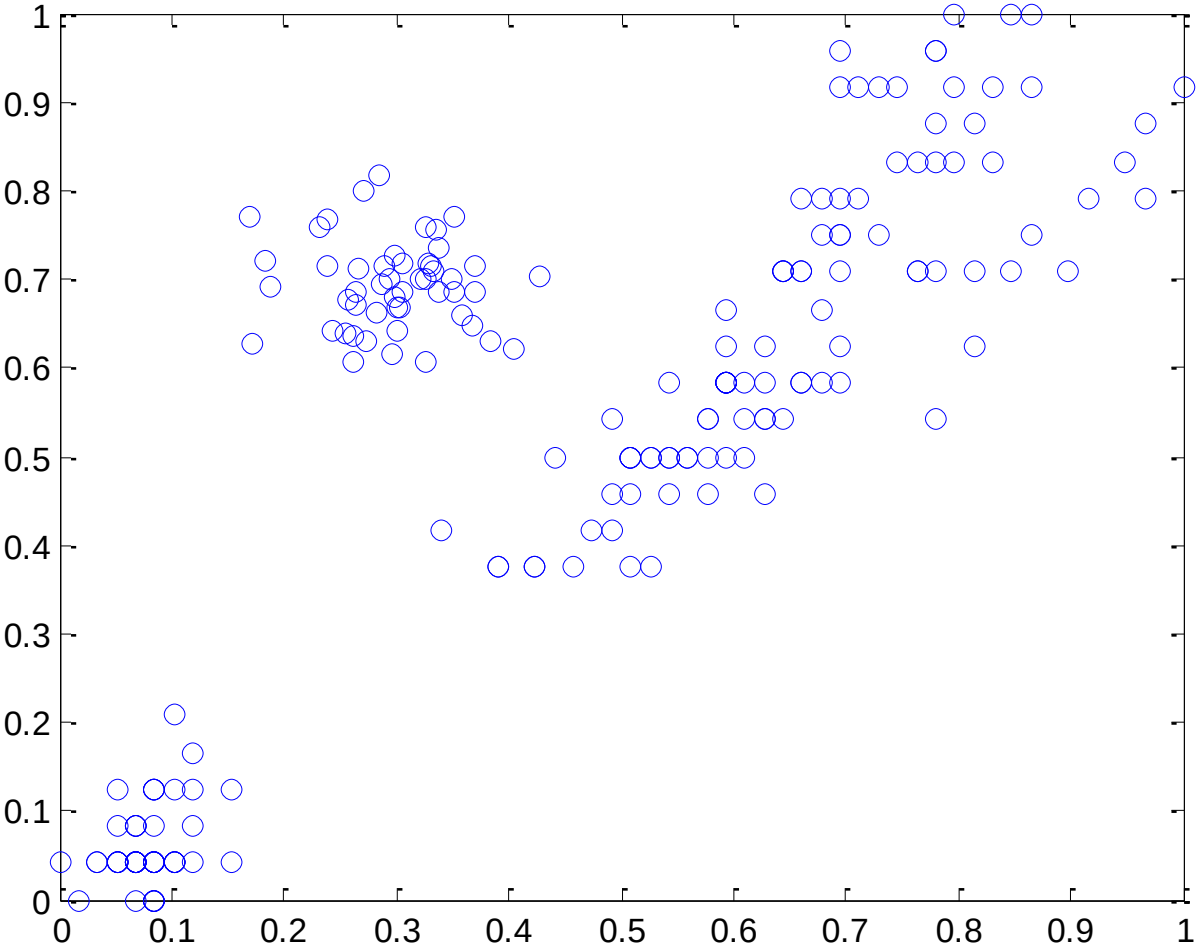
