Analiza skupień (I)

Marcin Szeląg

Zakład ISWD, Instytut Informatyki, Politechnika Poznańska

18.12.2019

WYBRANE PROBLEMY UCZENIA NIENADZOROWANEGO

---

1. Definicja uczenia nienadzorowanego
2. Co to jest analiza skupień? (Cele grupowania danych)
3. Metody numeryczne w analizie skupień
4. Wybrane numeryczne algorytmy tworzenia skupień:
   - hierarchiczne
   - dynamiczne
5. Ograniczenia metod numerycznych
6. Metody konceptualne budowy skupień (*Conceptual Clustering*)
7. Przykład algorytmu *CLUSTER/2*,
8. Algorytm COBWEB
9. Problemy otwarte

GRUPOWANIE
ANALIZA SKUPIEŃ
TAKSONOMIA

- Grupowanie (ang. Clustering)
  - Proces podziału zbioru danych (obiektów) na podzbiory, nazywane klasami (skupieniami)
    - ang. clusters
- Dwa aspekty:
  - podział na skupienia, tj wyodrębnienie jednorodnych grup obiektów
  - pomoc użytkownikowi w zrozumieniu skupień i struktury zbioru danych
- Skupienie (Klasa, ang. Cluster)
  - zbiór obiektów, które są „podobne" do siebie i mogą być traktowane zbiorczo jako jednorodna grupa

Tworzenie klasyfikacji

- **Ludzie i zwierzęta aby przeżyć muszą posiadać zdolność analizowania świata zewnętrznego:**
  - Tworzenie klasyfikacji i grupowanie obserwacji jest niezbędnym elementem budowania reprezentacji świata
  - Ludzie posiadają naturalne zdolności dostrzegania różnic i podobieństw miedzy obiektami, identyfikowania obiektów oraz ich strukturalizowania w pewne kategorie

- **Działalność badawczo-poznawcza - liczne przykłady tworzenia klasyfikacji:**
  - Taksonomia roślin w botanice, podział zwierząt na różne klasy, klasyfikacje w chemii, itp....
  - "*Wszechobecnym problemem w nauce jest tworzenie klasyfikacji obserwowanych obiektów czy sytuacji. Klasyfikacje muszą posiadać właściwe znaczenie interpretacyjne*" (*R.Michalski et al. 1983*)

"Dobre" i "złe" klasyfikacje ?

- Istnieje wiele sposobów definiowania klasyfikacji **!**
- Znaczenie klasyfikacji może być definiowane na poziomie semantycznym

"Problemem nie jest otrzymanie rezultatu prawdziwego lub fałszywego ale uzyskanie rezultatu użytecznego lub bezużytecznego" (Lance, Williams 77)

❒ **Niektóre kryteria użyteczności tworzonych klasyfikacji** (Bisson 94):

- Umożliwia odkrywanie nowych i nieoczekiwanych pojęć (Mendelejew 1869)
- Ułatwia analizę i "zrozumienie" danych
- Ocenia "jakość" atrybutów użytych do opisu danych

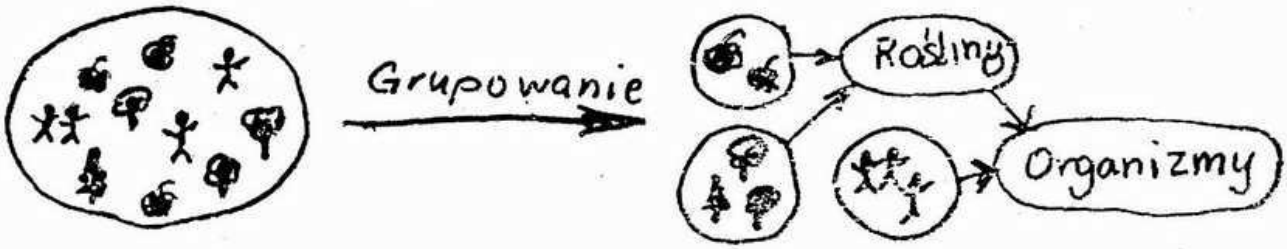
Co to jest analiza skupień?

Terminologia:

- Analiza skupień (ang. Clustor analysis)
  Grupowanie / Taksonomia

- Od Danych do WIEDZY...

Podział danego zbioru obiektów na podzbiory (skupienia, klasy, grupy) z punktu widzenia określonego kryterium opartego na cechach klasyfikowanych obiektów.



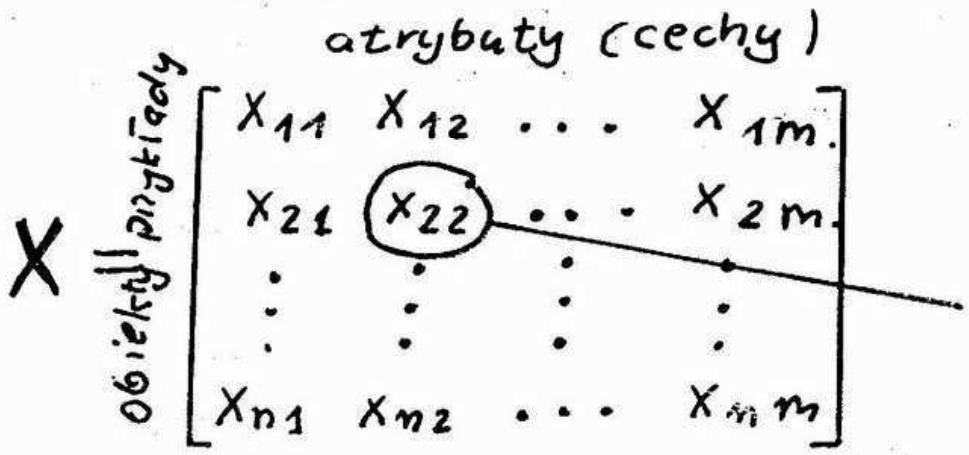
- Zadanie związane z analizą skupień:
  - strukturalizacja zbioru obiektów.
  - tworzenie hierarchii pojęć (semantycznie)
  - kompresja danych
  - przydatność dla predykcji

„DOBRE" GRUPOWANIE

- DOBRE GRUPOWANIE jest metodą tworzącą skupienia charakteryzujące się:
  - wysokim podobieństwem wzajemnym obiektów wewnątrz skupień (ang. high intra-class similarity)
  - niskim podobieństwem obiektów z różnych skupień (ang. low inter-class similarity)
- JAKOŚĆ grupowania zależy zarówno od miary bliskości (ang. similarity measure) jak i samej metody łączenia/podziału zbiorów obiektów.
- Jakość jest także związana z możliwością odkrywania części lub wszystkich wzorców informacyjnych ukrytych w danych.

Co jest potrzebne do tworzenia skupień (klasyfikacji)?

- ☐ Zbiór obiektów



- ☐ Doprowadzenie do porównywalności atrybutów
  - normalizacja
  - standaryzacja

  macierz standaryzowanych obserwacji →

- ☐ Wybór miary odległości

Metryka Minkowskiego

$$L_p(x_i, y_j) = \left(\sum_{k=1}^{m} w_k |x_{ik} - y_{jk}|^p \right)^{1/p}$$

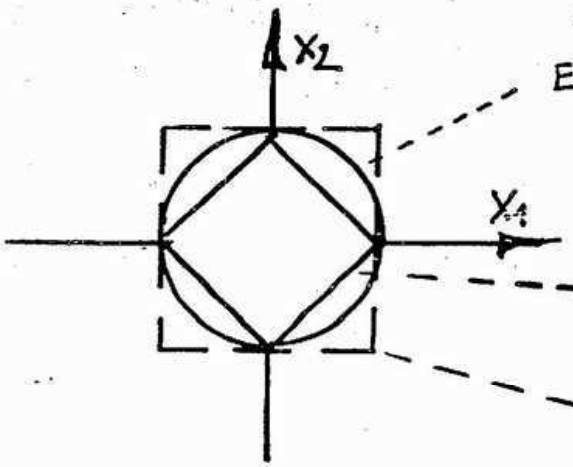
$p = 1$ odległość "miejska"

$p = 2$ klasyczna odległość euklidesowa

$p \to \infty$ $L_\infty = \max_k |x_{ik} - y_{jk}|$

metryka Czebyszewa

MIARY ODLEGŁOŚCI – cd.



