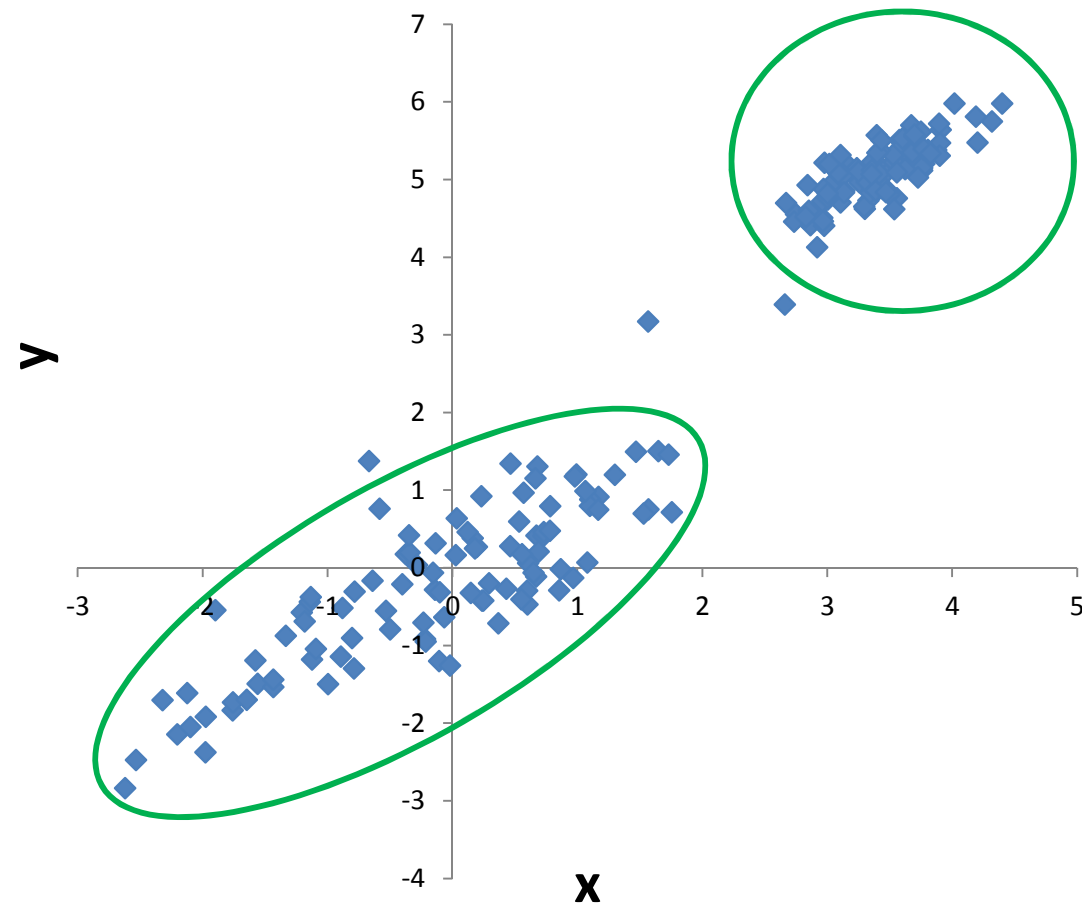




Grupowanie

Grupowanie



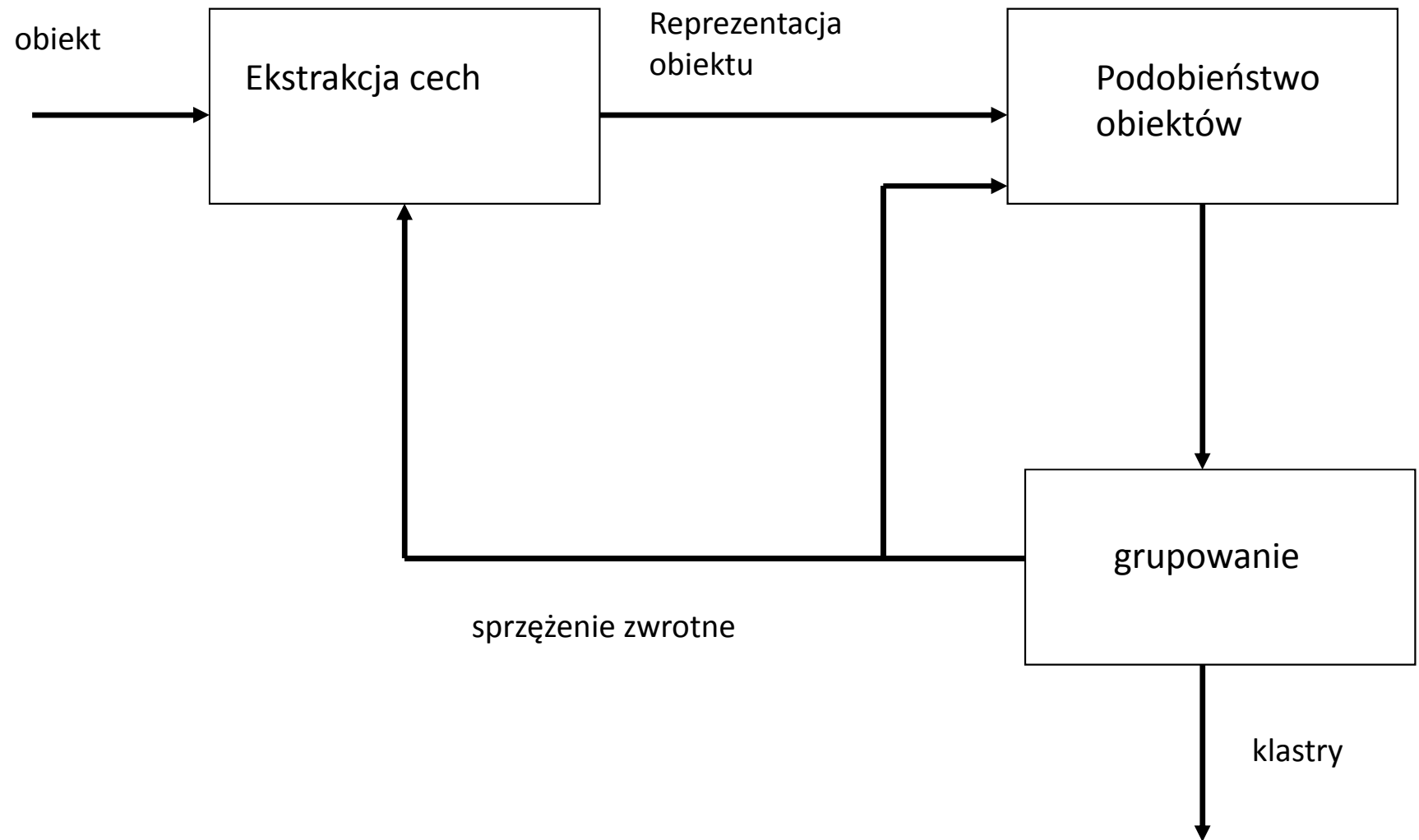
Wprowadzenie

- Celem procesu grupowania jest podział zbioru obiektów, fizycznych lub abstrakcyjnych, na klasy obiektów o podobnych cechach, nazywane klastrami lub skupieniami

