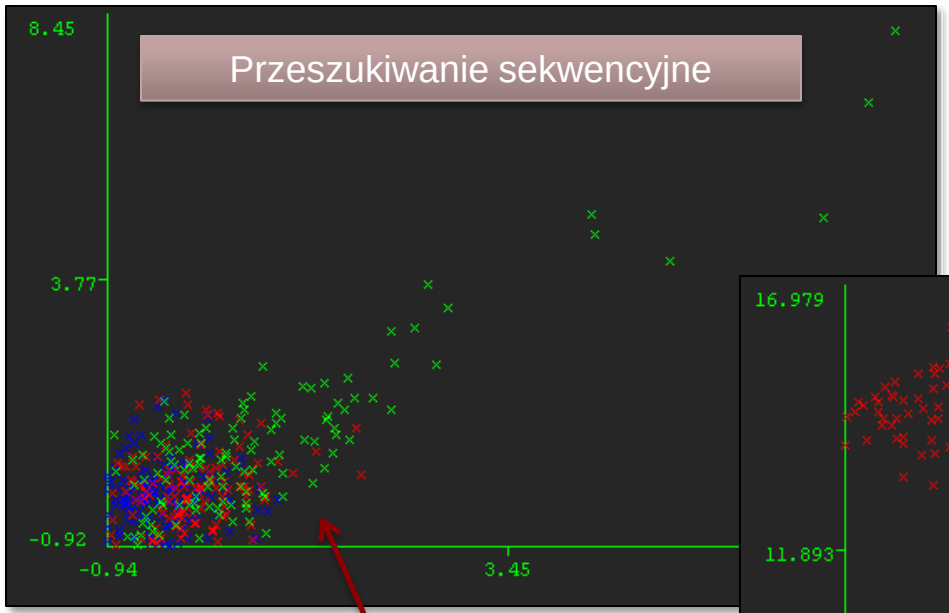
Classes to Clusters:

```
  0   1   2  <-- assigned to cluster
122   4   2 | grass
 68  60   0 | herringbone_weave
 13   0 115 | plastic_bubbles
```

Incorrectly clustered instances : 87.0 22.6563 %

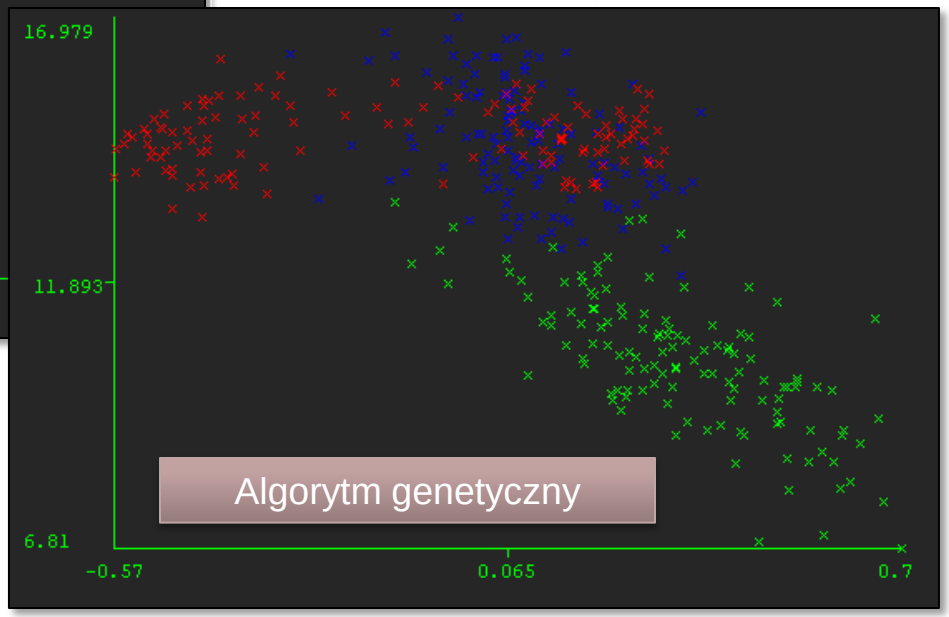
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Wizualizacja wyznaczonych przestrzeni cech



