KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

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"*

dr inż. Artur Klepaczko

Eksploracja danych
Analiza skupień

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

Politechnika Łódzka
Instytut Elektroniki

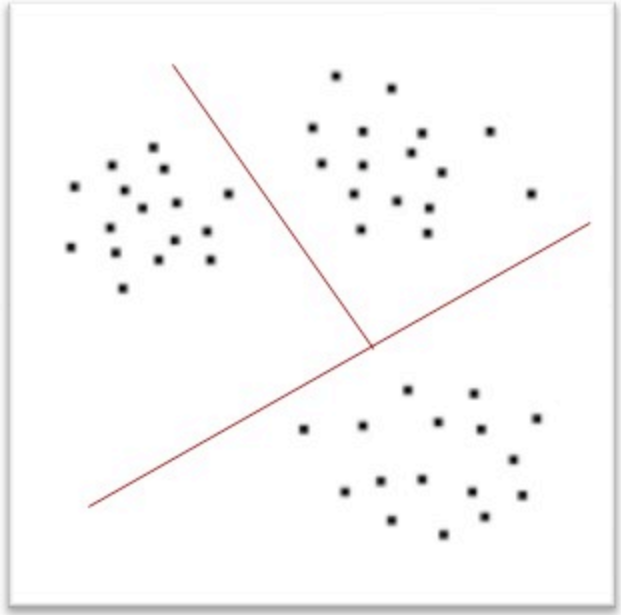
90-924 Łódź, ul. Żeromskiego 116,
tel. 042 631 28 83
www.kapitalludzki.p.lodz.pl

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Co to jest analiza skupień?

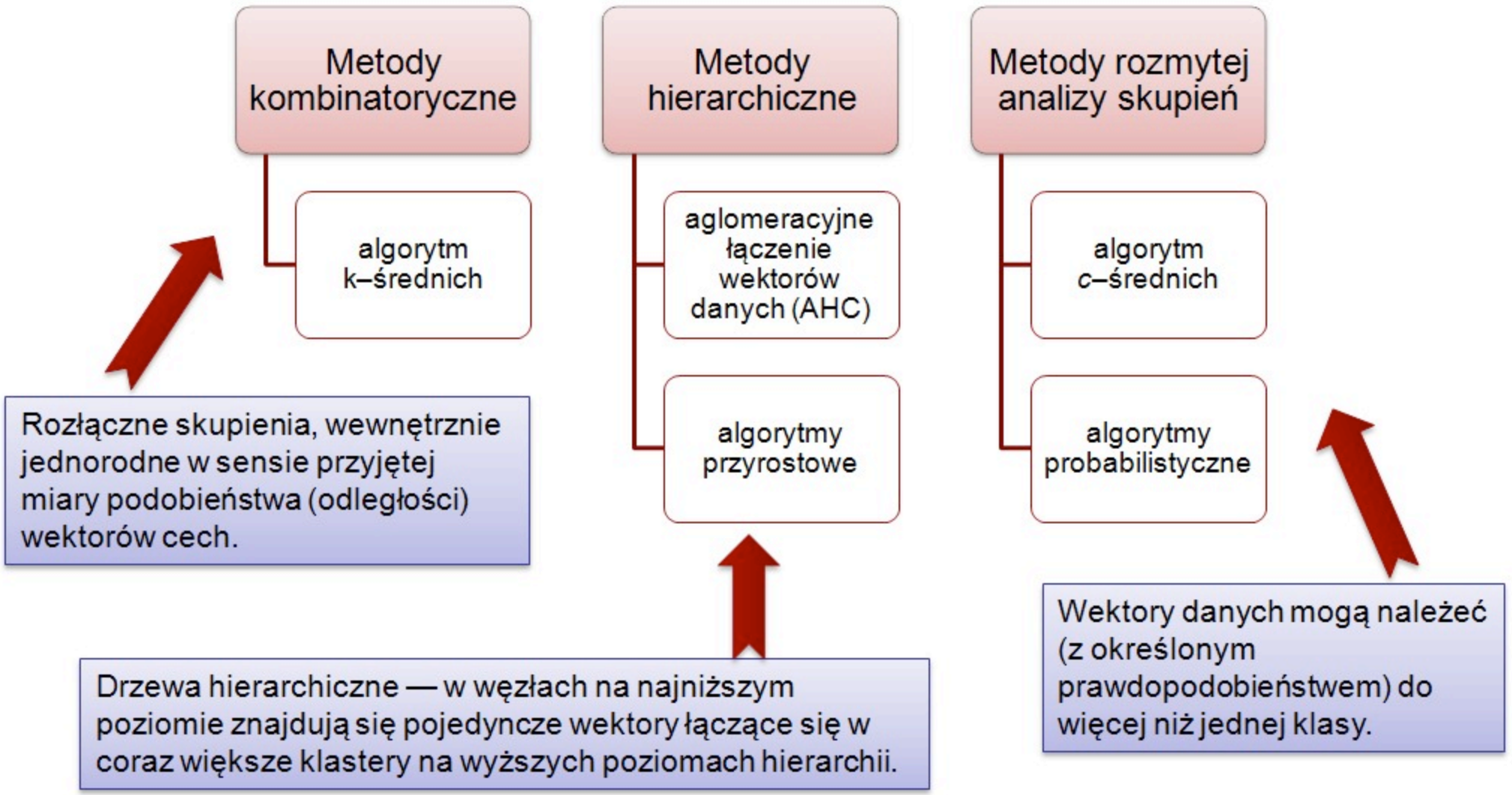
Podział zbioru danych na skupienia (grupy, klastery), wewnątrz których wektory danych wykazują większe podobieństwo wobec siebie niż wobec wektorów z innych skupień.

- Wykrycie nieznanej struktury zbioru danych
- Wyodrębnienie „naturalnych” klas wektorów cech
- Nienadzorowana klasyfikacja
- Grupowanie
- Klasteryzacja



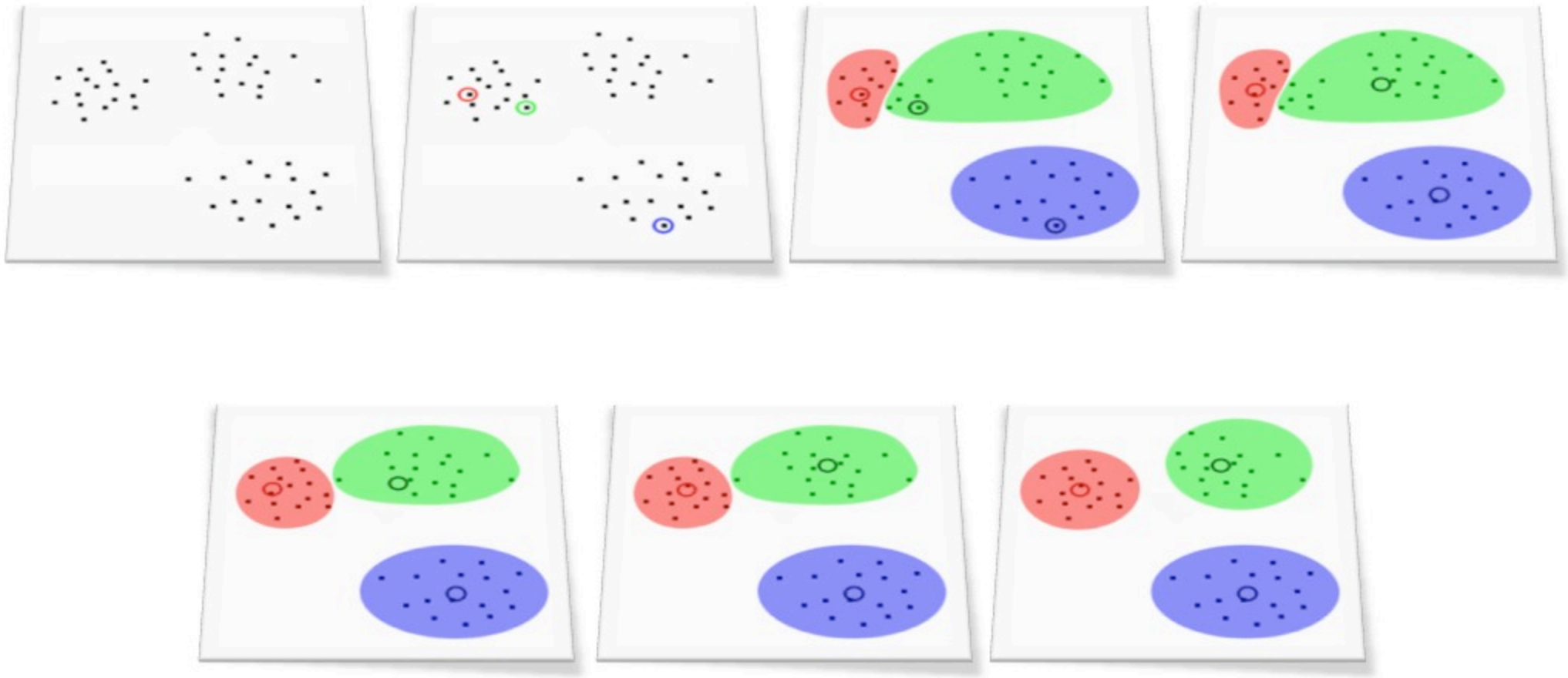
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytmy analizy skupień



