WYKŁAD VIII: ANALIZA SKUPIEŃ

MiNI PW, semestr letni 2013/2014

Analiza skupień

Dotąd problemy dotyczące analizy dyskryminacyjnej (klasyfikacji pod nadzorem): określonych g populacji (klas) oraz dana próba ucząca zawierająca informacje na temat wektora atrybutów jak i przynależności do klas.

Analiza skupień (klasyfikacja bez nadzoru): populacje (klasy) nie są określone i nie mamy próby uczącej. Na postawie obserwacji $\mathbf{x}_1, \ldots, \mathbf{x}_n$ chcemy podzielić je na skupienia i, na podstawie tych skupień, określić odpowiadające populacje.

Podstawowy problem: Nie ma całkowicie satysfakcjonującej, formalnej definicji skupienia, którą można by wykorzystać w konstrukcji algorytmu. Popularne podejście oparte na **odmiennościach** między obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_j$.

Odmienności powinny mieć następujące własności:

- $d(x_i, x_j) \geqslant 0$
- $d(x_i, x_i) = 0$ (czasami $d(x_i, x_j) = 0 \equiv x_i = x_j$
- $d(x_i, x_j) = d(x_j, x_i)$

Jeśli odmienność spełnia dodatkowo warunek trójkąta to jest to **odmienność metryczna**. Jeśli spełnia

$$d(x_i, x_j) \leqslant \max\{d(x_i, x_k), d(x_j, x_k)\}$$

jest **odmiennością ultrametryczną**.
Typowe odmienności: odległość euklidesowa i jej kwadrat, odległość Minkowskiego, metryka Manhattan, również dla $x, y \in R^p$

$$d(x, y) = 1 - r(x, y),$$

gdzie $r(x, y)$ empiryczny współczynnik korelacji (używania w skupianiu zmiennych).

Cel analizy skupień

Cel analizy skupień (w tym podejściu): podział zbioru obserwacji na grupy (skupienia) w taki sposób, żeby odmienność obserwacji w tej samej grupie była z reguły mniejsza niż odmienność obserwacji w różnych grupach.

– metody kombinatoryczne;
– metody hierarchiczne;
– inne metody (modelowanie mieszaninami rozkładów, mapy samoorganizujące się, itp.)

Metody kombinatoryczne

Analiza skupień dla obserwacji $\mathbf{x}_1, \ldots \mathbf{x}_n$ w R^p:
$\mathbf{x}_i, i = 1, 2, \ldots, n$ dzielimy na na K skupień (K ustalone z góry (na razie)).

Suma kwadratów odległości elementów próby

$$T = \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'}$$

gdzie $d_{ii'} = d(\mathbf{x}_i, \mathbf{x}_{i'})$ jest kwadratem odległości euklidesowej między obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_{i'}$.

Rozkład całkowitej sumy kwadratów

Załóżmy, że dokonaliśmy podziału na K skupień.

$C(i) = k$ – gdy x_i należy do $k-$tego skupienia.

T – suma kwadratów odległości między elementami tego samego skupienia i różnych skupień

$$T = W + B$$

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d_{ii'}$$

$$B = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'}$$

(T – total, W – within, B – between)

Minimalizacja W, czyli minimalizacja rozrzutu punktów wewnątrz skupień $\equiv$ maksymalizacji rozrzutu punktów między skupieniami.

Optymalizacja kombinatoryczna

Kwestia znalezienia odpowiedniego podziału – zadanie optymalizacji kombinatorycznej.

Niemożliwe sprawdzenie wszystkich podziałów !

$S(n, K)$ – liczba podziałów $(=K!^{-1} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^N)$
$S(10, 4) = 34105$
$S(19, 4) \simeq 10^{10}$

Iteracyjne szukanie minimum W ze względu na funkcję C.

$$\frac{1}{2} \sum_{C(i)=k} \sum_{C(i')=k} \underbrace{\| \mathbf{x}_i - \mathbf{x}_{i'} \|^2}_{d(\mathbf{x}_i, \mathbf{x}_{i'})} = \sum_{C(i)=k} \| \mathbf{x}_i - \mathbf{m}_k \|^2 \, n_k, \qquad n_k = \#(C_k)$$

Zatem

$$W = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) n_k, \qquad \mathbf{m}_k = \frac{1}{n_k} \sum_{C(i)=k} \mathbf{x}_i.$$

Metoda k-średnich

W praktyce: dążymy do takiego podziału, który minimalizuje sumy kwadratów odległości obserwacji od środka swojego skupienia

$$\bar{W} = \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

(pomijamy ważenie przez liczność skupienia !)

Metoda K–średnich (K–means). Zachłanne algorytmy minimalizacji $\bar{W}$.

