Analiza Skupień - Grupowanie
Zaawansowana Eksploracja Danych

JERZY STEFANOWSKI
Inst. Informatyki PP
Wersja dla TPD 2013

Część II

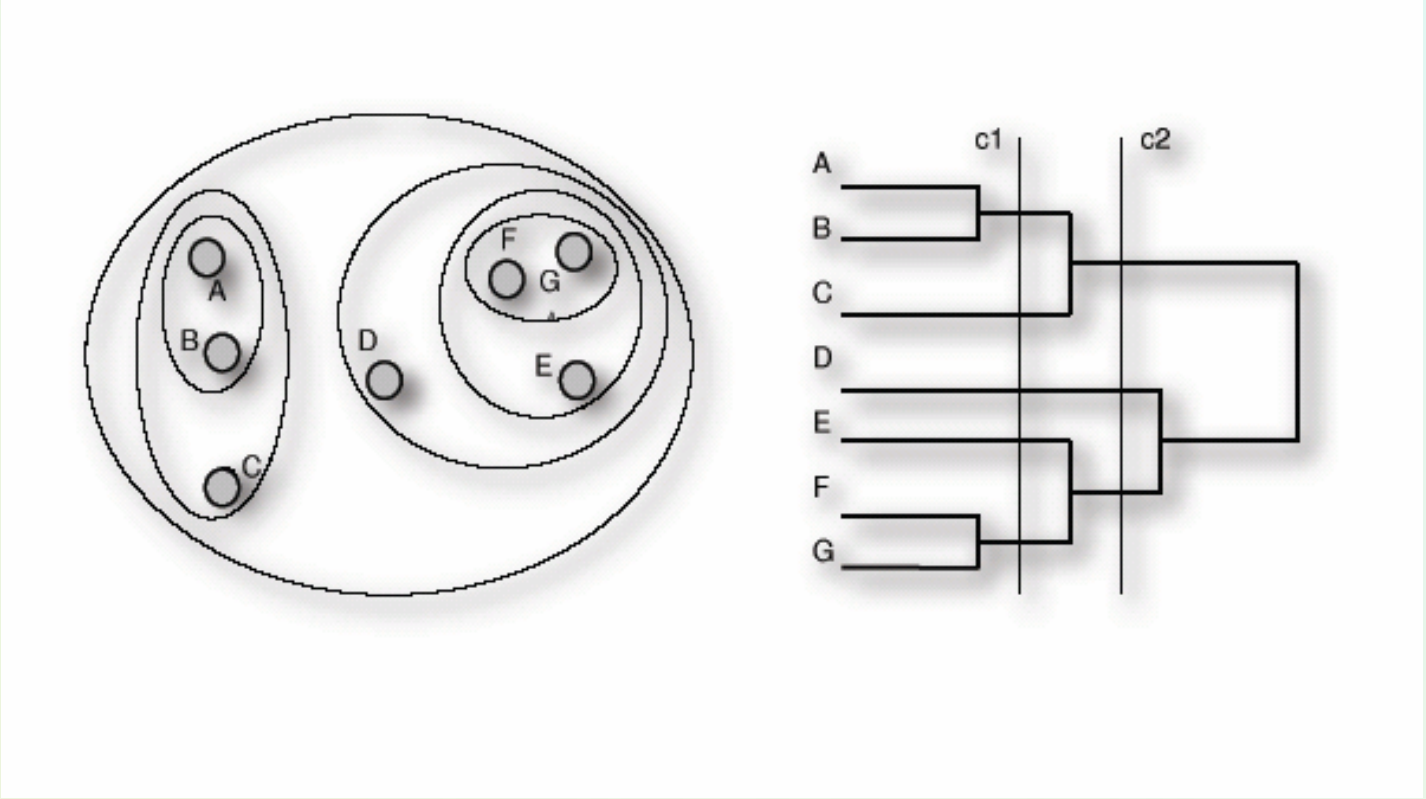
Organizacja wykładu

- Przypomnienie wyboru liczby skupień
- Studium przypadku użycia
- Rozszerzenia dla analizy danych o większych rozmiarach.
- Podsumowanie

Referencje do literatury (przykładowe)

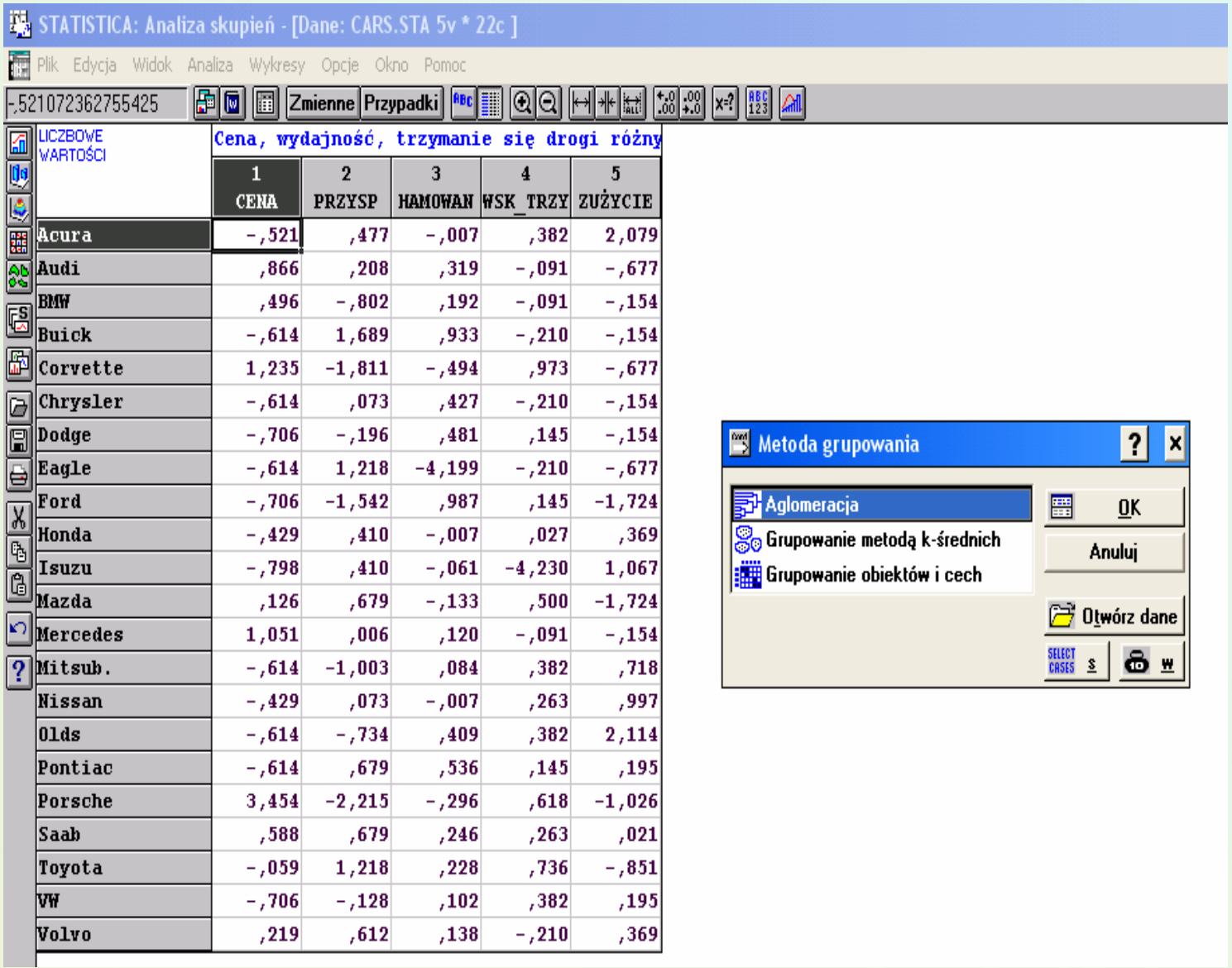
- Koronacki J. Statystyczne systemy uczące się, WNT 2005.
- Pociecha J., Podolec B., Sokołowski A., Zając K. „Metody taksonomiczne w badaniach społeczno-ekonomicznych”. PWN, Warszawa 1988,
- Stąpor K. „Automatyczna klasyfikacja obiektów” Akademicka Oficyna Wydawnicza EXIT, Warszawa 2005.
- Hand, Mannila, Smyth, „Eksploracja danych”, WNT 2005.
- Larose D: „Odkrywania wiedzy z danych”, PWN 2006.
- T.Morzy: „Eksploracja danych”, PWN 2013.
- Kucharczyk J. „Algorytmy analizy skupień w języku ALGOL 60” PWN Warszawa, 1982,
- Materiały szkoleniowe firmy Statsoft.
- Inne polecane przez …. wykładowcę

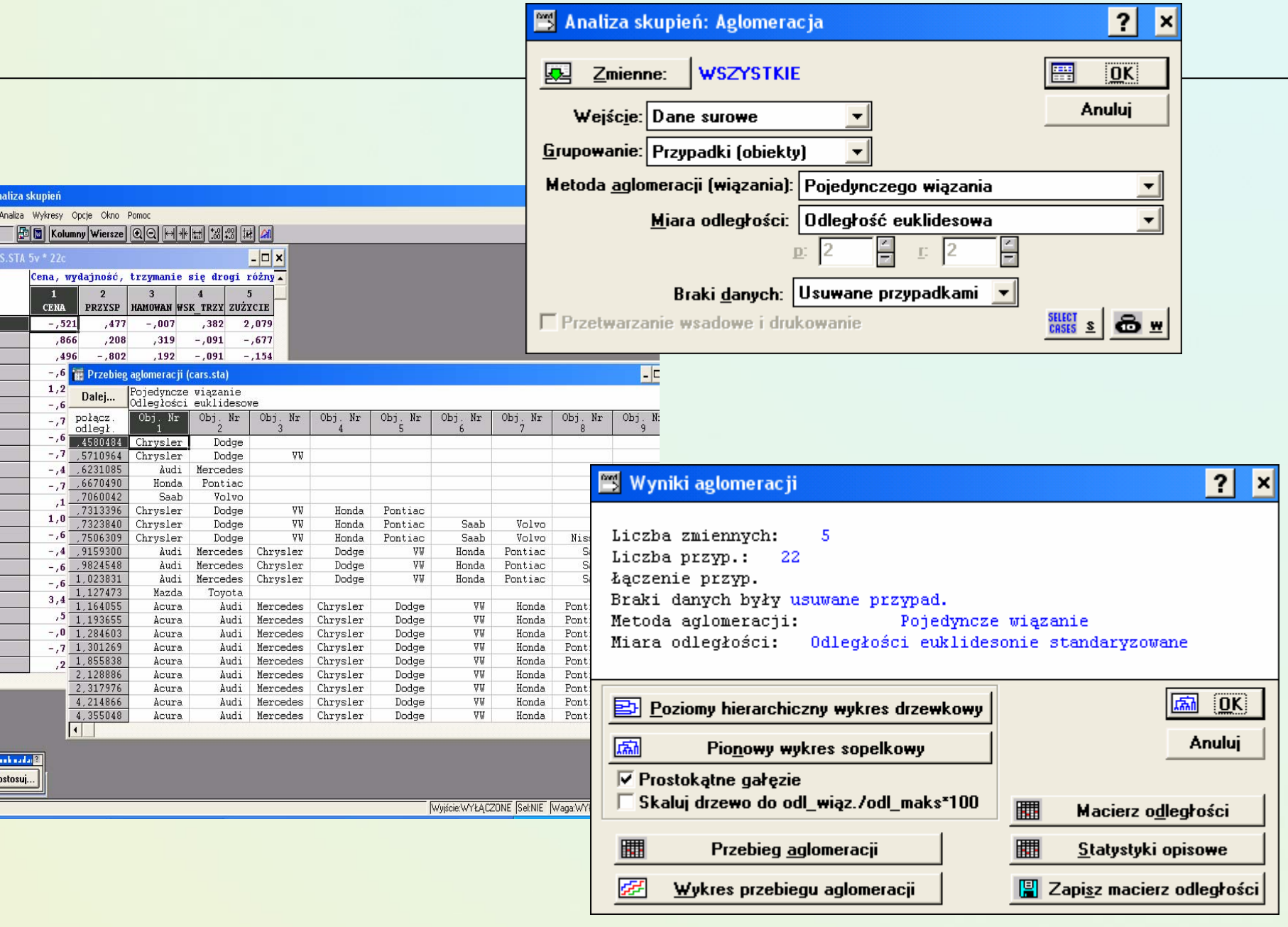
Jak wybrać liczbę skupień?