Euklidesowa odległość
$\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

odległość Hamminga
$\sum_{i=1}^{n} |x_i - y_i|$

odległość Czebyszewa
$\max_{i=1,..n} |x_i - y_i|$

Interpretacja graficzna odległości Minkowskiego

Inne odległości:

odległość Mahalanobisa

$$\|\underline{x} - \underline{y}\| = (\underline{x} - \underline{y})^T M^{-1} (\underline{x} - \underline{y})$$

gdzie M – macierz reprezentująca wzajemne zależności pomiędzy zmiennymi

np. zastosowania statystyczne S macierz kowariancji

Zalety:
- uwzględnienie zależności pomiędzy zmiennymi
- ujednolicenie ze względu na jednostkę miary i rząd wielkości.

MIARY ODLEGŁOŚCI zależne od Typów atrybutów

Zmienne binarne

- Tablica „kontyngencji"

		Obiekt j		
		1	0	suma
Obiekt i	1	a	b	a+b
	0	c	d	c+d
	suma	a+c	b+d	p

- „Simple matching" (jeśli istnieje symetria)

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

ZMIENNE RÓŻNEGO TYPU

- Dane zawierają zmienne różnego typu (mieszane), np. binarne (symetryczne, asymetryczne), nominalne, porządkowe, ilościowe (przedziałowe, porządkowe).

- Stosuje się odległość ważoną integrującą różne czynniki składowe

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - jeśli f jest binarna lub nominalna

$$d_{ij}^{(f)} = 0 \quad \text{gdy} \quad x_i = x_j$$
$$= 1 \quad \text{w przeciwnym razie}$$

  - f porządkowa; oblicz rangi r_{if}. oraz

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

i traktuj z_{if} jako zmienną przedziałową

  - f przedziałowa - stosuj znormalizowaną odległość

WYMAGANIA DO GRUPOWANA Z PUNKTU WIDZENIA DATA MINING

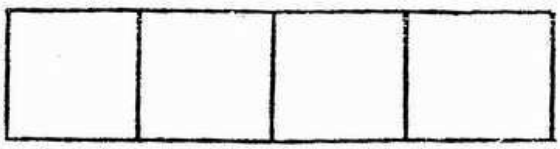
- Skalowalność
- uwzględnianie bardzo dużej liczby atrybutów
- uwzględnianie różnego typu atrybutów
- Ograniczenie wymogów co do posiadania wiedzy dziedzinowej dla doboru parametrów metod
- zdolność analizy niedoskonałych danych (m.in „szum" i obserwacje samotnicze)
- Odkrywanie skupień o różnorodnych (niewypukłych) kształtach
- „Niewrażliwość" na porządek prezentacji przykładów
- Interpretowalność i użyteczność tworzonych skupień

Metody numeryczne w analizie skupień

- Próby historyczne (Andanson 1757 taksonomia w botanice)

"Proponuje się łączyć obiekty w rodziny zgodnie z ich podobieństwem wzajemnym, a następnie łączyć ze sobą podobne rodziny"

- Analiza danych
  - Metody podziału.
    - klasyczne (zgodne z teorią mnogości)
    - "rozmyte"



Podział

Nakładanie się

  - Metody hierarchiczne

Ścisła hierarchia

'Piramida'

Nakładające się hierarchie.

3. Metody numeryczne w analizie skupień

Podstawowe metody:

- algorytmy dynamicznego grupowania - niehierarchiczne (*k*-means, metoda Forgy'ego i Jancey'a, metoda Wisharta, ISODATA)
- algorytmy hierarchiczne aglomeracyjne:
  - *kombinatoryczne* (najbliższego sąsiedztwa, najdalszego sąsiedztwa, centroidalna, średniej skupieniowej, skupiania parami, metoda Warda)
  - *niekombinatoryczne* (średniej odległości między skupieniami, średniej odległości wewnątrz skupień)
- algorytmy hierarchiczne podziału
- algorytmy oparte na uporządkowaniu obiektów

Algorytmy hierarchiczne - 1

- Początkowo wykonuje się transformację

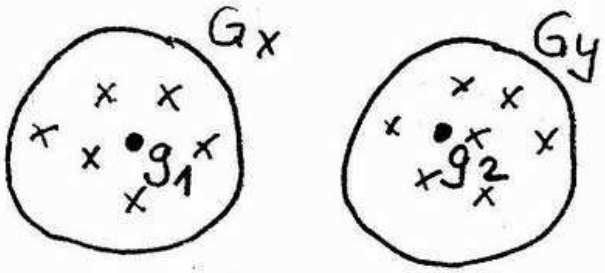
➤ Macierz obserwacji zamienia się do macierzy odległości / podobieństwa między skupieniami

	V_1	V_2	V_3	V_4	V_5
A	α	β	γ	ψ	δ
B	χ	δ	β	ω	β
C	α	β	α	ω	δ
D	ω	χ	φ	α	γ

→ Miara odległości

	A	B	C	D
A	–			
B	4	–		
C	5	6	–	
D	1	3	4	–

➤ Miara odległości zależna od procedury



- Single linkage.
  $d(G_x, G_y) = MIN(d(x_i, y_i))$
- Complete Linkage
- Centroid
  $d(G_x, G_y) = d(g_1, g_2)$

Algorytmy hierarchiczne - 2

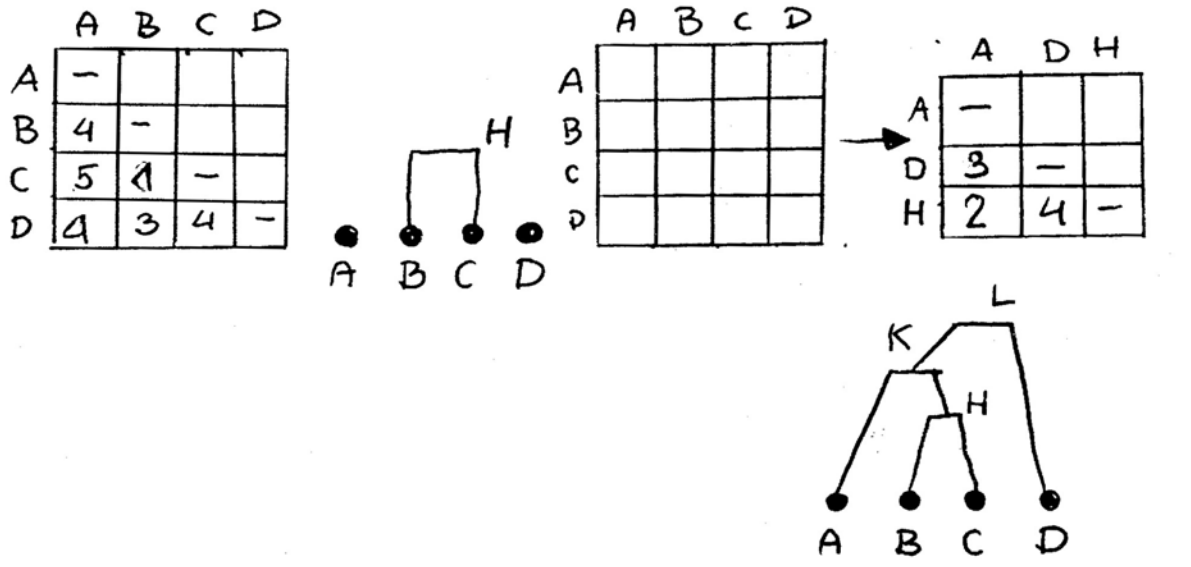
Procedura aglomeracyjna:

- Początkowo każdy obiekt tworzy jednoelementowe skupienie

Dopóki istnieją choć 2 skupienia **powtarzaj**

1. Znajdź w macierzy odległości dwa najbliższe skupienia ***I*** oraz ***J***
2. Połącz skupienia ***I*** oraz ***J*** w nowe skupienie ***H***, nadając mu numer ***I*** oraz usuwając skupienie o numerze ***J***. Zmniejsz numery skupień większe od ***J*** oraz liczbę skupień o jeden
3. Przekształć macierz odległości stosownie do wybranej metody aglomeracji