Alternatywne definicje

- Zbiór obiektów, które są “podobne”
- Zbiór obiektów, takich, że odległość pomiędzy dwoma dowolnymi obiektami należącymi do klastra jest mniejsza aniżeli odległość pomiędzy dowolnym obiektem należącym do klastra i dowolnym obiektem nie należącym do tego klastra
- Spójny obszar przestrzeni wielowymiarowej, charakteryzujący się dużą gęstością występowania obiektów

Proces grupowania



Niepodobieństwo obiektów

- Niepodobieństwo (podobieństwo) obiektów opisujemy za pomocą macierzy niepodobieństwa (podobieństwa)
- Danych jest N obiektów, z których każdy jest opisany wartościami p atrybutów $A_1, \dots, A_p$ (nazywanymi zmiennymi)
- Macierz niepodobieństwa obiektów D , typu $N \times N$ opisuje niepodobieństwo pomiędzy każdą parą obiektów:

$$D = \begin{bmatrix} 0 & D(x_1, x_2) & \dots & D(x_1, x_N) \\ D(x_2, x_1) & 0 & \dots & D(x_2, x_N) \\ \vdots & \vdots & 0 & \vdots \\ D(x_N, x_1) & D(x_N, x_2) & \dots & 0 \end{bmatrix}$$

- Gdzie $D(x_i, x_j)$ oznacza niepodobieństwo obiektów

Odległość pomiędzy obiektami

- Konieczne jest znalezienie miary odległości
- Bardzo często używane miary odległości są metrykami:
 - $D(x, y) > 0$
 - $x=y \iff D(x, y) = 0$
 - $D(x, y) = D(y, x)$
 - $D(x, y) \leq D(x, z) + D(z, y)$ (nierówność trójkąta)

Metryka Minkowskiego

$$d(X, Y) = \left(\sum_{i=1}^n \underbrace{|x_i - y_i|}_{}^h \right)^{1/h}$$

Odległość ze
względem na i-ty
atrybut
 $d(x_i, y_i)$

- Specjalne przypadki
 - Manhattan: $h = 1$
 - Euklidesowa: $h = 2$
 - Chebysheva: $h = \infty$

$$\text{cheb}(X, Y) = \max |x_i - y_i|$$

Atrybuty kategoryczne

- Odległość Jaccarda
- (ew. znormalizowana) odległość Hamminga
- Macierz niepodobieństwa:

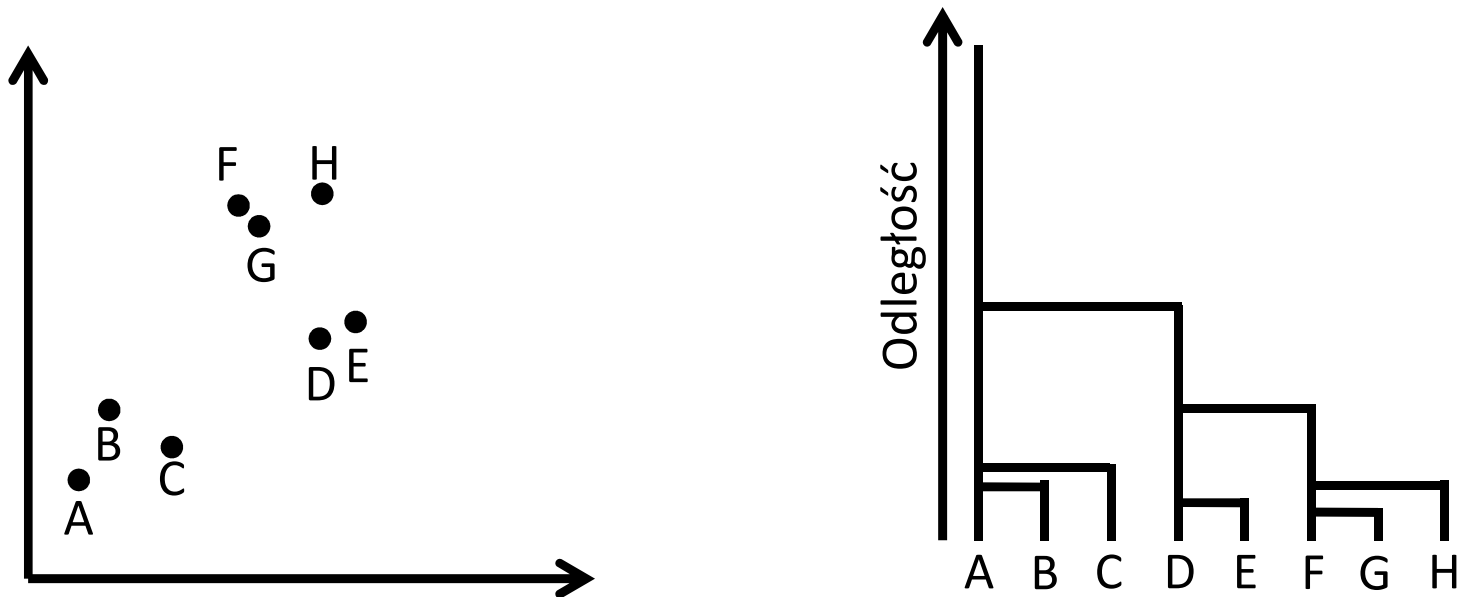
	Honda	Fiat	Volvo	Nissan
Honda	0	1	1	0
Fiat	1	0	1	1
Volvo	1	1	0	1
Nissan	0	1	1	0

Klasyfikacja metod grupowania

- metody hierarchiczne
- metody iteracyjno-optymalizacyjne
- metody gęstościowe
- metody gridowe

Metody hierarchiczne

Aglomeracyjne/Podziałowe



Metody hierarchiczne

- Metoda grupowania hierarchicznego polega na sekwencyjnym grupowaniu obiektów - drzewo klastrów (tzw. dendrogram).
 - 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 są na kolejne mniejsze klastry
 - Podejście aglomeracyjne (bottom-up): początkowo, każdy obiekt stanowi osobny klaster, następnie, w kolejnych iteracjach, klastry są łączone w większe klastry

Odległość pomiędzy zgrupowaniami

- Oznaczenia:
 - $D(x_i, x_j)$ – odległość pomiędzy obiektami x_i i x_j
 - m_i – średnia wartość zgrupowania C_i
 - n_i – liczba obiektów w zgrupowaniu C_i