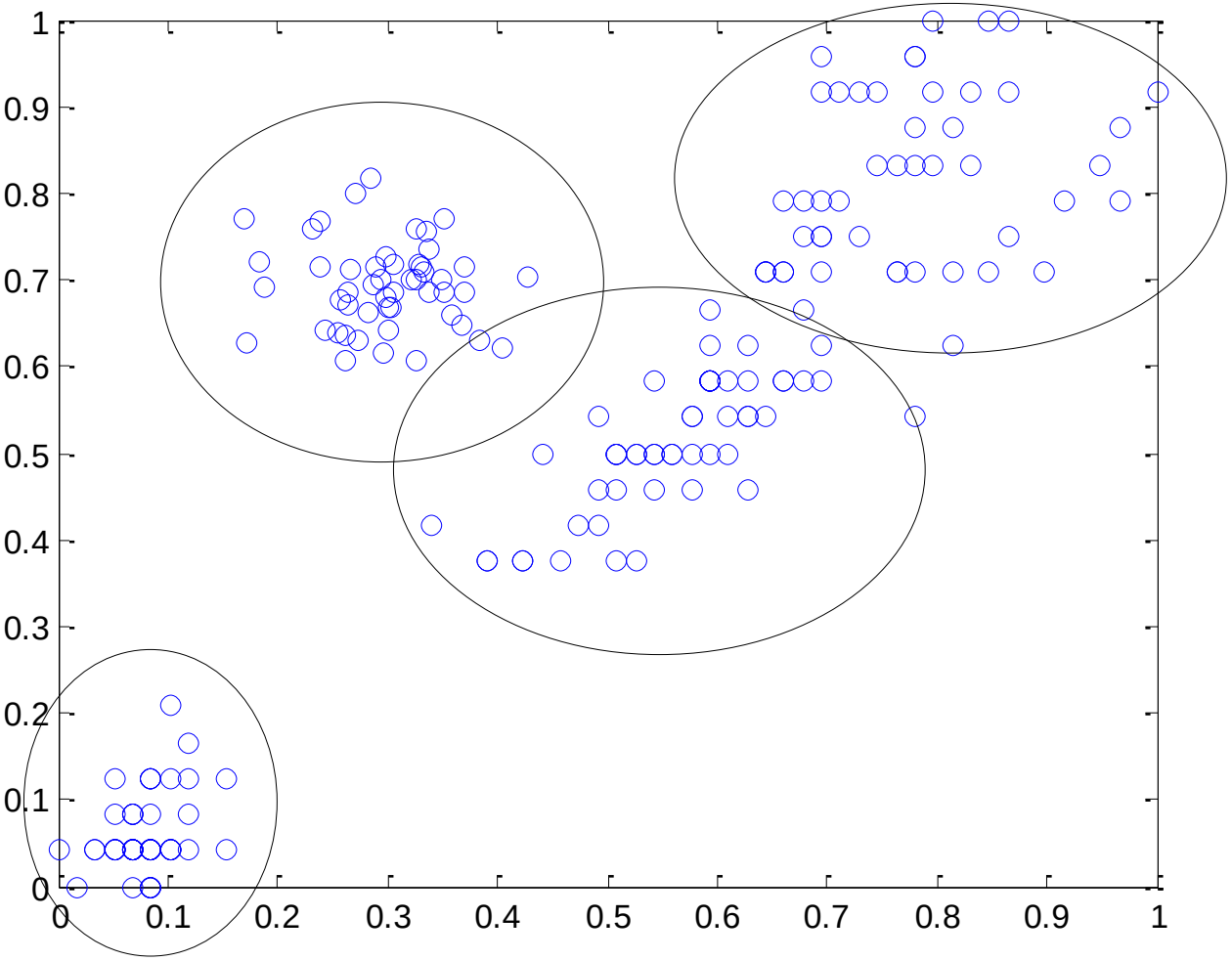
Grupowanie danych