Podejście 1 (wsadowe)

- Wybór początkowych środków $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K$;
  określenie $C(i)$: dla obserwacji $\mathbf{x}_i$ wybieramy środek najbliższy;
  wszystkie punkty podzielone na skupienia.
- Wyznaczamy nowe $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K$ – środki powstałych skupień ;
  wyznaczamy nową funkcje $C(i)$.
- stop – gdy w kolejnym kroku żaden punkt próby nie został przeniesiony z jednego skupienia do drugiego.

Podejście 2 (sekwencyjne - zależy od kolejności $\mathbf{x}_i$).

Różnica z podejściem 1: po przydzieleniu $\mathbf{x}_i$ do średniej $\mathbf{m}_k$ od razu przeliczamy średnią w skupieniu.

Podejście 3 (z podziałem początkowym)

Zamiast $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K$ mamy podział próby,
z niego liczymy średnie – dalej tak samo.

Wybór początkowych środków $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K$ (i) losowanie K elementów próby;
(ii) metoda iteracyjna: $\mathbf{m}_1$ – srednia $\mathbf{x}_i$,
$\mathbf{m}_2$: minimalizuje $\bar{W}$ dla $k = 2$ i z wartością $\mathbf{m}_1 =$ średniej z obserwacji;
$\mathbf{m}_3$: minimalizuje $\bar{W}$ dla $k = 3$ z ustalonymi wartościami $\mathbf{m}_1$ i $\mathbf{m}_2$ itd.

Kwantyzacja wektorowa

Zastosowanie: Kompresja sygnału – kwantyzacja wektorowa
zdjęcie czarno-białe 1024×1024 pikseli
każdy piksel – jeden z 256 odcieni szarości
piksele podzielone na większe kwadraty po 4 piksele (traktowane jako elementy R^4)

$$(512)^2 \text{ wektorów z } R^4$$
$$\Downarrow$$
$$\text{analiza skupień z ustalonym } K$$
$$\Downarrow$$

przesłaniu podlegają średnie skupień oraz współrzędne kwadratów przyporządkowane każdemu skupieniu.

Wybór liczby skupień K

– nie ma jednego uniwersalnego algorytmu. Najczęściej metody oparte na zmianie $\bar{W}_k$

$$\bar{W}_k = \sum_{i=1}^{n} d(x_i, m_{C(i)}), \qquad C(i) \in \{1, 2, \ldots, k\}$$

Prawdziwa liczba skupień K^*

intuicja: dla $k < K^*$, $\bar{W}_k - \bar{W}_{k+1}$ – „duże” (**)

dla $k \geqslant K^*$, $\bar{W}_k - \bar{W}_{k+1}$ – „małe”

(dla $k < K^*$ muszą istnieć skupienia zawierające odległe od siebie punkty, podział takich skupień – istotna zmiana $\bar{W}$)

Kwestia ustalenia progów odpowiadających (**).

Statystyka odstępu (gap statistic): oparta na zmianach $\log \bar{W}_k$ i porównaniu ich z analogicznymi wielkościami dla próby z rozkładu jednostajnego na hiperkostce opisującej zbiór obserwacji.

Inne metody.

Dla ustalonego K podział na skupienia traktujemy jak przydział do klas w problemie dyskryminacji pod nadzorem. Właściwe K – wartość dająca mały estymowany błąd klasyfikacji.

Ocena jakości podziału na skupienia

$C_1, \ldots, C_k$ podział na skupienia. Niech $\mathbf{x}_i \in C_l$. Definiujemy średnią odmienność $\mathbf{x}_i$ od elementów swego skupienia:

$$a(\mathbf{x}_i) = \frac{\sum_{\mathbf{u} \in C_l} d(\mathbf{x}_i, \mathbf{u})}{n_l}$$

oraz średnia odmienność $\mathbf{x}_i$ od skupienia C

$$d(\mathbf{x}_i, C) = \frac{\sum_{\mathbf{u} \in C} d(\mathbf{x}_i, \mathbf{u})}{|C|},$$

gdzie $|C|$ - liczba elementów w C. Odległość $\mathbf{x}_i$ od 'najbliższego' skupienia (innego niż C_l)

$$b(\mathbf{x}_i) = min_{C \neq C_l} d(\mathbf{x}_i, C)$$

Sylwetka (*silhouette*) obserwacji $\mathbf{x}_i$

$$sil(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max(b(\mathbf{x}_i), a(\mathbf{x}_i))}$$

