Analiza Skupień - Grupowanie
Zaawansowana Eksploracja Danych

JERZY STEFANOWSKI
Inst. Informatyki PP
Wersja dla TPD 2009,
Aktualizacja 2010
Email: Jerzy.Stefanowski@cs.put.poznan.pl

Elementy terminologiczne

Troche 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.

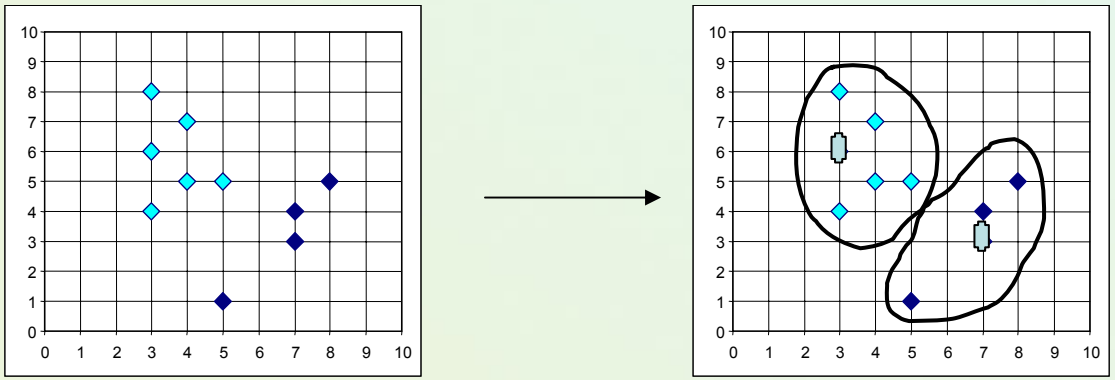
Przykłady zastosowań analizy skupień

- Zastosowania ekonomiczne:
  - Identyfikacja grup klientów bankowych (np. właścicieli kart kredytowych wg. sposobu wykorzystania kart oraz stylu życia, danych osobowych, demograficznych) → 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

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

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

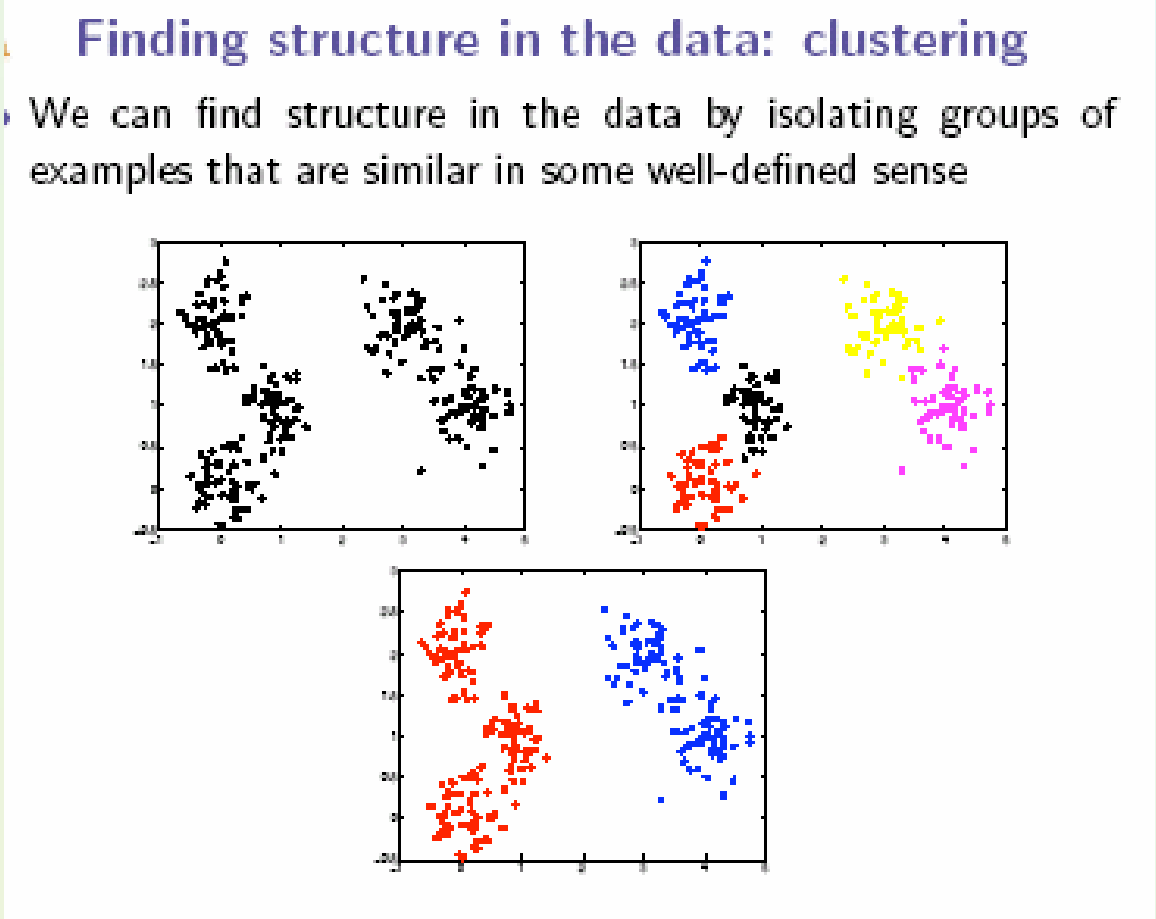


Czym jest skupienie?

1. Zbiorem najbardziej podobnych obiektów
2. Podzbiór obiektów, dla których odległość jest mniejsza niż ich odległość od obiektów z innych skupień.
3. Podobszar wielowymiarowej przestrzeniu zawierający odpowiednio dużą gęstość obiektów

Poszukiwanie „zrozumiałych struktur” w danych

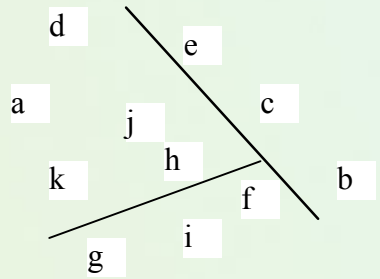
- Ma ułatwiać odnalezienie pewnych spójnych podobszarów danych



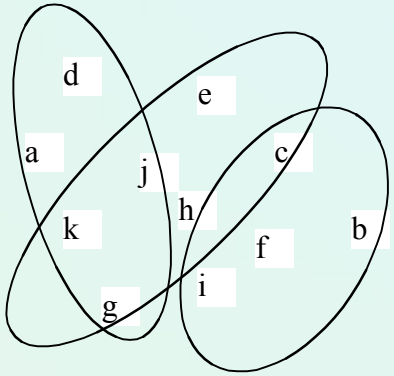
- Lecz nadal wiele możliwości

Różne sposoby reprezentacji skupień

(a)



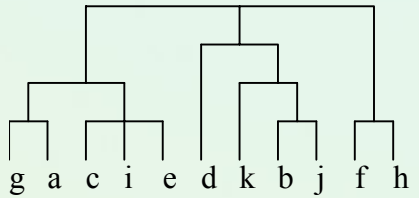
(b)



(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)



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

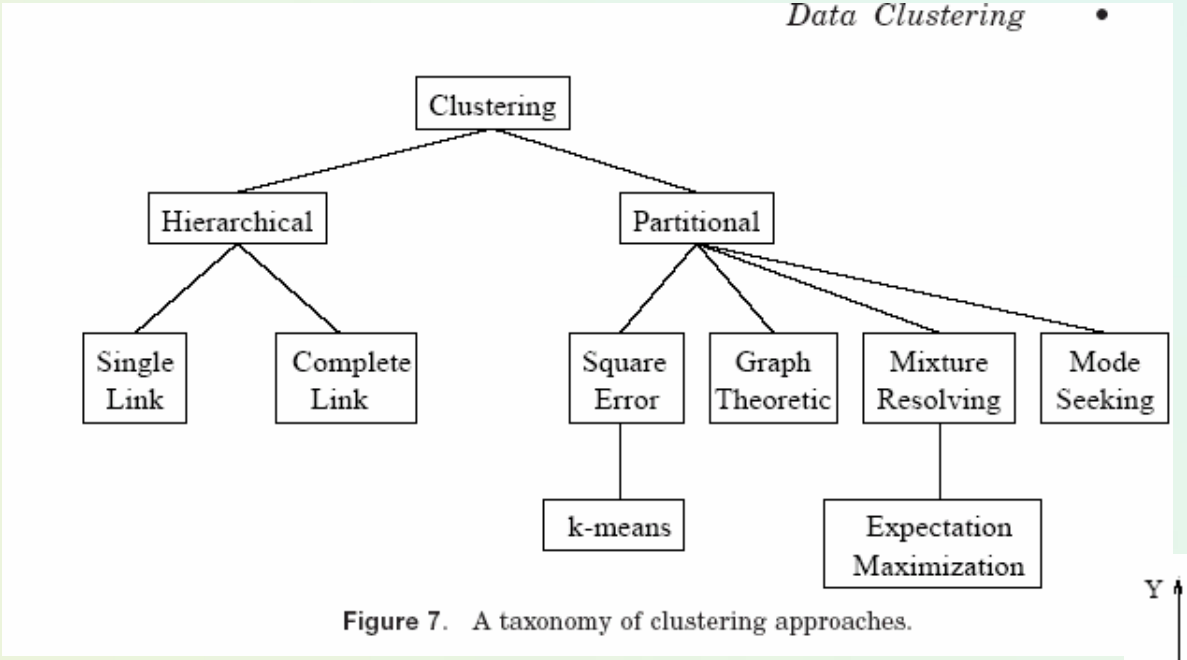


Figure 7. A taxonomy of clustering approaches.

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

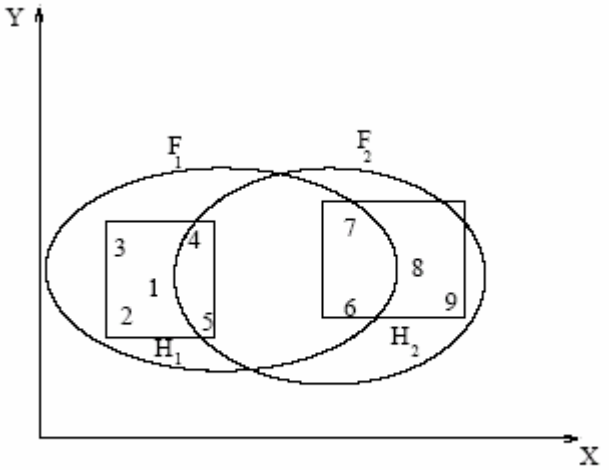


Figure 16. Fuzzy clusters.

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	$d_{(rs)} = \left[\sum_{k=1}^{p} \left(x_{(r)k} - x_{(s)k} \right)^2 \right]^{\frac{1}{2}}$	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).

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 $\rightarrow$ $O(Knl)$

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*).

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:

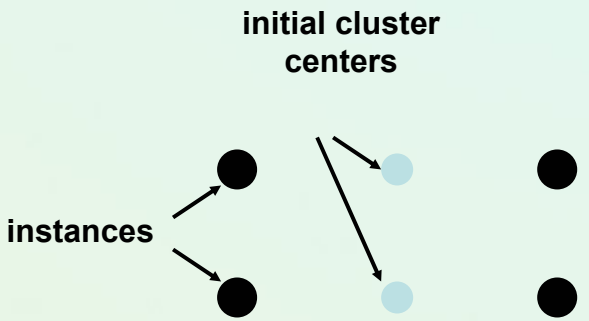
$$k_{\min} \leq k \leq k_{\max}$$

Możliwe są różne podejścia:

1. Arbitralny sposób np. przyjmuje się współrzędne pierwszych *k* obiektów (nie zawierające braków danych) 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

K-means krótkie podsumowanie

Zalety

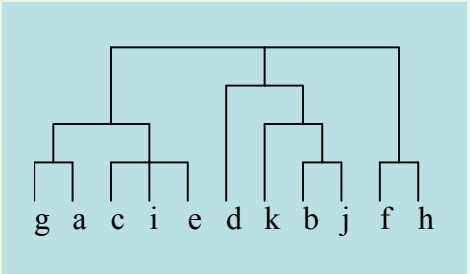
- Proste i łatwe do zrozumienia
- Reprezentacja skupień jako centroidy

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*

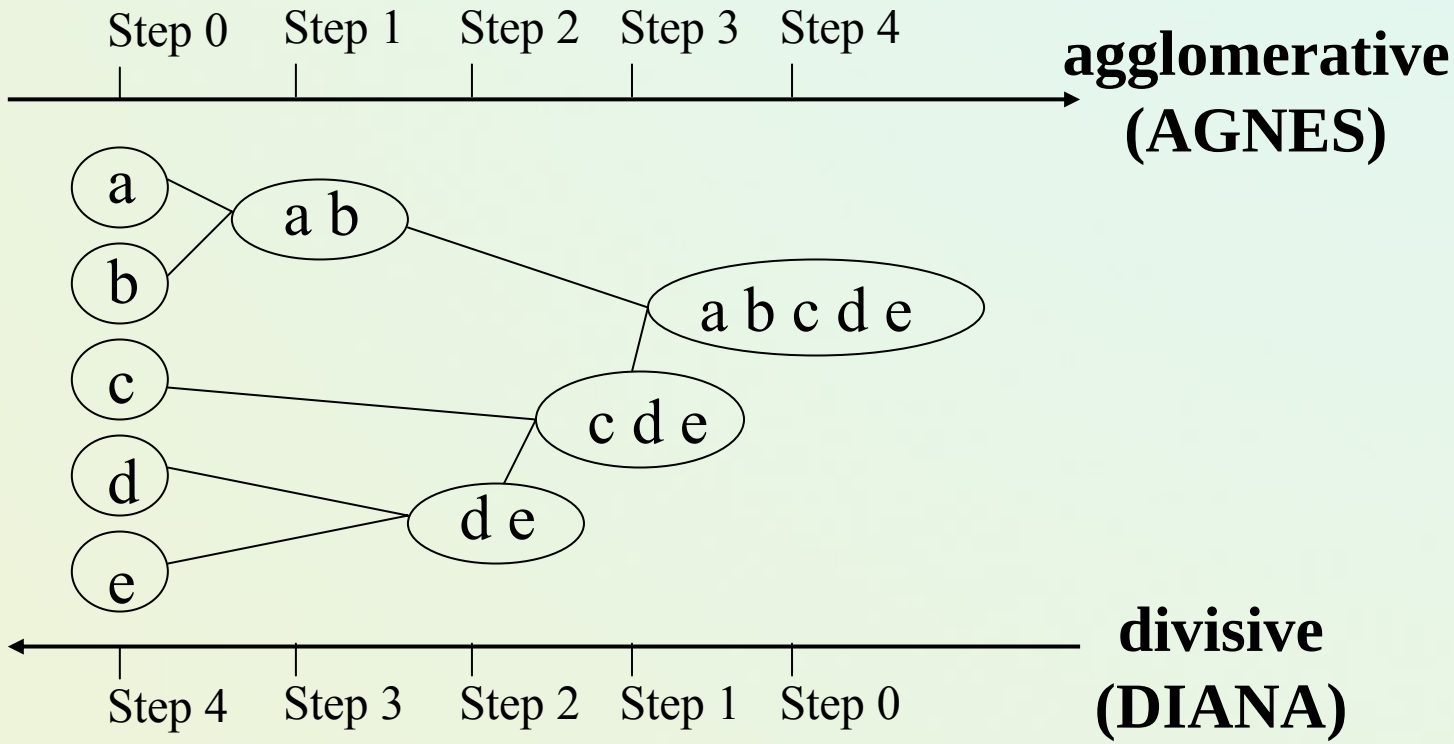


Hierarchiczne metody aglomeracyjne - algorytm

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.