Odległość pomiędzy zgrupowaniami

- Minimalna odległość:

$$d_{min} = \min_{x_i \in C_i, x_j \in C_j} D(x_i, x_j)$$

- Maksymalna odległość:

$$d_{max} = \max_{x_i \in C_i, x_j \in C_j} D(x_i, x_j)$$

- Średnia odległość

$$d_{avg} = \frac{1}{n_i n_j} \sum_{x_i \in C_i} \sum_{x_j \in C_j} D(x_i, x_j)$$

- Odległość średnich

$$d_{mean} = D(m_i, m_j)$$

Hierarchiczny algorytm grupowania

- Wejście: baza danych D n obiektów.
 - Wyjście: dendrogram reprezentujący sekwencję grupowania obiektów
1. umieść każdy obiekt w osobnym klastrze;
 2. skonstruuj macierz odległości międzyklastrowej dla wszystkich par klastrów;
 3. korzystając z macierzy odległości międzyklastrowych, znajdź najbliższą parę klastrów i połącz znalezione klastry, tworząc nowy klaster;
 4. uaktualnij macierz odległości międzyklastrowych po operacji połączenia;

Hierarchiczny algorytm grupowania

5. if wszystkie obiekty należą do jednego klastra then
6. zakończ procedurę grupowania;
7. else
8. przejdź do kroku 3;
9. end if
10. return dendrogram reprezentujący sekwencje grupowania obiektów;

Alg. iteracyjno-optymalizacyjne

- Dane: k – liczba zgrupowań
- Utwórz wstępny podział na k zgrupowań
- W kolejnych iteracjach przesuwaj obiekty pomiędzy zgrupowaniami tak długo jak poprawia to jakość grupowania.
- Optymalne rozwiązanie można osiągnąć jedynie po przeanalizowaniu wszystkich możliwych podziałów.
- W praktyce stosuje się jedną z wielu technik optymalizacji kombinatorycznej: iterative improvement, tabu search, simulating annealing, genetic algorithms, ant algorithms, itp

Miary oceny zgrupowania

- Dwa aspekty grupowania:
 - Klastry powinny być zwarte
 - Klastry powinny być maksymalnie rozłączne
- Odchylenie wewnątrzklastrowe - $wc(C)$
- Odchylenie międzyklastrowe – $bc(C)$
- Średnia klastra (mean) - m_k

$$m_k = \frac{1}{n_k} \sum_{x \in C_k} x$$

– Gdzie:

- n_k - liczba obiektów należących do k-tego klastra C_k

Miary oceny zgrupowania

- Miary $wc(C)$:

$$wc(C) = \sum_{i=1}^k wc(C_i) = \sum_{i=1}^k \sum_{x_j \in C_i} d(x_j, m_i)^2$$

$$wc(C) = \max_i \min_{y_j \in C_k} \{d(x_i, y_j) \mid x_i \in C_k, x_i \neq y_j\}$$

- Miara $bc(C)$:

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

Algorytm k-średnich

- Algorytm zachłanny
- Optymalizuje miarę:

$$wc(C) = \sum_{i=1}^k wc(C_i) = \sum_{i=1}^k \sum_{x_j \in C_i} d(x_j, m_i)^2$$

– Gdzie

- $m_k = \frac{1}{n_k} \sum_{x \in C_k} x$

- Zgrupowanie jest reprezentowane przez średnią
- Obiekt należy do tego zgrupowania, którego środek jest najbliżej

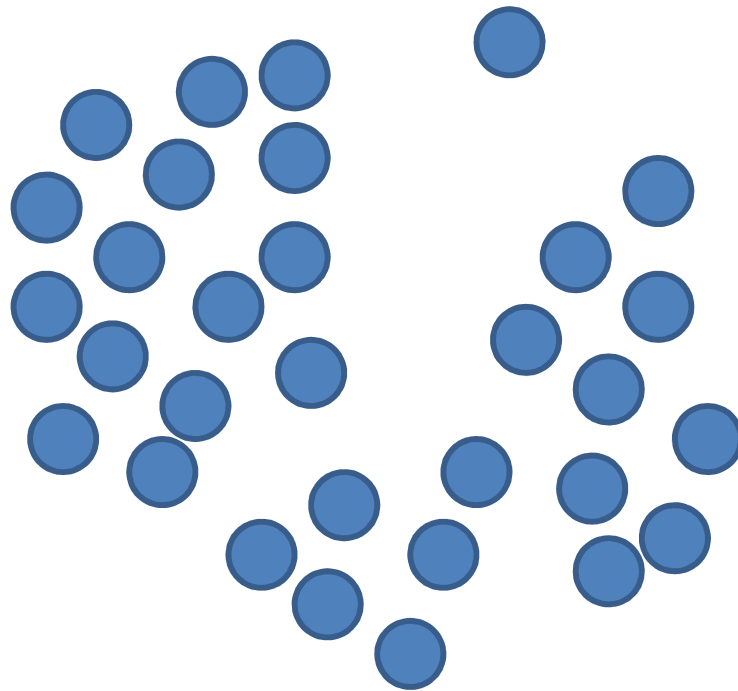
Algorytm k-średnich

Wejście: liczba zgrupowań k , baza danych n obiektów

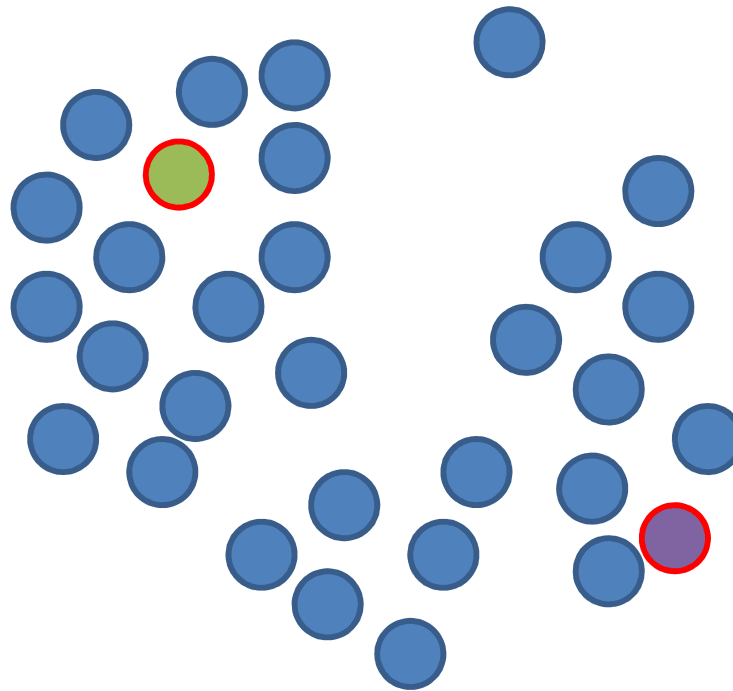
Wyjście: zbiór k zgrupowań minimalizujący kryterium błędu średniokwadratowego

1. wybierz losowo k obiektów, jako środki skupień
2. while przydział obiektów się zmienił do
3. każdy obiekt przydziel do zgrupowania, którego środek jest najbliżej
4. oblicz nowe środki zgrupowań - m_k
5. end while