Przykład zastosowania metody "*single linkage*" (najbliższego sąsiedztwa**)**

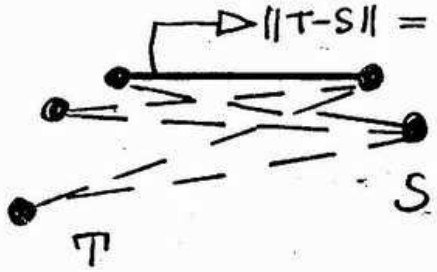


PRZYKŁADY MIAR ODLEGŁOŚCI MIĘDZY SKUPIENIAMI

---

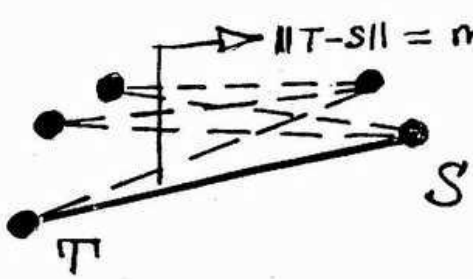
single linkage

$$\|T-S\| = \min_{\substack{x\in T\\ y\in S}} \|x-y\|$$



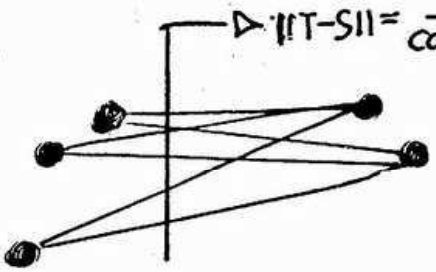
complete linkage

$$\|T-S\| = \max_{\substack{x\in T\\ y\in S}} \|x-y\|$$



average linkage

$$\|T-S\| = \frac{1}{card(T)\,card(S)} \sum_{\substack{y\in S\\ x\in T}} \|x-y\|$$



Przykładowy dendrogram

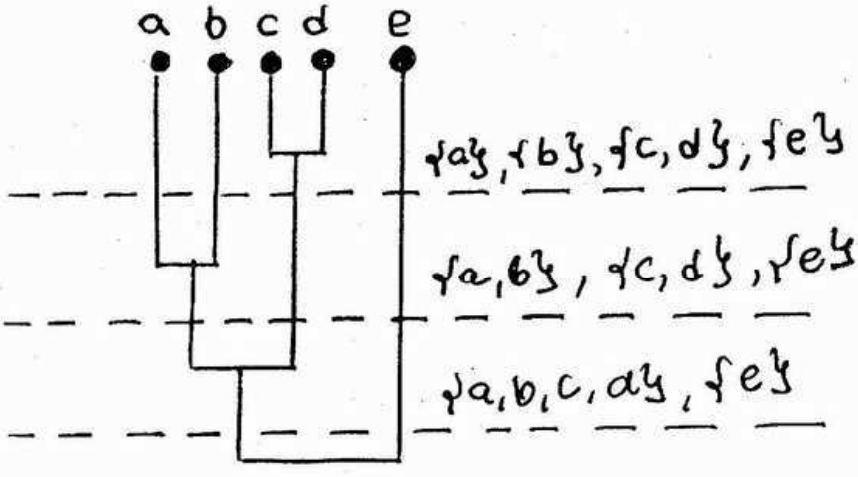
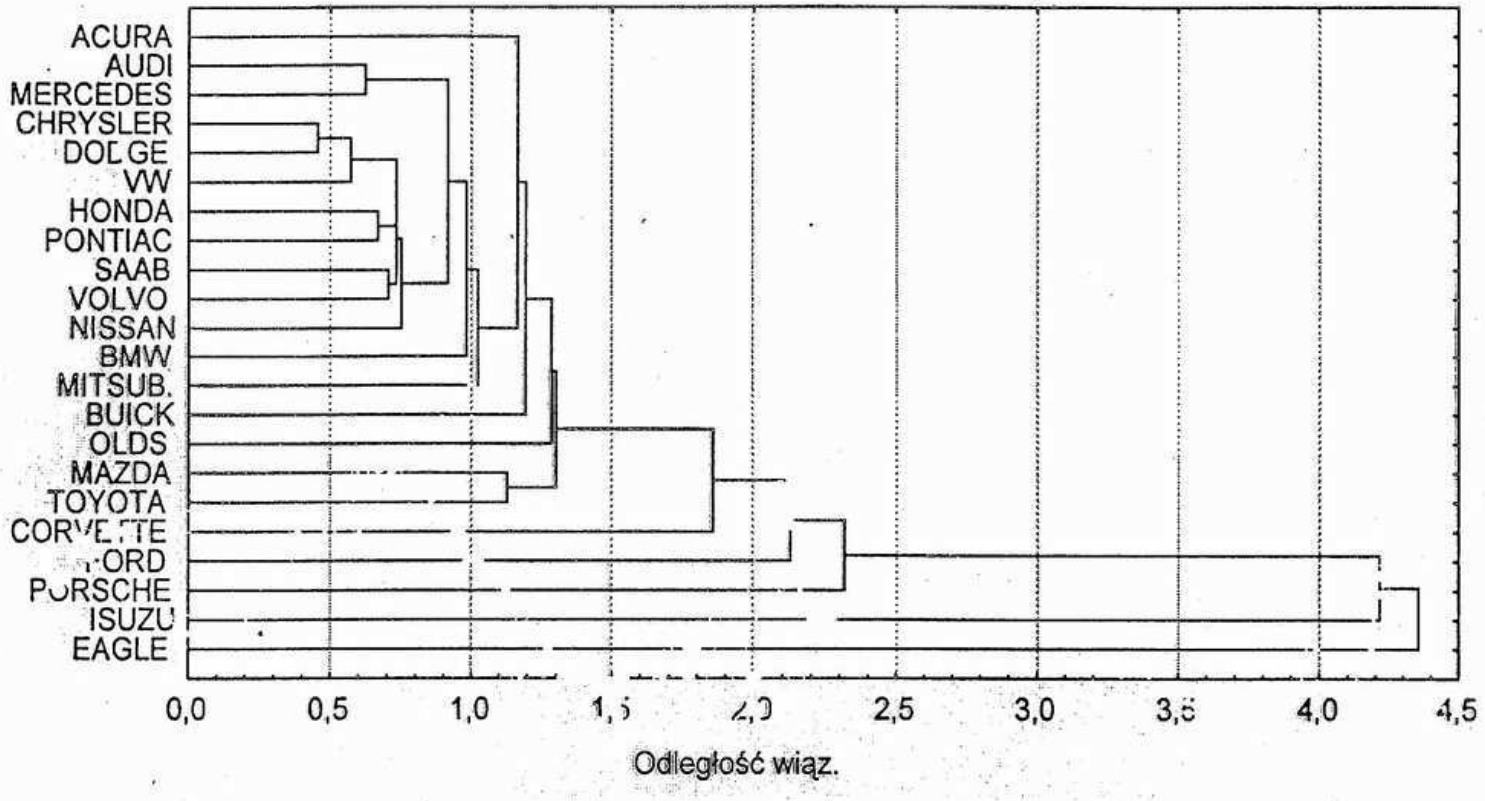


Diagram dla 22 przyp.
Pojedyncze wiązanie
Odległości euklidesowe



STATISTICA: Analiza skupień - [Przebieg aglomeracji (cars.sta)]

Plik Edycja Widok Analiza Wykresy Opcje Okno Pomoc

Dodge | Kolumny | Wiersze

ANALIZA SKUPIEŃ
poł¹cz. odległ³.

