Analiza Skupień - Grupowanie

Zaawansowana Eksploracja Danych

JERZY STEFANOWSKI
Inst. Informatyki PP
Wersja dla TPD 2009

Poprawiona 2012

Organizacja wykładu

- Wprowadzenie i możliwe zastosowania
- Podstawy (odległości, …) .
- Dobór parametrów algorytmów:
  - Hierarchiczne (AHC)
  - k - średnich
- Studium przypadku użycia
- Rozszerzenia dla analizy danych o większych rozmiarach.
- Podsumowanie

Elementy terminologiczne

Trochę uwag:

- Cluster Analysis → Analiza skupień, grupowanie.
- Numerical taxonomy → Metody taksonomiczne (ekonomia)
  - Uwaga: znaczenie taksonomii w biologii może mieć inny kontest (podział systematyczny oparty o taksony)
- Cluster→ Skupienie, skupisko, grupa / klasa / pojęcie
- Nigdy nie mów: klaster, klastering, klastrowanie!

…

Polski elementy w rozwoju analizy skupień

- **Jan Czekanowski** (1882-1965) - wybitny polski antropolog, etnograf, demograf i statystyk, profesor Uniwersytetu Lwowskiego (1913 – 1941) oraz Uniwersytetu Poznańskiego (1946 – 1960).
  - Nowe odległości i metody przetwarzania macierzy odległości w algorytmach, …, tzw. metoda Czekanowskiego.
  - Kontynuacja Jerzy Fierich (1900-1965) Kraków
- **Hugo Steinhaus**, (matematycy Lwów i Wrocław)
  - Wrocławska szkoła taksonomiczna (metoda dendrytowa)
- **Zdzisław Hellwig** (Wrocław)
  - wielowymiarowa analizą porównawcza, i inne …
- Współcześnie …
- „ Sekcja Klasyfikacji i Analizy Danych” (SKAD) Polskiego Towarzystwa Statystycznego

Referencje do literatury (przykładowe)

- Koronacki J. Statystyczne systemy uczące się, WNT 2005.
- Pociecha J., Podolec B., Sokołowski A., Zając K. „Metody taksonomiczne w badaniach społeczno-ekonomicznych”. PWN, Warszawa 1988,
- Stąpor K. „Automatyczna klasyfikacja obiektów” Akademicka Oficyna Wydawnicza EXIT, Warszawa 2005.
- Hand, Mannila, Smyth, „Eksploracja danych”, WNT 2005.
- Larose D: „Odkrywania wiedzy z danych”, PWN 2006.
- Kucharczyk J. „Algorytmy analizy skupień w języku ALGOL 60” PWN Warszawa, 1982,
- Materiały szkoleniowe firmy Statsoft.

Przykłady zastosowań analizy skupień

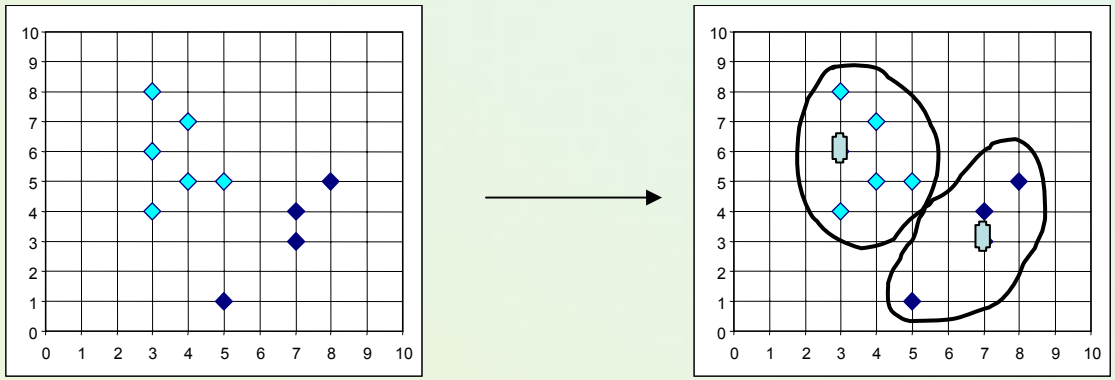
- Zastosowania ekonomiczne:
  - Identyfikacja grup klientów bankowych (np. właścicieli kart kredytowych wg. sposobu wykorzystania kart oraz stylu życia, danych osobowych, demograficznych) $\rightarrow$ cele marketingowe.
  - Systemy rekomendacji produktów i usług.
  - Rynek usług ubezpieczeniowych (podobne grupy klientów).
  - Analiza sieci sprzedaży (np. czy punkty sprzedaży podobne pod względem społecznego sąsiedztwa liczby personelu, itp., przynoszą podobne obroty).
  - Poszukiwanie wspólnych rynków dla produktów.
  - Planowanie, np. nieruchomości
- Badania naukowe (biologia, medycyna, nauki społeczne)
- Analiza zachowań użytkowników serwisów WWW
- Rozpoznawanie obrazów, dźwięku
- Wiele innych

Web Search Result Clustering – Carrot2

Wielowymiarowe statystyczne spojrzenie

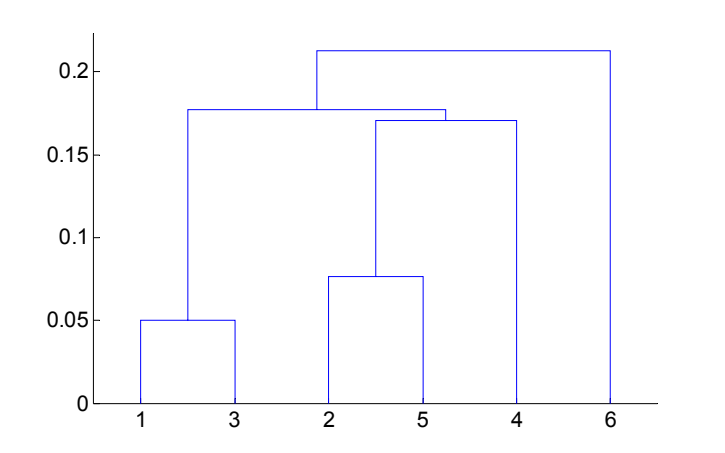
- Obiekt opisany za pomocą *n* zmiennych X_1, X_2,…X_n jest punktem x=(x_1,…,x_n) w *n*-wymiarowej przestrzeni Ω
- Cel podziału na grupy (S) $\rightarrow$ obiekty podobne (reprezentowane przez punkty znajdujące się blisko siebie w przestrzeni) przydzielone do tej samej grupy, a obiekty niepodobne (reprezentowane przez punkty leżące w dużej odległości w przestrzeni) znajdują się w różnych grupach

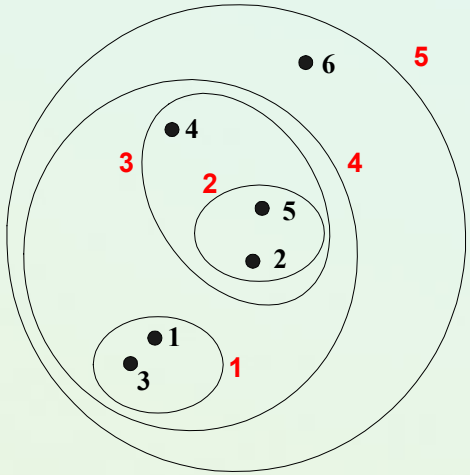
$$S_i \cap S_j = \varnothing \quad (i \neq j;\ i,j = 1,\ldots p) \qquad \bigcup_{i=1}^{p} S_i = \Omega$$



Grupowanie hierarchiczne

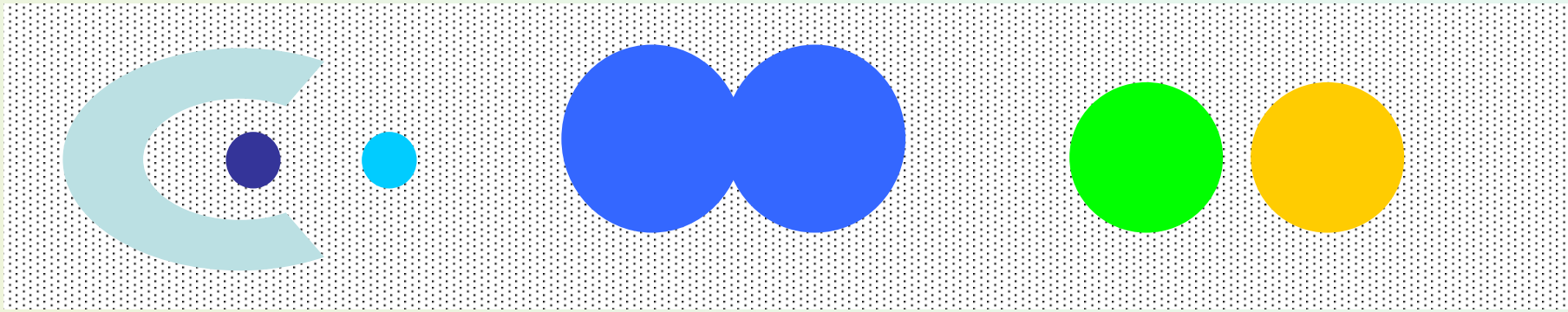
- Tworzy się stopniowo hierarchię zawierających się skupisk
  - Połączenie lub podział podzbiorów obiektów
- Wizualizacja – struktura drzewa nazwana dendrogramem





Spojrzenie gęstościowe: Density-Based Clusters

- Za Tan et al. Introduction to data mining



6 density-based clusters

Inne spojrzenie na skupiska

- Crisp vs. soft clusters (Fuzzy clustering)

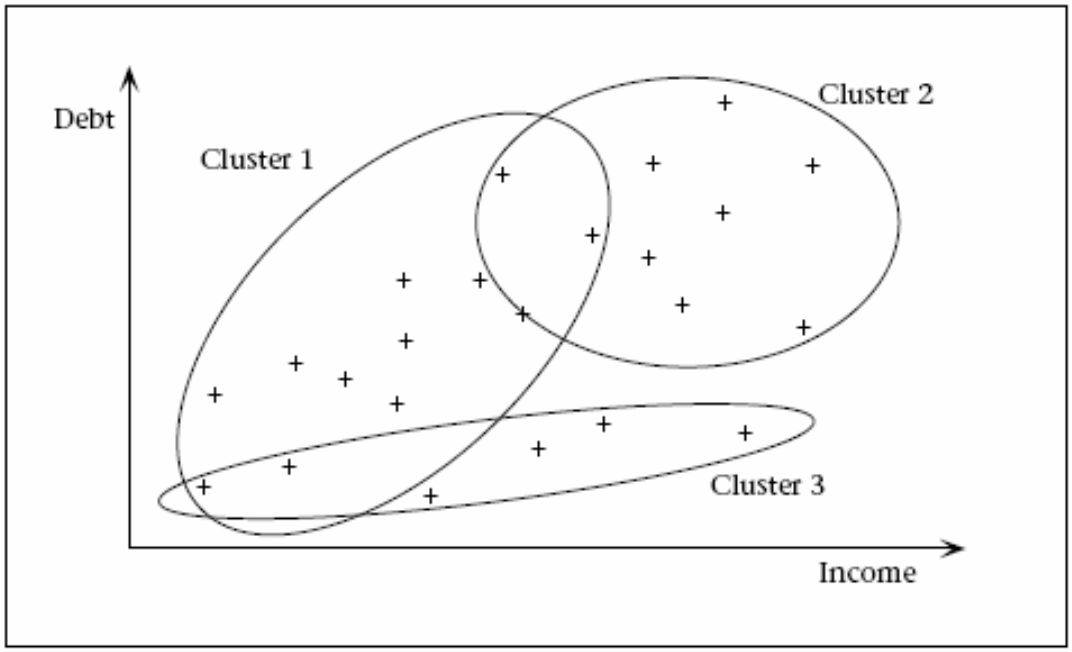


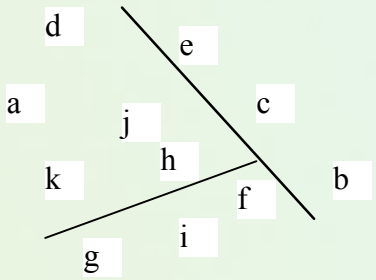
Figure 5. A Simple Clustering of the Loan Data Set into Three Clusters.
Note that original labels are replaced by a +.

Czym jest skupienie?

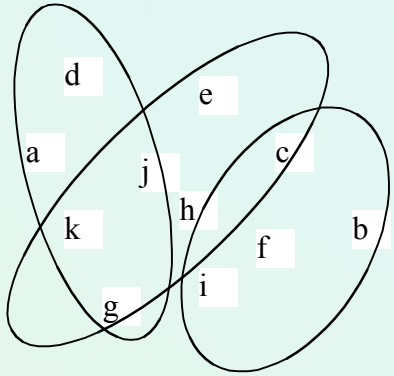
1. Zbiorem najbardziej podobnych obiektów
2. Podzbiór obiektów, dla których odległość jest mniejsza niż ich odległość od obiektów z innych skupień.
3. Podobszar wielowymiarowej przestrzeni zawierający odpowiednio dużą gęstość obiektów, oddzielony od innych podobszarów o dużej gęstości strefą rzadkiego występowania obiektów

Różne sposoby reprezentacji skupień

(a)



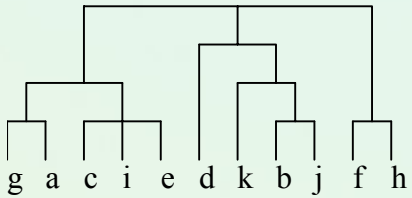
(b)



(c)

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1
…			

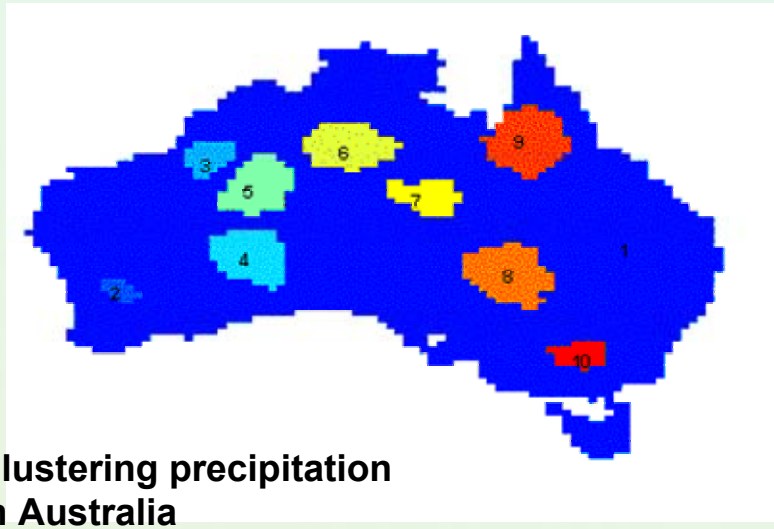
(d)



Możliwe cele analizy skupień

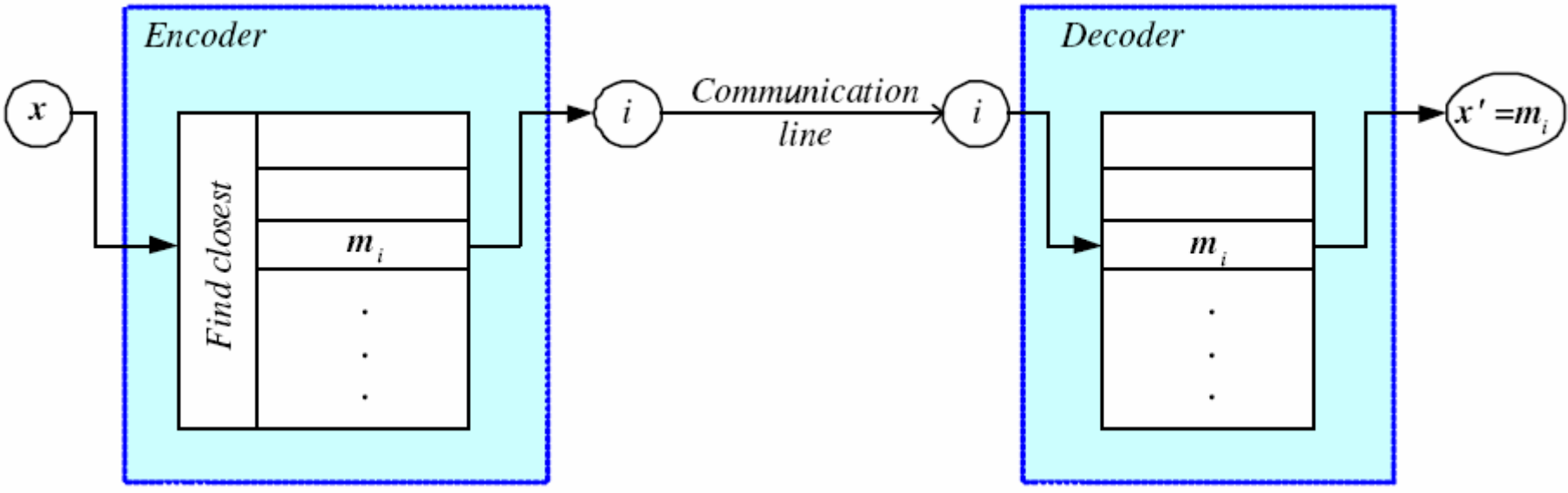
- **Opis / zrozumienie struktury danych**
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations
- **Podsumowywanie danych Summarization**
  - Zredukowanie rozmiaru danych

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Clustering precipitation in Australia

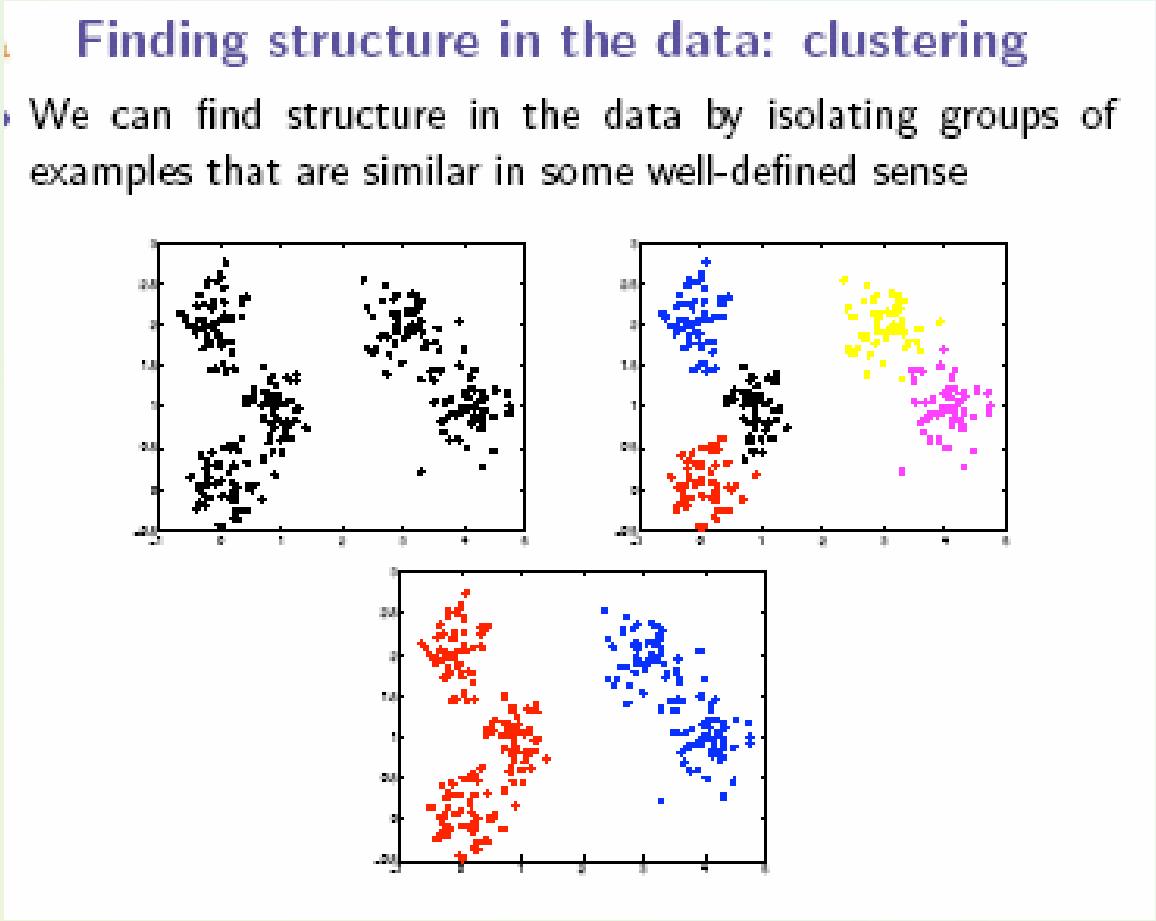
Encoding/Decoding



$$b_i^t = \begin{cases} 1 & \text{if } \left\| \boldsymbol{x}^t - \boldsymbol{m}_i \right\| = \min_j \left\| \boldsymbol{x}^t - \boldsymbol{m}_j \right\| \\ 0 & \text{otherwise} \end{cases}$$

Poszukiwanie „zrozumiałych struktur” w danych

- Ma ułatwiać odnalezienie pewnych spójnych podobszarów danych



- Lecz nadal wiele możliwości

Podział znanych metod

- Podziałowo-optymalizacyjne: Znajdź podział na zadaną liczbę skupień wg. zadanego kryterium.
- Metody hierarchiczne: Zbuduj drzewiastą strukturę skupień.
- Gęstościowo (Density-based): Poszukuj obszarów o większej gęstości występowania obserwacji
- Grid-based: wykorzystujące wielowymiarowy podział przestrzeni siatką ograniczeń
- Model-based: hipoteza co do własności modelu pewnego skupienia i procedura jego estymacji.

Jeszcze inny podział

- Za Jain's tutorial

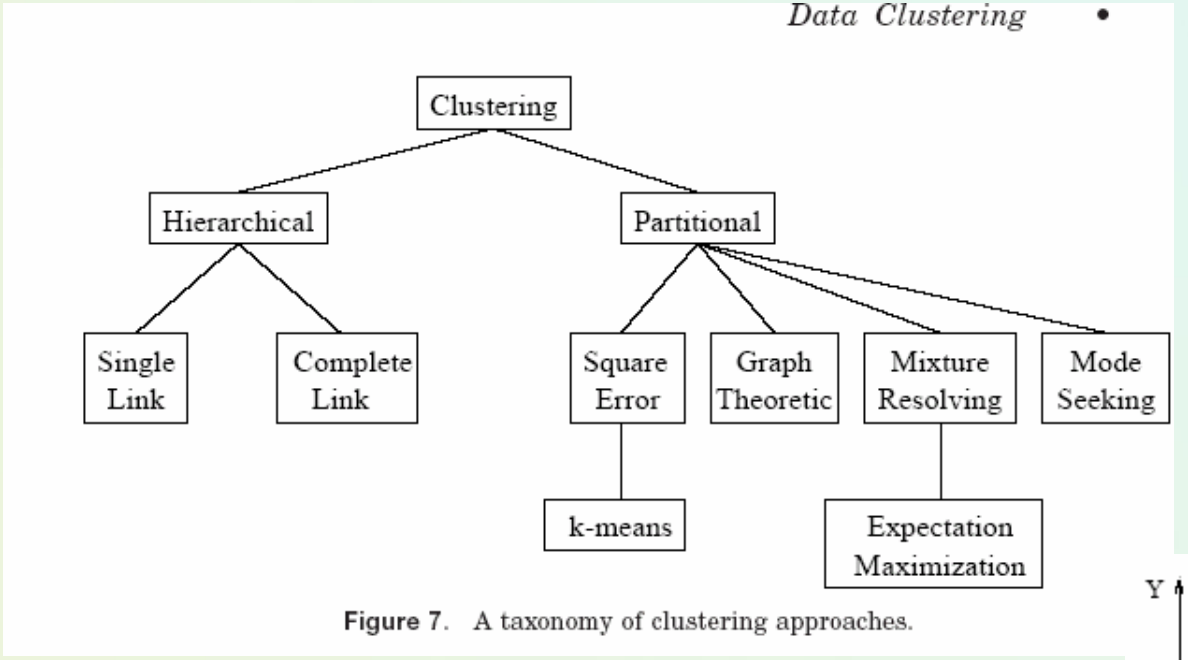


Figure 7. A taxonomy of clustering approaches.