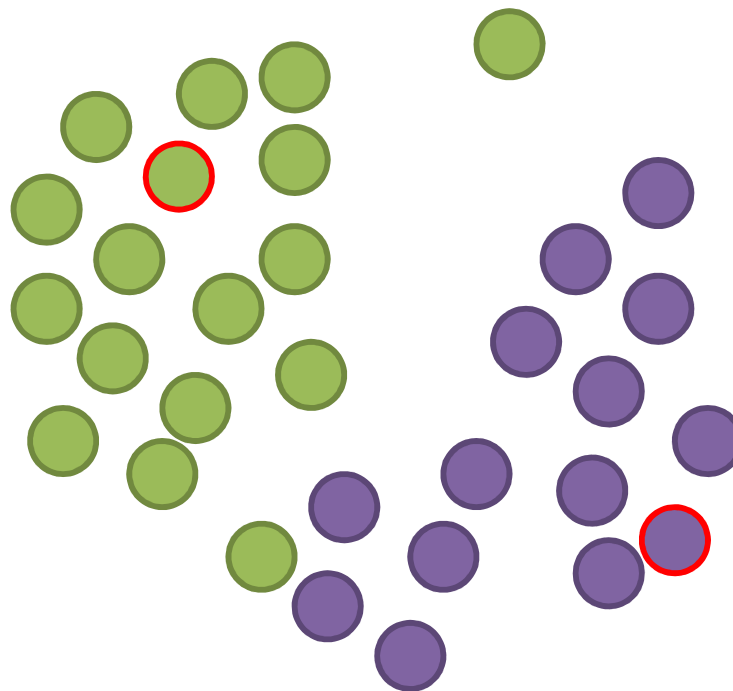
Grupowanie - KMeans



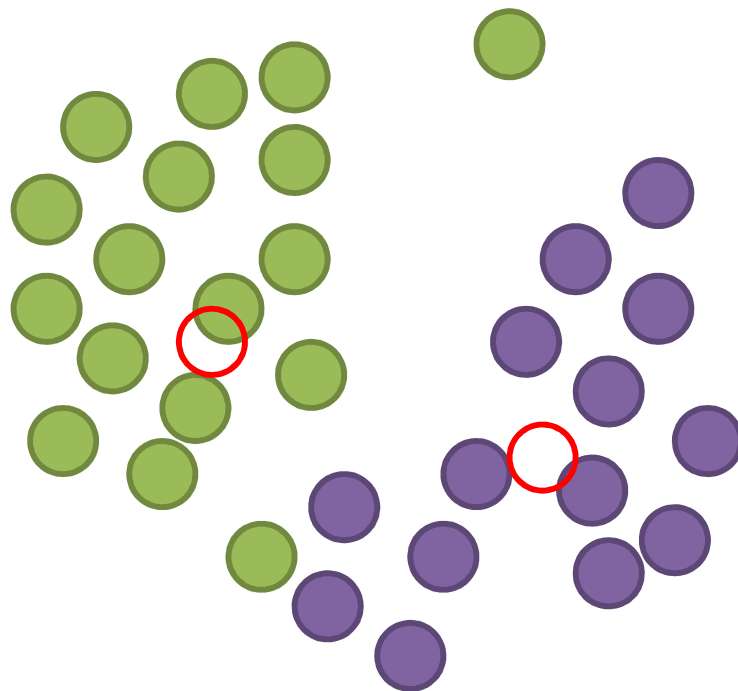
Grupowanie - KMeans



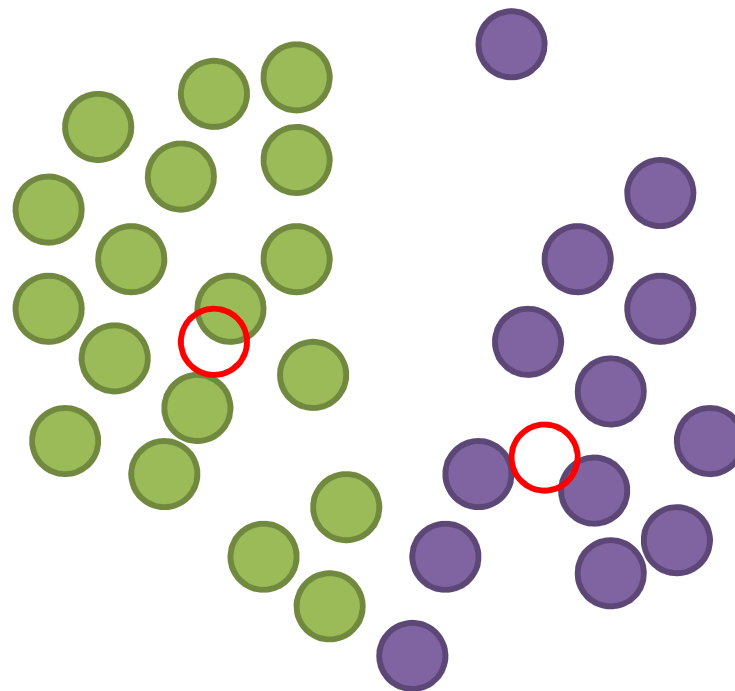
Grupowanie - KMeans



Grupowanie - KMeans



Grupowanie - KMeans



Problem początkowego podziału

- Wynik działania algorytmu (tj. ostateczny podział obiektów pomiędzy klastrami) silnie zależy od początkowego podziału obiektów
- Algorytm łatwo wpada w minimum lokalne
- W celu zwiększenia szansy znalezienia optimum globalnego należy kilkakrotnie uruchomić algorytm dla różnych podziałów początkowych

Problem punktów osobliwych

- Wadą algorytmu k-średnich jest jego czułość na występowanie punktów osobliwych
- Punktami osobliwymi nazywamy obiekty, które znacząco różnią się od pozostałych grupowanych obiektów
- Punkty osobliwe reprezentują najczęściej dwa przypadki:
 - błędy w danych
 - obiekty o bardzo specyficznych wartościach atrybutów
- Najprostsze rozwiązanie polega na usunięciu obiektów, które znacząco odbiegają od środków klastrów

Przykład 1

- Dane są punkty:

$A = (2, 1)$	$B = (4, 2)$	$C = (4, 3)$	$D = (10, 5)$
$E = (11, 5)$	$F = (9, 9)$	$G = (12, 9)$	

- Przyjmujemy $k=3$.
- Jako początkowe środki przyjmujemy współrzędne punktów A, B i C

Przykład 1

Odległości

	A	B	C	D	E	F	G
(2,1)	0	2,24	2,83	8,94	9,85	10,63	12,81
(4,2)	2,24	0	1	6,71	7,62	8,6	10,63
(4,3)	2,838	1	0	6,32	7,28	7,81	10