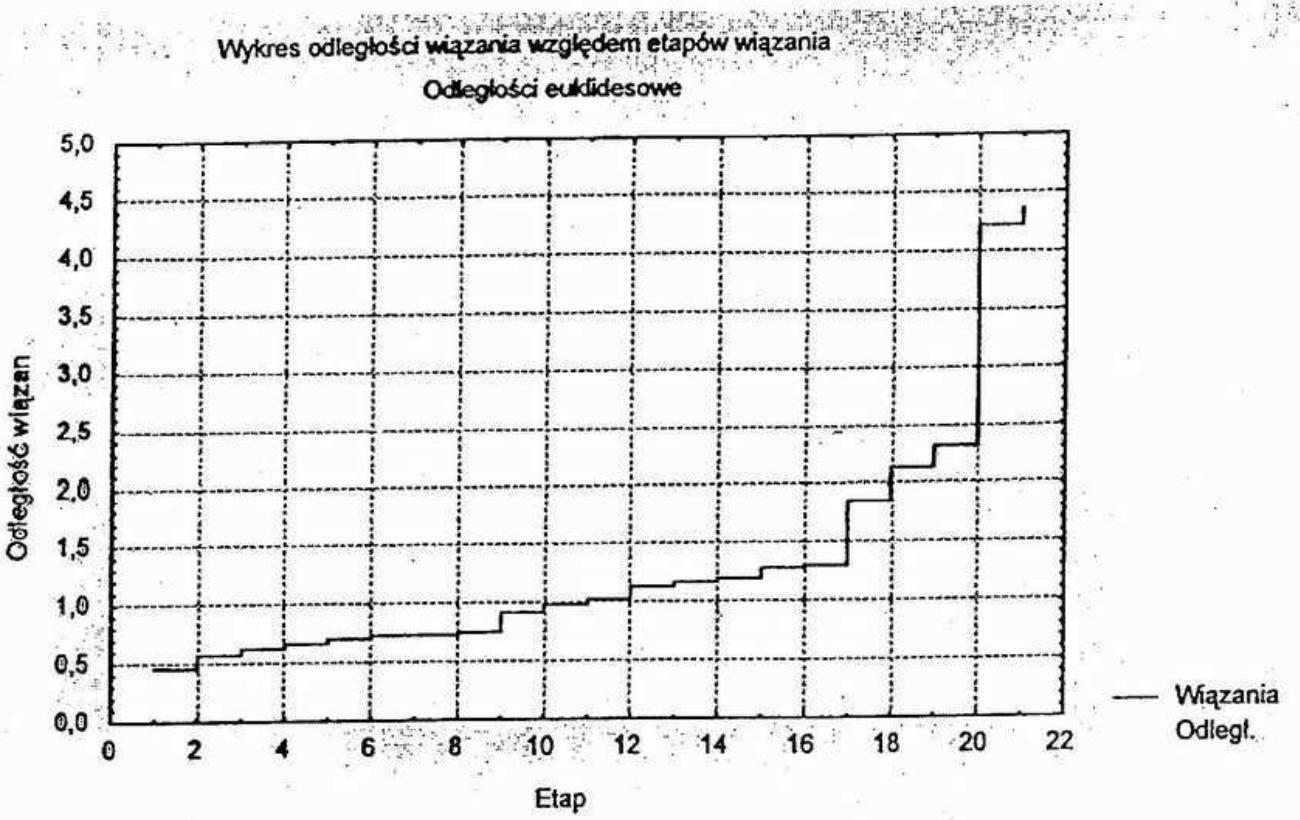
Pojedyncze wi¹zanie
Odleg³oœci euklidesowe

Obj. Nr 1	Obj. Nr 2	Obj. Nr 3	Obj. Nr 4	Obj. Nr 5
,5580484	Chrysler	Dodge		
,5710964	Chrysler	Dodge	VW	
,6231085	Audi	Mercedes		
,6670490	Honda	Pontiac		
,7060042	Saab	Volvo		
,7313896	Chrysler	Dodge	VW	Honda	Por
,7323840	Chrysler	Dodge	VW	Honda	Por
,7506309	Chrysler	Dodge	VW	Honda	Por
,9159300	Audi	Mercedes	Chrysler	Dodge	
,9824548	Audi	Mercedes	Chrysler	Dodge	
1,023831	Audi	Mercedes	Chrysler	Dodge	
1,127473	Mazda	Toyota			
1,164055	Acura	Audi	Mercedes	Chrysler	Do
1,193655	Acura	Audi	Mercedes	Chrysler	Do
1,284603	Acura	Audi	Mercedes	Chrysler	Do
1,301269	Acura	Audi	Mercedes	Chrysler	Do
1,855838	Acura	Audi	Mercedes	Chrysler	Do
2,128886	Acura	Audi	Mercedes	Chrysler	Do
2,317976	Acura	Audi	Mercedes	Chrysler	Do
4,	Acura	Audi	Mercedes	Chrysler	Do
4,	Acura	Audi	Mercedes	Chrysler	Do

Przyciski zadań | Dostosuj...

Gotowy | Wyłącz WYŁĄCZONE | Set NIE | Waga WYŁĄCZONA

Start | Windows Commander 4.0... | STATISTICA: Analiza... | Document3 - Microsoft W... | 16:40


Wykres odległości wiązania względem etapów wiązania
Odległości euklidesowe
Odległość wiązan
5,0
4,5
4,0
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
12
14
16
18
20
22
Etap
Wiązania
Odległ.

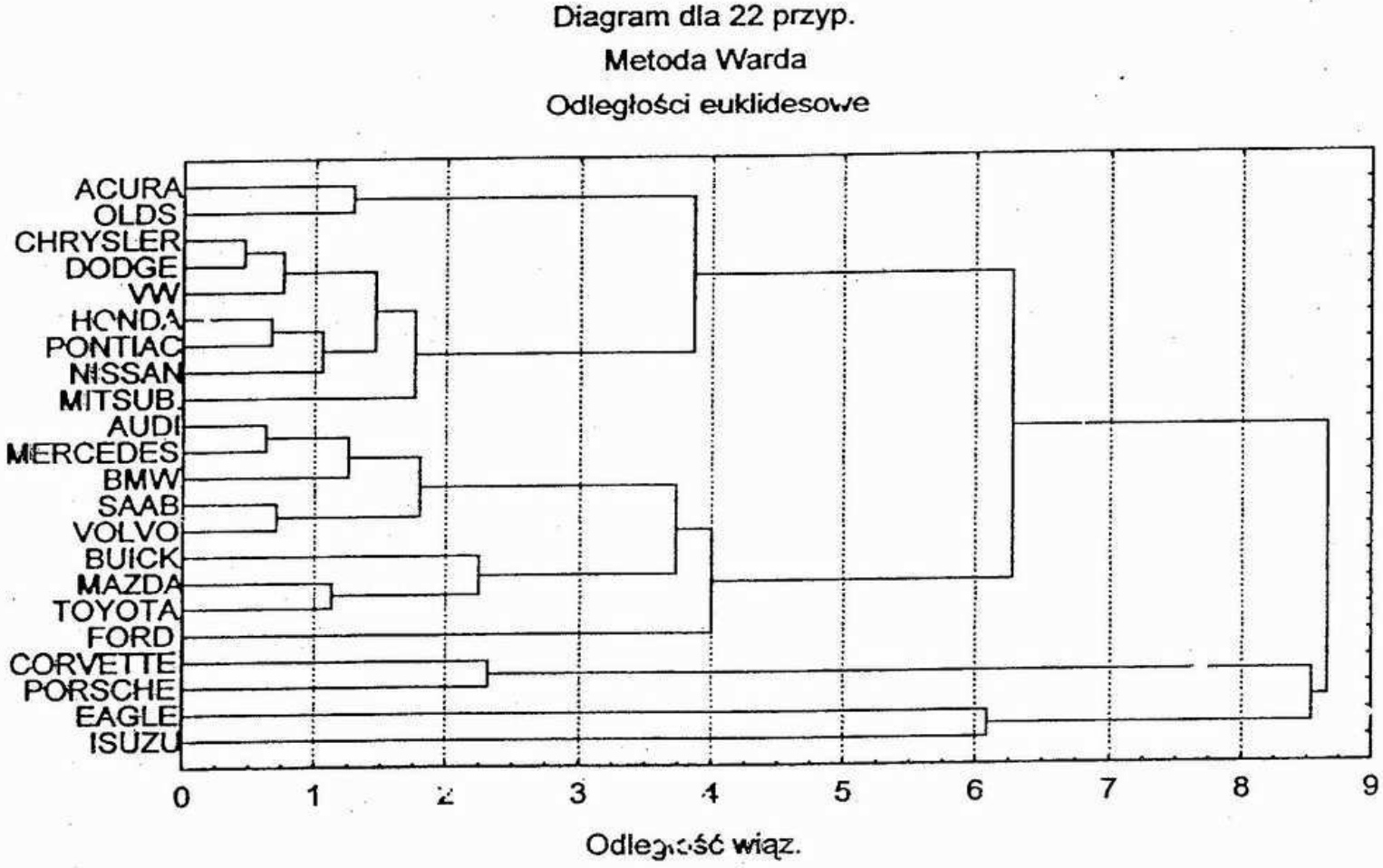


Diagram dla 22 przyp.
Metoda Warda
Odległości euklidesowe
ACURA
OLDS
CHRYSLER
DODGE
VW
HONDA
PONTIAC
NISSAN
MITSUB.
AUDI
MERCEDES
BMW
SAAB
VOLVO
BUICK
MAZDA
TOYOTA
FORD
CORVETTE
PORSCHE
EAGLE
ISUZU
0
1
2
3
4
5
6
7
8
9
Odległość wiąz.