Co to jest grupowanie



Szukanie grup, obszarów stanowiących lokalne gromady punktów

Co to jest grupowanie



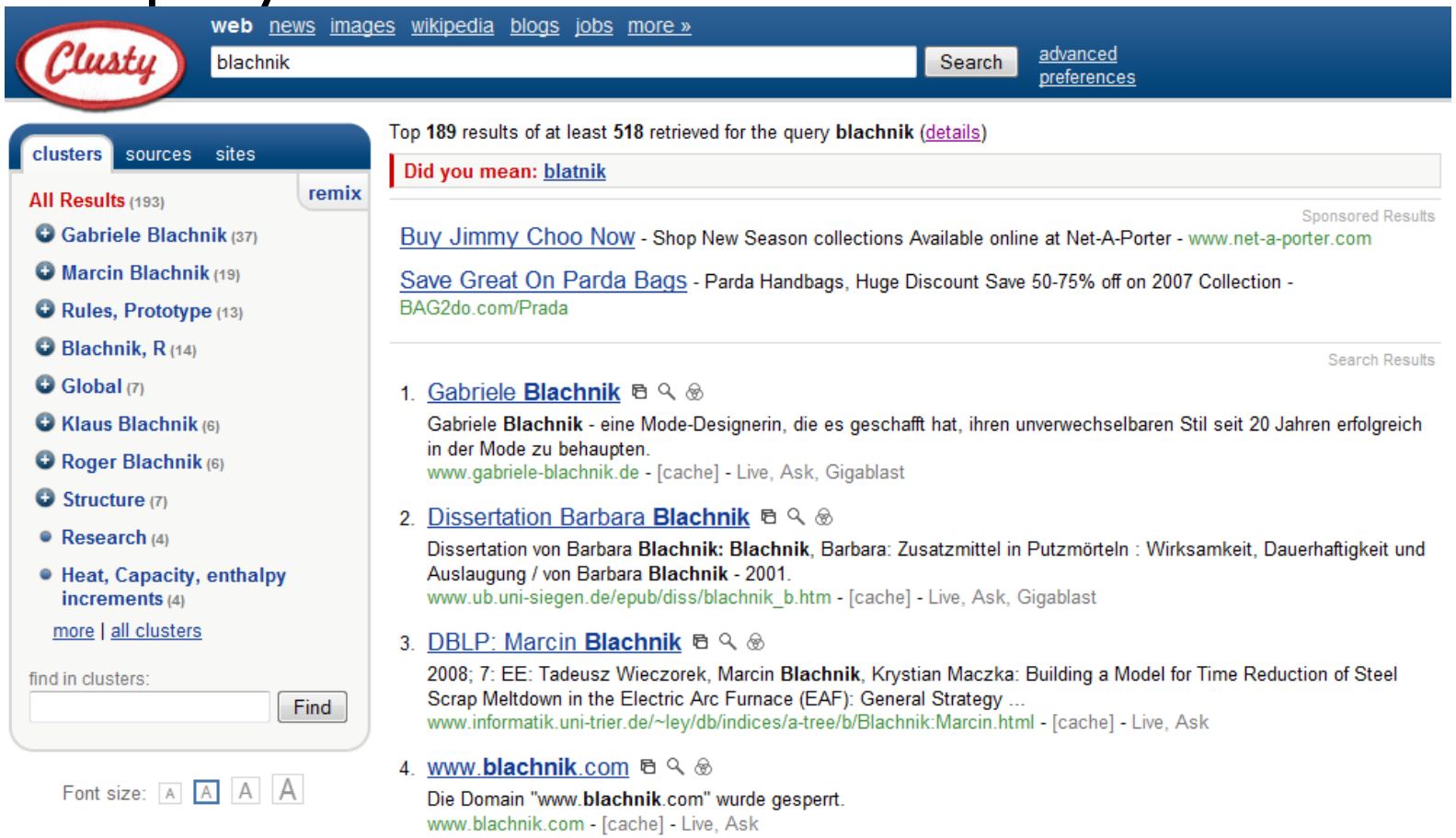
Grupow nie nienad

- analiza owanie danyc
  - Ucza o czego mam
  przykła honena,
  Przykła ny co się n ewno czeg

  Przykła kwiatów np. i ha oraz dług ę dowi rysa dysp

Co to jest grupowanie - przykłady

- Analiza dokumentów np.. www.clusty.com -> przykład Blachnik:

