Grupowanie

Dany jest zbiór obiektów (rekordów). Chcemy znaleźć naturalne pogrupowanie obiektów w klasy (klastry, skupienia) obiektów o podobnych cechach.

Proces grupowania: grupowanie obiektów, rzeczywistych bądź abstrakcyjnych, w klasy, nazywane klastrami lub skupieniami, o podobnych cechach.

Problem

Grupowanie może dotyczyć zarówno obiektów rzeczywistych np.:

- pacjentów
- sekwencji DNA
- dokumentów tekstowych

Grupowane mogą być również obiekty abstrakcyjne np.:

- sekwencja dostępów do stron WWW
- grafy reprezentujące dokumenty XML

Klaster

Obiekty grupujemy w klastry (skupienia).

Istnieje wiele definicji pojęcia klastra.

- Zbiór obiektów, które sa “podobne”.
- Zbiór obiektów, takich, że odległość pomiędzy dwoma dowolnymi obiektami należącymi do klastra jest mniejsza niż odległość pomiedzy dowolnym obiektem należącym do klastra i dowolnym obiektem nie należącym do tego klastra.
- Spójny obszar przestrzeni wielowymiarowej, charakte-ryzujący sie dużą gęstością występowania obiektów.

Klaster

Przykład

Rozważmy zbiór dokumentów.

Interesują nas zbiory punktów w przestrzeni wielowymiarowej, w której pojedyńczy wymiar odpowiada jednemu słowu z określonego słownika.

Współrzędne dokumentu w przestrzeni są definiowane (względną) częstością występowania słów ze słownika.

Klastry dokumentów odpowiadaja grupom dokumentów dotyczących podobnej tematyki.

Proces grupowania

W procesie grupowania wyróżniamy następujące etapy:

- Reprezentacja obiektów w tym ekstrakcja/selekcja cech obiektów
- Definicja miary podobieństwa pomiędzy obiektami (zależy od dziedziny zastosowań)
- Grupowanie obiektów (klastry)
- Znajdowanie charakterystyki klastrów

Kluczowa w procesie grupowania jest odpowiednia miara podobieństwa.

Miary odległości

Pojęcie podobieństwa (bliskości) związane jest z pojęciami odległości i metryki. Przypomnijmy:

1) $d(i, j) \geq 0$ dla każdego i oraz j, ponadto $d(i, j) = 0$ wtedy i tylko wtedy, gdy $i = j$;
2) $d(i, j) = d(j, i)$ dla każdego i oraz j;
3) $d(i, j) \leq d(i, k) + d(k, j)$ dla każdego i, j oraz k.

Miary odległości

Niestety, w przypadku, gdy obiekty nie poddają się transformacji do przestrzeni euklidesowej, proces grupowania wymaga zdefiniowania innych miar odległosci (podobieństwa).

Dotyczy to takich obiektów jak np.:

sekwencje dostępów do stron WWW, sekwencje DNA, sekwencje zbiorów, zbiory atrybutów kategorycznych, dokumenty tekstowe, XML, grafy etc.

Inne miary odległości

Przykład 1

Rozważmy zbiór stron WWW.

Reprezentujemy je jako punkty w przestrzeni wielowymiarowej, w której pojedyńczy wymiar odpowiada jednemu słowu z określonego słownika.

Podobienstwo (odległosc) d(x,y) stron x i y możemy zdefiniować jako znormalizowany iloczyn skalarny wektorów reprezentujacych strony x i y tzn. iloczyn skalarny x i y podzielony przez iloczyn ich długości.

Inne miary odległości

Przykład 1 (cd)

Przyjmijmy, że słownik składa się z 4 słów.

Rozważmy dwa dokumenty:

$$x=[2, 0, 3, 1] \text{ i } y=[5, 3, 2, 0]$$

Iloczyn skalarny $x \cdot y$ wynosi 16.

Długości wektorów x i y wynoszą odpowiednio $\sqrt{14}$ i $\sqrt{38}$.

Podobieństwo dokumentów wynosi zatem $\approx 0{,}7$.

Inne miary odległości

Przykład 2

Rozważmy sekwencje DNA.

Definicja odległości (podobieństwa) sekwencji symboli, powinna uwzględniać fakt, że sekwencje mogą miec różną długość oraz różne symbole na tych samych pozycjach np.: x=abcde i y=bcdxye.

Miara odległości: $D(x, y) = |x| + |y| - 2|LCS(x, y)|$

gdzie LCS oznacza najdłuszą wspólna podsekwencję (*ang. longest common subsequence*) (LCS(x,y) = bcde). D(x, y) = 3

Algorytmy grupowania

Istnieje wiele różnych metod i typów grupowania:

- Dla danych liczbowych i/lub danych symbolicznych
- Deterministyczne i probabilistyczne
- Rozłaczne i przecinające się
- Hierarchiczne i płaskie
- Monoteiczne i politeiczne
- Przyrostowe i nieprzyrostowe

Metody hierarchiczne

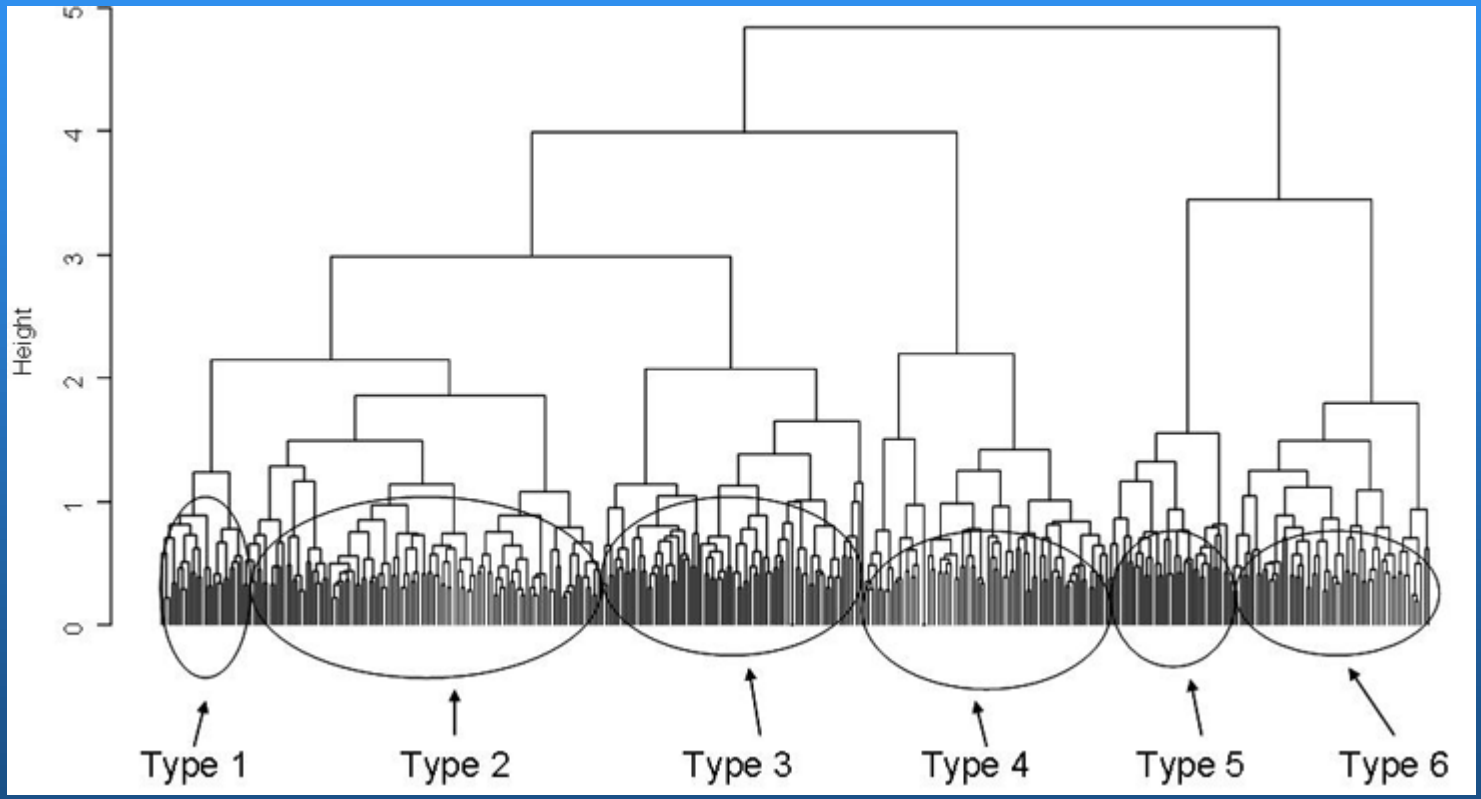
Jedno z dwóch podstawowych podejść do procesu grupowania obiektów to metody hierarchiczne.

Metody te generują zagnieżdżoną sekwencję podziałów zbiorów obiektów w procesie grupowania.

Otrzymujemy w ten sposób drzewo klastrów (tzw. dendrogram).

