Algorytmy eksploracji danych: Wykład 5

Wojciech Kempa

Politechnika Śląska
Wydział Matematyki Stosowanej

Wybór wymiaru przestrzeni rzutowania (1)

- Podobnie jak w przypadku analizy składowych głównych czy też analizy czynnikowej, wybór wymiaru przestrzeni rzutowania (liczby osi układu współrzędnych) jest określany za pomocą miernika udziału danej osi w wyjaśnieniu całkowitej inercji oryginalnego układu.

Wybór wymiaru przestrzeni rzutowania (1)

- Podobnie jak w przypadku analizy składowych głównych czy też analizy czynnikowej, wybór wymiaru przestrzeni rzutowania (liczby osi układu współrzędnych) jest określany za pomocą miernika udziału danej osi w wyjaśnieniu całkowitej inercji oryginalnego układu.
- Ogólnie, wielkość

$$\frac{\lambda_k}{\sum\limits_{i=1}^{rz(\mathbf{P})} \lambda_i} \cdot 100\%,$$

gdzie $rz(\mathbf{P}) = \min(r-1, c-1)$ oznacza rząd macierzy korespondencji, wyraża procent inercji całkowitej wyjaśniony za pomocą k-tej osi głównej (k-tego wymiaru).

Wybór wymiaru przestrzeni rzutowania (1)

- Podobnie jak w przypadku analizy składowych głównych czy też analizy czynnikowej, wybór wymiaru przestrzeni rzutowania (liczby osi układu współrzędnych) jest określany za pomocą miernika udziału danej osi w wyjaśnieniu całkowitej inercji oryginalnego układu.
- Ogólnie, wielkość

$$\frac{\lambda_k}{\sum_{i=1}^{rz(\mathbf{P})} \lambda_i} \cdot 100\%,$$

gdzie $rz(\mathbf{P}) = \min(r-1, c-1)$ oznacza rząd macierzy korespondencji, wyraża procent inercji całkowitej wyjaśniony za pomocą k-tej osi głównej (k-tego wymiaru).
- Jeżeli zatem np. dwie pierwsze osie główne wyjaśniają 80% inercji całkowitej, możemy poprzestać na dwuwymiarowym opisie powiązań pomiędzy kategoriami (ze stratą informacji rzędu 20%).

Wybór wymiaru przestrzeni rzutowania (2)

- Innym sposobem określania wymiaru przestrzeni rzutowania jest analiza łamanej przedstawiającej kolejne wartości własne λ_k macierzy AA^T i A^TA. W odróżnieniu od PCA i FA, stosuje się tu nazwę **„kryterium łokcia”** (ang. *the elbow criterion*) zamiast analizy wykresu osypiska.

Wybór wymiaru przestrzeni rzutowania (2)

- Innym sposobem określania wymiaru przestrzeni rzutowania jest analiza łamanej przedstawiającej kolejne wartości własne λ_k macierzy AA^T i $A^T A$. W odróżnieniu od PCA i FA, stosuje się tu nazwę **„kryterium łokcia"** (ang. *the elbow criterion*) zamiast analizy wykresu osypiska.
- W ostatniej metodzie (zaproponował ją Greenacre w 1984 roku) jako optymalny wymiar przestrzeni rzutowania wybieramy minimalne k, dla którego spełniony jest warunek

$$\lambda_{Z,k} > \frac{1}{q},$$

gdzie $\lambda_{Z,k}$ jest k-tą największą wartością własną macierzy znaczników Z, a q oznacza liczbę analizowanych zmiennych (tu $q = 2$). Macierz znaczników Z jest tu rozumiana jako macierz, w której kolejne wiersze odpowiadają kolejnym elementom próby losowej (jest ich n), a kolumny - kolejnym kategoriom kolejnych zmiennych.

Interpretacja nowych współrzędnych (1)

- Po wyznaczeniu nowych współrzędnych można oszacować „korelację" punktu odpowiadającego danej kategorii oryginalnej cechy z którąś z osi przestrzeni rzutowania.