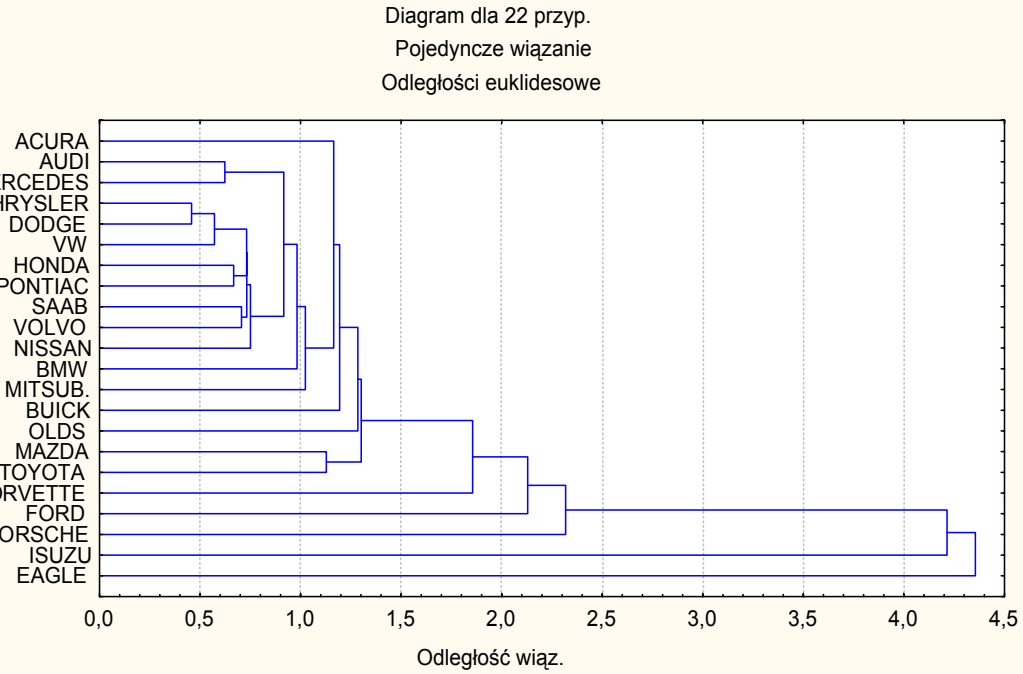
Analiza skupień w Statsoft -Statistica

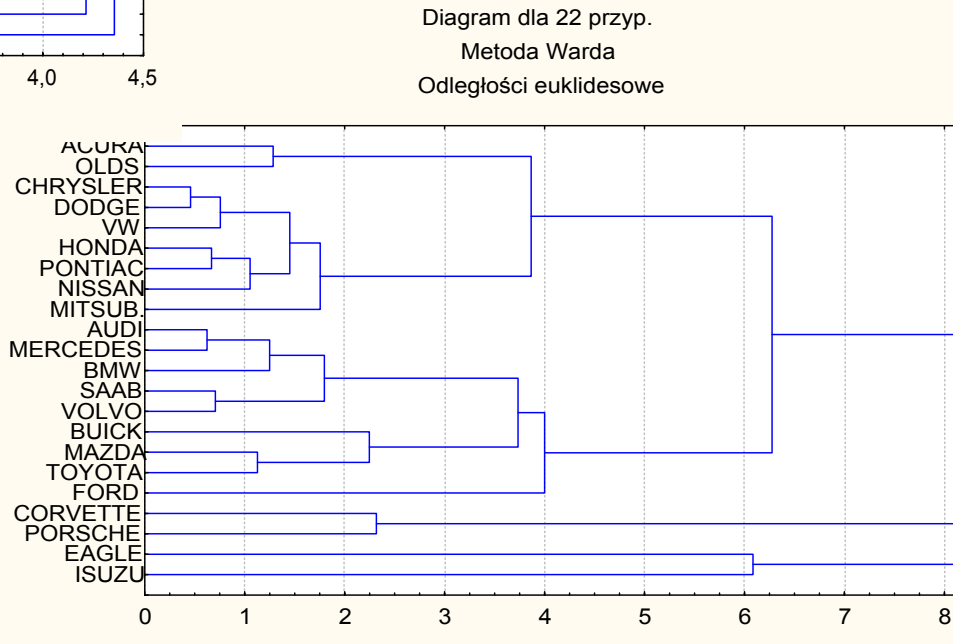
Analiza Skupień – Statistica; więcej na www.statsoft.com. Przykład analizy danych o parametrach samochodów



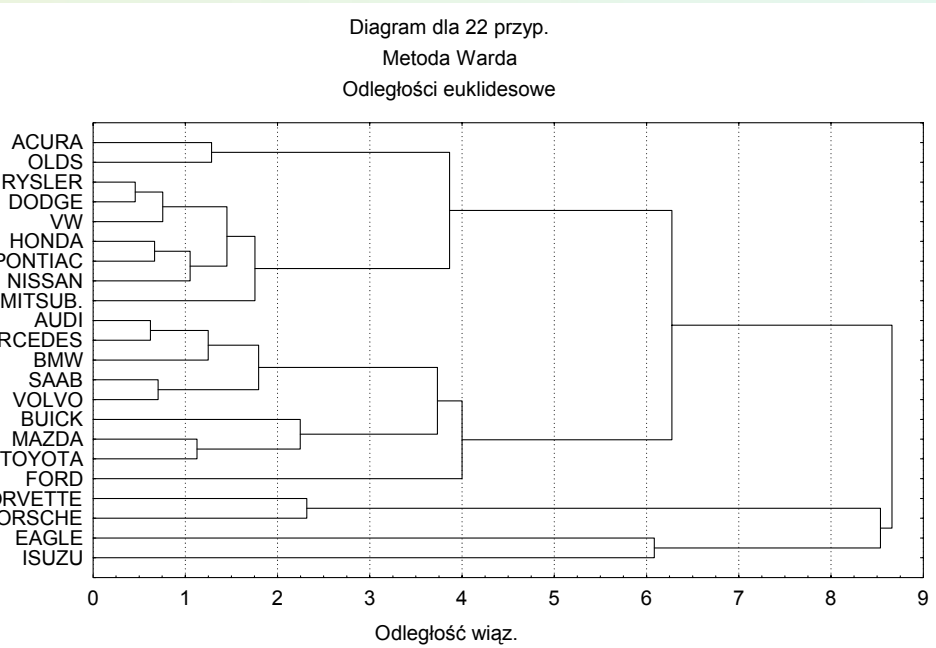

Analiza skupień: Aglomeracja
Zmienne: WSZYSTKIE
OK
Anuluj
Wejście: Dane surowe
Grupowanie: Przypadki (obiekty)
Metoda aglomeracji (wiązania): Pojedynczego wiązania
Miara odległości: Odległość euklidesowa
p: 2
r: 2
Braki danych: Usuwane przypadkami
Przetwarzanie wsadowe i drukowanie
SELECT CASES s
w
aliza 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