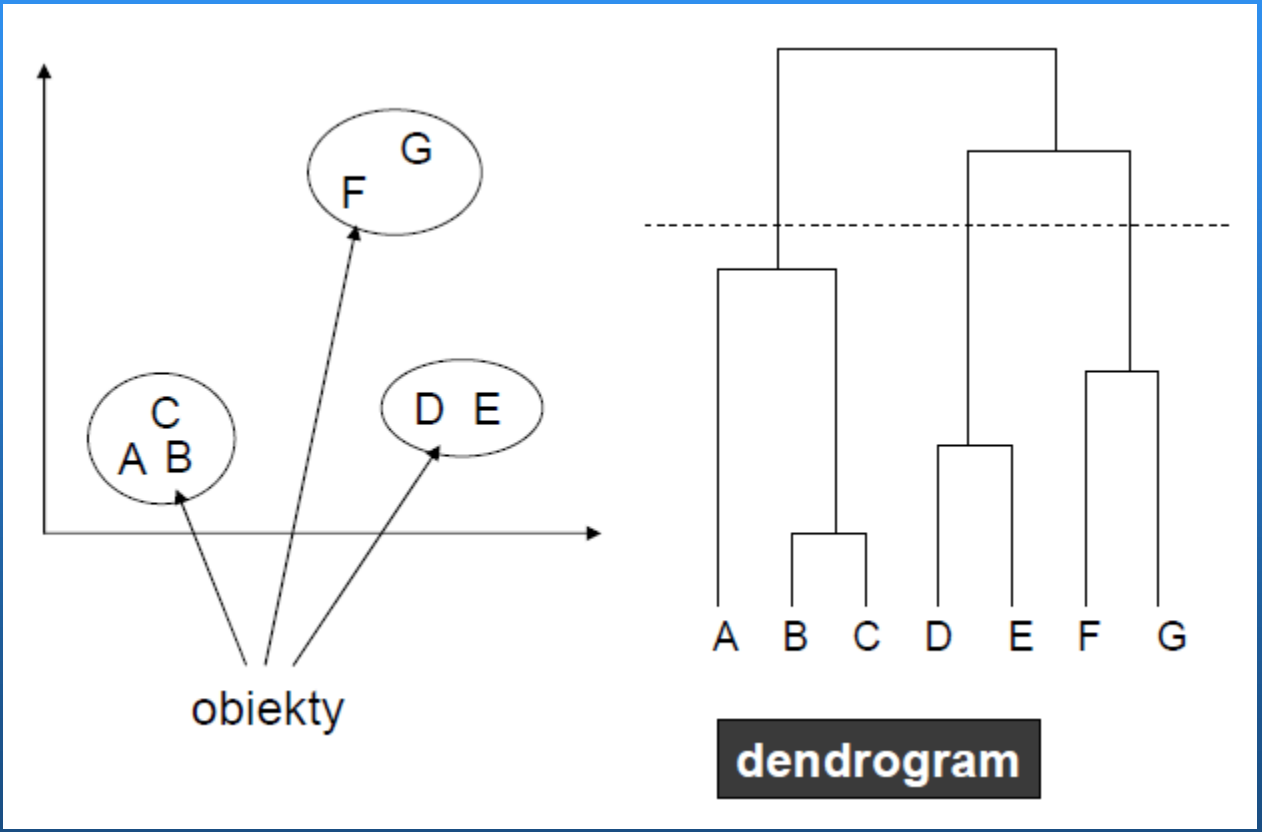
Metody hierarchiczne

Przykład



Metody hierarchiczne

Przykład



Metody hierarchiczne

Dwa podejścia:

Podejście podziałowe (top-down) - początkowo, wszystkie obiekty przypisujemy do jednego klastra; następnie, w kolejnych iteracjach, klaster jest dzielony na mniejsze klastry, które, z kolei, dzielone sa na kolejne mniejsze klastry.

Podejście aglomeracyjne (bottom-up) - początkowo, każdy obiekt stanowi osobny klaster, następnie, w kolejnych iteracjach, klastry sa łączone w większe klastry.

Metody hierarchiczne

W metodach hierarchicznych stosowanych jest kilka miar odległości między klastrami.

Wprowadźmy oznaczenia:

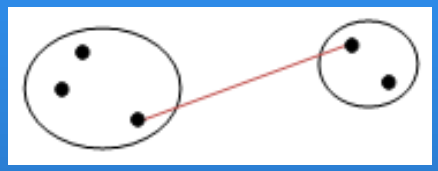
- $|| p - p' ||$ - oznacza odległość pomiędzy dwoma obiektami
- n_i - oznacza liczbę obiektów należących do klastra C_i

Metody hierarchiczne

Miary odległości

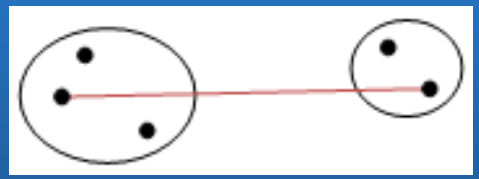
Minimalna odległość:

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



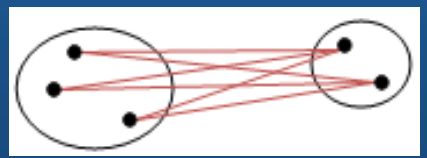
Maksymalna odległość:

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



Średnia odległość:

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

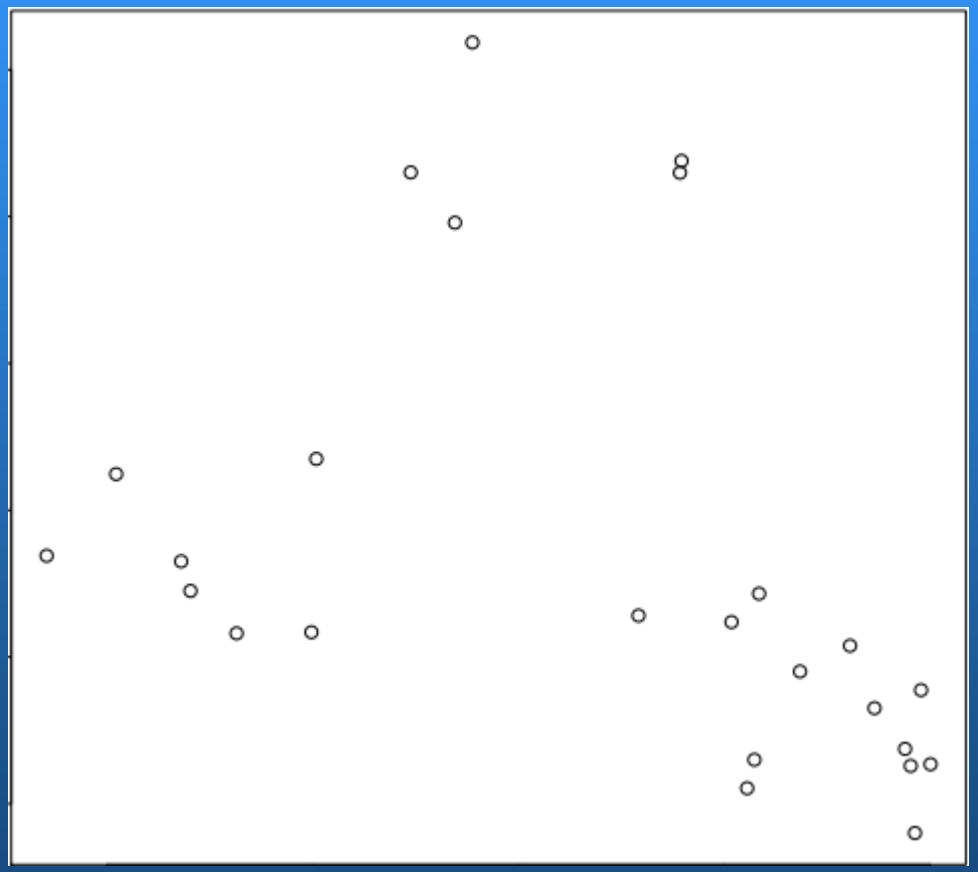


Algorytm hierarchiczny aglomeracyjny

- Umieszczamy każdy obiekt w osobnym klastrze. Tworzymy macierz przyległości zawierającą odległości pomiędzy każdą parą klastrów.
- Korzystając z macierzy przyległości znajdujemy najbliższą parę klastrów. Łączymy znalezione klastry tworząc nowy klaster. Uaktualniamy macierz przyległości po operacji połączenia.
- Jeżeli jeżeli uzyskaliśmy porządana ilość klastrów lub obiektów należących do jednego klastra, kończymy procedure grupowania, w przeciwnym razie przechodzimy do punktu 2.