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Klasyczny algorytm k–średnich



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytm *k*–średnich — formalnie

1. Dane wejściowe: **X** – macierz danych $Q \times N$, Q – liczba wektorów danych, N – liczba cech; k – liczba klastrów
2. Całkowita suma odległości:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{Q} \left\| x_i^{(j)} - m_j \right\|^2$$

1. Wybierz losowo k wektorów z **X** jako początkowe centra klastrów $m_1, \ldots, m_k$, $J = +\text{Inf}$.
2. Dla każdego wektora $\mathbf{x}_i$, $i=1..Q$, określ przynależność do właściwego klastra: $C(\mathbf{x}_i) = \arg\min_j d(\mathbf{x}_i, m_j)$.
3. Wyznacz nowe centra klastrów jako średnie punktów do nich przypisanych i ponownie oblicz J.
4. Jeśli wartość J się zmniejszyła powtórz kroki 2 i 3.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytm *k*–średnich w zastosowaniu do zbioru *Iris*

```
kMeans
======
Number of iterations: 6
Within cluster sum of squared errors: 6.9981140048267605
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute     Full Data         0         1         2
                (150)      (61)      (50)      (39)
=============================================================
sepallength     5.8433    5.8885     5.006    6.8462
sepalwidth       3.054    2.7377     3.418    3.0821
petallength     3.7587    4.3967     1.464    5.7026
petalwidth      1.1987     1.418     0.244    2.0795

Clustered Instances
0      61 ( 41%)
1      50 ( 33%)
2      39 ( 26%)
```

Grupowanie *k*-średnich

- zbiór ostatecznych centrów skupień stanowi regułę klasyfikującą dla nowych przykładów
- wynik grupowania metodą *k*-średnich zależy od wyboru początkowych centrów klasterów
- w przypadku klasteryzacji nie ma narzędzi do jednoznacznej oceny prawidłowości wyniku

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Grupowanie a prawidłowa klasyfikacja

W przypadku uczenia nienadzorowanego, wektory danych treningowych nie zawierają etykiet kategorii → trudność w jednoznacznej ocenie wyniku klasyfikacji.

Zbiór danych niezaetykietowanych

Wektory danych posiadają prawidłowe etykiety kategorii

Algorytm grupowania

Miary jakości grupowania:
- heurystyczne
- teorioinformacyjne
- metody probabilistyczne

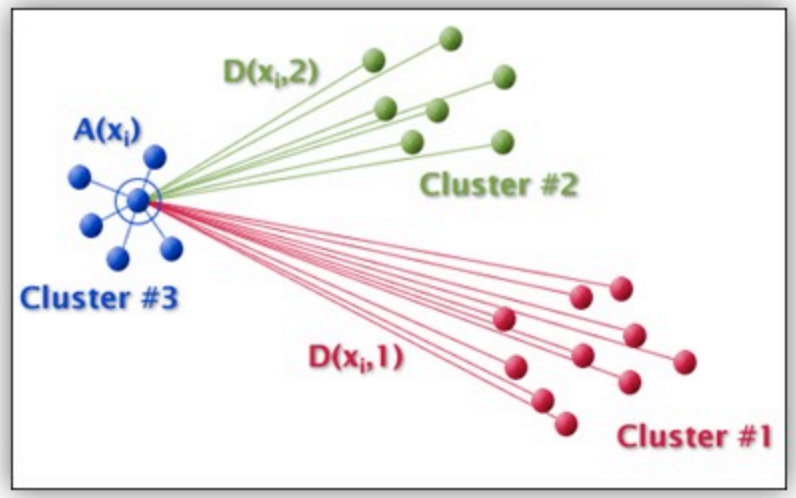
Tabela klas–klasterów

	Klaster 1	Klaster 2	Klaster 3	Klaster 4
Klasa 1	1	20	-	4
Klasa 2	25	-	-	-
Klasa 3	10	-	14	1

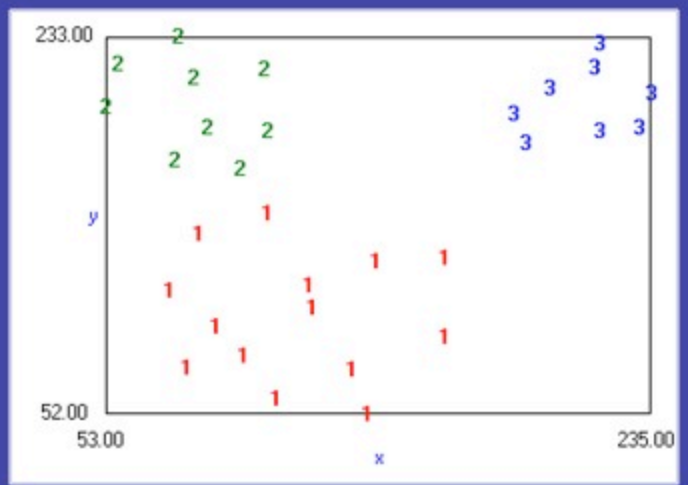
Szacowanie wiarygodności klasteryzacji

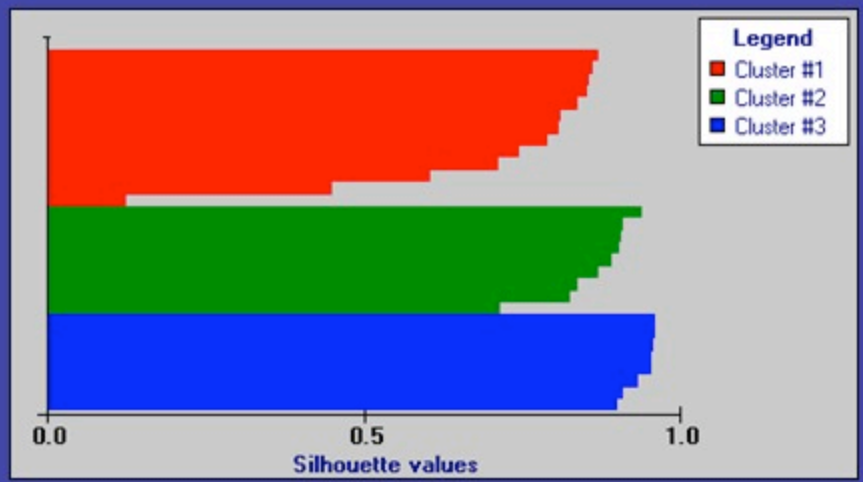
Miara podobieństwa (*silhouette value*)

$$S_i = \frac{\min_{j \neq C(x_i)} D(x_i, j) - A(x_i)}{\max\left(A(x_i), \min_{j \neq C(x_i)} D(x_i, j)\right)}$$