Ocena jakości podziału na skupienia cd

$$sil(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max(b(\mathbf{x}_i), a(\mathbf{x}_i))}$$

$$-1 \leqslant sil(\mathbf{x}_i) \leqslant 1$$

$sil(\mathbf{x}_i) \approx 1$, to $\mathbf{x}_i$ dobrze zaklasyfikowany $(a(\mathbf{x}_i)) << b(\mathbf{x}_i))$.
$sil(\mathbf{x}_i) \approx -1$, to $\mathbf{x}_i$ mniej odmienny od obserwacji w najbliższym skupieniu niż w skupieniu, do którego został zaklasyfikowany.
Często interesujące obserwacje, bo odpowiadają brzegom skupień.
Współczynnik jakości podziału oparty na sylwetkach (Kaufman, Roussseuv)

$$SC = \mathrm{ave}(\bar{s}_k),$$

gdzie $\bar{s}_k$ średnia sywetek k-tego skupienia.

Metody hierarchiczne

Metody hierarchiczne

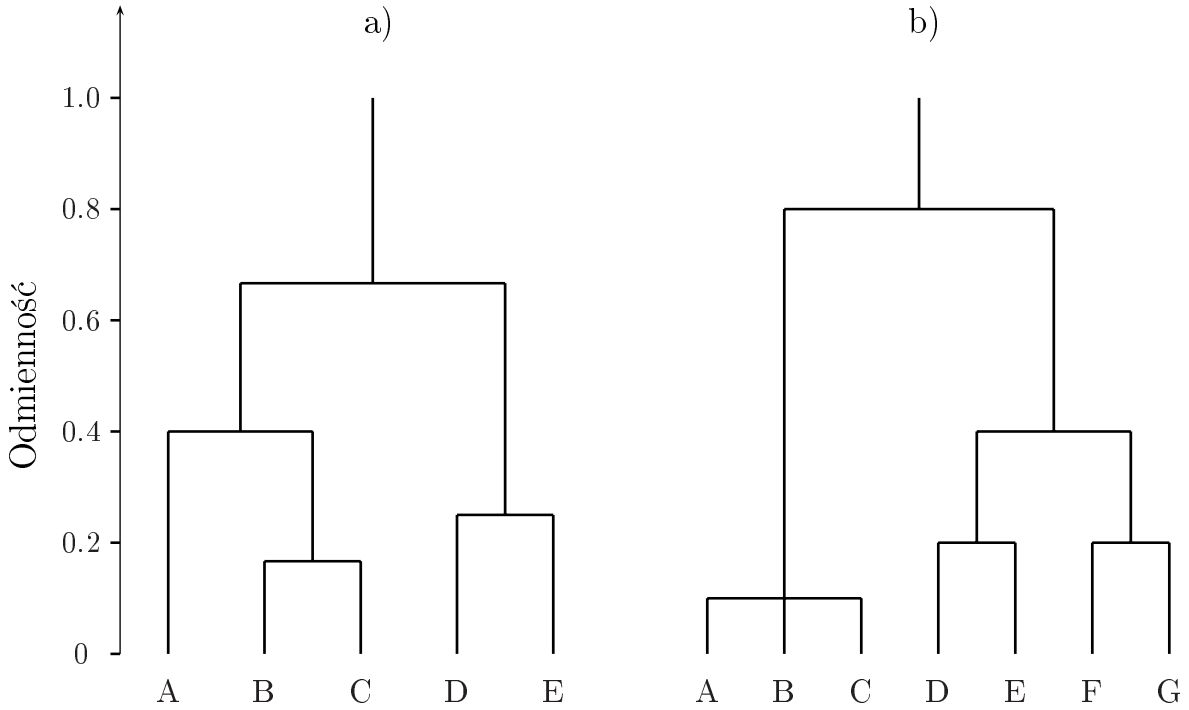
↓ ↓

aglomeracyjne oparte na dzieleniu zbioru danych

– oparte na koncepcji odmienności między dwoma zbiorami obserwacji (powinna sprowadzać się do zwykłej odmienności, gdy zbiory redukują się do obserwacji)

Metoda aglomeracyjna

początek: każdy z n punktów stanowi oddzielne skupienie;
– łączymy w skupienie dwa punkty najmniej odmienne ;
– łączymy w skupienie dwa najmniej odmienne skupienia, uzyskane w poprzednim kroku.
Za każdym razem otrzymujemy o jedno skupienie mniej – na końcu jedno skupienie.



Rys. 9.4. Dendrogramy.

Próba $A, B, C, D, E \longrightarrow A, \{B, C\}, D, E \longrightarrow A, \{B, C\}, \{D, E\} \longrightarrow \{A, B, C\}, \{D, E\} \longrightarrow \{A, B, C, D, E\}$

Metoda oparta na dzieleniu – startujemy z jednego skupienia
$\longrightarrow$ podział na dwa podzbiory najbardziej odmienne od siebie $\longrightarrow$ itd

większa złożoność obliczeniowa

Obie metody nie wymagają określenia liczby skupień – powstały dendrogram możemy przeciąć na dowolnym poziomie odmienności.