- Ponadto:
  - Crisp vs. Fuzzy
  - Inceremental vs. batch

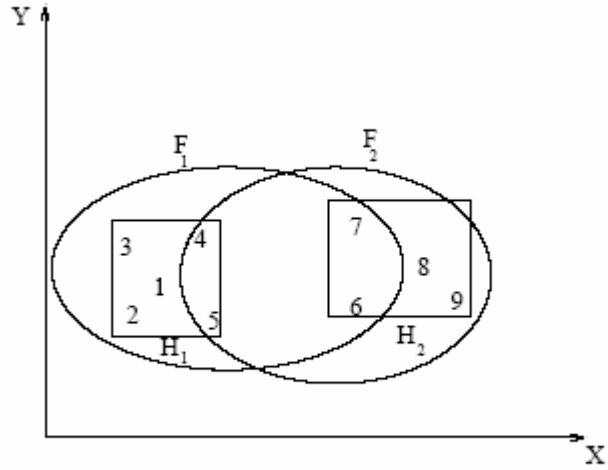
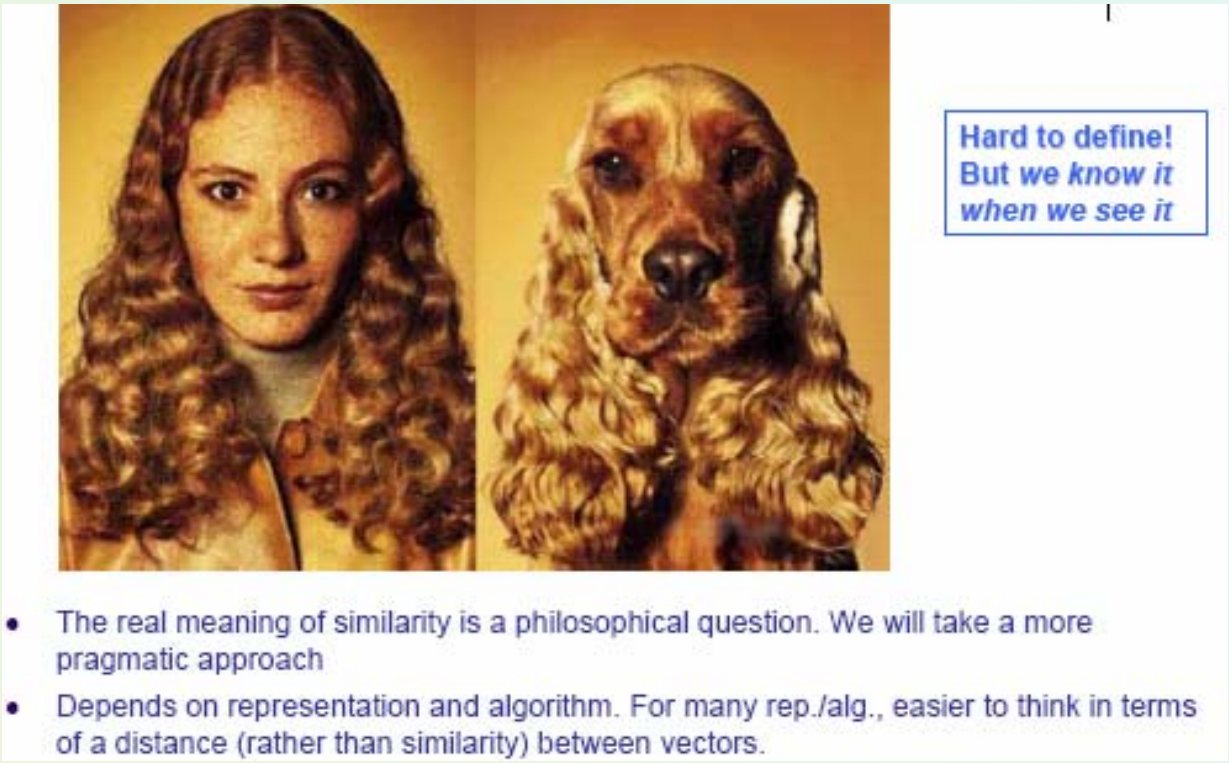


Figure 16. Fuzzy clusters

Podobieństwo – podstawa analizy skupień

- Ciągle dyskusyjne – zwłaszcza w nietechnicznych zastosowaniach



- Prościej mówić o odległościach między obserwacjami
  - Zwłaszcza jak są matematycznie dobrze zdefiniowane
  - Metryka odległości?

Problemy do rozstrzygnięcia

- Jak odwzorować obiekty w przestrzeni?
  - Wybór zmiennych
  - Normalizacja zmiennych
- Jak mierzyć odległości między obiektami?
- Jaką metodę grupowania zastosować?

Trudności z różnym zakresem danych liczbowych

- Normalizacja ma na celu doprowadzenie obiektów lub zmiennych do porównywalnych wielkości. Problem ten dotyczy zmiennych mierzonych w różnych jednostkach (np. sztuki, czas, waluta).

Przykład

- Rozważmy 3 obiekty i dwie zmienne: wiek osoby mierzony w latach i jej dochód mierzony w złotych lub tys. zł.

Zmienna ->	X	Y1	Y2
	Wiek	Dochód	Dochód
Osoba	(w latach)	(w zł)	(w tys. zł)
A	35	12000	12,0
B	37	6700	6,7
C	45	7000	7,0

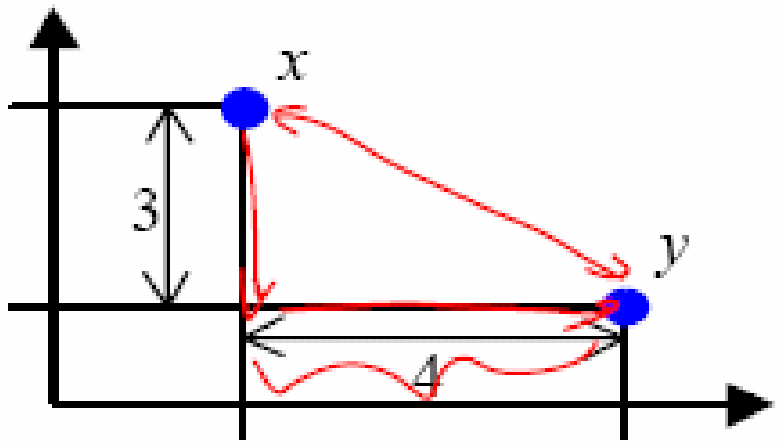
Najczęściej stosowane miary odległości

Nazwa miary odległości	Definicja miary	Uwagi
Odległość euklidesowa	$d_{(rs)} = \left[\sum_{k=1}^{p} \left(x_{(r)k} - x_{(s)k} \right)^2 \right]^{\frac{1}{2}}$	Jest to odległość geometryczna w przestrzeni wielowymiarowej.
Kwadrat odległość euklidesowej	$d_{(rs)} = \left[\sum_{k=1}^{p} \left(x_{(r)k} - x_{(s)k} \right)^2 \right]$	Odległość euklidesową podnosi się do kwadratu, aby przypisać większą wagę obiektom, które są bardziej oddalone
Odległość miejska (Manhattan, City block)	$d_{(rs)} = \sum_{k=1}^{p} \left\| x_{(r)k} - x_{(s)k} \right\|$	Jest to przeciętna różnica mierzoną wzdłuż wymiarów. W większości przypadków ta miara odległości daje podobne wyniki, jak zwykła odległość euklidesowa. W przypadku tej miary, wpływ pojedynczych dużych różnic (przypadków odstających) jest stłumiony (ponieważ nie podnosi się ich do kwadratu).

Charakterystka miar

Odległość Czebyszewa	$d_{(rs)} = \max\left\|x_{(r)k} - x_{(s)k}\right\|$	Stosowna jest w przypadkach, w których chcemy zdefiniować dwa obiekty jako "inne" wtedy, gdy różnią się one w jednym dowolnym wymiarze.
Kwadrat potęgowa	$d_{(rs)} = \left[\sum_{k=1}^{p}\left(x_{(r)k} - x_{(s)k}\right)^{a}\right]^{\frac{1}{b}}$ a, b – parametry określane przez użytkownika	Stosowana, gdy chcemy zwiększyć lub zmniejszyć wzrastającą wagę, która jest przypisana do wymiarów, na których odpowiednie obiekty bardzo się różnią. Parametr a steruje wzrastającą wagą, która jest przypisana różnicom w poszczególnych wymiarach, parametr b steruje wzrastającą wagą, która jest przypisana większym różnicom między obiektami. Jeśli a i b są równe 2, to odległość ta jest równa odległości euklidesowej.
Niezgodność procentowa	Liczba obserwacji, dla których: $\frac{X_{(r)k} \neq X_{(s)k}}{k}$	Jest szczególnie przydatna wtedy, gdy dane dla wymiarów objętych analizą są z natury dyskretne (oraz w skali nominalnej).

Prosty przykład obliczeń



1 : Euclidean distance : $\sqrt[2]{4^2+3^2}=5.$

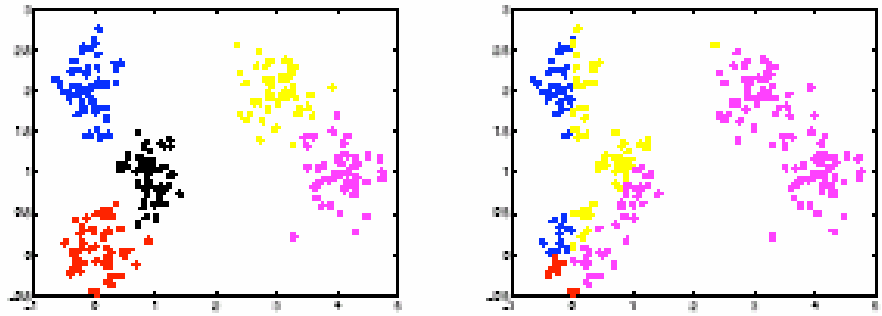
2 : Manhattan distance : $4+3=7.$

3 : "sup" distance : $\max\{4,3\}=4.$

Problem doboru miary odległości / podobieństwa

- Nietrywialny i silnie wpływa na wynik
- Różne miary odległości

Clustering: metric



- Clustering results are crucially dependent on the measure of similarity (or distance) between the "points" to be clustered

Algorytmy podziałowo - optymalizacyjne

- Zadanie: Podzielenie zbioru obserwacji na *K* zbiorów elementów (skupień C), które są jak najbardziej jednorodne
- Jednorodność – funkcja oceny
- Intuicja $\rightarrow$ zmienność wewnątrzskupieniowa *wc*(C) i zmienność międzyskupieniowa *bc*(C)
  - Możliwe są różne sposoby zdefiniowania
  - np. wybierzmy środki skupień $\mathbf{r}_k$ (centroidy) $\mathbf{r}_k = \frac{1}{n_k}\sum_{\mathbf{x}\in C_k}\mathbf{x}$
  - Wtedy

$$wc(C) = \sum_{k=1}^{K}\sum_{\mathbf{x}\in C_k} d(\mathbf{x},\mathbf{r}_k)^2$$

$$bc(C) = \sum_{1\le j<k\le K} d(\mathbf{r}_j,\mathbf{r}_k)^2$$

Podstawowe algorytmy podziałowe

- Metoda *K* - średnich $\rightarrow$ minimalizacja *wc*(*C*)
- Przeszukiwanie przestrzeni możliwych przypisań $\rightarrow$ bardzo kosztowne (oszacowanie w ks. Koronackiego)
- Problem optymalizacji kombinatorycznej $\rightarrow$ systematyczne przeszukiwanie metodą iteracyjnego udoskonalania:
  - Rozpocznij od rozwiązania początkowego (losowego).
  - Ponownie przypisz punkty do skupień tak, aby otrzymać największą zmianę w funkcji oceny.
  - Przelicz zaktualizowane środki skupień, …
  - Postępuj aż do momentu, w którym nie ma już żadnych zmian w funkcji oceny lub w składzie grup.
- Zachłanne przeszukiwanie $\rightarrow$ proste i prowadzi do co najmniej lokalnego minimum. Różne modyfikacje, np. rozpoczynania od kilku rozwiązań startowych
- Złożoność algorytmy K - średnich $\rightarrow$ *O*(*Knl*)

Przykład (z macierzą początkową)

- Zbiór danych: $x_1 = \begin{bmatrix}1\\2\end{bmatrix}$ $x_2 = \begin{bmatrix}1\\3\end{bmatrix}$ $x_3 = \begin{bmatrix}1\\4\end{bmatrix}$ $x_4 = \begin{bmatrix}2\\1\end{bmatrix}$ $x_5 = \begin{bmatrix}3\\1\end{bmatrix}$

  $x_6 = \begin{bmatrix}3\\3\end{bmatrix}$ $x_7 = \begin{bmatrix}4\\1\end{bmatrix}$ $x_8 = \begin{bmatrix}5\\2\end{bmatrix}$ $x_9 = \begin{bmatrix}5\\3\end{bmatrix}$ $x_{10} = \begin{bmatrix}5\\4\end{bmatrix}$

- Początkowy przydział $B(0) = \begin{bmatrix}1&0&0&0&1&0&0&0&1&0\\0&1&0&1&0&1&0&1&0&1\\0&0&1&0&0&0&1&0&0&0\end{bmatrix}$
- Przeliczenie centroidów $R(0) = \begin{bmatrix}3&3.2&2.5\\2&2.6&2.5\end{bmatrix}$
- Przeliczenie odległości

$$D(1) = \begin{bmatrix}2&2.24&2.83&1.41&1&1&1.41&2&2.24&2.83\\2.28&2.24&2.61&2&1.61&0.45&1.79&1.9&1.84&2.28\\1.58&1.58&2.12&1.58&1.58&0.71&2.12&2.55&2.55&2.91\end{bmatrix}$$

- Ponowny przydział do skupień:

$$B(1) = \begin{bmatrix}0&0&0&\mathbf{1}&1&0&0&0&0&0\\0&0&0&0&0&1&0&1&\mathbf{1}&1\\\mathbf{1}&\mathbf{1}&1&0&0&0&1&0&0&0\end{bmatrix}$$

Przykład cd.

- Przeliczenie centroidów $R(1) = \begin{bmatrix} 3 & 5 & 1.5 \\ 1 & 3 & 3 \end{bmatrix}$
- Ponowny przydział do skupień: $B(2) = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$
- Warunek końcowy: $B(2) = B(1)$

$$G_1 = \{x_4, x_5, x_7\} \quad G_2 = \{x_8, x_9, x_{10}\} \quad G_3 = \{x_1, x_2, x_3, x_6\}$$

Metody optymalizacyjno-iteracyjne (*k*-średnich)

- Jednocześnie obliczana jest funkcja błędu podziału - ogólna suma kwadratów odległości wewnątrzgrupowych liczonych od środków ciężkości grup: tzn.

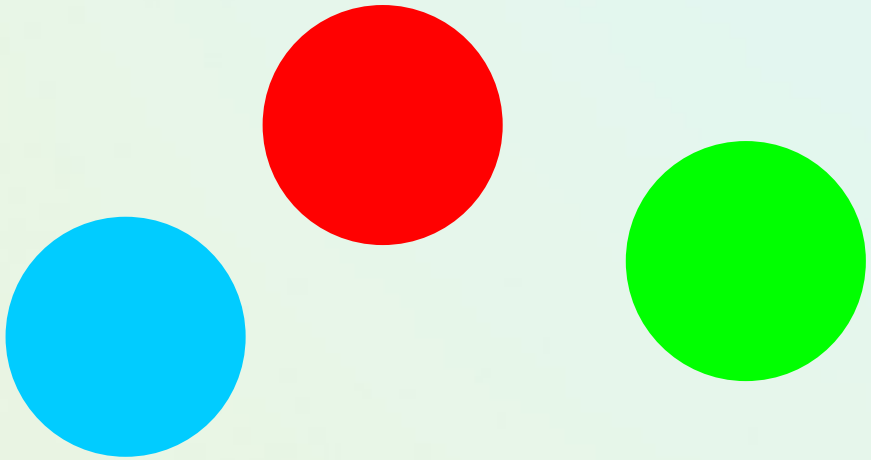
$$F = \sum_{j=1}^{k} \sum_{O_i \in S_j} d\left(O_i, M_j\right)^2$$

gdzie *d* jest odległością euklidesową.

W praktyce proces jest zbieżny po kilku lub kilkunastu iteracjach. Ponieważ w ogólności algorytm nie musi być zbieżny, ustala się maksymalną liczbę iteracji (*L*).

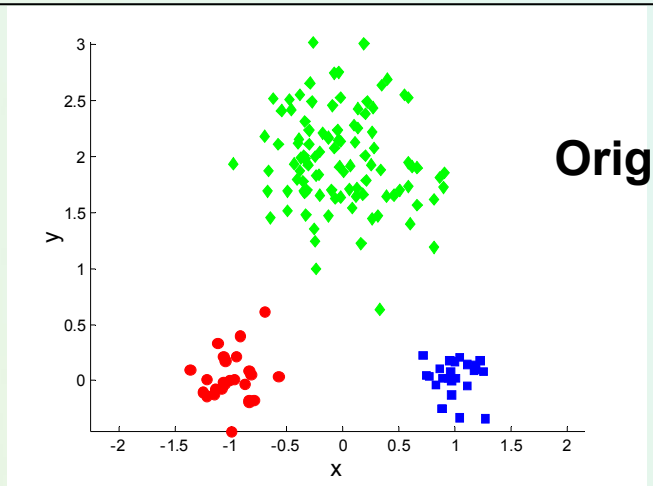
Pewne ukierunkowanie K-średnich

- Tworzy się kuliste kształty skupień

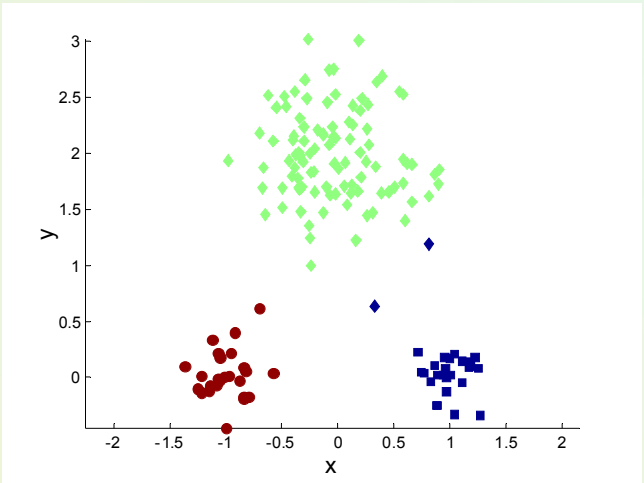


- Co z obserwacjami odstającymi i nieregularnymi kształtami skupień?

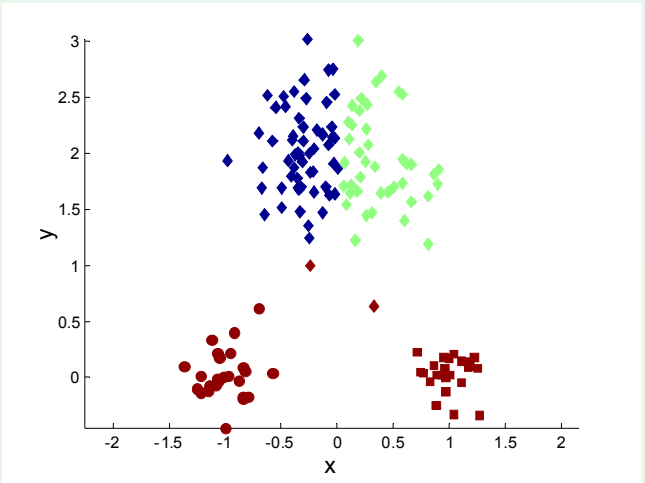
Czy łatwo określić liczbę skupień



Original Points



Optimal Clustering



Sub-optimal Clustering

Ustalanie liczby skupień i startowych centroidów

Liczbę skupień wybiera się na podstawie przesłanek merytorycznych albo szacuje się je metodami hierarchicznymi. Można dokonać obliczeń dla wszystkich wartości *k* z ustalonego przedziału:

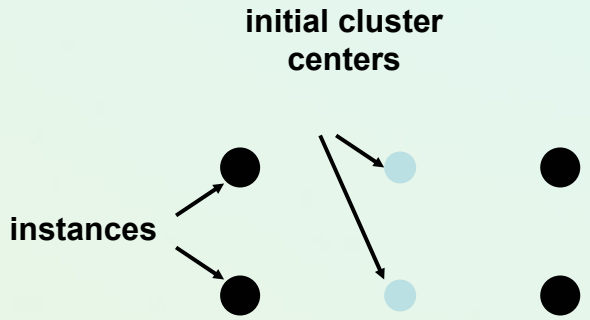
$$k_{\min} \leq k \leq k_{\max}$$

Możliwe są różne podejścia:

1. Arbitralny sposób np. przyjmuje się współrzędne pierwszych *k* obiektów jako zalążki środków ciężkości
2. Losowy wybór środków ciężkości, przy czym może to być losowy wybór *k* obiektów ze zbioru danych albo losowy wybór *k* punktów przestrzeni niekoniecznie pokrywających się z położeniem obiektów
3. Wykorzystanie algorytmu optymalizującego w pewien sposób położenie początkowych środków ciężkości np. przez uwzględnianie *k* obiektów leżących daleko względem siebie
4. Przyjęcie jako początkowych środków ciężkości uzyskanych na podstawie podziału otrzymanego inna metodą, głównie jedną z metod hierarchicznych

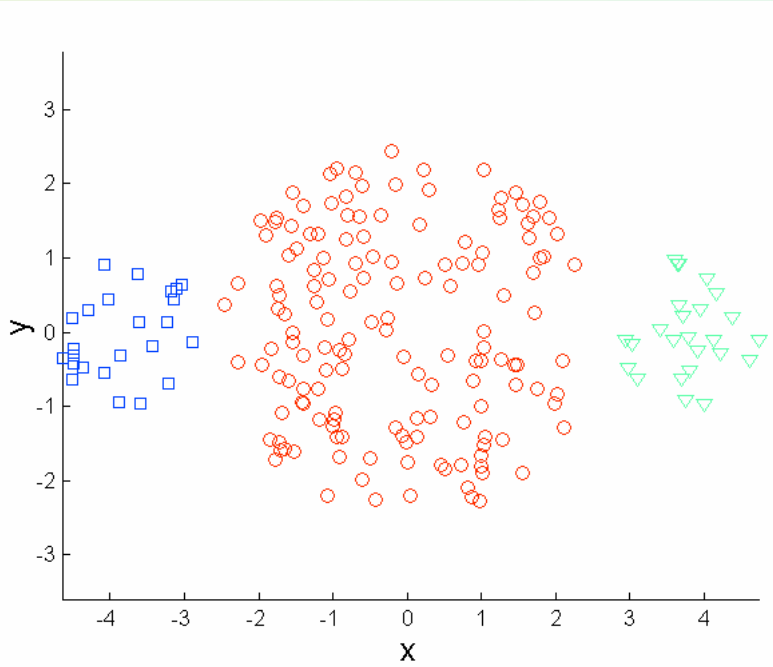
Uwagi nt. podziałowo-optymalizacyjnych

- Dobór parametru *k*
- Kwestia heurystyki wyboru początkowych ziaren
- Czy jest zawsze zbieżny do „optimum globalnego”
  - Przykład:

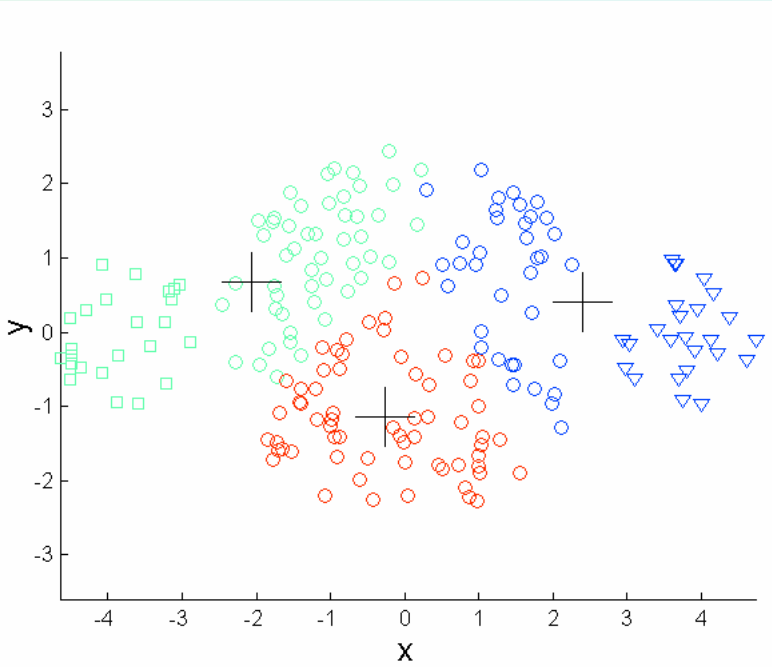


- Możesz próbować kilka uruchomień z różnymi parametrami

Ograniczenia – nietypowe rozkłady o różnej liczności

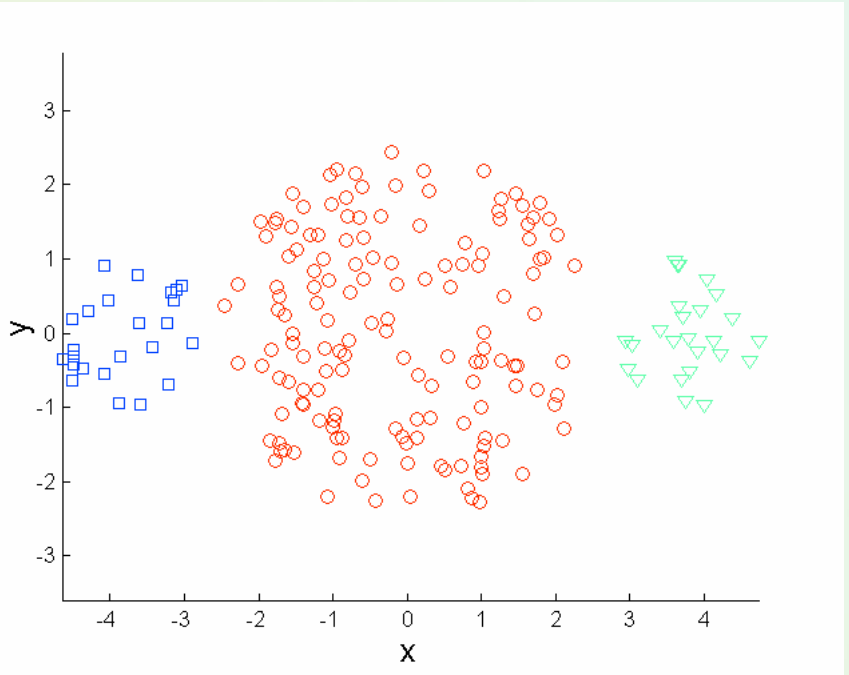


Original Points

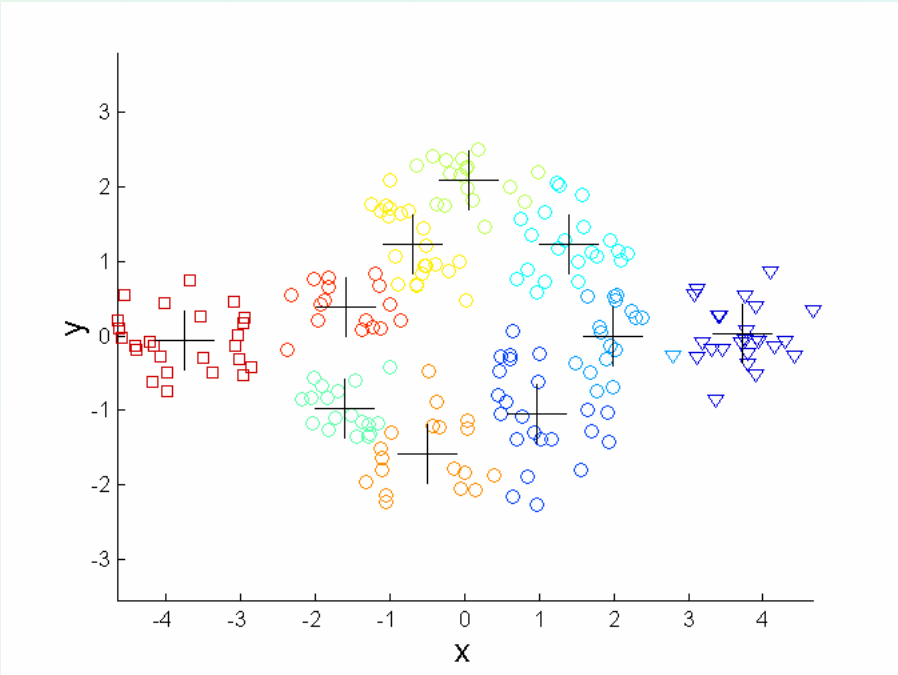


K-means (3 skupienia)

Próby rozwiązań – graj różną ilością skupień



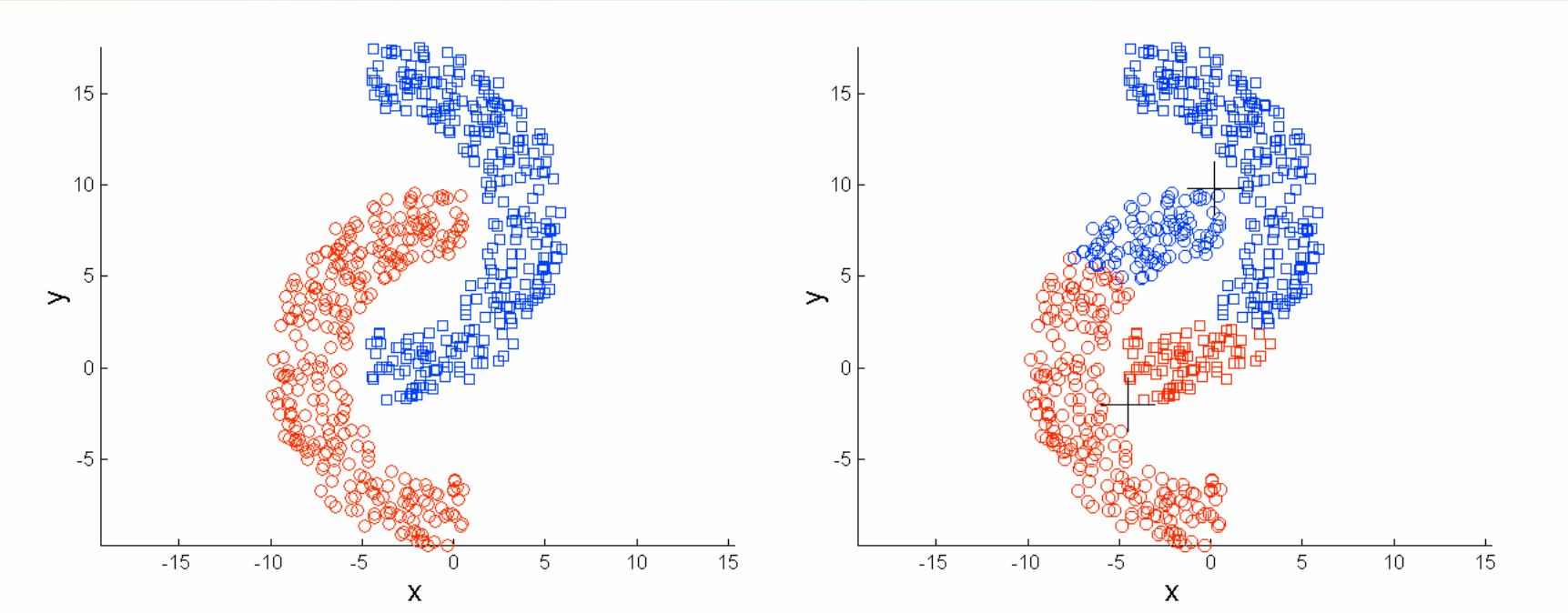
Original Points



K-means Clusters

One solution is to use many clusters.
Find parts of clusters, but need to put together.

Ograniczenia K-średnich: Niesferyczne kształty



Original Points

K-means (2 skupienia)

K-means krótkie podsumowanie

Zalety

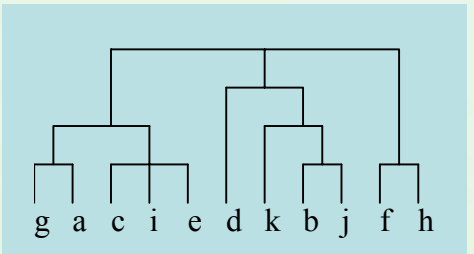
- Proste i łatwe do zrozumienia
- Reprezentacja skupień jako centroidy (centra, obserwacje centralne)

Wady

- Jawne podanie liczby skupień
- Wszystkie przykłady muszą być przydzielone do skupień
- Problem z outliers (za duża wrażliwość)
- Ukierunkowanie na jednorodne „sferyczne” kształty skupień

Metody hierarchiczne

- Bottom up (aglomerative)
  - Start with single-instance clusters
  - At each step, join the two closest clusters
  - Design decision: distance between clusters
    - e.g. two closest instances in clusters vs. distance between means
- Top down (divisive approach)
  - Start with one universal cluster
  - Find two clusters
  - Proceed recursively on each subset
  - Can be very fast
- Both methods produce a *dendrogram*

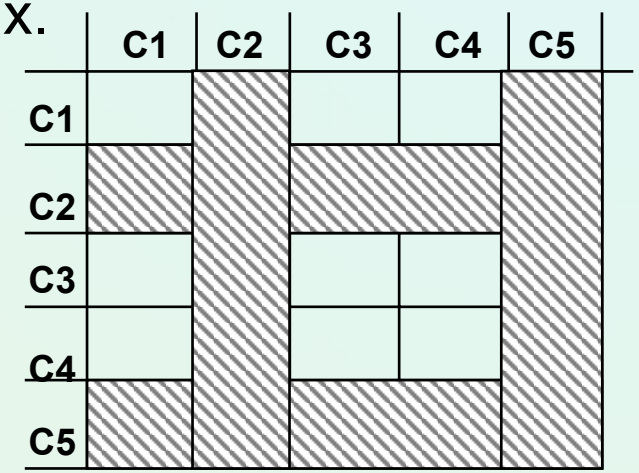


Hierarchiczne metody aglomeracyjne - algorytm

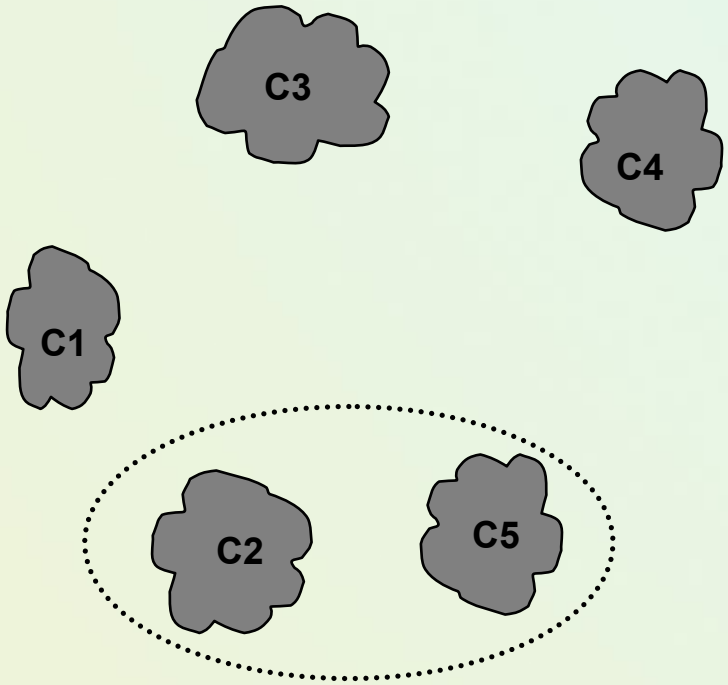
1. W macierzy odległości znajduje się parę skupień najbliższych sobie.
2. Redukuje się liczbę klas łącząc znalezioną parę
3. Przekształca się macierz odległości metodą wybraną jako kryterium klasyfikacji
4. Powtarza się kroki 1- 3 dopóki nie powstanie jedna klasa zawierająca wszystkie skupienia.

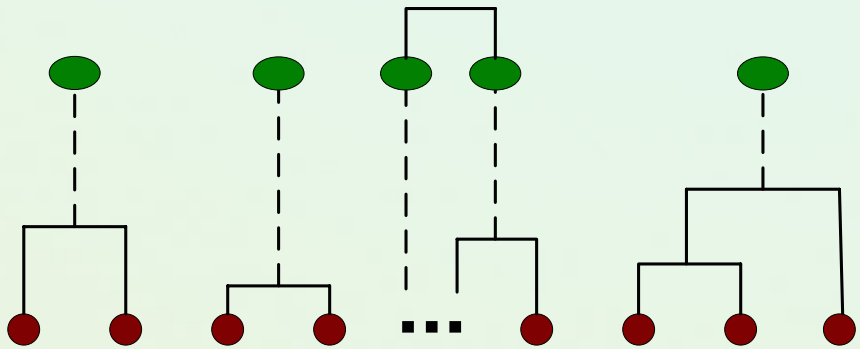
Jak przeliczać macierz odległości?

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



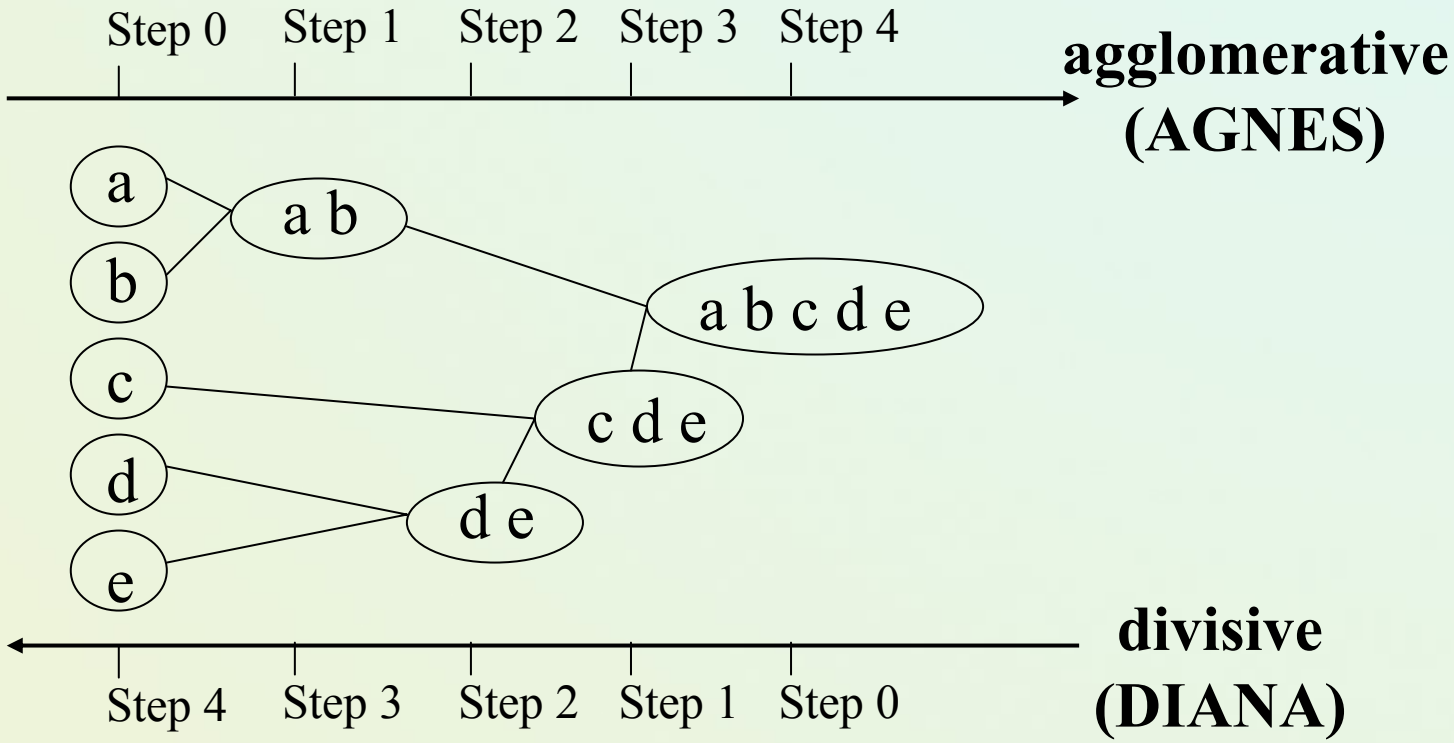
Proximity Matrix





AHC - komentarz

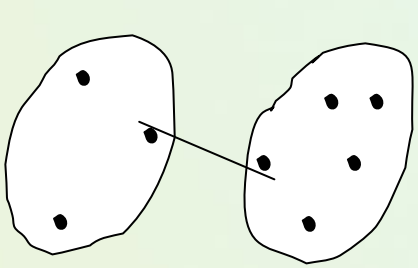
- Użycie i przekształcanie macierzy odległości.
- Nie ma predefiniowanej liczby klas ***k*** jako parametr wejściowy, ale czy zawsze budujemy pełne drzewo. Parametry: podstawowa odległość i metoda łączenia



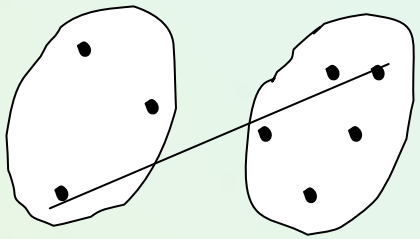
AHC – wybór metody łączenia

1. Najbliższego sąsiedztwa (*Single linkage*, *Nearest neighbor*)
2. Najdalszego sąsiedztwa (*Complete linkage*, *Furthest neighbor*)
3. Mediany (*Median clustering*)
4. Środka ciężkości (*Centroid clustering*)
5. Średniej odległości wewnątrz skupień (*Average linkage within groups*)
6. Średniej odległości między skupieniami (*Average linkage between groups*)
7. Minimalnej wariancji Warda (*Ward's method*)

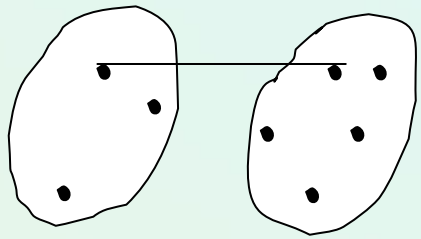
Porównanie sposobu wyznaczania odległości między skupieniami w wybranych metodach aglomeracyjnych



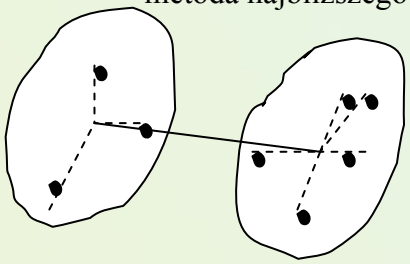
metoda najbliższego sąsiedztwa



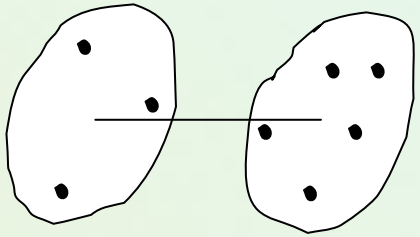
metoda najdalszego sąsiedztwa



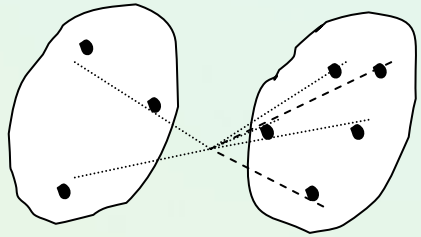
metoda mediany



metoda środka ciężkości



metoda średniej grupowej



metoda Warda

Odległości między skupieniami

Single linkage
minimum distance:

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \left\| p - p' \right\|$$

Complete linkage
maximum distance:

$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \left\| p - p' \right\|$$

mean distance:

$$d_{mean}(C_i, C_j) = \left\| m_i - m_j \right\|$$

average distance:

$$d_{ave}(C_i, C_j) = 1/(n_i n_j) \sum_{p \in C_i} \sum_{p' \in C_j} \left\| p - p' \right\|$$

m_i is the mean for cluster C_i n_i is the number of points in C_i

Single Link Agglomerative Clustering

- Użyj maksymalnego podobieństwa dwóch obiektów:

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- Prowadzi do „(long and thin) clusters due to *chaining effect*” (efekt łańcuchowy); prowadzić do formowania grup niejednorodnych (heterogenicznych);
  - Dogodne w specyficznych zastosowaniach
- Pozwala na wykrycie obserwacji odstających, nie należących do żadnej z grup, i warto przeprowadzić klasyfikację za jej pomocą na samym początku, aby wyeliminować takie obserwacje i przejść bez nich do właściwej części analizy

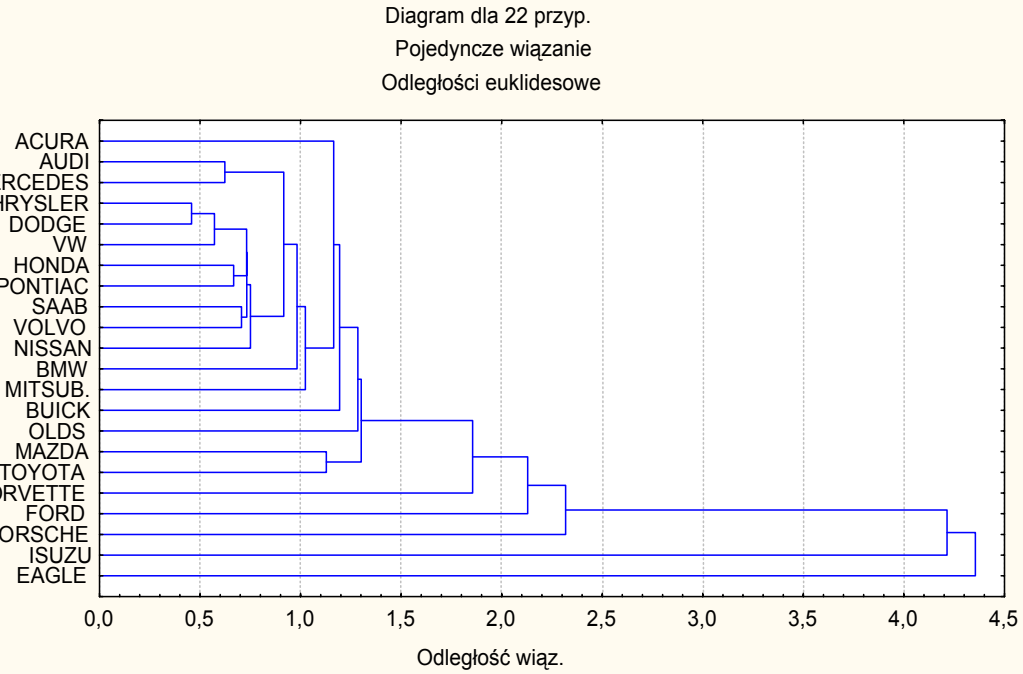
Complete Link Agglomerative Clustering

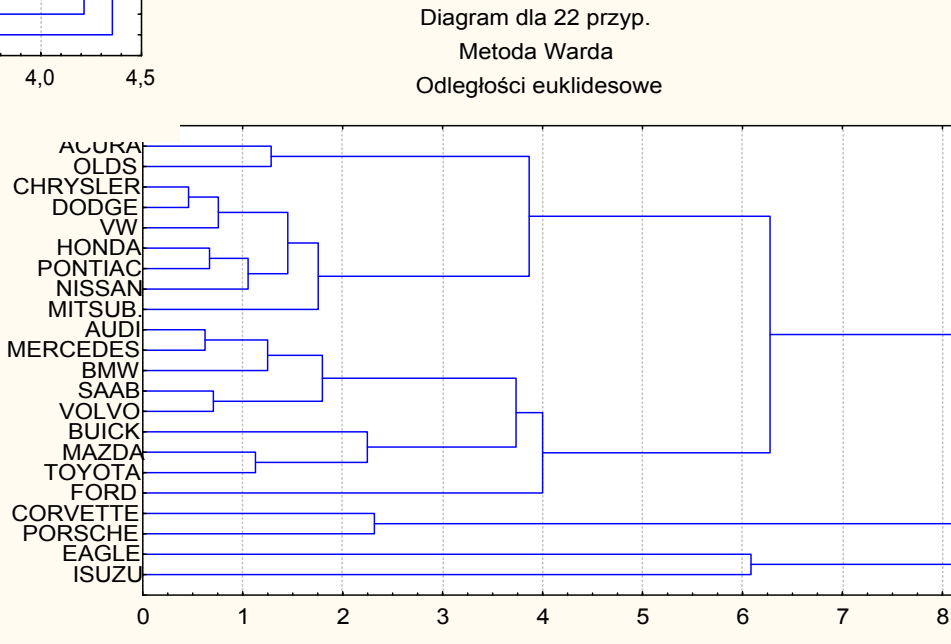
- Użyj maksymalnej odległości – minimalnego podobieństwa