Wrażliwość na dobór metod łączenia skupień

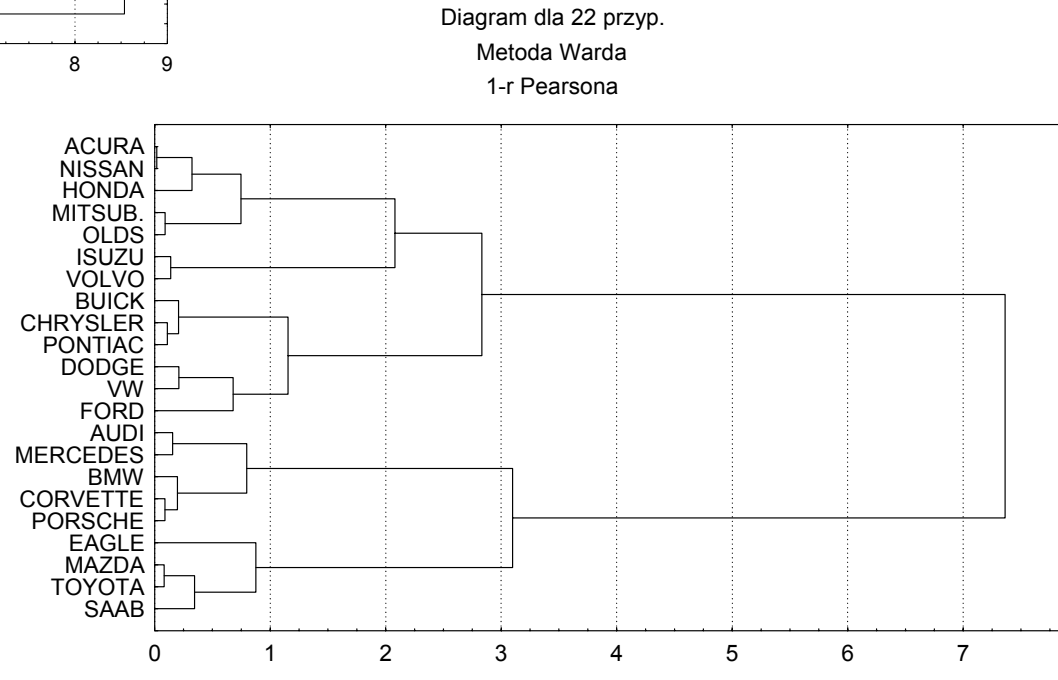




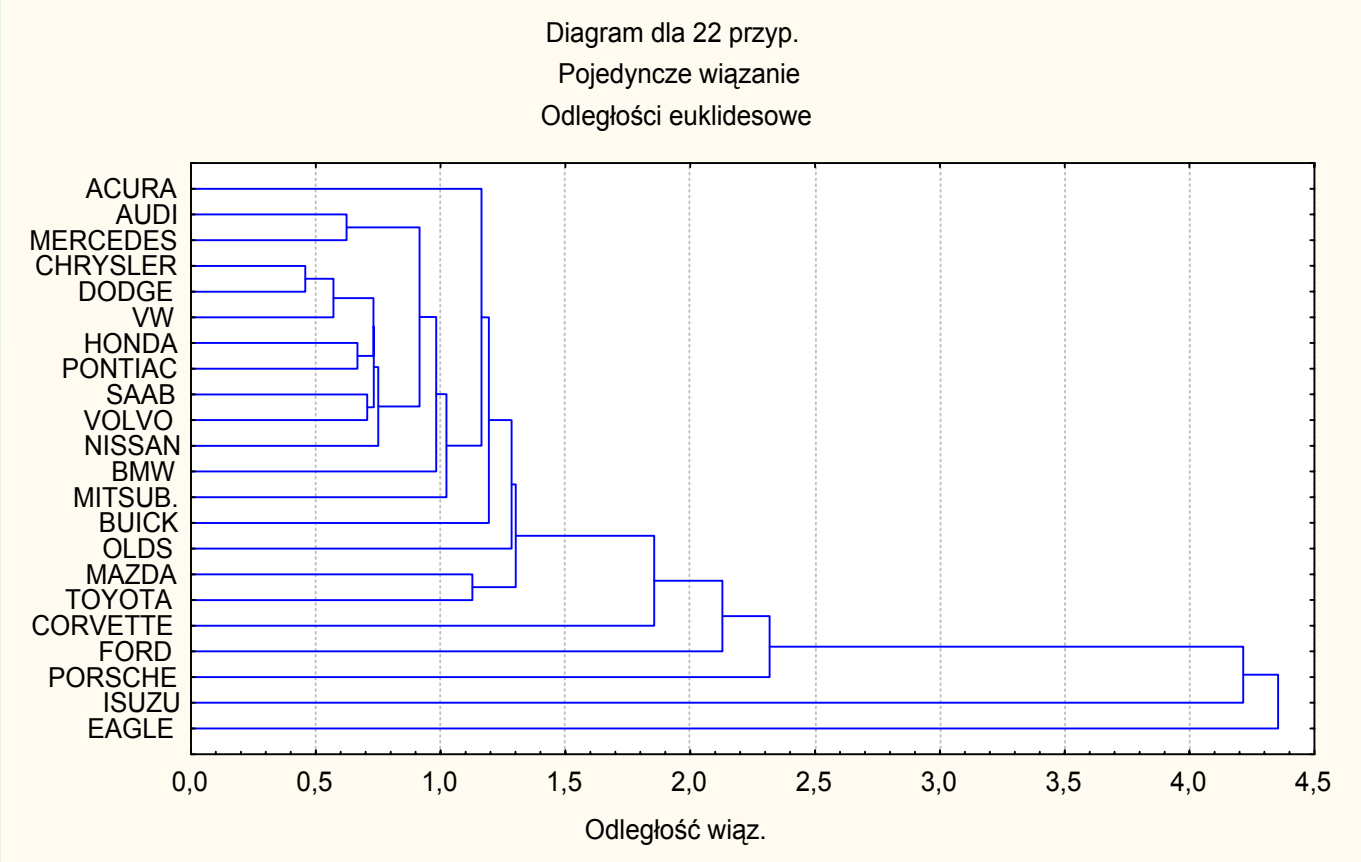
Przykłady użycia metody Warda



Cars data

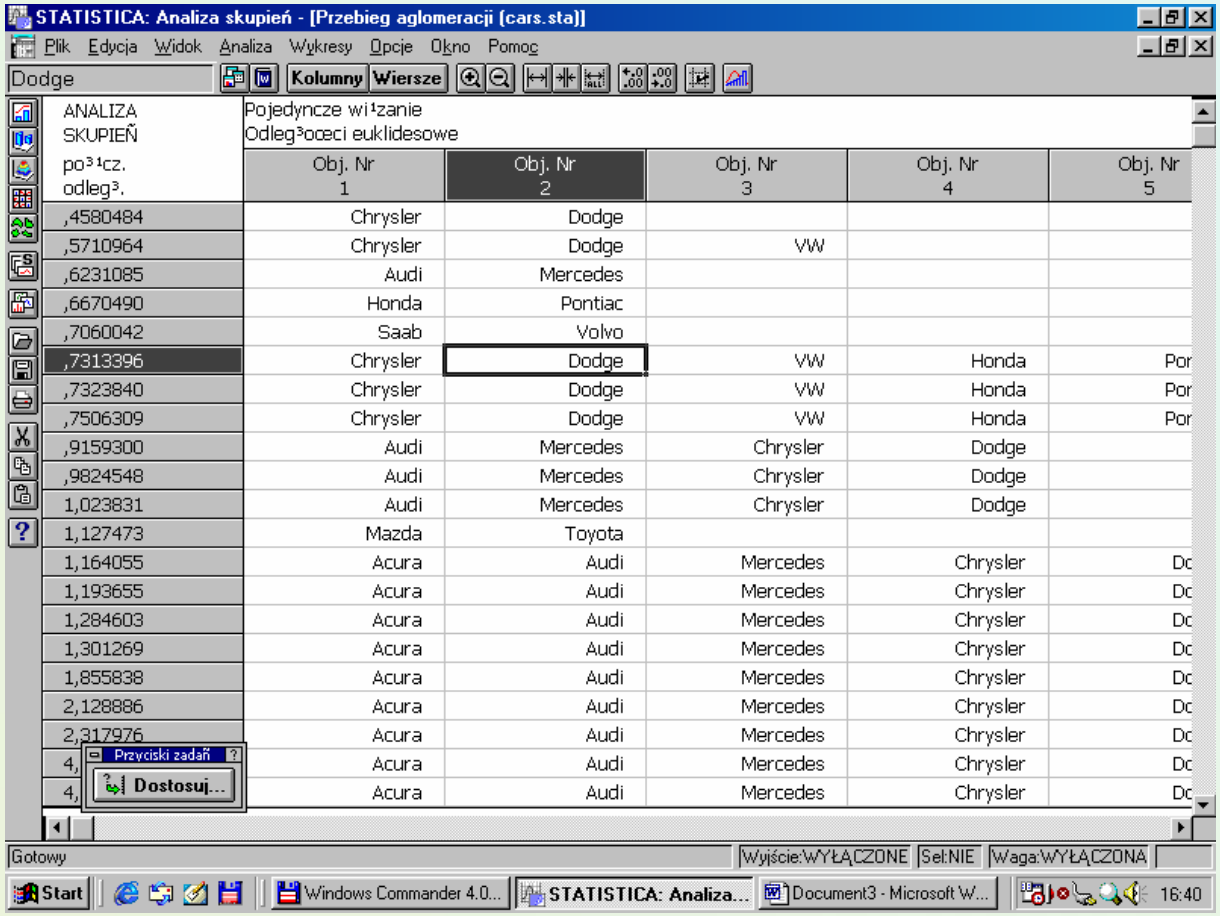


Dendrogram for „Single Linkage”