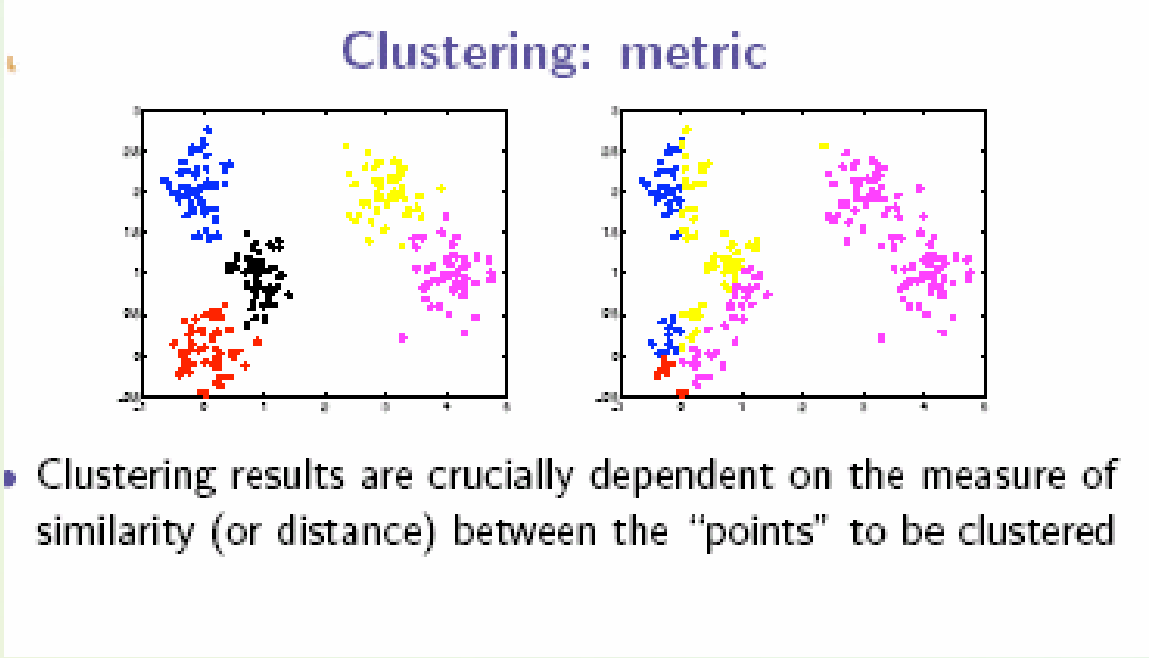
AHC - komentarz

- 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



Problem doboru miary odległości / podobieństwa

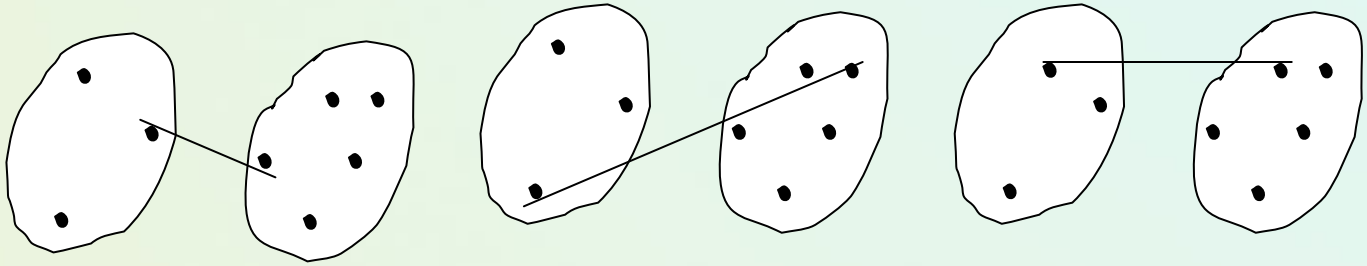
- Nietrywialny i silnie wpływa na wynik
- Różne miary odległości



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*)

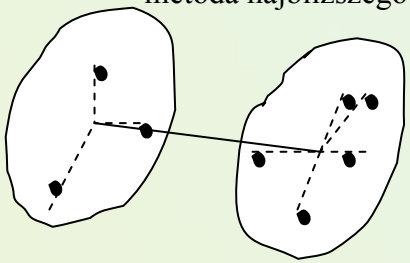
Porównanie sposobu wyznaczania odległości między skupieniami w wybranych metodach aglomeracyjnych



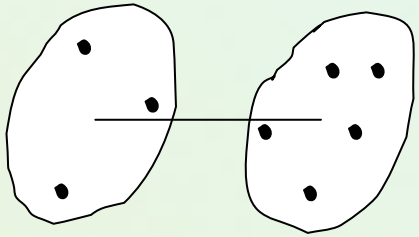
metoda najbliższego sąsiedztwa

metoda najdalszego sąsiedztwa

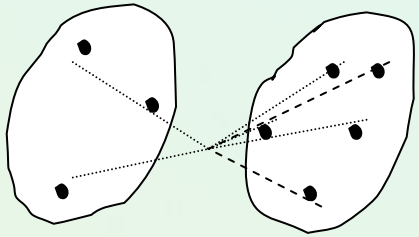
metoda mediany



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

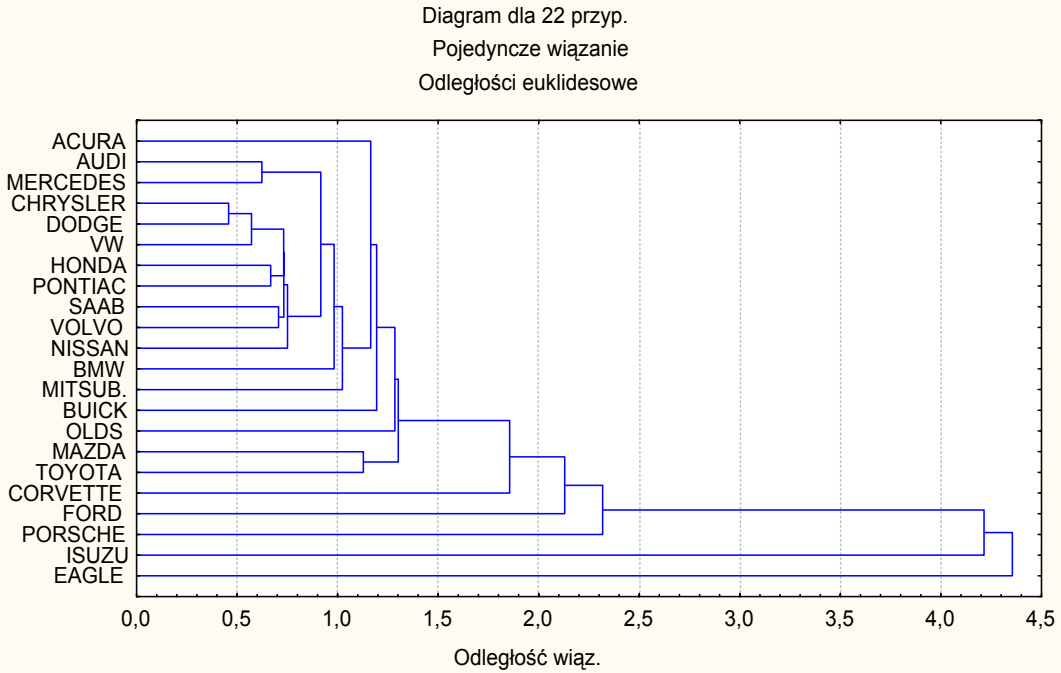
Complete Link Agglomerative Clustering

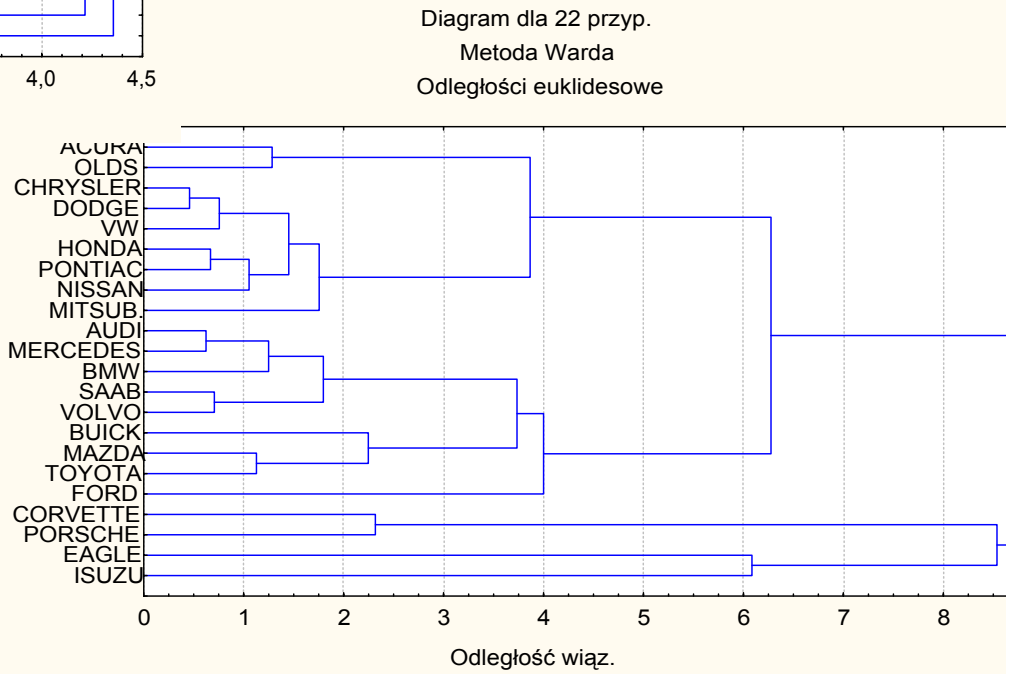
- Użyj maksymalnej odległości – minimalnego podobieństwa

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

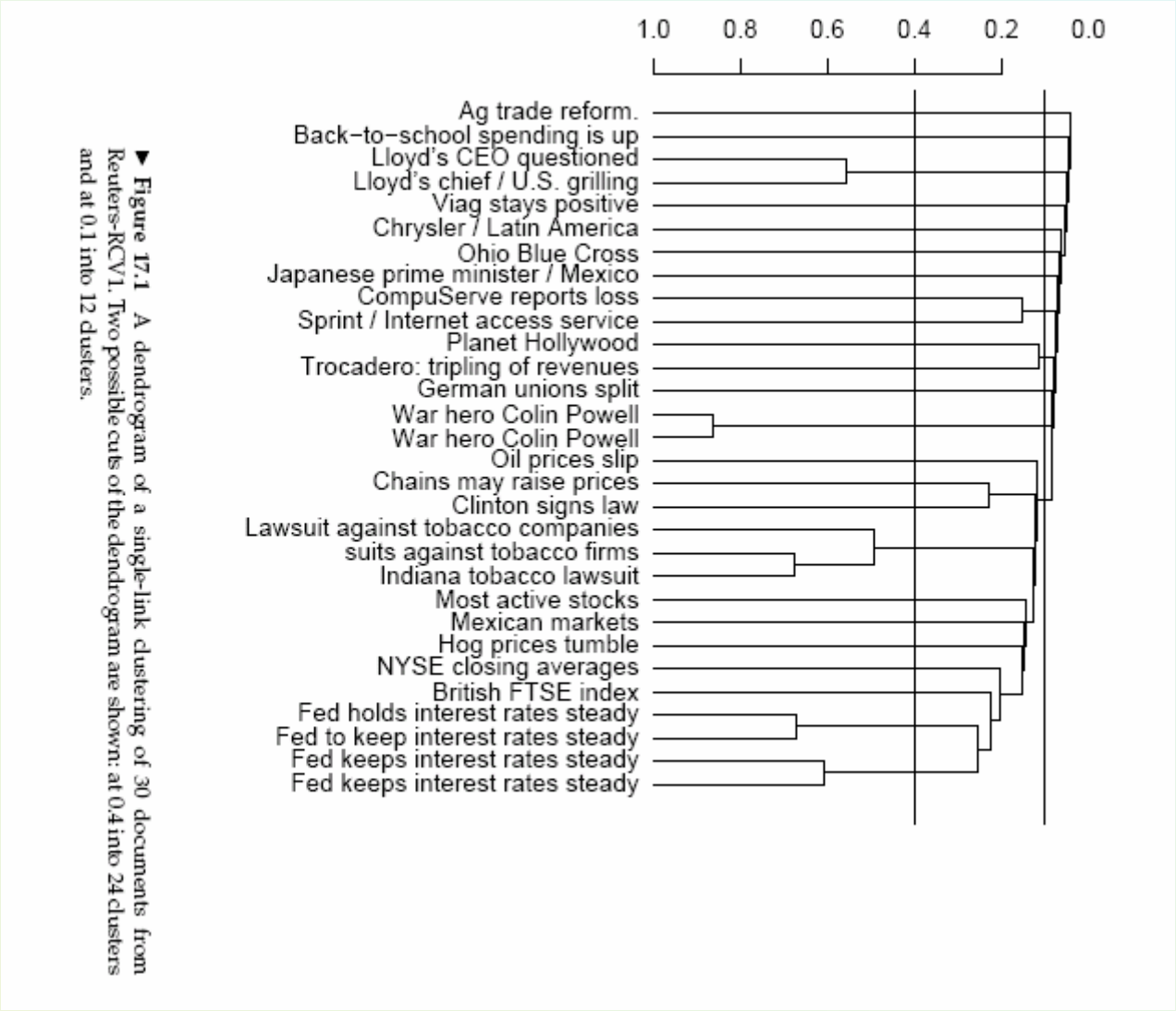
- Ukierunkowana do “tight,” spherical clusters
- Metoda zalecana gdy, kiedy obiekty faktycznie formują naturalnie oddzielone "kępki". Metoda ta nie jest odpowiednia, jeśli skupienia są w jakiś sposób wydłużone lub mają naturę "łańcucha".

Wrażliwość na dobór metod łączenia skupień



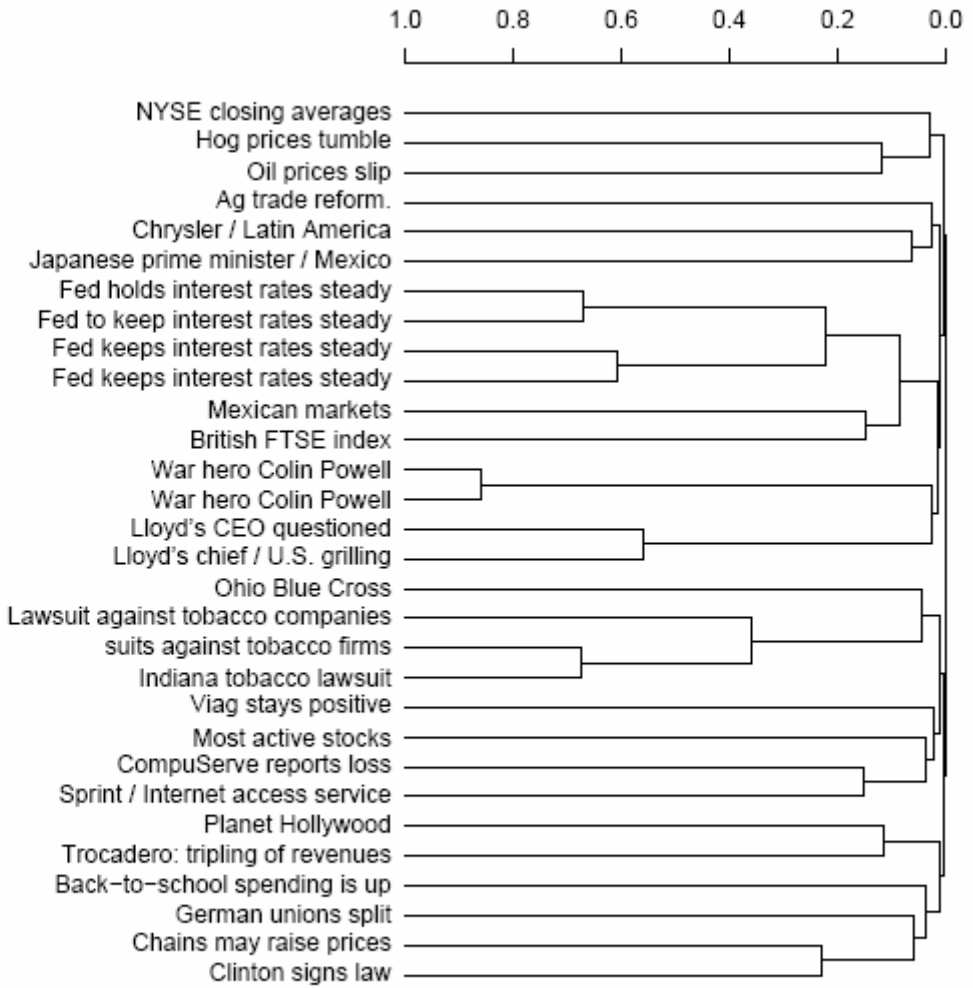


Maning – texts – single linkage



▶ 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.

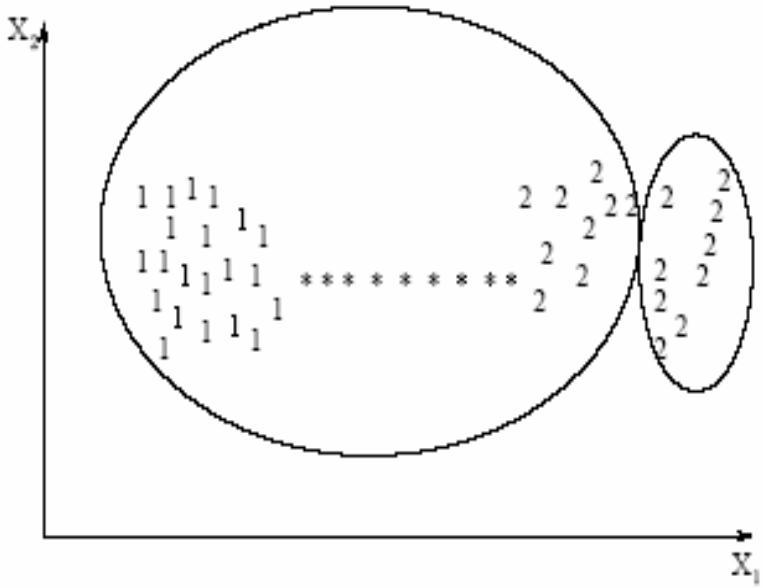


Figure 12. A single-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

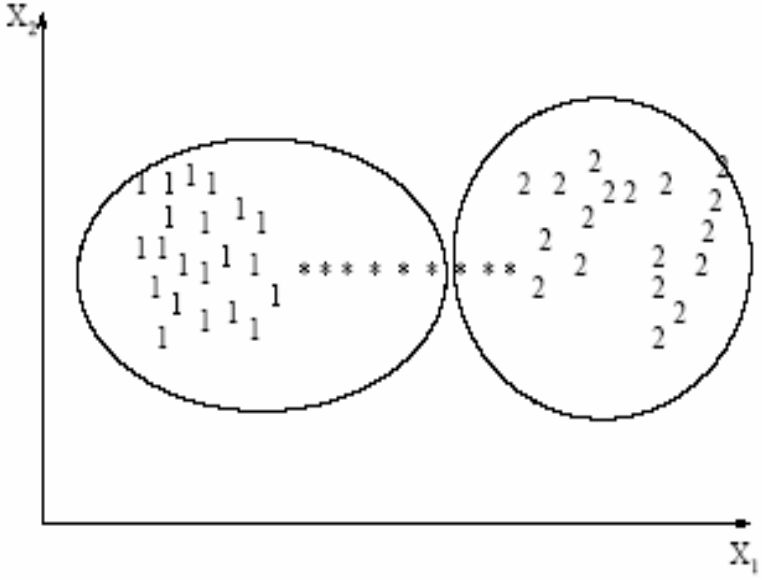
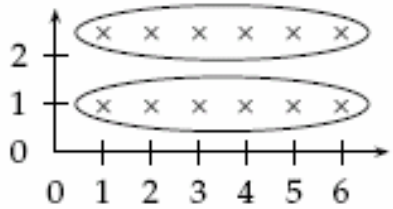
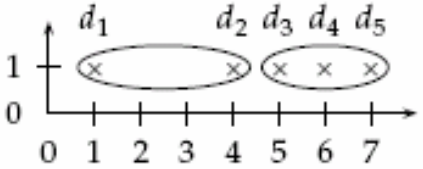


Figure 13. A complete-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

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 skupień.

