ZAAWANSOWANE METODY EKSPLORACJI

GRUPOWANIE DANYCH

Prof. dr hab. inż. Grzegorz Dudek
Wydział Matematyki i Informatyki
Uniwersytet Łódzki

INFORMACJE WSTĘPNE

- Grupowanie danych zwane inaczej grupowaniem pojęciowym, klasteryzacją lub analizą skupień (*cluster analysis*) to przykład uczenia bez nadzoru.

- Celem grupowania danych jest wykrycie w zbiorze przykładów skupień i podział przykładów na grupy.
- Zbiór uczący składa się z przykładów nieetykietowanych.
- Uczeń dzieli przykłady na grupy/kategorie i konstruuje opis każdej grupy. Opis stanowi hipotezę pozwalającą przydzielać do grup nowe przykłady.
- Przynależność do grupy opiera się na podobieństwie przykładów (dla atrybutów ciągłych i porządkowych podobieństwo opiera się na mierze odległości pomiędzy przykładami).
- W obrębie każdej grupy przykłady powinny być do siebie maksymalnie podobne. Przykłady należące do różnych grup powinny być maksymalnie niepodobne.

ZASTOSOWANIA

Grupowanie używane jest zarówno jako wstępny krok analizy danych, jak i pełnoprawne narzędzie badawcze. Do najczęstszych zastosowań należy:

- eksploracja danych *(data mining),* gdzie grupowanie używane jest np. do podziału klientów na pewne podgrupy
- segmentacja obrazu *(image segmentation),* czyli podział obrazu na regiony homogeniczne pod względem pewnej własności obrazu (kolor, tekstura, intensywność). Taki uproszczony obraz jest prostszy do obróbki np. przez algorytmy rozpoznawania obrazu
- rozpoznawanie obrazów (*pattern recognition*)
- ekstrakcja informacji *(information retrieval),* mająca za zadanie uporządkowanie i uproszczenie dostępu do informacji. Do klasycznych zastosowań należy stworzenie klasyfikacji książek, czy stron internetowych
- grupowanie zadań w problemie harmonogramowania tak, by zadania intensywnie ze sobą komunikujące się trafiły do tej samej grupy. Taka grupa zostanie w następnym kroku przypisana do wykonania na jednym procesorze

TYPY ALGORYTMÓW GRUPOWANIA

Algorytmy grupowania danych dzieli się na kilka kategorii:

- Algorytmy oparte na podziałach (*partitioning algorithms*) – konstruują różne podziały i oceniają je za pomocą funkcji kryterialnej. Zadanie sprowadza się do optymalizacji tej funkcji
- Algorytmy hierarchiczne – tworzą hierarchiczną reprezentację zbioru danych
- Algorytmy oparte na gęstościach (*density-based algorithms*) – oparte na funkcji gęstości i na lokalnych połączeniach; produkują grupy o dowolnych kształtach
- Inne, np. sekwencyjne, jądrowe, grafowe

KROKI ALGORYTMU GRUPOWANIA

W algorytmie grupowania wyodrębnia się następujące kroki:

1. Wybór reprezentacji przykładów (włączając selekcję i ekstrakcję atrybutów (*feature selection, feature extraction*)
2. Wybór modelu – reprezentacji klastrów i pewnych warunków, jakie muszą być spełnione w wyniku (ograniczeń)
3. Zdefiniowanie funkcji będącej miarą podobieństwa pomiędzy klastrem a przykładem
4. Zdefiniowanie funkcji oceny grupowania (i funkcji oceny klastra, z której korzysta funkcja oceny grupowania)
5. Grupowanie, czyli przeglądanie przestrzeni w jakiej opisane są klastry w celu znalezienia rozwiązania optymalnego z punktu widzenia funkcji zdefiniowanej w 4
6. Abstrakcja rezultatów
7. Ocena rezultatów

MIARY PODOBIEŃSTWA/NIEPODOBIEŃSTWA

- Artybuty ciągłe – miary podobieństwa/niepodobieństwa opierają się na miarach odległości lub na mierze korelacji:

$$\rho(\mathbf{x}_a, \mathbf{x}_b) = \frac{\sum_{j=1}^{n}(x_{a,j} - \bar{x}_j)(x_{b,j} - \bar{x}_j)}{\sqrt{\sum_{j=1}^{n}(x_{a,j} - \bar{x}_j)^2 \sum_{j=1}^{n}(x_{b,j} - \bar{x}_j)^2}}$$

- Artybuty porządkowe – wymagają konwersji na wartości numeryczne i zastosowania miar podobieństwa/niepodobieństwa takich jak dla atrybutów ciągłych.

- Atrybuty nominalne – miara podobieństwa/niepodobieństwa opiera się na odległości Hamminga. Odległość Hamminga pomiędzy przykładami $\mathbf{x}_a$ i $\mathbf{x}_b$ z nominalnymi atrybutami równa jest liczbie atrybutów o różnych wartościach w tych przykładach.