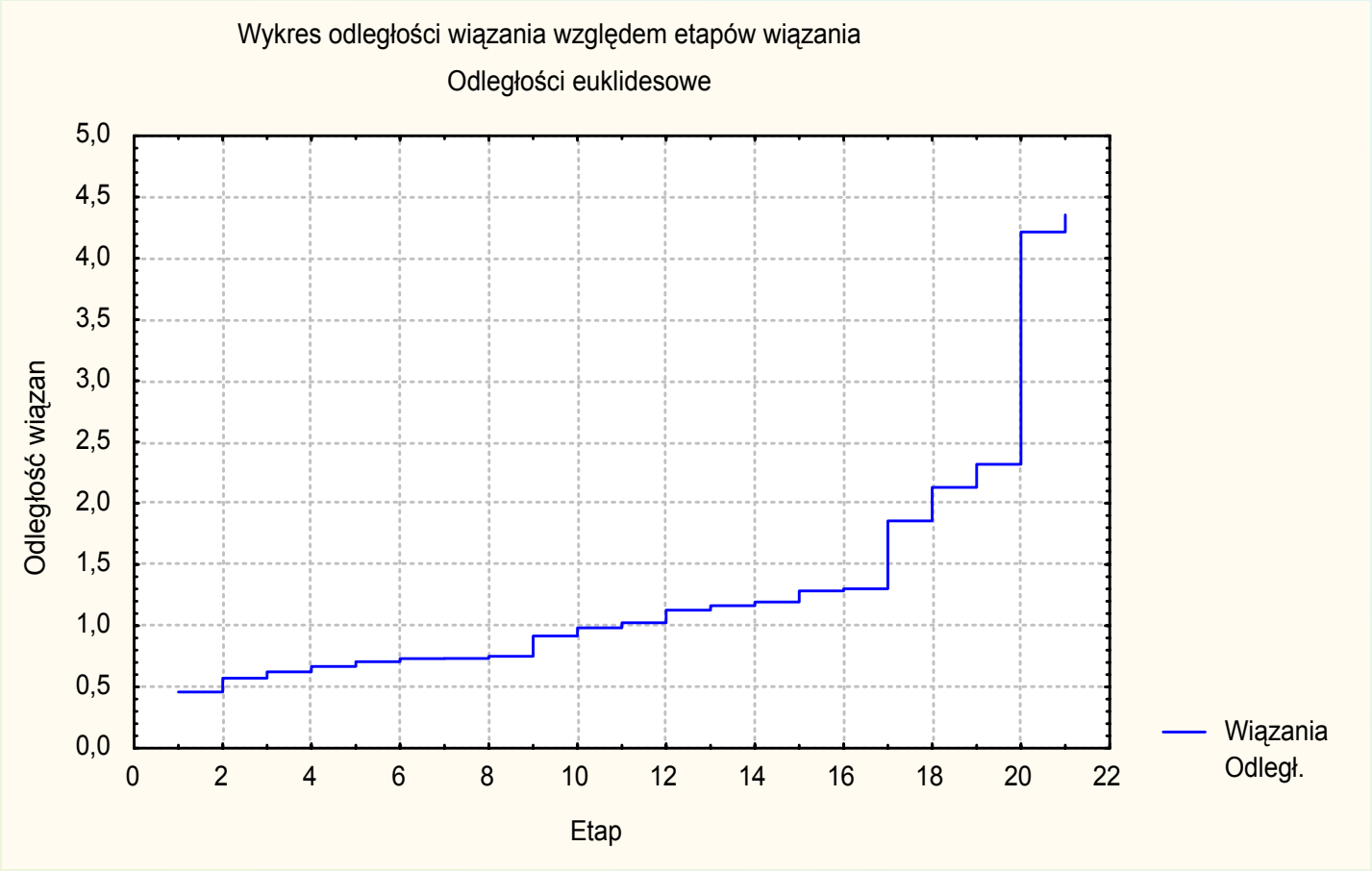
Opis tworzenia dendrogramu

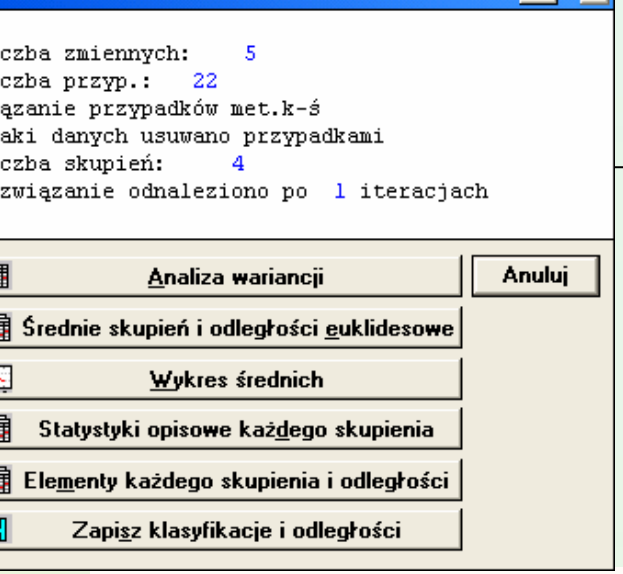
- Łączenie obiektów w kolejnych krokach

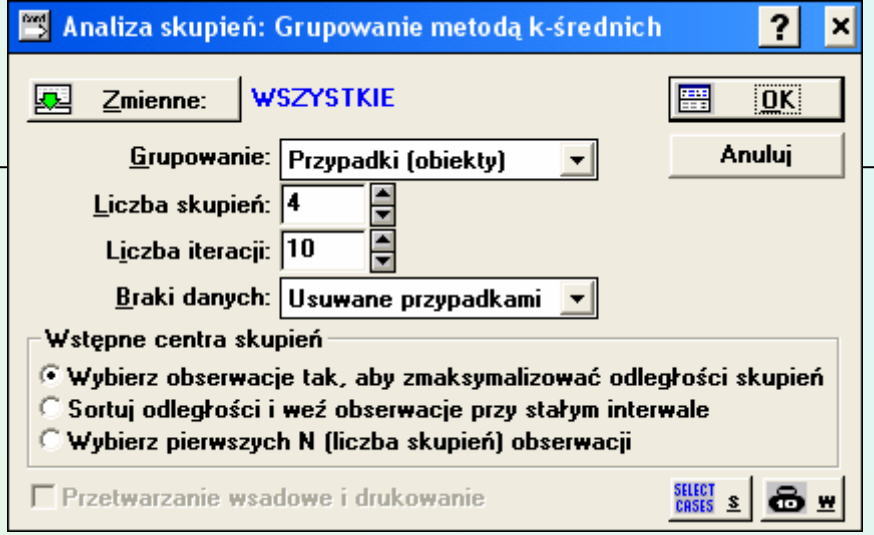


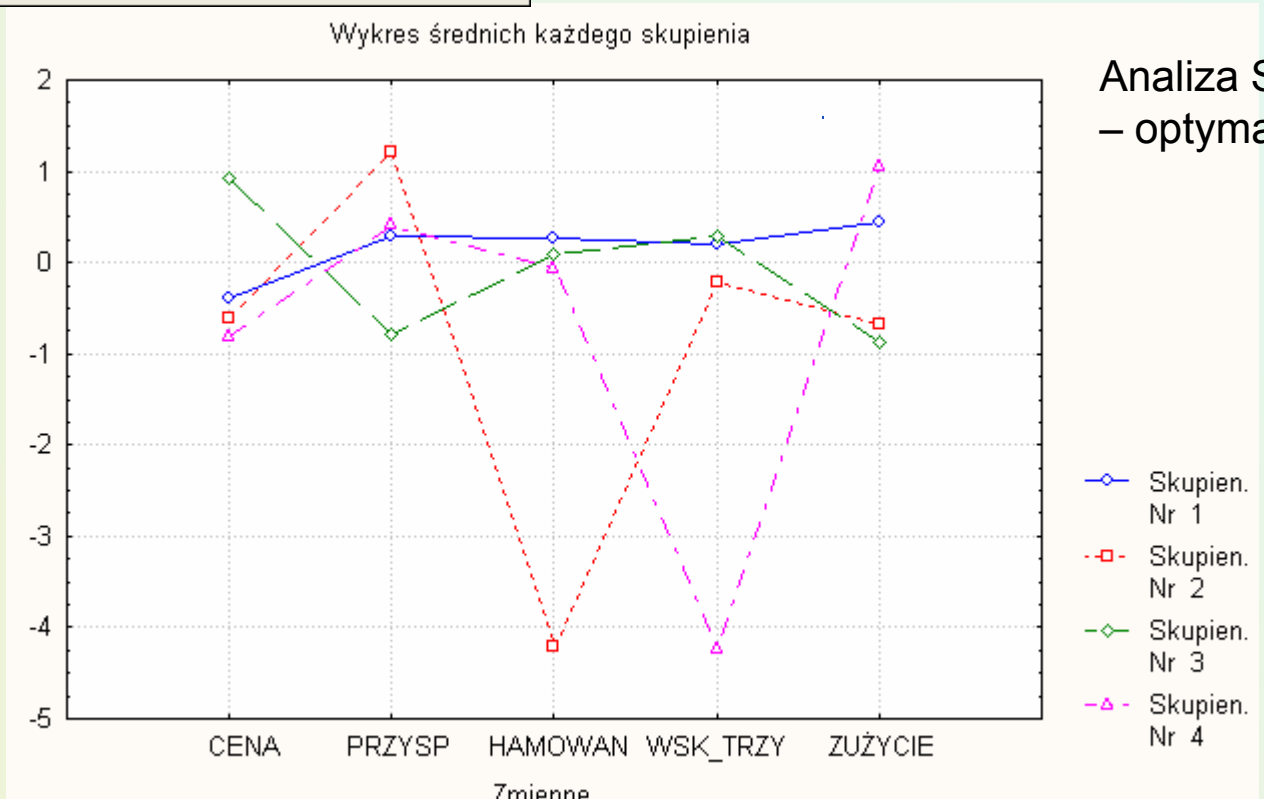
Analiza procesu łączenia

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









Analiza Skupień – optymalizacja k-średnich

Wsparcie do charakterystyki skupień

Odległości euklidesowe skupień (cars.sta)

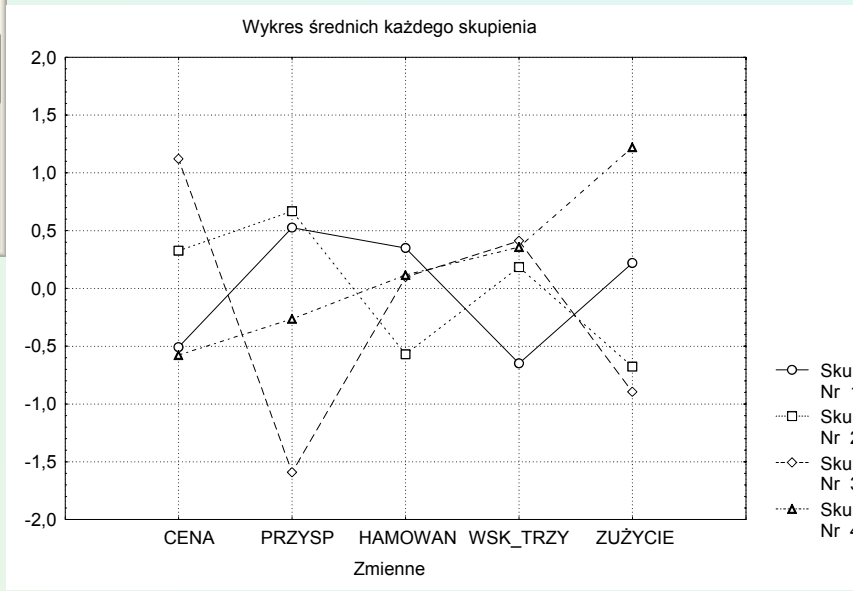
Dalej...

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



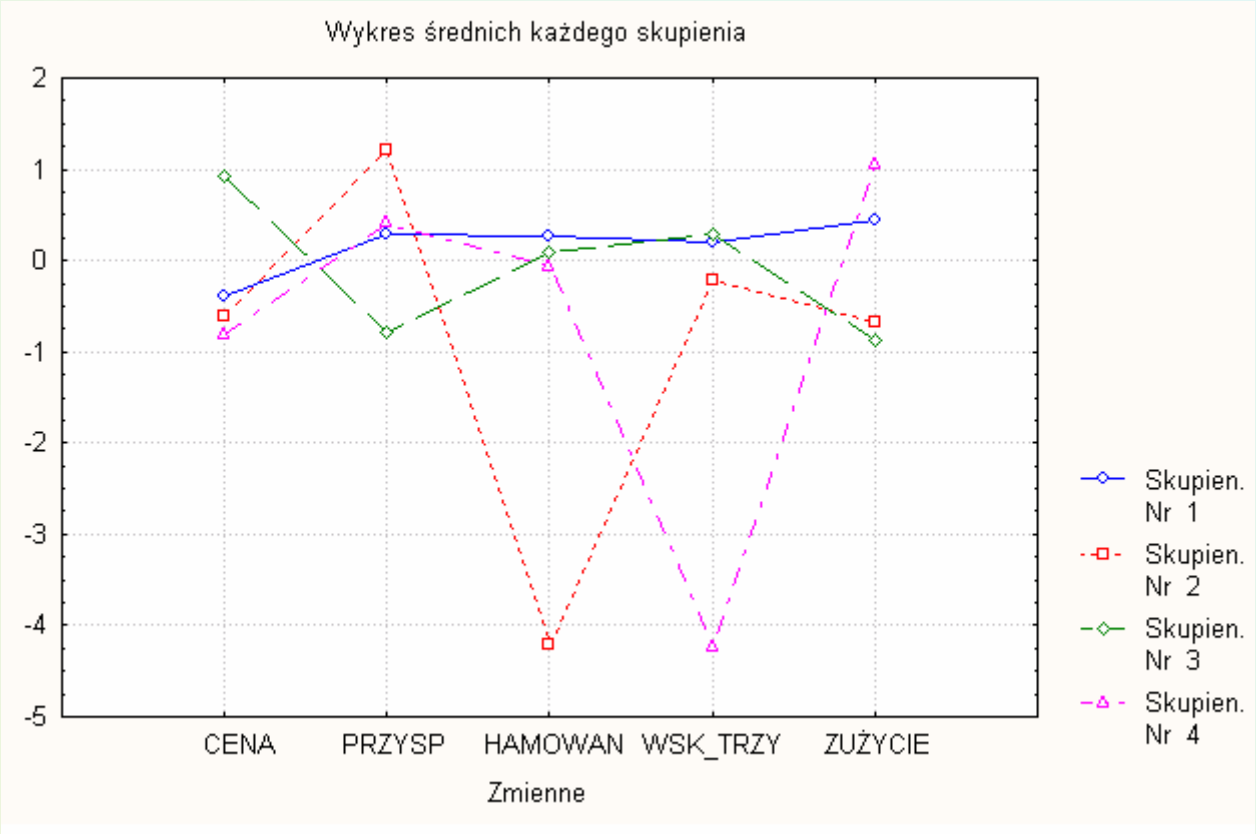
Elementy skupienia numer 2 (cars.sta)

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

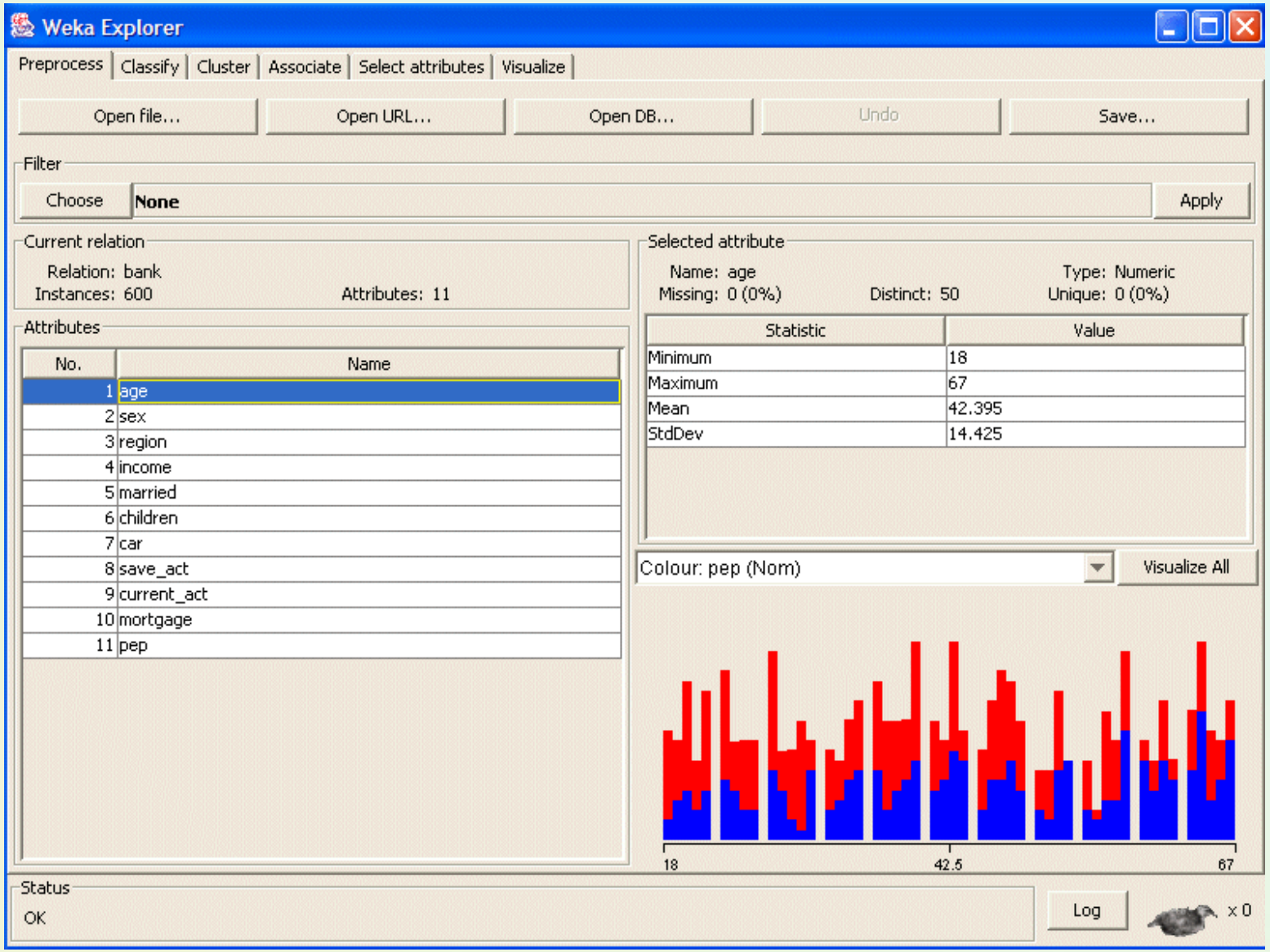
Przykład 1. Bank data in WEKA

- Prosty przykład – sample of customers of the bank
  - Bank data (bank-data.cvs -> bank.arff)
  - Pre-processing wykonany na wersji cvs
  - 600 klientów opisanych przez 11 atrybutó

```
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
……..
…….
```