org. zmienne były standaryzowane

stosowano odl. euklidesową

Tabela 5.2. Hierarchia łączenia rynków sprzedaży żywności w metodzie kompletnego połączenia

Numer kroku	Numery łączonych klas	Odległość międzyklasowa	Numer kroku	Numery łączonych klas	Odległość międzyklasowa
1	27—28	0,0426	28	47—49	0,4881
2	20—21	0,0957	29	41—44	0,5350
3	36—37	0,1173	30	22—26	0,5470
4	45—46	0,1227	31	5—6	0,6908
5	43—44	0,1347	32	28—33	0,7061
6	51—52	0,1486	33	50—54	0,7148
7	21—22	0,1500	34	10—11	0,7354
8	32—33	0,1511	35	3—4	0,8184
9	40—41	0,1557	36	38—39	0,8480
10	25—26	0,1597	37	34—37	0,8568
11	48—49	0,1600	38	13—15	0,8711
12	9—10	0,1912	39	44—49	0,9215
13	42—44	0,2058	40	6—8	1,0221
14	46—47	0,2222	41	18—26	1,0720
15	19—22	0,2371	42	33—37	1,1908
16	29—30	0,2416	43	1—2	1,2818
17	17—18	0,2545	44	49—54	1,4940
18	35—37	0,2612	45	12—15	1,5088
19	53—54	0,2713	46	8—11	1,7215
20	24—26	0,3011	47	26—37	1,8005
21	31—33	0,3013	48	3[illegible]—54	2,1453
22	7—8	0,3390	49	4—11	2,6378
[illegible]	23—26	0,3923	50	11—15	3,6166
24	14—15	0,3991	51	37—54	3,8431
25	30—33	0,4198	52	15—54	5,6142
26	16—18	0,4229	53	2—54	9,2645
27	52—54	0,4442			

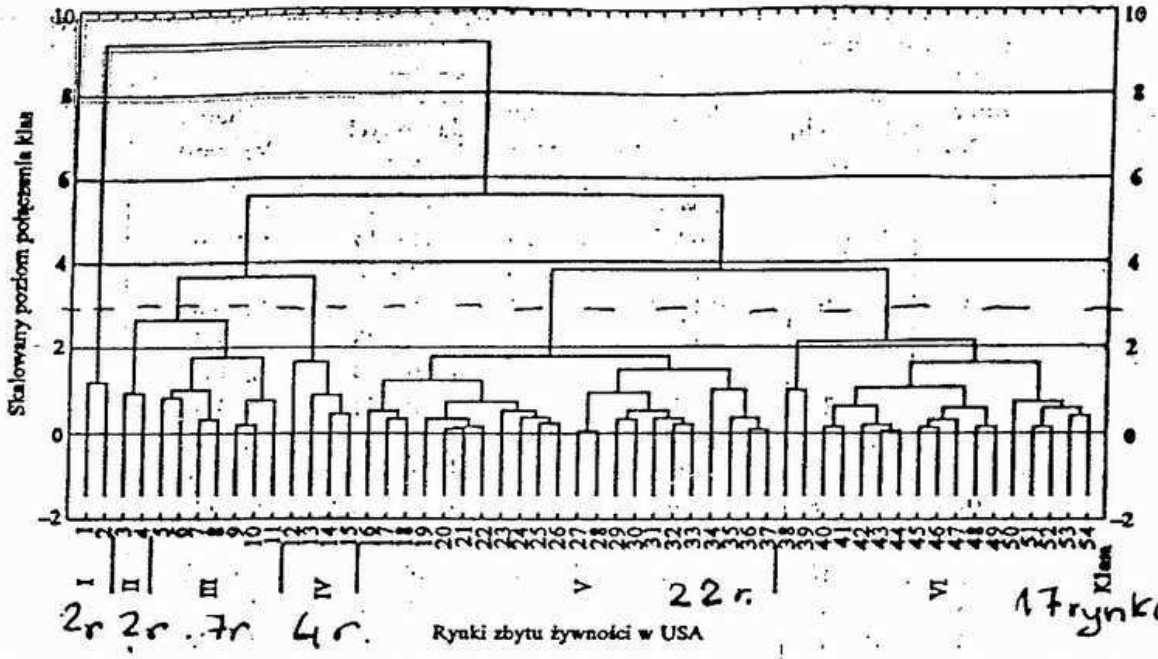
wybór wg metody Mojena

$h_{e+1} > 2.58$

Źródło: Opracowanie własne na podstawie obliczeń wykonanych z użyciem pakietu statystycznego SPSS for Windows.

Analiza rynków sprzedaży żywności w USA (1991).

Cel: wyodrębnienie relatywnie jednorodnych klas rynków w celu testowania produktów.



wybór.

2r 2r 7r 4r 22r 17 rynków.

Rynki zbytu żywności w USA

1 New York	19 Portland (Oregon)	37 Birmingham, Montgomery, Huntsville
2 Los Angeles, San Diego	20 Oklahoma City, Tulsa	38 Quad Cities
3 Philadelphia	21 Norfolk, Richmond	39 Hartford, New Haven, Springfield
4 Jacksonville, Orlando, Tampa	22 El Paso, Albuquerque, Lubbock	40 Syracuse
5 Minneapolis, St. Paul	23 Louisville, Lexington	41 Indianapolis
6 Houston	24 Salt Lake City, Boise	42 Raleigh, Greensboro, Winston-Salem
7 Detroit	25 Omaha, Des Moines	43 Nashville, Knoxville
8 Dallas, Fort Worth	26 Charlotte	44 Buffalo, Rochester
9 Miami	27 Peoria, Springfield	45 St. Louis
10 Cleveland	28 Green Bay	46 Seattle, Tacoma
11 Cincinnati, Dayton, Columbus	29 Scranton, Wilkes Barre	47 Denver
12 San Francisco	30 Greenville, Spartanburg, Asheville	48 New Orleans
13 Chicago	31 Portland, Me., Concord	49 Atlanta
14 Boston, Providence	32 Spokane, Yakima	50 Wichita
15 Baltimore, Washington	33 Charleston, Savannah	51 Milwaukee
16 Pittsburgh	34 Charleston, Huntington	52 Kansas City
17 San Antonio, Corpus Christi	35 Memphis, Little Rock	53 Grand Rapids, Kalamazoo
18 Phoenix, Tuscon	36 Shreveport, Jackson	54 Albany, Schenectady, Troy

Rys. 5.1. Dendrogram 54 rynków sprzedaży żywności w USA w 1989 r. uzyskany metodą kompletnego połączenia z zastosowaniem pakietu statystycznego Statistica

Źródło: Opracowanie własne

... Zmienna 1 – Liczba ludności (tys)

Zmienna 2 – Liczba gospod. domowych (tys)

Zm 3 – udział sprzed. żywności na dan. rynku w sprzedaży żywności ogółem w USA (%)

Zmien. 4 – dochód realny przypadający na gospod. domowe (USD)

Tabela 5.3. Średnie arytmetyczne i odchylenia standardowe zmiennych w klasach rynków sprzedaży żywności w USA