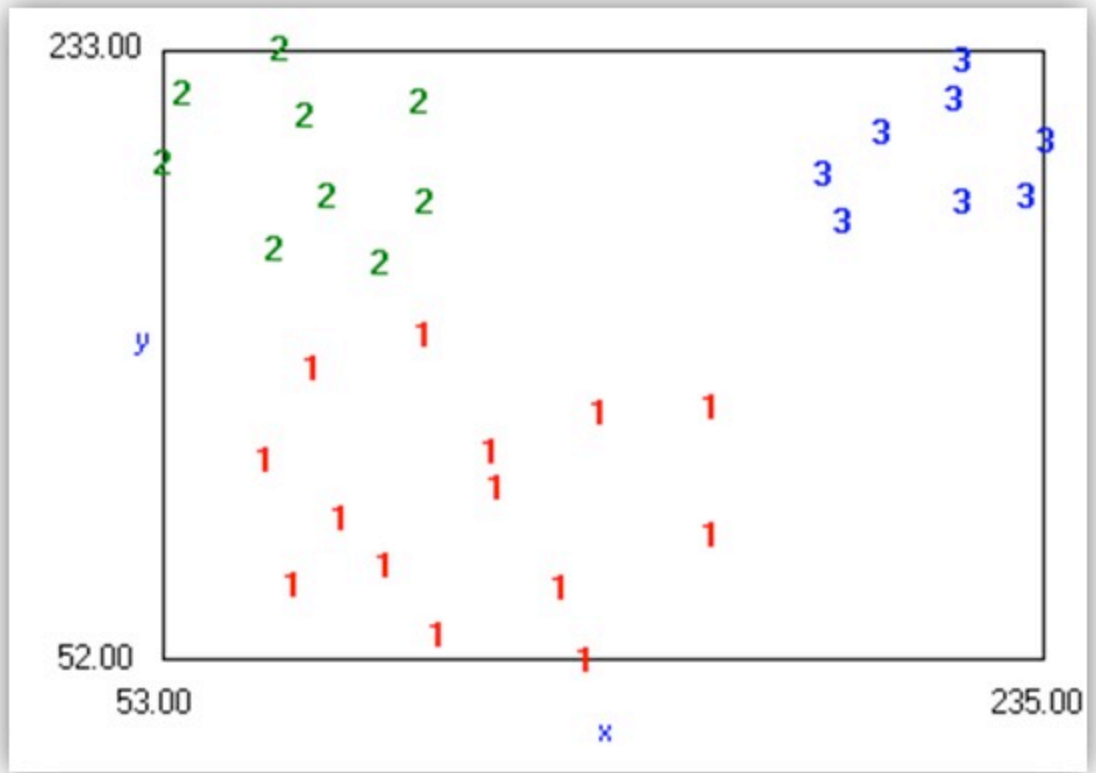
Przykład





Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

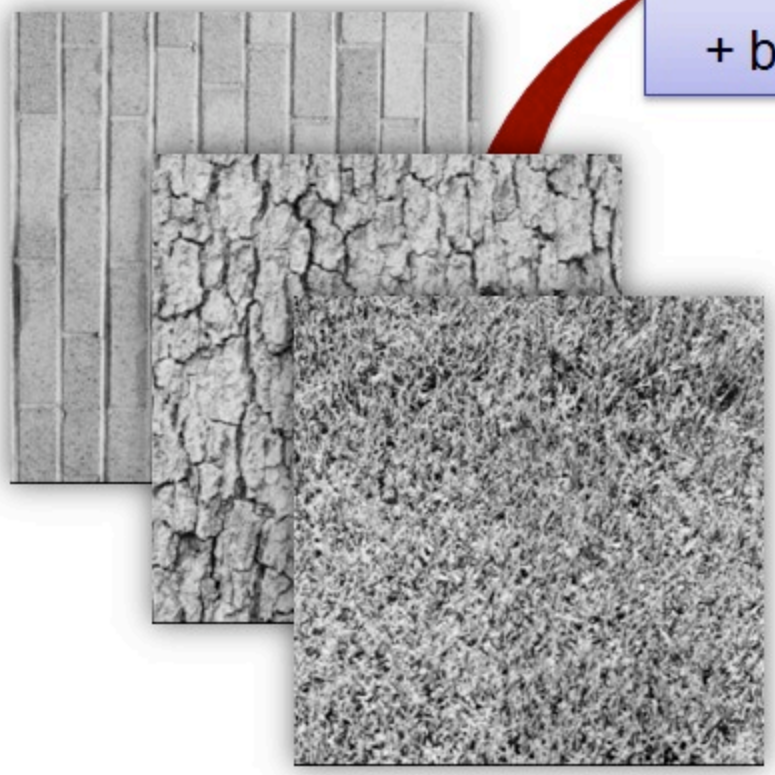
Ile klasterów?



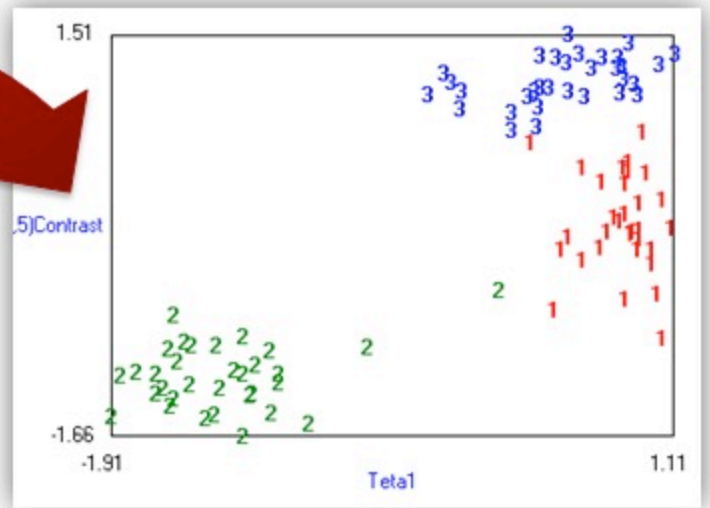
Liczba klastrów	Average silhouette value
2	0.69
3	0.79
4	0.70
5	0.68

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Badanie liczby klasterów dla zbioru tekstur

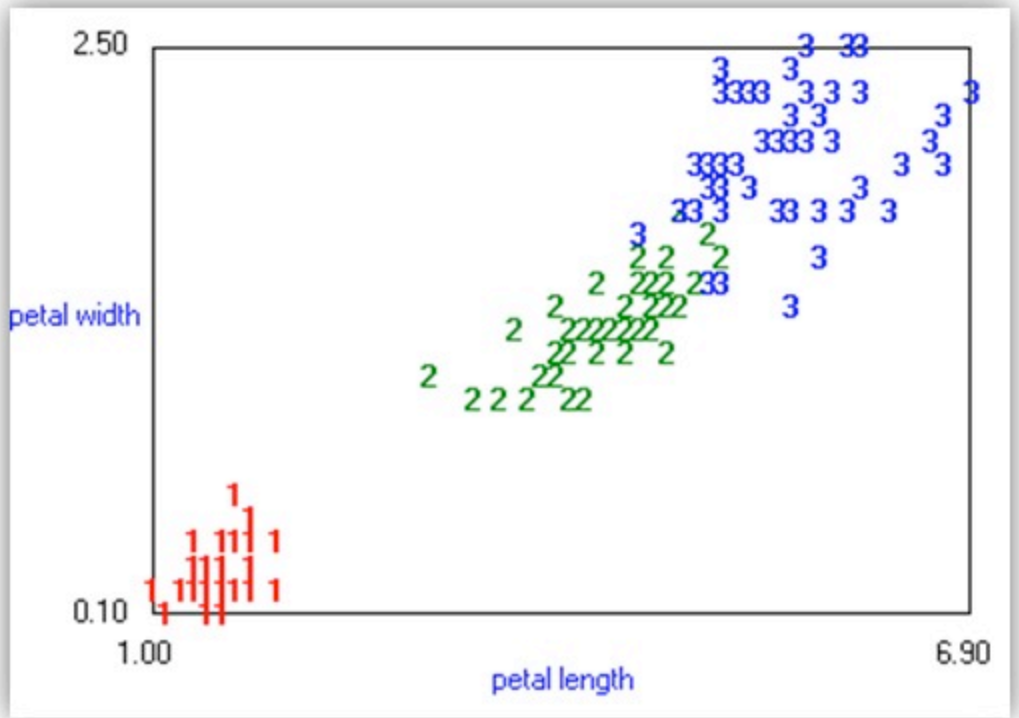


3 tekstury naturalne z albumu Brodatza



Liczba klastrów	Average silhouette value
2	0.71
3	0.86
4	0.76
5	0.69

Badanie liczby klastrów dla zbioru *Iris*



Liczba klastrów	Average silhouette value
2	0.87
3	0.83
4	0.76
5	0.75

Potrzebna jest inna miara jakości albo inny algorytm grupowania…

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Długość zakodowanego komunikatu