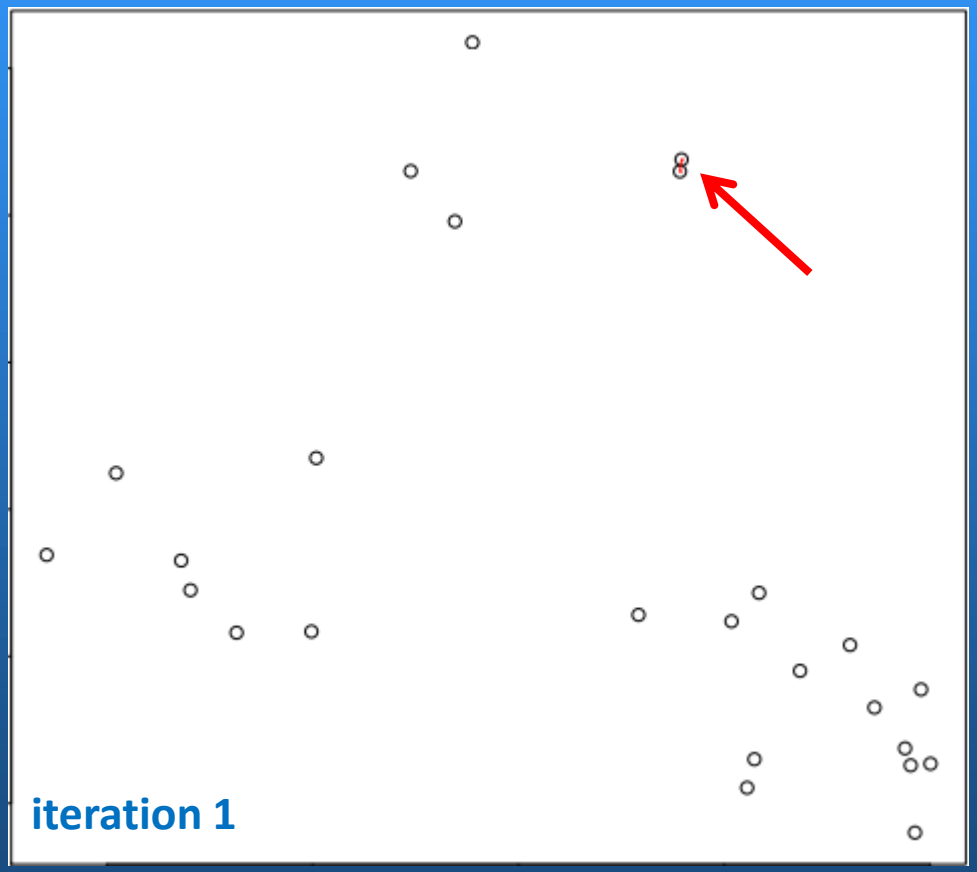
Algorytm hierarchiczny aglomeracyjny

Przykład



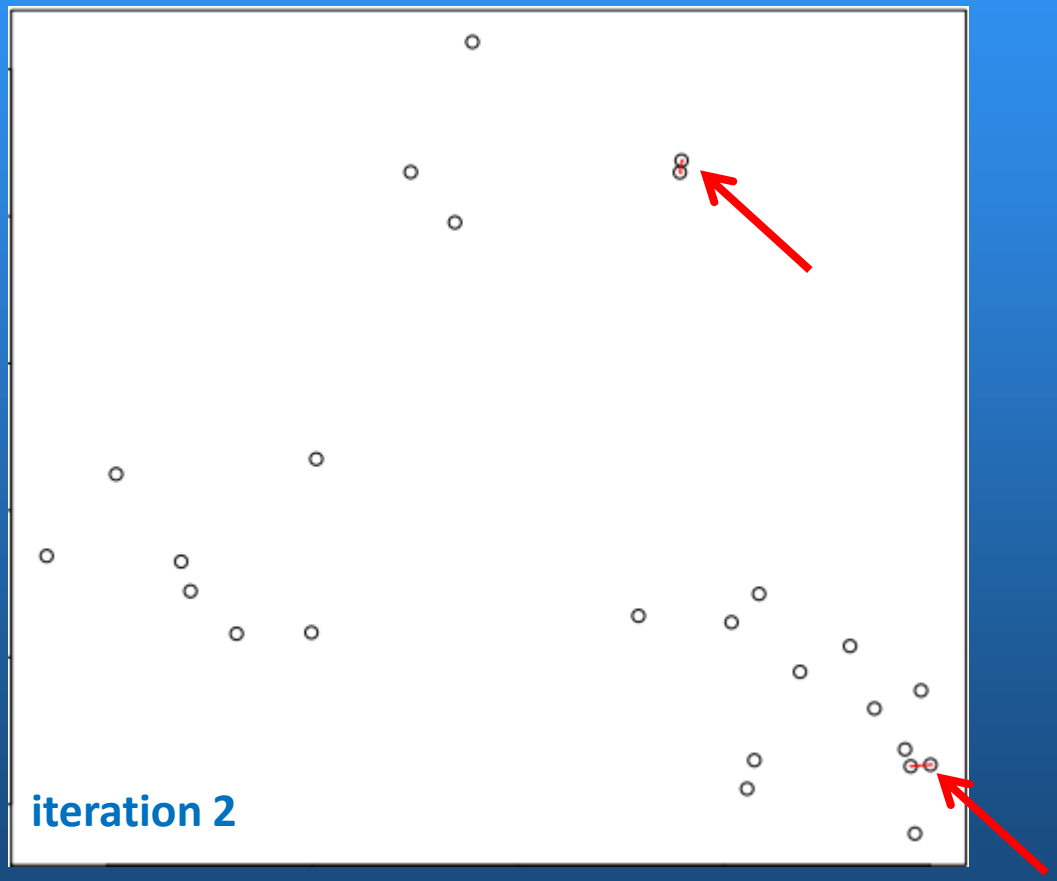
Algorytm hierarchiczny aglomeracyjny

Przykład



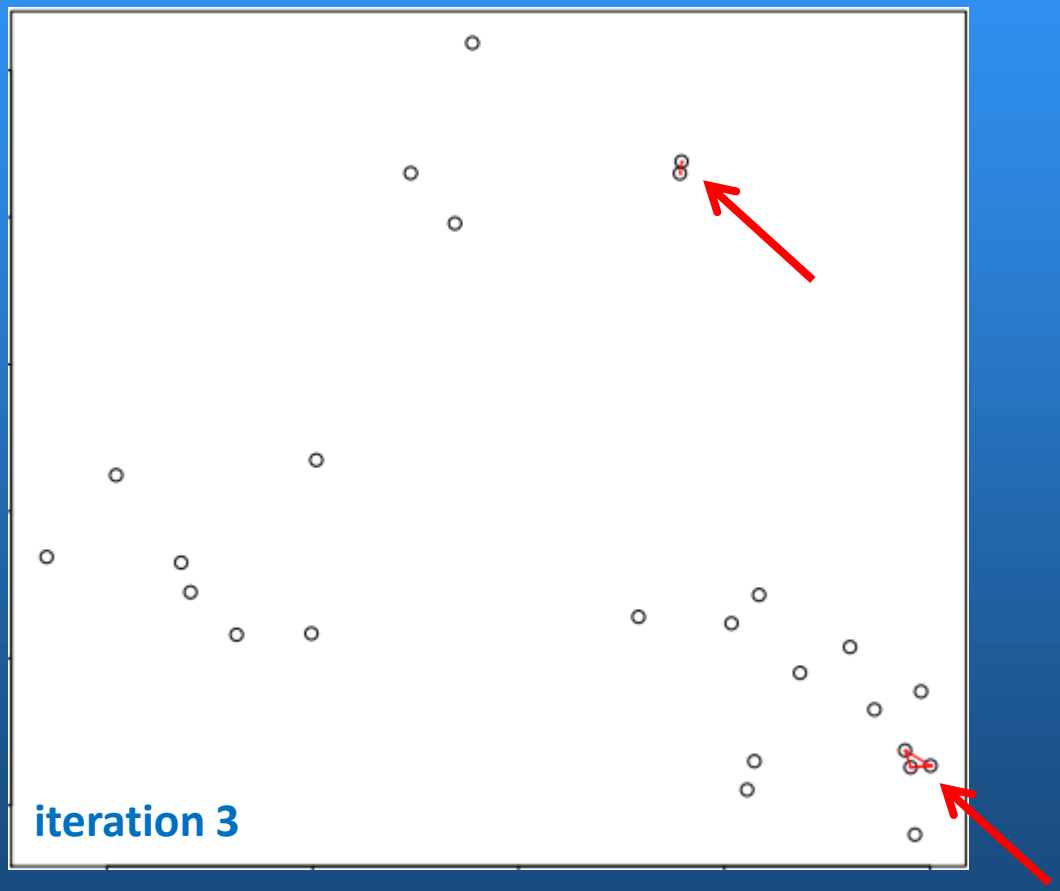
Algorytm hierarchiczny aglomeracyjny

Przykład



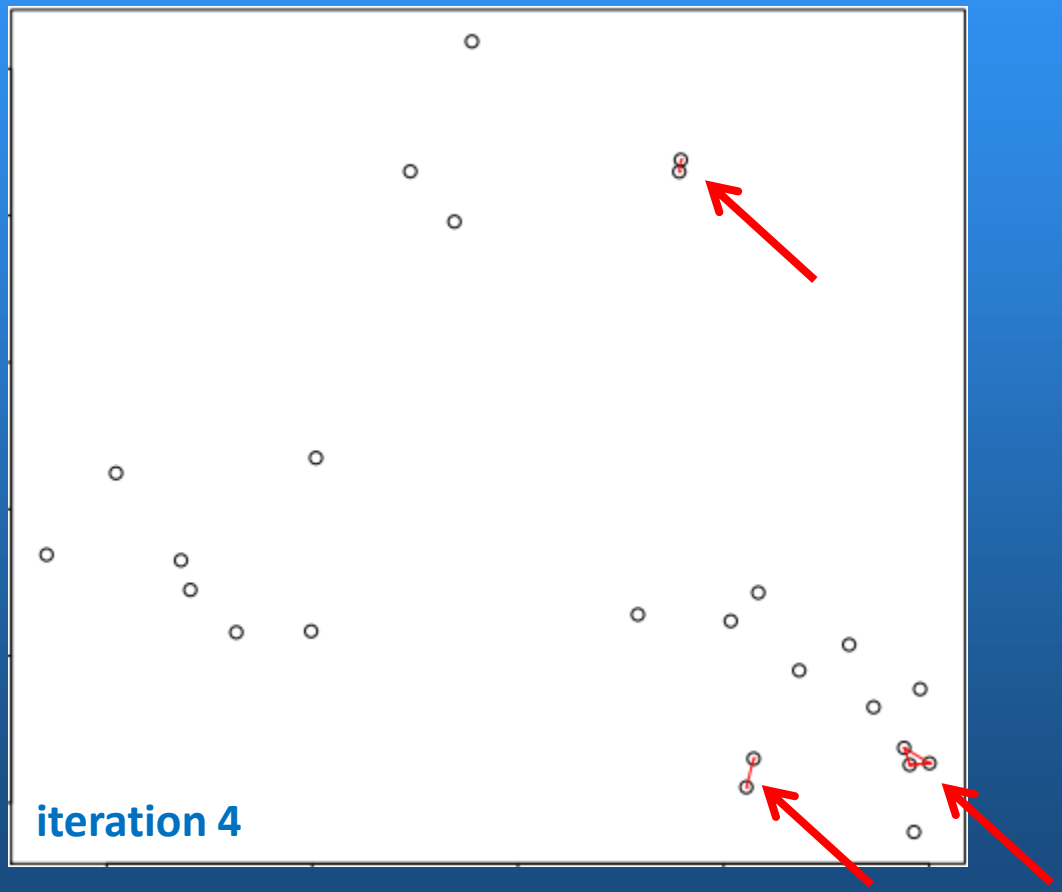
Algorytm hierarchiczny aglomeracyjny

Przykład



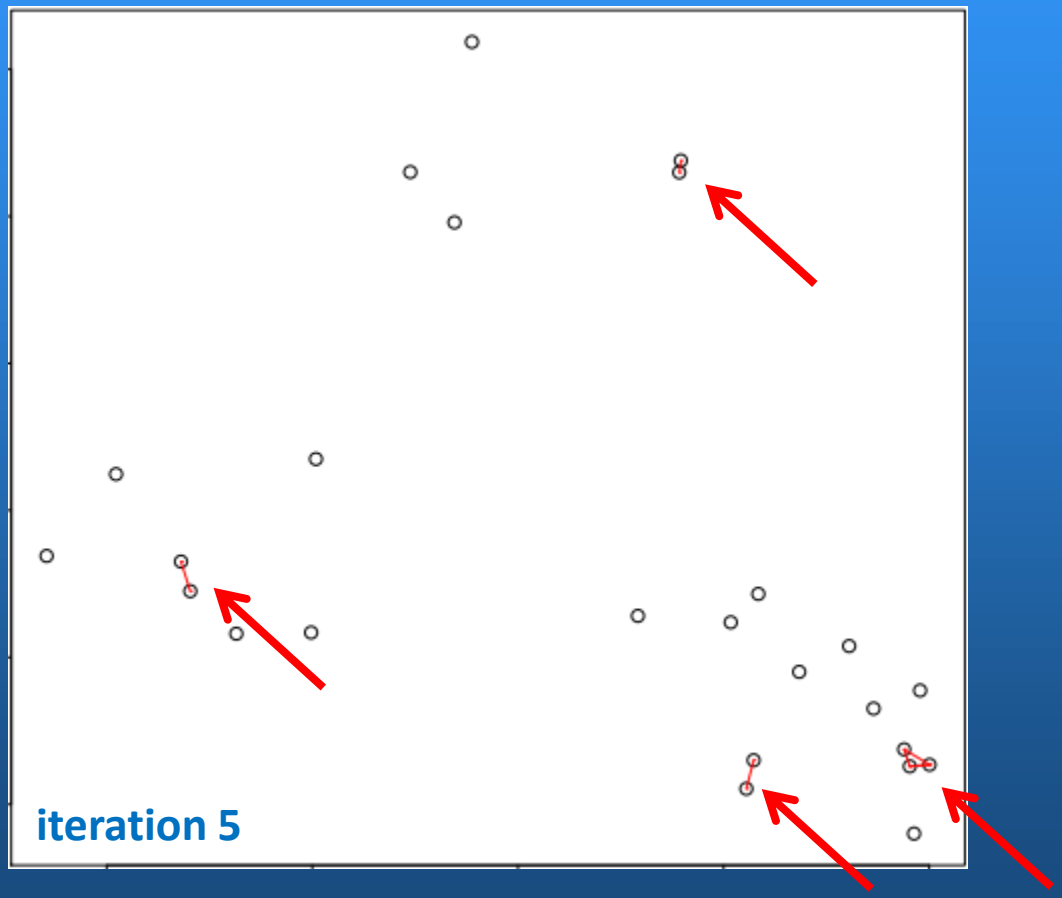
Algorytm hierarchiczny aglomeracyjny

Przykład



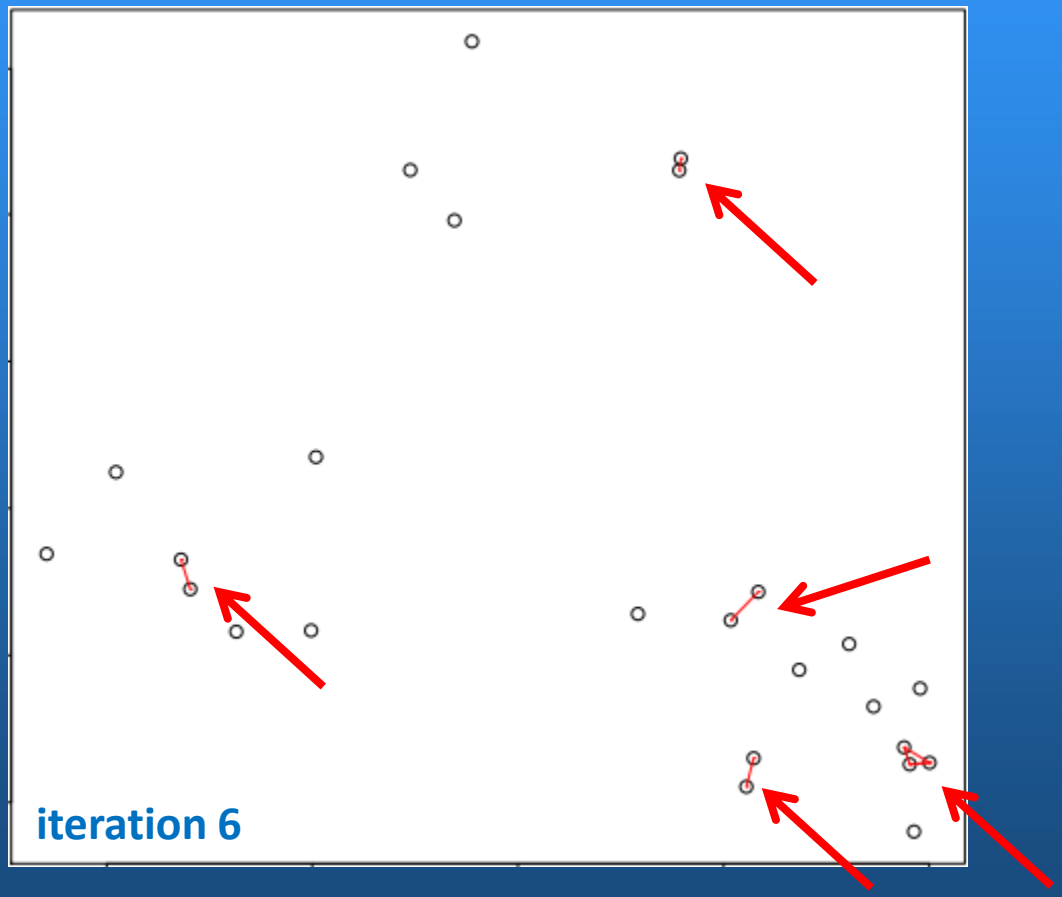
Algorytm hierarchiczny aglomeracyjny

Przykład



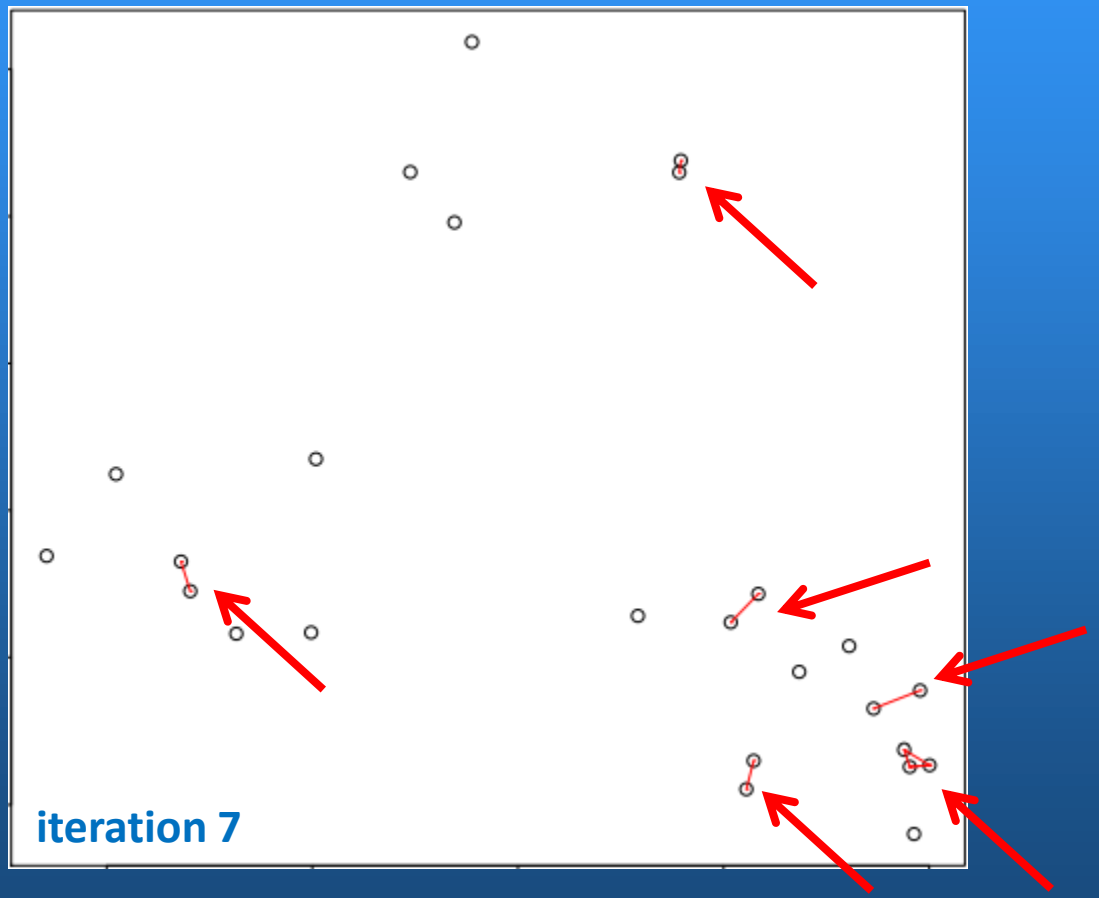
Algorytm hierarchiczny aglomeracyjny

Przykład



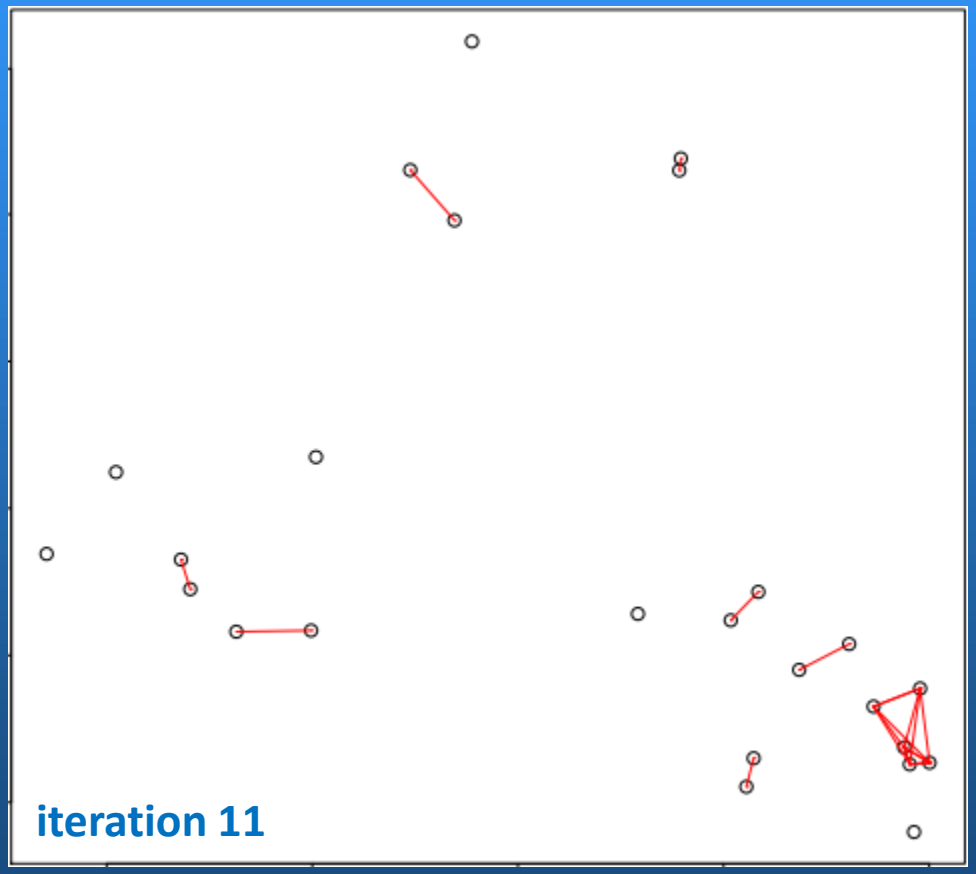
Algorytm hierarchiczny aglomeracyjny

Przykład



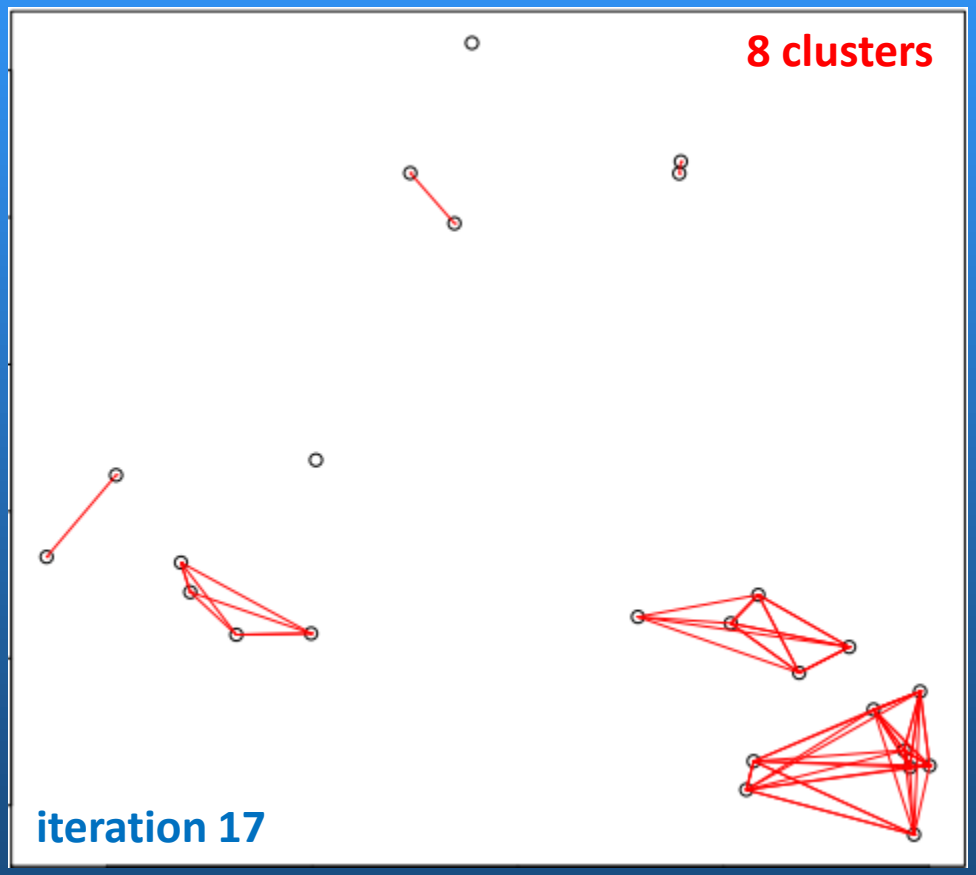