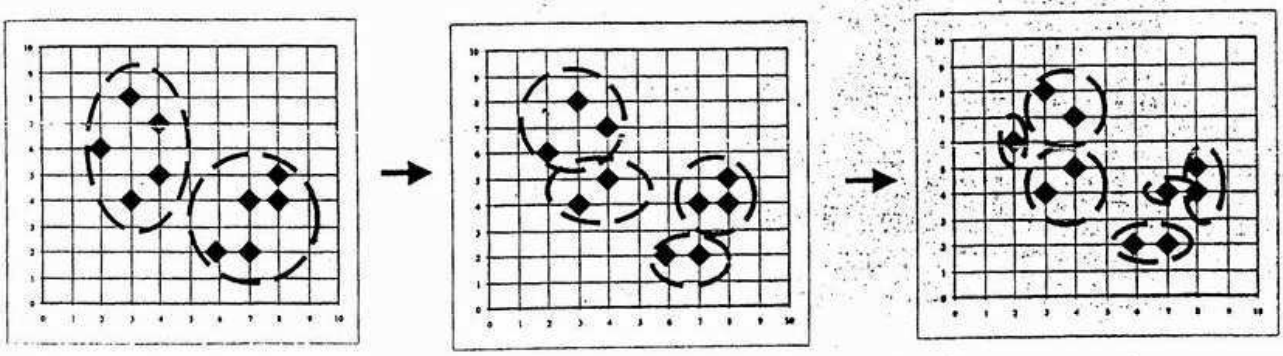
Klasa/zmienna		1	2	3	4
I	A	16945,0	6237,0	7,27	34464
	B	1392,0	636,7	0,47	1773
II	A	7005,6	2680,9	2,98	28866
	B	779,9	138,1	0,27	1727
III	A	4486,5	1668,1	1,92	33028
	B	755,9	273,3	0,26	2116
IV	A	8307,5	3058,2	3,58	36481
	B	1225,3	473,6	0,59	525
V	A	2391,2	866,9	0,97	27450
	B	886,5	323,4	0,38	1809
VI	A	2630,0	960,0	1,09	31748
	B	690,8	244,6	0,32	3066

A — średnia arytmetyczna, B — odchylenie standardowe.

Źródło: Opracowanie własne.

Metody deglomeracyjne - podziału

- **Odwrotny kierunek działania algorytmu niż w przypadku metod aglomeracyjnych.**
- **Stopniowy podział z góry na dół. Na początku wszystkie obiekty tworzą pojedyncze skupienia, następnie dane skupienie jest dzielone na (dwa) kolejne skupienia..., podział jest powtarzany aż każdy obiekt utworzy własne skupienia lub osiągnie się warunek zatrzymania.**
- **Przykład warunku zatrzymania: odległość między najbliższymi skupieniami powyżej zadanego progu.**
- **Zaimplementowane w niektórych pakietach statystycznych, np. S+.**



4. Wybrane algorytmy numerycznego tworzenia skupień

Algorytmy Dynamiczne: *k - means*

Inicjalizacja: Wykonać wstępny podział obiektów na *k* skupień (możliwe do wykonania na wiele sposobów)

Przebieg algorytmu :

1. Dla każdego skupienia oblicza się jego centroid
2. Rozważa się kolejno wszystkie obiekty i przydziela do najbliższego centroidu (tworząc skupienia)
3. Jeżeli nie osiągnięto warunku stabilizacji (np. jeśli obiekt lub obiekty w rezultacie kroku 2 przechodzą do innego skupienia) wyznacza się (poprawia się) nowe centroidy dla każdego ze skupień
4. Kroki 2 oraz 3 powtarza się aż do osiągnięcia stabilizacji (tzn. np. gdy nie ma zmian w przydziale obiektów do skupień) lub do osiągnięcia maksymalnej dopuszczalnej liczby iteracji.

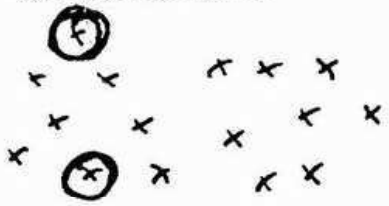
Liczba skupień *k* oraz maksymalna liczba iteracji są parametrami procedury. Jako odległości najczęściej używa się kwadratu odległości euklidesowej. Parametr oceniający - macierz zmienności wewnątrzskupieniowej.

ALGORYTMY DYNAMICZNE 2

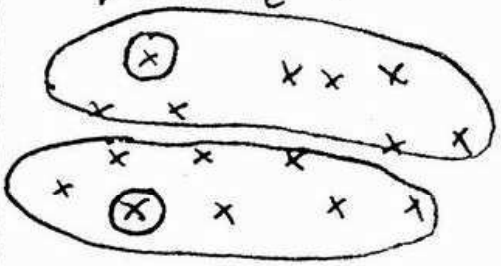
- Algorytmy wykorzystują dwie funkcje.
  - F: Funkcja przypisująca obserwacje do skupienia (miara bliskości)
  - G: Funkcja charakteryzujące skupienie
- Idea algorytmu

Początkowy zbiór obserwacji

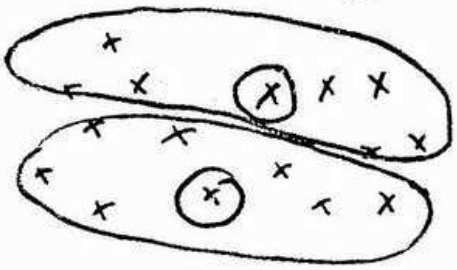
Wybór 2 (k=2) centroidów



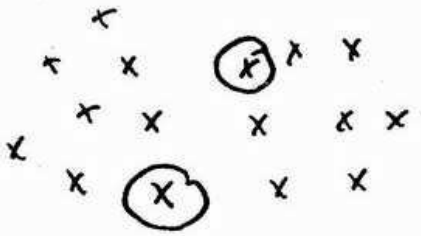
Agregacja za pomocą F



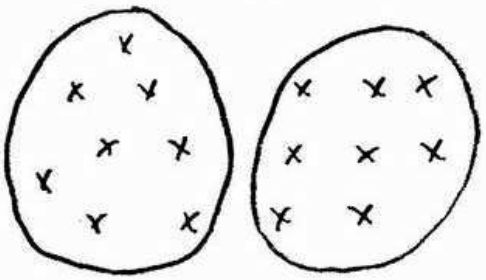
G tworzy nowe 2 centroidy



Nowa iteracja



Po wielu iteracjach



ALGORYTMY DYNAMICZNE 3

- Elastyczna metoda zależna od definicji F oraz G
  - Funkcja F: przypisanie do najbliższego centroidu za pomocą miary bliskości (Minkowski)
  - Funkcja G: Ukierunkowanie na najbardziej centralny element → na ogół "uśredniony" element.
- Metoda inspirująca niektóre techniki konceptualne
  - CLUSTER/2 : F funkcja zbieżności, G metoda uogólniania.
  - WITT : używa się k-means dla uzyskania wstępnych skupień
- ograniczenia metod
  - Użytkownik decyduje o liczbie K tworzonych skupień
  - Należy zapewnić właściwy wybór początkowych skupień
- Istnieje rozszerzenie "rozmyte" algorytmu tzw. FUZZY K-MEANS (Bezdek)

Algorytm K-Medoids

- Rozszerzenie idei algorytmu k-means (mniejsze? koszty obliczeniowe)
- Cel: znaleźć reprezentatywne obiekty dla skupień, nazywane medoid-ami

→ odległości pomiędzy obiektami są wejściem do algorytmu i obliczane są jednokrotnie (!)

- Algorytm PAM (partitioning Around Medoids, 1987) – realizacja metody k-medoids
- Rozszerzenia dla większych rozmiarów danych

  CLARA (1990)

  CLARENS (1994)

  - możliwości analizy "spatial data" (1995)

Algorytm PAM

- Rozpocznij ze zbiorem początkowym „medoids", następnie iteracyjnie wymieniaj jeden z punktów medoidalnych z punktem niewybranym tak, aby polepszać jakość grupowania

- Jakość grupowania (złożona jakość grupowania w każdym z medoidów)
  - uśredniona odległość pomiędzy medoidem, a obiektami należącymi do skupienia

- TC_{ih} - koszt wymiany pomiędzy obiektem wybranym O_i (medoid) a obiektem niewybranym O_h (kandydatem)

$$TC_{ih} = \sum_j C_{jih}.$$

C_{jih} jest kosztem wymiany O_i i O_h względem innego nie wybranego obiektu O_j.