Twierdzenie Bayesa

$$\Pr(h \mid D) = \frac{\Pr(h)\Pr(D \mid h)}{\Pr(D)}$$

$$-\log \Pr(h \mid D) = -\log \Pr(h) - \log \Pr(D \mid h) + \log \Pr(D)$$

$\log \Pr(\cdot)$

Prawdopodobieństwo logarytmiczne (ang. *log–likelihood*)

$-\log \Pr(\cdot)$

Długość kodu dla komunikatu o prawdopodobieństwie $\Pr(\cdot)$

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Długość kodu — interpretacja składników

$$-\log \Pr(h \mid D) = -\log \Pr(h) + -\log \Pr(D \mid h) - -\log \Pr(D)$$

Liczba bitów potrzebna do przesłania zakodowanej informacji:

- $-\log \Pr(h \mid D)$: o hipotezie *h* po zaobserwowaniu zbioru danych *D*
- $-\log \Pr(h)$: o hipotezie (reguła decyzyjna, struktura skupień)
- $-\log \Pr(D \mid h)$: o zbiorze danych *D*, przy założeniu, że obowiązuje hipoteza *h* (na ile dane pasują do hipotezy; informacja o wyjątkach od reguły)
- $-\log \Pr(D)$: o zbiorze danych *D*, składnik niezależny od przyjętej hipotezy

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Brzytwa Ockhama

William of Ockham (1300–1350) — filozof, teolog, franciszkanin; głosił konsekwentny nominalizm, według którego pojęcia ogólne, będące skutkiem abstrakcji, nie mają odbicia w rzeczywistości; sformułował postulat metodologiczny, zwany brzytwą Ockhama.

Kaczorowski M. (red.) *Nowa Encyklopedia Powszechna PWN*, Warszawa 2004

Proste pojęcia są lepsze i bardziej prawdopodobne niż teorie rozbudowane

Zastosowanie w analizie skupień:
– preferowane rozwiązania o dużym prawdopodobieństwie
– im mniej parametrów (np. liczba klasterów), tym lepiej

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Zasada minimalnej długości kodu (MDL)

Optymalne kodowanie: minimalna długość zakodowanego komunikatu.

Przyjmując oznaczenie:

$$-\log \Pr(\cdot) = \mathrm{L}(\cdot)$$

Porównywanie hipotez na podstawie wyrażenia:

$$\mathrm{L}(h \mid D) = \mathrm{L}(h) + \mathrm{L}(D \mid h)$$

Pomijając składnik

$$-\log \Pr(D)$$

Duże, jeśli hipoteza/reguła jest zbyt skomplikowana.

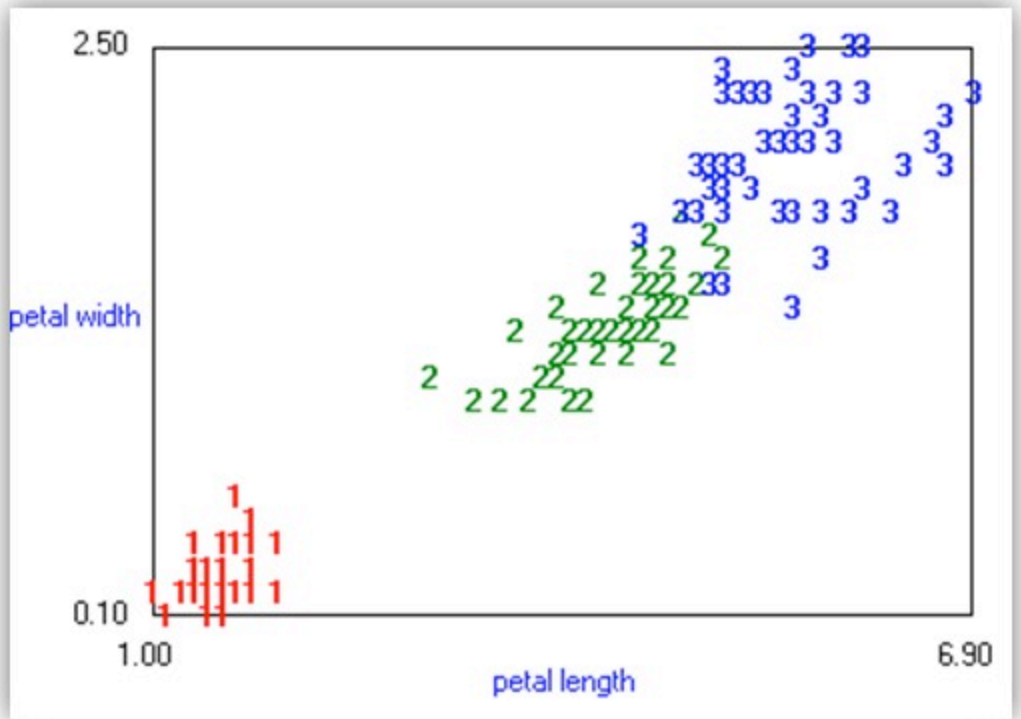
Małe, gdy hipoteza *idealnie* pasuje do zbioru danych.

Nadmierne dopasowanie!

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Jakość grupowania według MDL — przykład zbioru *Iris*

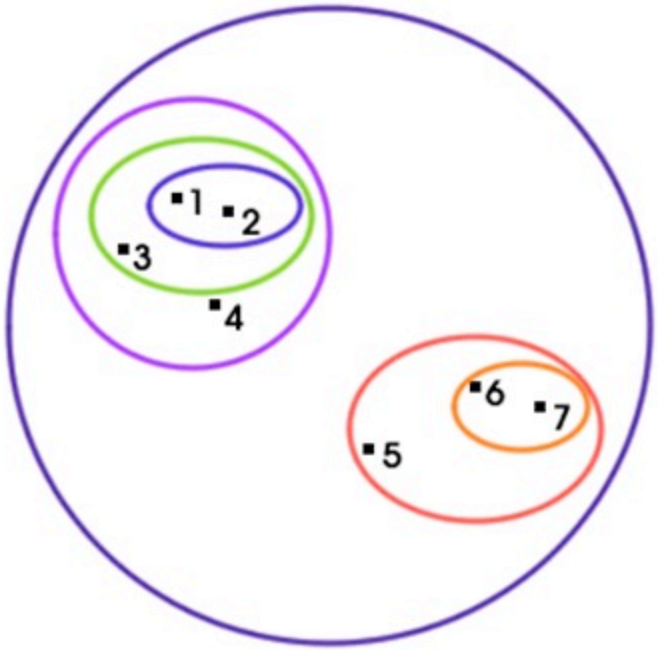
Najlepsze grupowanie powinno zapewnić najbardziej efektywne kodowanie.