Algorytm hierarchiczny aglomeracyjny

Przykład



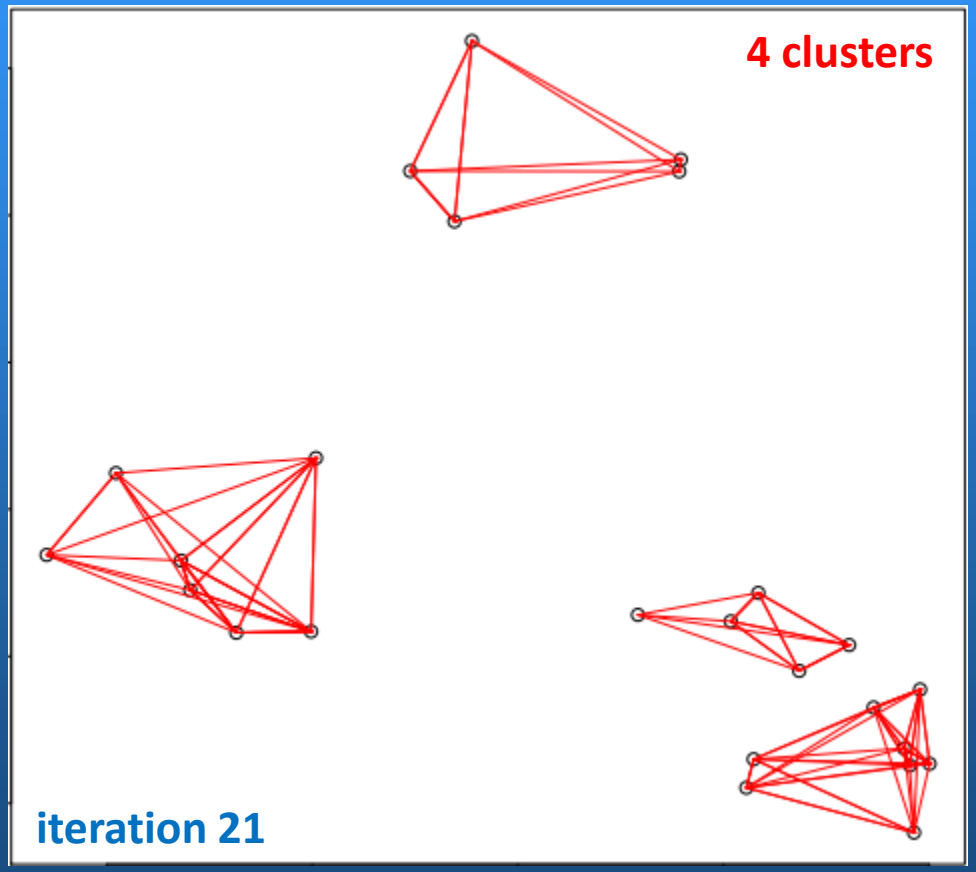
Algorytm hierarchiczny aglomeracyjny

Przykład



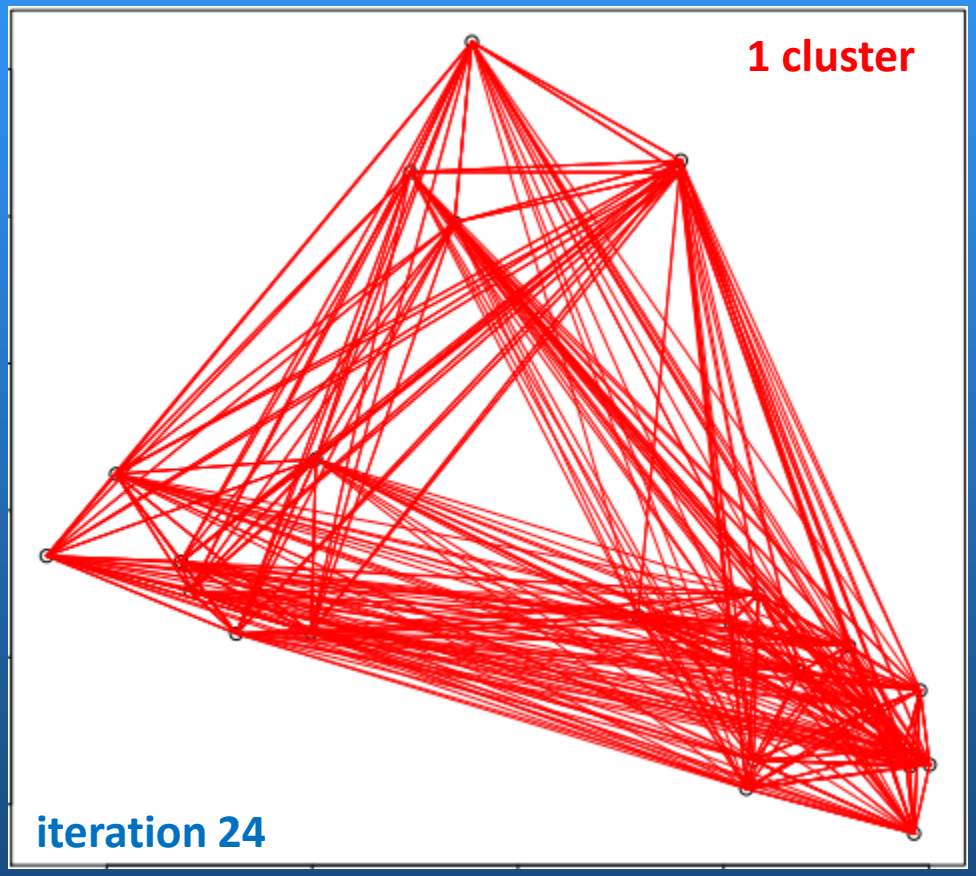
Algorytm hierarchiczny aglomeracyjny

Przykład

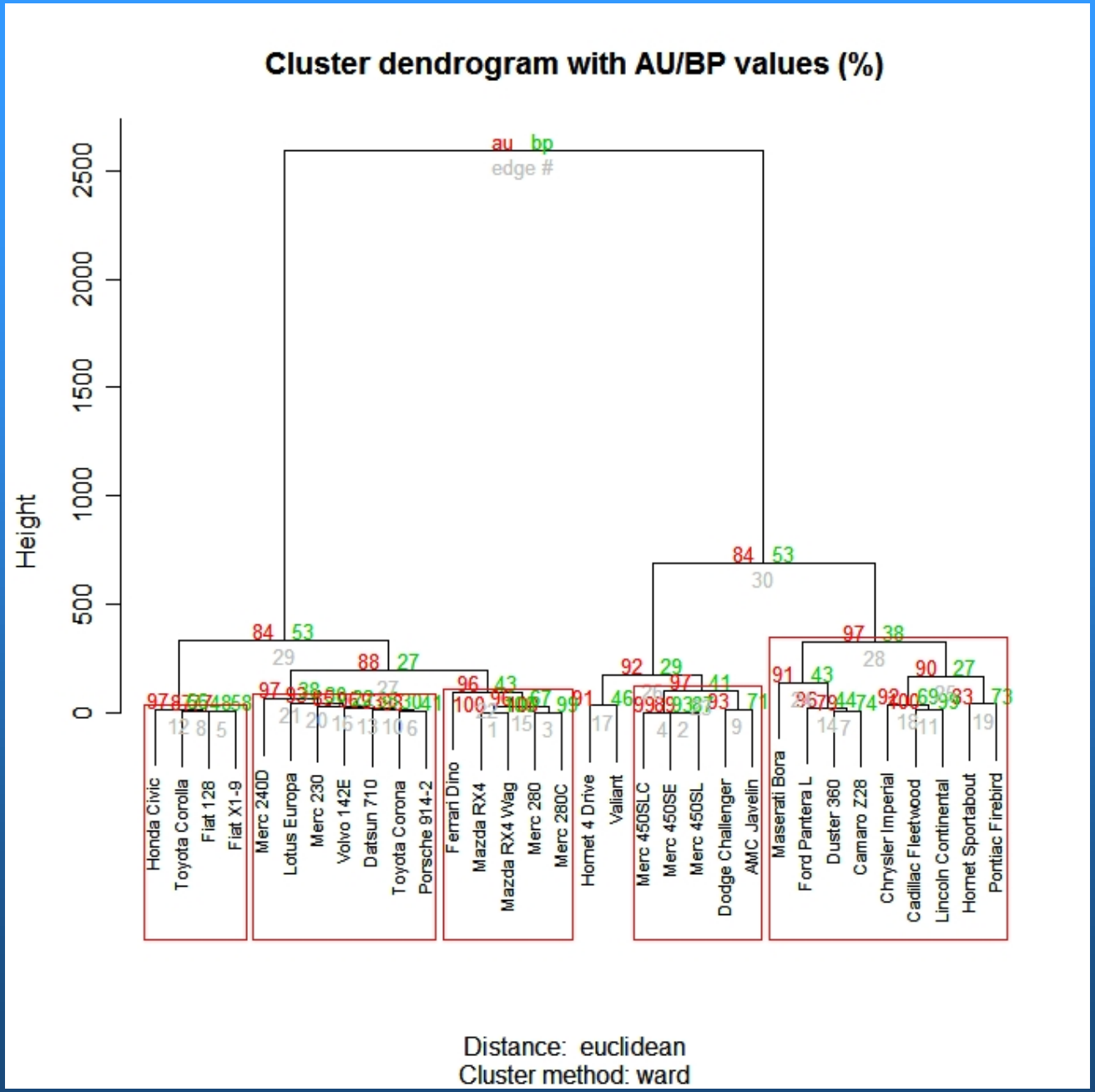


Algorytm hierarchiczny aglomeracyjny

Przykład



Algorytm…


Cluster dendrogram with AU/BP values (%)
au bp
edge #
Height
0
500
1000
1500
2000
2500
Honda Civic
Toyota Corolla
Fiat 128
Fiat X1-9
Merc 240D
Lotus Europa
Merc 230
Volvo 142E
Datsun 710
Toyota Corona
Porsche 914-2
Ferrari Dino
Mazda RX4
Mazda RX4 Wag
Merc 280
Merc 280C
Hornet 4 Drive
Valiant
Merc 450SLC
Merc 450SE
Merc 450SL
Dodge Challenger
AMC Javelin
Maserati Bora
Ford Pantera L
Duster 360
Camaro Z28
Chrysler Imperial
Cadillac Fleetwood
Lincoln Continental