Nowe środki

$$\frac{1}{1} * (2,1) = (2,1)$$

$$\frac{1}{1} * (4,2) = (4,2)$$

$$\frac{1}{5} * ((4,3) + (10,5) + (11,5) + (9,9) + (12,9)) = \frac{1}{5} * (46,31) = (9.2,6.2)$$

Przykład 1

Odległości

	A	B	C	D	E	F	G
(2,1)	0	2,24	2,83	8,94	9,85	10,63	12,81
(4,2)	2,24	0	1	6,71	7,62	8,6	10,63
(9.2,6.2)	8,88	6,68	6,11	1,44	2,16	2,81	3,96

Nowe środki

$$\frac{1}{2} * ((2,1) + (4,3)) = (3,2)$$

$$\frac{1}{1} * (4,2) = (4,2)$$

$$\frac{1}{4} * ((10,5) + (11,5) + (9,9) + (12,9)) = \frac{1}{4} * (42,28) = (10.5,7)$$

Przykład 1

Odległości

	A	B	C	D	E	F	G
(3,2)	1,41	1	1,41	7,62	8,54	9,22	11,4
(4,2)	2,24	0	1	6,71	7,62	8,6	10,63
(10.5,7)	10,4	8,2	7,63	2,06	2,06	2,5	2,5

Nowe środki

$$\frac{1}{1} * (2,1) = (2,1)$$

$$\frac{1}{2} * ((4,2) + (4,3)) = (4,2.5)$$

$$\frac{1}{4} * ((10,5) + (11,5) + (9,9) + (12,9)) = \frac{1}{4} * (42,28) = (10.5,7)$$

Przykład 1

Odległości

	A	B	C	D	E	F	G
(2,1)	0	2,24	2,83	8,94	9,85	10,63	12,81
(4,2.5)	2,5	0,5	0,5	6,5	7,43	8,2	10,31
(10.5,7)	10,4	8,2	7,63	2,06	2,06	2,5	2,5

Koniec, ponieważ przypisanie obiektów do zgrupowań się nie zmieniło

Przykład 2

- Wykonamy teraz przebieg dla innego początkowego podziału.
- Jako początkowe środki przyjmujemy współrzędne punktów A, D i F

Przykład 2

Odległości

	A	B	C	D	E	F	G
(2,1)	0	2,24	2,83	8,94	9,85	10,63	12,81
(10,5)	8,94	6,71	6,32	0	1	4,12	4,47
(9,9)	10,63	8,6	7,81	4,12	4,47	0	3

Nowe środki

$$\frac{1}{3} * ((2,1) + (4,2) + (4,3)) = (\frac{10}{3}, 3)$$

$$\frac{1}{2} * ((10,5) + (11,5)) = (10.5, 5)$$

$$\frac{1}{2} * ((9,9) + (12,9)) = \frac{1}{2} * (21, 18) = (10.5, 9)$$

Przykład 2

Odległości

	A	B	C	D	E	F	G
(3.33,3)	2,4	1,2	0,67	6,96	7,92	8,25	10,54
(10.5,5)	9,39	7,16	6,8	0,5	0,5	4,27	4,27
(10.5,9)	11,67	9,55	8,85	4,03	4,03	1,5	1,5

Koniec, ponieważ przypisanie obiektów do zgrupowań się nie zmieniło

Przykład

- Porównajmy jakość jednego i drugiego grupowania:

– Przykład 1

- Zgrupowanie 1: (2,1) - A
- Zgrupowanie 2: (4,2.5) – B, C
- Zgrupowanie 3: (10.5,7) – D, E, F, G
- $wc(C) = 2 * \frac{1^2}{2} + 2 * 2.06^2 + 2 * 2.5^2 \approx 21.49$

– Przykład 2

- Zgrupowanie 1: (3.333,2) – A, B, C
- Zgrupowanie 2: (10.5,5) – D, E
- Zgrupowanie 3: (10.5,9) – F, G
- $wc(C) = 2.4^2 + 1.2^2 + 0.67^2 + 2 * \frac{1^2}{2} + 2 * 4.27^2 \approx 12.65$

Grupowanie gęstościowe

- Poprzednie metody grupowania tworzą „sferyczne zgrupowania”
- Proces grupowania w metodach grupowania gęstościowego jest oparty na pojęciu *gęstości* (ang. *density*)
- Zgrupowaniem obiektów jest obszar w przestrzeni obiektów charakteryzujący się dużą gęstością obiektów.
- Zgrupowania obiektów są odseparowane od siebie obszarami o małej gęstości obiektów

Podstawowa idea

- Wybierz 1 obiekt jako „załączek” zgrupowania
- Dokładaj do niego kolejne obiekty tak długo jak dokoła jest wystarczające zagęszczenie obiektów
- Wybierz kolejny obiekt jeszcze nie przydzielony jako kolejny „załączek” i... patrz poprzedni punkt 😊
- Postępuj w ten sposób tak długo aż wszystkie obiekty zostaną przydzielone

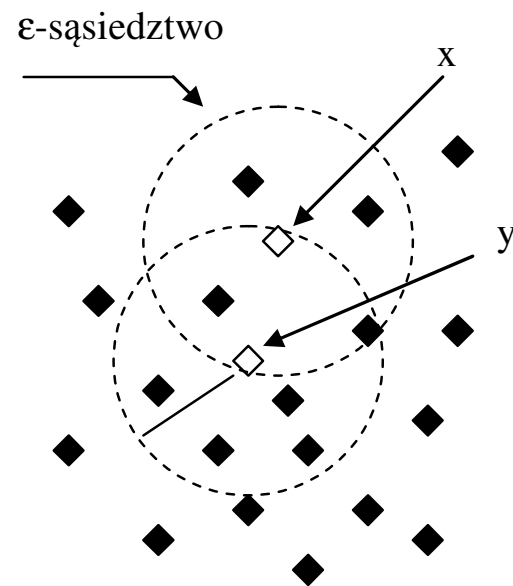
Podstawowe terminy

- Oznaczenia:
 - D – zbiór grupowanych obiektów
 - $D(x_i, x_j)$ – odległość pomiędzy obiektami x_i i x_j
- $N_\varepsilon(x)$ – ε – sąsiedztwo obiektu x – zbiór obiektów nie dalej niż ε od obiektu x :
$$N_\varepsilon(x) = \{y \in D \mid D(x, y) \leq \varepsilon\}$$
- Obiekt *centralny* – obiekt, którego ε – sąsiedztwo jest większe niż minPts