Przeszukiwanie sekwencyjne

Widać odrębne skupienia, lecz jeden z klasterów zawiera wektory z dwóch klas



Algorytm genetyczny

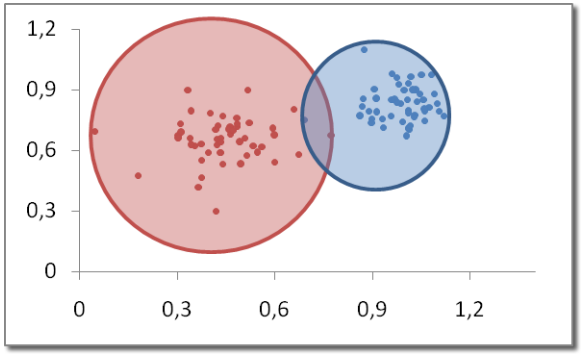
Wektory nie tworzą odrębnych klasterów

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

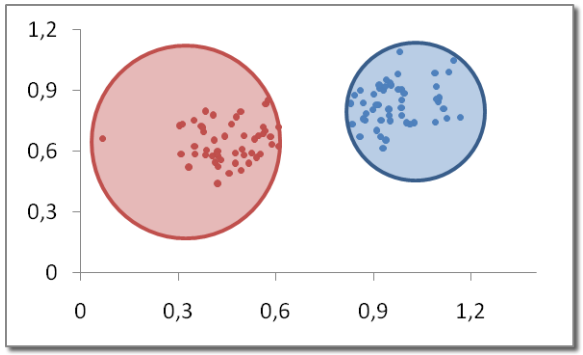
(Nie)stabilność grupowania

Jakość grupowania nie jest uniwersalnym kryterium selekcji cech. Konieczna jest inna nienadzorowana miara istotności atrybutów. Istnieją zbiory danych, w których prawdziwe klasy wektorów cech nie tworzą skupień dobrej jakości.

Niestabilność grupowania — dwie niezależne próby wektorów danych należących do tej samej dziedziny tworzą skupienia o odmiennej strukturze.



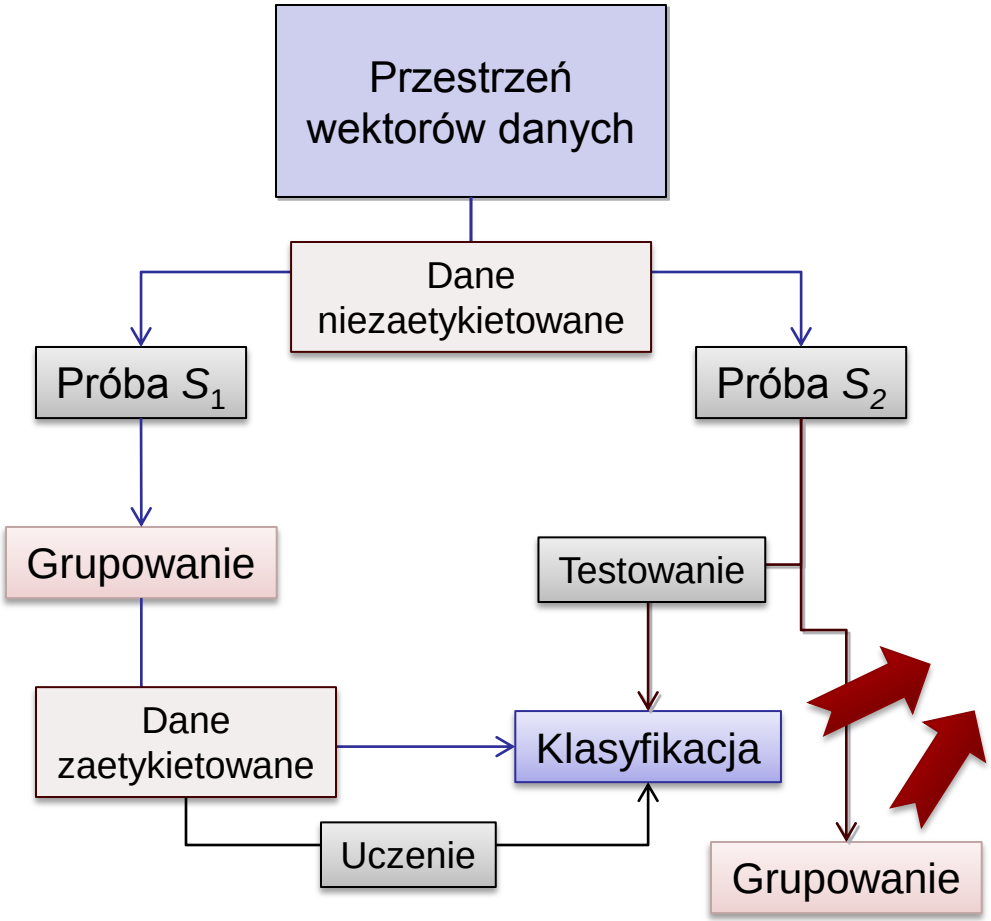
Próba S_1



Próba S_2

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Szacowanie stopnia niestabilności grupowania



Dla każdej pary wektorów danych:

Wektory należą do tej samej klasy	Wektory należą do tego samego klastera	
	TAK	NIE
TAK	$s_{ij} = 0$	$s_{ij} = 1$
NIE	$s_{ij} = 1$	$s_{ij} = 0$