Hornet Sportabout
Pontiac Firebird
Distance: euclidean
Cluster method: ward

Metody iteracyjno-optymalizacyjne

Metody iteracyjno-optymalizacyjne grupowania tworzą jeden podział zbioru obiektów (partycję) w miejsce hierarchicznej struktury podziałów, jak ma to miejsce w przypadku algorytmów hierarchicznych.

Tworzony jest początkowy podział obiektów (zbiór klastrów), a następnie, stosując technikę iteracyjnej realokacji obiektów pomiędzy klastrami, podział ten jest modyfikowany w taki sposób, aby uzyskać poprawę podziału zbioru obiektów pomiedzy klastry.

Algorytm k-średnich

- Dane wejściowe: liczba klastrów k oraz baza danych n obiektów
- Dane wyjściowe: zbiór k klastrów minimalizujący np. kryterium błędu średniokwadratowego.

Idea:

Wybieramy losowo początkowy podział zbioru n obiektów na k klastrów, a następnie realokujemy obiekty pomiędzy klastrami. Początkowy podział jest modyfikowany w taki sposób, aby uzyskać poprawę funkcji kryterialnej aż do osiagniecia warunku stopu. Jest to algorytm zachłanny.

Algorytm k-średnich

1. Wybieramy losowo k obiektów jako początkowe środki k klastrów

2. Dopóki występują zmiany przydziału obiektów do klastrów wykonuj:

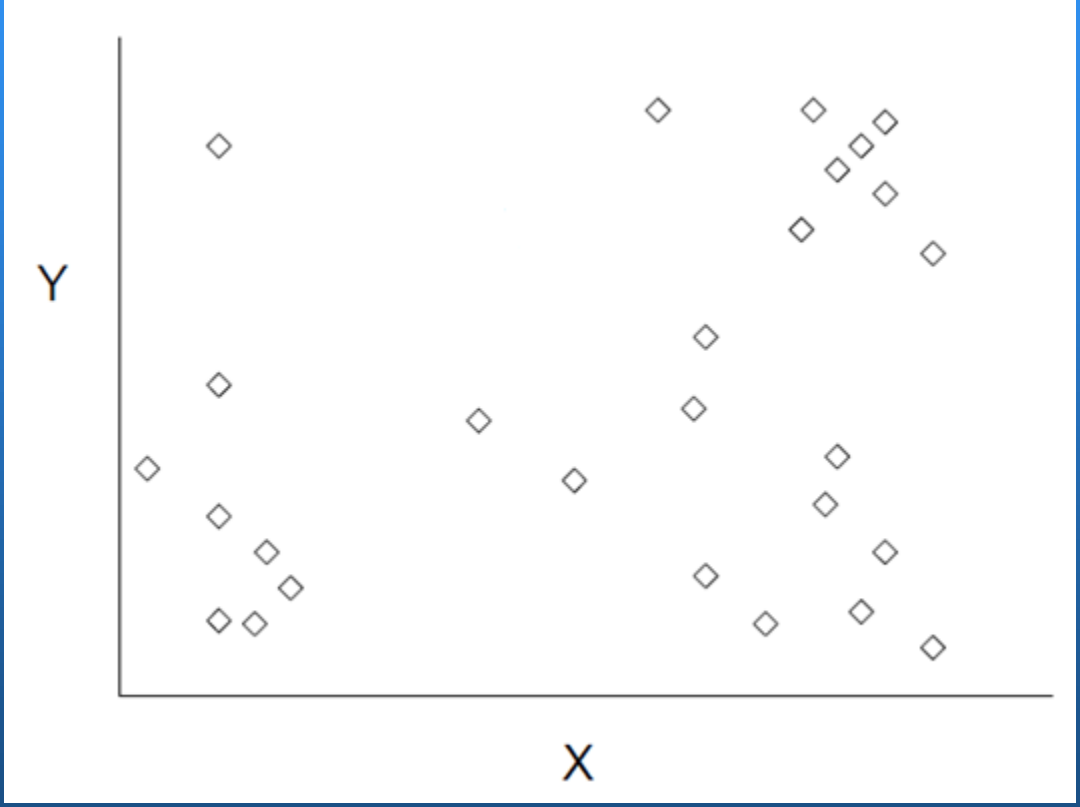
- Przydziel każdy obiekt do tego klastra, dla którego odległość obiektu od środka (średniej) klastra jest najmniejsza.

Średnia klastra: $r_k = \frac{1}{n_k} \sum x$

- Uaktualnij środki klastrów – środkiem klastra jest wartość średniej danego klastra

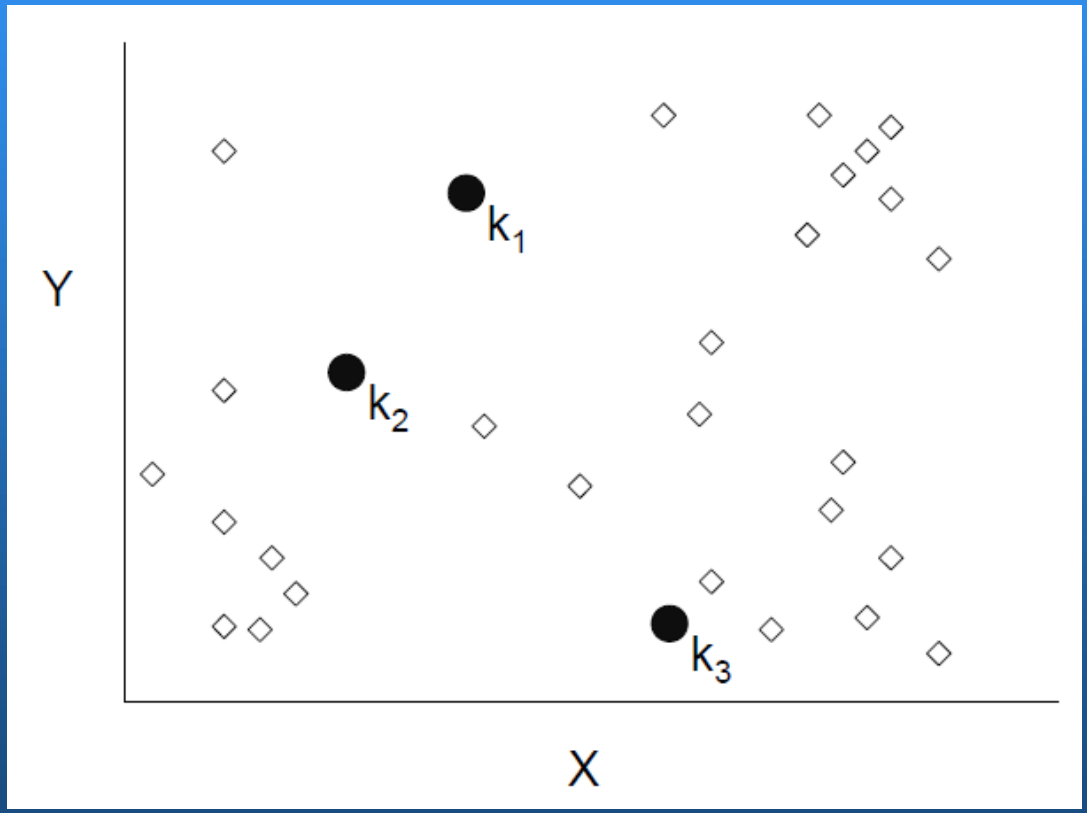
Algorytm k-średnich

Rozważmy następujący zbiór danych:



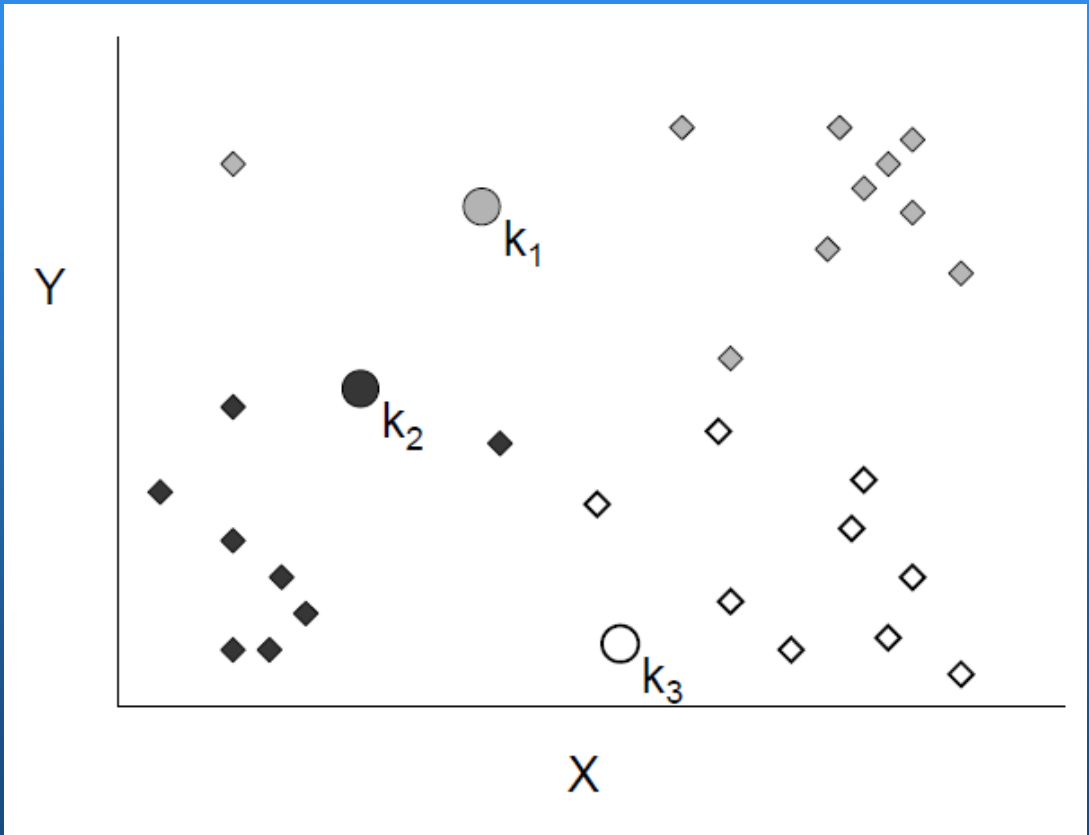
Algorytm k-średnich

Interesują nas k = 3 klastry. Wybieramy losowo początkowe środki klastrów.



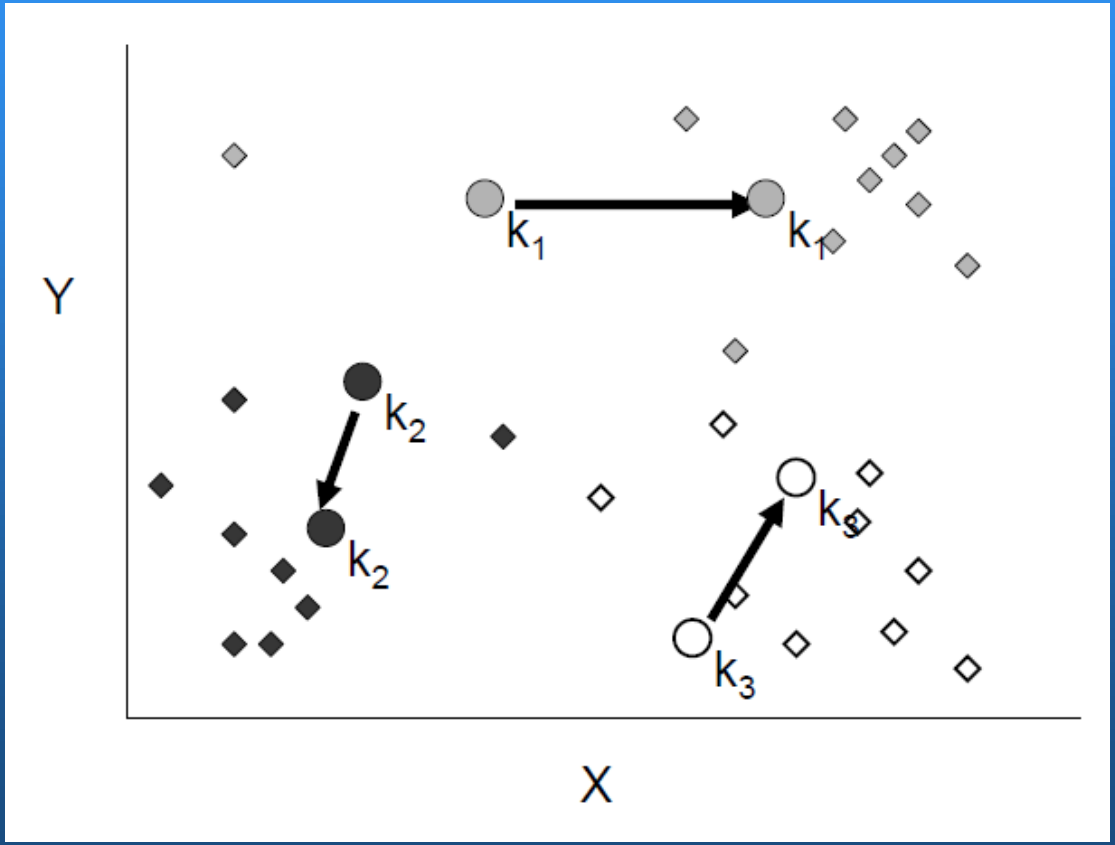
Algorytm k-średnich

Przydzielamy każdy obiekt do klastra w oparciu o odległość obiektu od środka klastra



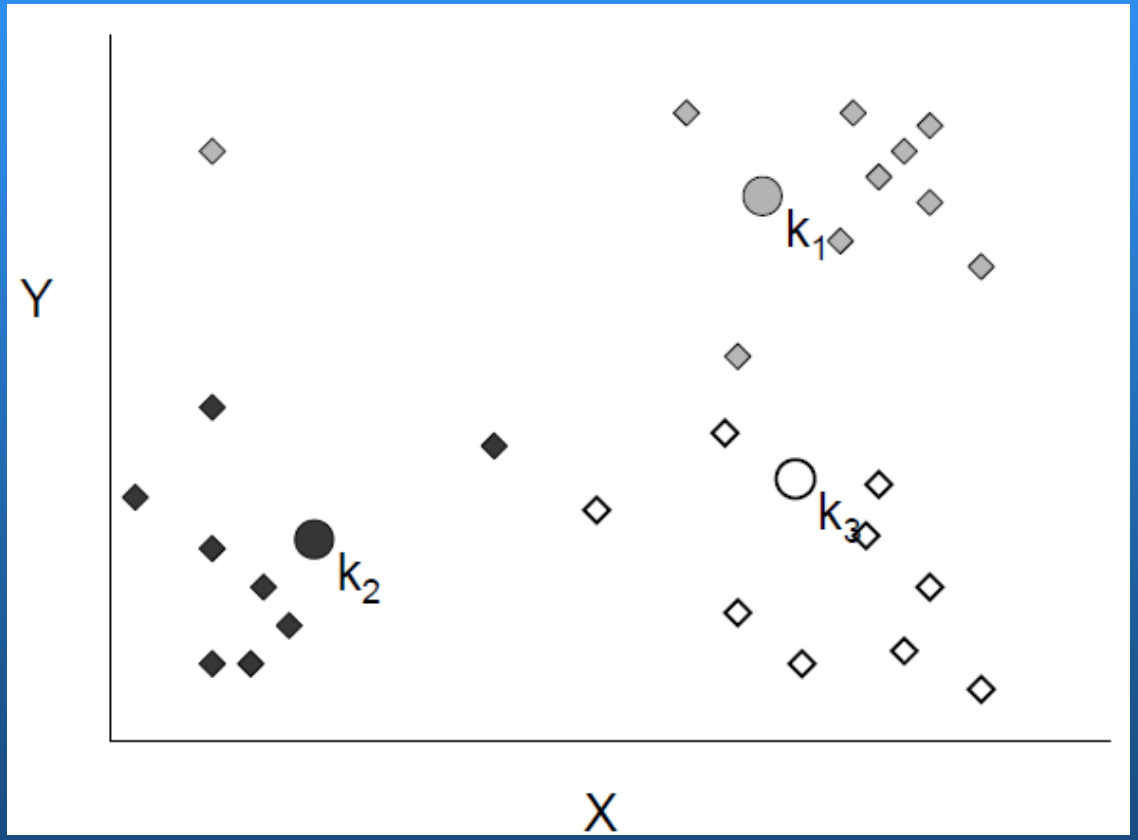
Algorytm k-średnich

Uaktualniamy środek (średnią) każdego klastra.



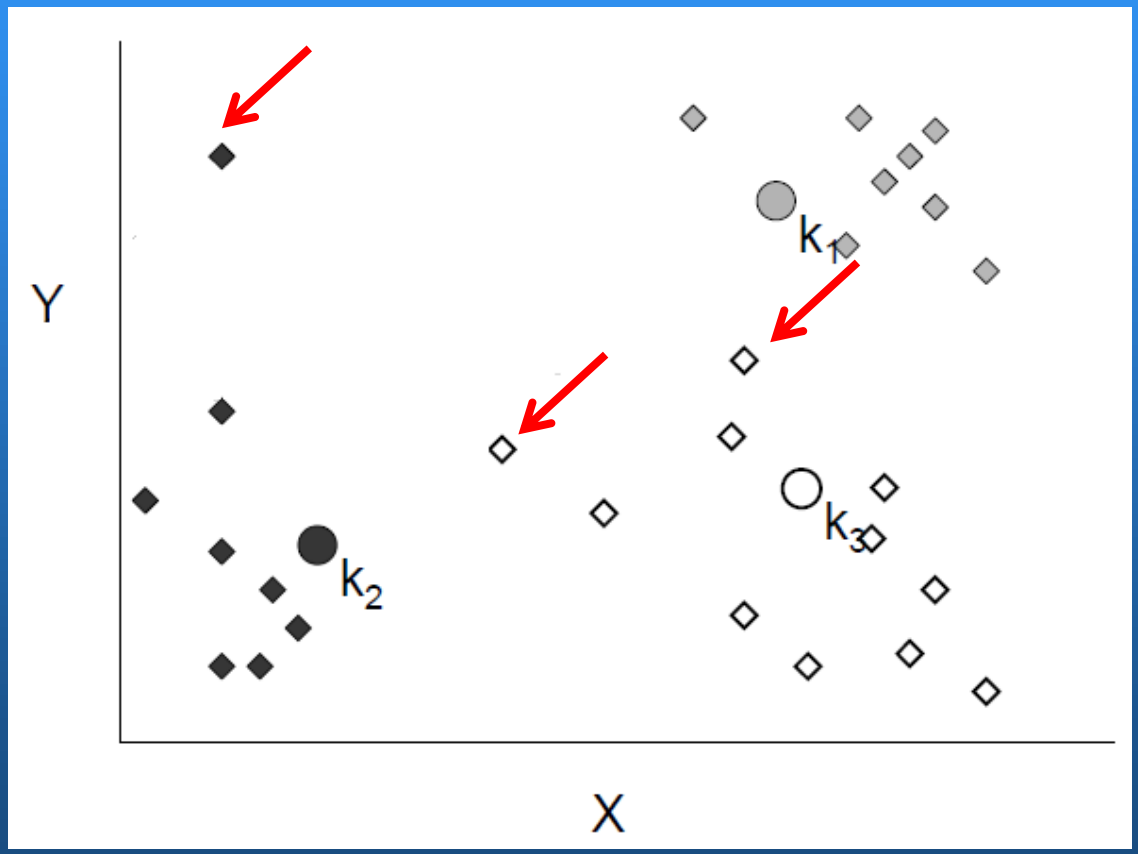
Algorytm k-średnich

Ponownie przypisujemy obiekty do klastrów.