Odmienności między „skupieniami".

D_{ij} – odmienność między skupieniem i a skupieniem j
1) odmienność najbliższego sąsiada (single linkage)

$$\min d_{kk'}$$

k – odpowiada obserwacji z i–tego skupienia
k' – odpowiada obserwacji z j–tego skupienia

2) odmienność najdalszego sąsiada (complete linkage)

$$\max d_{kk'}$$

3) odmienność średnia (average linkage): $\bar{d}_{..}$

Jak zmieniają się powyższe odmienności przy połączeniu dwóch skupień?

Łączymy skupienie i–te i j–te

$D_{k \cdot ij}$ – odmienność k–tego skupienia od połączonego i–tego i j–tego

Odmienność najbliższego sąsiada

$$D_{k \cdot ij} = \min(D_{ki}, D_{kj}) = \frac{1}{2}(D_{ki} + D_{kj} - |D_{ki} - D_{kj}|)$$

Odmienność najdalszego sąsiada

$$D_{k \cdot ij} = \max(D_{ki}, D_{kj}) = \frac{1}{2}(D_{ki} + D_{kj} + |D_{ki} - D_{kj}|)$$

Odmienność średnia

$$D_{k \cdot ij} = \frac{n_i}{n_i + n_j} D_{ki} + \frac{n_j}{n_i + n_j} D_{kj}$$

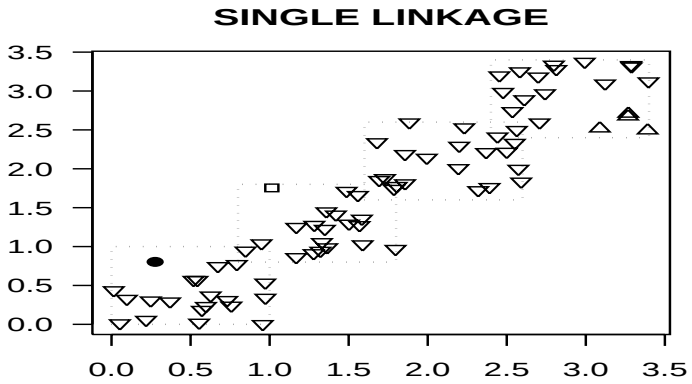
Analiza oparta na odmienności najbliższego sąsiada:
– tendencja do otrzymywania wąskich i wydłużonych skupień (efekt łańcuchowy)
(por. pierwszy z rysunków, gdzie jedno skupienie zawiera (mylnie !) większość punktów z trzech kwadratów)

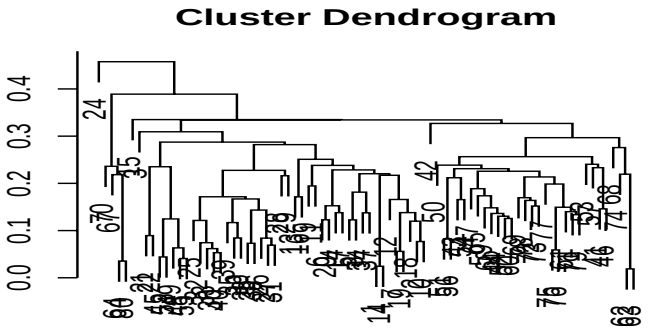
Najdalszy sąsiad: tendencja do tworzenia małych, „kulistych" skupień

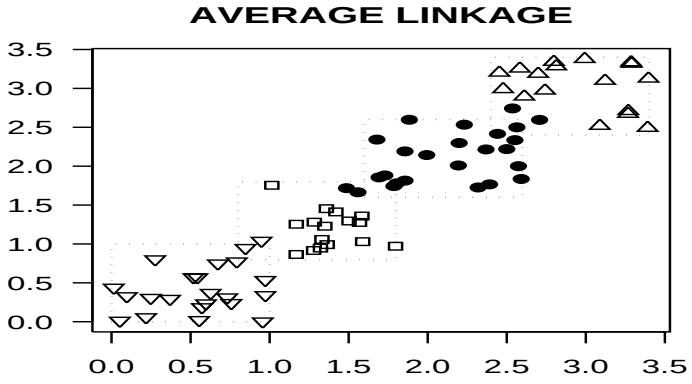
Odmienność średnia: kompromis pomiędzy dwiema pierwszymi.