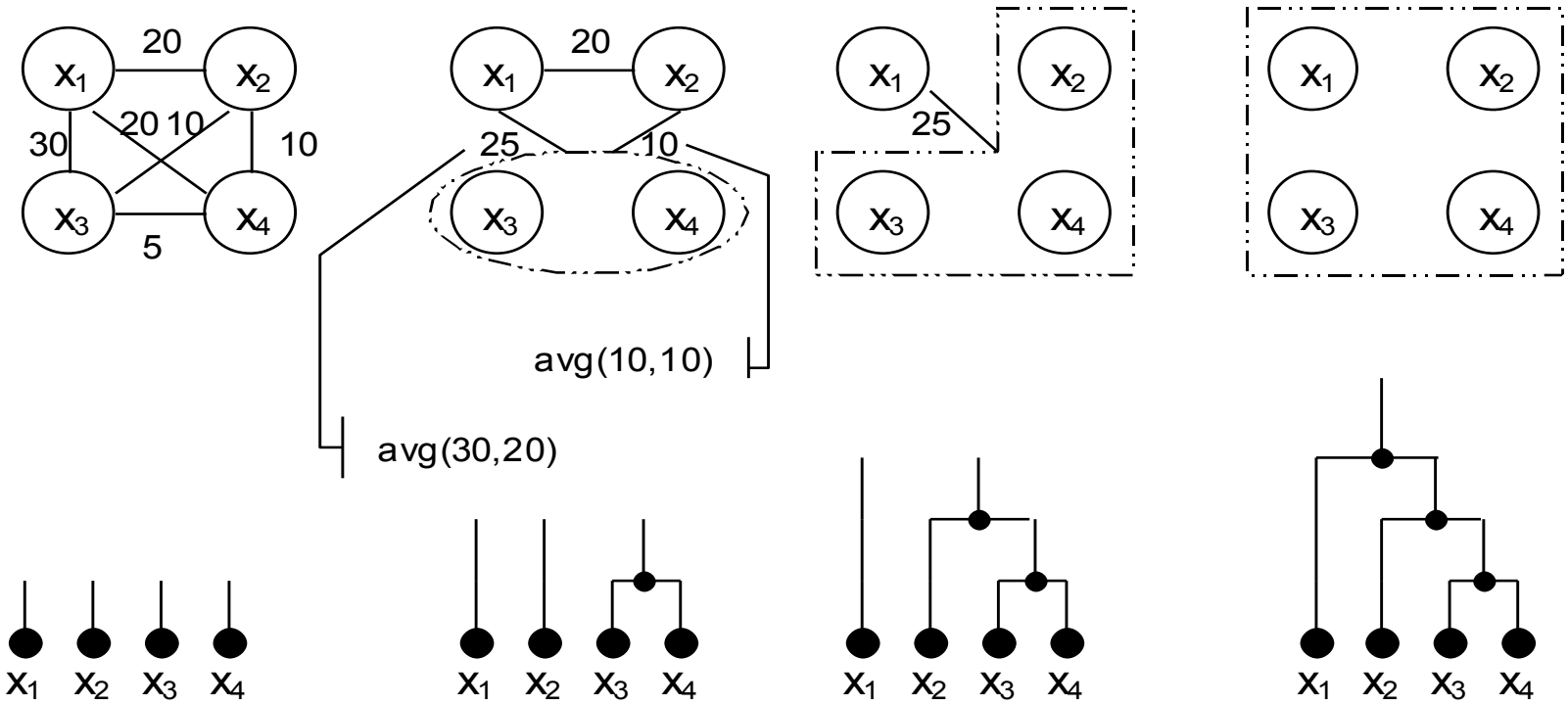


Podział metod grupowania danych

- Metody hierarchiczne
- Metody oparte o dekompozycję rozkładów prawdopodobieństwa
- Metody bazujące na minimalizacji skalarnego współczynnika jakości
- Metody oparte na teorii grafów,

Grupowanie chierarchiczne

Grupowanie chierarchiczne



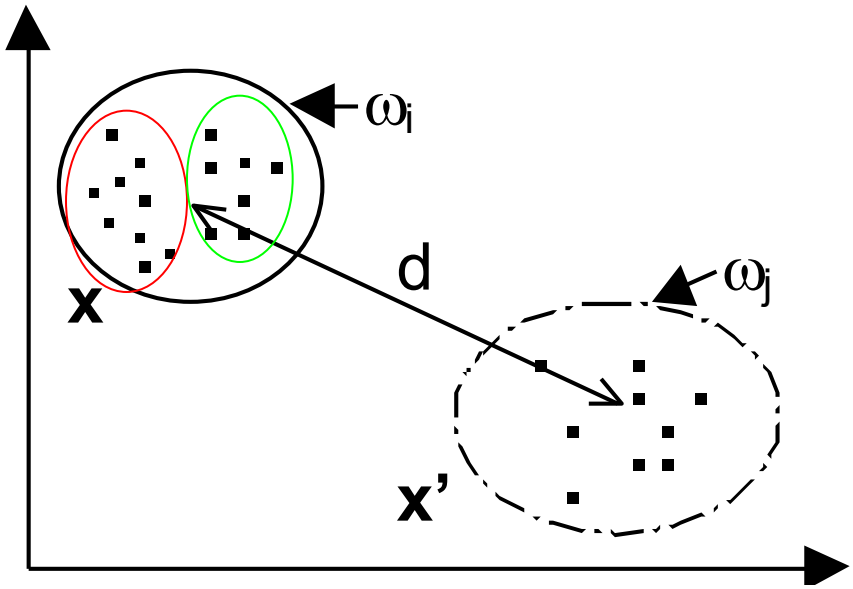
Odległości pomiędzy skupiskami

- Minimum – minimalna odległość pomiędzy elementami zbiorów **x** i **x'** $d_{\min}(\omega_i,\omega_j)=\min_{\mathbf{x}\in\omega_i,\ \mathbf{x}'\in\omega_j}\|\mathbf{x}-\mathbf{x}'\|$
- Maksimum - maksymalna odległość pomiędzy elementami zbiorów **x** i **x'**