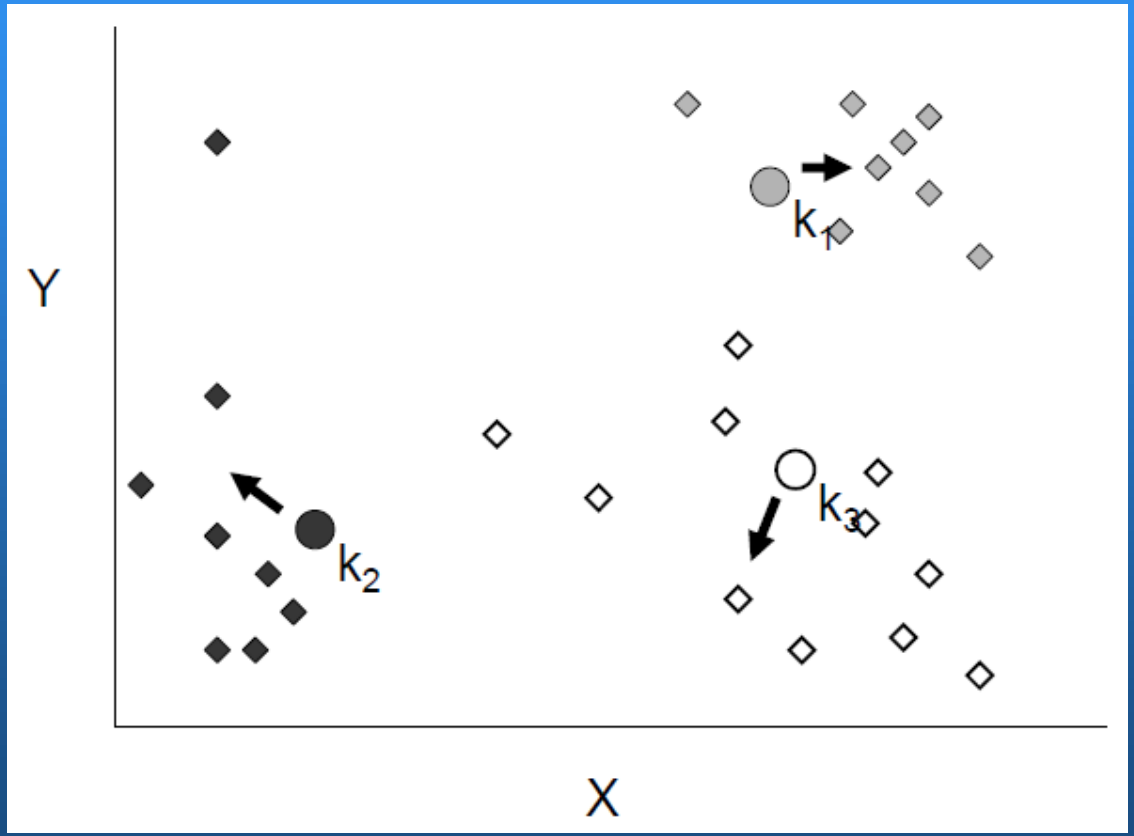
Algorytm k-średnich

Niektóre obiekty zmieniły przynależność do klastrów.



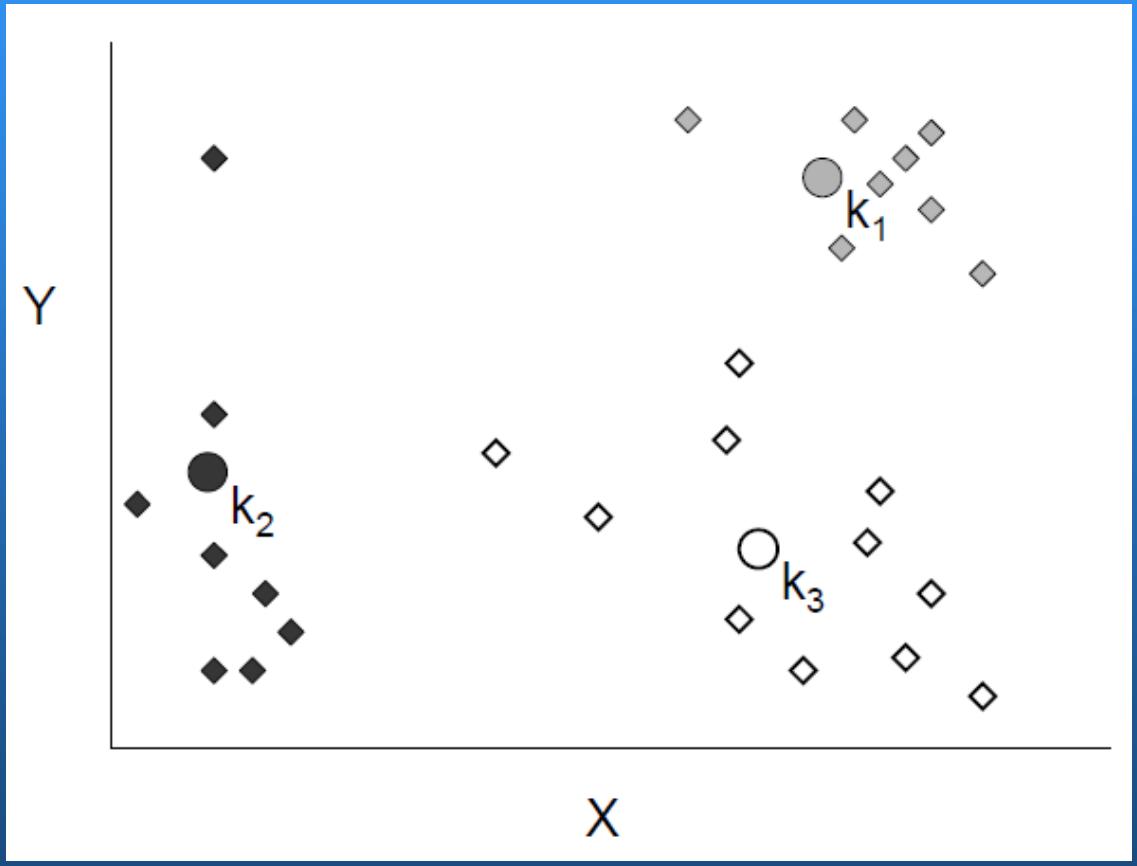
Algorytm k-średnich

Obliczmy nowe średnie klastrów.



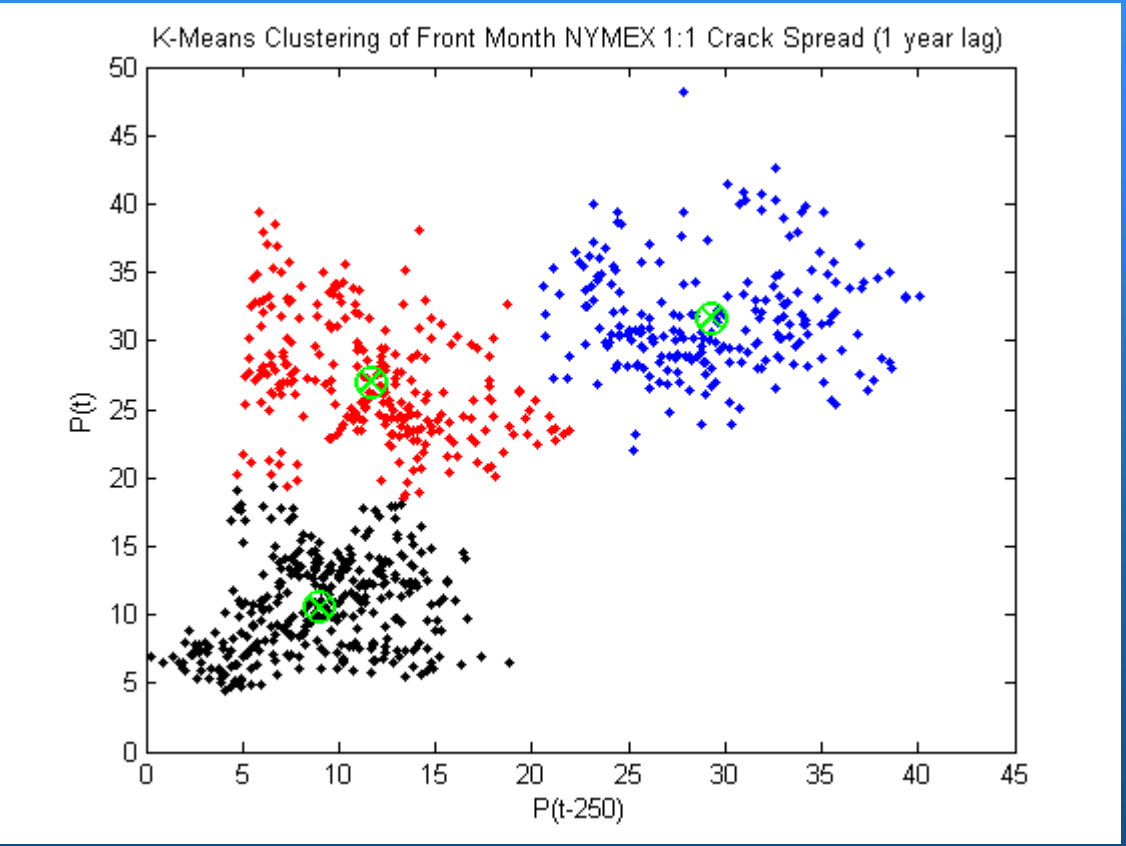
Algorytm k-średnich

Przesuwamy środki klastrów.



Algorytm k-średnich

Przykład



Algorytm k-średnich

Chcemy aby:

- Klastry były zwarte – jako miarę „zwartości” wykorzystujemy odchylenie wewnatrzklastrowe:

$$wc(C) = \sum_{j=1}^{k} wc(C_j) = \sum_{j=1}^{k} \sum_{x(i) \in C_j} d(x(i), r_j)^2$$

- Klastry były maksymalnie rozłączne – jako miarę „rozłączności” wykorzystujemy odchylenie międzyklastrowe:

$$bc(C) = \sum_{1 \le i < j \le k} d(r_j, r_j)^2$$