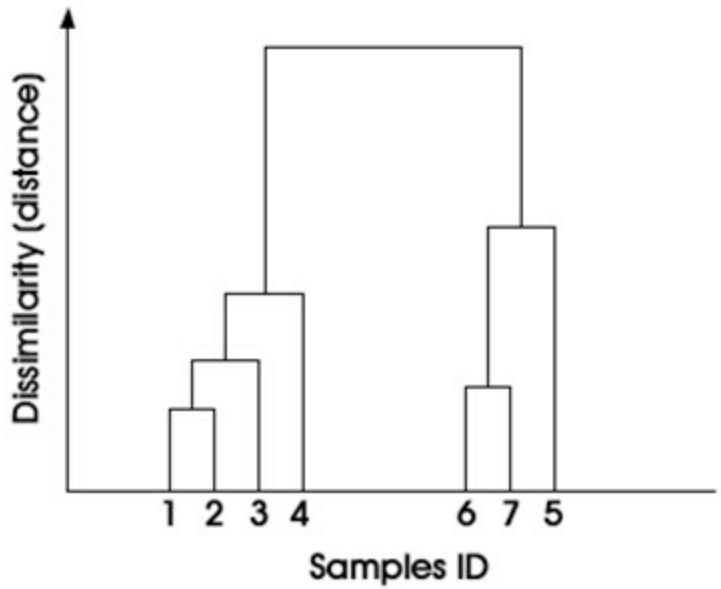


Liczba klastrów	Średnia długość kodu
2	4.06
3	3.79
4	4.04
5	4.10

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Grupowanie hierarchiczne





- Zagnieżdżona struktura klasterów (drzewo hierarchiczne–dendrogram)
- Strategia scalania (ang. *agglomerative hierarchical clustering*)
- Strategia rozdzielania (ang. *divisive hierarchical clustering*)

Algorytm AHC

1. Umieść wszystkie wektory danych w osobnych klasterach.
2. Oblicz odległości między wszystkimi parami klasterów.
3. Znajdź najbliższą parę klasterów i połącz je w jeden nowy klaster.
4. Jeśli pozostały jakieś klastery do połączenia powróć do kroku 2.

Jak liczyć odległość między klasterami?

Strategia łączenia:

średnie odległości → *average-link*

najbliższa para → *single-link*

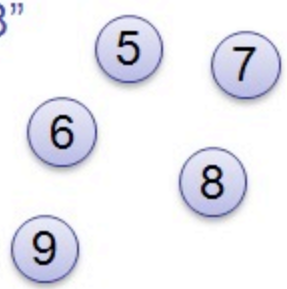
najdalsza para → *complete-link*

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Odległość między klasterami

Klaster „A”

Klaster „B”



Single-link

$$Dist(A,B) = \mathrm{d}(\mathbf{x}_4, \mathbf{x}_5)$$

Complete-link

$$Dist(A,B) = \mathrm{d}(\mathbf{x}_3, \mathbf{x}_9)$$

Average-link

$$S_1 = \mathrm{d}(\mathbf{x}_1, \mathbf{x}_5) + \mathrm{d}(\mathbf{x}_1, \mathbf{x}_6) + \mathrm{d}(\mathbf{x}_1, \mathbf{x}_7) + \mathrm{d}(\mathbf{x}_1, \mathbf{x}_8) + \mathrm{d}(\mathbf{x}_1, \mathbf{x}_9)$$

$$Dist(A,B) = \frac{1}{4 \cdot 5}\left(S_1 + S_2 + S_3 + S_4\right)$$

$\mathrm{d}(\cdot,\cdot)$ — przyjęta miara odległości, np. euklidesowa

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Euklidesowa miara odległości

Wektor $\mathbf{x}_1$

$$\mathbf{x}_1 = \langle x_1^1, x_1^2, x_1^3 \rangle$$

Wektor $\mathbf{x}_2$

$$\mathbf{x}_2 = \langle x_2^1, x_2^2, x_2^3 \rangle$$

Odległość euklidesowa

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(x_1^1 - x_2^1)^2 + (x_1^2 - x_2^2)^2 + (x_1^3 - x_2^3)^2}$$

Różna skala atrybutów

$\mathbf{x}_1 = \langle 2,4,1,3,1 \rangle$ $\mathbf{x}_2 = \langle 3,4,1,3,201 \rangle$ $\mathbf{x}_3 = \langle 1,3,2,7,101 \rangle$

$d(\mathbf{x}_1, \mathbf{x}_2) \approx 200$ $d(\mathbf{x}_1, \mathbf{x}_3) \approx 101$

$\mathbf{x}_1$ i $\mathbf{x}_2$ różnią się dwoma atrybutami

$\mathbf{x}_1$ i $\mathbf{x}_3$ różnią się pięcioma atrybutami

Atrybuty o małej skali tracą znaczenie przy obliczaniu odległości

Konieczność normalizacji lub standaryzacji danych

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Normalizacja/standaryzacja danych

Normalizacja

Wartość minimalna $x_{\min}$

Wartość maksymalna $x_{\max}$

$$\tilde{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Wartości wszystkich atrybutów w tej samej skali → między 0 a 1.

Standaryzacja

Wartość średnia

$$\bar{x} = \frac{1}{Q}\sum_{i=1}^{Q} x_i$$

Odchylenie standardowe

$$\sigma = \sqrt{\frac{1}{Q-1}\sum_{i=1}^{Q}(x_i - \bar{x})^2}$$

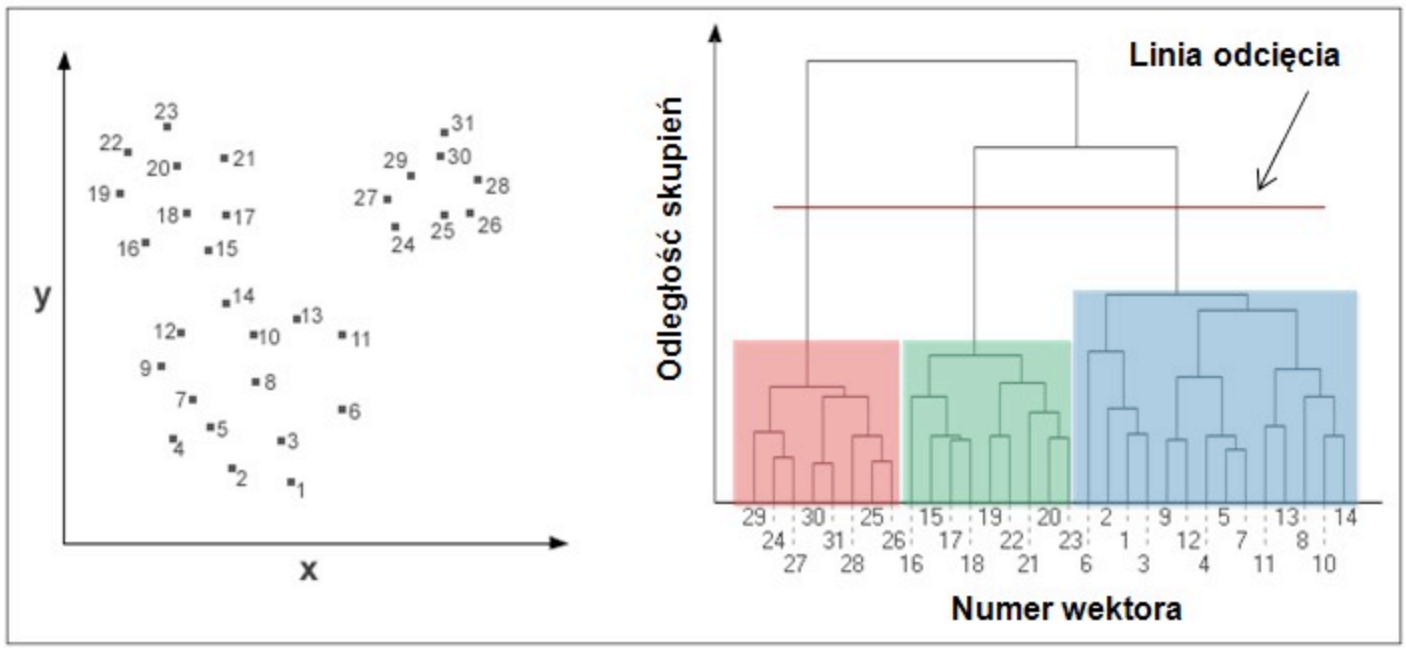
$$\hat{x}_i = \frac{x_i - \bar{x}}{\sigma}$$

Wartości średnie wszystkich atrybutów → 0

Wartości odchylenia standardowego dla wszystkich atrybutów → 1

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Podział struktury hierarchicznej na skupienia