Algorytm PAM (Kaufman, Rousseeuw 87)

1. Wybierz k reprezentatywnych obiektów.
2. Oblicz. TC_{ih} dla wszystkich par obiektów. $O_i O_h$ gdzie O_i jest obiektem wybranym, a O_h jest obiektem dotychczas niewybranym
3. Wybierz parę $O_i O_h$ z $\min_{O_i O_h} TC_{ih}$.
   Jeśli $TC_{ih} < 0$ zastąp O_i przez O_h
   i wróć do kroku (2)
   w przeciwnym razie
4. Przydziel każdy niewybrany obiekt do skupienia reprezentowanego przez najbliższy medoid
5. STOP

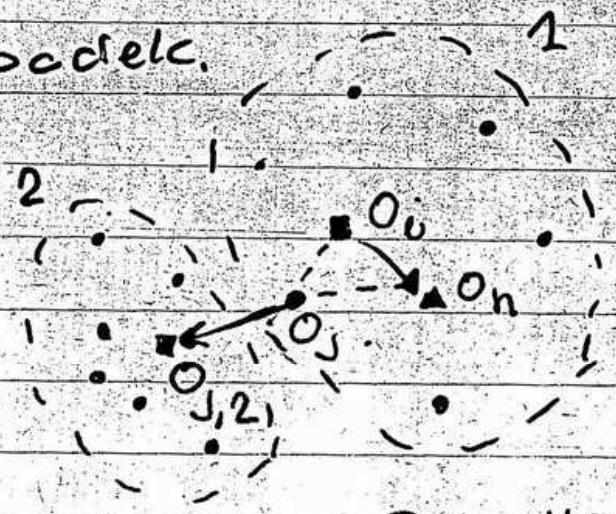
Uwagi
Złożoność $O(k(n-k)^2$ iteracji)
Problemy dla dużych zbiorów

PAM - obliczanie kosztu wymiany

$C_{jih} = ?$

Różne sposoby obliczania

1. Przypadek



O_j e skupienie (O_i)

O_{j2} - następny bliski medoid do O_j

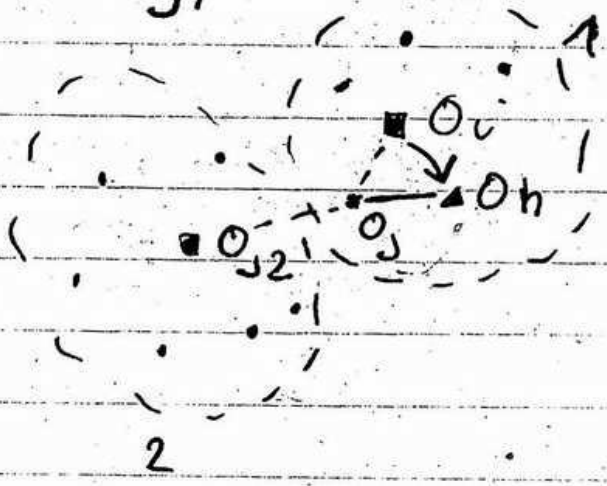
Ponadto

$$d(O_j, O_h) > d(O_j, O_{j2})$$

Jeśli $O_i \to O_h$ to $O_j \to O_{j2}$

$$C_{jih} = d(O_j, O_{j2}) - d(O_j, O_i)$$

Przypadek 2



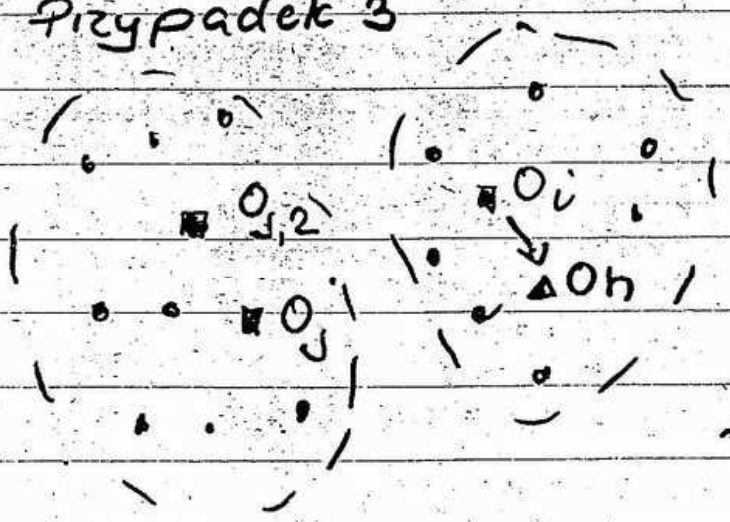
$$d(O_j, O_h) < d(O_j, O_{j2})$$

O_j pozostaje w skupieniu 2

$$C_{jih} = d(O_j, O_h) - d(O_j, O_i)$$

PAM - obliczanie kosztów wymiany

Przypadek 3



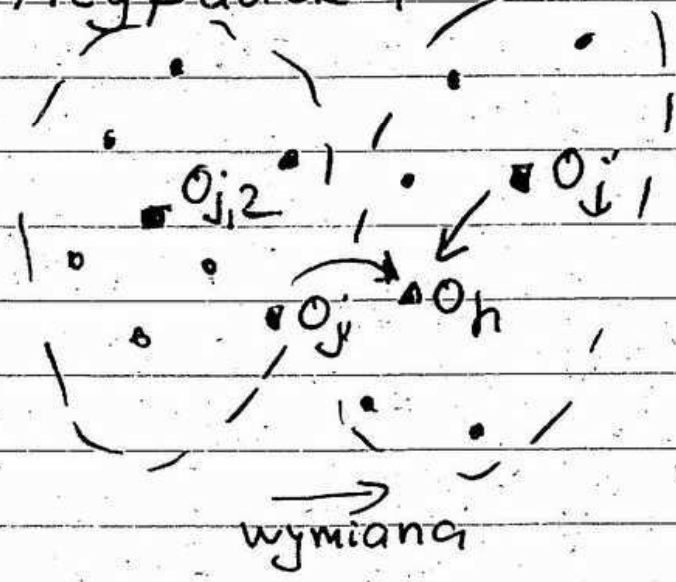
$O_j \in$ skupienie $(O_{j,2})$

$d(O_j, O_{j,2}) \leq d(O_j, O_h)$

$C_{jih} = 0$

Brak wymiany

Przypadek 4



$O_j \notin$ skupienie $(O_{j,2})$

$d(O_j, O_{j,2}) > d(O_j, O_h)$

Jeśli $O_i \rightarrow O_h$ to

$O_j \rightarrow$ skupienie (O_h)

$$C_{jih} = d(O_j, O_h) - d(O_j, O_{j,2})$$

CLARA (Clustering Large Applications)

- Opiera się na idei próbkowania mniejszych podzbiorów z całej bazy danych. (wiele podzbiorów!)
- Algorytm PAM używa się na podzbiorze generując k medoids → skupień
- Dla wielu podzbiorów i osiągniętych skupień wybiera się najlepsze
- Wyniki eksperymentalne → dla danych rzędu tysięcy obiektów rozmiar próby losowej ~ $40+2k$
- Złożoność $\Theta(k(40+k)^2 + k(n-k))$ możliwość analizy większych zbiorów danych
- Ograniczenia: Dobry dobór skupień na próbie niekoniecznie reprezentuje dobre skupienia na całości (efekt optimum lokalnego)

CLARANS – reprezentacja grafu przeszukiwań i heurystyki sterujące wyborem prób wokół sąsiadów w grafie