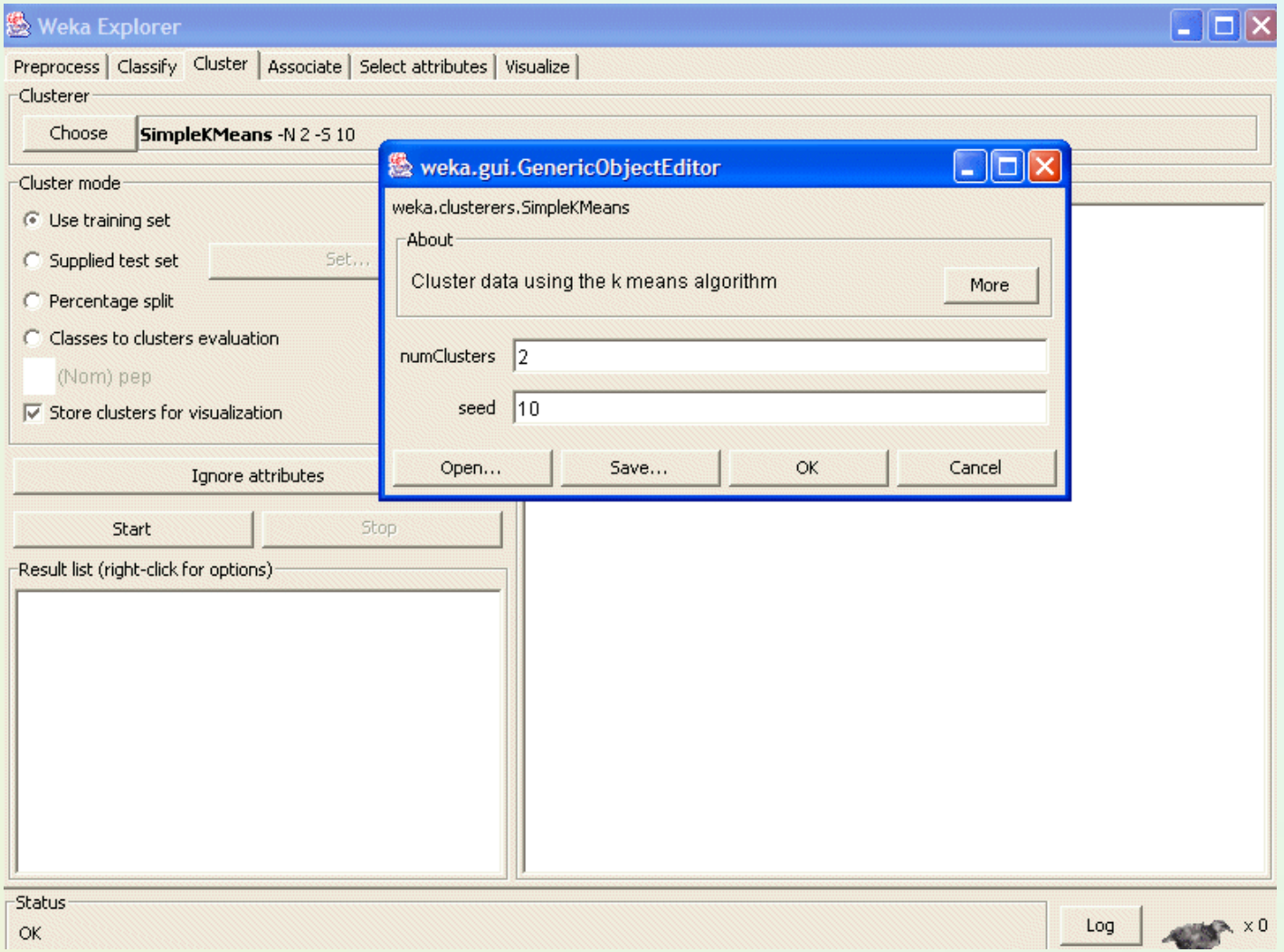
  - Skorzystaj z k-means

Charakterystyka danych

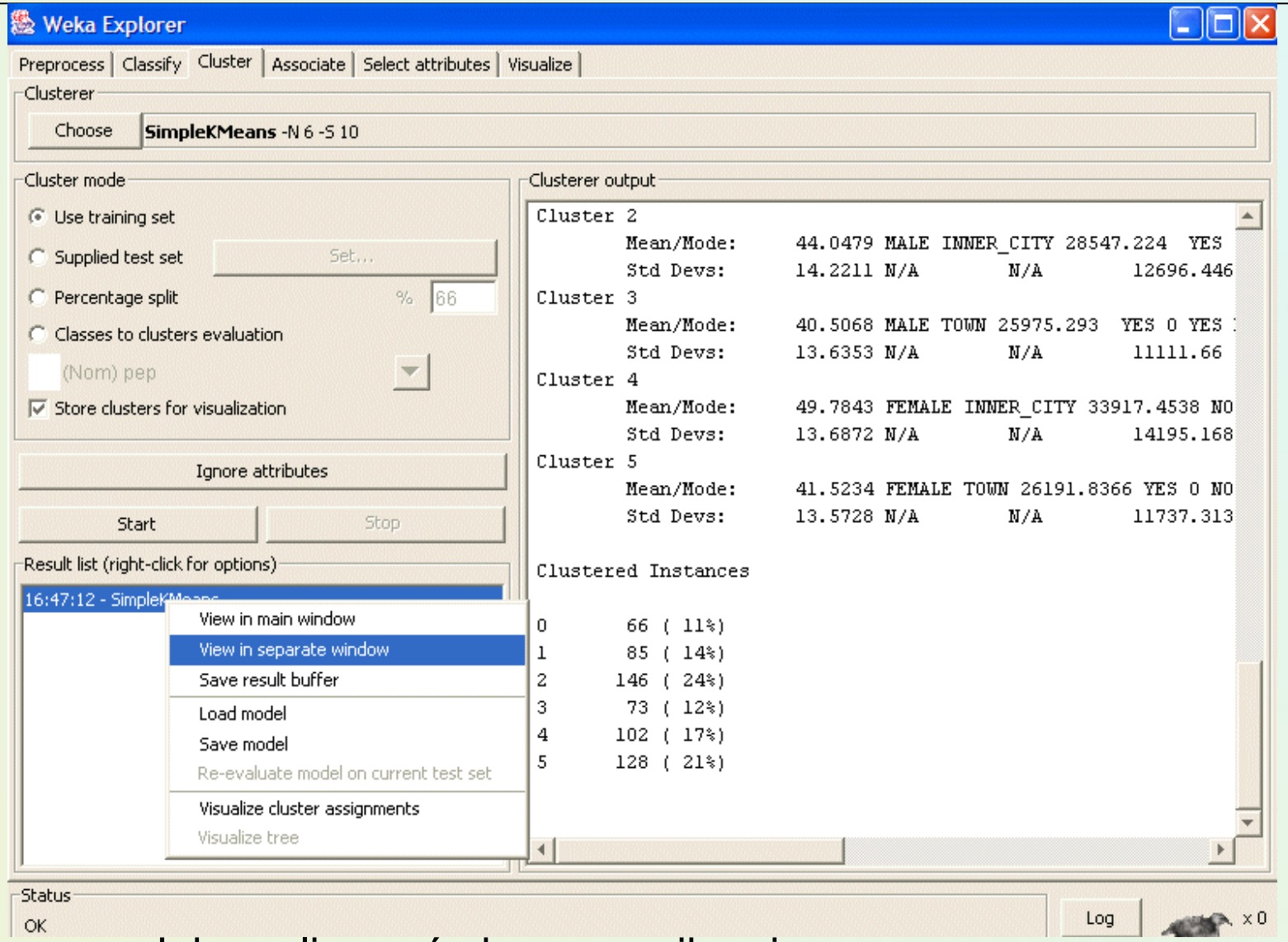


Simple k-means

- Dobór parametrów



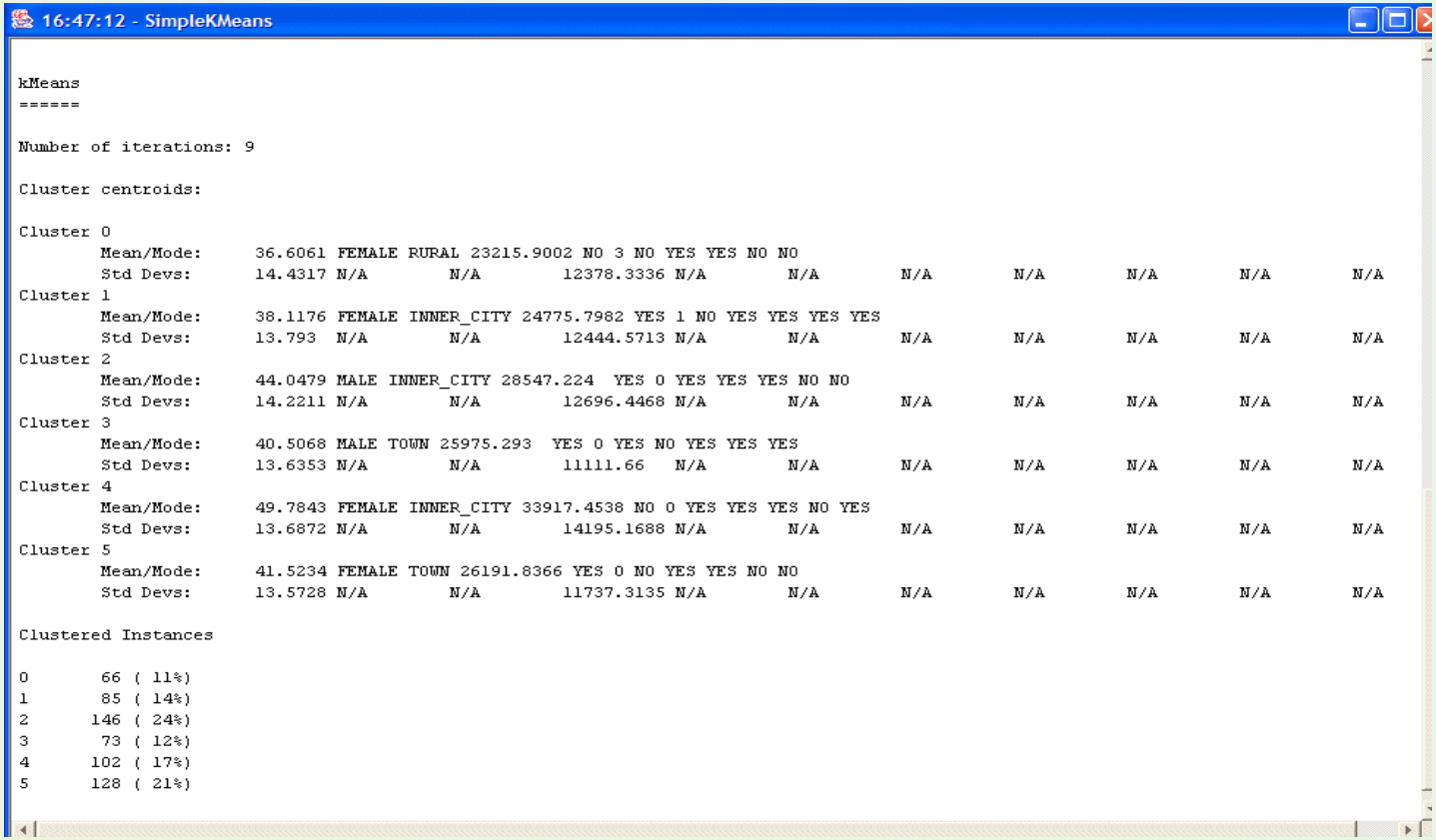
Clustering results



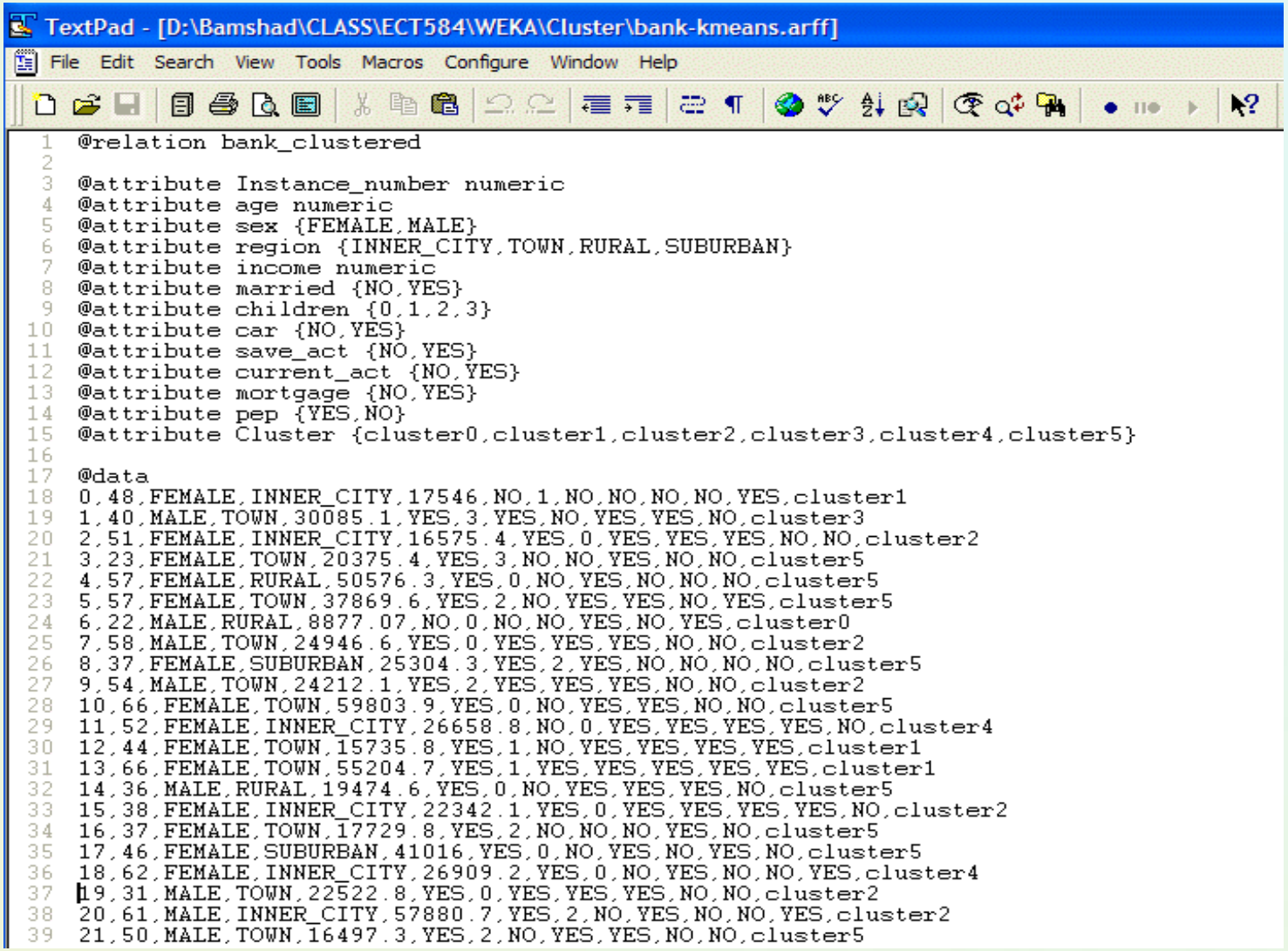
- Jak analizować okno z wynikami

Okno wynikowe

- Co można odnaleźć
- Jak opisano centra skupień?