Interpretacja nowych współrzędnych (1)

- Po wyznaczeniu nowych współrzędnych można oszacować „korelację" punktu odpowiadającego danej kategorii oryginalnej cechy z którąś z osi przestrzeni rzutowania.
- Jeżeli $f_{i,k}$ jest współrzędną i-tego punktu na k-tej osi, d_i - odległością i-tego punktu od centrum rzutowania (środka nowego układu współrzędnych), natomiast $\alpha_{i,k}$ oznacza kąt pomiędzy promieniem wodzącym i-tego punktu a k-tą osią, to prawdziwy jest następujący wzór:

$$\cos^2 \alpha_{i,k} = \frac{f_{i,k}^2}{d_i^2}.$$

Interpretacja nowych współrzędnych (1)

- Po wyznaczeniu nowych współrzędnych można oszacować „korelację” punktu odpowiadającego danej kategorii oryginalnej cechy z którąś z osi przestrzeni rzutowania.
- Jeżeli $f_{i,k}$ jest współrzędną i-tego punktu na k-tej osi, d_i - odległością i-tego punktu od centrum rzutowania (środka nowego układu współrzędnych), natomiast $\alpha_{i,k}$ oznacza kąt pomiędzy promieniem wodzącym i-tego punktu a k-tą osią, to prawdziwy jest następujący wzór:

$$\cos^2 \alpha_{i,k} = \frac{f_{i,k}^2}{d_i^2}.$$

- Wielkość $\cos^2 \alpha_{i,k}$ wskazuje na stopień „wyjaśnienia” danej kategorii oryginalnej zmiennej (reprezentowanej przez i-ty punkt) przez k-tą oś główną przestrzeni rzutowania (można tę wielkość podawać również w ujęciu procentowym).

Interpretacja nowych współrzędnych (2)

Poniższe zasady są pomocne we właściwej interpretacji konkretnego położenia punktu (reprezentującego daną kategorię zmiennej) w przestrzeni rzutowania (w nowym układzie współrzędnych):

- Im bliżej centrum rzutowania położony jest dany punkt, tym bardziej jego profil jest zbliżony do profilu średniego (np. jeżeli kategoria X_2 cechy X jest położona najbliżej centrum rzutowania ze wszystkich punktów odpowiadających kategoriom cechy X, oznacza to, że kategoria ta jest najbardziej zbliżona do średniego profilu wierszowego).

Interpretacja nowych współrzędnych (2)

Poniższe zasady są pomocne we właściwej interpretacji konkretnego położenia punktu (reprezentującego daną kategorię zmiennej) w przestrzeni rzutowania (w nowym układzie współrzędnych):

- Im bliżej centrum rzutowania położony jest dany punkt, tym bardziej jego profil jest zbliżony do profilu średniego (np. jeżeli kategoria X_2 cechy X jest położona najbliżej centrum rzutowania ze wszystkich punktów odpowiadających kategoriom cechy X, oznacza to, że kategoria ta jest najbardziej zbliżona do średniego profilu wierszowego).
- Im dalej punkt jest położony od centrum rzutowania, tym większy „udział" tego punktu (czyli kategorii) w ewentualnym odrzuceniu hipotezy o niezależności cech.

Interpretacja nowych współrzędnych (3)

- W przypadku tej samej zmiennej: bliskie względem siebie położenie punktów (np. X_1 i X_2) świadczy o podobieństwie ich profili.

Interpretacja nowych współrzędnych (3)

- W przypadku tej samej zmiennej: bliskie względem siebie położenie punktów (np. X_1 i X_2) świadczy o podobieństwie ich profili.
- W przypadku różnych zmiennych: bliskie względem siebie położenie punktów (np. X_1 i Y_2) świadczy o istnieniu powiązań pomiędzy kategoriami tych zmiennych.

Efekt Guttmana (1)

- W przestrzeni rzutowania (szczególnie dwuwymiarowej) można stosunkowo często zaobserwować tzw. **efekt Guttmana,** nazywany także **efektem podkowy** (ang. *horseshoe effect*), kiedy punkty odpowiadające poszczególnym kategoriom (obydwu cech) układają się w dość wyraźny łuk.

Efekt Guttmana (1)

- W przestrzeni rzutowania (szczególnie dwuwymiarowej) można stosunkowo często zaobserwować tzw. **efekt Guttmana,** nazywany także **efektem podkowy** (ang. *horseshoe effect*), kiedy punkty odpowiadające poszczególnym kategoriom (obydwu cech) układają się w dość wyraźny łuk.
- Efekt ten można także zaobserwować, gdy po zmianie kolejności poszczególnych kategorii cech X i Y w tablicy kontyngencji na taką, która odpowiada kolejności pojawiania się punktów w przestrzeni rzutowania po zrzutowaniu ich na pierwszą oś główną, zaobserwowane liczebności grupują się wokół „pasa” wyznaczonego przez główną przekątną tablicy.

Efekt Guttmana (2)

- Efekt Guttmana świadczy o **dominacji pierwszej osi głównej**.

Efekt Guttmana (2)

- Efekt Guttmana świadczy o **dominacji pierwszej osi głównej**.
- Im wyraźniejszy ów łuk w przestrzeni rzutowania, tym udział pierwszej osi (współrzędnej) w wyjaśnieniu inercji całkowitej jest bliższy 100%. W konsekwencji analizę korespondencji można wówczas przeprowadzić jednowymiarowo.