Do określania odległości pomiędzy punktami stosuje się:

- odległość euklidesową:

$$d(\mathbf{x}_a, \mathbf{x}_b) = [(\mathbf{x}_a - \mathbf{x}_b)^{\mathsf{T}} (\mathbf{x}_a - \mathbf{x}_b)]^{1/2} = \sqrt{\sum_{t=1}^{T} (x_{a,t} - x_{b,t})^2}$$

- odległość miejską (*city-block*, *Manhattan*):

$$d(\mathbf{x}_a, \mathbf{x}_b) = \sum_{t=1}^{T} | x_{a,t} - x_{b,t} |$$

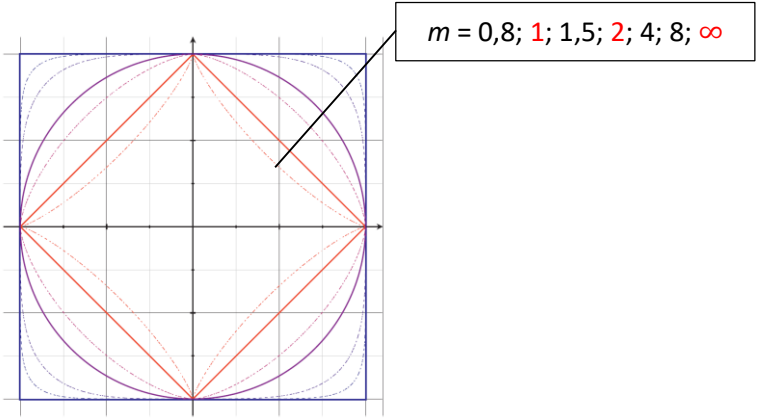
- odległość Czebyszewa:

$$d(\mathbf{x}_a, \mathbf{x}_b) = \max_{1 \le t \le T} | x_{a,t} - x_{b,t} |$$

Każda z tych funkcji stanowi szczególny przypadek odległości Minkowskiego:

$$d(\mathbf{x}_a, \mathbf{x}_b) = \left(\sum_{t=1}^{T} | x_{a,t} - x_{b,t} |^m \right)^{1/m} .$$

Miary oparte na odległości Minkowskiego nie są niezmiennicze względem skali wartości zmiennych. Zmiana skali powoduje zmianę odległości pomiędzy obserwacjami. Aby temu zapobiec, zaleca się wcześniejszą normalizację/standaryzację obserwacji.



FUNKCJA CELU W GRUPOWANIU

Sumę kwadratów odległości pomiędzy parami punktów (przykładów) *T* możemy rozłożyć na sumę kwadratów odległości między parami punktów należących do tej samej grupy *W* oraz sumę kwadratów odległości między parami punktów należących do różnych grup *B*: $T = W + B$, gdzie*:

$$T = \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} d^2(\mathbf{x}_i, \mathbf{x}_{i'})$$

$$W = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(\mathbf{x}_i)=k}\sum_{C(\mathbf{x}_i)=k} d^2(\mathbf{x}_i, \mathbf{x}_{i'})$$

$$B = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(\mathbf{x}_i)=k}\sum_{C(\mathbf{x}_i)\neq k} d^2(\mathbf{x}_i, \mathbf{x}_{i'})$$

$d^2(\mathbf{x}_i, \mathbf{x}_{i'})$ – kwadrat odległości pomiędzy dwoma przykładami, $C(\mathbf{x}_i)$ numer grupy przypisanej przykładowi $\mathbf{x}_i$, *K* – liczba grup.

Zmieniając podział punktów na *K* grup zmieniamy *W* i *B*. *T* pozostaje stałe.

Celem grupowania jest minimalizacja *W* (rozrzutu wewnątrzgrupowego) lub, równoważnie, maksymalizacja *B* (rozrzutu międzygrupowego).

* Koronacki J., Ćwik J.: Statystyczne systemy uczące się. WNT 2005.