Gdzie jest przydział do skupień?




```
@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 „trudne” rady

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

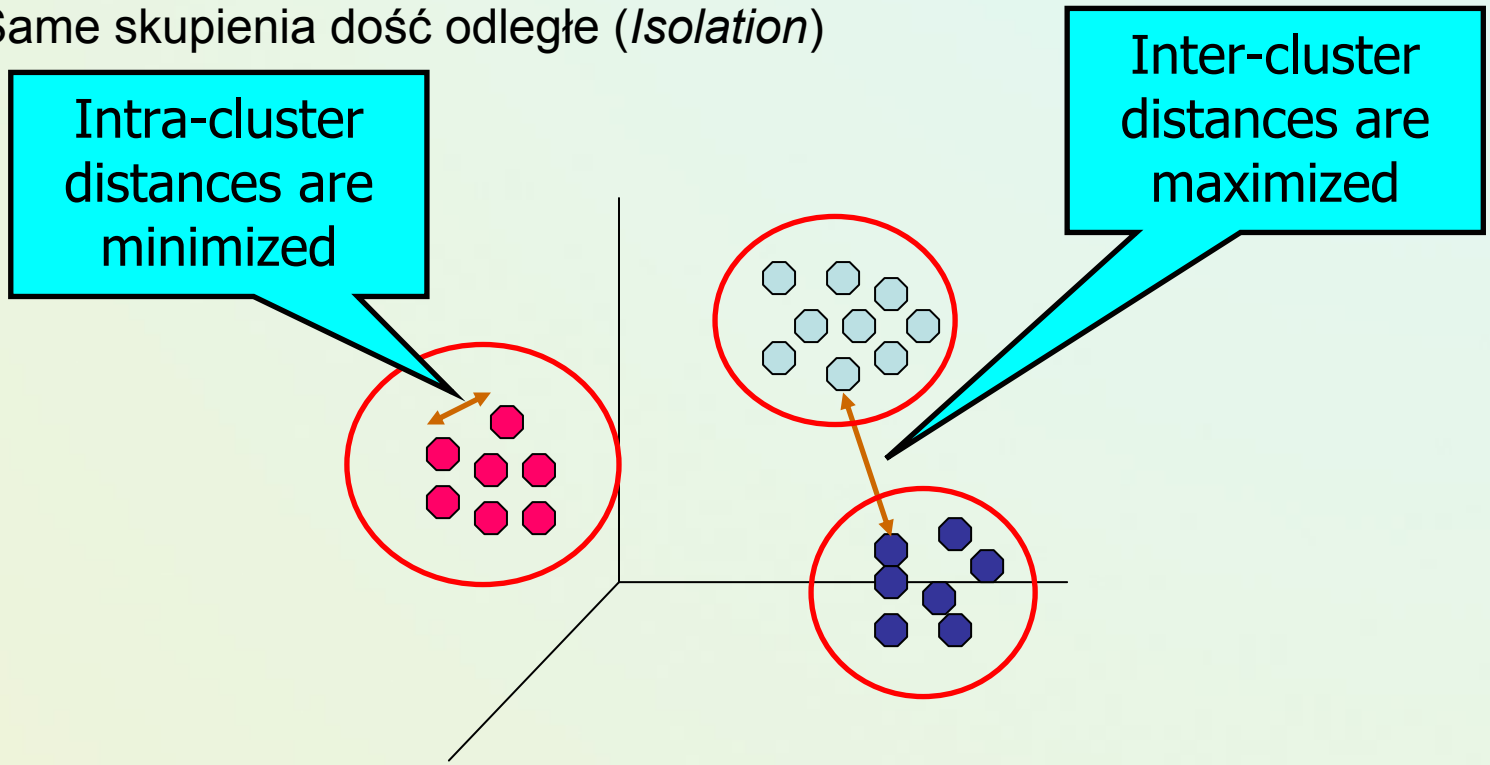
"The nature of processes leading to useful classifications remains little understood, despite considerable effort in this direction."
— R. Michalski, R. Stepp [MS83]

"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

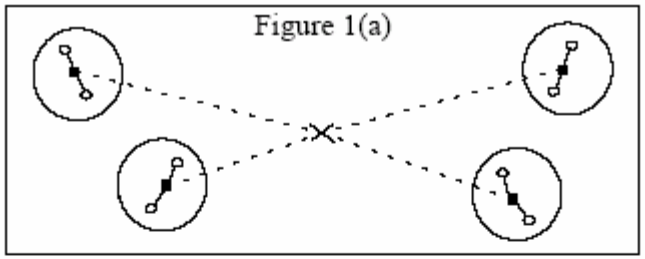
- Compactness → determining the weakest connection within the cluster, i.e., the largest distance between two objects *Ri* and *Rk* within the cluster.
- **Isolation** → determining the strongest connection of a cluster to another cluster, i.e., the smallest distance between a cluster centroid and another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



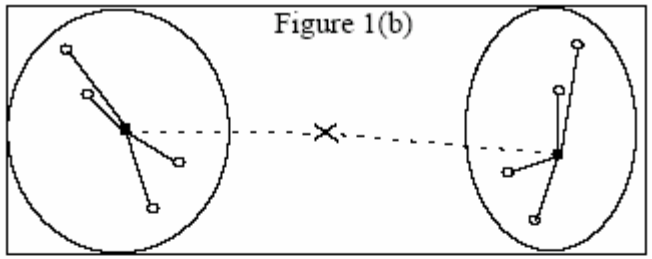
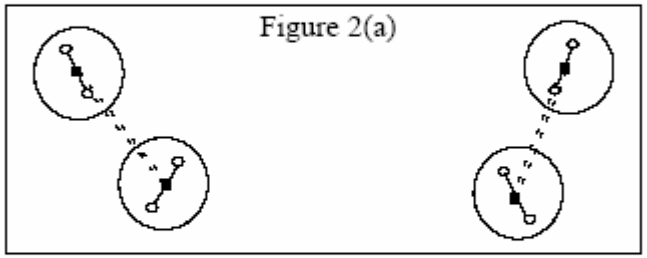


Figure 1: Minimum Total Distance Criterion



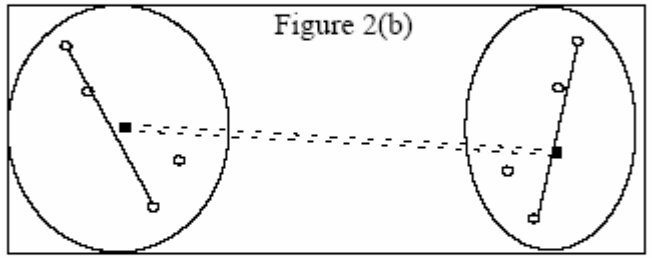
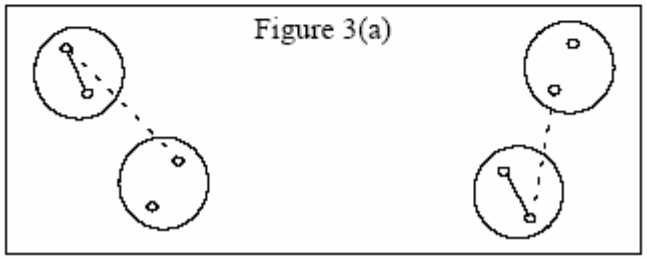


Figure 2: Separated Clusters Criterion



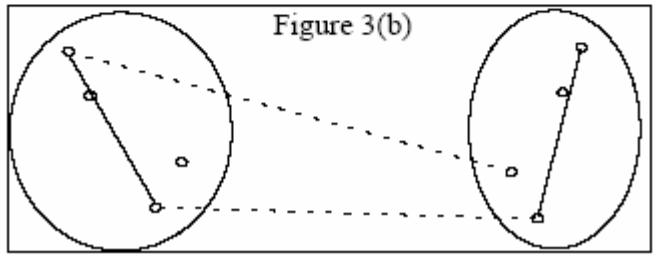


Figure 3: Object Positioning Criterion

Dodatkowe zewnętrzne informacje

- Dysponujemy referencyjnymi etykietami

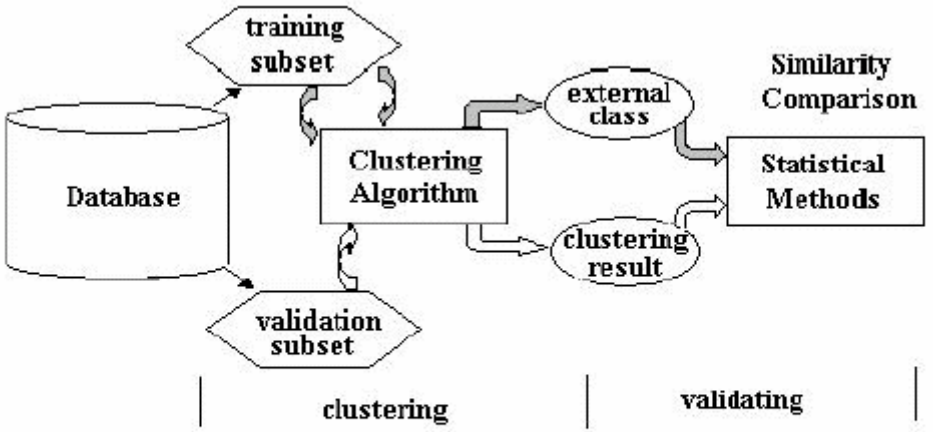


Figure 3-4 External criteria based validation [ZOZ07a]

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

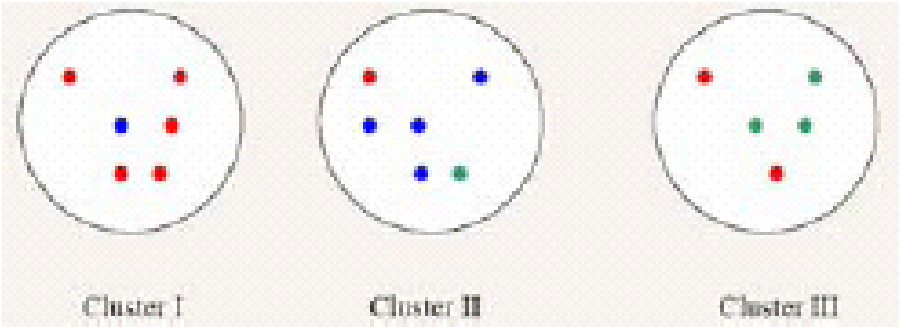
- Homogeneity - Jednorodności
  - Każde skupienie zawiera przykłady z jak najmniejszej liczby etykiet klas
  - Idealnie – tylko jedna klasa
- „Completeness”
  - Każda klasy reprezentowana w możliwe najmniejszej liczbie skupień
- Typowe miary
  - Purity
  - F-miara

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



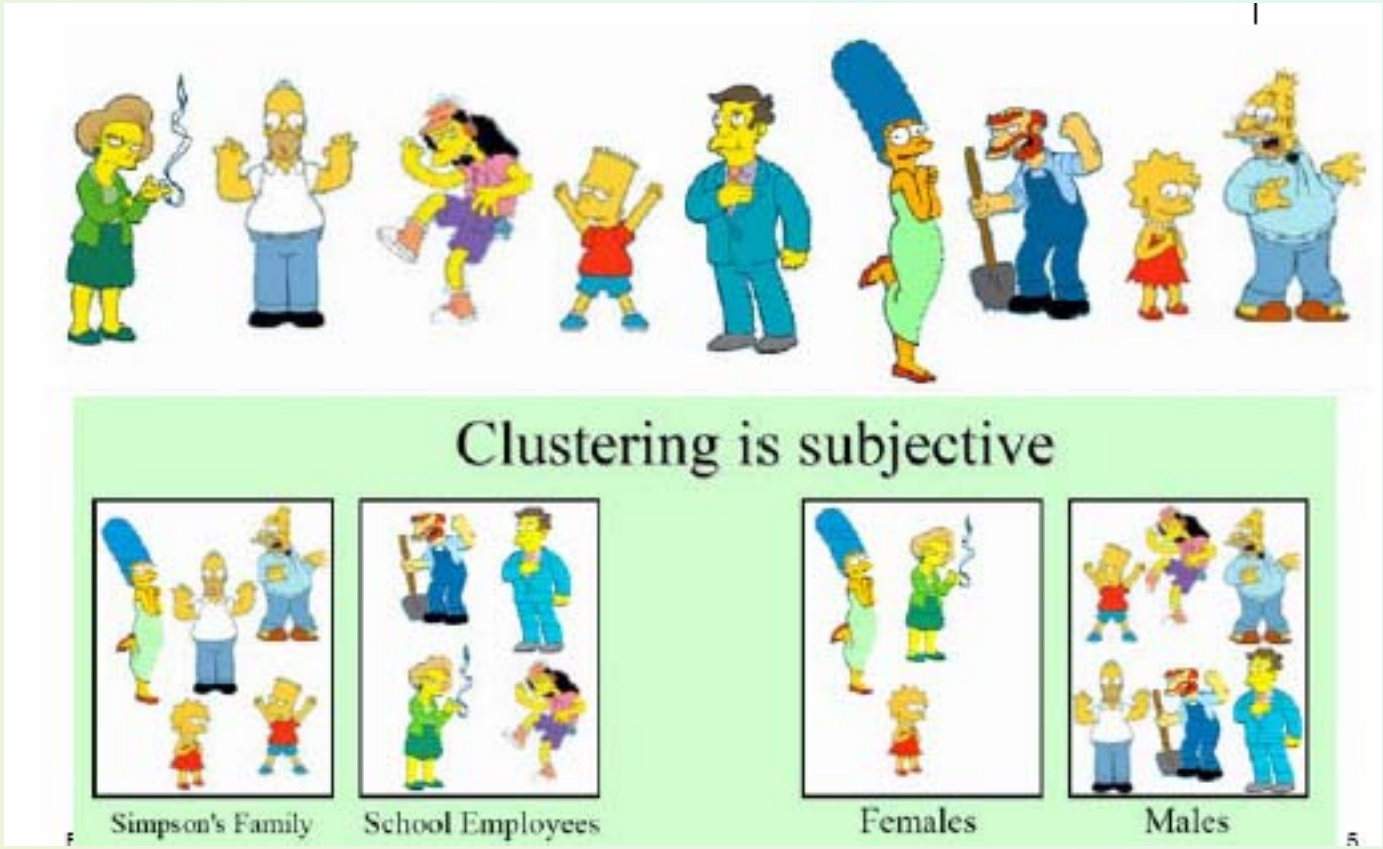
Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

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

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

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)

Grupowanie Dużych Repozytoriów

- Skalowalność (przykłady lecz także „High dimensionality”)
- Uwzględnianie złożonych typów danych
- Konstruowanie dowolnych kształtów skupień
- Wspomaganie parametryzacji (jak dobrać *k*, parametry DBSCAN, itp.)
- Odporność na szum i obserwacje nietypowe
- Grupowanie przyrostowe
- Przetwarzanie strumieni danych
- „Concept drift”

Problemy i wyzwania

- Znaczący postęp w zakresie skalowalnych algorytmów:
  - Partitioning: *k*-means, *k*-medoids, PAM, CLARANS
  - Hierarchical: BIRCH, CURE
  - Density-based: DBSCAN, CLIQUE, OPTICS
  - Grid-based: STING, WaveCluster.
  - Model-based: Autoclass, Denclue, Cobweb.
- Obecne techniki ciągle nie spełniają wystarczająco dobrze stawianych wymagań.
- Otwarte problemy i wyzwania badawcze; zwłaszcza dla nietypowych i złożonych danych.

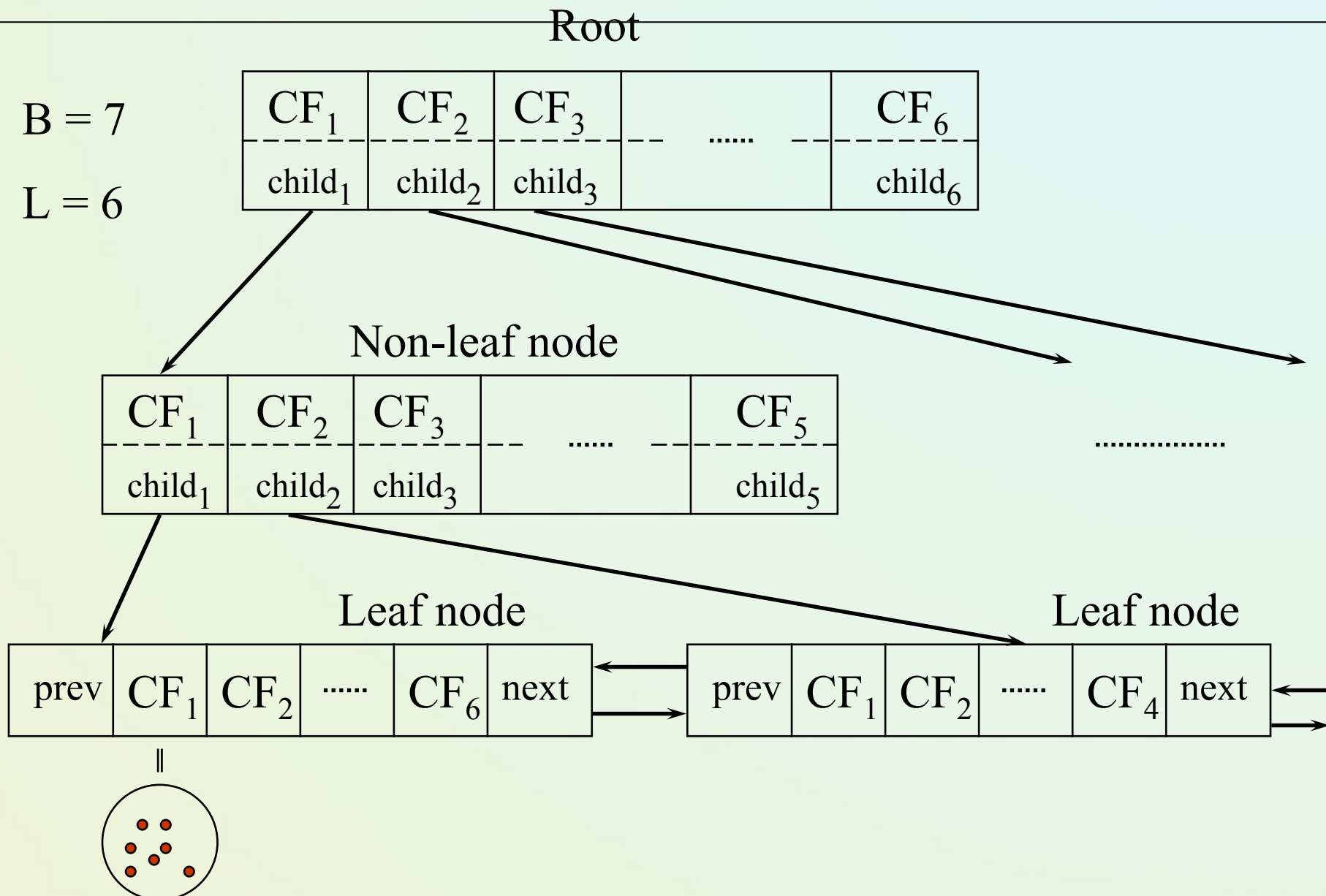
Metody hierarchiczne dla dużych zbiorów danych

- Niektóre z ograniczeń metod aglomeracyjnych:
  - słaba skalowalność: złożoność czasowa przynajmniej $O(n^2)$, gdzie n jest liczbą obiektów,