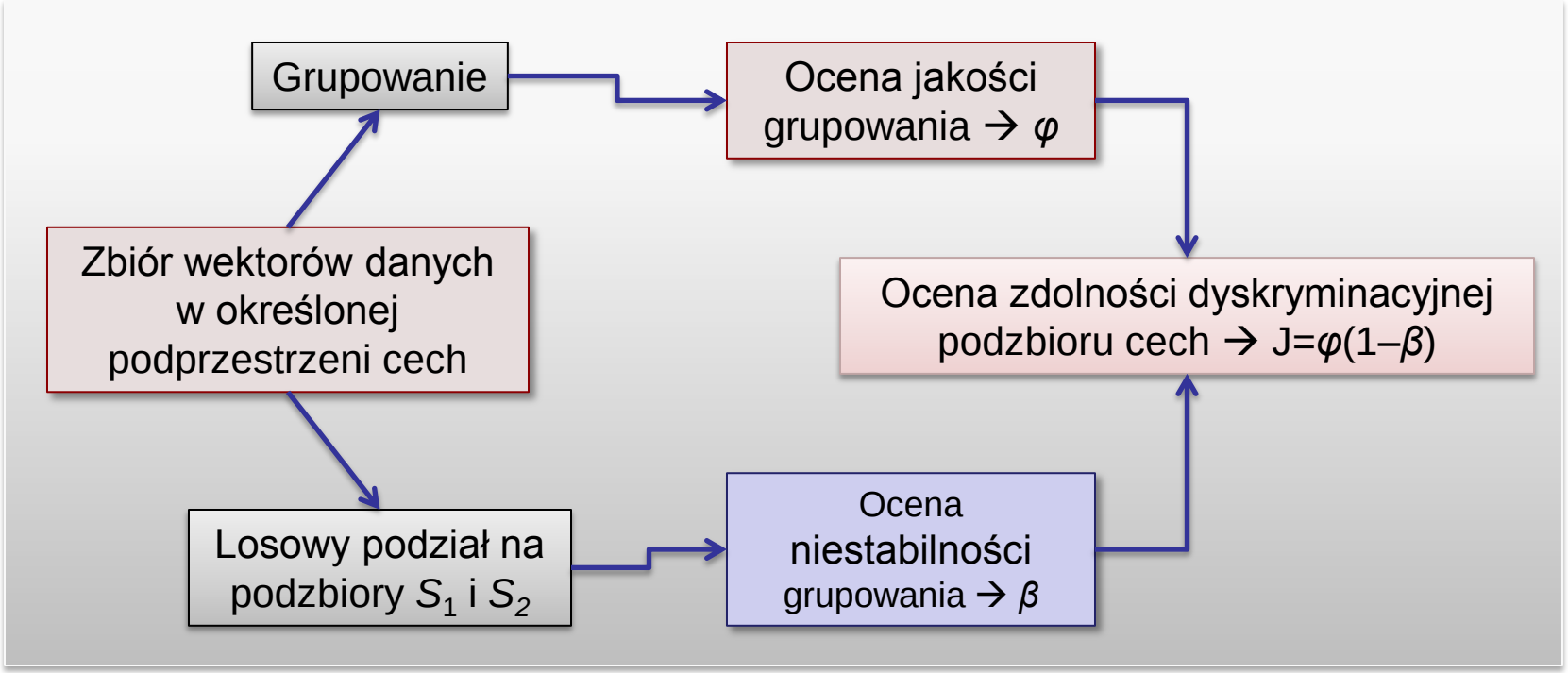
Miara niestabilności

$$\beta = \frac{2}{Q(Q-1)} \sum_{i=1}^{Q} \sum_{j=1}^{Q} s_{ij}$$

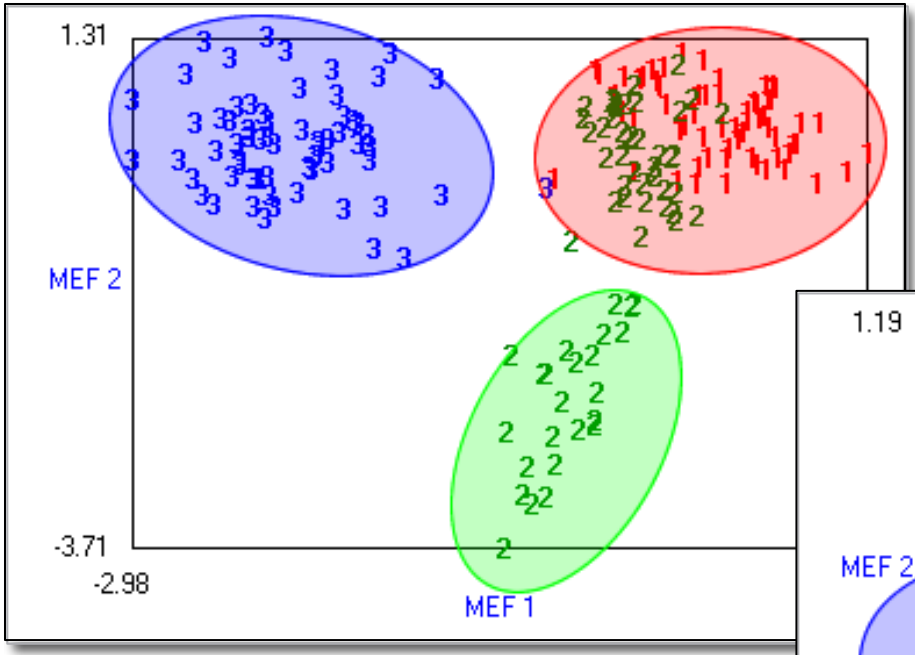
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Selekcja cech z uwzględnieniem stabilności grupowania

Cechy zawierające istotną informację dla poprawnej klasyfikacji wektorów danych zapewniają większą stabilność grupowania niż parametry nieznaczące.