Metody łączenia cd. Ward

- 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

Ocena jakości skupień

- Ocena ekspercka
- „Benchmarking on existing labels”
  - Comparing clusters with ground-truth categories (partition P)
- Miary oceny oparte na danych (internal measures)
  - Oparte na odległościach lub …
  - Duże podobieństwo obiektów wewnątrz skupienia (*Compactness*)
  - Same skupienia dość odległe (*Isolation*)

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]

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

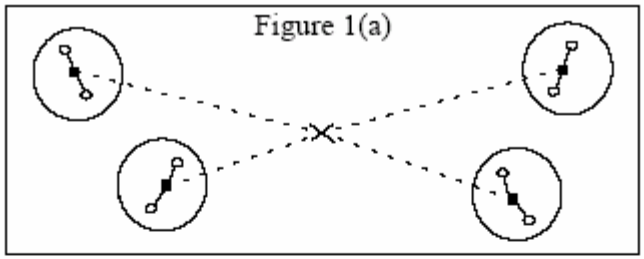
- Compactness $\rightarrow$ determining the weakest connection within the cluster, i.e., the largest distance between two objects *Ri* and *Rk* within the cluster.
- **Isolation** $\rightarrow$ determining the strongest connection of a cluster to another cluster, i.e., the smallest distance between a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



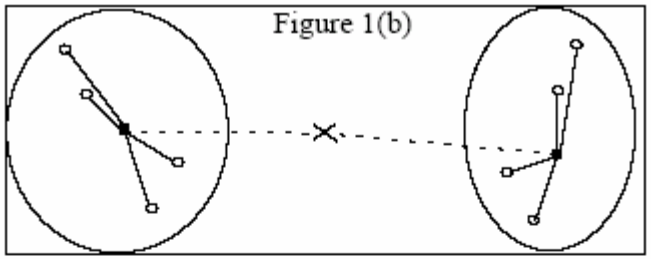
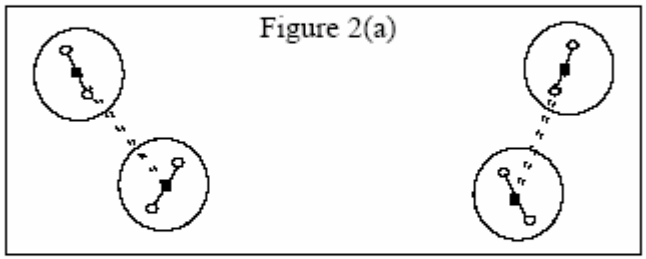


Figure 1: Minimum Total Distance Criterion



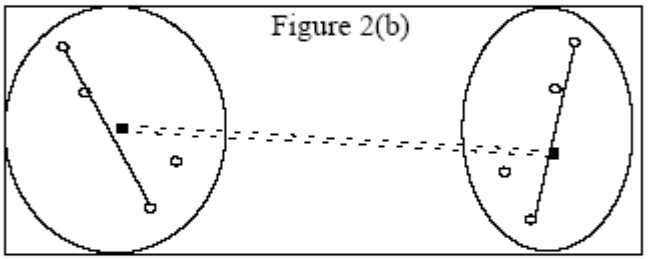
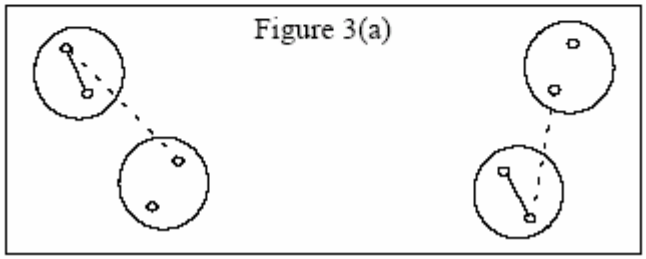


Figure 2: Separated Clusters Criterion



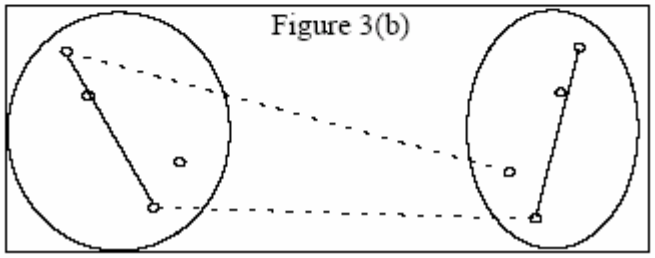


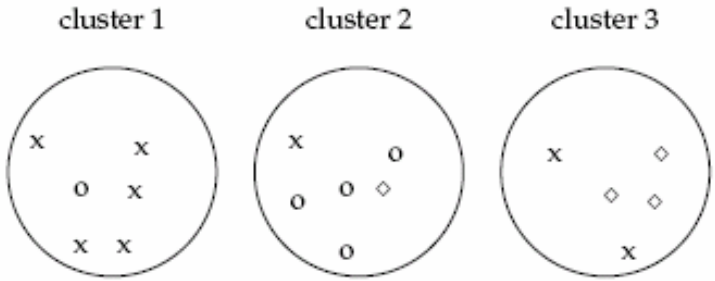
Figure 3: Object Positioning Criterion

Odniesienie do zewnętrznego podziału

- Różne podejścia
  - Measured by manually labeled data
    - We manually assign tuples into clusters according to their properties (e.g., professors in different research areas)
  - Accuracy of clustering: Percentage of pairs of tuples in the same cluster that share common label
    - This measure favors many small clusters
    - We let each approach generate the same number of clusters

Oceny poprzez porównanie do zbioru referencyjnego (ground truth)

- Przykład tekstowy



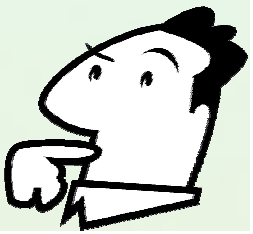
► **Figure 16.4** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ◇, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

	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.

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)