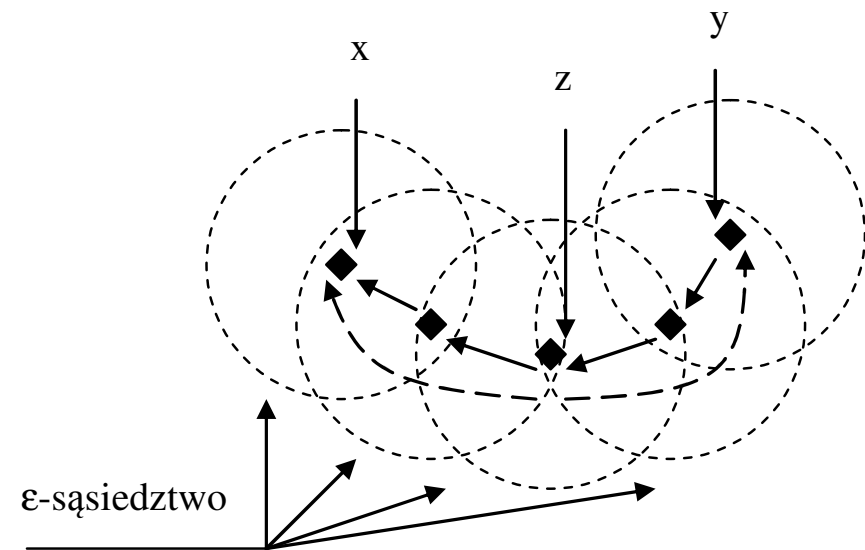
Podstawowe terminy

- Mówimy, że obiekt x jest bezpośrednio gęstościowo osiągalny z obiektu y , jeżeli:
 - $x \in N_\varepsilon(y)$
 - Y jest obiektem centralnym
- Mówimy, że obiekt x jest gęstościowo osiągalny z obiektu y , jeżeli istnieje łańcuch obiektów $p_1, p_2, \dots, p_n$, że:
 - $p_1=y, p_n=x$
 - p_{i+1} jest bezpośrednio gęstościowo osiągalny z p_i ($1 \leq i < n$)

Przykład



(a)



(b)

Podstawowe terminy

- Obiekt x jest *gęstościowo połączony* z obiektem y , względem parametrów ε i MinPts, jeżeli istnieje obiekt $z \in D$, taki że oba obiekty x i y są gęstościowo osiągalne z obiektu z względem parametrów ε i MinPts

Zgrupowanie

- Zbiór $C \subseteq D$ taki, że:
 - dla dowolnych obiektów $x, y \in D$: jeżeli $x \in C$ i y jest gęstościowo osiągalny z obiektu x , względem parametrów ε i MinPts, wówczas $y \in C$
 - dla dowolnych obiektów $x, y \in C$: obiekt x jest gęstościowo połączony z obiektem y , względem parametrów ε i MinPts

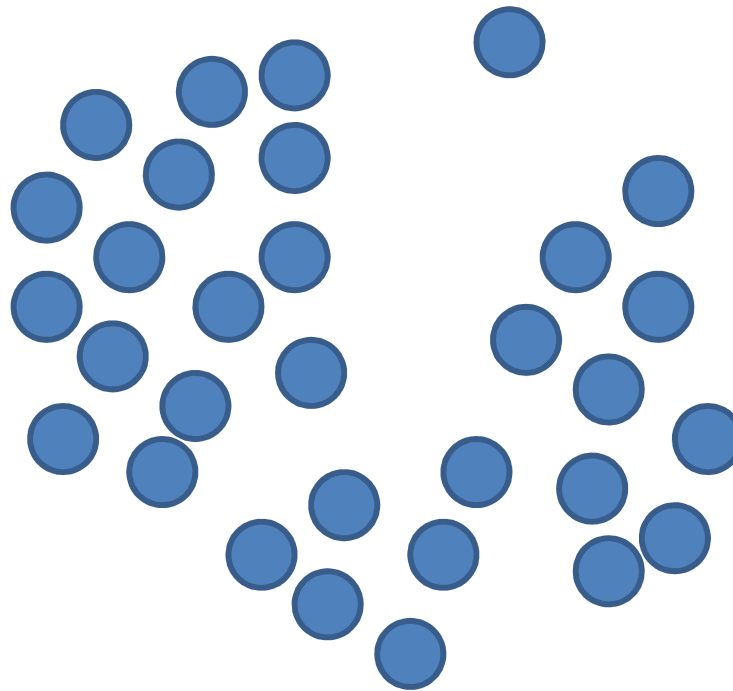
Oznaczenia

- Niech $C_1, \dots, C_k$ oznacza zbiór klastrów zbioru obiektów D względem parametrów ε i MinPts
- Zbiorem punktów osobliwych jest taki podzbiór obiektów zbioru D , który nie należy do żadnego klastra C_i , $i = 1, \dots, k$
- Obiekty nie należące do żadnego zgrupowania to obiekty osobliwe

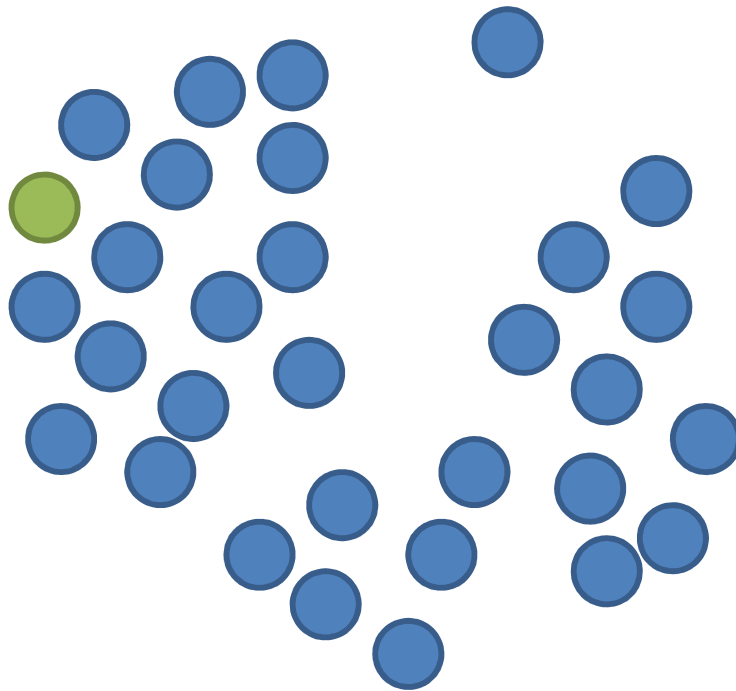
Algorytm DBSCAN

- Algorytm jest realizowany w trzech krokach:
 1. rozpoczynamy od dowolnego obiektu $x \in D$,
 2. jeżeli ε -sąsiedztwo obiektu p spełnia warunek minimalnej gęstości, to jest $|N_\varepsilon(x)| \geq \text{MinPts}$, wówczas tworzony jest klaster C i wszystkie obiekty gęstościowo osiągalne z obiektu x są dołączane do klastra C ; w przeciwnym razie (obiekt x nie jest obiektem centralnym), wracamy do kroku (1) i wybieramy następny obiekt zbioru D ,
 3. proces grupowania jest kontynuowany tak długo, aż zostaną przetworzone wszystkie obiekty zbioru D
- Obiekty, które nie zostały zaklasyfikowane do żadnego z klastrów, tworzą zbiór punktów osobliwych