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

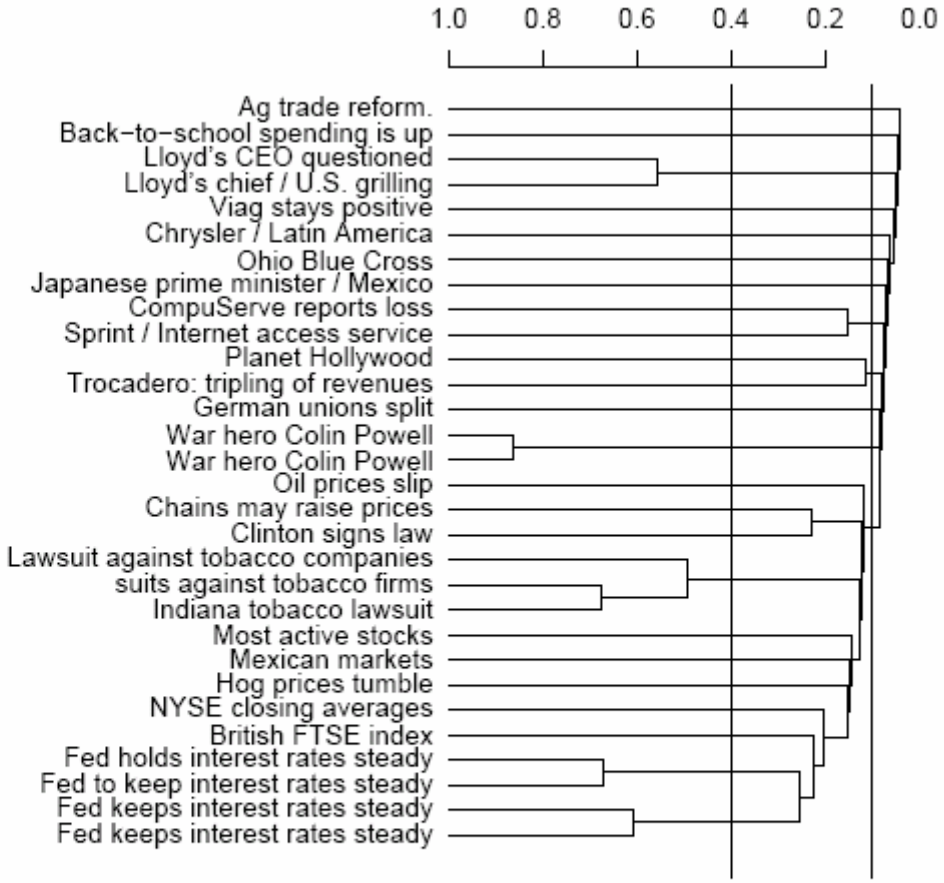
- Ukierunkowana do “tight,” spherical clusters
- Metoda zalecana gdy, kiedy obiekty faktycznie formują naturalnie oddzielone "kępki". Metoda ta nie jest odpowiednia, jeśli skupienia są w jakiś sposób wydłużone lub mają naturę "łańcucha".

Wrażliwość na dobór metod łączenia skupień



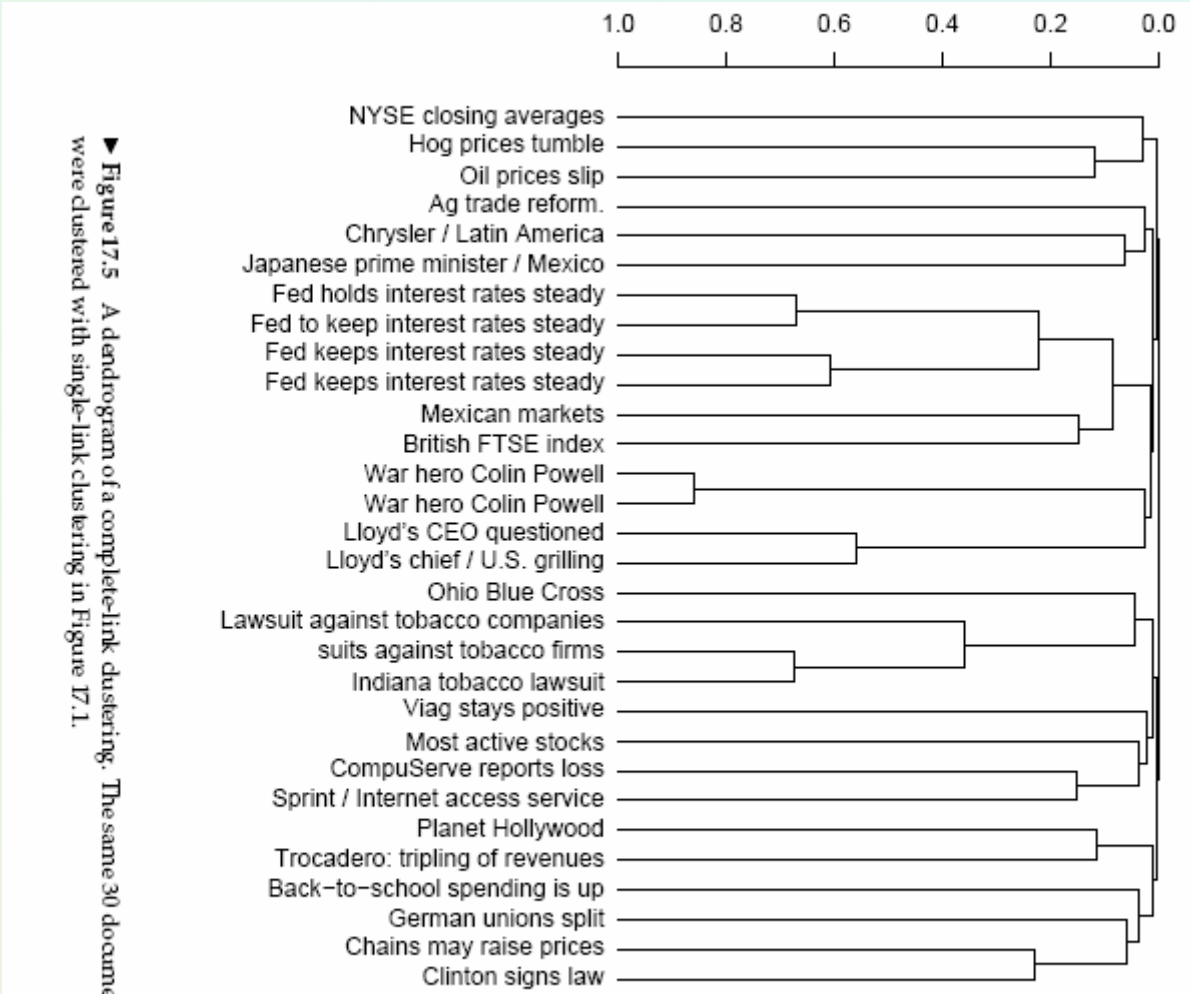


Maning – texts – single linkage



▼ **Figure 17.1** A dendrogram of a single-link clustering of 30 documents from Reuters-RCV1. Two possible cuts of the dendrogram are shown: at 0.4 into 24 clusters and at 0.1 into 12 clusters.

Maning – texts / complete linkage



▶ **Figure 17.5** A dendrogram of a complete-link clustering. The same 30 documents were clustered with single-link clustering in Figure 17.1.

Single vs. Complete Linkage

- A.Jain et al.: Data Clustering. A Review.

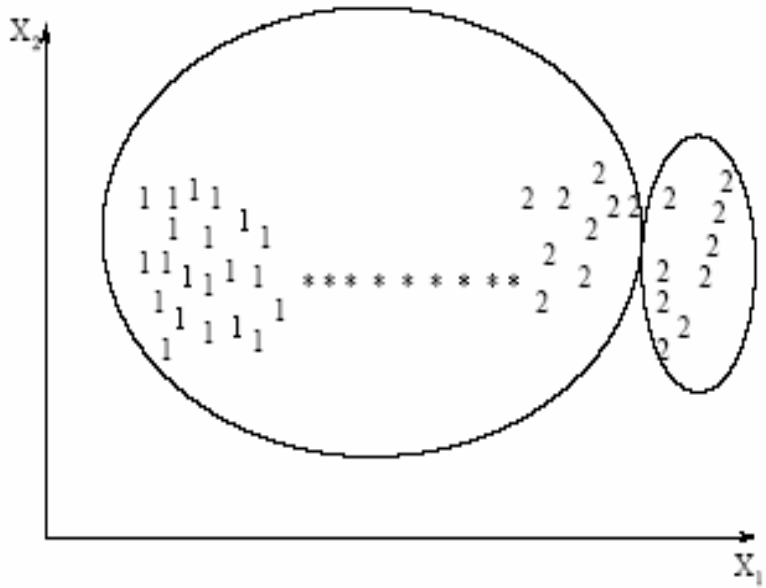


Figure 12. A single-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

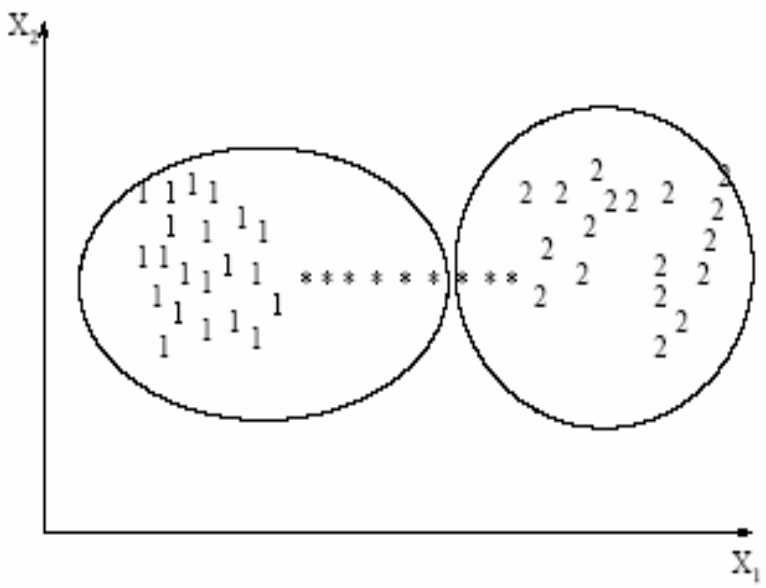
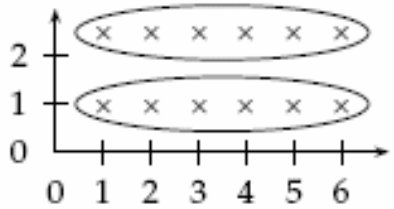
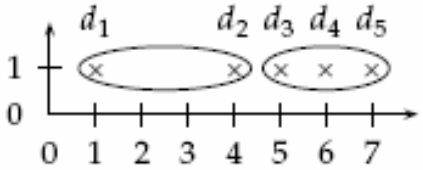


Figure 13. A complete-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

Single vs. complete linkage



► **Figure 17.6** Chaining in single-link clustering. The local criterion in single-link clustering can cause undesirable elongated clusters.



► **Figure 17.7** Outliers in complete-link clustering. The five documents have the x-coordinates $1 + 2\epsilon, 4, 5 + 2\epsilon, 6$ and $7 - \epsilon$. Complete-link clustering creates the two clusters shown as ellipses. The most intuitive two-cluster clustering is $\{\{d_1\}, \{d_2, d_3, d_4, d_5\}\}$, but in complete-link clustering, the outlier d_1 splits $\{d_2, d_3, d_4, d_5\}$ as shown.

Metoda średnich połączeń [Unweighted pair-group average]

- W metodzie tej odległość między dwoma skupieniami oblicza się jako średnią odległość między wszystkimi parami obiektów należących do dwóch różnych skupień
- Metoda ta jest efektywna, gdy obiekty formują naturalnie oddzielone "kępki", ale zdaje także egzamin w przypadku skupień wydłużonych, mających charakter "łańcucha"

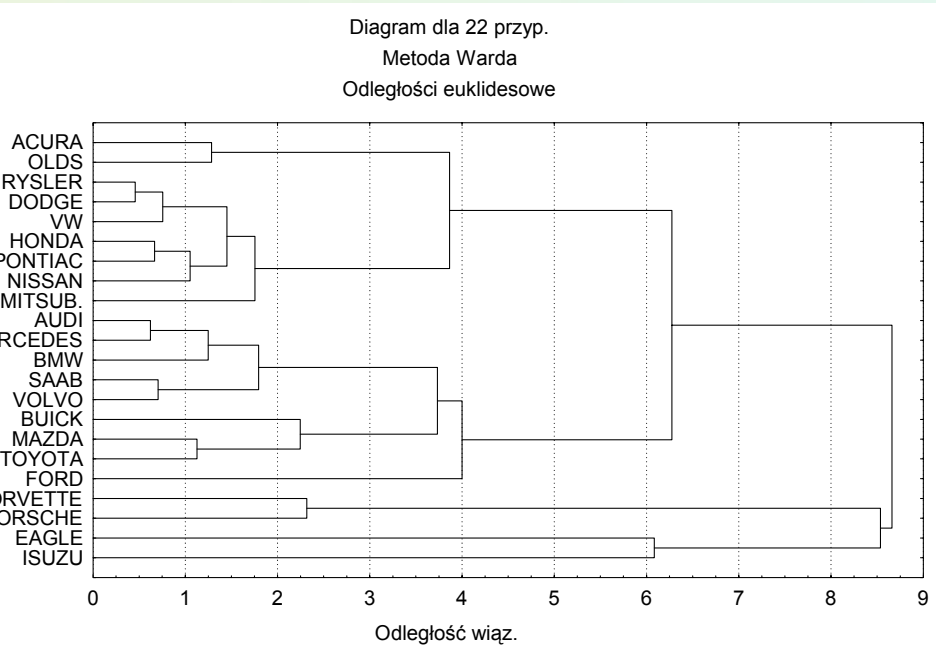
Metoda ważonych środków ciężkości (mediany) [Weighted pair-group centroid].

- Jest to metoda podobna jak poprzednia, z tym wyjątkiem, że w obliczeniach wprowadza się „ważenie”, aby uwzględnić różnice między wielkościami skupień (tzn. liczbą zawartych w nich obiektów).
- Zatem, metoda ta jest lepsza od poprzedniej w sytuacji, gdy istnieją (lub podejrzewamy, że istnieją) znaczne różnice w rozmiarach (liczności) skupień

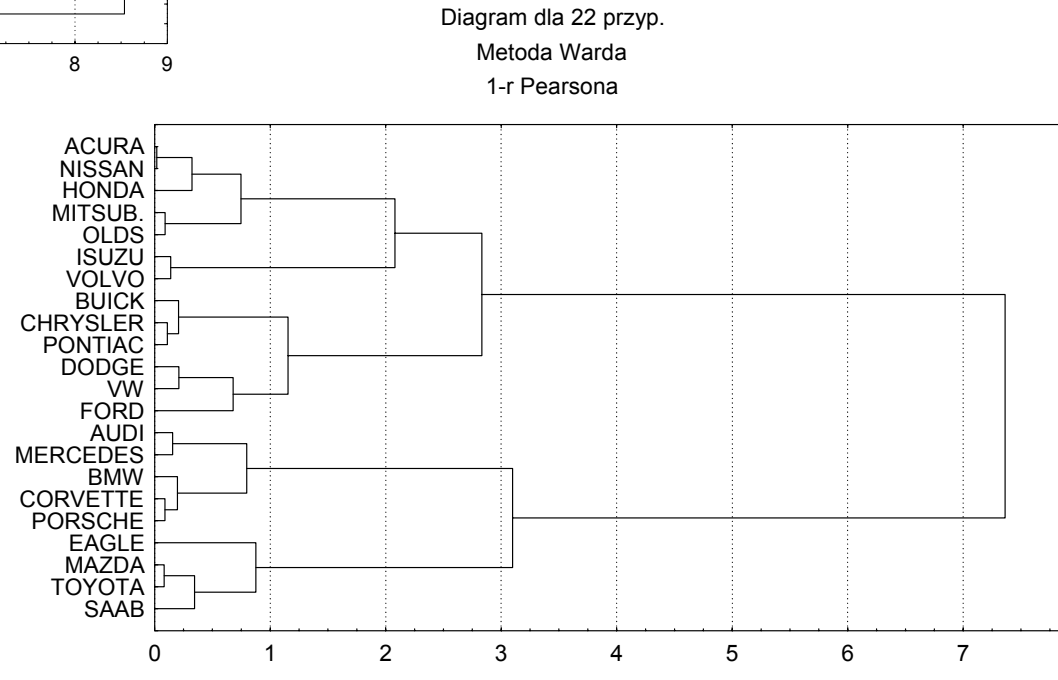
Metody łączenia – Ward method

- Gdy powiększamy jedno ze skupień *C*k, wariancja wewnątrzgrupowa (liczona przez kwadraty odchyleń od średnich w zbiorach *C*k) rośnie.
- Metoda polega na takim powiększaniu zbiorów *C*k, która zapewnia najmniejszy przyrost tej wariancji dla danej iteracji.
- Kryterium grupowania jednostek: minimum zróżnicowania wektorów cech **x**j tworzących zbiór *C*k (*k* = 1,....., *K*) względem wartości średnich w tych zbiorach.
- Ogólnie, metoda ta jest traktowana jako bardzo efektywna, chociaż zmierza do tworzenia skupień o małej wielkości $\rightarrow$ zrównoważone drzewa o wielu elemenatach
- Ważne – powiązanie z miarą odległości między obiektami (Pearson vs. inne)

Przykłady użycia metody Warda

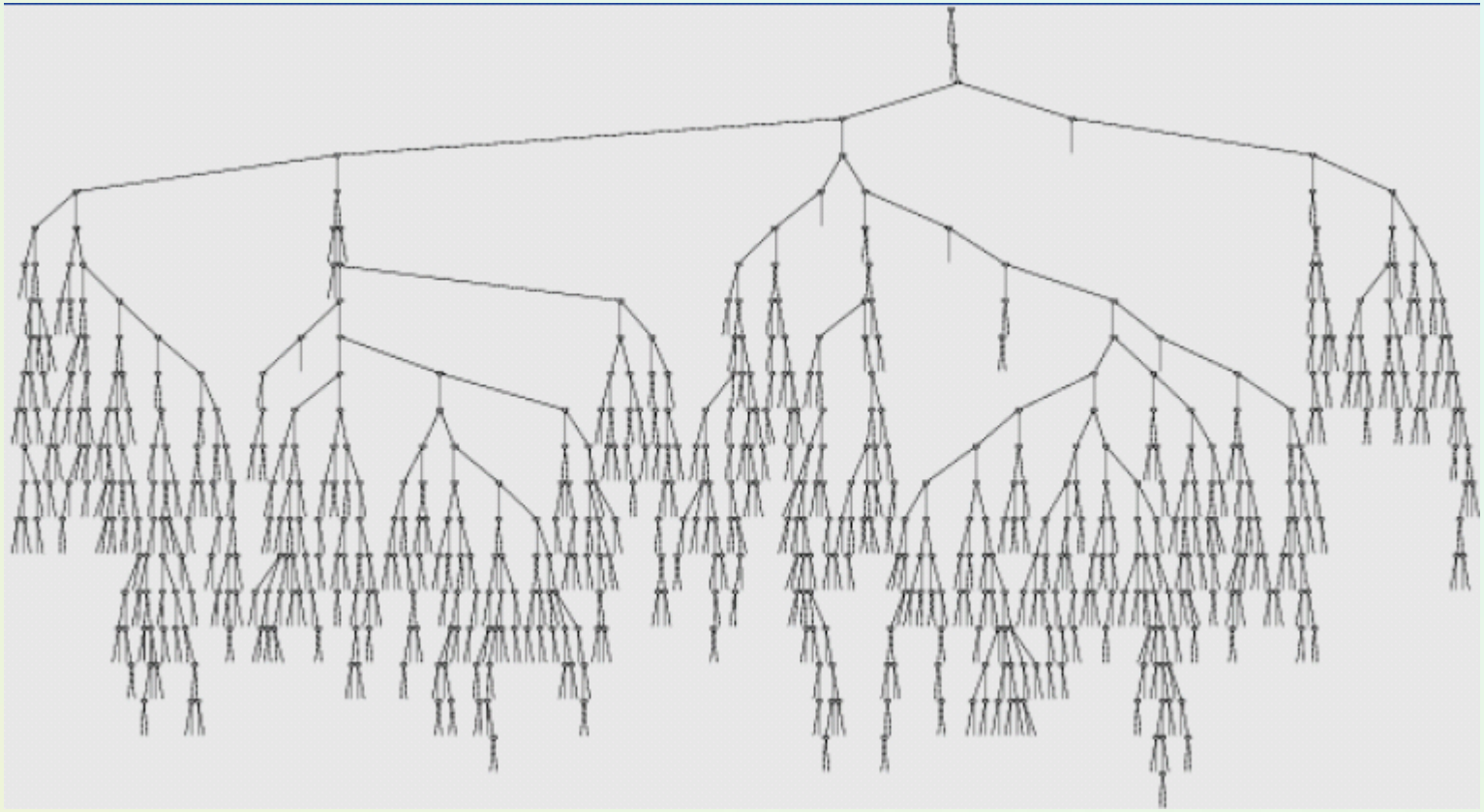


Cars data

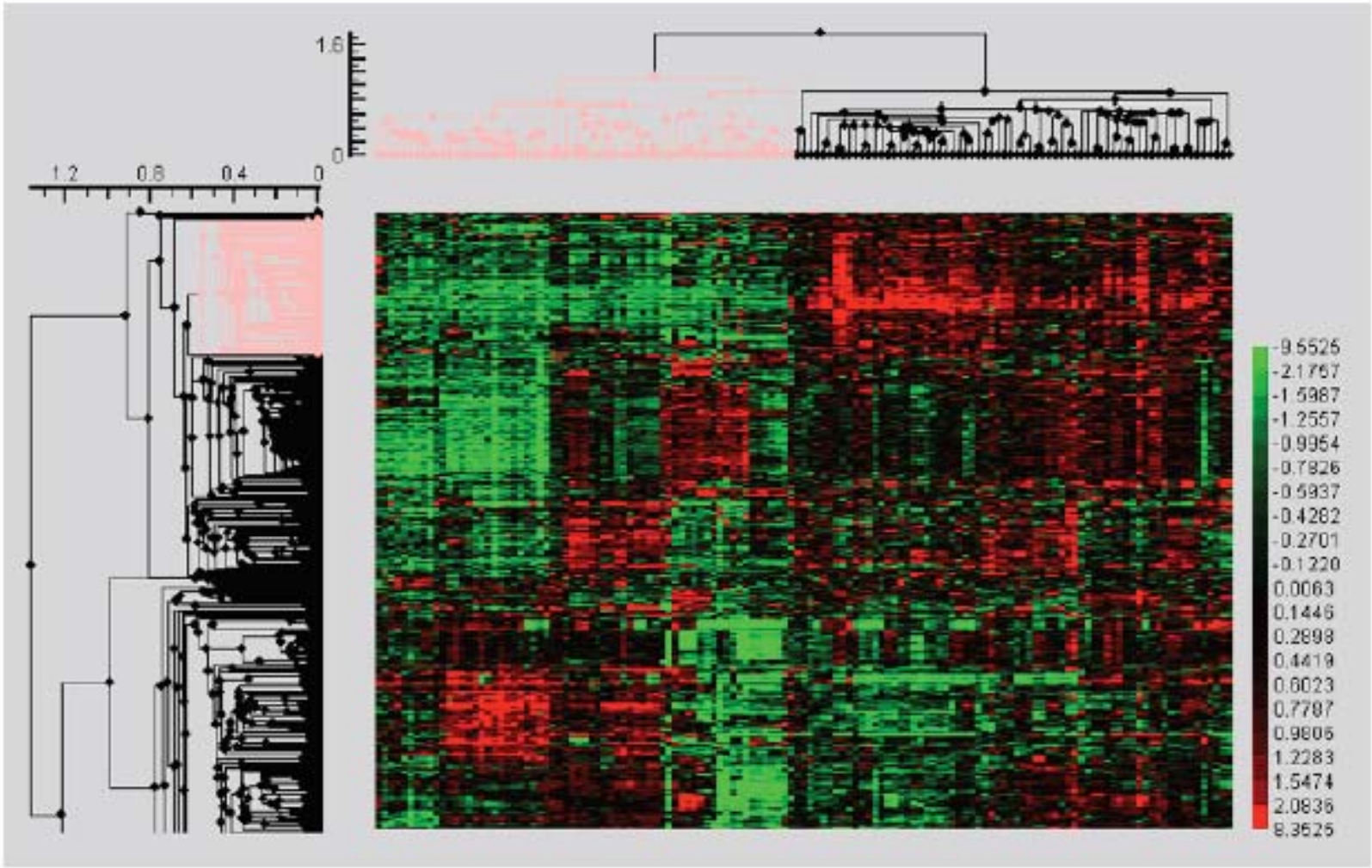


Czy struktura drzewa skupień jest czytelna?