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

LICZBOWE WARTOŚCI

Cena, wydajność, trzymanie się drogi różny

	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

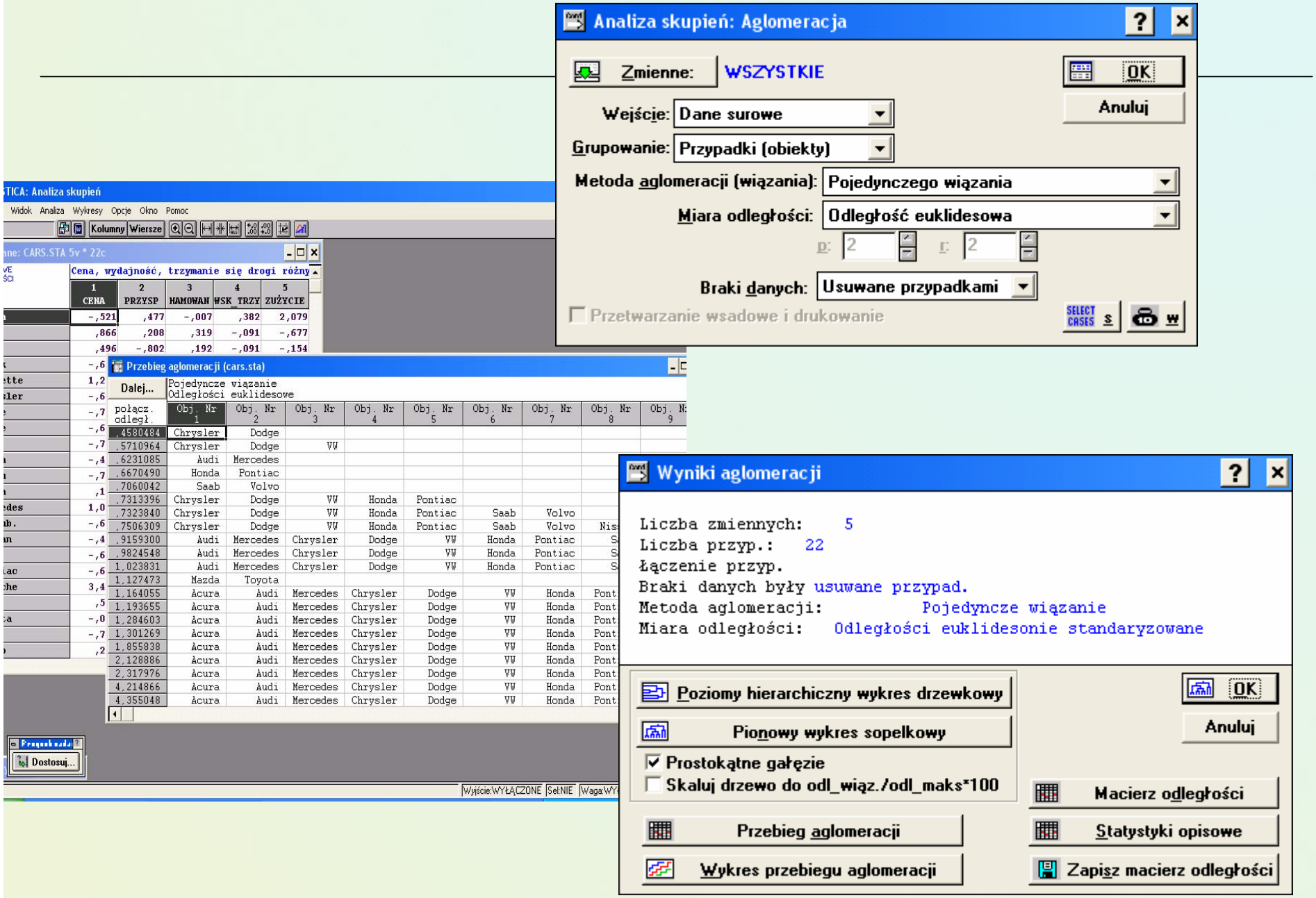
Metoda grupowania

- Aglomeracja
- Grupowanie metodą k-średnich
- Grupowanie obiektów i cech

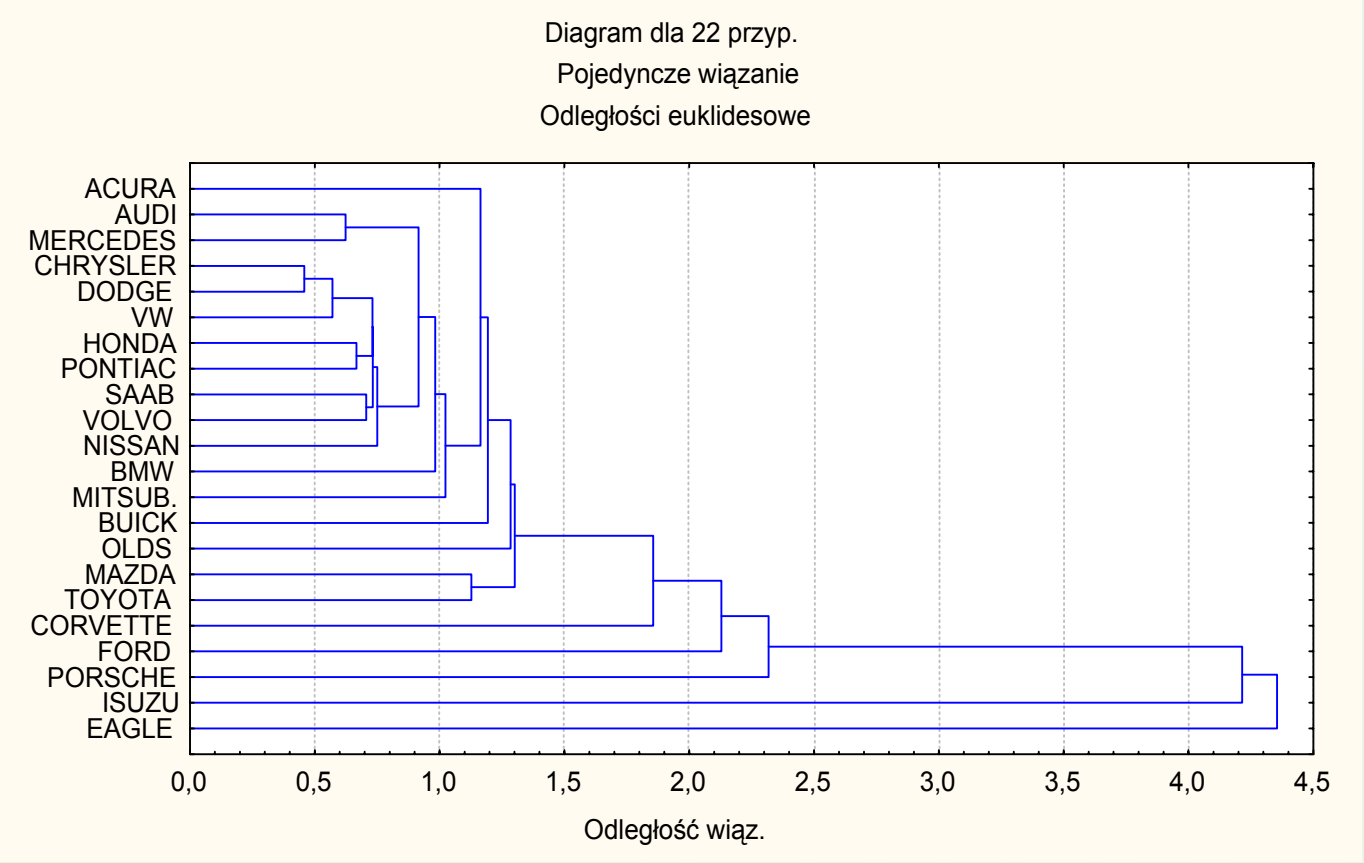
OK

Anuluj

Otwórz dane

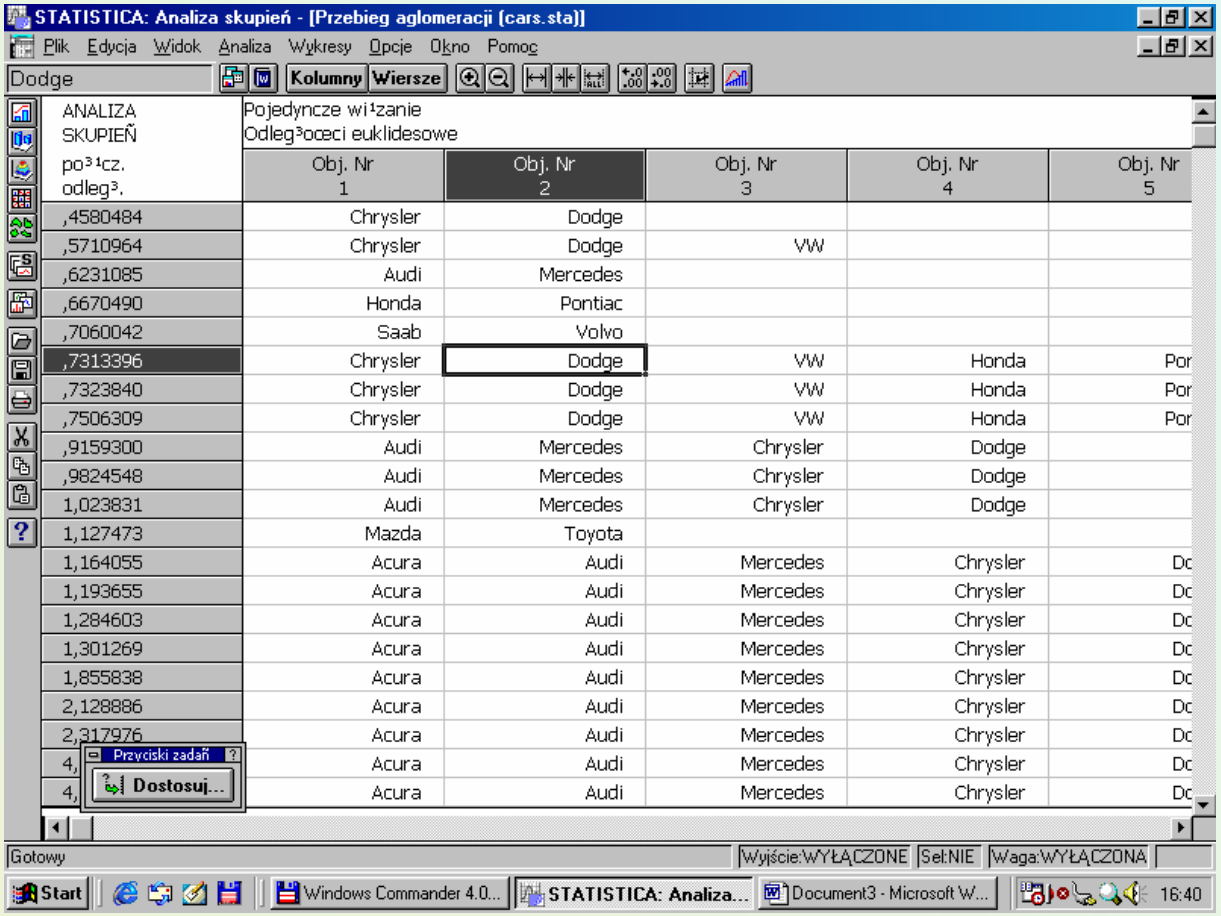


Dendrogram for Single Linkage



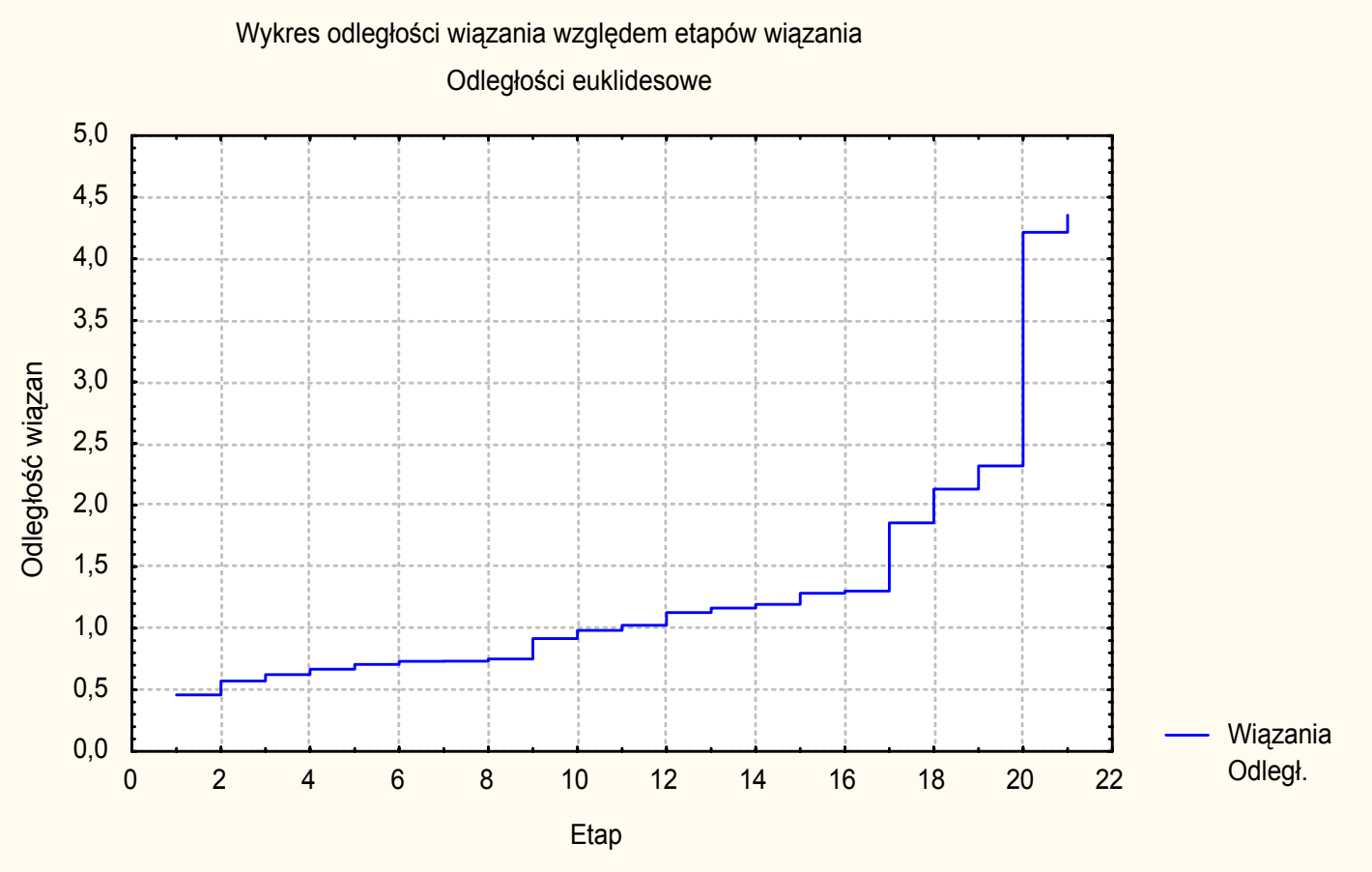
Opis tworzenia dendrogramu

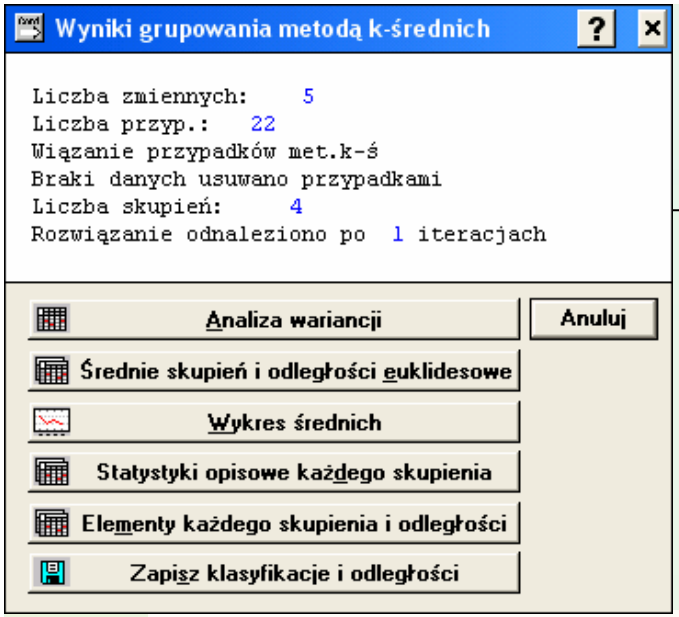
- Łączenie obiektów w kolejnych krokach

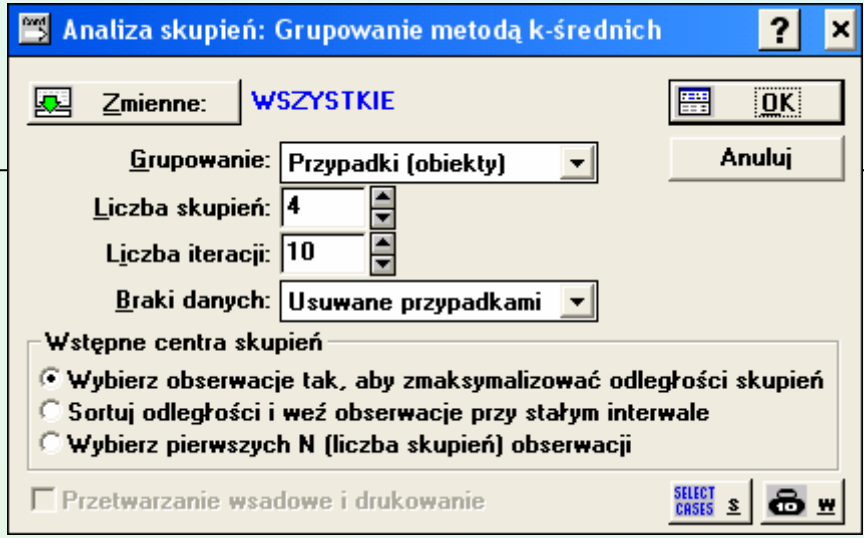