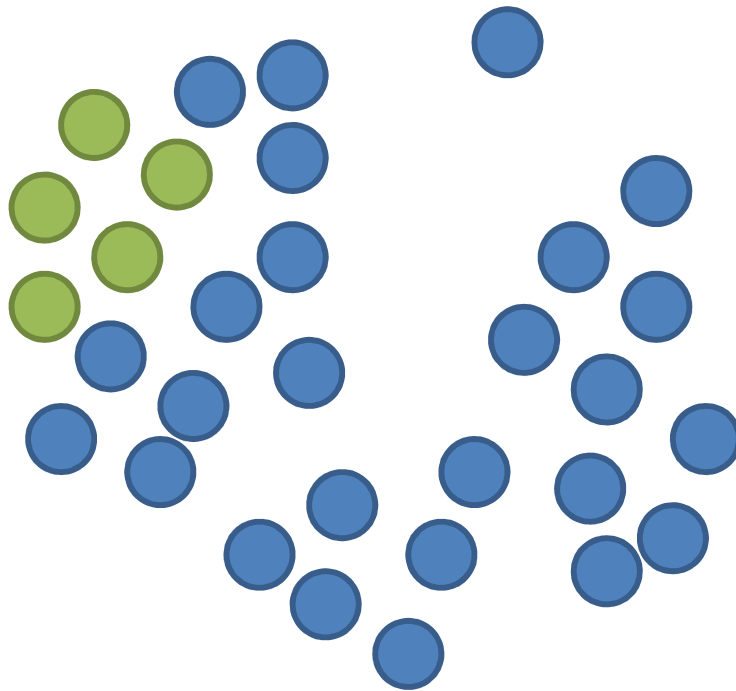
Grupowanie - DBSCAN



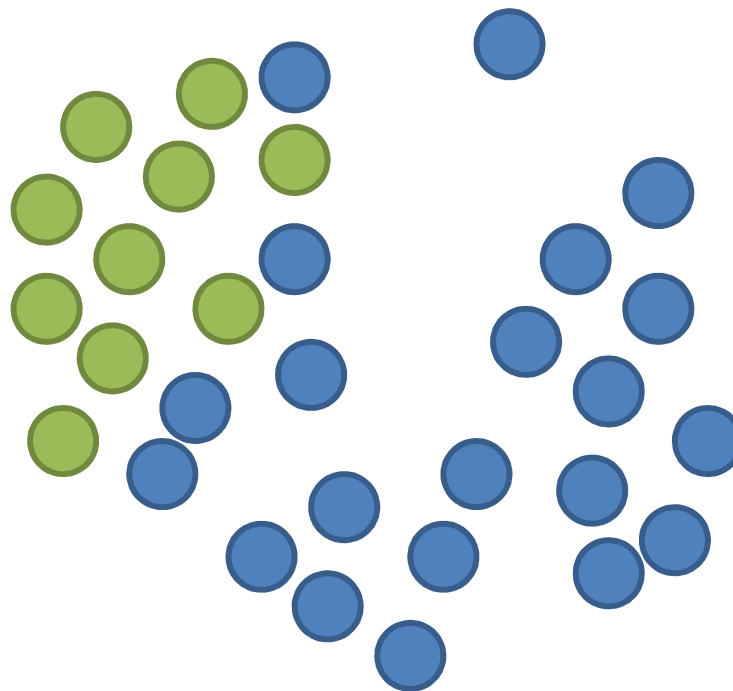
Grupowanie - DBSCAN



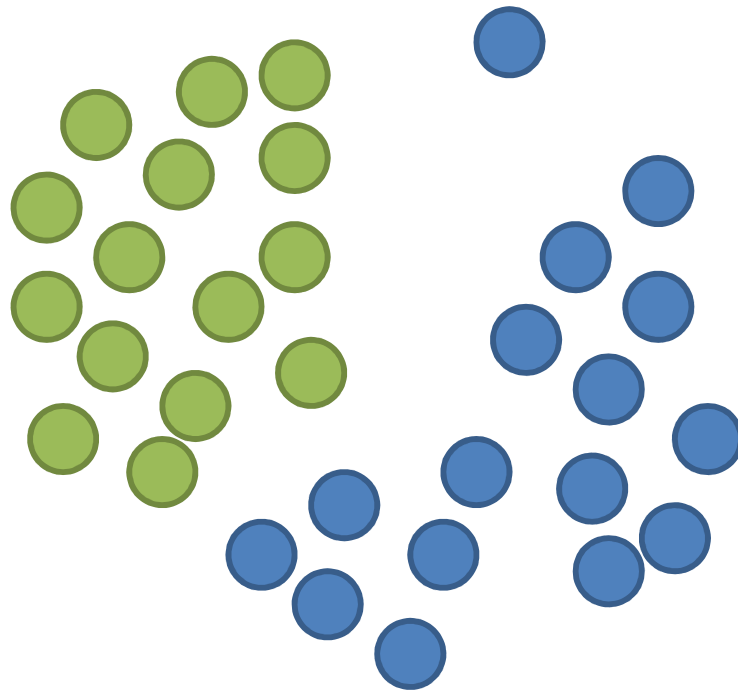
Grupowanie - DBSCAN



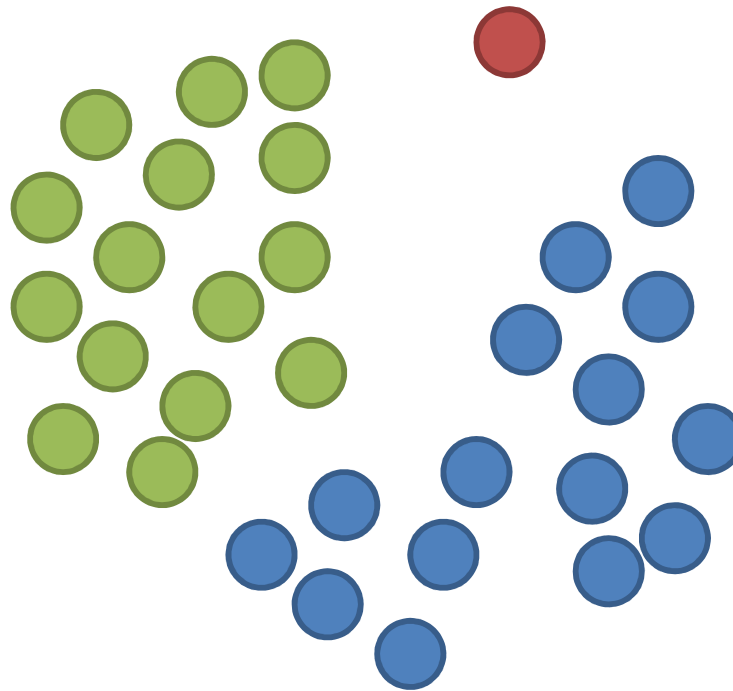
Grupowanie - DBSCAN



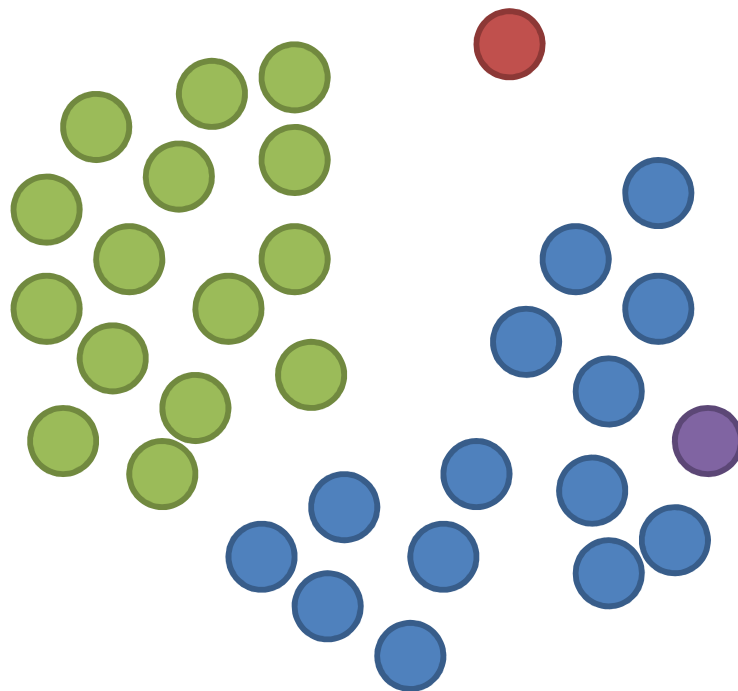