ALGORYTM *K*-ŚREDNICH

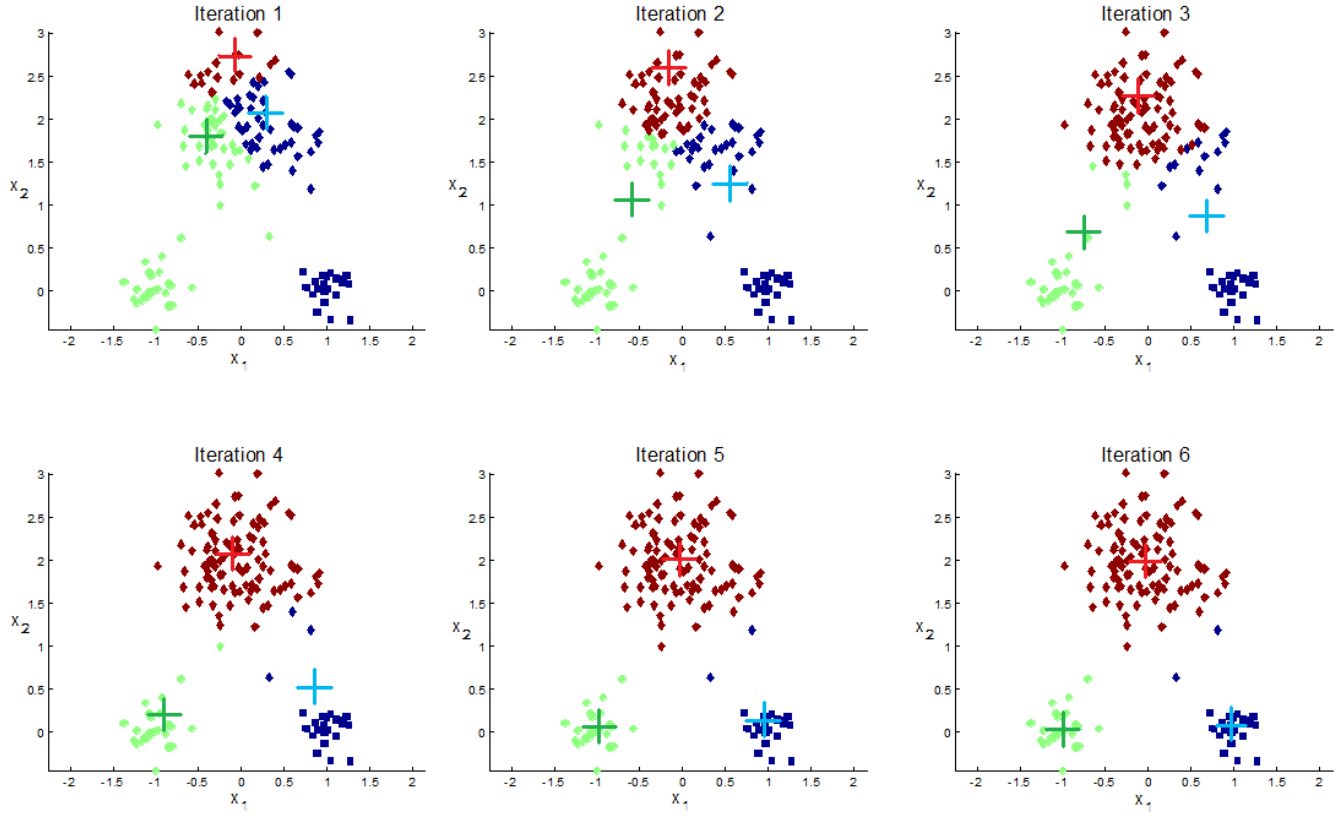
Jednym z najbardziej popularnych algorytmów grupowania danych jest algorytm *K*-średnich (*K-means*), w którym reprezentantem *k*-tej grupy jest wektorowa średnia (centroid) punktów należących do tej grupy – $\mathbf{m}_k$. Rozrzut wewnątrzgrupowy wyrażony jest tu w postaci:

$$W = \sum_{k=1}^{K} \sum_{C(\mathbf{x}_i)=k} d^2(\mathbf{x}_i, \mathbf{m}_k)$$

Algorytm *k*-średnich można zapisać w następujących krokach:

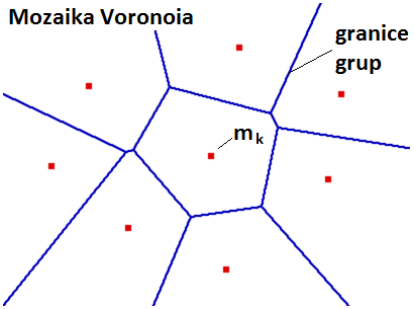
1. Zainicjuj *k* środków grup $\mathbf{m}_k$ (np. losując *k* przykładów ze zbioru trenującego)
2. Przydziel każdy przykład trenujący do najbliższego środka $\mathbf{m}_k$.
3. Wyznacz nowe środki grup $\mathbf{m}_k$ jako średnie wektorowe przykładów przydzielonych do tych grup.
4. Jeśli pozycje środków $\mathbf{m}_k$ nie zmieniają się, zakończ, w przeciwnym przypadku idź do punktu 2.

ALGORYTM *K*-ŚREDNICH



ALGORYTM *K*-ŚREDNICH

Algorytm *K*-średnich jest szybki, efektywny i zbieżny, ma małe wymagania pamięciowe, choć może dawać rozwiązania nieoptymalne. Aby tego uniknąć algorytm uruchamia się wielokrotnie z różnych punktów startowych. Podział przestrzeni w wyniku grupowania odpowiada mozaice Voronoia.



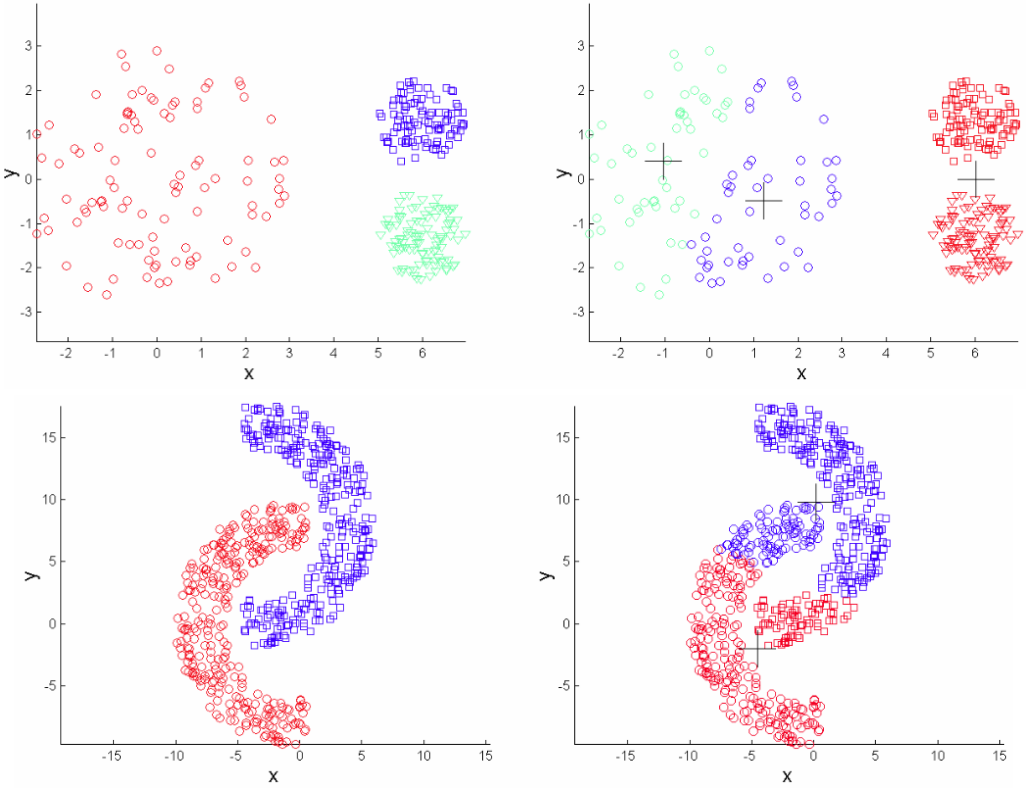
Wady tego algorytmu:

- problem z atrybutami nominalnymi
- liczba grup *K* musi być podana a priori
- wrażliwy na punkt startowy
- wrażliwy na błędne dane (odstające obserwacje)
- problemy z grupowaniem, gdy grupy różnią się wielkością, gęstością lub nie są wypukłe (patrz rysunek na następnym slajdzie)

Podobne algorytmy grupowania:

- algorytm *K*-medoidów, w którym każda grupa jest reprezentowana przez jeden z przykładów
- algorytm *K*-median, w którym zamiast wektorów średnich stosujemy wektory median do reprezentacji grup

ALGORYTM *K*-ŚREDNICH



GRUPOWANIE HIERARCHICZNE

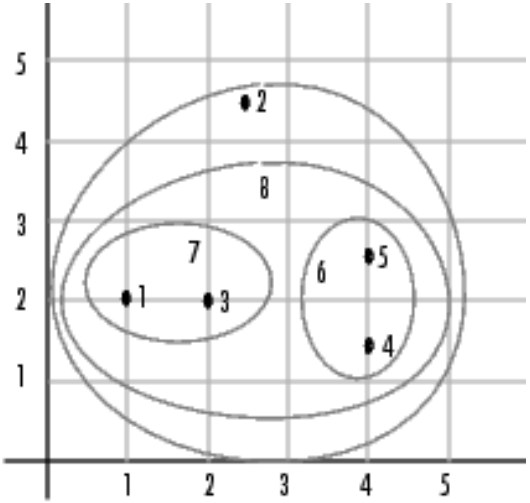
Metody hierarchiczne należą do najbardziej popularnych algorytmów grupowania i eksploracji danych. W wyniku działania tworzą hierarchię podziałów punktów na grupy, w której grupy na poziomie *l* formowane są z grup utworzonych na poziomie *l*–1. Punkty zgrupowane razem na poziomie *l*, pozostają w tej samej grupie na wyższych poziomach. Liczba grup nie musi być tutaj zadawana a priori.

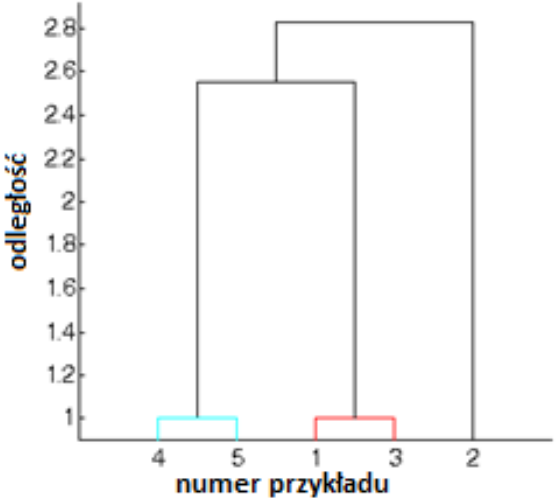
Proces grupowania hierarchicznego przebiega najczęściej według algorytmu aglomeracyjnego:

1. Przyjmuje się, że każdy przykład stanowi odrębną grupę. Otrzymujemy więc N grup jednoelementowych.
2. Powtarzamy N-1 razy:
   2.1. Wyznaczamy macierz odległości pomiędzy grupami (w pierwszym kroku będzie to macierz odległości pomiędzy przykładami $\boldsymbol{D} = [d_{ij}]$, $i, j = 1, 2, \ldots, N$).
   2.2. Znajdujemy parę najbliższych sobie grup.
   2.3. Grupy najbliżej położone łączy się, redukując liczbę grup o jeden.

GRUPOWANIE HIERARCHICZNE

Po wykonaniu algorytmu wszystkie klastry zostaną połączone w jeden. Proces grupowania można zwizualizować dendrogramem – drzewem hierarchicznym, którego liście stanowią przykłady **x**, węzły – grupy powstałe przez połączenie dwóch grup z niższego poziomu, a gałęzie obrazują łączone klastry i odległości pomiędzy nimi.





Proces grupowania przerywa się w momencie osiągnięcia założonej liczby grup (jeśli jest znana), przekroczenia progowej wartości odległości pomiędzy łączonymi grupami lub spełnienia przyjętego warunku jakości grupowania. W tym ostatnim przypadku wykorzystuje się miary wariancji wewnątrzgrupowej i międzygrupowej.