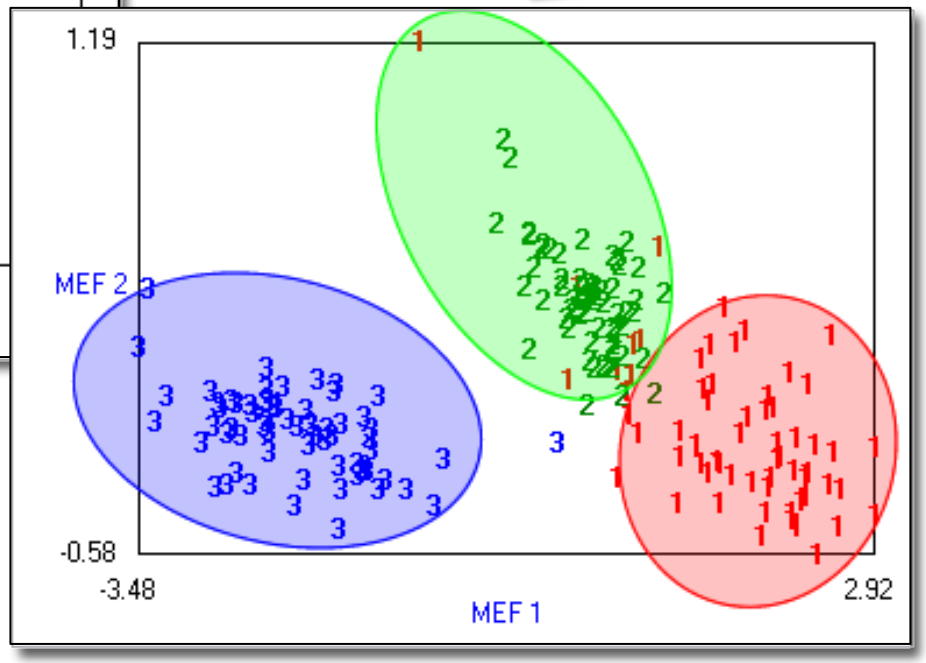


Przykład — zbiory tekstur raz jeszcze



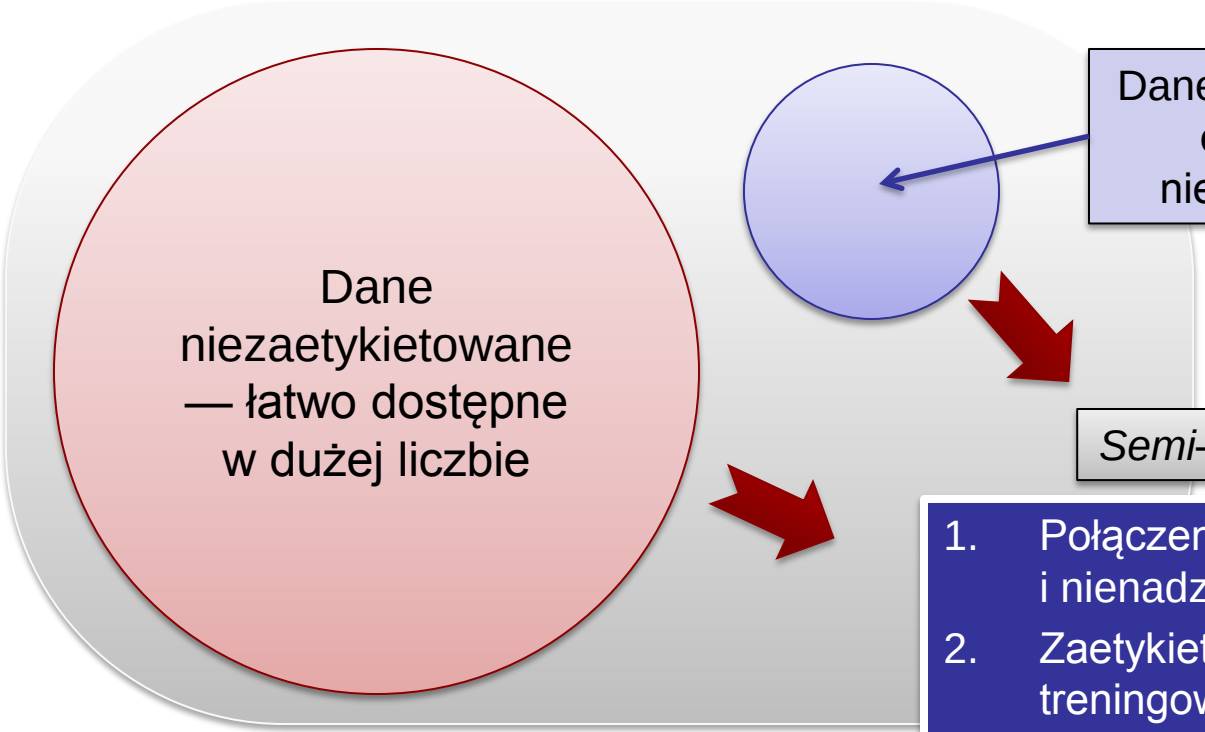
Selekcja cech tylko na podstawie jakości grupowania