Grupowanie - DBSCAN



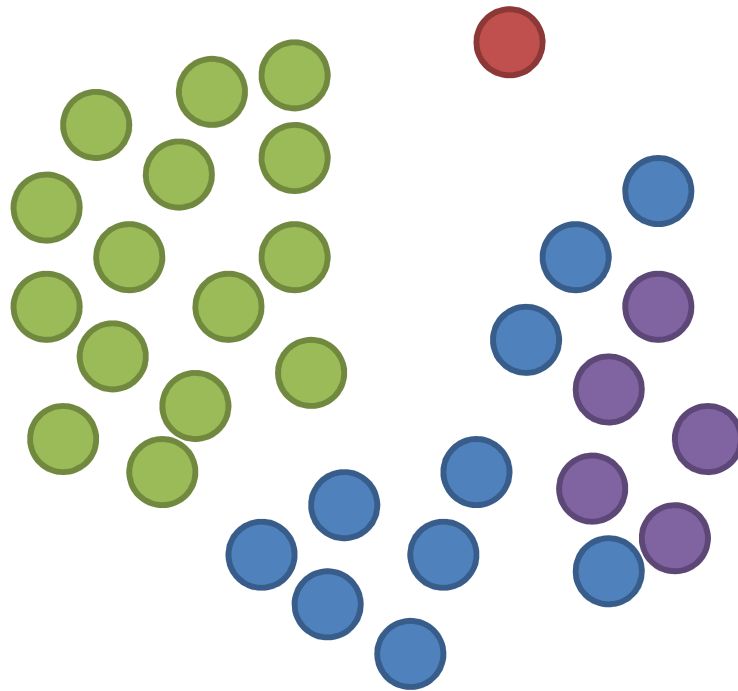
Grupowanie - DBSCAN



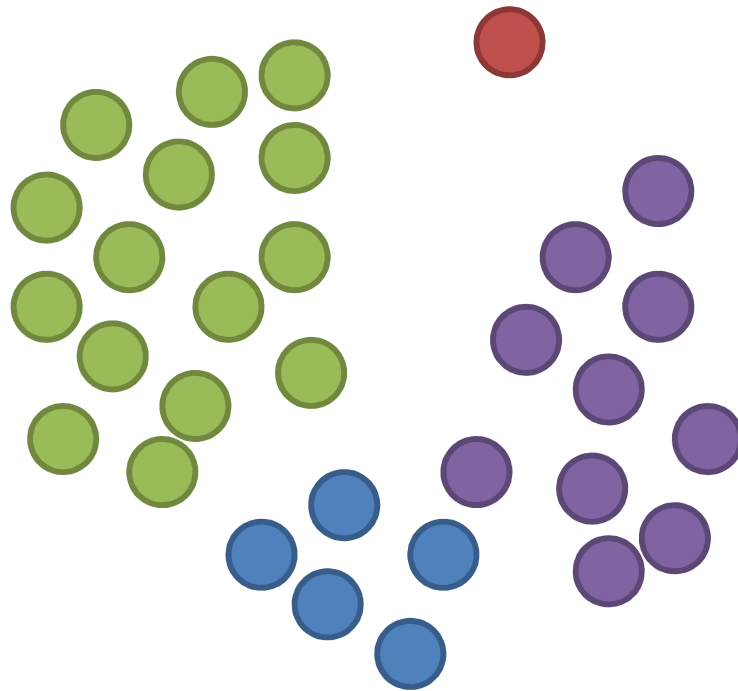
Grupowanie - DBSCAN



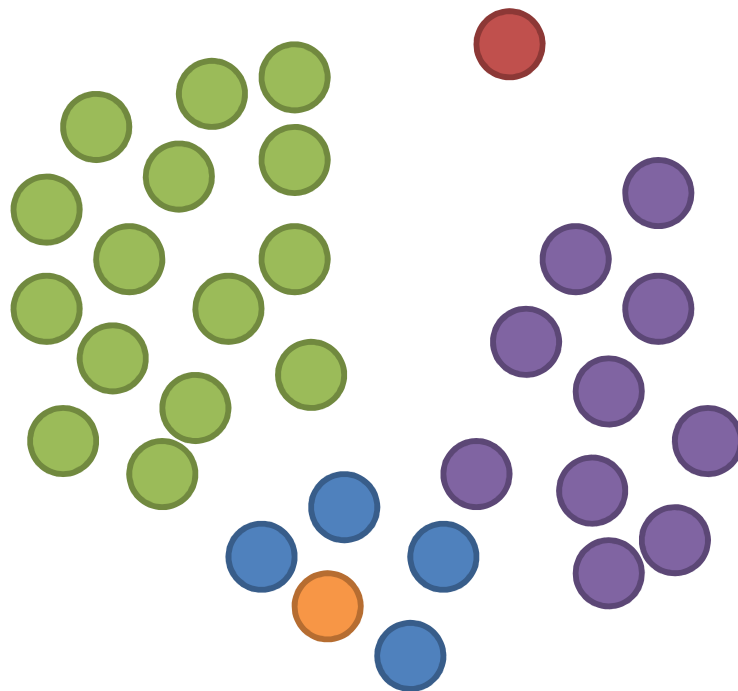
Grupowanie - DBSCAN



Grupowanie - DBSCAN



Grupowanie - DBSCAN



Grupowanie - DBSCAN