Proces grupowania hierarchicznego może zachodzić także wg algorytmu dzielącego, który wychodząc od jednej grupy obejmującej wszystkie przykłady, dzieli ją na dwie. Proces podziału powtarza się do momentu, gdy uzyskane grupy będą dostatecznie różne.

Odległości pomiędzy klastrami G i H, o liczebnościach odpowiednio N_G i N_H, definiuje się następująco:

- metoda pojedynczego wiązania (najbliższego sąsiada, *single link*)
  – minimalna odległość między przykładami należącymi do tych grup:

$$d(G,H) = \min_{i \in G,\ j \in H} d(\mathbf{x}_i, \mathbf{x}_j)$$

  (wrażliwa na dane odstające; ma tendencję do tworzenia długich, „cienkich” grup)

- metoda pełnego wiązania (najdalszego sąsiada, *complete link*)
  – maksymalna odległość między przykładami należącymi do tych grup:

$$d(G,H) = \max_{i \in G,\ j \in H} d(\mathbf{x}_i, \mathbf{x}_j)$$

  (pozwala uniknąć problemu łączenia (łańcuchowania) klastrów, wrażliwa na odstające dane)

- metoda średniego wiązania (średniej międzygrupowej, *average link*)
  – średnia arytmetyczna odległości między każdą parą przykładów należących do tych grup:

$$d(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d(\mathbf{x}_i, \mathbf{x}_j)$$

(niewrażliwa na odstające dane)

- metoda Warda
  – wyznaczana na bazie odległości pomiędzy środkami klastrów (centroidami):

$$d(G,H) = \sqrt{\frac{2N_G N_H}{N_G + N_H}} d_{Eukl}(\overline{\mathbf{x}_G}, \overline{\mathbf{x}_H})$$

(niewrażliwa na odstające dane)

GRUPOWANIE HIERARCHICZNE – PRZYKŁAD

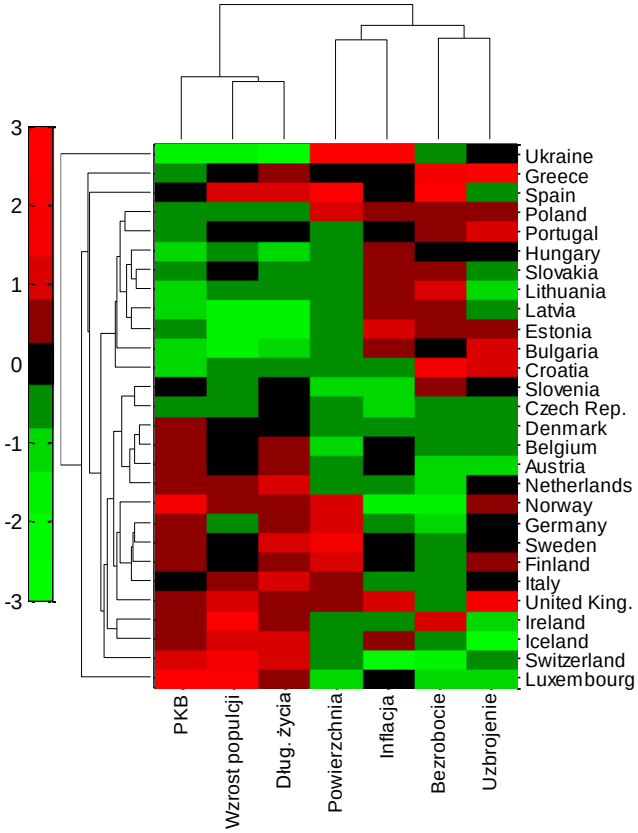
Dane są wskaźniki opisujące państwa europejskie:

Państwo	Powierzchnia	PKB	Inflacja	Dług. życia	Uzbrojenie	Wzrost populacji	Bezrobocie
Austria	83871	41600	3.5	79.91	0.8	0.03	4.2
Belgium	30528	37800	3.5	79.65	1.3	0.06	7.2
Bulgaria	110879	13800	4.2	73.84	2.6	-0.8	9.6
Croatia	56594	18000	2.3	75.99	2.39	-0.09	17.7
Czech Republic	78867	27100	1.9	77.38	1.15	-0.13	8.5
Denmark	43094	37000	2.8	78.78	1.3	0.24	6.1
Estonia	45228	20400	5	73.58	2	-0.65	12.5
Finland	338145	36000	3.3	79.41	2	0.07	7.8
Germany	357022	38100	2.5	80.19	1.5	-0.2	6
Greece	131957	26300	3.3	80.05	4.3	0.06	17.4
Hungary	93028	19600	3.9	75.02	1.75	-0.18	10.9
Iceland	103000	38100	4	81	0	0.67	7.4
Ireland	70273	40800	2.6	80.32	0.9	1.11	14.4
Italy	301340	30500	2.9	81.86	1.8	0.38	8.4
Latvia	64589	16800	4.4	72.93	1.1	-0.6	12.8
Lithuania	65300	19100	4.1	75.55	0.9	-0.28	15.4
Luxembourg	2586	80600	3.4	79.75	0.9	1.14	5.7
Netherlands	41543	42000	2.3	80.91	1.6	0.45	4.4
Norway	323802	53400	1.3	80.32	1.9	0.33	3.3
Poland	312685	20200	4.2	76.25	1.9	-0.08	12.4
Portugal	92090	23400	3.7	78.7	2.3	0.18	12.7
Slovakia	49035	23300	3.9	76.03	1.08	0.1	13.2
Slovenia	20273	28800	1.8	77.48	1.7	-0.19	11.8
Spain	505370	30500	3.1	81.27	1.2	0.65	21.7
Sweden	450295	40700	3	81.18	1.5	0.17	7.5
Switzerland	41277	44500	0.2	81.17	1	0.92	2.8
Ukraine	603550	7200	8	68.74	1.4	-0.63	7.9
United Kingdom	243610	36500	4.5	80.17	2.7	0.55	8.1