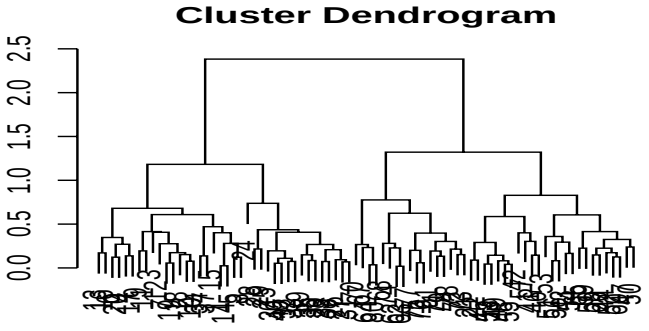
Zamiast ustalać liczbę skupień można ustalić wartość progową odmienności, po przekroczeniu której zaprzestaje się łączenia skupień.
Uwaga Otrzymane skupienia silnie zależą od wyjściowej definicji

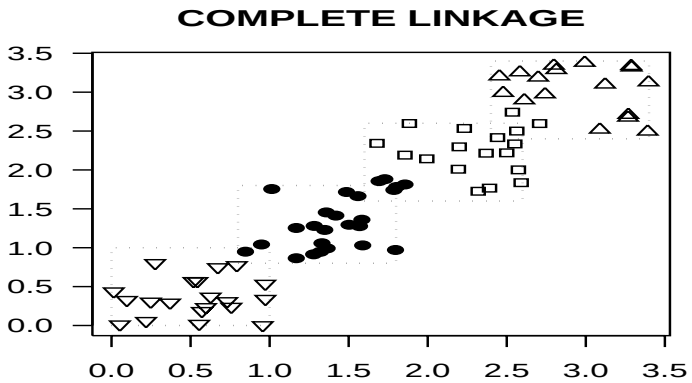
odmienności dla par pojedyńczych obserwacji, szczególnie dla obserwacji zawierających zmienne liczbowe. Jeśli składowa odmienności związana z taką zmienną nie jest unormowana (jak np odległośc euklidesowa) to odmienność silnie zależy od jednostek, w których została wyrażona ta zmienna.