Selekcja uwzględniająca stabilność i jakość grupowania



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Uczenie częściowo nadzorowane



Dane zaetykietowane — bardzo duży koszt pozyskania; niedostępne w dużej liczbie

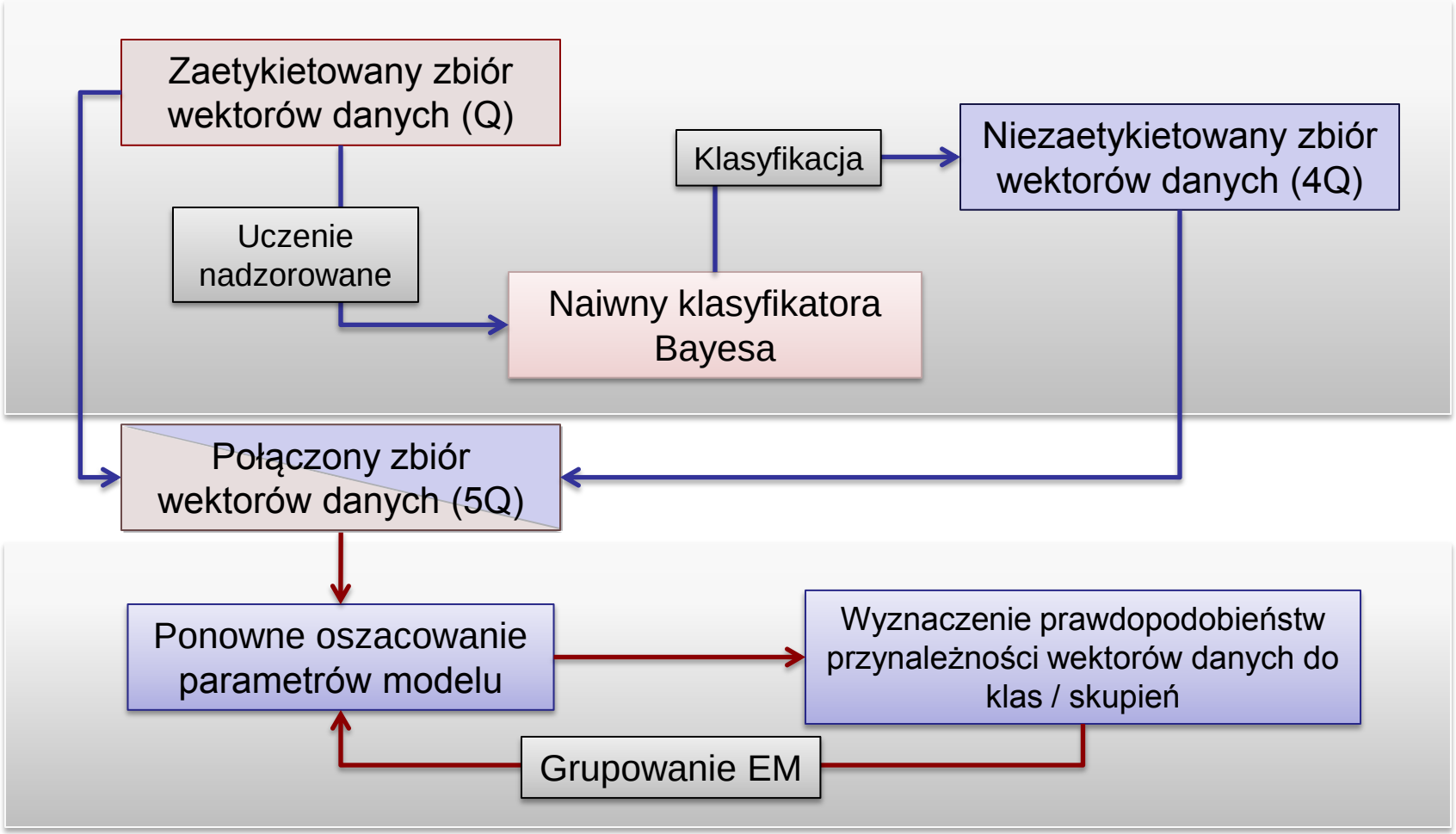
Semi–supervised learning

1. Połączenie uczenia nadzorowanego i nienadzorowanego
2. Zaetykietowany zbiór danych treningowych jest zbyt mały, aby utworzyć wiarygodny klasyfikator
3. Uwzględnienie informacji zawartej w zbiorze niezaetykietowanym pozwala poprawić jakość klasyfikacji