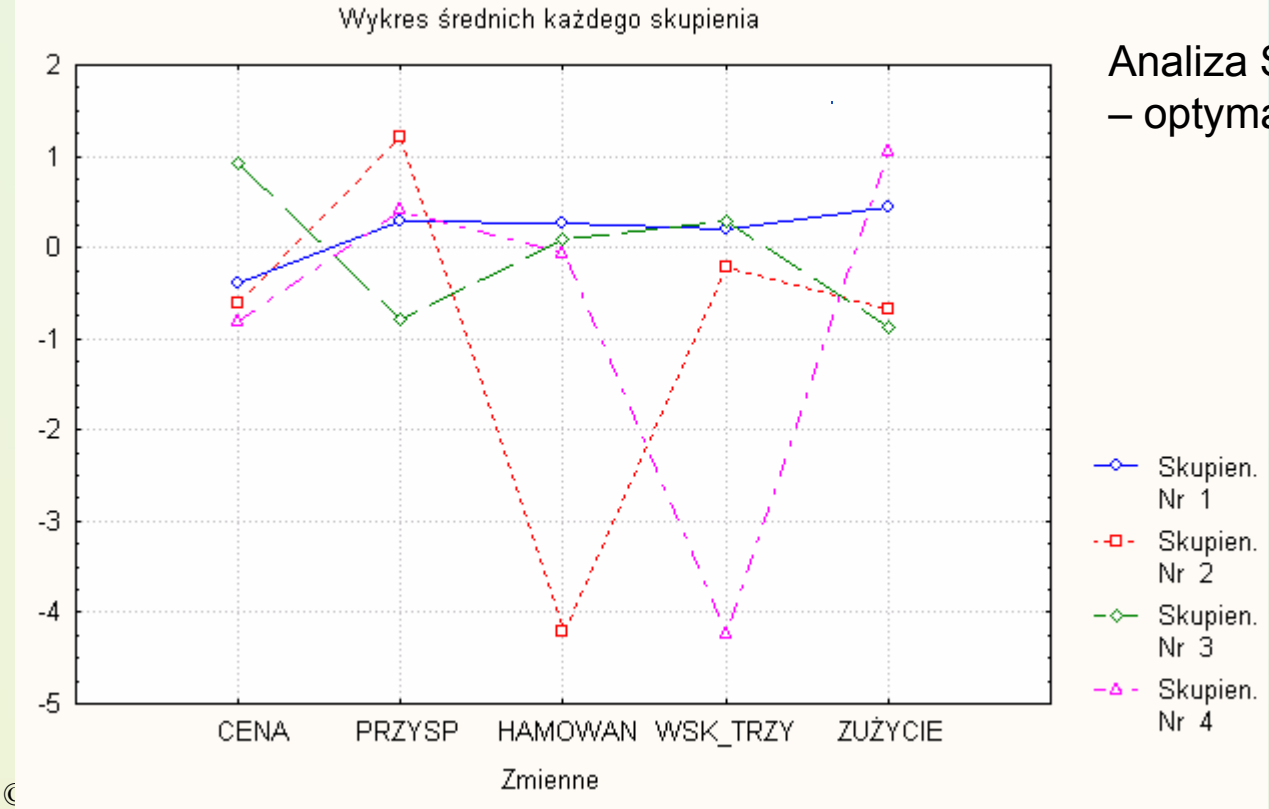
Analiza procesu łączenia

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



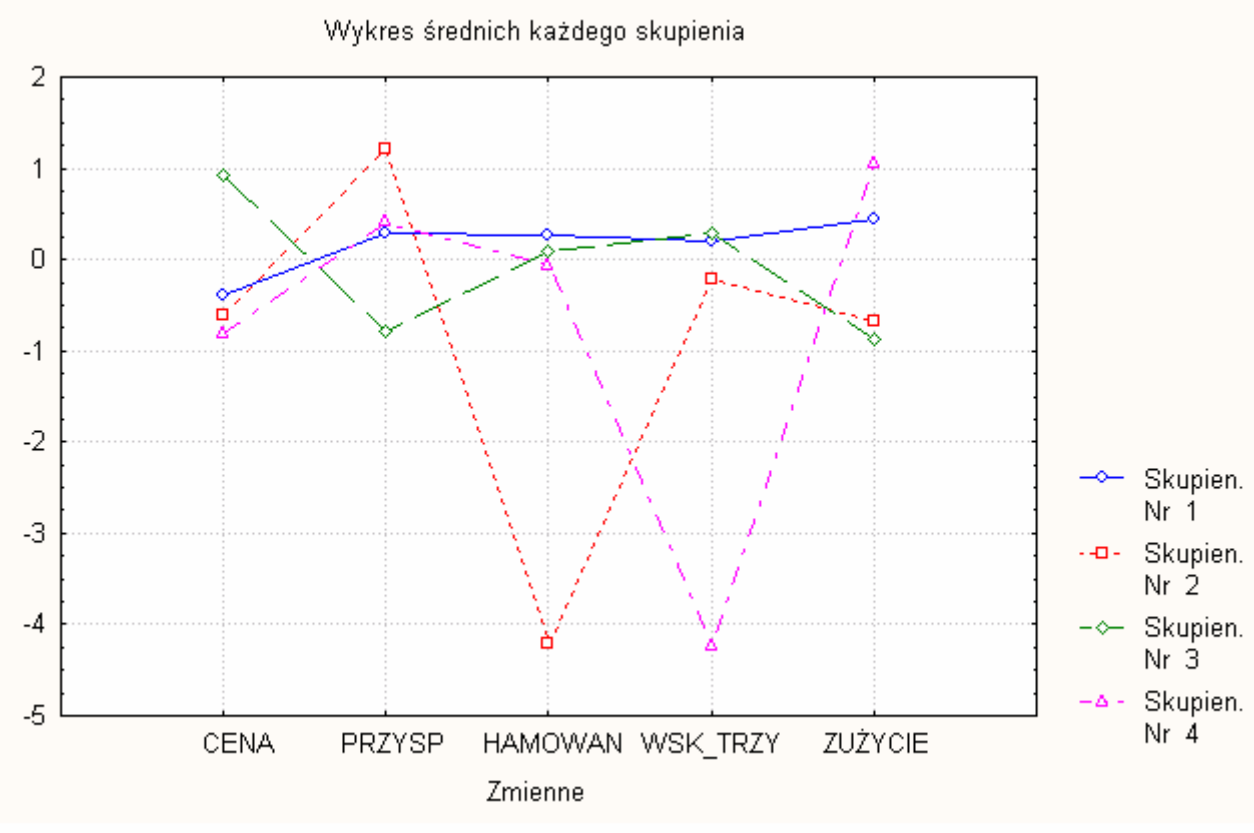






Analiza Skupień – optymalizacja k-średnich

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

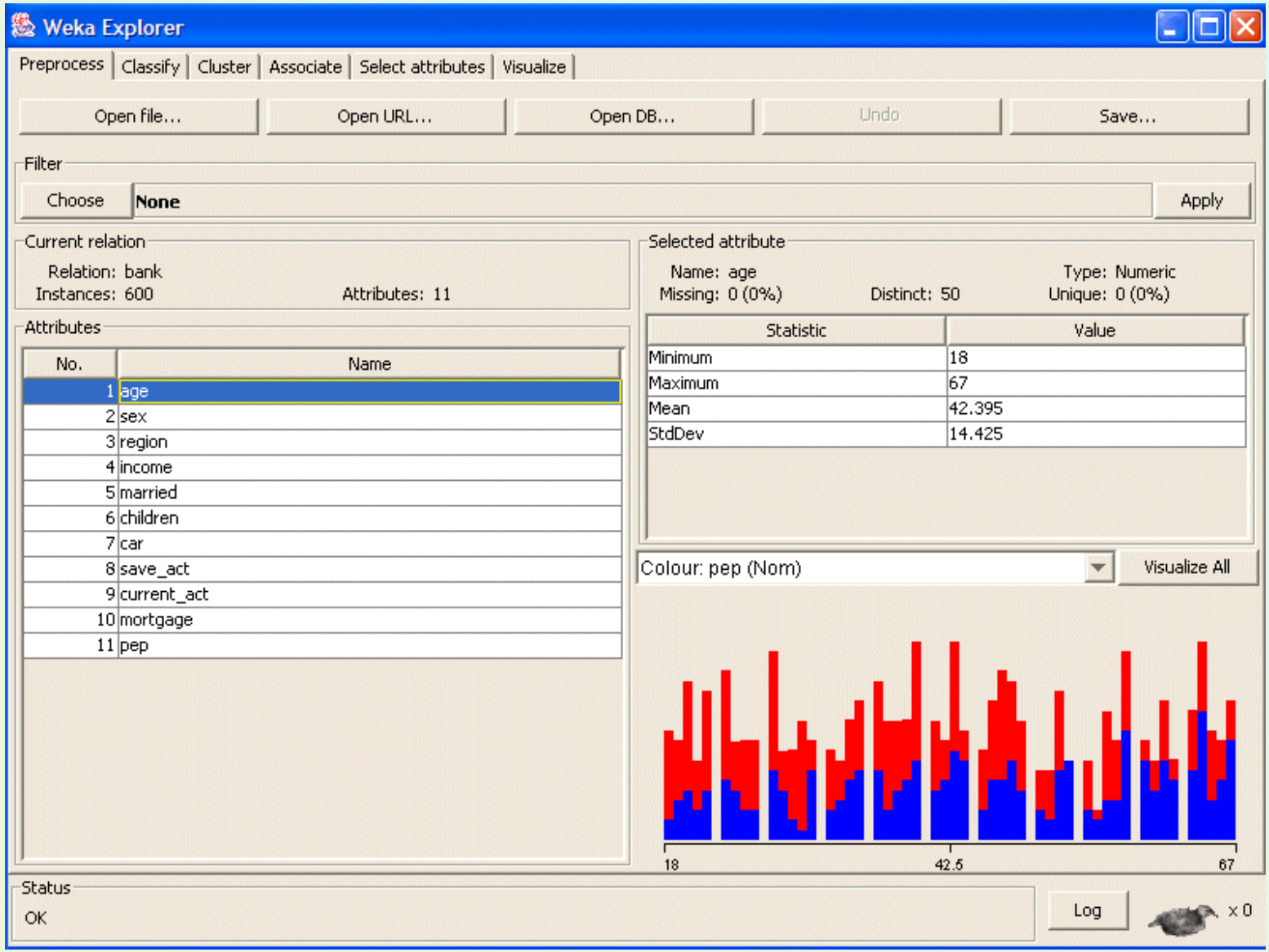
Exercise 1. K-means clustering in WEKA

- 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

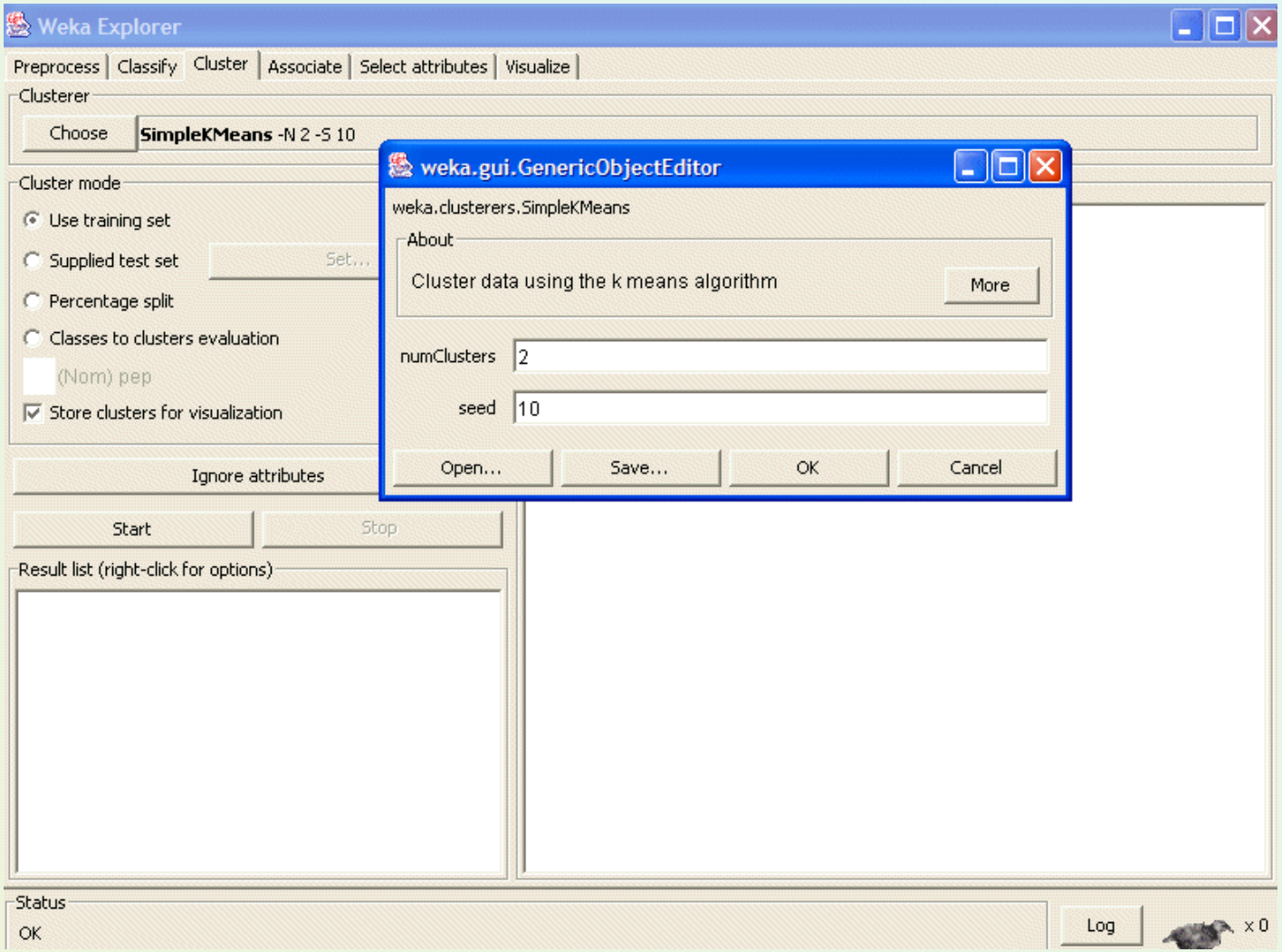


Preprocessing for clustering

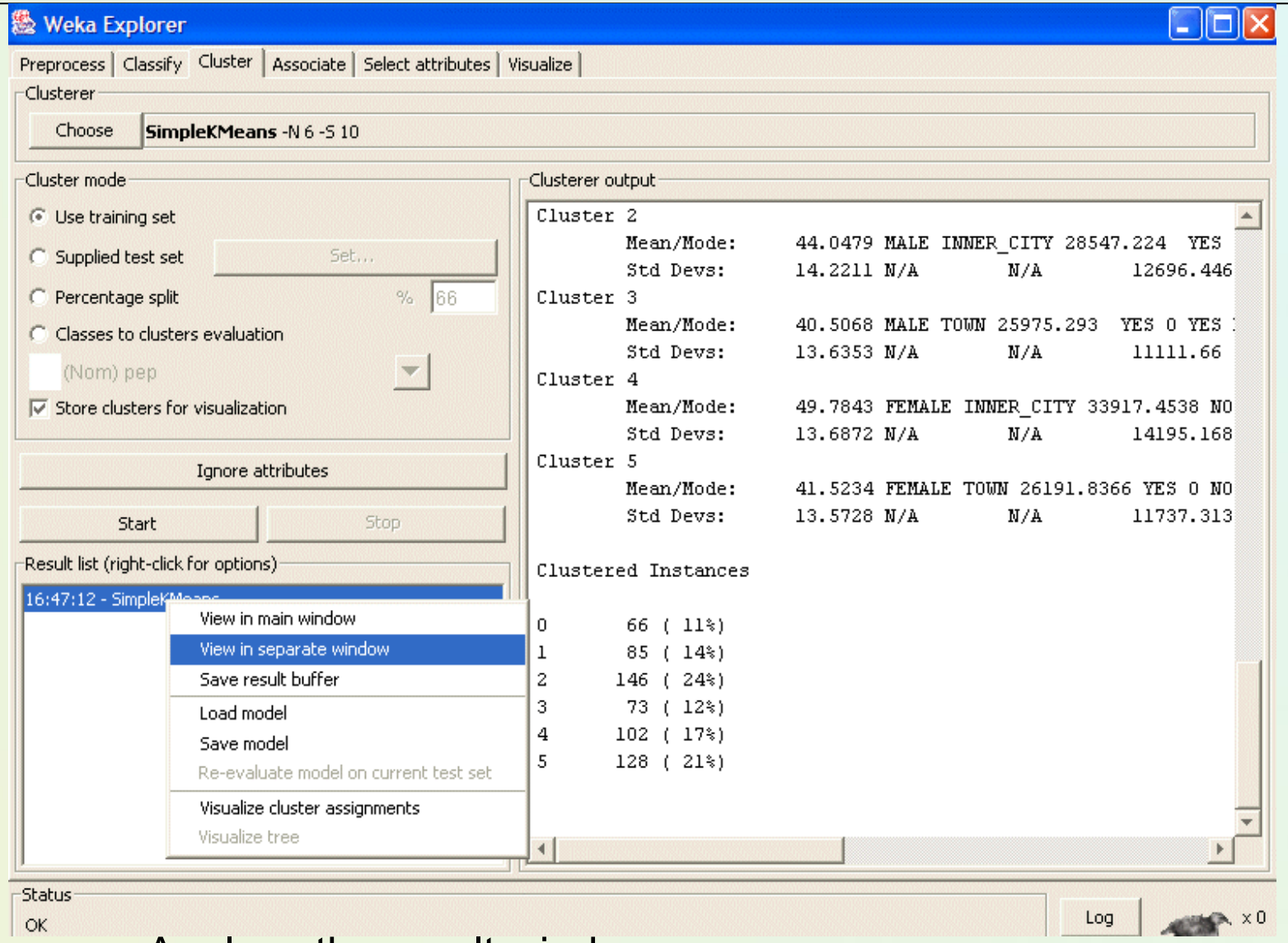
- What about non-numerical attributes?
  - Remember about Filters
- Should we normalize or standarize attributes?
- How it is handled in WEKA k-means?