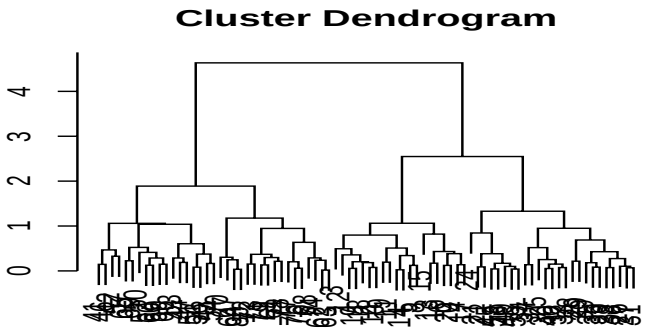

SINGLE LINKAGE
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5

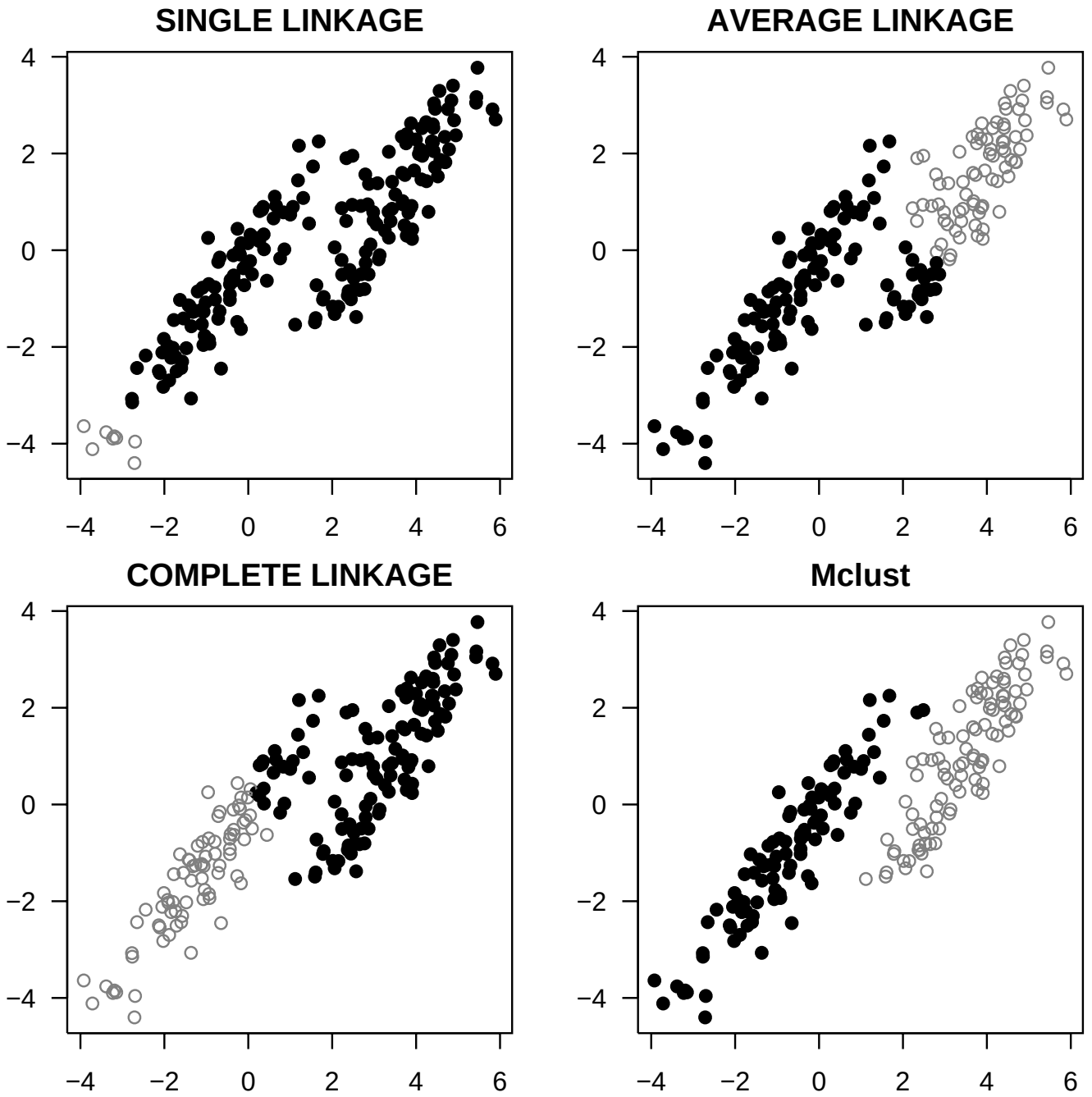

Cluster Dendrogram
0.0 0.1 0.2 0.3 0.4


AVERAGE LINKAGE
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5


Cluster Dendrogram
0.0 0.5 1.0 1.5 2.0 2.5


COMPLETE LINKAGE
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5


Cluster Dendrogram
0 1 2 3 4


SINGLE LINKAGE
AVERAGE LINKAGE
COMPLETE LINKAGE
Mclust
4
2
0
−2
−4
−4
−2
0
2
4
6

Dotąd: odmienność dla wektorów $\in R^p$ (np. odległość euklidesowa, jej kwadrat, dowolna inna metryka).

Dla wektorów o wsp. binarnych $(0, 1)$
$x = (0, 1, 0, \ldots, 0), \quad y = (1, 1, 0, \ldots, 0)$

$a = \#\{x_i = 1 \wedge y_i = 1\}$, $b = \#\{x_i = 0 \wedge y_i = 1\}$
$c = \#\{x_i = 1 \wedge y_i = 0\}$, $d = \#\{x_i = 0 \wedge y_i = 0\}$

	x=1	x=0
y=1	a	b
y=0	c	d

Miary odmienności

$$d_{ij} = 1 - \frac{a+d}{a+b+c+d} = \frac{b+c}{a+b+c+d}$$

$$\text{Jaccard:} \quad d_{ij} = 1 - \frac{a}{a+b+c} = \frac{b+c}{a+b+c}$$

nie uwzględniamy zgodnych wystąpień braku cechy (np. uzasadnienie ekologiczne: fakt, że dwa miejsca nie mają pewnej własności nie czyni ich bardziej podobnymi).

Czekanowski: $d_{ij} = 1 - \frac{2a}{2a+b+c}$

Dane jakościowe o więcej niż 2 poziomach:

$$1 - \frac{\#\text{ współrzędnych o tych samych wartościach}}{\#\text{ współrzędnych}}$$

Zmienne na skali porządkowej

M wartości zmiennych. Częste zastępowanie przez

$$\frac{i - \frac{1}{2}}{M}, \quad i = 1, 2, \ldots, M$$

i traktowane jako wartości rzeczywiste.