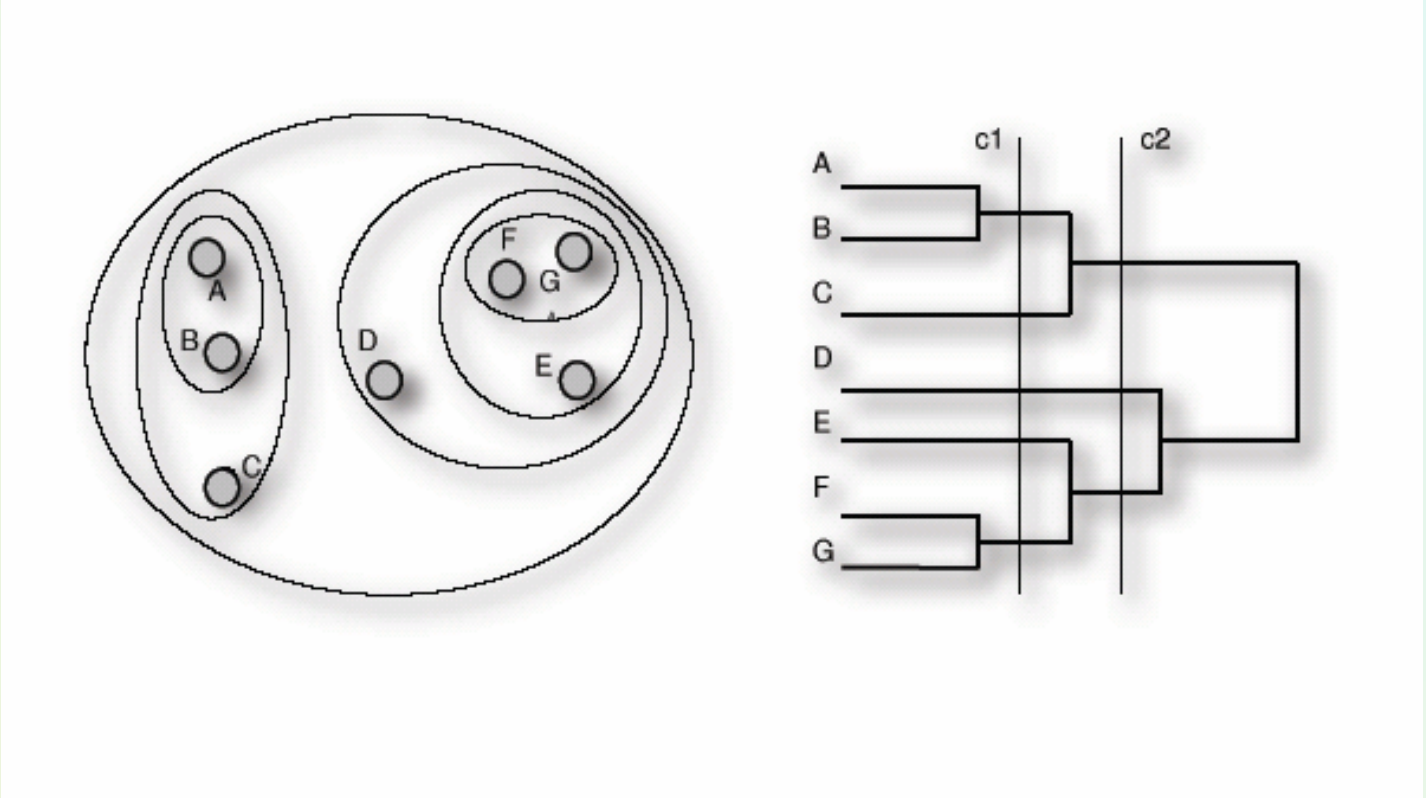
- Przykład biochemiczny



Hierarchical Clustering in Biology

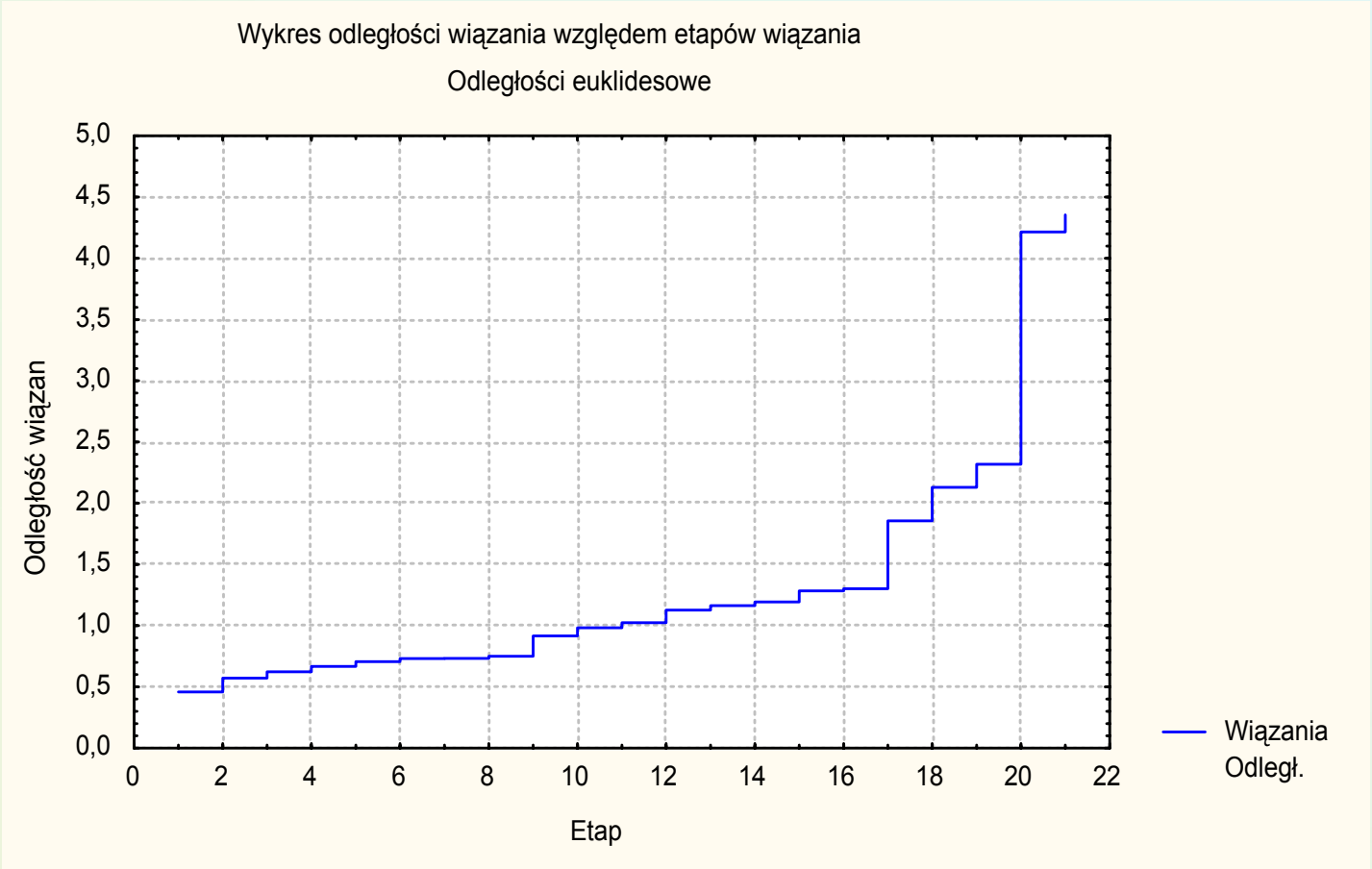


Jak wybrać liczbę skupień?



AHC – jak odnaleźć liczbę skupień?

- Find a cut point („kolanko” wykresu/ knee)



Analiza skupień w Statsoft -Statistica

Analiza Skupień – Statistica; więcej na www.statsoft.com. Przykład analizy danych o parametrach samochodów

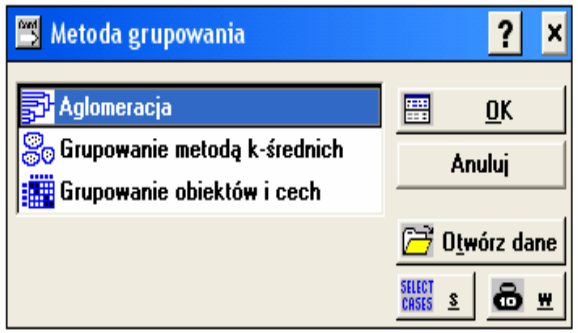
STATISTICA: Analiza skupień - [Dane: CARS.STA 5v * 22c]

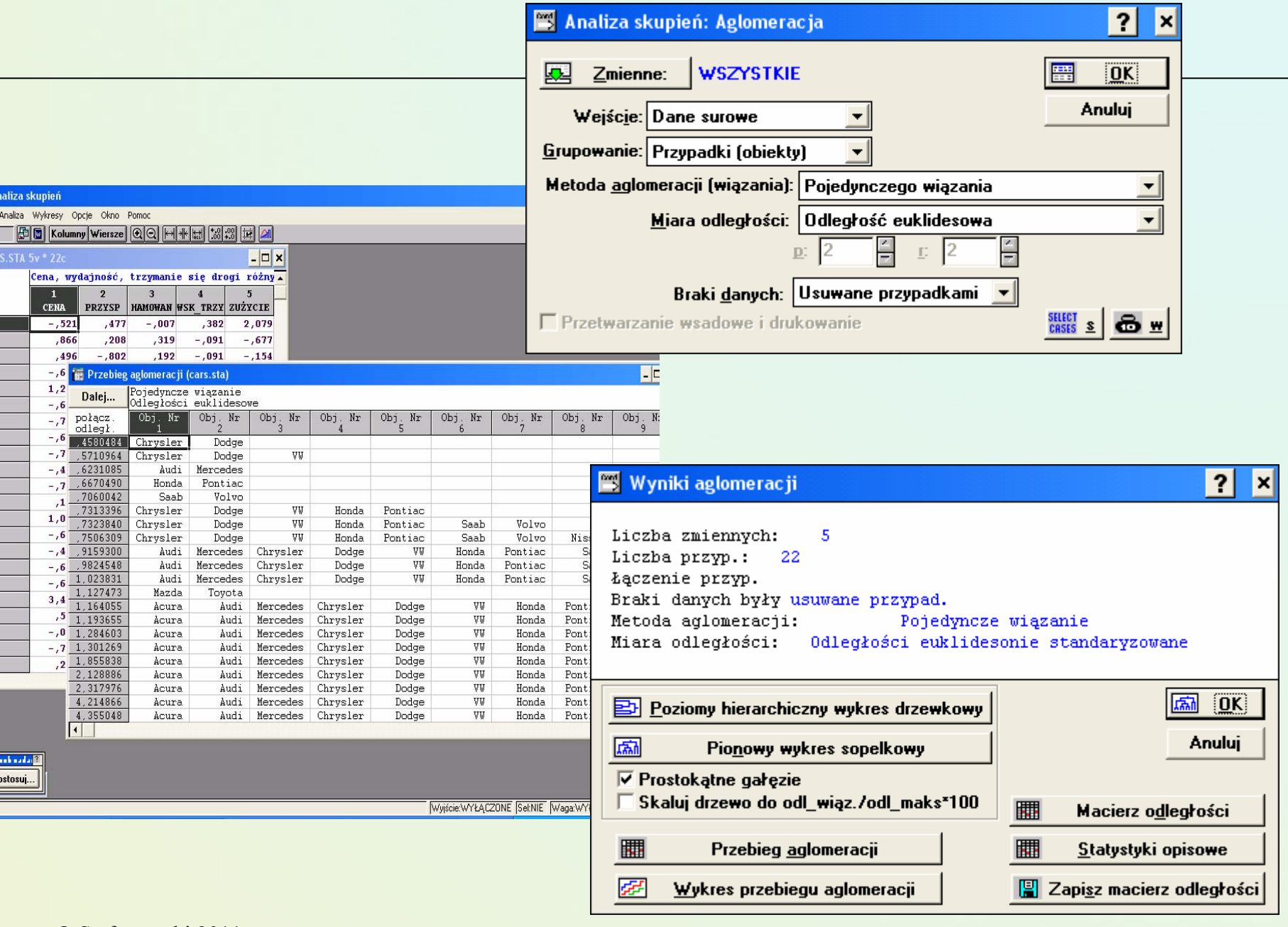
Plik Edycja Widok Analiza Wykresy Opcje Okno Pomoc

-,521072362755425 Zmienne Przypadki

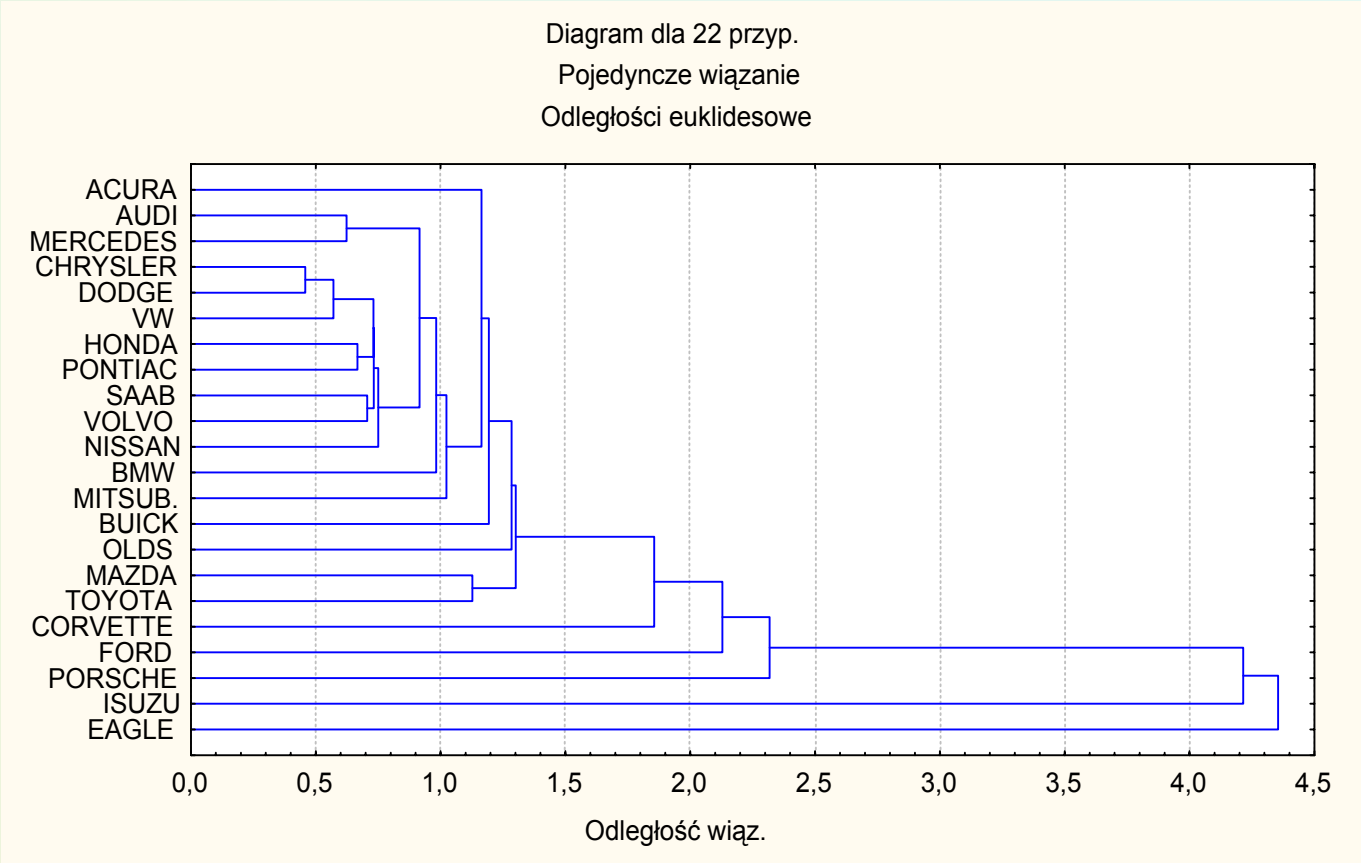
Cena, wydajność, trzymanie się drogi różny

LICZBOWE WARTOŚCI	1 CENA	2 PRZYSP	3 HAMOWAN	4 WSK_TRZY	5 ZUŻYCIE
Acura	-,521	,477	-,007	,382	2,079
Audi	,866	,208	,319	-,091	-,677
BMW	,496	-,802	,192	-,091	-,154
Buick	-,614	1,689	,933	-,210	-,154
Corvette	1,235	-1,811	-,494	,973	-,677
Chrysler	-,614	,073	,427	-,210	-,154
Dodge	-,706	-,196	,481	,145	-,154
Eagle	-,614	1,218	-4,199	-,210	-,677
Ford	-,706	-1,542	,987	,145	-1,724
Honda	-,429	,410	-,007	,027	,369
Isuzu	-,798	,410	-,061	-4,230	1,067
Mazda	,126	,679	-,133	,500	-1,724
Mercedes	1,051	,006	,120	-,091	-,154
Mitsub.	-,614	-1,003	,084	,382	,718
Nissan	-,429	,073	-,007	,263	,997
Olds	-,614	-,734	,409	,382	2,114
Pontiac	-,614	,679	,536	,145	,195
Porsche	3,454	-2,215	-,296	,618	-1,026
Saab	,588	,679	,246	,263	,021
Toyota	-,059	1,218	,228	,736	-,851
VW	-,706	-,128	,102	,382	,195
Volvo	,219	,612	,138	-,210	,369



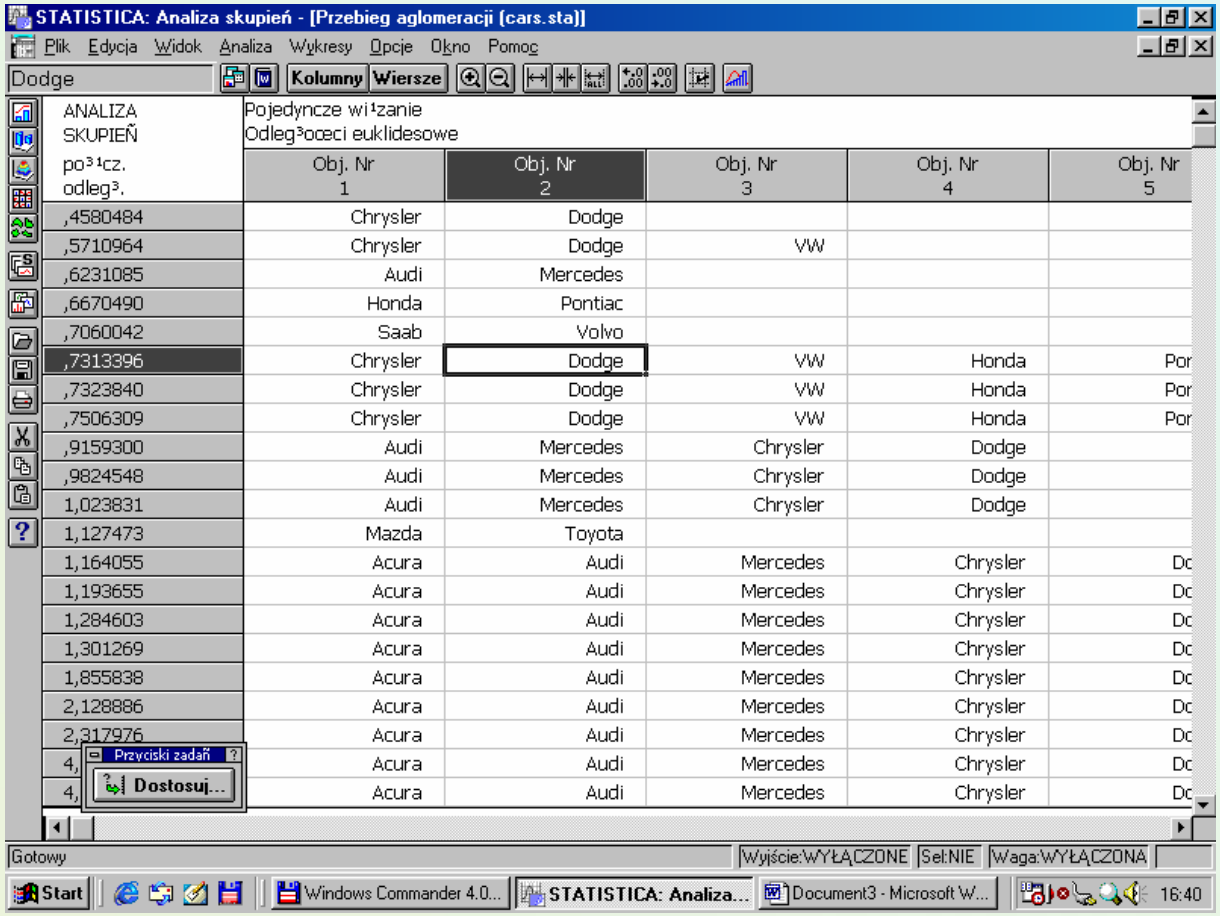

Analiza skupień: Aglomeracja
Zmienne: WSZYSTKIE
OK
Anuluj
Wejście: Dane surowe
Grupowanie: Przypadki (obiekty)
Metoda aglomeracji (wiązania): Pojedynczego wiązania
Miara odległości: Odległość euklidesowa
p: 2
r: 2
Braki danych: Usuwane przypadkami
Przetwarzanie wsadowe i drukowanie
SELECT CASES s
w
aliza skupień
Analiza Wykresy Opcje Okno Pomoc
Kolumny Wiersze
S.STA 5v * 22c
Cena, wydajność, trzymanie się drogi różny
1 CENA
2 PRZYSP
3 HAMOWAN
4 WSK_TRZY
5 ZUŻYCIE
-,521 ,477 -,007 ,382 2,079
,866 ,208 ,319 -,091 -,677
,496 -,802 ,192 -,091 -,154
Przebieg aglomeracji (cars.sta)
Dalej...
Pojedyncze wiązanie
Odległości euklidesowe
połącz. odległ.
Obj. Nr 1
Obj. Nr 2
Obj. Nr 3
Obj. Nr 4
Obj. Nr 5
Obj. Nr 6
Obj. Nr 7
Obj. Nr 8
Obj. Nr 9
,4580484 Chrysler Dodge
,5710964 Chrysler Dodge VW
,6231085 Audi Mercedes
,6670490 Honda Pontiac
,7060042 Saab Volvo
,7313396 Chrysler Dodge VW Honda Pontiac
,7323840 Chrysler Dodge VW Honda Pontiac Saab Volvo
,7506309 Chrysler Dodge VW Honda Pontiac Saab Volvo
,9159300 Audi Mercedes Chrysler Dodge VW Honda Pontiac
,9824548 Audi Mercedes Chrysler Dodge VW Honda Pontiac
1,023831 Audi Mercedes Chrysler Dodge VW Honda Pontiac
1,127473 Mazda Toyota
1,164055 Acura Audi Mercedes Chrysler Dodge VW Honda
1,193655 Acura Audi Mercedes Chrysler Dodge VW Honda
1,284603 Acura Audi Mercedes Chrysler Dodge VW Honda
1,301269 Acura Audi Mercedes Chrysler Dodge VW Honda
1,855838 Acura Audi Mercedes Chrysler Dodge VW Honda
2,128886 Acura Audi Mercedes Chrysler Dodge VW Honda
2,317976 Acura Audi Mercedes Chrysler Dodge VW Honda
4,214866 Acura Audi Mercedes Chrysler Dodge VW Honda
4,355048 Acura Audi Mercedes Chrysler Dodge VW Honda
Wyjście:WYŁĄCZONE
Sel:NIE
Waga:WY
Wyniki aglomeracji
Liczba zmiennych: 5
Liczba przyp.: 22
Łączenie przyp.
Braki danych były usuwane przypad.
Metoda aglomeracji: Pojedyncze wiązanie
Miara odległości: Odległości euklidesonie standaryzowane
Poziomy hierarchiczny wykres drzewkowy
Pionowy wykres sopelkowy
Prostokątne gałęzie
Skaluj drzewo do odl_wiąz./odl_maks*100
OK
Anuluj
Macierz odległości
Statystyki opisowe
Przebieg aglomeracji
Wykres przebiegu aglomeracji
Zapisz macierz odległości