Choosing Simple k-means

- Tune proper parameters



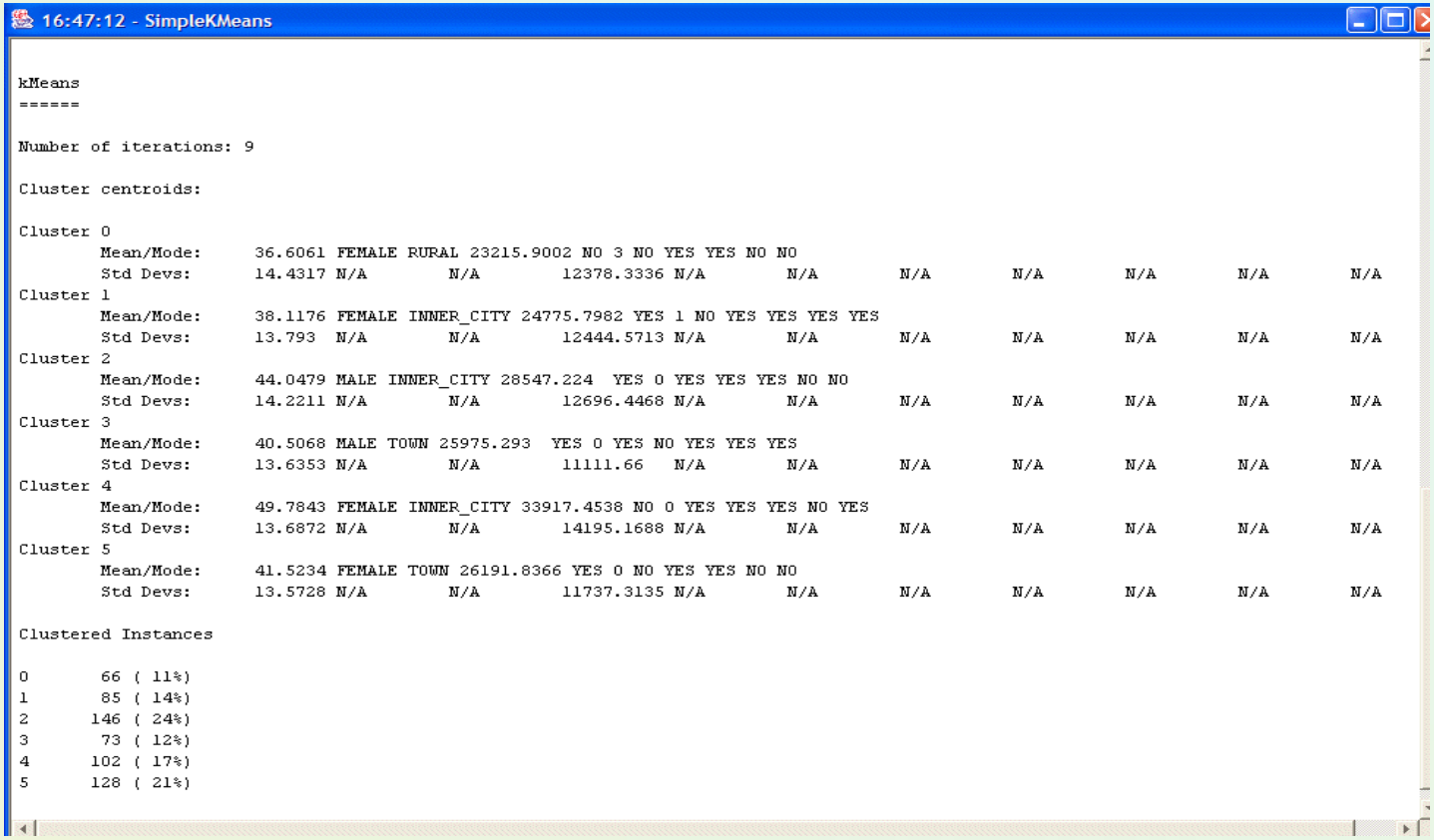
Clustering results



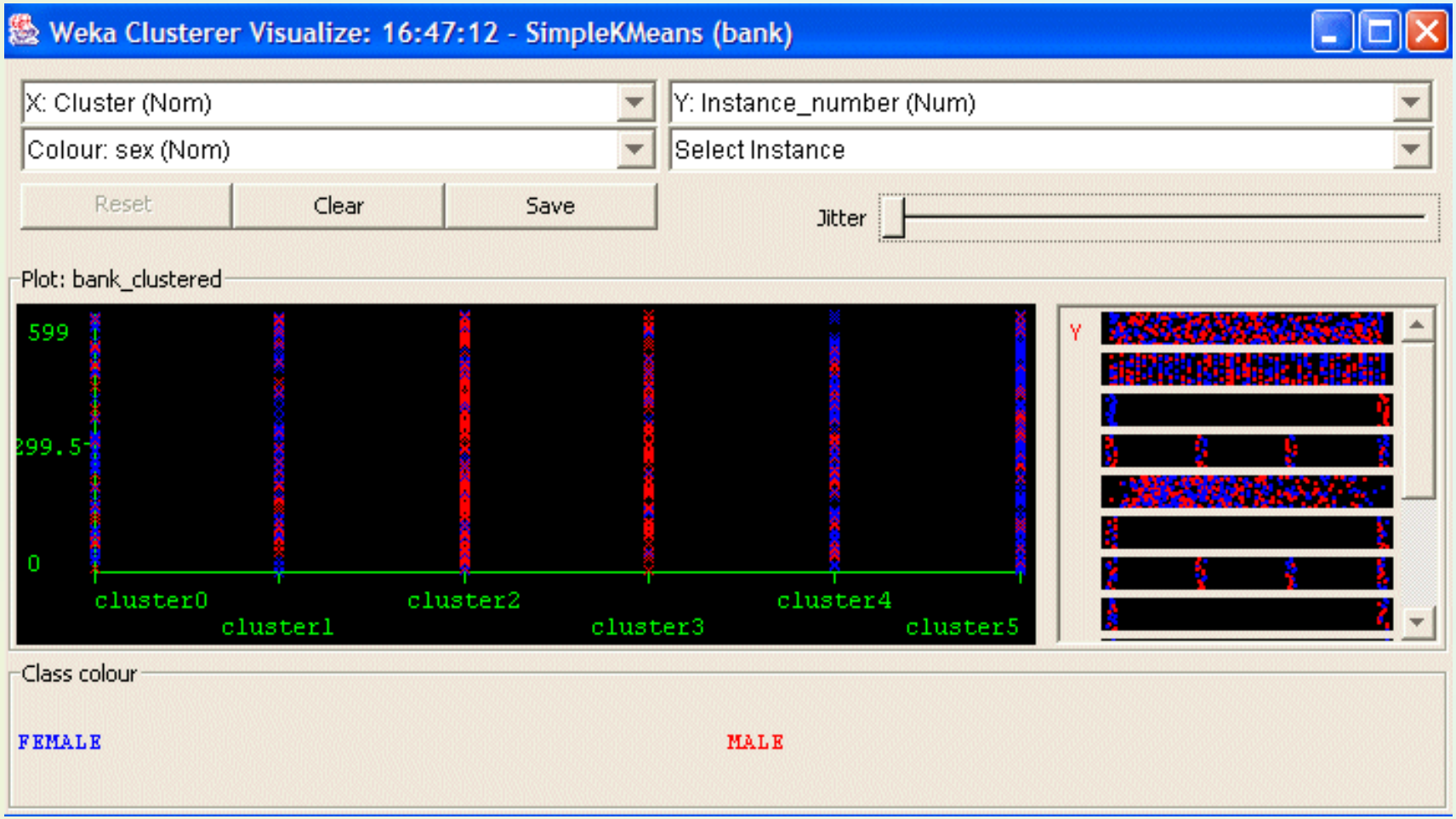
- Analyse the result window

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
```

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).

Pamiętaj o doborze cech

- Za Jain's tutorial

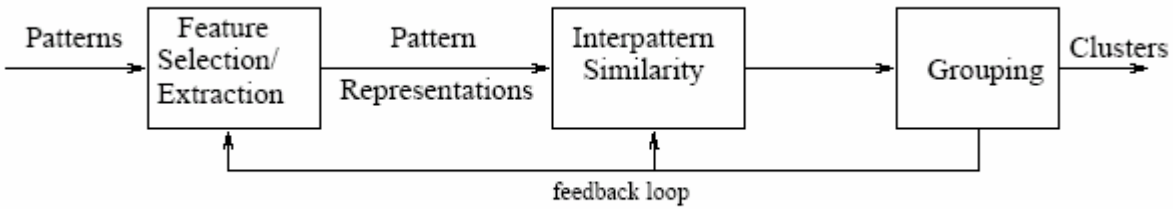


Figure 2. Stages in clustering.

Algorytmy analizy skupień dla baz danych o wielkich rozmiarach

- Scalability
- Dealing with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Interpretability and usability.

Algorytm *K-Medoids*

- Znajdź *medoidy* - obiekty reprezentujące skupienia
  - Wykorzystuje się odległości par obiektów.
- *PAM* (Partitioning Around Medoids, 1987)
  - Zacznij od początkowego zbioru medoidów i krokowo zamieniaj jeden z obiektów - medoidów przez obiekt, nie będący aktulanie medoidem, jeśli to polepsza całkowitą odległość skupień.
  - *PAM* – efektywny dla małych zbiorów, lecz nieskalowalny dla dużych zbiorów przykładów.
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): „Randomized sampling”.

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.

BIRCH - Clustering Feature Vector

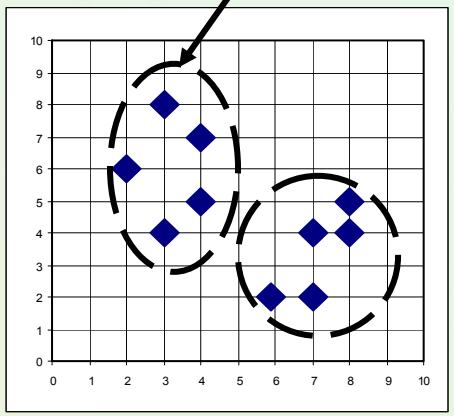
Clustering Feature: $CF = (N, \overrightarrow{LS}, SS)$

N: **liczba grupowanych obiektów**

$LS: \sum_{i=1}^{N} = \overrightarrow{X_i}$

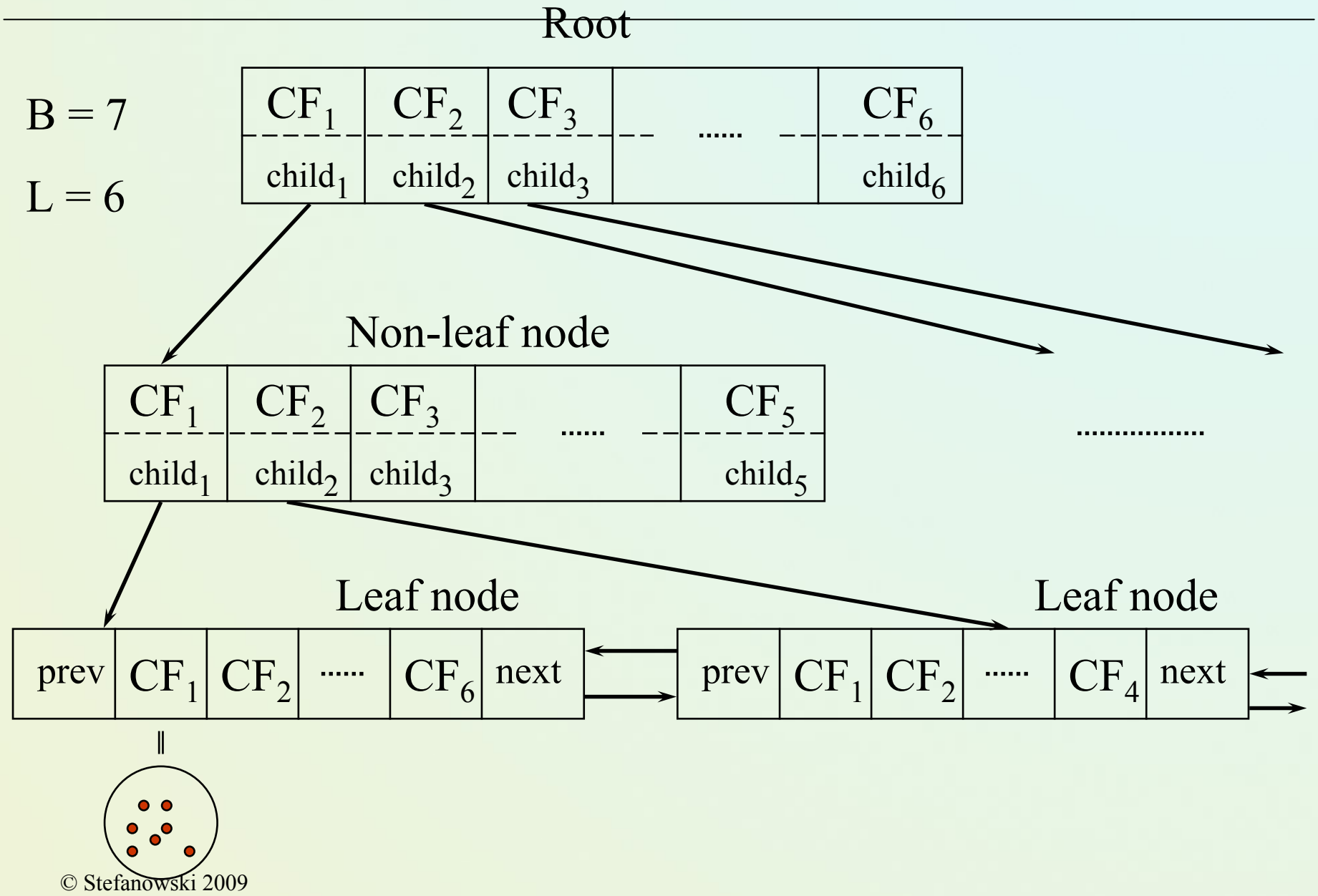
$SS: \sum_{i=1}^{N} = \overrightarrow{X_i^2}$

CF = (5, (16,30),(54,190))



(3,4)
(2,6)
(4,5)
(4,7)
(3,8)

CF Tree



Wstawianie elementów do drzewa CF

- Parametery drzewa CF:
  - „Branching factor” B – max. liczba potomków w węźle (non-leaf node),
  - „Threshold” T – max. średnica podskupienia skojarzonego z liściem drzewa.
- Kolejno wczytywany obiekt przydziel do najbliższego podskupienia (leaf entry)
  - zgodnie z wybraną miarą odległości
- Dokonaj modyfikacji liści
  - jeśli średnica liścia > T, to dokonaj podziału (być może także inne liście powinny być podzielone)
- Dokonaj modyfikacji ścieżki do liścia
  - Uaktualnij wektory CF w węzłach drzewa. Dokonaj podziału węzłów jeśli to konieczne.
- Jeżeli rozmiar drzewa CF jest zbyt duży w stosunku do dostępnej pamięci operacyjnej, dokonuje się modyfikacji parametrów algorytmu.

BIRCH: dalsze szczegóły algorytmu

1: Utwórz drzewo CF w pamięci operacyjnej

- przeczytaj przyrostowo DB i buduj początkową strukturę drzew

2: (Opcjonalnie) Dokonaj kompresji poprzez budowę mniejszego drzewa CF zmieniając zakres parametrów

- przejrzyj „leaf entries” w początkowym drzewie CF w celu zbudowania mniejszego drzewa.

3: Globalne grupowanie

- grupuj wykorzystując „leaf entries” drzewa

4: (opcjonalne, „off line”) Ulepszanie struktury skupień

- Użyj centroidów z etapu 3, rozdzielając obiekty

Complexity: ***O(N) czy O(N^2)?***

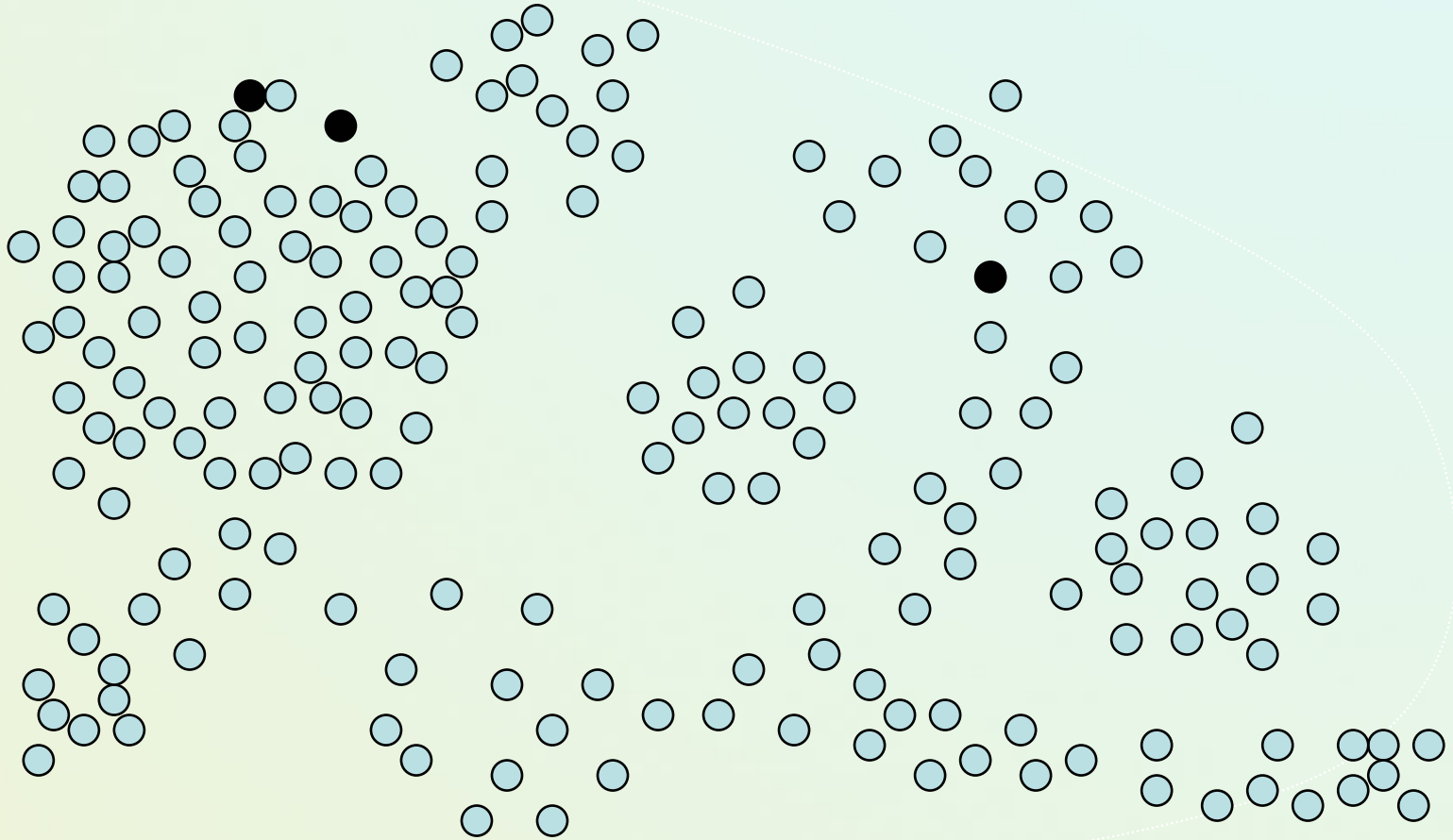
Metody gęstościowe

- 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)

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

Bardziej nieregularne kształty



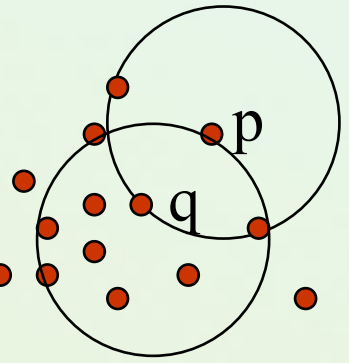
DBSCAN: Podstawowe pojęcia

- 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:

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

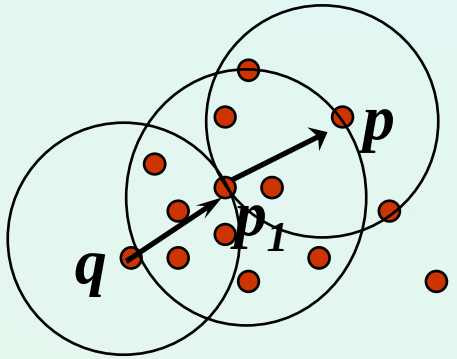


MinPts = 5

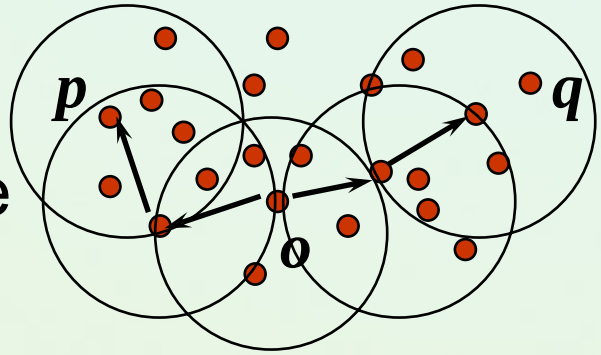
Eps = 1 cm

DBSCAN: Podstawowe pojęcia (II)

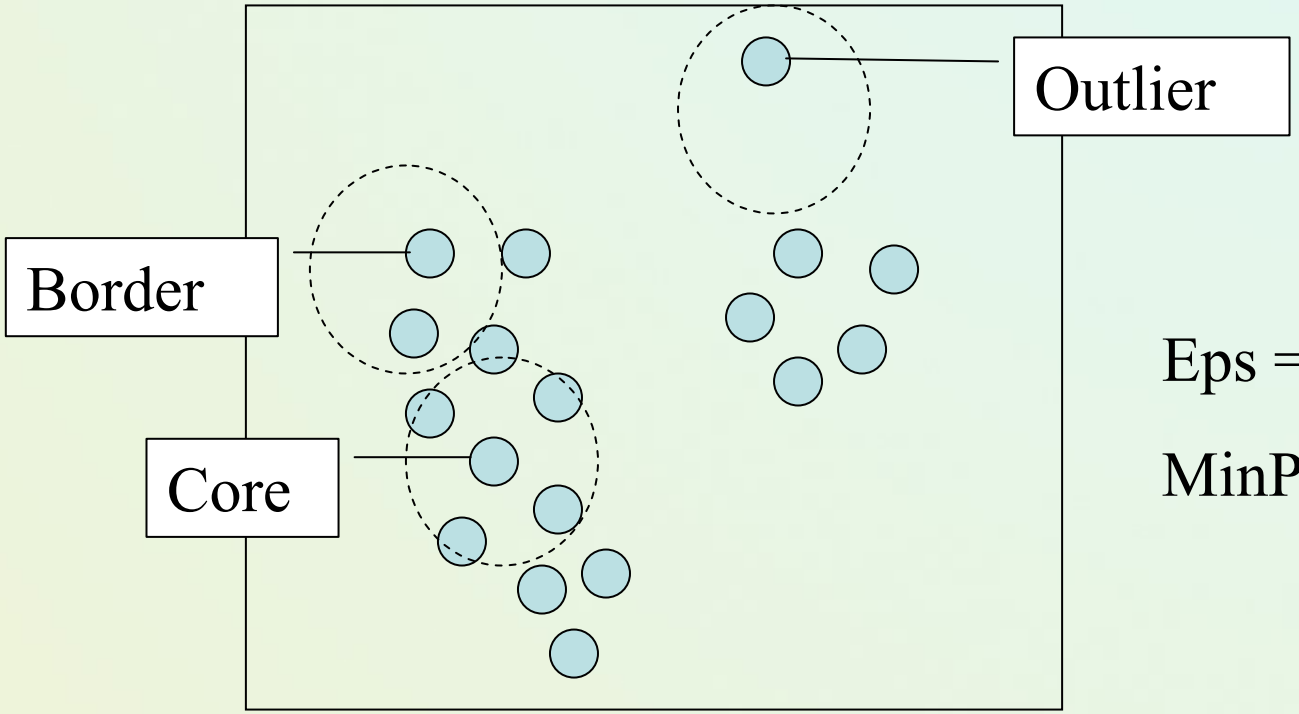
- Density-reachable:
  - A point *p* is density-reachable from a point *q* wrt. *Eps*, *MinPts* if there is a chain of points p_1, …, p_n, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i



- 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: jak poszukuje skupień?



Eps = 1cm

MinPts = 5

DBSCAN: Zarys algorytmu

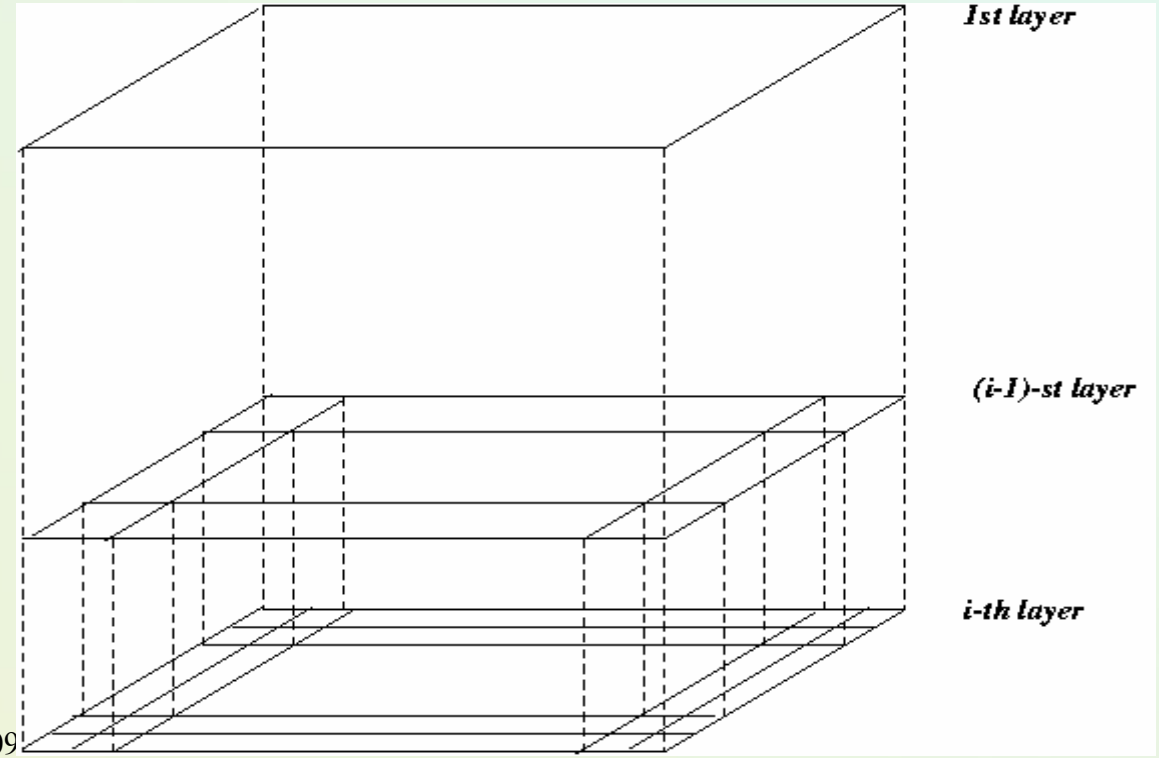
- Wybierz punkt startowy ***p***
- Odnajdź wszystkie punkty do gęstościowego osiągnięcia z *p* (density-reachable from ***p*** wrt ***Eps*** and ***MinPts)***.
- Jeśli ***p*** jest rdzeniem (*core point*), utwórz skupienie.
- Jeśli ***p*** jest punktem granicznym (border point) i żadne punkty nie są 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).

Grid-Based Clustering Method

- Using multi-resolution grid data structure
- Several interesting methods
  - STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
    - A multi-resolution clustering approach using wavelet method
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

STING: A Statistical Information Grid Approach

- Wang, Yang and Muntz (VLDB'97)
- The spatial area area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution

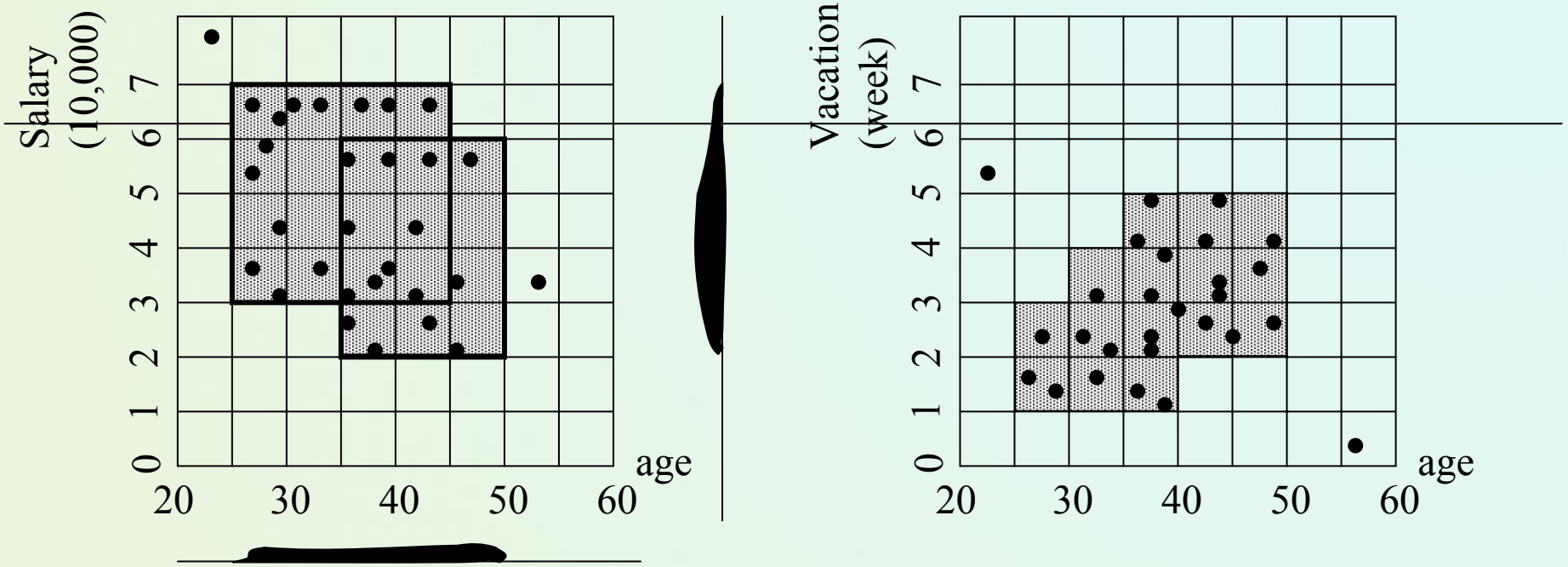


CLIQUE (Clustering In QUEst)

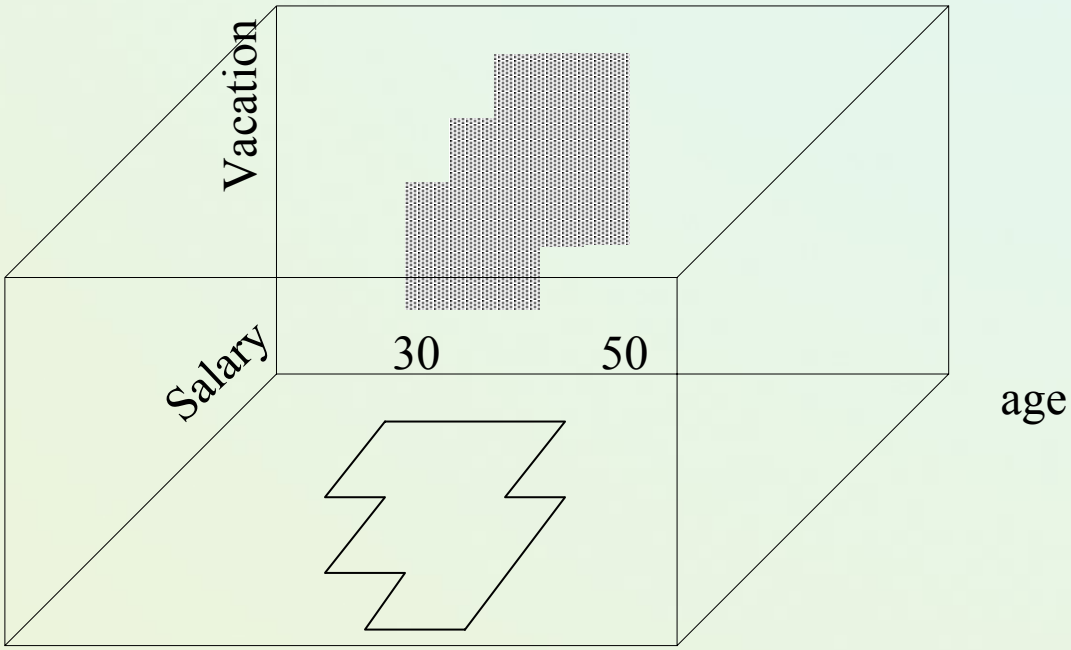
- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98).
- Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space