GRUPOWANIE HIERARCHICZNE – PRZYKŁAD

Wyznacz dendrogram grupujący państwa według ww. wskaźników.

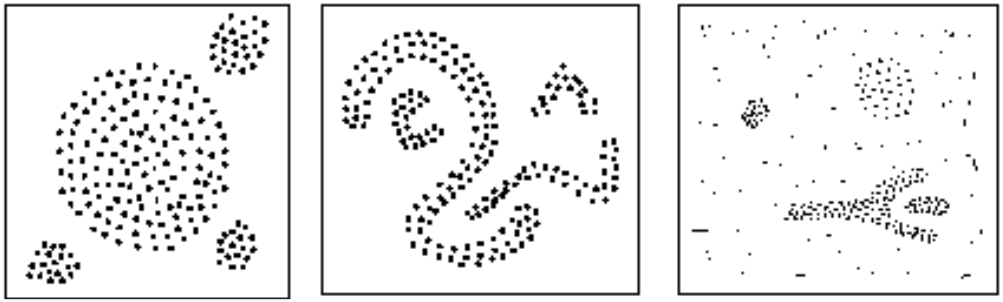
Przed utworzeniem dendrogramu wskaźniki (atrybuty) poddano standaryzacji.



ALGORYTMY GRUPOWANIA OPARTE NA GĘSTOŚCIACH

Główne cechy:

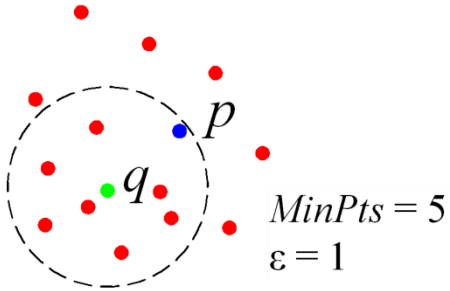
- bazują na wyszukiwaniu punktów gęsto ułożonych (regiony pokrywające grupy mają wyższą gęstość niż regiony na zewnątrz grup)
- tworzą grupy o dowolnych kształtach
- wyłapują szum i punkty oddalone
- jednokrotnie przeglądają zbór danych



ALGORYTMY GRUPOWANIA OPARTE NA GĘSTOŚCIACH

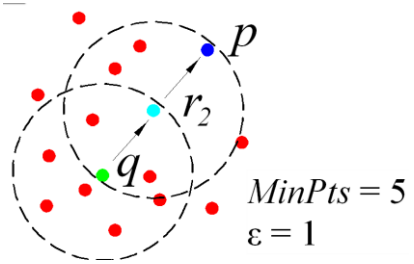
Podstawowe pojęcia:

- ε – promień definiujący otoczenie punktu
- ε-otoczenie punktu p – $N_\varepsilon(p)$: $\{q \in D \mid d(p,q) \leq \varepsilon\}$, gdzie $d(p,q)$ jest odległością pomiędzy p i q
- *MinPts* – minimalna liczba punktów w ε-otoczeniu
- Rdzeń – obiekt, który ma co najmniej *MinPts* punktów w swoim ε-otoczeniu
- Punkt brzegowy – punkt, który ma mniej niż *MinPts* punktów w swoim ε-otoczeniu.
- Punkty niepołączone z żadnym skupiskiem to tzw. punkty oddalone (*outliers*)
- Punkt p jest bezpośrednio wyprowadzalny z punktu q, jeśli:
  - $p \in N_\varepsilon(q)$
  - $|N_\varepsilon(q)| \geq MinPts$

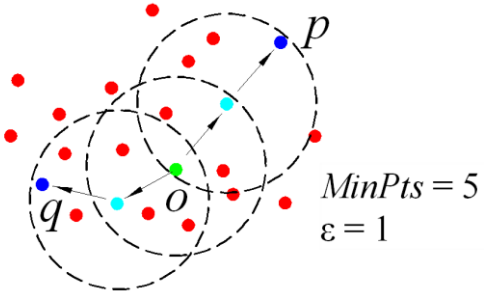


ALGORYTMY GRUPOWANIA OPARTE NA GĘSTOŚCIACH

- Punkt p jest wyprowadzalny z punktu q, jeśli istnieje ciąg punktów $r_1, \ldots, r_n$ taki, że $r_1 = q$, $r_n = p$ i r_{i+1} jest bezpośrednio osiągalny z r_i



- Punkt p i q są połączone, jeśli istnieje punkt o taki, że p i q są wyprowadzalne z o



- Grupa – maksymalny zbiór punktów połączonych

ALGORYTMY GRUPOWANIA OPARTE NA GĘSTOŚCIACH

Przykład algorytmu opartego na gęstości:

1. Wybierz dowolny punkt *p*
2. Wyszukiwać zbiór *G* wszystkich punktów osiągalnych z punktu *p* dla ustalonych ε i *MinPts*
3. Jeśli *p* jest rdzeniem, zwróć *G* (tworzymy grupę)
4. Jeśli *p* jest punktem brzegowym, to sprawdź następny nieodwiedzony punkt
5. Jeśli *p* jest punktem oddalonym, to zaznaczamy go jako takiego
6. Kontynuuj, aż wszystkie punkty zostaną odwiedzone

Silhouette

Wartość silhouette przykładu jest wyznaczana na podstawie średniej odległości tego przykładu od przykładów w tej samej grupie (a_i) i minimalnej średniej odległości tego przykładu od przykładów w pozostałych grupach (b_i):

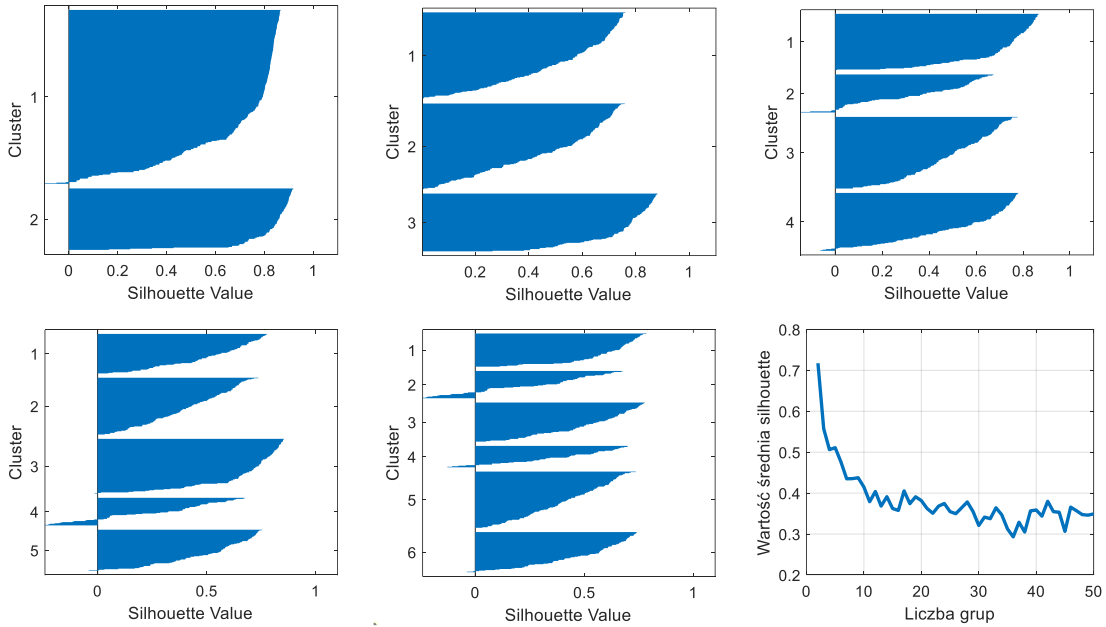
$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)} \in [-1, 1]$$

Wysoka wartość s_i wskazuje, że przykład jest dobrze dopasowany do własnej grupy, a słabo do innych grup.

Grupowanie jest poprawne, gdy większość przykładów ma wysoką wartość s_i. Jeśli wiele przykładów ma niską lub ujemną wartość s_i, oznacza to, że liczba grup może być źle dobrana.

OCENA GRUPOWANIA

Przykład: dane carbig (atrybuty: Acceleration, Displacement, Horsepower, Weight, MPG)



ANALIZA GŁÓWNYCH SKŁADOWYCH

Analiza głównych składowych (*Principal Component Analysis*, PCA) metoda redukcji wymiarowości danych, wykorzystywana do wizualizacji danych wielowymiarowych. Często okazuje się, że niewielka liczba głównych składowych zawiera nieomal tyle samo informacji o strukturze danych, co atrybuty oryginalne.

PCA na podstawie wektorów oryginalnych **x** tworzy nowe wektory **x**' o składowych wzajemnie nieskorelowanych (tzw. składowe główne). Składowe główne wyjaśniają w maksymalnym stopniu całkowitą wariancję składowych oryginalnych.

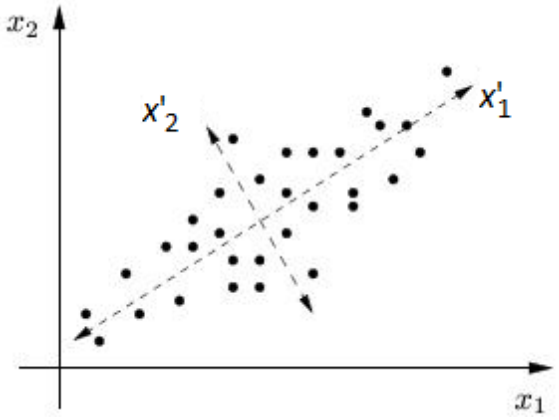
Składowe wektora **x**' są liniowymi kombinacjami składowych wektora **x** (zakłada się, że wektory **x** mają zerową średnią):

$$x'_{i,j} = \sum_{l=1}^{n} a_{j,l} x_{i,l} = \mathbf{a}_j^{\mathsf{T}} \mathbf{x}_i, \quad j = 1,2,\ldots,n \qquad (*)$$

gdzie wektor współczynników $\mathbf{a}_j$ jest wektorem charakterystycznym odpowiadającym kolejnym największym wartościom własnym macierzy kowariancji z próby uczącej. Składowe główne są unormowane i wzajemnie ortogonalne, tj. $\mathbf{a}_j^{\mathsf{T}} \mathbf{a}_j = 1$ i $\mathbf{a}_j^{\mathsf{T}} \mathbf{a}_k = 0$ dla wszystkich $j \neq k$.