Dendrogram for Single Linkage



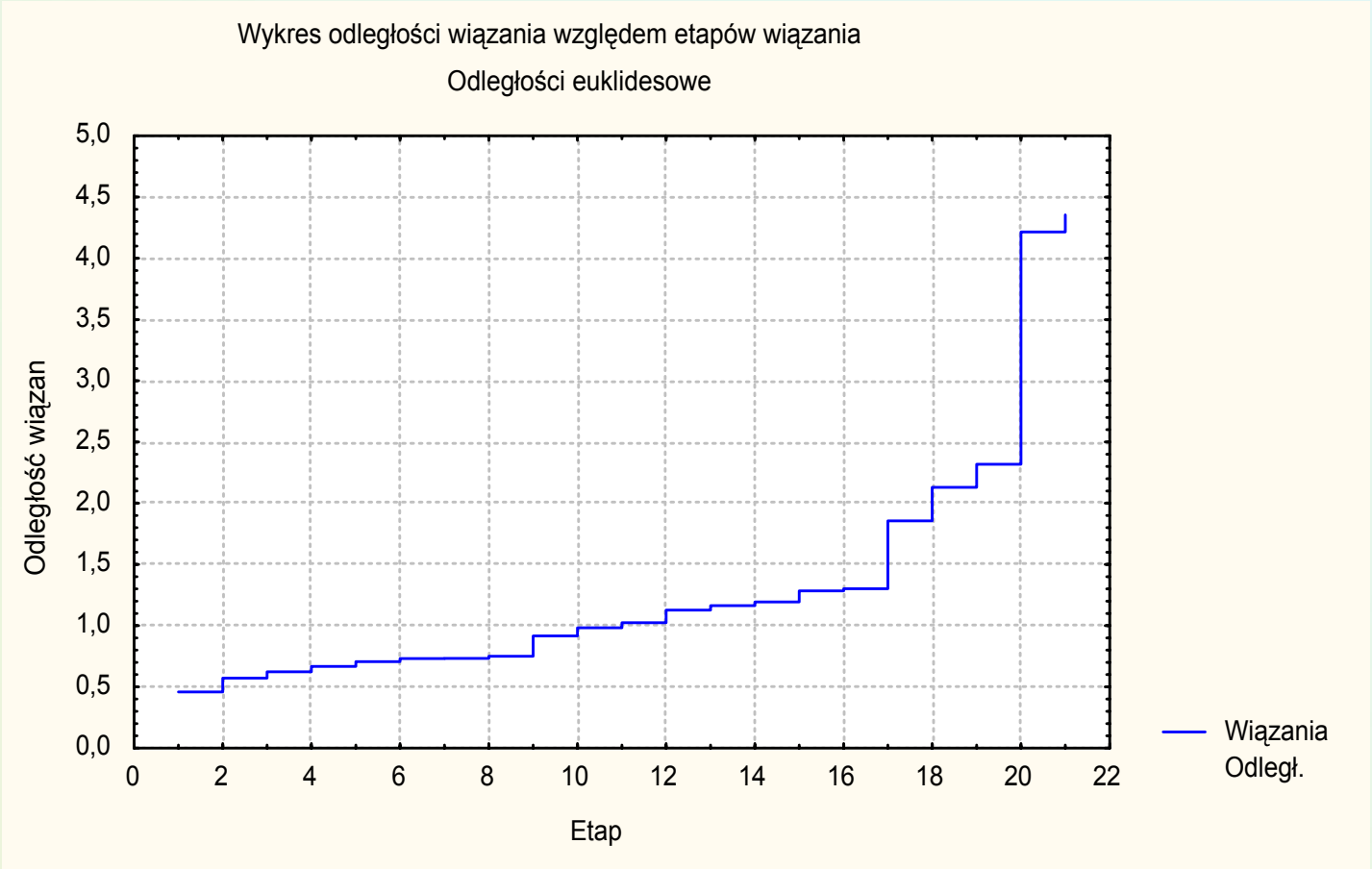
Opis tworzenia dendrogramu

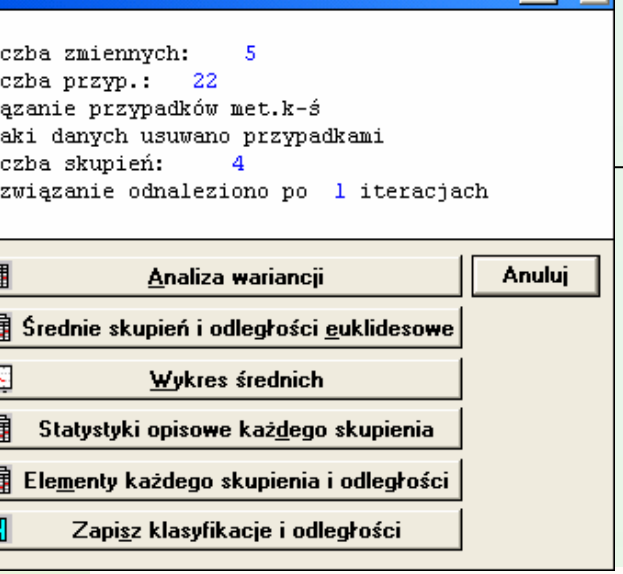
- Łączenie obiektów w kolejnych krokach

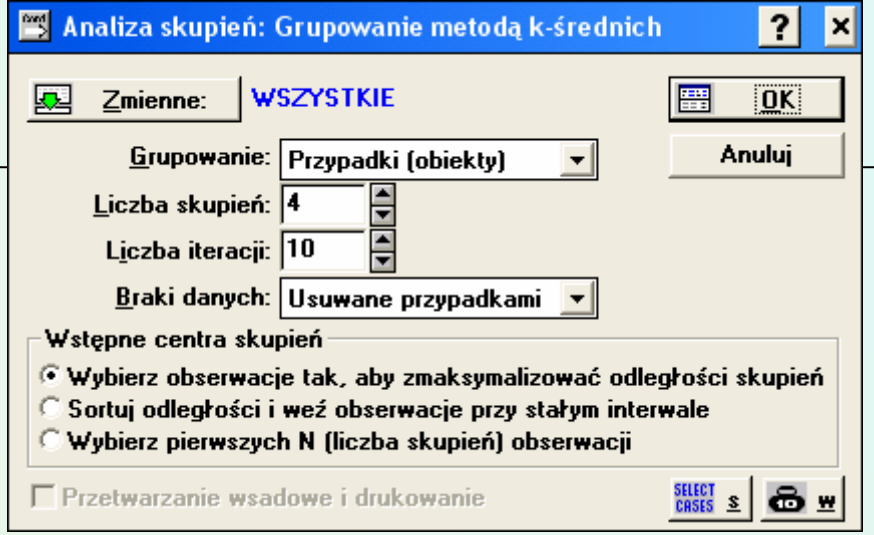


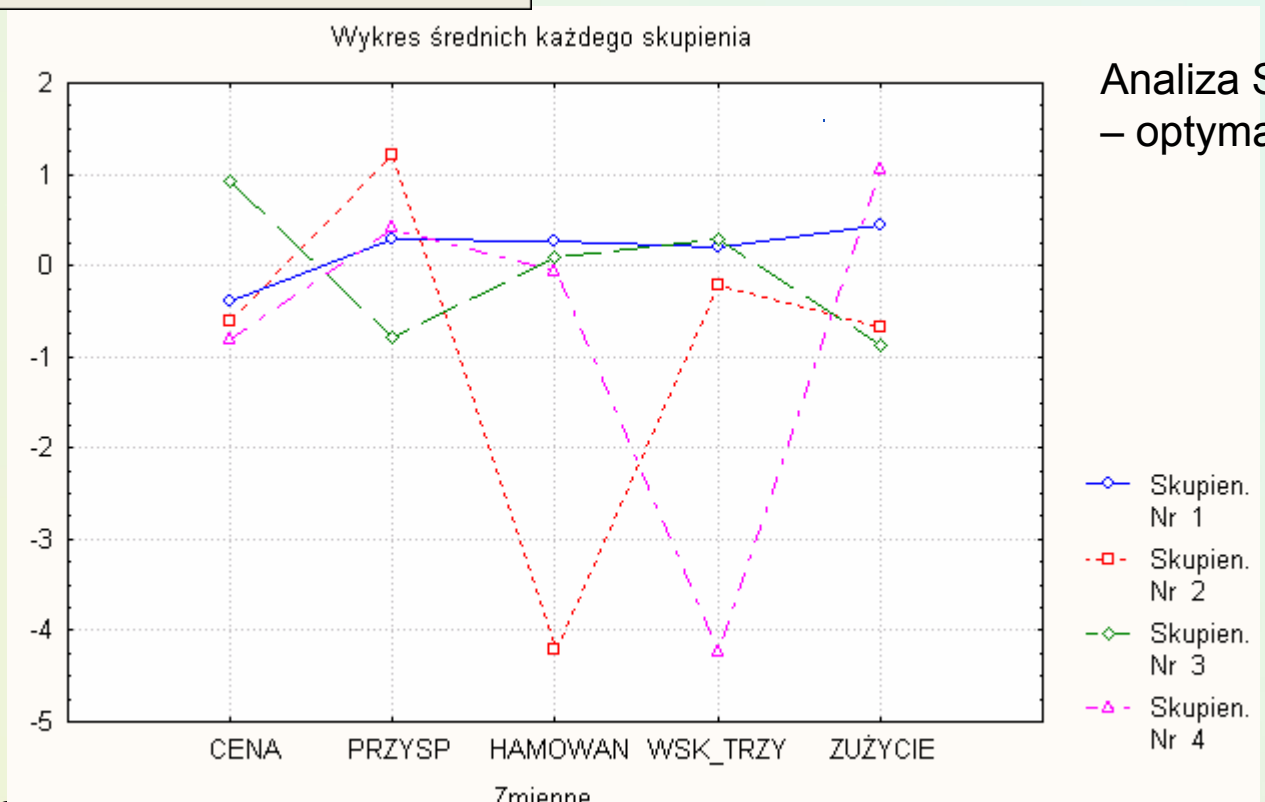
Analiza procesu łączenia

- Wykres kolankowy – a cut point („kolanko” / knee)









Analiza Skupień – optymalizacja k-średnich

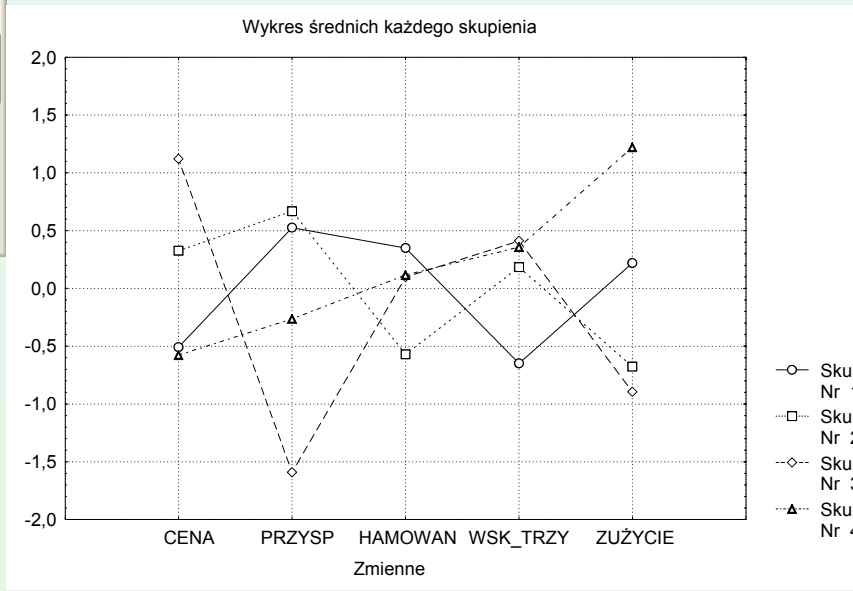
Wsparcie do charakterystyki skupień

Odległości euklidesowe skupień (cars.sta)

Dalej...

Odległości pod przekątną
Kwadr. odległości nad przekątną

Skupien. Numer	Nr 1	Nr 2	Nr 3	Nr 4
Nr 1	0,000000	,612157	1,912964	,539154
Nr 2	,782405	0,000000	1,256528	1,156810
Nr 3	1,383100	1,120950	0,000000	1,824839
Nr 4	,734271	1,075551	1,350866	0,000000



Średnie skup. (cars.sta)

ANALIZA SKUPIEŃ	Skupien. Nr 1	Skupien. Nr 2	Skupien. Nr 3	Skupien. Nr 4
CENA		,326371	1,11989	-,576541
PRZYSP	,525328	,667950	-1,59237	-,263101
HAMOWAN	,349673	-,569764	,09733	,116307
WSK_TRZY	-,648829	,184539	,41118	,357969
ZUŻYCIE	,219949	-,677062	-,89508	1,220614

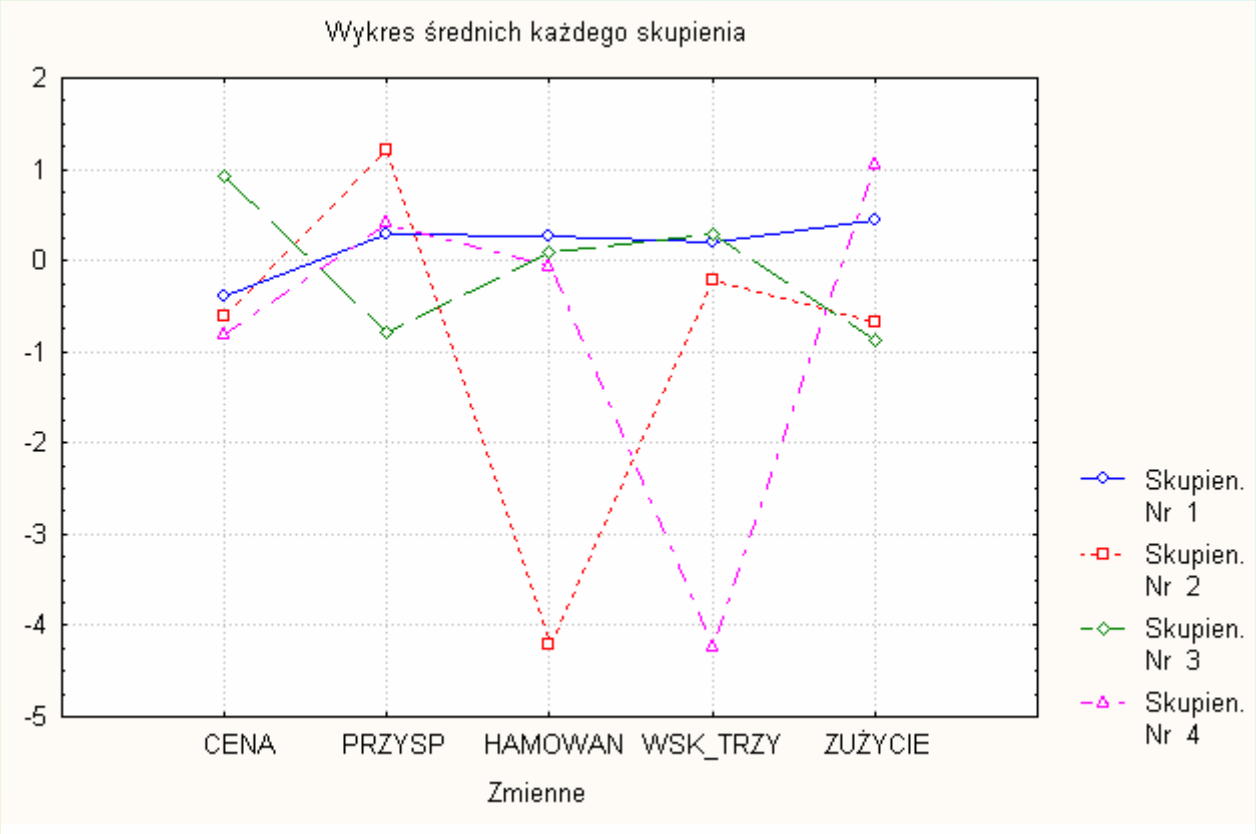
Elementy skupienia numer 2 (cars.sta)

ANALIZA SKUPIEŃ

i odległości od środka właściwego skupienia
W skupieniu jest 6 przyp

	Audi	Eagle	Mazda	Mercedes	Saab	Toyota
Odległ.	,523036	1,703612	,533962	,598004	,495550	,533369

Wizualizacja centroidów



Analiza skupień w WEKA

WEKA zakładka Clustering

- Stopniowy przyrost implementacji
  - *k*-Means
  - EM
  - Cobweb
  - X-means
  - FarthestFirst…
  - DbScann
  - Oraz nowe
- Możliwości prostej wizualizacji i ew. porównania przydziałów do „wzorcowej klasyfikacji” – jeśli jest dostepna w pliku arff

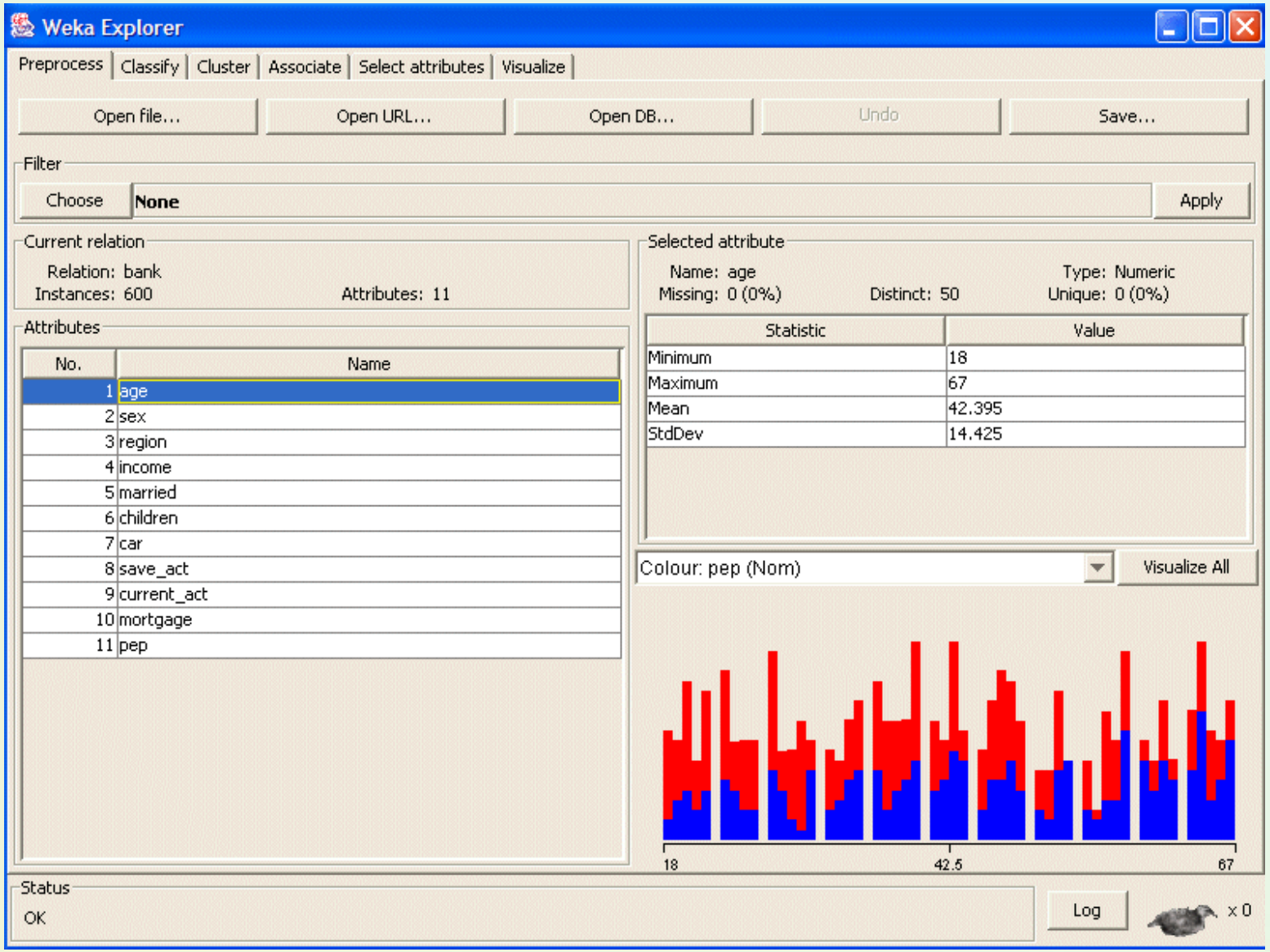
Exercise 1. K-means clustering in WEKA

- The exercise illustrates the use of the k-means algorithm.
- The example – sample of customers of the bank
  - Bank data (bank-data.cvs -> bank.arff)
  - All preprocessing has been performed on cvs
  - 600 instances described by 11 attributes

```
id,age,sex,region,income,married,children,car,save_act,current_act,mortgage,pep
ID12101,48,FEMALE,INNER_CITY,17546.0,NO,1,NO,NO,NO,NO,YES
ID12102,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO
ID12103,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO
ID12104,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO
ID12105,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO
……..
…….
```