- CLIQUE can be considered as both density-based and grid-based
  - It partitions each dimension into the same number of equal length interval
  - It partitions an m-dimensional data space into non-overlapping rectangular units
  - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
  - A cluster is a maximal set of connected dense units within a subspace

CLIQUE: The Major Steps

- Partition the data space and find the number of points that lie inside each cell of the partition.
- Identify the subspaces that contain clusters using the Apriori principle
- Identify clusters:
  - Determine dense units in all subspaces of interests
  - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters
  - Determine maximal regions that cover a cluster of connected dense units for each cluster
  - Determination of minimal cover for each cluster



$\tau = 3$

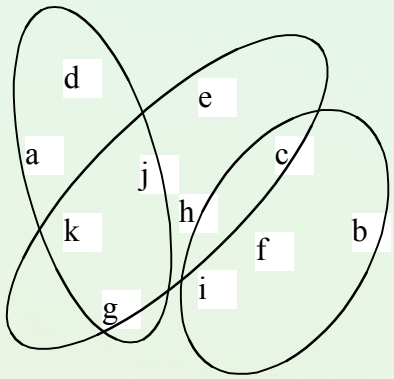


Strength and Weakness of *CLIQUE*

- Strength
  - It *automatically* finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces
  - It is *insensitive* to the order of records in input and does not presume some canonical data distribution
  - It scales *linearly* with the size of input and has good scalability as the number of dimensions in the data increases
- Weakness
  - The accuracy of the clustering result may be degraded at the expense of simplicity of the method

Soft Clustering

- Clustering typically assumes that each instance is given a "hard" assignment to exactly one cluster.
- Does not allow uncertainty in class membership or for an instance to belong to more than one cluster.
- *Soft clustering* gives probabilities that an instance belongs to each of a set of clusters.
- Each instance is assigned a probability distribution across a set of discovered categories (probabilities of all categories must sum to 1).



Expectation Maximization (EM Algorithm)

- Probabilistic method for soft clustering.
- Direct method that assumes *k* clusters:$\{c_1, c_2, \ldots c_k\}$
- Soft version of *k*-means.
- Assumes a probabilistic model of categories that allows computing $P(c_i \mid E)$ for each category, c_i, for a given example, *E*.
- For text, typically assume a naïve-Bayes category model.
  - Parameters $\theta = \{P(c_i), P(w_j \mid c_i): i \in \{1,\ldots k\}, j \in \{1,\ldots,|V|\}\}$

Expectation-Maximization (EM)

- Log likelihood with a mixture model

$$L(\Phi \mid X) = \log \prod_t p\left(\mathbf{x}^t \mid \Phi\right)$$

$$= \sum_t \log \sum_{i=1}^{k} p\left(\mathbf{x}^t \mid G_i\right) P(G_i)$$

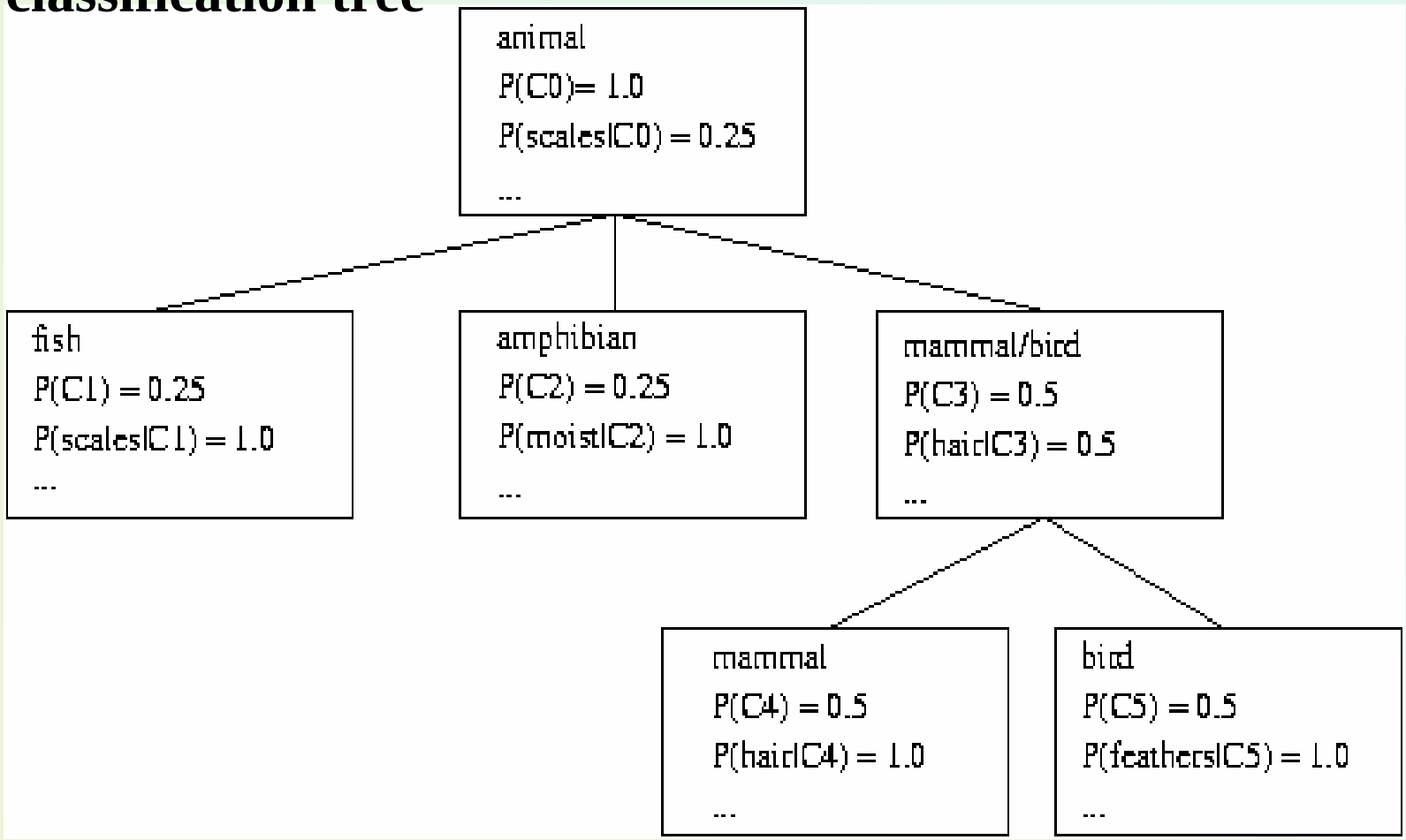
- Assume hidden variables *z*, which when known, make optimization much simpler
- Complete likelihood, $L_c(\Phi \mid X,Z)$, in terms of ***x*** and ***z***
- Incomplete likelihood, $L(\Phi \mid X)$, in terms of ***x***

Model-Based Clustering Methods

- Attempt to optimize the fit between the data and some mathematical model
- Statistical and AI approach
  - Conceptual clustering
    - A form of clustering in machine learning
    - Produces a classification scheme for a set of unlabeled objects
    - Finds characteristic description for each concept (class)
  - COBWEB (Fisher'87)
    - A popular a simple method of incremental conceptual learning
    - Creates a hierarchical clustering in the form of a classification tree
    - Each node refers to a concept and contains a probabilistic description of that concept

COBWEB Clustering Method

A classification tree



*Incremental clustering (COBWEB based)

- Heuristic approach (COBWEB/CLASSIT)
- Form a hierarchy of clusters incrementally
- Start:
  - tree consists of empty root node
- Then:
  - add instances one by one
  - update tree appropriately at each stage
  - to update, find the right leaf for an instance
  - May involve restructuring the tree
- Base update decisions on *category utility*

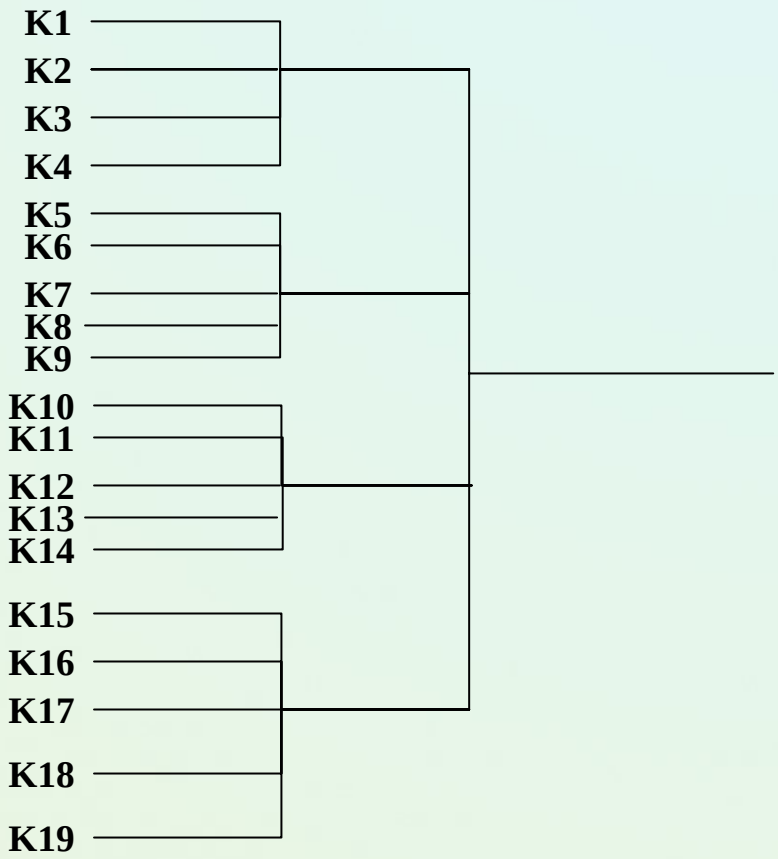
World countries data

Kraj	C1	C2	C3	C4	C5	C6	C7	C8	C9
Afganistan	M	AP	S	N	N	N	N	N	S
Argentyna	K	AL.	U	N	S	W	W	W	N
Armenia	O	SW	SM	S	S	W	W	W	N
Australia	P	OECD	S	N	S	W	W	W	N
Austria	K	OECD	U	N	W	W	W	W	N
Azerbejdżar	M	SW	S	N	W	W	W	W	N
Belgia	K	OCED	U	W	S	W	W	W	N
Białoruś	O	EW	U	N	W	W	S	S	N
Boliwia	K	A	SM	N	W	S	S	S	S
...	...	...	...	...	...	...	...	...	...

COBWB results

Selected classes

- K1: Rosja, Portugalia, Polska, Litwa, Łotwa, Węgry, Grecja, Gruzja, Estonia, Czechy, Chorwacja
- K2: USA, Szwajcaria, Hiszpania, Norwegia, Holandia, Włochy, Irlandia, Niemcy, Francja, Dania, Belgia, Austria
- K3: Szwecja, Korea Płd., Nowa Zelandia, Finlandia, Kanada, Australia, Islandia
- ...
- K17: Somalia, Gambia, Etiopia, Kambodża
- K18: Uganda, Tanzania, Ruanda, Haiti, Burundi
- ...



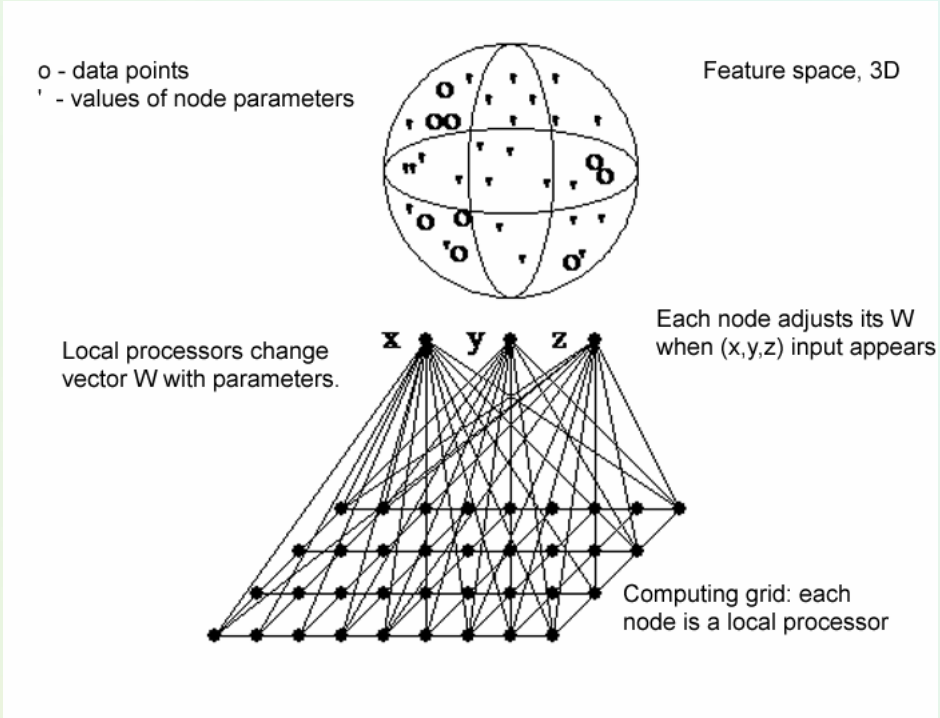
Other Model-Based Clustering Methods

- Neural network approaches