  - „krytyczne” znaczenie decyzji o wyborze punktu połączenia kolejnych skupień w trakcie budowania drzewa hierarchii,
  - algorytmy nie zmieniają, ani nie poprawiają, wcześniej podjętych decyzji.
- Rozwinięcia algorytmów hierarchicznych oraz ich integracja z metodami gęstościowymi:
  - BIRCH (1996): użycie drzew „CF-tree”, uczenie przyrostowe i stopniowa poprawa jakości pod-skupień.
  - CURE (1998): wybór losowy odpowiednio rozproszonych punktów, wstępne grupowanie z określeniem ich punktów reprezentatywnych, łączenie grup w nowe skupienia wraz z przesuwaniem punktów reprezentatywnych w stronę środków tworzonego skupienia zgodnie z „shrinking factor α”; eliminacja wpływu „outliners”.

BIRCH – ang. Balanced Iterative Reducing and Clustering using Hierarchies – Zhang et al. (1996)

- Wykorzystuje hierarchiczne drzewo CF (Clustering Feature)
- Działanie algorytmu:
  - Faza 1: przyrostowo przeczytaj raz DB w celu zbudowania w pamięci początkowej struktury drzewa CF (rodzaj wielopoziomowej kompresji danych zachowującej wewnętrzną strukturę zgrupowań danych).
  - Faza 2: zastosuj wybrany (inny) algorytm skupień dla lepszego pogrupowania obiektów w liściach drzewa CF.
- *Dobra skalowalność*: znajduje zadawalające grupowanie po jednokrotnym przeczytaniu bazy danych i ulepsza je wykorzystując niedużo dodatkowych operacji odczytu DB.
- *Ograniczenia:* zaproponowany dla danych liczbowych, wrażliwość wyników na kolejność prezentacji przykładów.

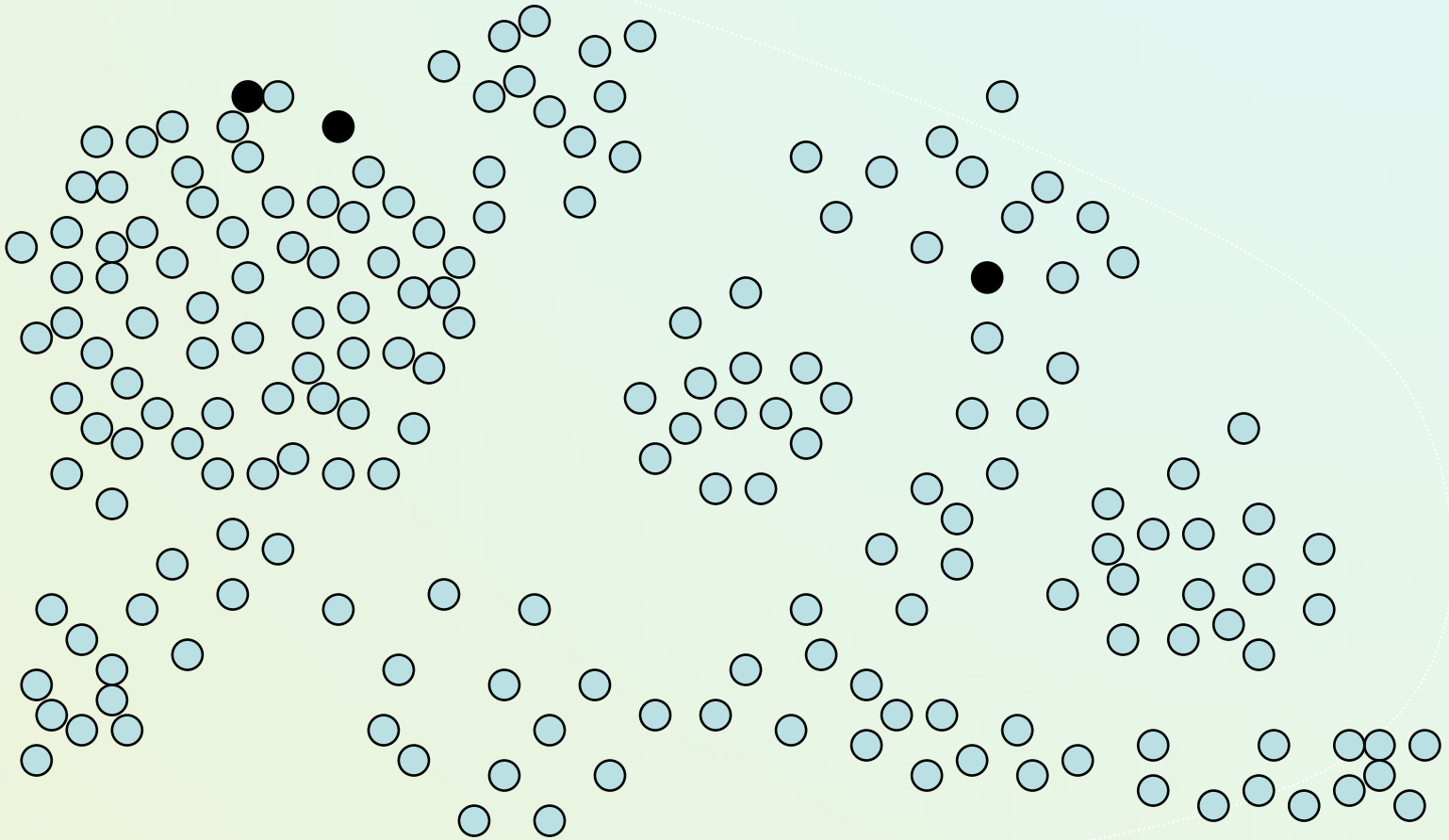
CF Tree


Root
B = 7
L = 6
CF1 CF2 CF3 ..... CF6
child1 child2 child3 child6
Non-leaf node
CF1 CF2 CF3 ..... CF5
child1 child2 child3 child5
Leaf node
Leaf node
prev CF1 CF2 ..... CF6 next
prev CF1 CF2 ..... CF4 next

Metody gęstościowe

- Podstawowe metody wykorzystują miary odległości między obiektami
- Inne metody wykorzystują pojęcie gęstości (ang. density) – lokalne sąsiedztwo punktu/skupienia, a także „gęsto” połączonych punktów
- Właściwości metod gęstościowych:
  - Wykrywanie skupień o dowolnych kształtach (niesferycznych)
  - Odporność na „szum informacyjny”
  - Jednokrotne przeglądanie DB
  - Potrzebna parametryzacja oceny gęstości i warunków zatrzymania
- Interesujące algorytmy:
  - DBSCAN: Ester, et al. (KDD’96)
  - OPTICS: Ankerst, et al (SIGMOD’99).
  - DENCLUE: Hinneburg & D. Keim (KDD’98)
  - CLIQUE: Agrawal, et al. (SIGMOD’98)

Modelowanie dowolnych kształtów



DBSCAN: Algorytm gęstościowy

- DBSCAN: Density Based Spatial Clustering of Applications with Noise.
  - Wykorzystuje pojęcie „*density-based cluster*”: Skupienie będące maksymalnym zbiorem punktów gęsto połączonych „*density-connected points*”.
  - Poszukuje się zgrupowań odpowiednio gęsto (blisko siebie) położonych obiektów (*dense regions/clusters*) odzielonych od siebie obszarami o niskiej gęstości („noise”)
  - Możliwość wykrywania skupień o dowolnym kształcie w obecności szumu informacyjnego (noise)

DBSCAN: Podstawowe pojęcia

Parametry**:**

- ***Eps***: Maksymalny promień sąsiedztwa