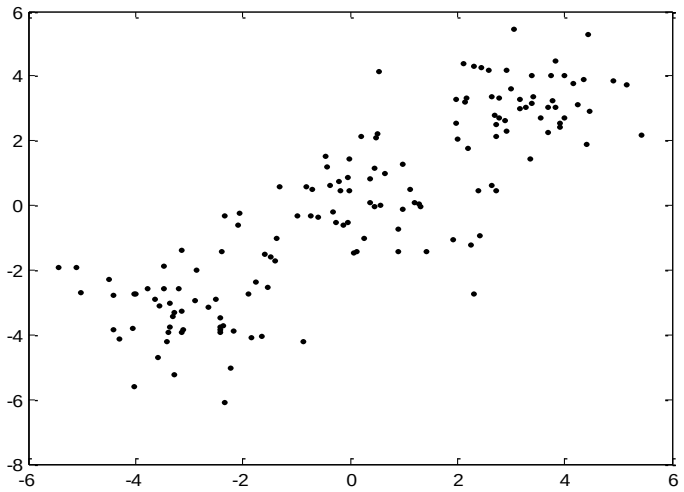
$$d_{\max}(\omega_i,\omega_j)=\max_{\mathbf{x}\in\omega_i,\ \mathbf{x}'\in\omega_j}\|\mathbf{x}-\mathbf{x}'\|$$

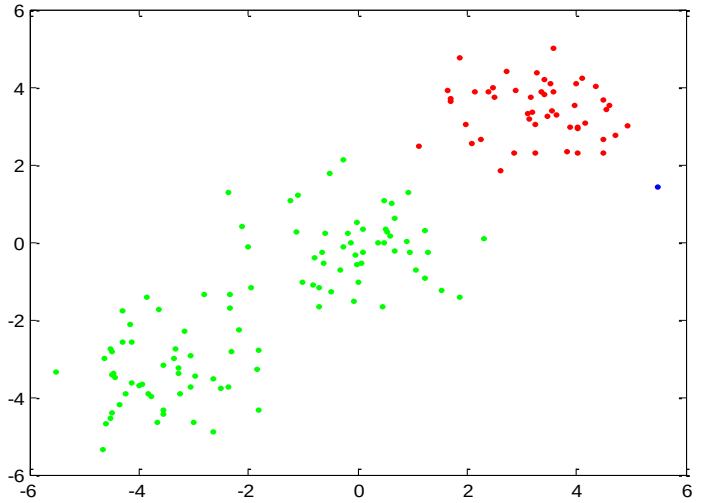
- Norma różnicy wartości średnich

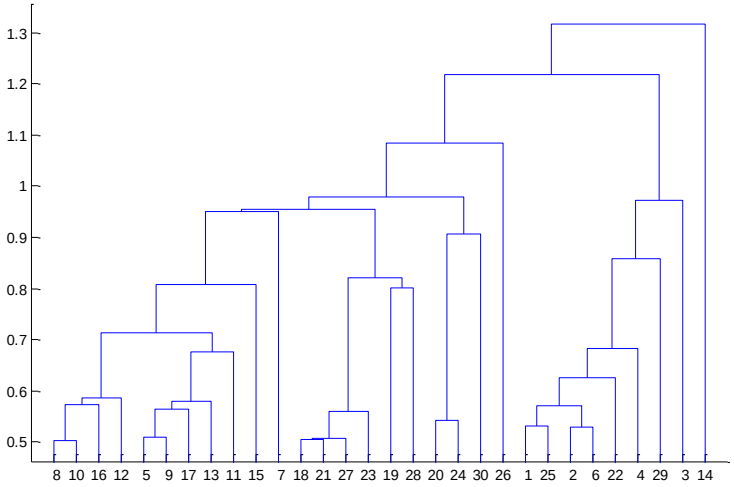
$$d_m(\omega_i,\omega_j)=\|m-m'\|$$



Przykład


6
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
6


6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6


1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
8 10 16 12 5 9 17 13 11 15 7 18 21 27 23 19 28 20 24 30 26 1 25 2 6 22 4 29 3 14

Metody oceny jakości grupowania

Cophenetic correlation coefficien

$$c = \frac{\Sigma_{i<j}(Y_{ij} - y)(Z_{ij} - z)}{\sqrt{\Sigma_{i<j}(Y_{ij} - y)^2 \Sigma_{i<j}(Z_{ij} - z)^2}}$$

Z - wektor odległości drzewa (m-1)x3

Y - wektor odległości obiektów niezgrupowanych (m-1)m/2 x 1

Y_{ij} – odległość pomiędzy obiektami i oraz j w Y.

Z_{ij} – odległość pomiędzy obiektami i oraz j w Z.

y i z to odpowiednio wartości średnie Y i Z.