  - Represent each cluster as an exemplar, acting as a "prototype" of the cluster
  - New objects are distributed to the cluster whose exemplar is the most similar according to some dostance measure
- Competitive learning (Kohonen, SOM)
  - Involves a hierarchical architecture of several units (neurons)
  - Neurons compete in a "winner-takes-all" fashion for the object currently being presented

Self-Organizing Feature Map (SOM)

- SOMs, also called topological ordered maps, or Kohonen Self-Organizing Feature Map (KSOMs)
- It maps all the points in a high-dimensional source space into a 2 to 3-d target space, s.t., the distance and proximity relationship (i.e., topology) are preserved as much as possible
- Similar to k-means: cluster centers tend to lie in a low-dimensional manifold in the feature space
- Clustering is performed by having several units competing for the current object
  - The unit whose weight vector is closest to the current object wins
  - The winner and its neighbors learn by having their weights adjusted
- SOMs are believed to resemble processing that can occur in the brain
- Useful for visualizing high-dimensional data in 2- or 3-D space

Self-Organizing Maps - more



Data: vectors $\mathbf{X}^T = (X_1, \ldots X_d)$ from d-dimensional space.

Grid of nodes, with local processor (called neuron) in each node.

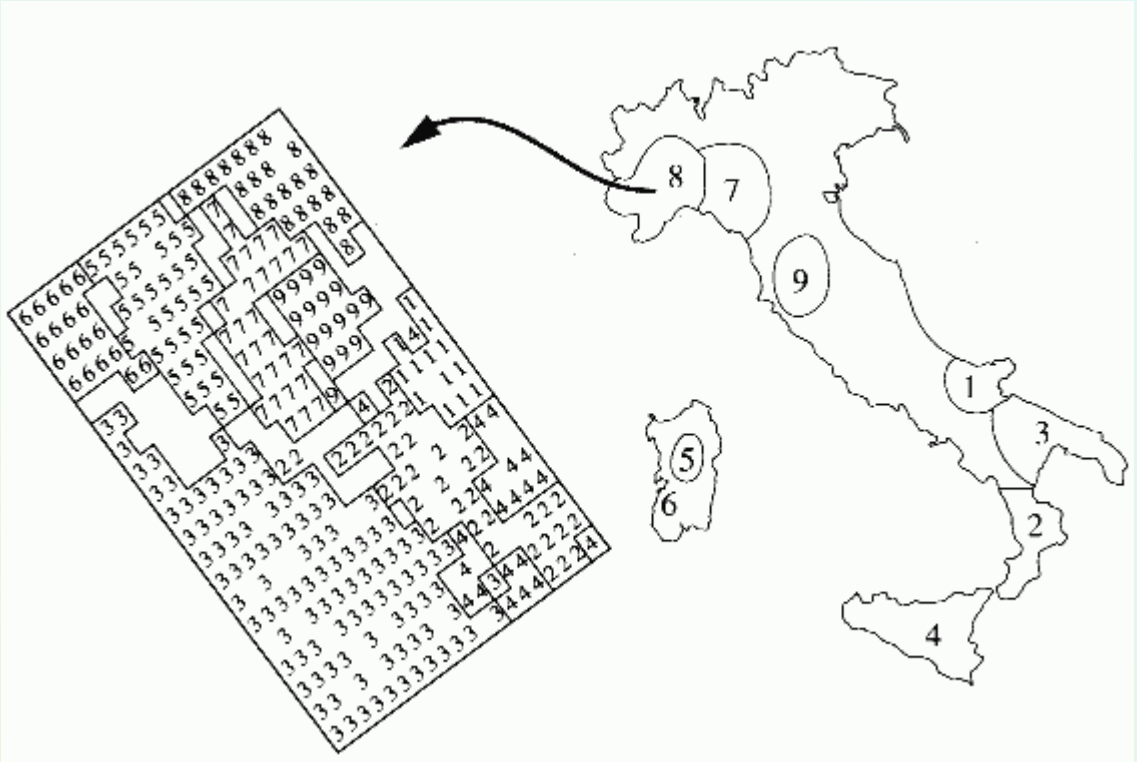
Local processor # *j* has *d* adaptive parameters $\mathbf{W}^{(j)}$.

Goal: change $\mathbf{W}^{(j)}$ parameters to recover data clusters in **X** space.

An example of analysing olive oil in Italy

An example of SOM application:

- 572 samples of olive oil were collected from 9 Italian provinces. Content of 8 fats was determine for each oil.
- SOM 20 x 20 network,
- Maps 8D => 2D.
- Classification accuracy was around 95-97%.



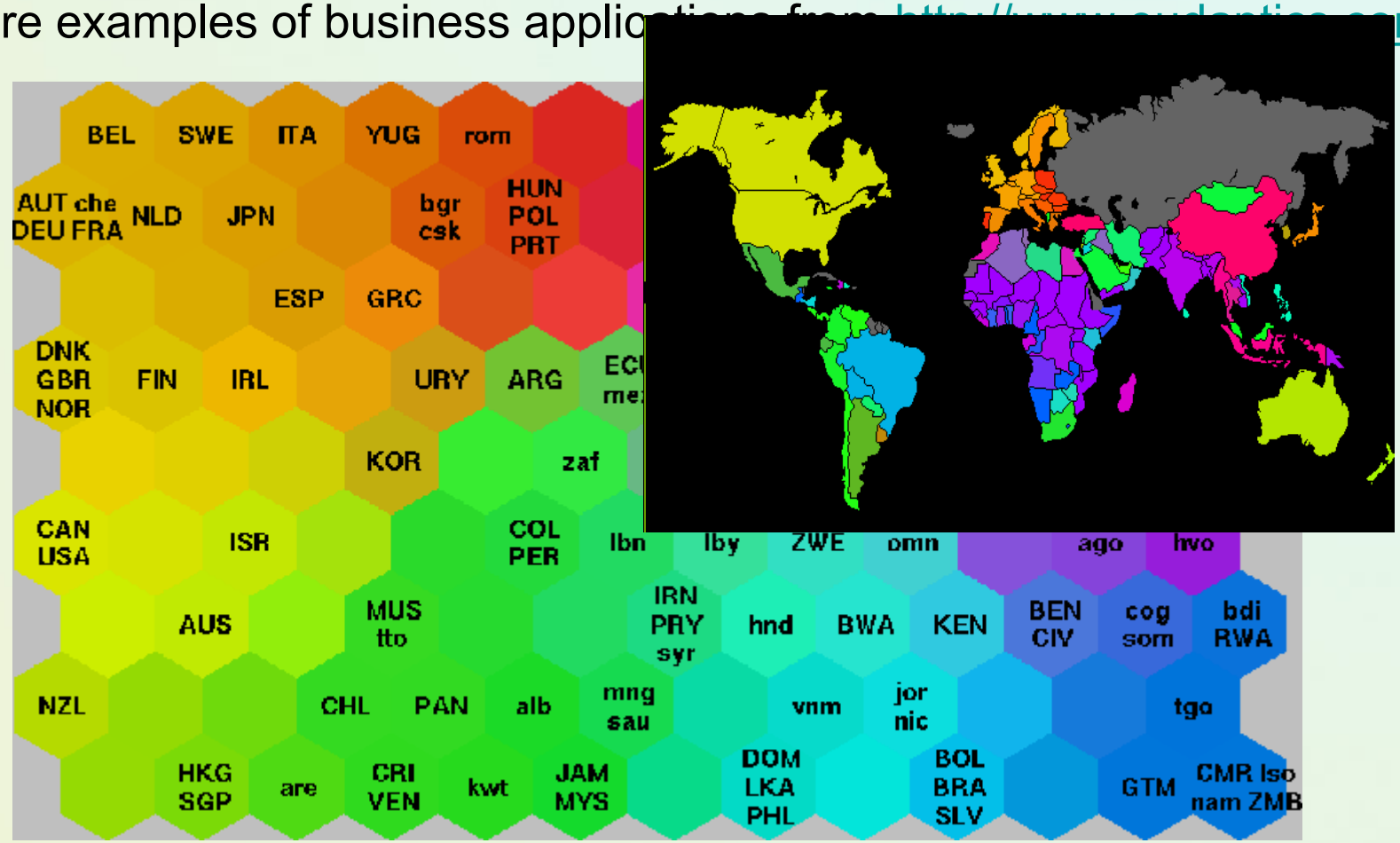
Note that topographical relations are preserved, region 3 is most diverse.

Quality of life data

WorldBank data 1992, 39 quality of life indicators.

SOM map and the same colors on the world map.

More examples of business applic[illegible]m/



Clustering High-Dimensional Data

- Clustering high-dimensional data
  - Many applications: text documents, DNA micro-array data
  - Major challenges:
    - Many irrelevant dimensions may mask clusters
    - Distance measure becomes meaningless—due to equi-distance
    - Clusters may exist only in some subspaces
- Methods
  - Feature transformation: only effective if most dimensions are relevant
    - PCA & SVD useful only when features are highly correlated/redundant
  - Feature selection: wrapper or filter approaches
    - useful to find a subspace where the data have nice clusters
  - Subspace-clustering: find clusters in all the possible subspaces
    - CLIQUE, ProClus, and frequent pattern-based clustering

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).

Problemy i wyzwania

- Znaczący postęp w zakresie skalowalnych algorytmów:
  - Partitioning: *k*-means, *k*-medoids, 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.

Clustering in Data Mining – przegląd

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*

Różne nietypowe zastosowanie AS

- Grupowanie złożonych typów danych
- Teksty, strony WWW, analiza zachowań użytkowników
- Szeregi czasowe
- Sekwencje zdarzeń
- Mikromacierze

Grupowanie opisów stron

- Komercyjne Vivisimo / Clusty

► Figure 16.2 Clustering of search results to improve recall. None of the top hits cover the animal sense of jaguar, but users can easily access it by clicking on the *cat* cluster in the *Clustered Results* panel on the left (third arrow from the top).

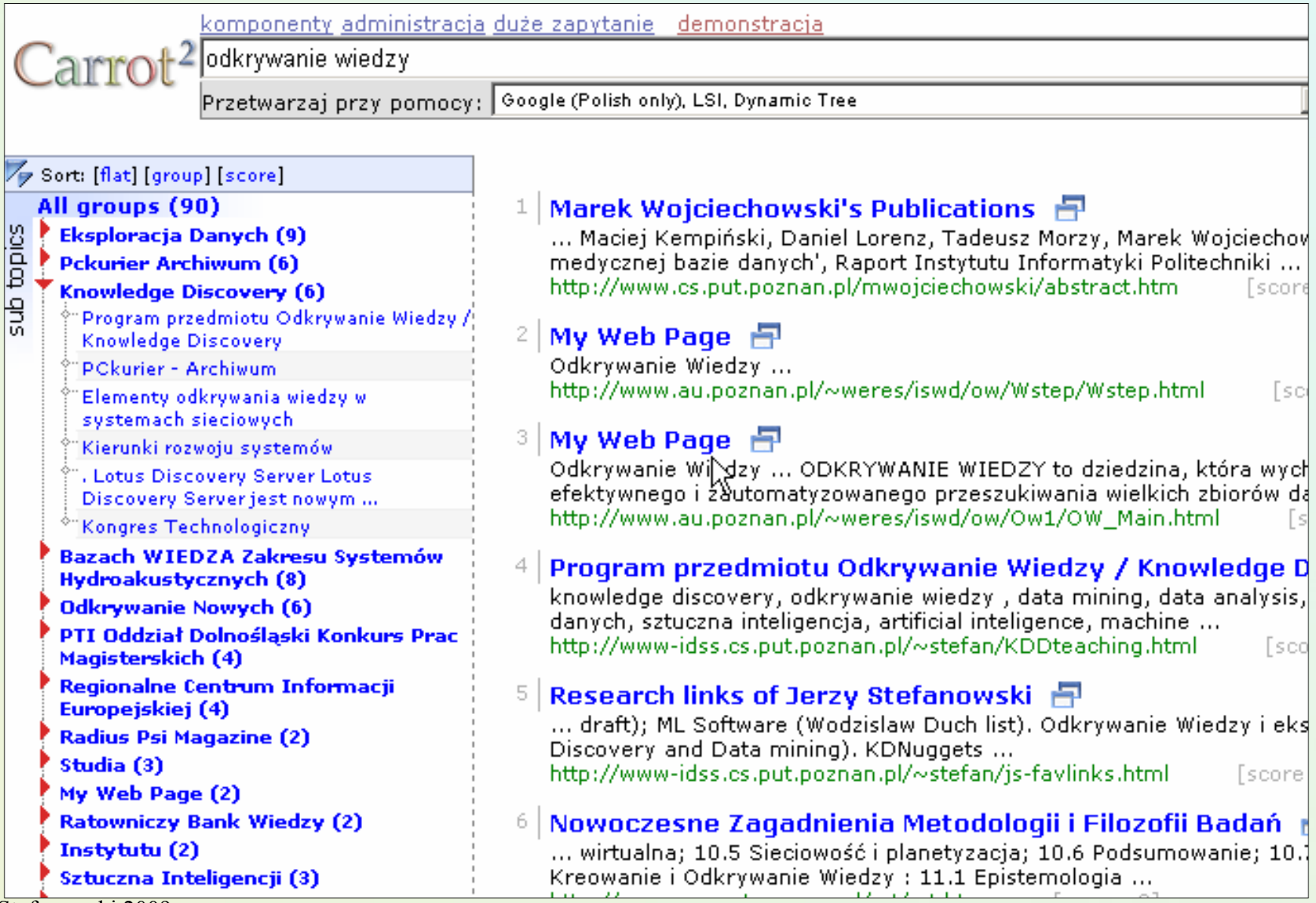
- Otwarte „Carrot” D.Weiss (S.Osinski + JS)

Projekt „Open source” Carrot 2

Carrot^2

- Projekt w Zakładzie Inteligentnych Systemów Wspomagania Decyzji
  - Paweł Kowalik
  - Stanisław Osiński
  - Jerzy Stefanowski
  - Dawid Weiss
  - Michał Wróblewski
- Strona główna: http://www.cs.put.poznan.pl/dweiss/carrot
- Demo: http://ophelia.cs.put.poznan.pl:2001

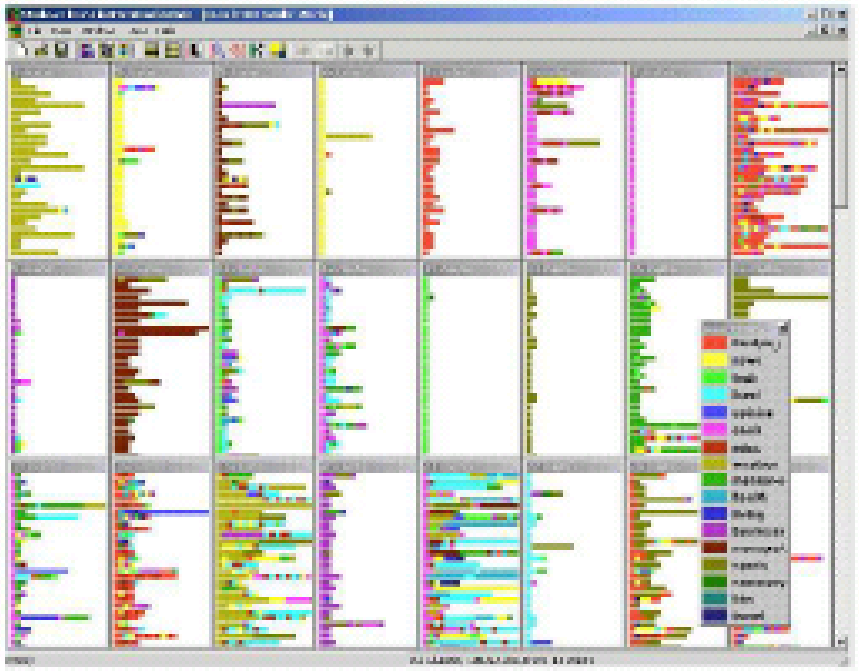
Web Search Result Clustering – Carrot2



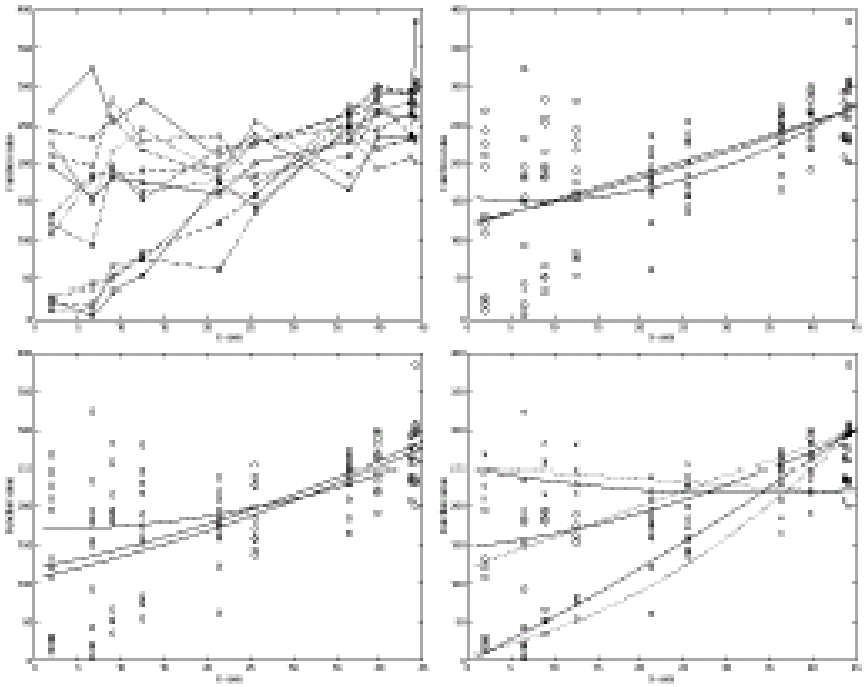
Specific data mining applications

More recently applications to non-vector data

Sequences (Web-usage)



Curves/Trajectories (Web-usage)



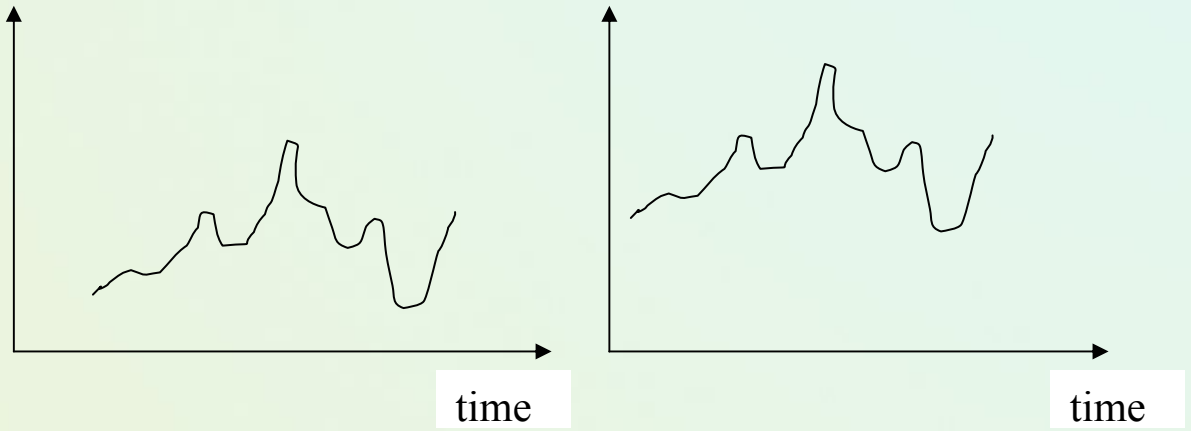
Trajectory clustering using mixtures of regression models
S. Gaffney and P. Smyth *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1999

Visualization of navigation patterns on a Web site using model-based clustering
I. Cadez, et˙ al. Technical Report MSR-TR-00-18, Microsoft Research, March 2000

Time-Series Similarities – specific data mining

Given a database of time-series.

Group “similar” time-series



Investor Fund A

Investor Fund B