  - Cluster customers and characterize the resulting customer segments

Loading the file and analysing the data

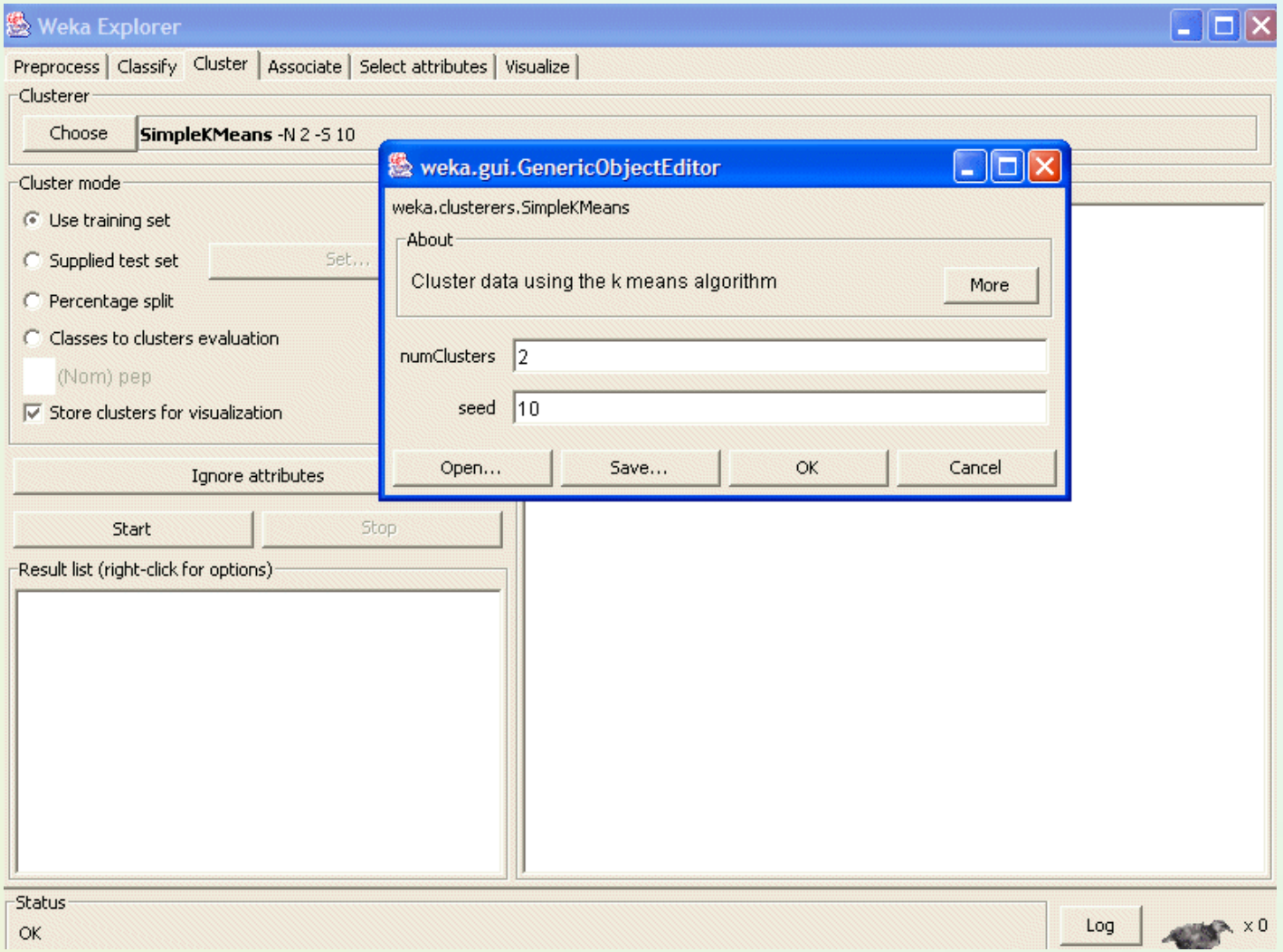


Preprocessing for clustering

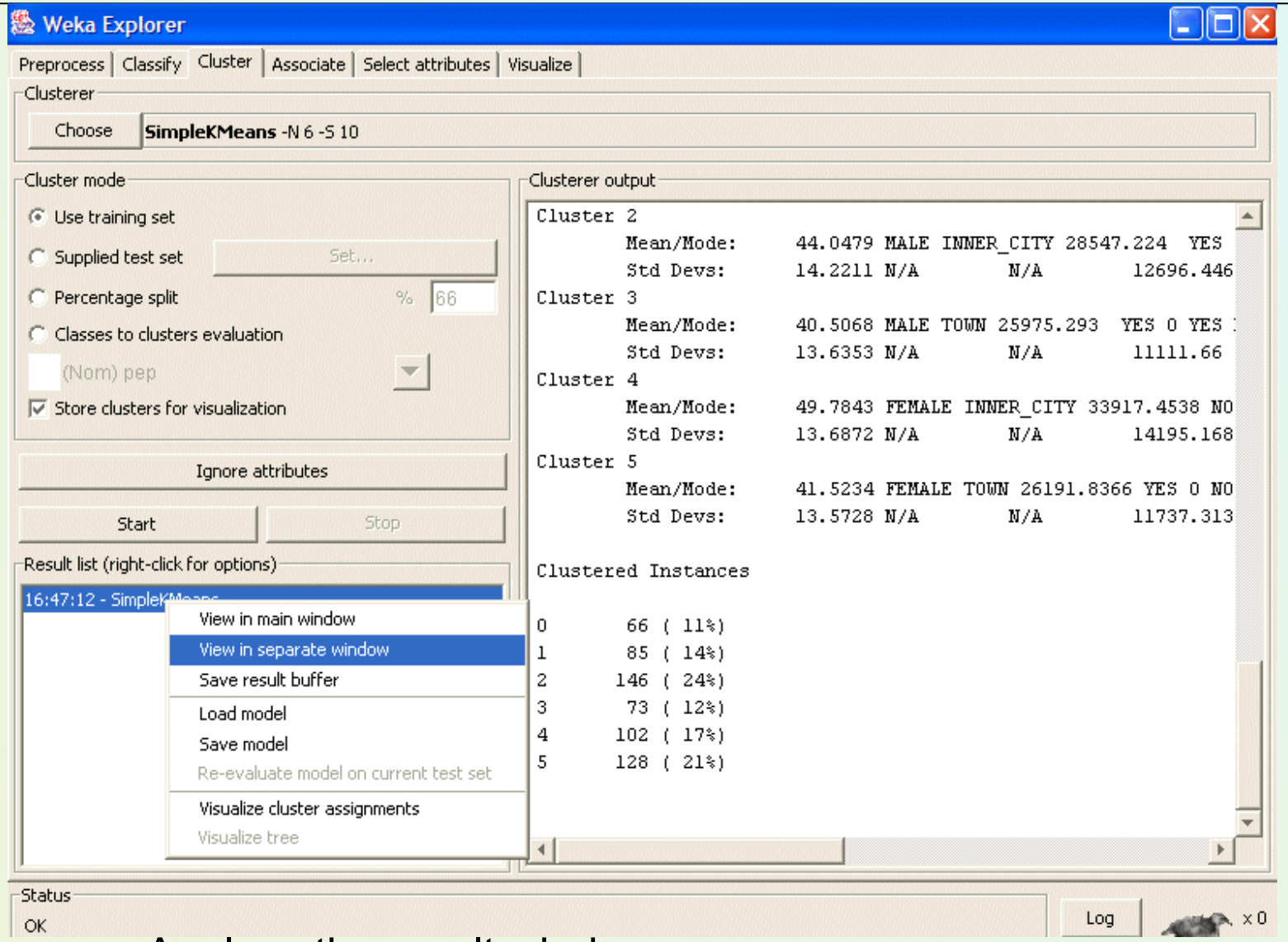
- What about non-numerical attributes?
  - Remember about Filters
- Should we normalize or standarize attributes?
- How it is handled in WEKA k-means?

Choosing Simple k-means

- Tune proper parameters



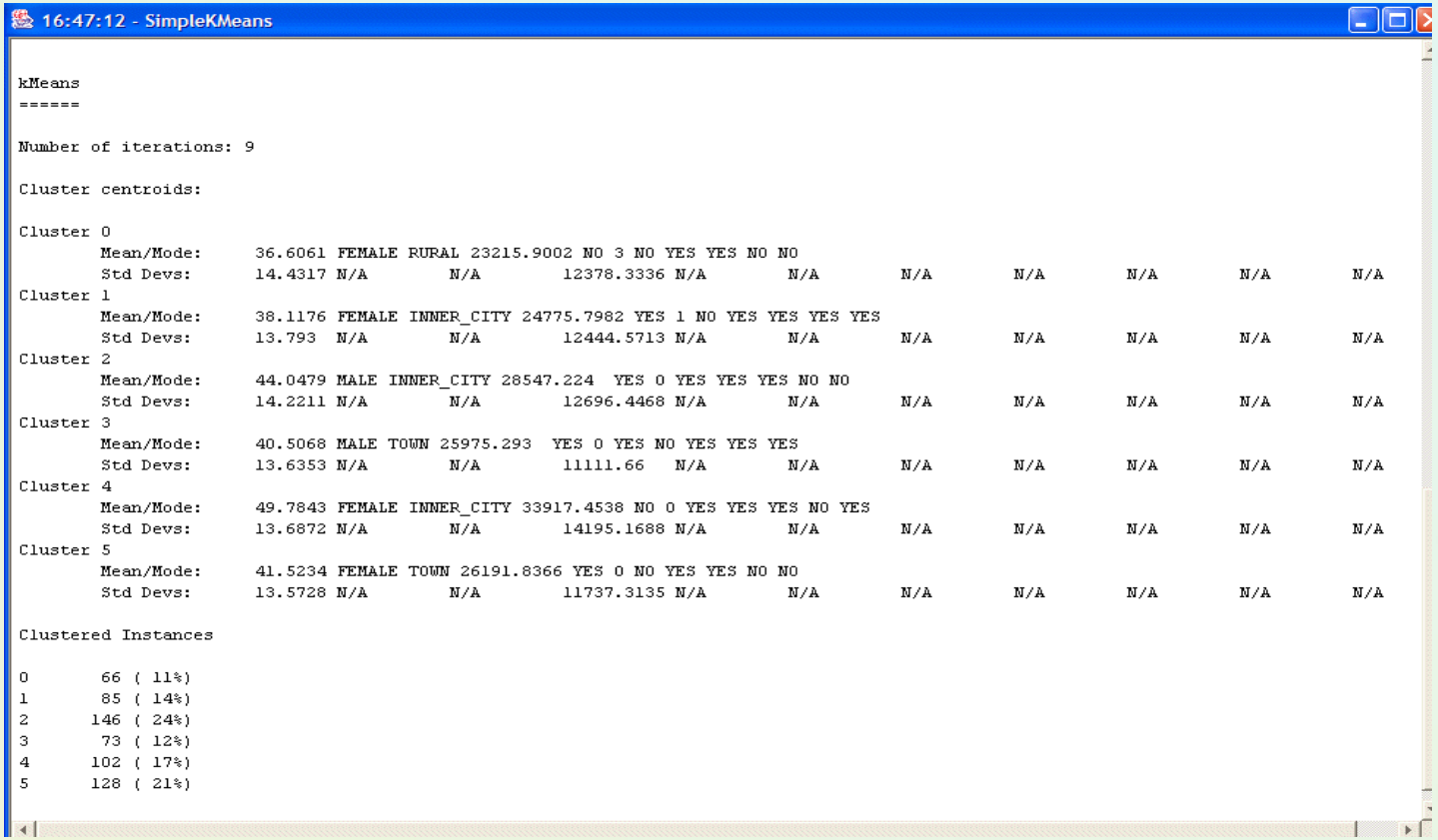
Clustering results



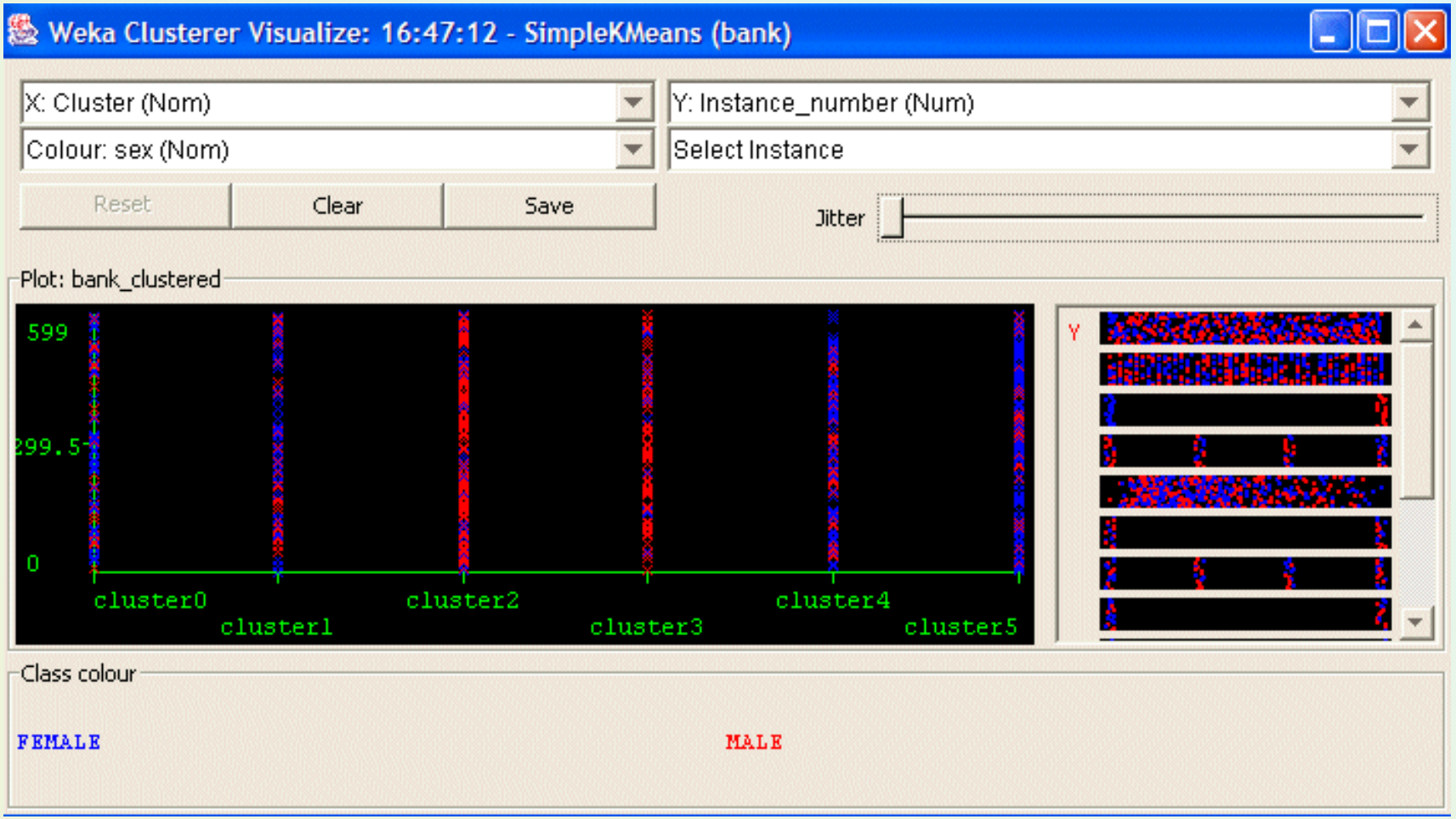
- Analyse the result window

Characterizing cluster

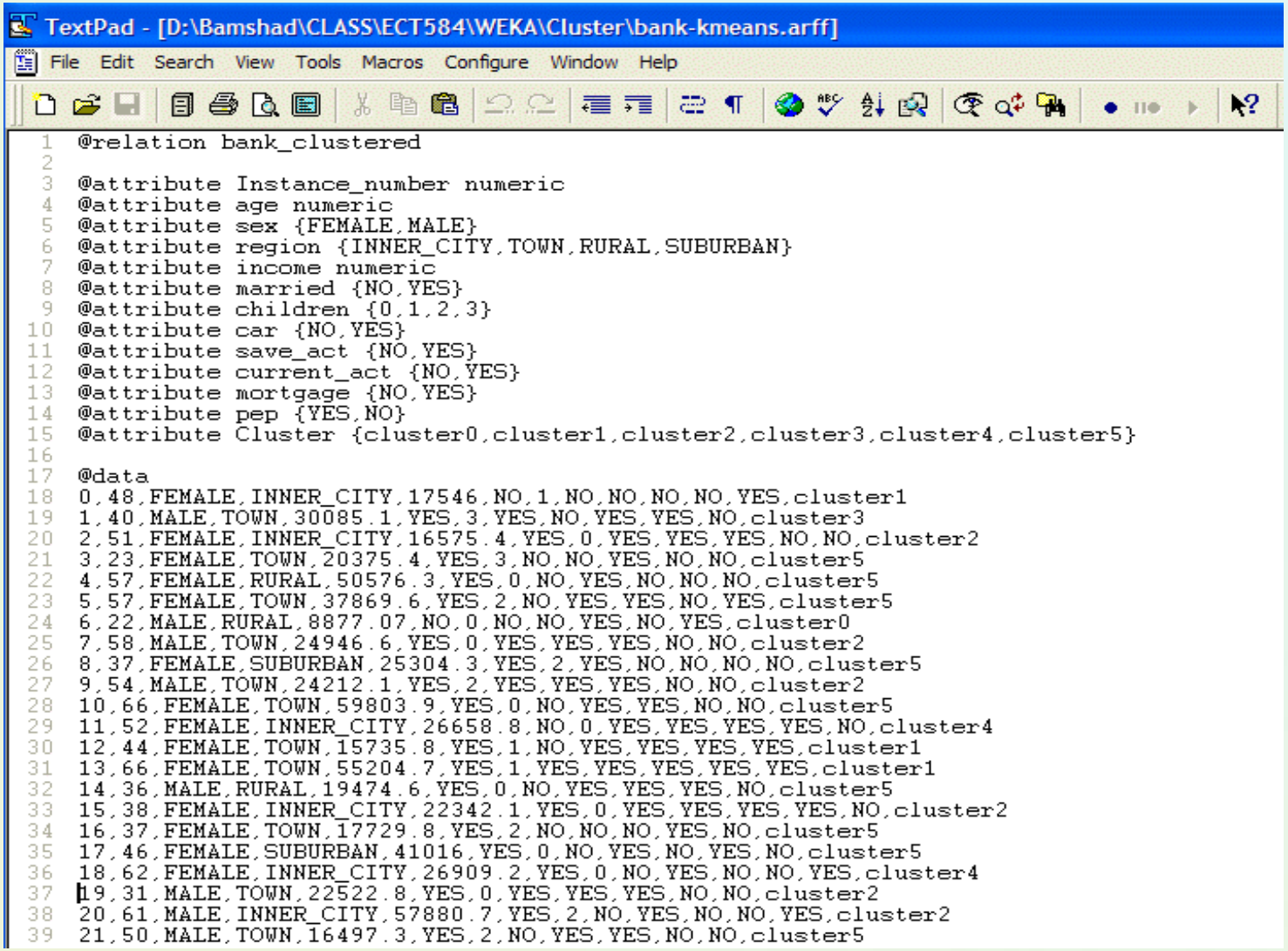
- How to describe clusters?
- What about descriptive statistics for centroids?



Understanding the cluster characterization through visualization



Finally, cluster assignments


TextPad - [D:\Bamshad\CLASS\ECT584\WEKA\Cluster\bank-kmeans.arff]

File Edit Search View Tools Macros Configure Window Help

```
@relation bank_clustered

@attribute Instance_number numeric
@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}
@attribute Cluster {cluster0,cluster1,cluster2,cluster3,cluster4,cluster5}

@data
0,48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES,cluster1
1,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO,cluster3
2,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO,cluster2
3,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO,cluster5
4,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO,cluster5
5,57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES,cluster5
6,22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES,cluster0
7,58,MALE,TOWN,24946.6,YES,0,YES,YES,YES,NO,NO,cluster2
8,37,FEMALE,SUBURBAN,25304.3,YES,2,YES,NO,NO,NO,NO,cluster5
9,54,MALE,TOWN,24212.1,YES,2,YES,YES,YES,NO,NO,cluster2
10,66,FEMALE,TOWN,59803.9,YES,0,NO,YES,YES,NO,NO,cluster5
11,52,FEMALE,INNER_CITY,26658.8,NO,0,YES,YES,YES,YES,NO,cluster4
12,44,FEMALE,TOWN,15735.8,YES,1,NO,YES,YES,YES,YES,cluster1
13,66,FEMALE,TOWN,55204.7,YES,1,YES,YES,YES,YES,YES,cluster1
14,36,MALE,RURAL,19474.6,YES,0,NO,YES,YES,YES,NO,cluster5
15,38,FEMALE,INNER_CITY,22342.1,YES,0,YES,YES,YES,YES,NO,cluster2
16,37,FEMALE,TOWN,17729.8,YES,2,NO,NO,NO,YES,NO,cluster5
17,46,FEMALE,SUBURBAN,41016,YES,0,NO,YES,NO,YES,NO,cluster5
18,62,FEMALE,INNER_CITY,26909.2,YES,0,NO,YES,NO,NO,YES,cluster4
19,31,MALE,TOWN,22522.8,YES,0,YES,YES,YES,NO,NO,cluster2
20,61,MALE,INNER_CITY,57880.7,YES,2,NO,YES,NO,NO,YES,cluster2
21,50,MALE,TOWN,16497.3,YES,2,NO,YES,YES,NO,NO,cluster5
```

Ocena jakości skupień

Dwa różne spojrzenia

- Wewnętrzne (ocena tylko charakterystyki skupień)
  - Brak dodatkowych źródeł informacji, np. zbioru odniesienia etykiet
  - Miary oceny oparte na danych (internal measures)
- Zewnętrzne
  - „Benchmarking on existing labels”
  - Porównanie skupień z tzw. ground-truth categories / partitions
- Ocena ekspercka

Czy można poszukiwać pojedynczej miary

- Pewne rady

"The problem of how to judge the quality of a clustering is difficult and there seems to be no universal answer to it."

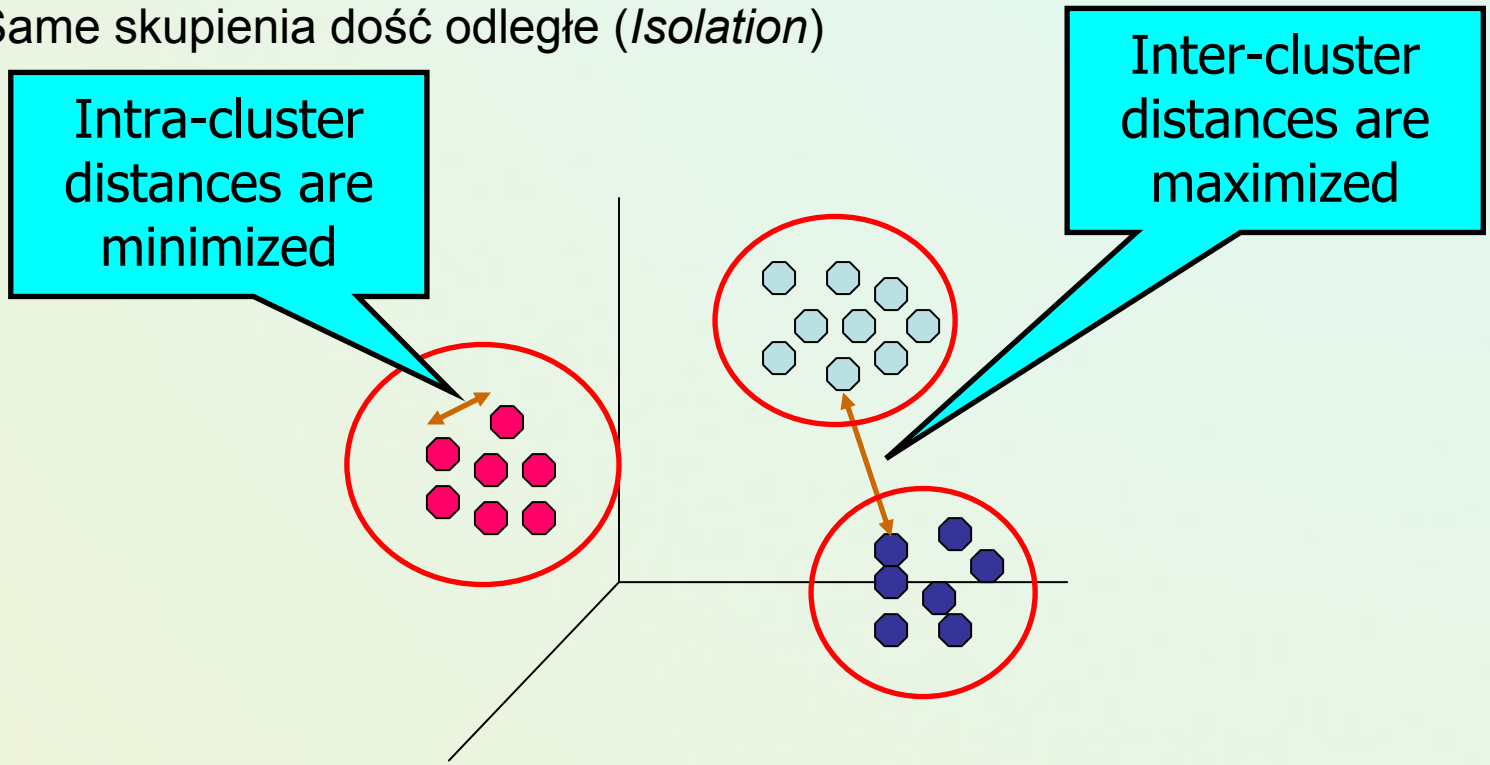
"The nature of processes leading to useful classifications remains little understood, despite considerable effort in this direction."
— R. Michalski, R. Stepp [MS83]

"How do you know the resulting classifications are any *good?"*
— D. Fisher [Fis87]

Ocena jakości skupień

Miary oceny oparte na danych (internal measures)

- Oparte na odległościach lub …
- Duże podobieństwo obiektów wewnątrz skupienia (*Compactness*)
- Same skupienia dość odległe (*Isolation*)



Typowe miary zmienności skupień

- Intuicja $\rightarrow$ „zmienność wewnątrz-skupieniowa” *wc*(C) i „zmienność między-skupieniowa” *bc*(C)
  - Można definiować różnymi sposobami
  - Wykorzystaj średni obiekt w skupieniu $\mathbf{r}_k$ (centroids)
  - Wtedy, np.

$$wc(C) = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \mathbf{r}_k) \qquad \mathbf{r}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

$$bc(C) = \sum_{1 \le j < k \le K} d(\mathbf{r}_j, \mathbf{r}_k)$$

  - Zamiast *bc* odległość od globalnego centrum danych (inter-class distance)

$$id = \sum_{C_j} d(\mathbf{r}_j, \mathbf{r}_{glob})$$

  - Preferencja dla zwartych, jednorodnych skupień dość odległych od centrum danych

Inne kryteria wewnętrznej jakości skupień

- Compactness → determining the weakest connection within the cluster, i.e., the largest distance between two objects *Ri* and *Rk* within the cluster.
- **Isolation** → determining the strongest connection of a cluster to another cluster, i.e., the smallest distance between a cluster centroid and another cluster centroid.

$$\left(\sum_{C_j}\left(\frac{max(D(R_i,R_k)) \text{ where } (R_i,R_k) \in C_j}{min(D(C_{jc},C_{mc})) \text{ where } C_m \neq C_j}\right)\right)^{-1}$$

- Object positioning → the quality of clustering is determined by the extent to which each object Rj has been correctly positioned in given clusters

$$\sum_{R_i}(max(D((R_i,R_k))) - min(D(R_i,R_m)))$$

where $(R_i,R_k) \in C_j$ and $R_m \notin C_j$.