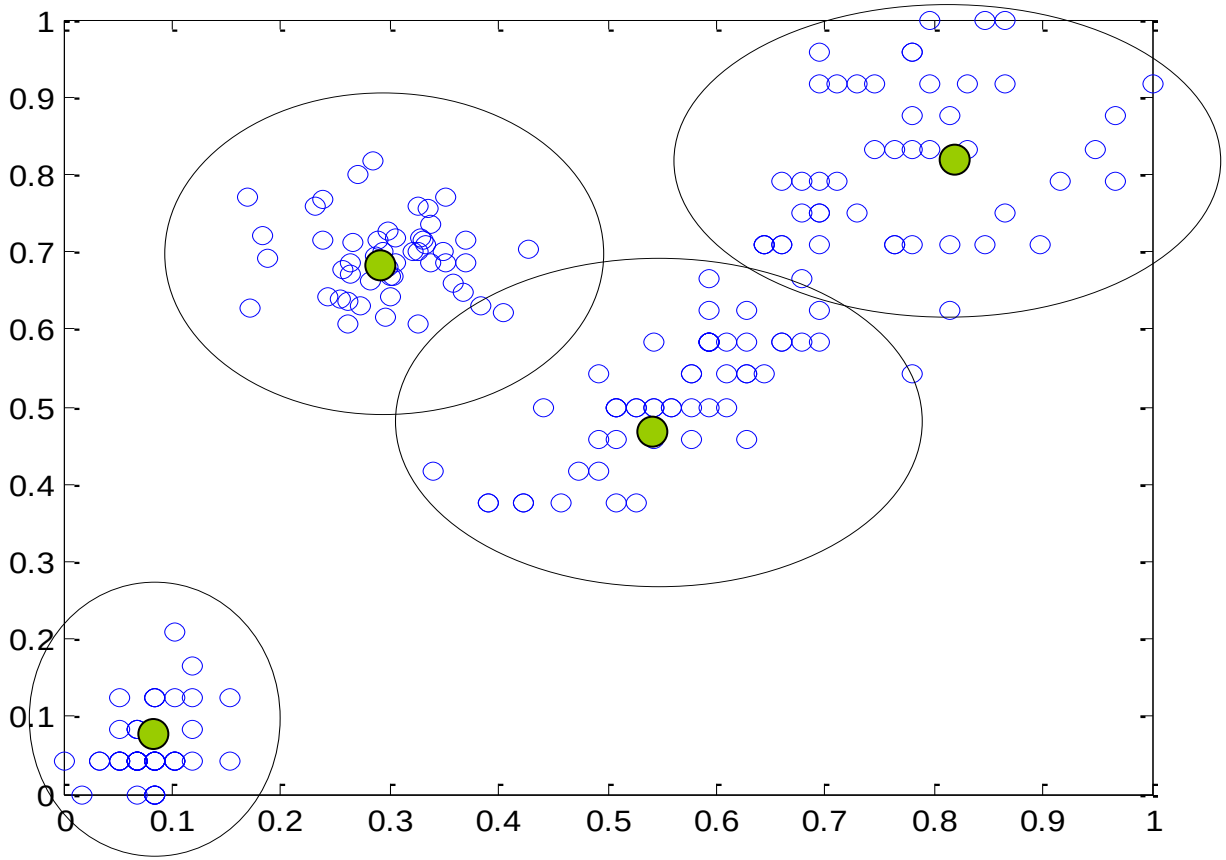
Grupowanie oparte o minimalizację skalarnego współczunnika jakości

Algorytm K-średnich

Co to jest skalarny wsp. jakości

- Skalarny współczynnik jakości oznacza funkcję kosztu która informuje jak „dobry” jest dany podział na grupy.
- Najprostsza postać skalarnego współczynnika jakości to suma odległości środka centrum klastra w stosunku do wszystkich wektorów w danej grupie.

Przykład



Oznaczenia

- K – liczba wektorów, obiektów
- C – liczba klasterów na które chcemy dokonać podziału
- x(k); k=1..K – k-aty element z wektora obiektów **X**
- ω_i ; i=1..C – i-ty element wektora klastrów $\boldsymbol{\omega}$
- v_i – centrum klastra = centrum grupy (grupy wektorów)

Założenia grupowania

Zbudować macierz podziału
$\mathbf{U}=[u_{ik}]$,dim($\mathbf{U}$)=C×K)

Warunki:

1° – każdy element macierzy u_{ik} należy do zbioru

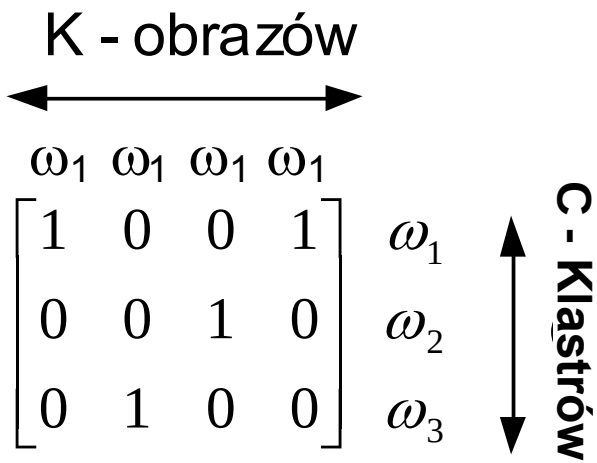
$$u_{ik} \in \{0,1\}$$

2° – w każdej kolumnie suma elementów równa jest 1

$$\sum_{i=1}^{C} u_{ik} = 1$$

3° – suma w wierszach należy do przedziału

$$\sum_{k-1}^{K} u_{ik} \in (0, K)$$



Algorytm k-średnich

Wskaźnik jakości:

$$\mathrm{J}(\mathbf{U}) = \sum_{k=1}^{K}\sum_{i=1}^{C} u_{ik} d_{ik}^2$$

gdzie:

$$d_{ik} = \left\| x(k) - v_i \right\|$$

v_i – zbiór (wektor) prototypów.

Algorytm k-średnich

1. Przyjmujemy macierz podziału **U** spełniającą trzy przedstawione uprzednio warunki
2. Wyznacza się położenie prototypów:

$$\forall_{1\le i\le C} v_i = \frac{\sum_{k=1}^{K} u_{ik}\mathbf{x}(k)}{\sum_{k=1}^{K} u_{ik}}$$

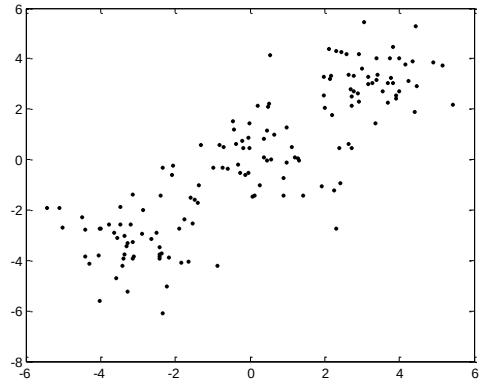
3. zwiększa się licznik iteracji z=z+1,
4. szukamy macierzy **U** tak, by wyznaczyć dla każdego elementu wektora danych **x** minimalną odległość od wzorców

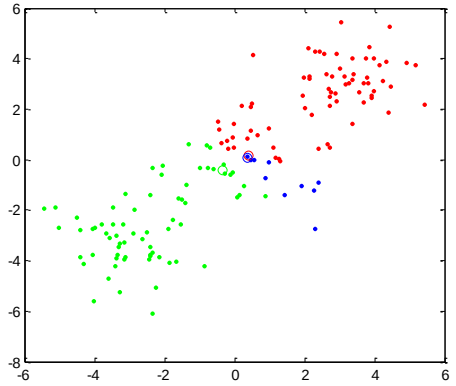
$$\forall_{1\le i\le C}\ \forall_{1\le k\le K} u_{ik} = \begin{cases} 1 & \min_j \left\| x(k) - v_j \right\| = \left\| x(k) - v_i \right\| \\ 0 & w\ \ pozostaych\ \ przypadkach \end{cases}$$

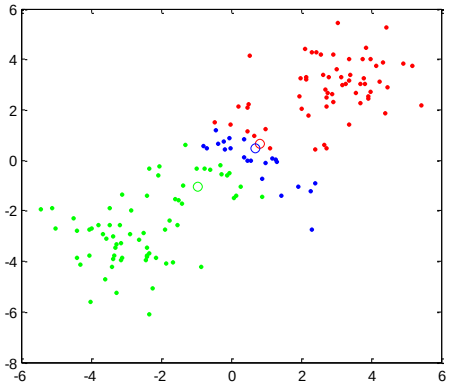
5. Sprawdzamy czy spełniony jest warunek

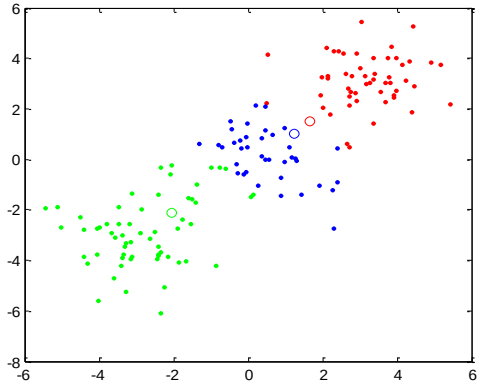
$$\left\| U^{(z)} - U^{(z-1)} \right\| < \varepsilon$$