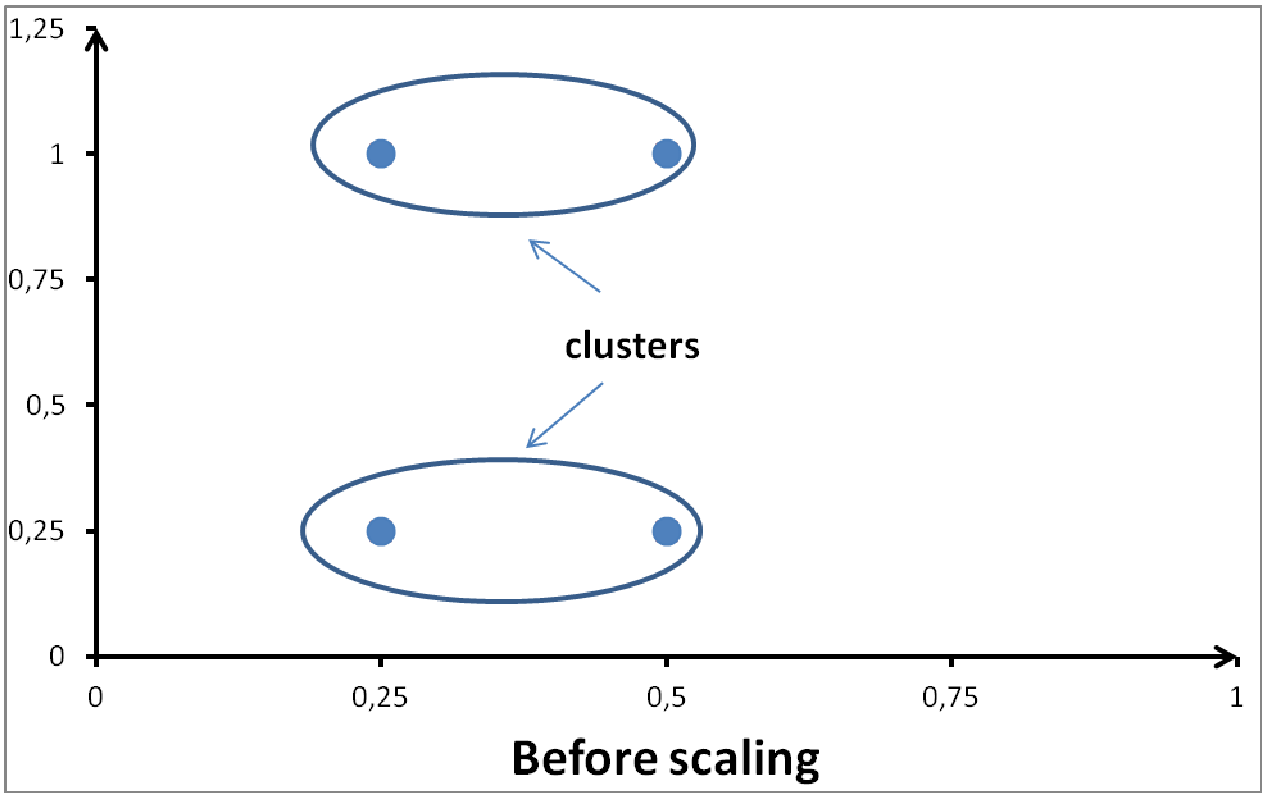
Algorytm CLARA

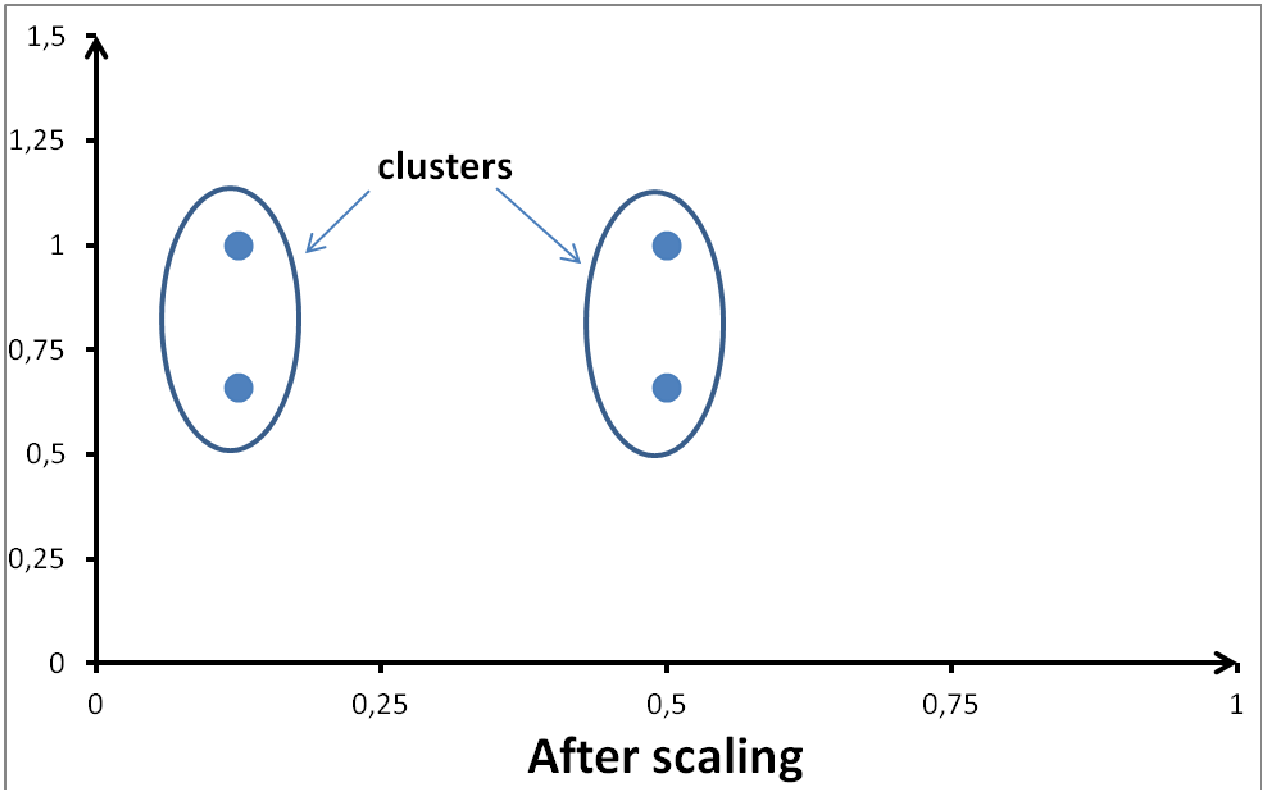
1. Dla zadanej liczby iteracji
2. Wylosuj próbę ~ 40+2k obiektów z całego zbioru danych[1] użyj algorytmu PAM dla znalezienia k medoidów
3. Dla każdego obiektu O_j z całego zbioru danych określ do którego k medoidu jest najbliższy
4. Oblicz średnią jakość grupowania. Jeśli jest mniejsza niż aktualne minimum, użyj jej jako nowego minimum, a k medoidy uznaj za „best set" dotychczas znaleziony
5. Powtórz kroki 2-4 dla kolejnej iteracji

[1] W Losowaniu można utrzymywać „pamięć" ostatnich najlepszych medoidów.

Uzupełnienie do wykładu z Analizy Skupień – metody numeryczne:

- k-means/k-medoid - metody iteracyjno-optymalizacyjne
- k-medoid - "uodpornienie" alg. k-średnich na występowanie obserwacji odstających w danych (zniekształcają wartości średnie)
- różne skalowanie atrybutów obiektów może prowadzić do różnych skupień:





- k-means:
    1. może się zdarzyć, że do jakiegoś centroidu nie zostaną przypisane żadne punkty (konieczność uwzględnienia takiej sytuacji w implementacji)
    2. algorytm zachłanny => ryzyko wpadnięcia w minimum lokalne - możliwość wielokrotnego uruchomienia z różnymi punktami startowymi, dla danego k
- adaptacyjne k-means (3 możliwości):
    1. wielokrotne uruchamianie algorytmu dla różnych k i wybór najlepszego z uzyskanych grupowań (większy nakład obliczeniowy) (SIC! zwiększanie k powoduje spadek błędu wewnątrz klastrów, ale też wzrost ryzyka przeuczenia)
    2. tworzenie grupowania hierarchicznego przez wielokrotne stosowanie algorytmu k-środków dla k=2; początkowo otrzymujemy podział na dwie grupy, potem każdą z nich dzielimy ponownie na dwie grupy (uzyskując łącznie cztery grupy) itd. (zachodzi konieczność określenia kryterium stopu lub kryterium oceny jakości, porównującego jakość jednej "dużej" grupy z dwiema mniejszymi, na które można ją podzielić)
    3. modyfikacja algorytmu umożliwiająca zmianę początkowej liczby grup w trakcie procesu grupowania - po zakończeniu etapu przypisywania przykładów trenujących do poszczególnych grup następuje weryfikacja czy: 1) żadna grupa nie jest zbyt zróżnicowana, 2) żadne dwie grupy nie są do siebie zbyt podobne

- hierarchical clustering – przykład na stronie http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html
- hierarchical clustering – demo Java na stronie http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html
- k-means – demo Java na stronie http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- fuzzy c-means (fuzzy sets!) – opis na stronie http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html
- fuzzy c-means (fuzzy sets!) - demo Java na stronie http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletFCM.html