Wielowymiarowa analiza korespondencji (1)

- **Wielowymiarowa analiza korespondencji** (ang. *Multiple Correspondence Analysis (MCA)*) jest przeprowadzana w sytuacji, w której obiekty opisane są za pomocą co najmniej trzech zmiennych o charakterze jakościowym.

- **Wielowymiarowa analiza korespondencji** (ang. *Multiple Correspondence Analysis (MCA)*) jest przeprowadzana w sytuacji, w której obiekty opisane są za pomocą co najmniej trzech zmiennych o charakterze jakościowym.
- Wprowadza się tu pojęcie tzw. **macierzy Burta**, zdefiniowanej w następujący sposób:

$$\mathbf{B} \stackrel{def}{=} Z^T Z,$$

gdzie Z jest macierzą znaczników (wiersze odpowiadają kolejnym elementom próby losowej, a kolumny wariantom kolejnych cech).

Wielowymiarowa analiza korespondencji (1)

- **Wielowymiarowa analiza korespondencji** (ang. *Multiple Correspondence Analysis (MCA)*) jest przeprowadzana w sytuacji, w której obiekty opisane są za pomocą co najmniej trzech zmiennych o charakterze jakościowym.
- Wprowadza się tu pojęcie tzw. **macierzy Burta**, zdefiniowanej w następujący sposób:

$$\mathbf{B} \stackrel{def}{=} Z^T Z,$$

gdzie Z jest macierzą znaczników (wiersze odpowiadają kolejnym elementom próby losowej, a kolumny wariantom kolejnych cech).

- Jeśli np. badane są trzy cechy: X, Y i U, to w macierzy znaczników każdy wiersz zawiera tylko trzy jedynki (na pozycjach odpowiadających kategoriom badanych cech, którymi charakteryzuje się konkretny element próby), a na pozostałych miejscach zera.

Wielowymiarowa analiza korespondencji (2)

- Łatwo sprawdzić, że macierz Burta utworzona dla trzech cech zawierać będzie dziewięć macierzy blokowych, z których każda będzie miała liczebność równą liczebności próby n.

Wielowymiarowa analiza korespondencji (2)

- Łatwo sprawdzić, że macierz Burta utworzona dla trzech cech zawierać będzie dziewięć macierzy blokowych, z których każda będzie miała liczebność równą liczebności próby n.
- Całkowita liczebność macierzy Burta wynosi zatem nq^2, gdzie q oznacza liczbę analizowanych zmiennych.

Wielowymiarowa analiza korespondencji (2)

- Łatwo sprawdzić, że macierz Burta utworzona dla trzech cech zawierać będzie dziewięć macierzy blokowych, z których każda będzie miała liczebność równą liczebności próby n.
- Całkowita liczebność macierzy Burta wynosi zatem nq^2, gdzie q oznacza liczbę analizowanych zmiennych.
- W wielowymiarowej analizie korespondencji macierz korespondencji jest zdefiniowana jako

$$\mathbf{P} = \frac{1}{nq^2}\mathbf{B},$$

zaś dekompozycji poddawana jest macierz

$$A = \mathbf{D_r}^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{r}^T)\mathbf{D_r}^{-1/2}.$$

Wprowadzenie do analizy skupień (1)

- **Grupowanie**, zwane inaczej **analizą skupień** lub **klasteryzacją** (ang. *data clustering*) stanowi zbiór technik i metod, których zasadniczym celem jest wyszukiwanie (przede wszystkim w dużych zbiorach danych), a następnie łączenie ze sobą w tzw. **skupienia** (zwane także **klastrami**) obiektów „podobnych".

Wprowadzenie do analizy skupień (1)

- **Grupowanie**, zwane inaczej **analizą skupień** lub **klasteryzacją** (ang. *data clustering*) stanowi zbiór technik i metod, których zasadniczym celem jest wyszukiwanie (przede wszystkim w dużych zbiorach danych), a następnie łączenie ze sobą w tzw. **skupienia** (zwane także **klastrami**) obiektów „podobnych".
- Utworzone w wyniku przeprowadzenia procesu grupowania skupienia powinny oczywiście być zbiorami rozłącznymi, ale także maksymalnie „oddzielonymi" od siebie, skupiającymi elementy mogące być podobnie scharakteryzowane.