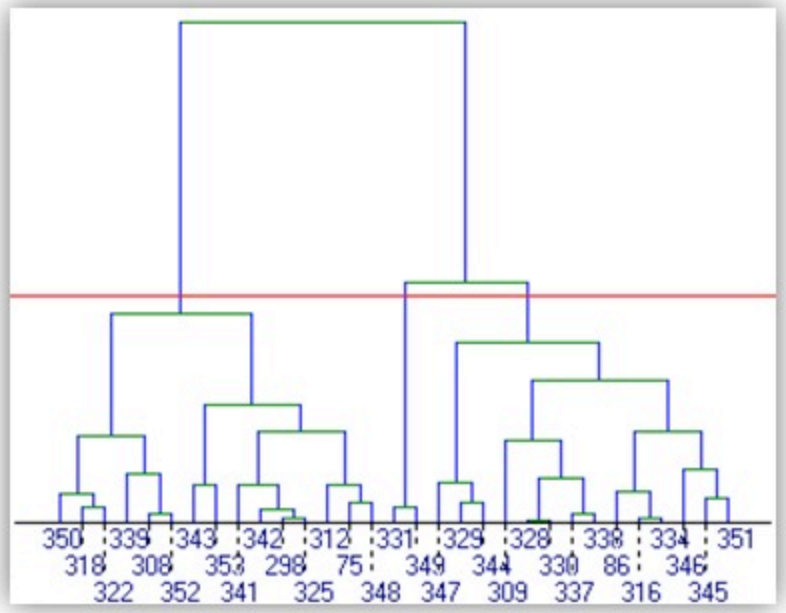
Współczynnik niespójności połączenia

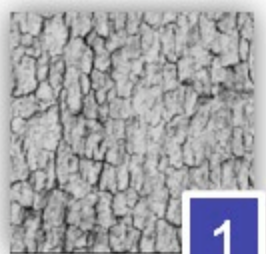
$$\psi_m = \frac{h_m - \mu_m}{\sigma_m}$$

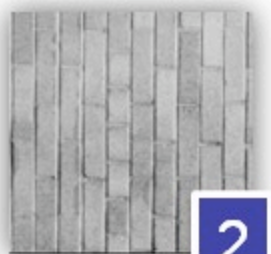
h_m – wysokość połączenia m

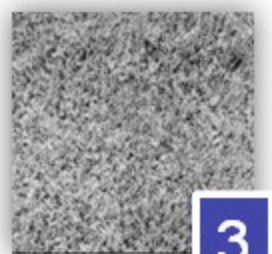
μ_m, σ_m– wartość średnia i odchylenie standardowe połączenia m i połączeń leżących poniżej

Algorytm AHC w zastosowaniu do zbioru tekstur









Wynik grupowania wobec prawidłowej klasyfikacji

	Klaster 1	Klaster 2	Klaster 3
Klasa 1		33	31
Klasa 2		14	50
Klasa 3	18	46	

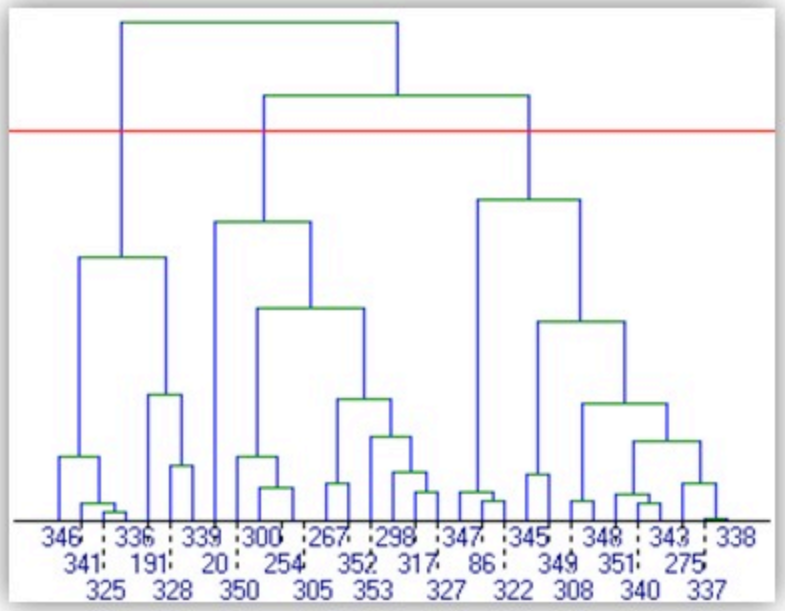
Klaster 1 zawiera 18 wektorów z klasy 3.

Klaster 2 zawiera 33 wektory z klasy 1, 14 wektorów z klasy 2 i 46 wektorów z klasy 3.

Klaster 3 zawiera 31 wektorów z klasy 1 i 50 wektorów z klasy 2.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Tekstury raz jeszcze — dane ustandaryzowane

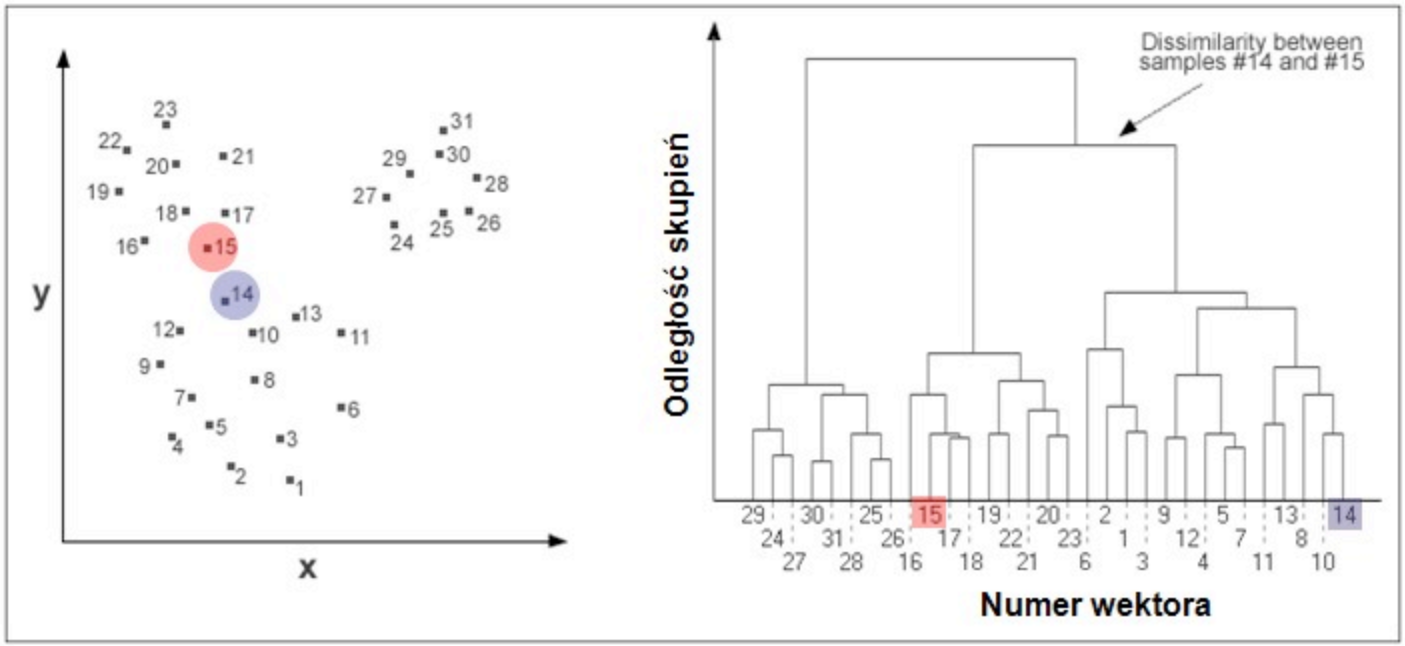


Wynik grupowania wobec prawidłowej klasyfikacji

	Klaster 1	Klaster 2	Klaster 3
Klasa 1	1	63	
Klasa 2	51	13	
Klasa 3			64

Trzy wyróżniające się poddrzewa – sugestia podziału hierarchii na skupienia.
Poszczególne klastery w większości zawierają wektory tylko z jednej klasy.