Dla danych mieszanych – system Gowera

Adaptacja algorytmu k–średnich na zmienne dowolnego typu

k–średnich $(\star)$ $\bar{W} = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} d(x_i, m_k)$, m_k – srednia k–tego skupienia

dla d – kwadrat odległości euklidesowej, średnia próbkowa ma wlasność

$$\bar{\mathbf{m}}_Z = \underset{y \in R^p}{\operatorname{argmin}} \sum_{\mathbf{x}_i \in Z} d(\mathbf{x}_i, \mathbf{y})$$

Modyfikacja ($\star$)

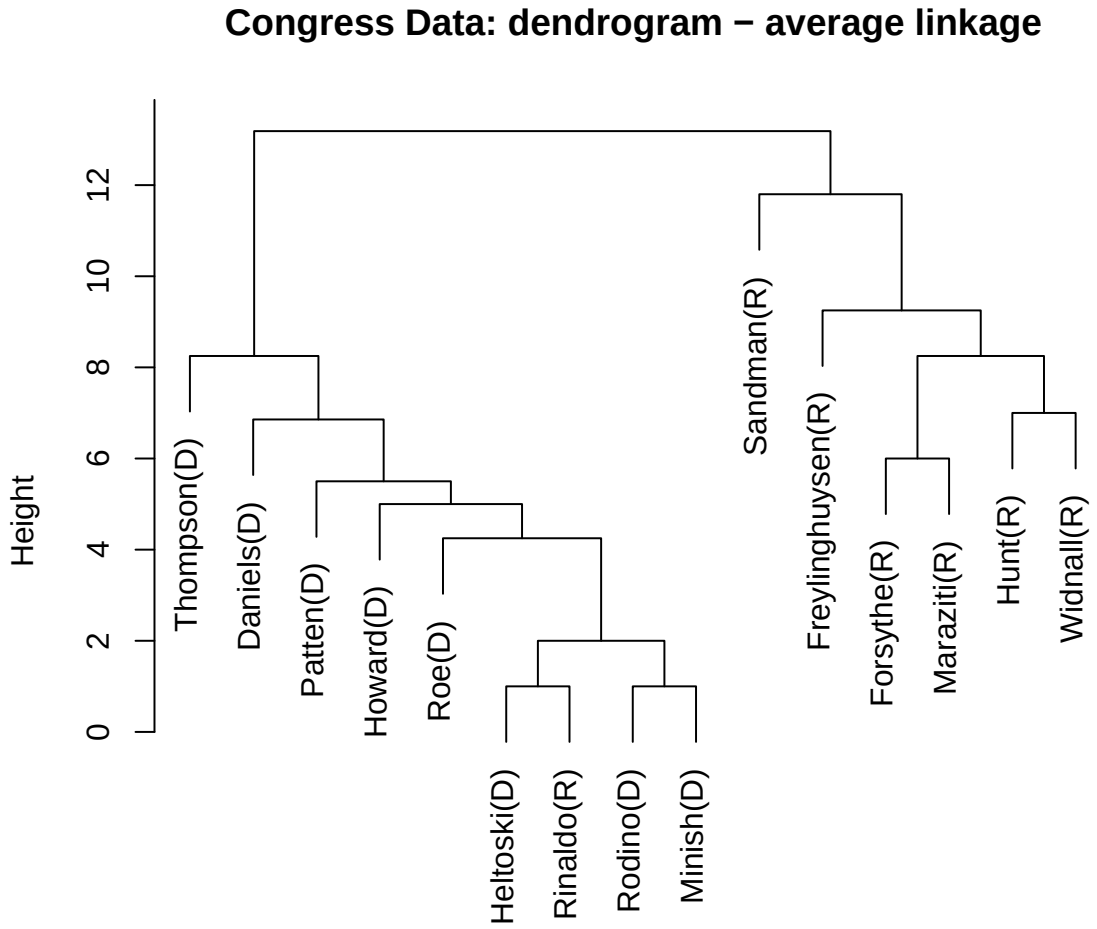
$$\underset{\{y_{C(i)}\}_1^K \in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} d(x_i, y_{C(i)})$$

dla odmienności w pierwszej potędze (np. odległości euklidesowej) i zmiennych dowolnego typu – algorytm k–medoidów lub k–median.

Inne metody wyznaczania skupień Modelowanie mieszaninami rozkładów. Polega na modelowaniu zbioru punktów jako próby z mieszanki wielowymiarowych gęstości $g(\mathbf{x}) = \sum_{i=1}^{K} \pi_i f(\mathbf{x}, \theta_k)$ (dla ustalonego K) i estymacji parametrów mieszanki przy użyciu iteracyjnego algorytmu EM. Procedura `mclust` wykorzystuje jako składniki mieszanki gęstości normalne.

W celu wstępnego zorientowania się jakie są skupienia w danych podanych przez zadanie ich macierzy odmienności, często używa się metody skalowania wielowymiarowego. Jest to metoda reprezentacji danych w niskowymiarowej przestrzeni euklidesowej możliwie wiernie odzwierciedlająca odmieności między obiektami. Jeśli odmienności spełniają warunek trójkąta, można stosować skalowanie metryczne. W R realizowane przez procedurę `cmdscale`.