Przykład: dokumenty tekstowe, strony internetowe

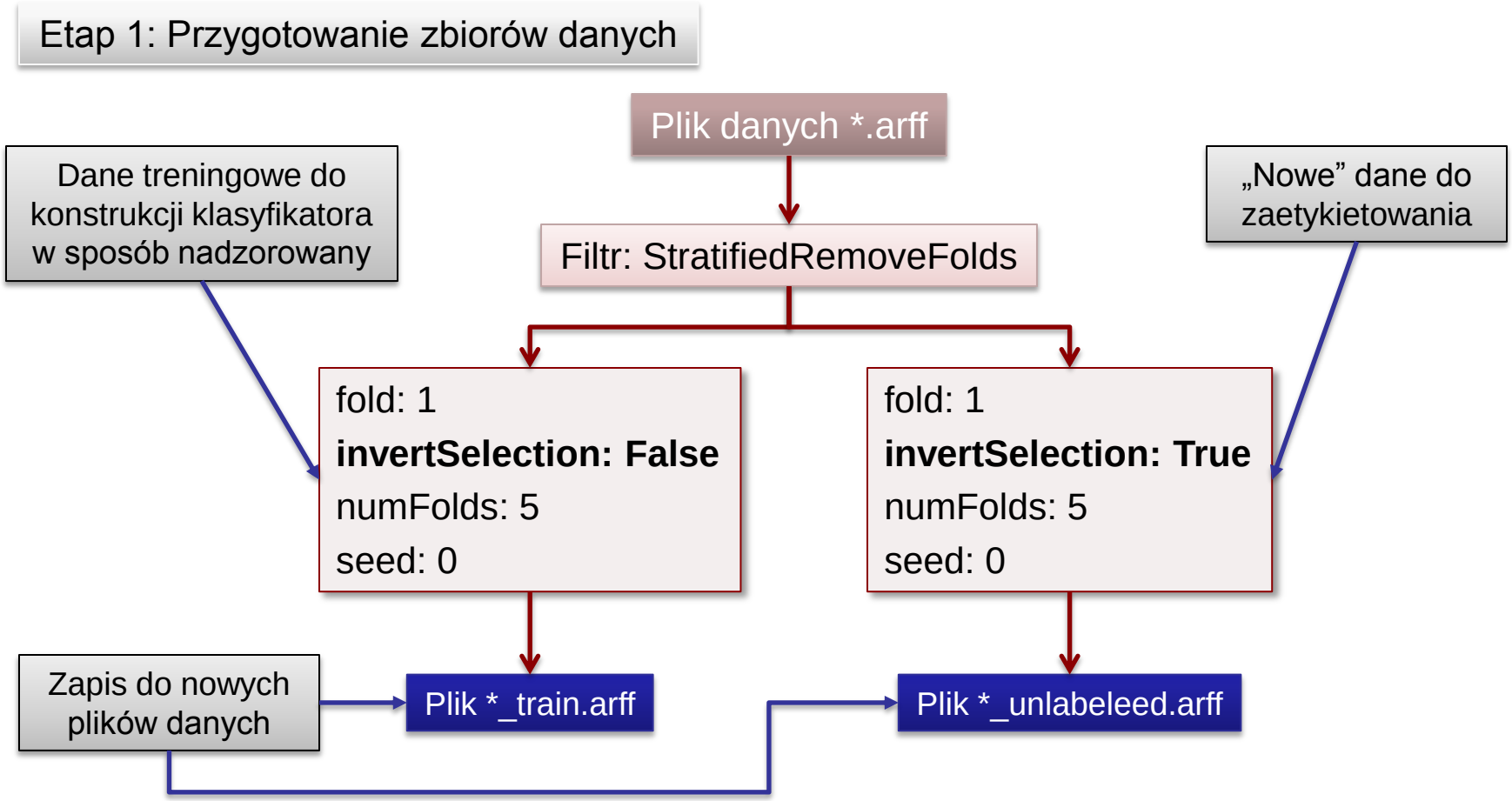
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Naiwny klasyfikator Bayesa + grupowanie EM