Pierwsza główna składowa $x'_{i,1}$ wykazuje największą wariancję, kolejne główne składowe mają wariancje coraz mniejsze.

Ponieważ najwięcej informacji przenoszą początkowe składowe, to końcowe składowe o najmniejszych wariancjach (często nieprzekraczających szumu pomiarowego) można odrzucić.

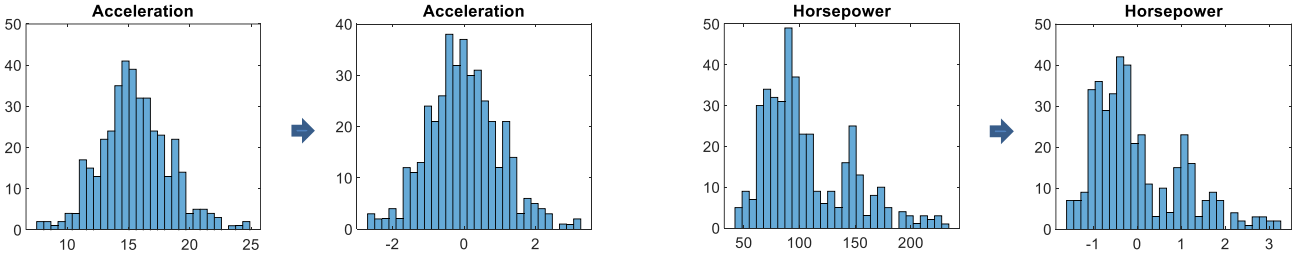


Proces analizy głównych składowych:

1) Standaryzacja danych (jeśli atrybuty mają różne zakresy)

$$x_s = \frac{x - \bar{x}}{\sigma_x}$$

Po standaryzacji wartość średnia atrybutu wynosi 0, a jego odchylenie standardowe wynosi 1.



2) Obliczanie macierzy kowariancji (korelacje między zmiennymi).

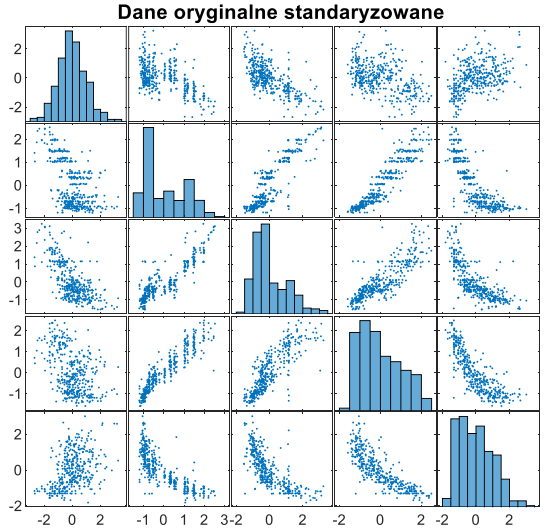
3) Obliczanie składowych głównych: Korzystając z macierzy kowariancji, oblicza się wektory własne i wartości własne.
   Wektory własne reprezentują kierunki składowych głównych (**a**), a wartości własne określają wariancję w tych kierunkach. Wektory własne są uporządkowywane malejąco według wartości własnych.

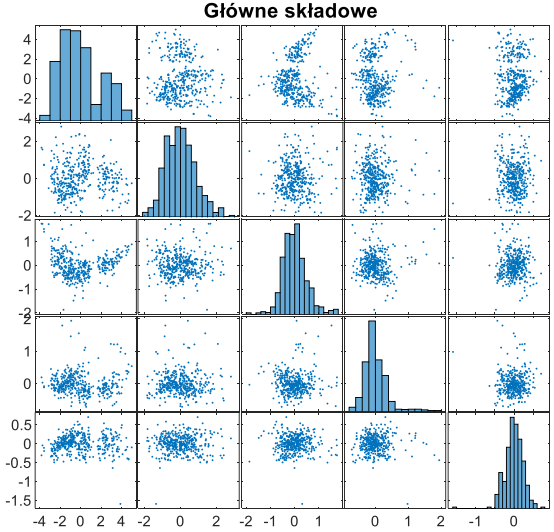
4) Wybór składowych głównych: Wybiera się pierwsze *k* składowych głównych, które wyjaśniają największą część wariancji w danych (można przyjąć pewien próg (np. 80%) wariancji, która ma być zachowana w danych reprezentowanych przez *k* składowych głównych).

5) Transformacja nowych danych: Dane oryginalne po standaryzacji są przekształcane wg (*).
   Ta transformacja pozwala uzyskać nowe reprezentacje danych, w których atrybuty są nieskorelowane i uporządkowane według ich wartości własnych.

ANALIZA GŁÓWNYCH SKŁADOWYCH

Przykład: dane carbig (atrybuty: Acceleration, Displacement, Horsepower, Weight, MPG)





Wariancja wyjaśniana przez główne składowe