Ocena jakości grupowania hierarchicznego



Współczynnik korelacji kofenetycznej

$$\varphi = \frac{\sum_{i<j}\left(\mathbf{P}_{ij} - \mu_{\mathbf{P}}\right)\cdot\left(\mathbf{D}_{ij} - \mu_{\mathbf{D}}\right)}{\sqrt{\sum_{i<j}\left(\mathbf{P}_{ij} - \mu_{\mathbf{P}}\right)^2 \cdot \sum_{i<j}\left(\mathbf{D}_{ij} - \mu_{\mathbf{D}}\right)^2}}$$

Jain A., Dubes R., *Algorithms for Clustering Data*, Prentice Hall, 1988.

P – macierz sąsiedztwa

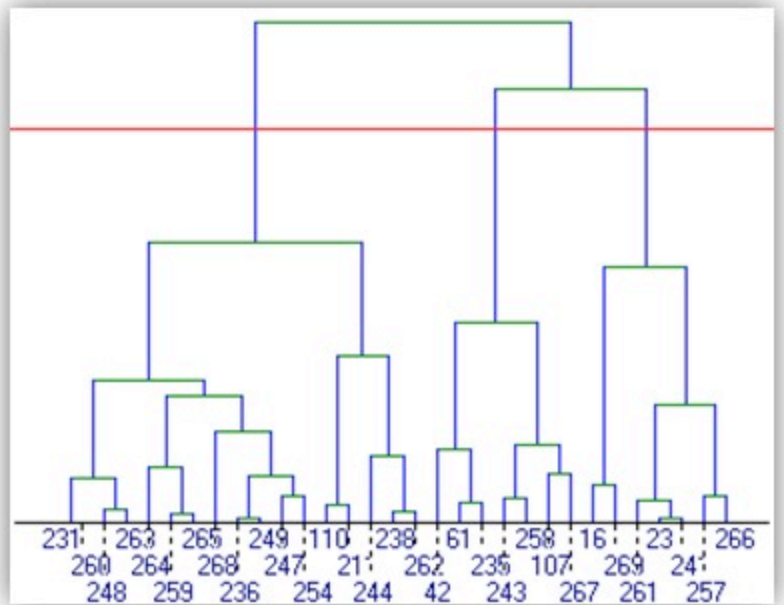
D – macierz odległości w drzewie

μ_P – wartość średnia w **P**

μ_D – wartość średnia w **D**

i, j – indeksy próbek

Algorytm AHC w zastosowaniu do zbioru *Iris*



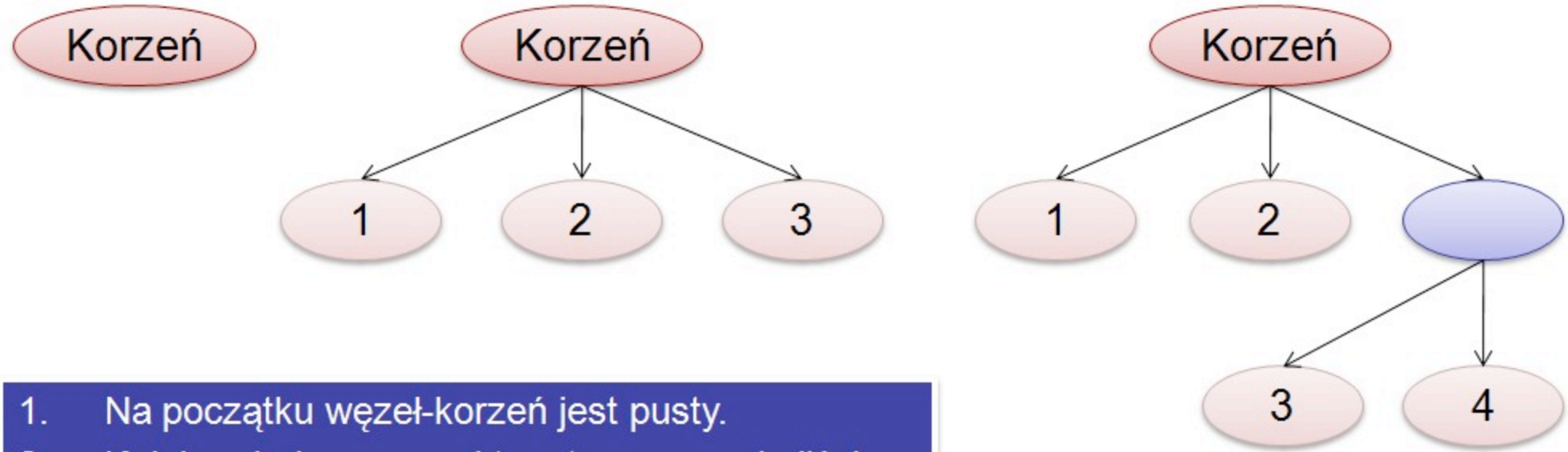
Wynik grupowania wobec prawidłowej klasyfikacji

	Klaster 1	Klaster 2	Klaster 3
Klasa 1	1	49	
Klasa 2	21		29
Klasa 3	2		48

Współczynnik korelacji: 0.75 (dane ustandaryzowane, *complete-link*).
Trzy odrębne klastery, ale nie w pełni pokrywające się z podziałem na klasy.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Grupowanie przyrostowe — algorytm Cobweb



1. Na początku węzeł-korzeń jest pusty.
2. Kolejno dodawane wektory tworzą węzły-liście.
3. W niektórych przypadkach nowy wektor dołączany jest do już dodanych – razem tworzą one węzły-kategorie.
4. Procedura kończy się, gdy wszystkie wektory danych zostały dołączone do drzewa.

Kiedy nowy wektor ma utworzyć oddzielny węzeł-liść, a kiedy ma zostać dołączony do istniejącej kategorii?

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Kryterium jakości kategorii

Funkcja kryterialna

$$\kappa(C_1, C_2, \ldots, C_k) = \frac{\sum_{l=1}^{k} \Pr[C_l] \sum_i \sum_j \left(\Pr[a_i = v_{ij} | C_l]^2 - \Pr[a_i = v_{ij}]^2 \right)}{k}$$

$\Pr[C_l]$ — Prawdopodobieństwo klastera C_l.

$\Pr[\alpha_i = v_{ij} | C_l]$ — Prawdopodobieństwo, że atrybut α_i wektora należącego do klastera C_l przyjmie wartość v_{ij}.

$\Pr[\alpha_i = v_{ij}]$ — Bezwarunkowe prawdopodobieństwo tego, że atrybut α_i przyjmie wartość v_{ij}.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Kryterium jakości kategorii – przykład

Atrybuty

$\alpha_1 \in$ (a, b, c)
$\alpha_2 \in$ (d, e, f, g)
$\alpha_3 \in$ (x, y)

Wektory danych

1: a, d, x, C_1
2: a, e, x, C_1
3: a, e, y, C_2
4: a, f, x, C_2
5: a, g, y, C_1
6: b, e, y, C_2
7: b, f, y, C_2
8: c, d, x, C_1
9: c, d, y, C_1
10: c, g, y: C_1