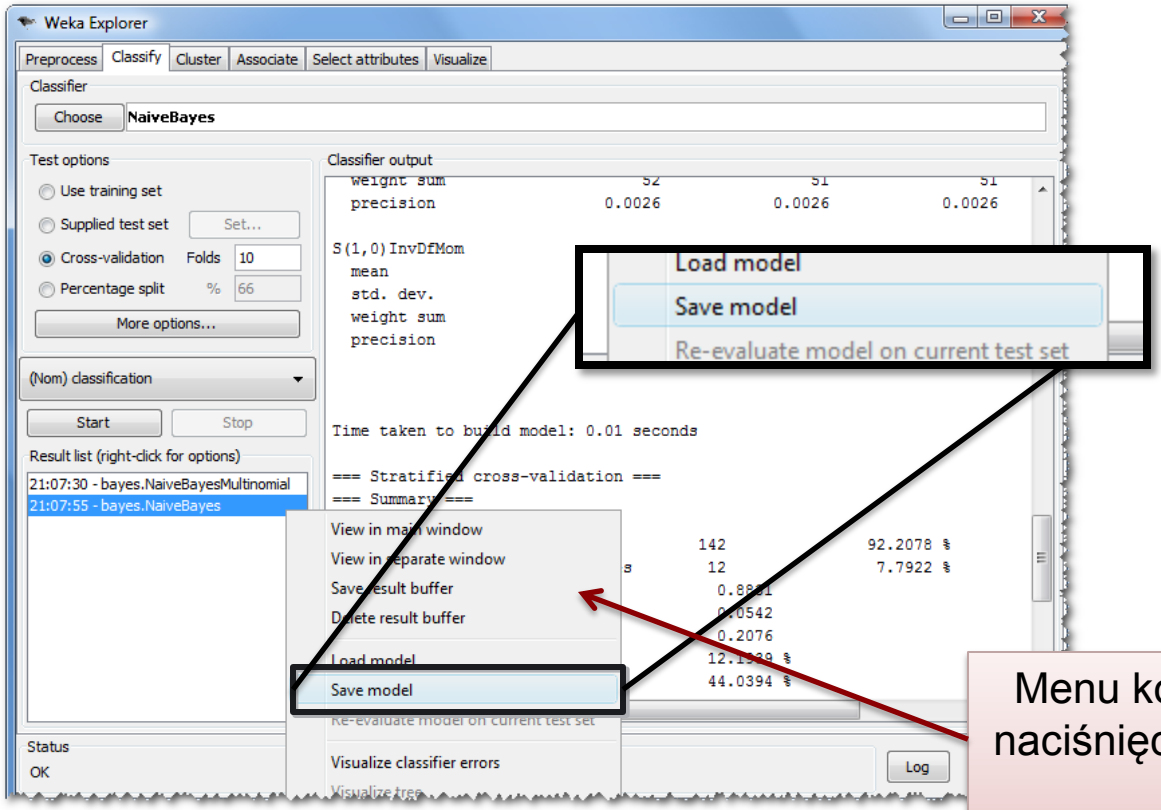
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Uczenie częściowo nadzorowane w programie *Weka* — #1



Uczenie częściowo nadzorowane w programie *Weka* — #2

Etap 2: Uczenie klasyfikatora



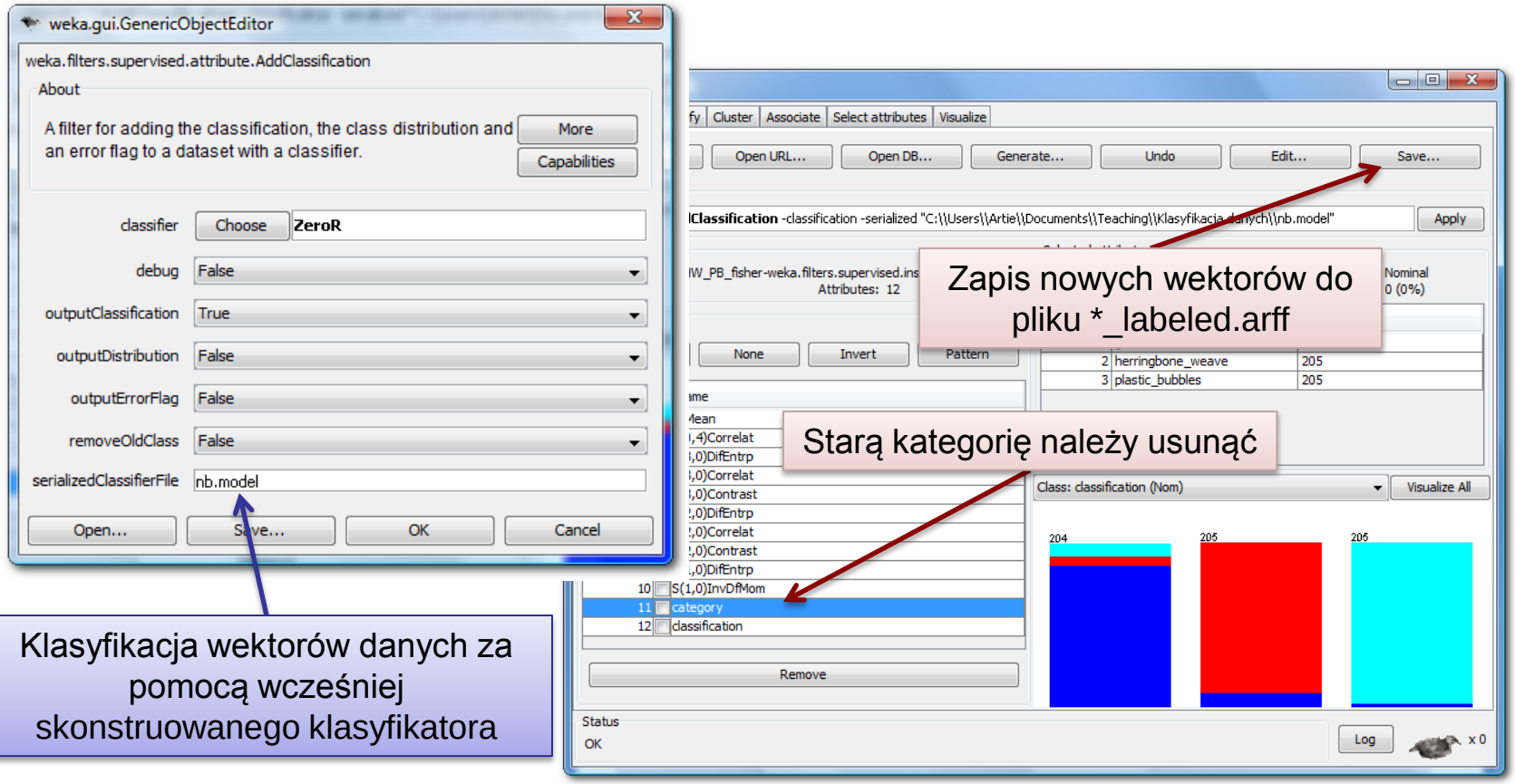
Zapis modelu wytrenowanego klasyfikatora do pliku *.model