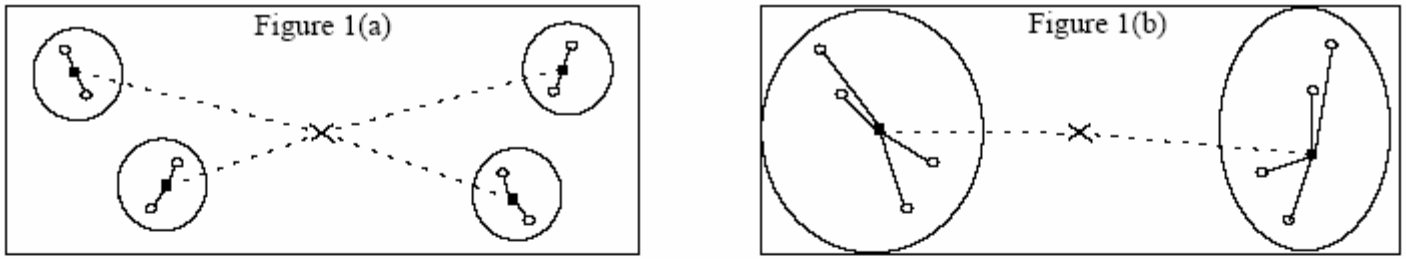


Figure 1: Minimum Total Distance Criterion

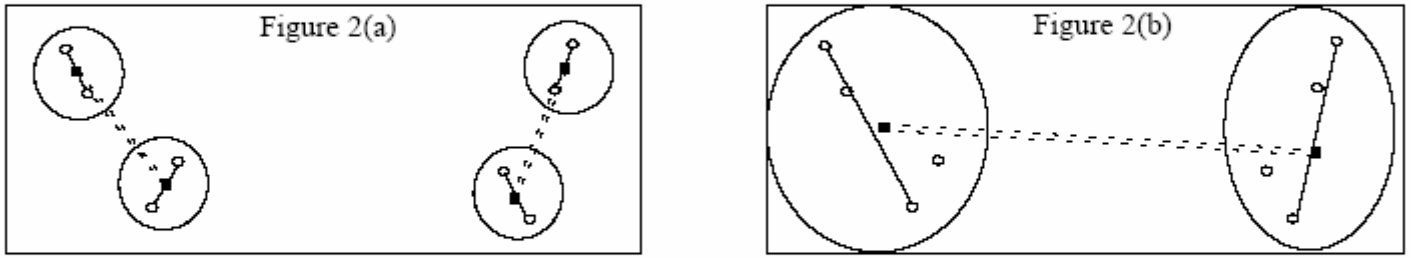


Figure 2: Separated Clusters Criterion

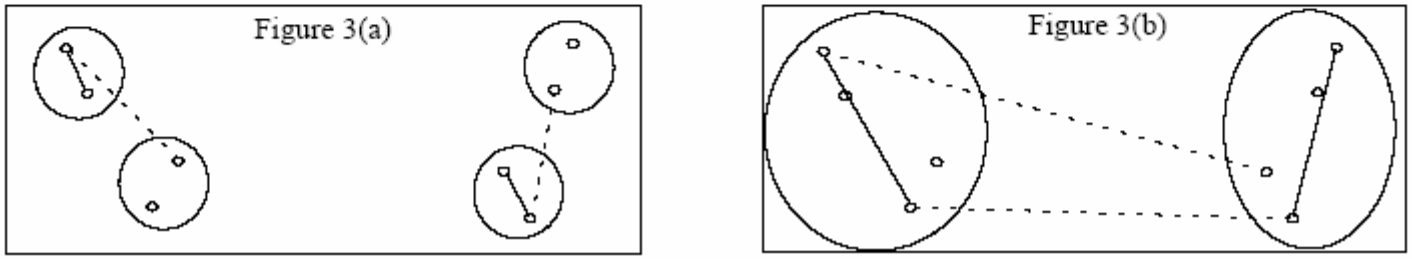


Figure 3: Object Positioning Criterion

Dodatkowe zewnętrzne informacje

- Dysponujemy referencyjnymi etykietami

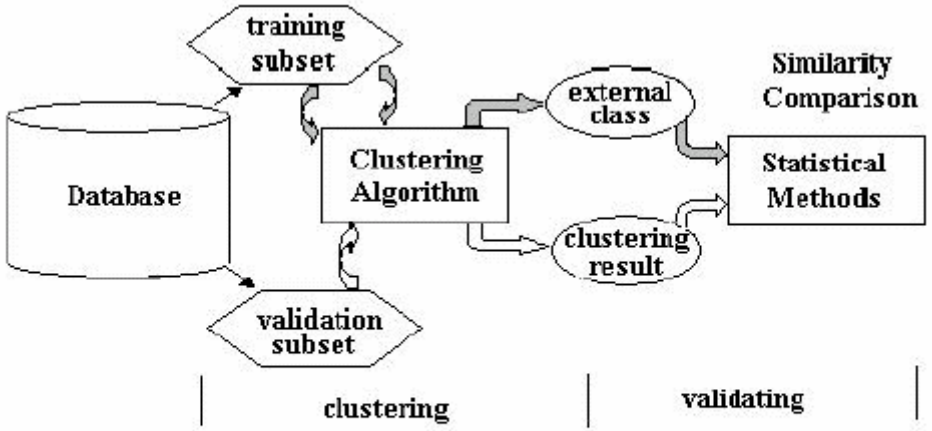


Figure 3-4 External criteria based validation [ZOZ07a]

Odniesienie do zewnętrznego podziału

- Dostępne referencyjne etykiety (manually labeled data)
  - Ekspert etykietuje w zależności od własności danych
  - Istnieją benchmarki TREC, Reuterds, itp..
  - Tryb sem-supervised
- Różne podejścia:
  - „Accuracy of clustering: Percentage of pairs of tuples in the same cluster that share common label” (etykiety w skupieniach_
    - Faworyzujemy małe „czyste” skupienia
    - Czy powinniśmy mieć zgodność liczby skupień i etykiet

Ogólne zasady

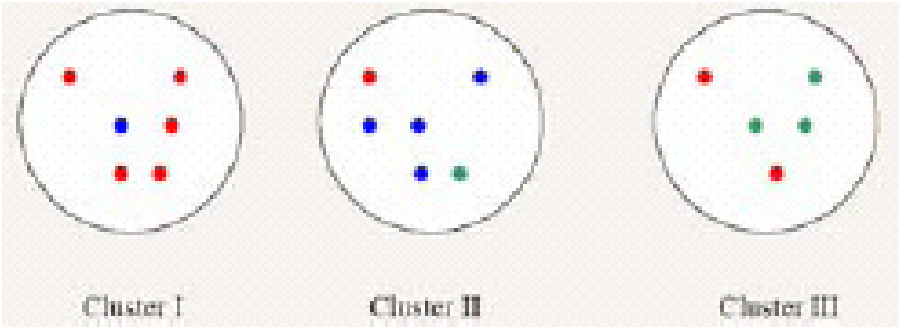
- Homogeneity - Jednorodności
  - Każde skupienie zawiera przykłady z jak najmniejszej liczby etykiet klas
  - Idealnie – tylko jedna klasa
- Completeness
  - Każda klasy reprezentowana w możliwe najmniejszej liczbie skupień
- Typowe miary
  - Purity
  - F-miara

Purity – najprostsza miara

- Simple measure: **purity**, the ratio between the dominant class in the cluster and the size of cluster
  - Assume documents with C gold standard classes, while our clustering algorithms produce K clusters, $\omega_1, \omega_2, \ldots, \omega_K$ with n_i members.

$$Purity(w_i) = \frac{1}{n_i} \max_j(n_{ij}) \quad j \in C$$

  - Example

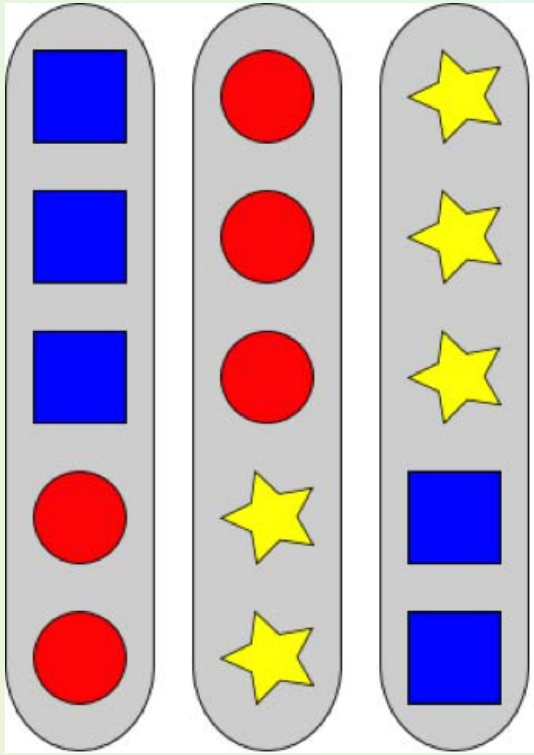


Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

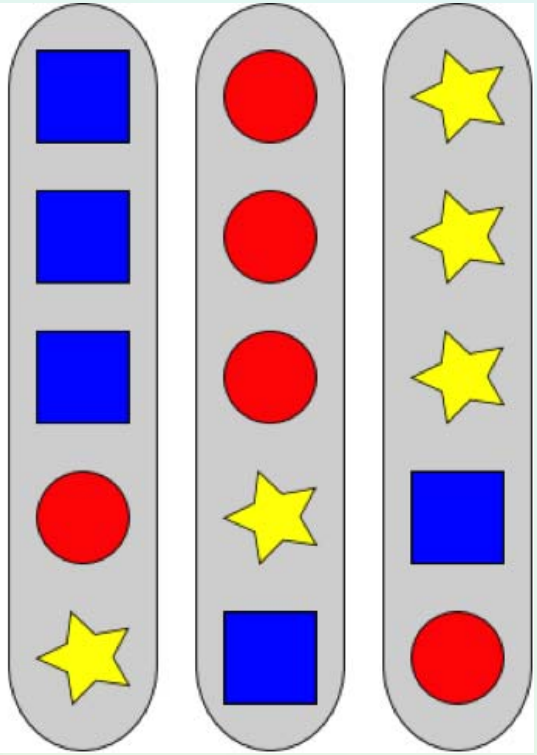
Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

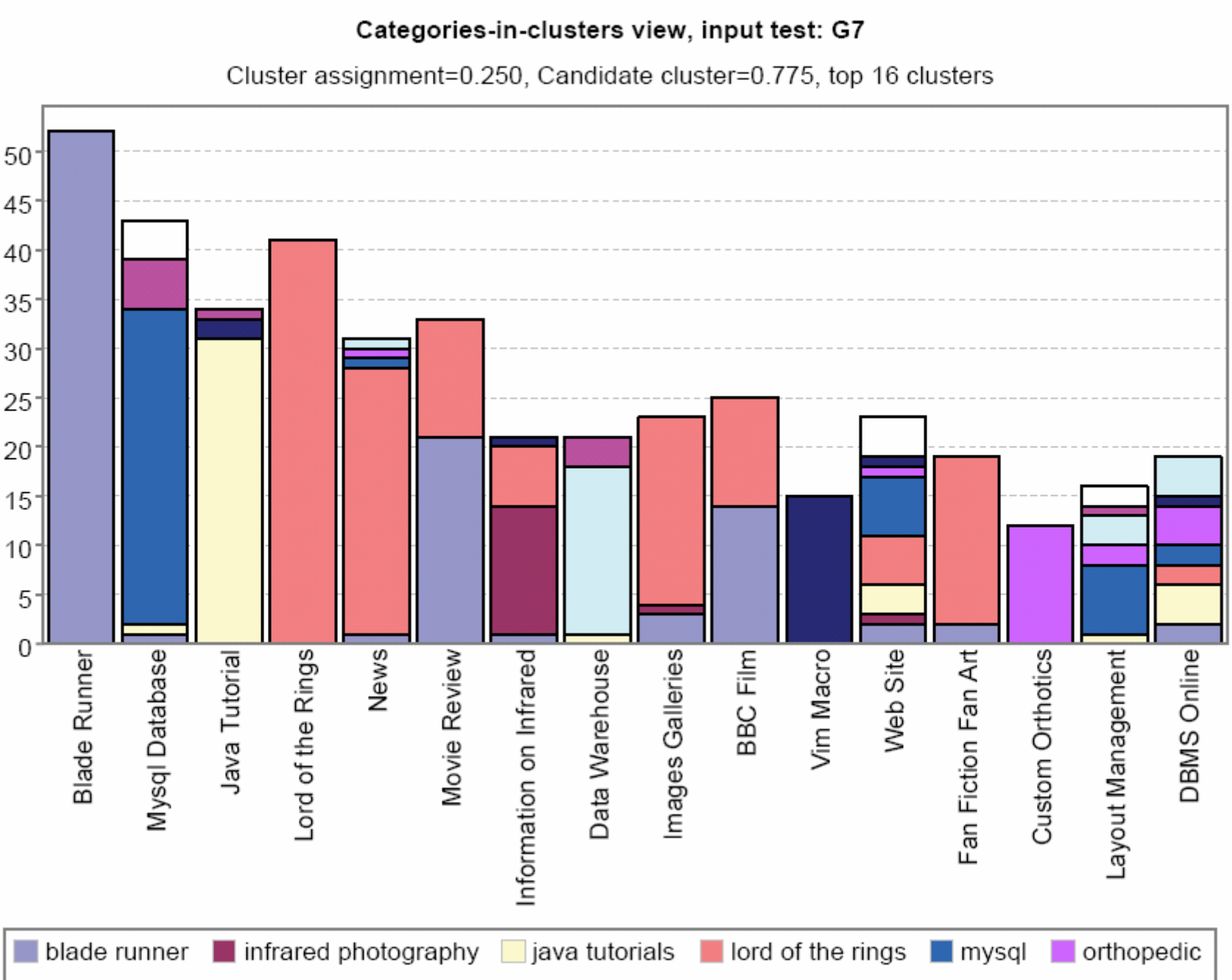
Pewne problemy w ocenie odwzorowań



F-measure: 0.6

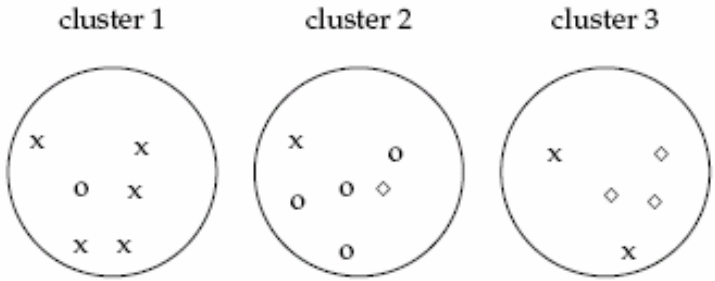


F-measure: 0.6


Categories-in-clusters view, input test: G7
Cluster assignment=0.250, Candidate cluster=0.775, top 16 clusters
50
45
40
35
30
25
20
15
10
5
0
Blade Runner
Mysql Database
Java Tutorial
Lord of the Rings
News
Movie Review
Information on Infrared
Data Warehouse
Images Galleries
BBC Film
Vim Macro
Web Site
Fan Fiction Fan Art
Custom Orthotics
Layout Management
DBMS Online
blade runner
infrared photography
java tutorials
lord of the rings
mysql
orthopedic

Oceny poprzez porównanie do zbioru referencyjnego (ground truth)

- Przykład tekstowy



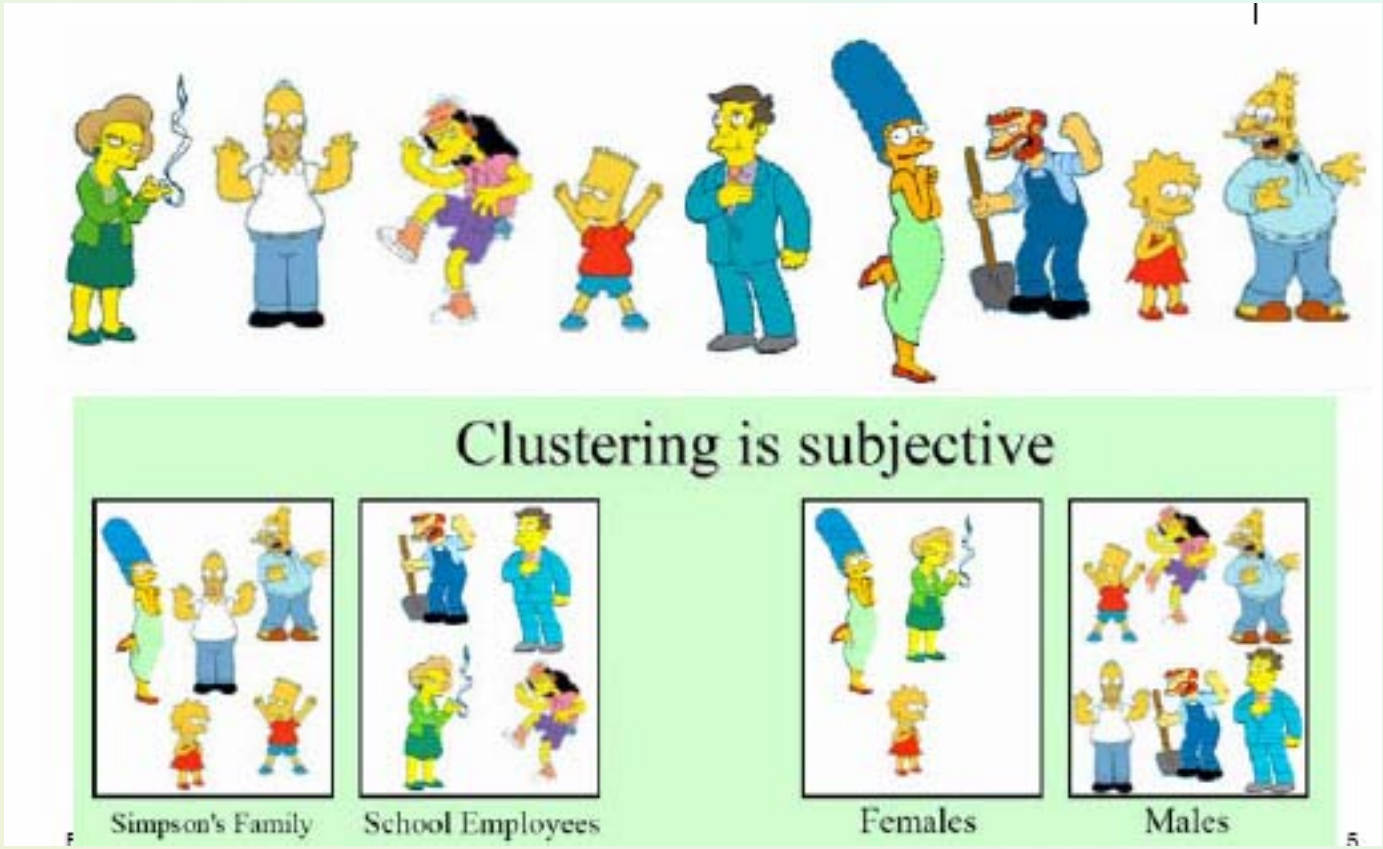
► **Figure 16.4** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ◇, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

	purity	NMI	RI	F_5
minimum	0.0	0.0	0.0	0.0
maximum	1	1	1	1
value for Figure 16.4	0.71	0.36	0.68	0.46

► **Table 16.2** The four external evaluation measures applied to the clustering in Figure 16.4.

Subiektywizm eksperta

- Możliwe różne punkty widzenia



Ocena grupowania

- Inna niż w przypadku uczenia nadzorowanego (predykcji wartości)
- Poprawność grupowania zależna od oceny obserwatora / analityka
- Różne metody AS są skuteczne przy różnych rodzajach skupień i założeniach, co do danych:
  - Co rozumie się przez skupienie, jaki ma kształt, dobór miary odległości $\rightarrow$ sferyczne vs. inne
- Dla pewnych metod i zastosowań:
  - Miary zmienności wewnątrz i między – skupieniowych
  - Idea zbiorów kategorii odniesienia (np.TREC)

Jain – zbiorcze porównanie

Algorithm	Property	Comments
K-means	Identifies hyperspherical clusters; could be modified to find hyper-ellipsoidal clusters using Mahalanobis distance; computationally efficient.	Need to specify K and the initial cluster centers. Additional parameters for creating new clusters, merging existing clusters and outlier detection can be provided.
Fuzzy K-means	Similar to K-means except that every pattern has a degree of membership into the K clusters (fuzzy partition).	Need to specify K, initial cluster centers and cluster membership function.
Minimum Spanning Tree (MST)	Clusters are formed by deleting inconsistent edges in the MST of the data.	Need to provide the definition of an inconsistent edge.
Mutual Neighborhood	Compute the mutual neighborhood value (MNV) for every pair of patterns. If x_j is the p^{th} near neighbor of x_i and x_i is the q^{th} near neighbor of x_j, then $MNV(x_i, x_j) = p + q$; $p, q = 1, \cdots, K$.	Need to specify the neighborhood depth, K.
Single-Link (SL)	A hierarchical clustering algorithm which accepts a $n \times n$ proximity matrix; output is a dendrogram or a tree structure; a single-link cluster is a maximally connected subgraph on the patterns.	Single-link clusters easily chain together and are often "straggly"; need a heuristic to cut the tree to form clusters (a partition).
Complete-Link (CL)	A hierarchical clustering algorithm which accepts a $n \times n$ proximity matrix; output is a dendrogram or a tree structure; a complete-link cluster is a maximally complete subgraph on the patterns.	Complete-link clusters tend to be small and compact which combine nicely into layer clusters even when such a hierarchy is not warranted; need a heuristic to form clusters (a partition).

Może pytanie lub komentarze?