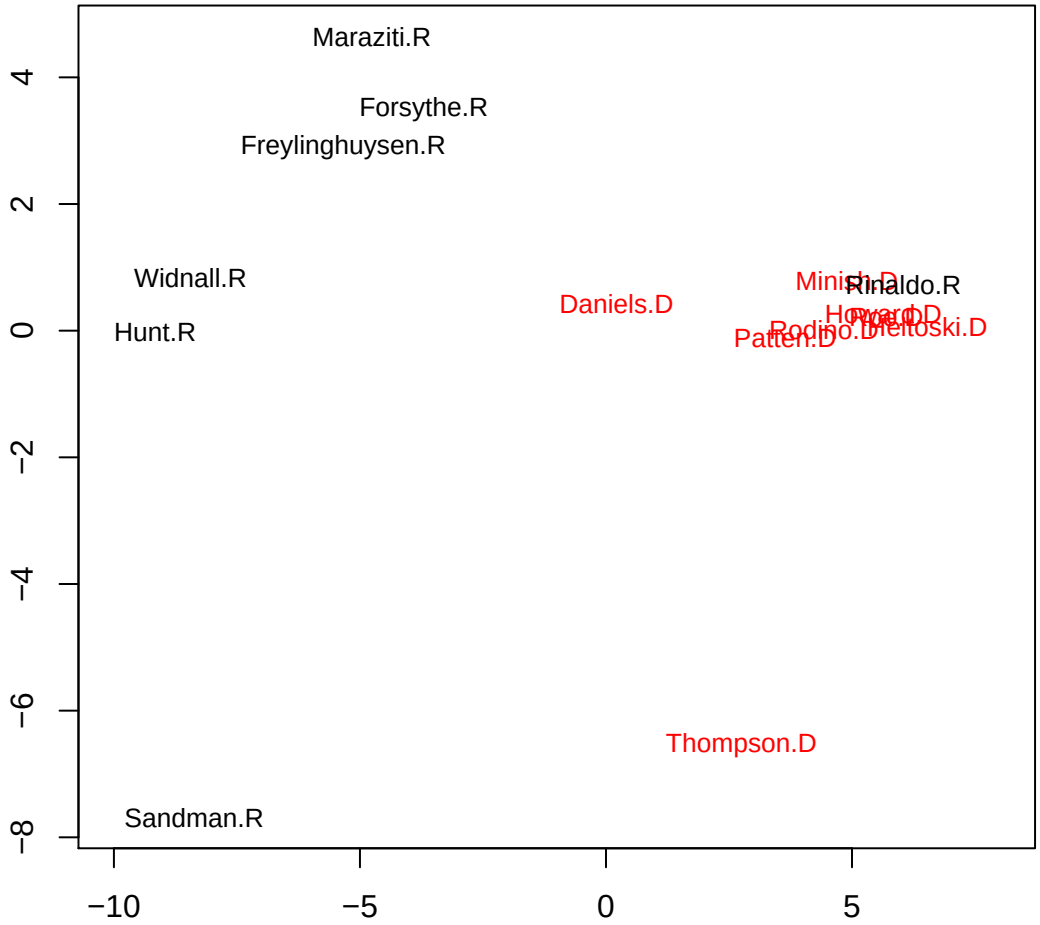
Przykład (CM (2009)) dane `congress.txt` zawierają macierz rozbieżności w głosowaniach dla 15 kongresmenów dotyczących 19 głosowań w sprawach ekologicznych (3 możliwe wyniki głosowania: za, przeciw, wstrzymanie się od głosu). Odległość między dwoma kongresmenami: liczba głosowań, w których głosowali różnie.



Funkcje hclust i cmdscale

```
congress <- read.table("congress.txt", header=TRUE)

c3.hc <- hclust(as.dist(congress), method="average")

plot(c3.hc, xlab="", main="Congress Data: dendrogram - average linkage"

loc <- cmdscale(as.dist(congress))
kolor <- c("black", "black", "red", "red", "black", "black", "black",
"red", "red", "red", "red", "black", "black", "red", "red")
x <- loc[,1]
y <- -loc[,2]
plot(x, y, type="n", xlab="", ylab="", xlim=c(-10,8),
     main="Congress Data: multidimensional scaling")

###   w kolorze
text(x, y, names(congress), cex=0.8, col=kolor)
```

Congress Data: multidimensional scaling



Republikanin o nazwisku Rinaldo znajduje się blisko kongresmenów z partii demokratycznej. Sandman (R) i Thompson (D) głosują wyraźnie inaczej niż ich koledzy partyjni.