- ***MinPts***: minimalna liczba punktów (obiektów) w Eps-sąsiedztwie badanego punktu

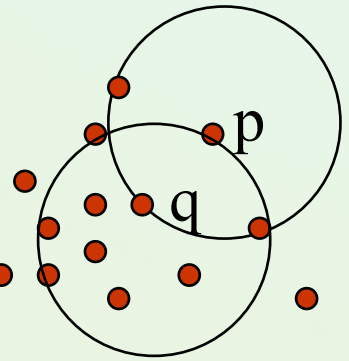
$N_{Eps}(p)$: **{punkt *q należy do D* | *dist(p,q)* <= *Eps*}**

Directly density-reachable**:** A point ***p*** is directly density-reachable from a point ***q*** wrt. ***Eps***, ***MinPts*** if

1) ***p*** belongs to $N_{Eps}(q)$

2) core point condition:

$$|N_{Eps}(q)| >= MinPts$$

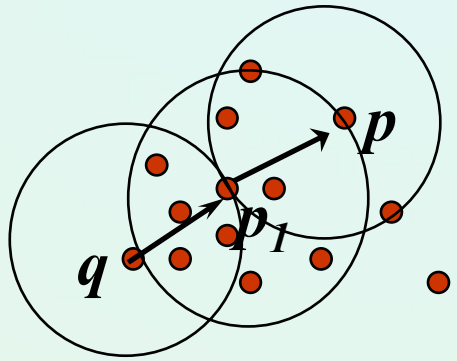


MinPts = 5

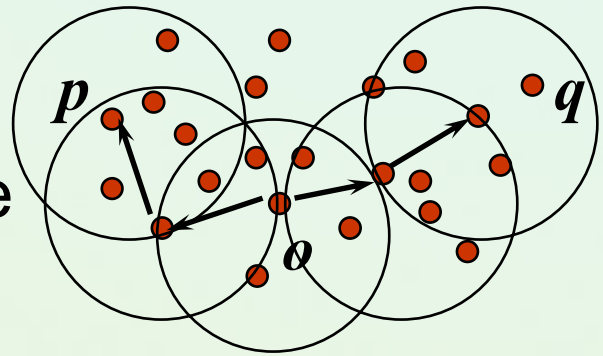
Eps = 1 cm

DBSCAN: Podstawowe pojęcia (II)

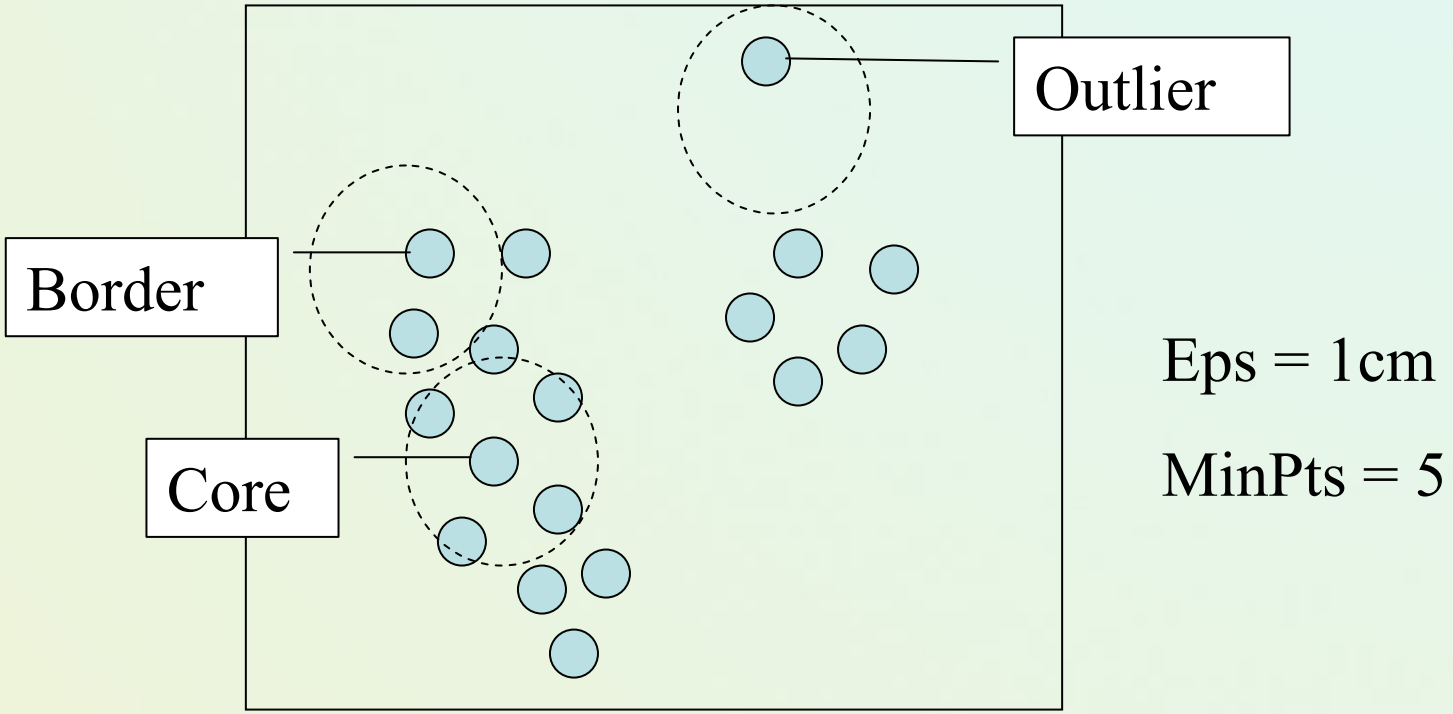
- Density-reachable:
  - A point *p* is density-reachable from a point *q* wrt. *Eps*, *MinPts* if there is a chain of points p_1, …, p_n, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i



- Density-connected
  - A point *p* is density-connected to a point *q* wrt. *Eps*, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* wrt. *Eps* and *MinPts*.



DBSCAN: General Ideas



DBSCAN: Zarys algorytmu

- Wybierz punkt startowy ***p***
- Odnajdź wszystkie punkty do gęstościowego osiągnięcia z *p* (density-reachable from ***p*** wrt ***Eps*** and ***MinPts)***.
- Jeśli ***p*** jest rdzeniem (*core point*), utwórz skupienie.
- Jeśli ***p*** jest punktem granicznym (border point) i żadne punkty nie są z niego gęstościowo osiągalne, DBSCAN wybiera następny punkt z bazy danych
- Proces jest konytuowany dopóki żaden nowy punkt nie może być dodany to dowolnego skupienia.
- Złożoność: O(*n*log *n*) w przypadku użycia specjalnego „spatial index”, w przeciwnym razie O(n^2).

Clustering in Data Mining – szukaj więcej!

Data Clustering: A Review

A.K. JAIN

Michigan State University

M.N. MURTY

Indian Institute of Science

AND

P.J. FLYNN

The Ohio State University

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorially, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur. This paper presents an overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. We present a taxonomy of clustering techniques, and identify cross-cutting themes and recent advances. We also describe some important applications of clustering algorithms such as image segmentation, object recognition, and information retrieval.

Categories and Subject Descriptors: I.5.1 [**Pattern Recognition**]: Models; I.5.3 [**Pattern Recognition**]: Clustering; I.5.4 [**Pattern Recognition**]: Applications—*Computer vision*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge acquisition*

Analiza skupień - podsumowanie

- Liczne i ważne zastosowanie praktyczne analizy skupień (AS).
- AS używana „samodzielnie” w zgłębianiu danych, lub jako jedno z narzędzi podczas wstępnego przetwarzania w procesie KDD.
- Jakość skupień i działanie wielu algorytmów związane są określeniem miary odległości obiektów.
- Podstawowe klasy metod:
  - hierarchiczne,
  - podziałowo/optymalizacyjne,
  - gęstościowe,
  - „grid-based”,
  - wykorzystujące modele matematyczne (np. probabilistyczne lub neuronowe)
- Ważne zagadnienie to także wykrywanie obiektów nietypowych (outliers discovery).

Może pytanie lub komentarze?