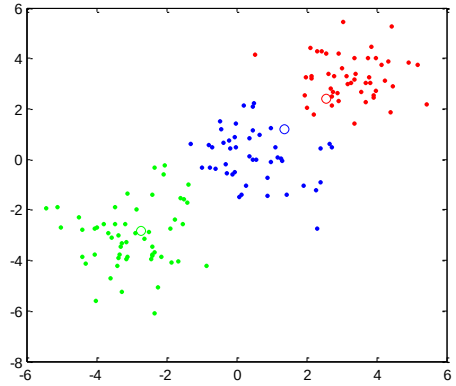
6. Jeśli różnica pomiędzy macierzami **U** w kolejnych iteracjach jest mniejsza od założonego ε to kończymy proces iteracji, jeśli nie to idź do 2

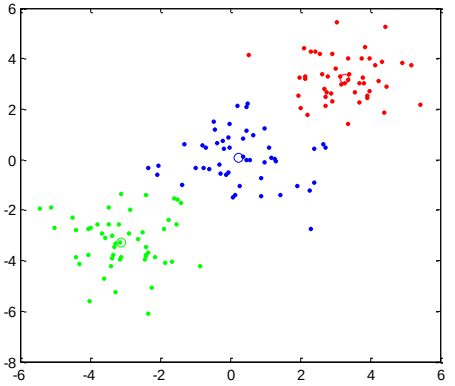












Przykład

Algorytm VQ

Algorytm VQ

- Algorytm kwantyzacji wektorów (VQ = ang. vector quantization)
- Podobna zasada działania do k-Średnich ale aktualizacji położenia wektora odbywa się po każdej prezentacji wektora uczącego
- W sieciach VQ centrum klastra nazywane jest również wektorem kodującym, które w terminologii sieci neuronowych nazywane są również neuronami.
- W sieciach VQ wagi neuronów określają położenie neuronu w przestrzeni wejściowej (podobnie jak w algorytmie k-Średnich)

Formuła aktualizacji

- Formuła aktualizacji położenia wektorów kodujących:

$$\mathbf{v}_i = \mathbf{v}_i + \alpha\left(\mathbf{x}_j - \mathbf{v}_i\right)$$

Gdzie:

- $\mathbf{v}_i$ wektor kodujący podlegający aktualizacji (wektor kodujący nażący najbliżej wektora $\mathbf{x}_j$)
- α Współczynnik uczenia – maleje z każdą iteracją programu
- $\mathbf{x}_j$ j-ty wektory uczący

Algorytm VQ

1. Zainicjuj położenie wektorów kodujących
2. Iteracyjnie *l*-razy
   1. Dla każdego wektora treningowego
      1. Znajdź najbliższy wektor kodujący (dla danej metryki)
      2. Dokonaj aktualizacji położenia (wag) neuronu zgodnie z zależnością (1)
   2. Dokonaj aktualizacji wsp. α wg. zależności (2)

Algorytm VQ

Inicjalizacja wektorów kodujących:

1. Przez wybór wektorów kodujących spośród wektorów uczących –
   - w tym celu najlepiej jest wykorzystać istniejące już wektory danych i losowo wybrać ze zbioru danych wektory których położenie będzie inicjowało położenie neuronów
2. Poprzez w pełni losową inicjalizację – zbadaj przestrzeń zmienności zmiennych wejściowych i wylosuj położenia wektorów kodujących
3. Poprzez PCA – dokonaj redukcji wymiarowości korzystając z PCA, wyznacz w zredukowanej przestrzeni położenia wektorów kodujących i przetransformuj je do przestrzeni oryginalnej

Algorytm VQ a k-Średnich

- Który lepszy?
- Algorytm k-Średnich – większa złożoność pamięciowa – konieczność przechowywania macierzy przynależności.
  Dla dużych danych i dużej liczby klasterów jest to istotne ograniczenie
- Algorytm VQ – mniejsza złożoność pamięciowa ale większa złożoność obliczeniowa –
  - każdy wektor treningowy powoduje konieczność aktualizacji położenia prototypu
  - Aktualizacja położenia prototypu powoduje konieczność aktualizacji – ponownego liczenia odległości.
  - Czuły na kolejność prezentacji wzorców – dlatego zalecane jest na wstępie losowe poukładanie wektorów.
- Z doświadczenia – algorytm VQ jest lepszy od k-Średnich dzięki pozytywnemu wpływowi chaosu!!!

Sieci SOM

Sieci SOM

- SOM – samoorganizujące mapy Kochonena
- Idea podobna do VQ z pewnymi dodatkami
- Dodatki to narzucenie wymogów sąsiedztwa na wektory kodujące
- Strategia typu WTM – Winner Take Most, w sieciach VQ strategia WTA – Winner Take All

Strategia WTM