Menu kontekstowe wywołane naciśnięciem prawego klawisza myszy

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Uczenie częściowo nadzorowane w programie *Weka* — #3

Etap 3: Etykietowanie zbioru nowych wektorów danych



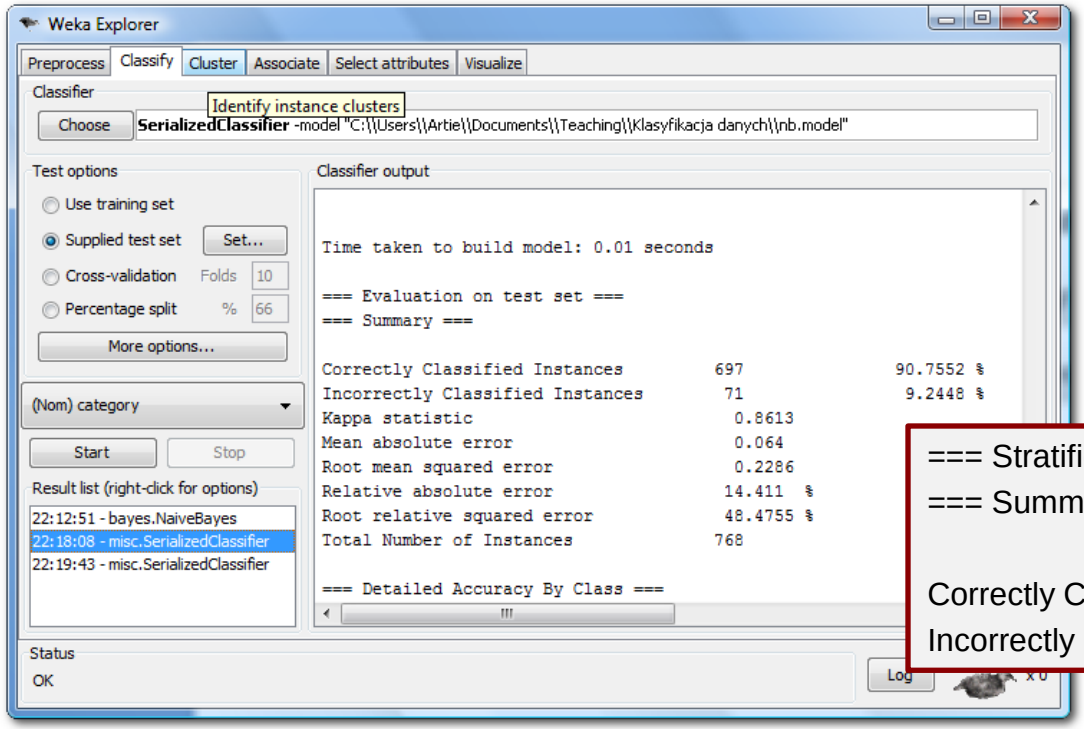
Klasyfikacja wektorów danych za pomocą wcześniej skonstruowanego klasyfikatora

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Uczenie częściowo nadzorowane w programie *Weka* — #4

Etap 4: Połączenie zbiorów danych — niestety, poza programem Weka

Etap 5: Konstrukcja nowego klasyfikatora, na podstawie połączonego zbioru danych



Wyniki klasyfikacji uzyskane przy użyciu całego zbioru danych

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

Correctly Classified Instances	704	91.6667 %
Incorrectly Classified Instances	64	8.3333 %

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Klasyfikacja tekstów

Często w przypadku danych tekstowych istnieją dwa alternatywne i niezależne opisy klasyfikowanych dokumentów.

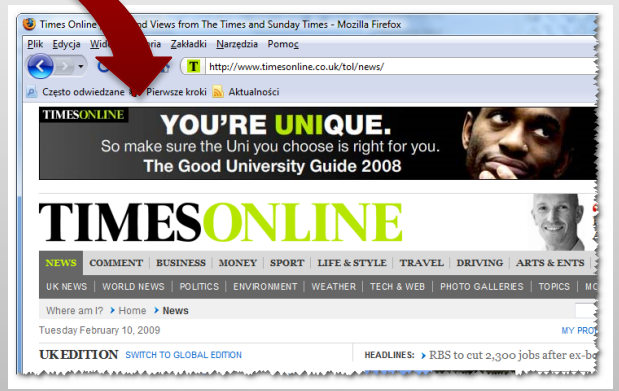
Przykład

Opis 1

Frazy występujące w dokumencie

Opis 2

Frazy występujące w innych tekstach zawierających odnośniki do właściwego dokumentu



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Wspólne uczenie klasyfikatorów — *co–training*