Prawdopodobieństwa

$|C_1|=6, |C_2|=4$
$\Pr[C_1]=0.6, \Pr[C_2]=0.4$

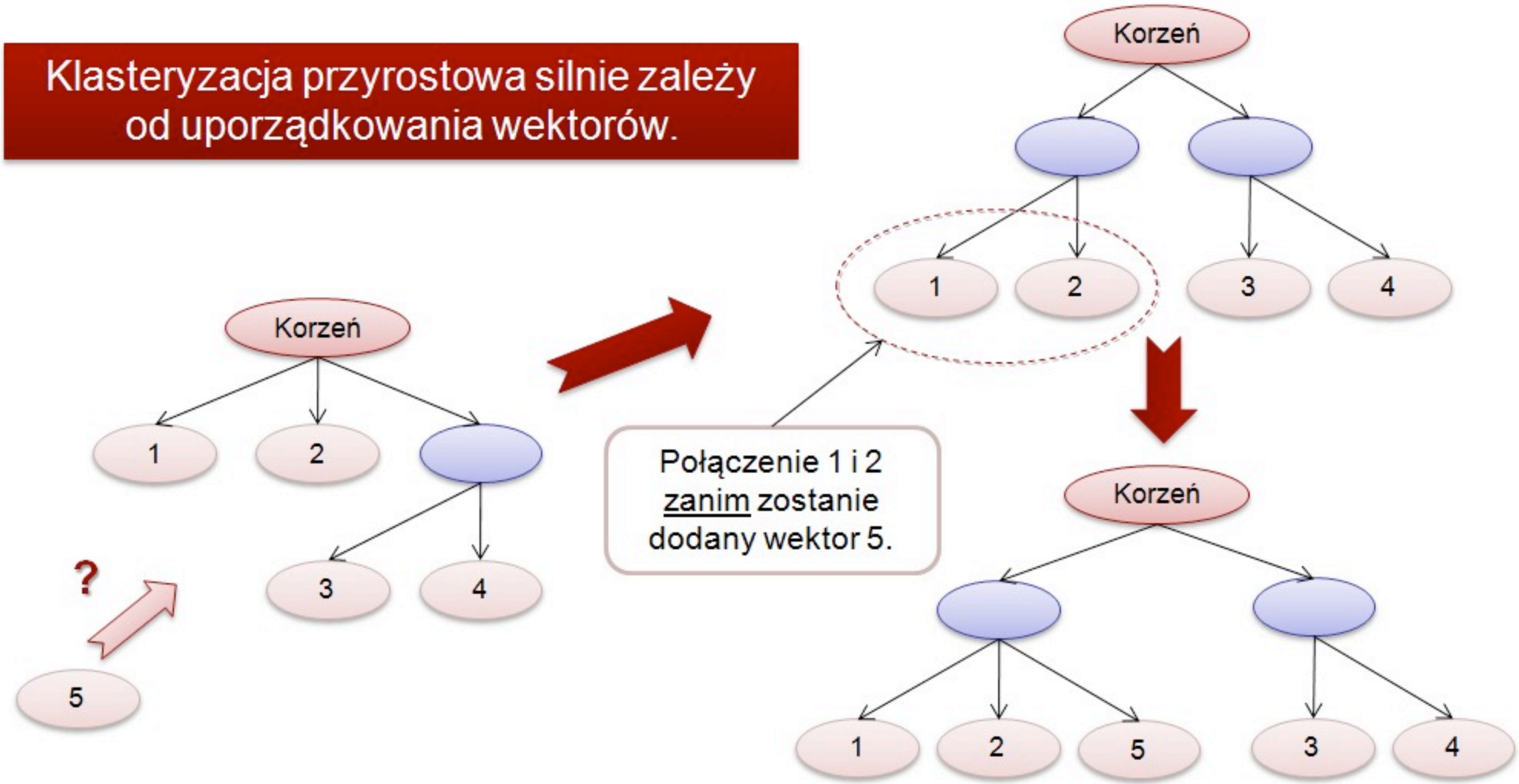
$\Pr[\alpha_1=\text{a}] = 0.5$
$\Pr[\alpha_1=\text{a}|C_1] = 0.6$
$\Pr[\alpha_1=\text{b}] = 0.2$
$\Pr[\alpha_1=\text{b}|C_2] = 1.0$
$\Pr[\alpha_1=\text{c}] = 0.3$
$\Pr[\alpha_1=\text{c}|C_2] = 0.0$

Utworzenie wspólnej kategorii dla wektorów danych ma miejsce wówczas, gdy wzrastają prawdopodobieństwa oszacowań wartości atrybutów.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Restrukturyzacja drzewa poprzez łączenie

Klasteryzacja przyrostowa silnie zależy od uporządkowania wektorów.



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Grupowanie przyrostowe — uwagi końcowe

Uogólnienie dla atrybutów ciągłych → algorytm Classit

$$\kappa(C_1, C_2, \ldots, C_k) = \frac{1}{k}\sum_{l=1}^{k} \Pr[C_l] \frac{1}{2\sqrt{\pi}} \sum_i \left(\frac{1}{\sigma_{il}} - \frac{1}{\sigma_i} \right)$$

- Założenie normalnego rozkładu prawdopodobieństwa dla wartości atrybutów
- σ_{il} – odchylenie standardowe atrybutu α_i wewnątrz skupienia C_l; w przypadku pojedynczych wektorów $\sigma_{il} = 0$ → konieczność określenia minimum (parametr *acuity*)
- σ_i – odchylenie standardowe atrybutu α_i w całym zbiorze danych

Powstrzymywanie rozrostu drzewa → parametr odcięcia (ang. *cutoff*)

Jeżeli utworzenie odrębnego węzła-liścia powoduje tylko nieznaczną poprawę kryterium jakości kategorii, to dany wektor dołączany jest do już istniejącej kategorii.

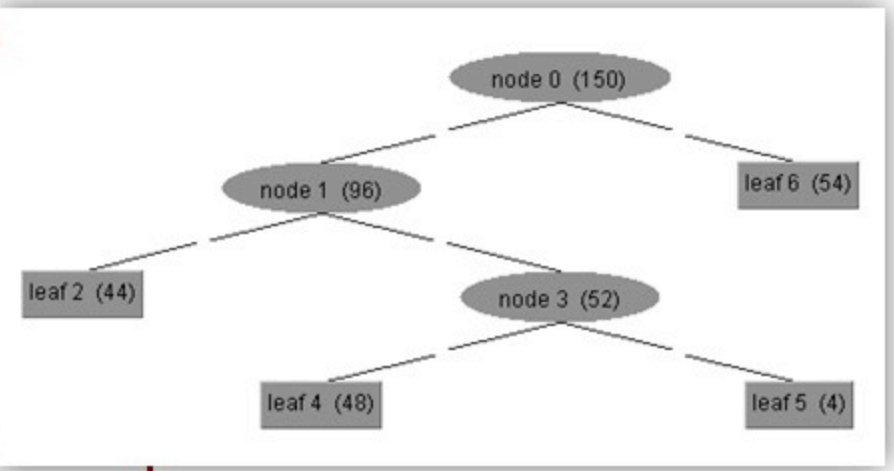
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytm Classit w zastosowaniu do zbioru *Iris*

```
=== Model and evaluation on training set ===
node 0 [150]
| node 1 [96]
|     |     leaf 2 [44]
|     node 1 [96]
|     |     node 3 [52]
|     |     |   leaf 4 [48]
|     |     node 3 [52]
|     |     |   leaf 5 [4]
node 0 [150]
|     leaf 6 [54]

Class attribute: class
Classes to Clusters:

  2  4  5  6    <-- assigned to cluster
  0  0  0 50   | Iris-setosa
 45  2  0  3   | Iris-versicolor
 10 39  1  0   | Iris-virginica
```



Algorytm Classit

- nie ma konieczności założenia z góry określonej liczby klasterów
- wyuczoną regułę klasyfikującą stanowi całe skonstruowane drzewo hierarchiczne
- uzyskanie względnie dobrego wyniku wymaga właściwego doboru parametrów algorytmu

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Grupowanie zbioru tekstur

Błąd klasyfikacji nienadzorowanej zbioru tekstur [%]

Algorytm	Bez standaryzacji	Dane ustandaryzowane
K-means	8.9	8.9
AHC (single-link)	32.8	33.3
AHC (complete-link)	32.8	7.3
Cobweb/Classit	21.9 (acuity=21)	24.5 (acuity=0.85)
Cobweb/Classit	40.1 (acuity=22)	31.2 (acuity=0.9)

1. Wynik grupowania wymaga wielostronnej weryfikacji.
2. Często „naturalny” podział zbioru danych nie jest zbieżny z podziałem zaproponowanym przez eksperta.
3. Uzyskanie prawidłowej klasyfikacji wymaga precyzyjnego doboru parametrów.
4. Klasteryzacja pozwala na obiektywizację procedury eksploracji danych.