Wprowadzenie do analizy skupień (1)

- **Grupowanie**, zwane inaczej **analizą skupień** lub **klasteryzacją** (ang. *data clustering*) stanowi zbiór technik i metod, których zasadniczym celem jest wyszukiwanie (przede wszystkim w dużych zbiorach danych), a następnie łączenie ze sobą w tzw. **skupienia** (zwane także **klastrami**) obiektów „podobnych”.
- Utworzone w wyniku przeprowadzenia procesu grupowania skupienia powinny oczywiście być zbiorami rozłącznymi, ale także maksymalnie „oddzielonymi” od siebie, skupiającymi elementy mogące być podobnie scharakteryzowane.
- Poszczególne obiekty podlegające procesowi klasteryzacji mogą być opisane zmiennymi różnego typu: ilościowymi, binarnymi, jakościowymi o skali porządkowej, jakościowymi opisanymi skalą nominalną.

Wprowadzenie do analizy skupień (2)

- Jako przykład może nam posłużyć baza danych o klientach pewnego operatora telekomunikacyjnego.

Wprowadzenie do analizy skupień (2)

- Jako przykład może nam posłużyć baza danych o klientach pewnego operatora telekomunikacyjnego.
- Poszczególne obiekty tej bazy (klienci) mogą być opisani np. za pomocą czterech zmiennych, zwanych **atrybutami**: X_1 - wiek, X_2 - status majątkowy, X_3 - liczba osób w gospodarstwie domowym, X_4 - miejsce zamieszkania.

Wprowadzenie do analizy skupień (2)

- Jako przykład może nam posłużyć baza danych o klientach pewnego operatora telekomunikacyjnego.
- Poszczególne obiekty tej bazy (klienci) mogą być opisani np. za pomocą czterech zmiennych, zwanych **atrybutami**: X_1 - wiek, X_2 - status majątkowy, X_3 - liczba osób w gospodarstwie domowym, X_4 - miejsce zamieszkania.
- Przykładowy obiekt może mieć zatem postać: (42, średni, 3, duże miasto), co oznacza osobę w wieku 42 lat, o średnim statusie majątkowym, której gospodarstwo domowe liczy łącznie 3 osoby, mieszkańca dużego miasta.

Wprowadzenie do analizy skupień (3)

- Załóżmy, że operator zamierza przygotować trzy różne oferty promocyjne „dedykowane” konkretnym klientom.

- Załóżmy, że operator zamierza przygotować trzy różne oferty promocyjne „dedykowane" konkretnym klientom.
- Jak podzielić całą bazę danych na trzy rozłączne podzbiory tak, by móc je następnie dobrze scharakteryzować i dostosować do nich konkretną ofertę?

Wprowadzenie do analizy skupień (3)

- Załóżmy, że operator zamierza przygotować trzy różne oferty promocyjne „dedykowane” konkretnym klientom.
- Jak podzielić całą bazę danych na trzy rozłączne podzbiory tak, by móc je następnie dobrze scharakteryzować i dostosować do nich konkretną ofertę?
- W odpowiedzi na tego typu pytania pomocna jest właśnie analiza skupień.

W analizie skupień punktem wyjścia jest tzw. **macierz niepodobieństwa** obiektów (zwana także **macierzą odległości**). Jeżeli przyjmiemy, że $X_1, ..., X_n$ są danymi obiektami, które zamierzamy pogrupować, ma ona następującą postać:

$$\mathbf{D} = \left[\begin{array}{cccc} 0 & D(X_1, X_2) & \dots & D(X_1, X_n) \\ D(X_2, X_1) & 0 & \dots & D(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ D(X_n, X_1) & D(X_n, X_2) & \dots & 0 \end{array} \right],$$

w której $D(X_i, X_j)$ oznacza miarę niepodobieństwa (odległość) obiektów X_i oraz X_j. Oczywiście, $\mathbf{D}$ jest macierzą symetryczną z zerami na głównej przekątnej.

Miary odległości obiektów (2)

Odległości $D(X_i, X_j)$ mogą być różnie definiowane. Jeżeli obiekty są opisane za pomocą atrybutów mierzalnych (ilościowych), ale nie binarnych, np. za pomocą p atrybutów, czyli w postaci $X_i = (X_{i,1}, ..., X_{i,p})$, $X_j = (X_{j,1}, ..., X_{j,p})$, wówczas stosuje się następujące miary:

- **odległość euklidesowa**: $D(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (X_{i,k} - X_{j,k})^2}$;

Miary odległości obiektów (2)

Odległości $D(X_i, X_j)$ mogą być różnie definiowane. Jeżeli obiekty są opisane za pomocą atrybutów mierzalnych (ilościowych), ale nie binarnych, np. za pomocą p atrybutów, czyli w postaci $X_i = (X_{i,1}, ..., X_{i,p})$, $X_j = (X_{j,1}, ..., X_{j,p})$, wówczas stosuje się następujące miary:

- **odległość euklidesowa**: $D(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (X_{i,k} - X_{j,k})^2}$;
- **kwadrat odległości euklidesowej**: jest wybierany jako miara odległości wówczas, gdy obiektom bardziej oddalonym od siebie chcemy przypisać większą wagę;

Miary odległości obiektów (2)

Odległości $D(X_i, X_j)$ mogą być różnie definiowane. Jeżeli obiekty są opisane za pomocą atrybutów mierzalnych (ilościowych), ale nie binarnych, np. za pomocą p atrybutów, czyli w postaci $X_i = (X_{i,1}, ..., X_{i,p})$, $X_j = (X_{j,1}, ..., X_{j,p})$, wówczas stosuje się następujące miary:

- **odległość euklidesowa**: $D(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (X_{i,k} - X_{j,k})^2}$;
- **kwadrat odległości euklidesowej**: jest wybierany jako miara odległości wówczas, gdy obiektom bardziej oddalonym od siebie chcemy przypisać większą wagę;
- **odległość Czebyszewa**: $D(X_i, X_j) = \max_{1 \leq k \leq p} |X_{i,k} - X_{j,k}|$;

Miary odległości obiektów (3)

- **odległość miejska (Manhattan)**:

$$D(X_i, X_j) = \sum_{k=1}^{p} |X_{i,k} - X_{j,k}|;$$

Miary odległości obiektów (3)

- **odległość miejska (Manhattan)**:
  $D(X_i, X_j) = \sum_{k=1}^{p} |X_{i,k} - X_{j,k}|;$
- **odległość Minkowskiego**:
  $D(X_i, X_j) = \left(\sum_{k=1}^{p} |X_{i,k} - X_{j,k}|^{\widehat{p}} \right)^{1/\widehat{p}}$, gdzie $\widehat{p} \geq 1$;

Miary odległości obiektów (3)

- **odległość miejska (Manhattan)**:
$D(X_i, X_j) = \sum_{k=1}^{p} |X_{i,k} - X_{j,k}|$;
- **odległość Minkowskiego**:
$D(X_i, X_j) = \left(\sum_{k=1}^{p} |X_{i,k} - X_{j,k}|^{\widehat{p}}\right)^{1/\widehat{p}}$, gdzie $\widehat{p} \geq 1$;
- **odległość Mahalanobisa**: definiowana jako

$$D(X_i, X_j) = \sqrt{\sum_{k=1}^{p} \sum_{l=1}^{p} (X_{i,k} - X_{j,k})(X_{i,l} - X_{j,l}) s_{i,j}},$$

gdzie $s_{i,j}$ jest kowariancją obiektu X_i oraz X_j (traktujemy je jak wektory).

Miary odległości obiektów (4)

- Zauważmy, że odległość euklidesowa oraz odległość miejska są szczególnymi przypadkami odległości Minkowskiego (odpowiednio, $\widehat{p} = 2$ oraz $\widehat{p} = 1$).

Miary odległości obiektów (4)

- Zauważmy, że odległość euklidesowa oraz odległość miejska są szczególnymi przypadkami odległości Minkowskiego (odpowiednio, $\widehat{p} = 2$ oraz $\widehat{p} = 1$).
- Odległość Mahalanobisa stosuje się w sytuacji dużego skorelowania obiektów X_i i X_j (traktowanych jako wektory współrzędnych).

- Zauważmy, że odległość euklidesowa oraz odległość miejska są szczególnymi przypadkami odległości Minkowskiego (odpowiednio, $\widehat{p} = 2$ oraz $\widehat{p} = 1$).
- Odległość Mahalanobisa stosuje się w sytuacji dużego skorelowania obiektów X_i i X_j (traktowanych jako wektory współrzędnych).
- W takiej sytuacji metryka euklidesowa może dawać wyniki mylące.

- W przypadku dużych różnic pomiędzy wartościami poszczególnych atrybutów dla różnych obiektów, powstaje **problem skalowalności** (np. obiekty $(1, 23)$ oraz $(1, 5023)$).

- W przypadku dużych różnic pomiędzy wartościami poszczególnych atrybutów dla różnych obiektów, powstaje **problem skalowalności** (np. obiekty $(1, 23)$ oraz $(1, 5023)$).
- Stosuje się wówczas **standaryzację** zmiennych (odejmujemy od wartości atrybutów ich średnią arytmetyczną i dzielimy przez odchylenie standardowe).

Miary odległości obiektów (5)

- W przypadku dużych różnic pomiędzy wartościami poszczególnych atrybutów dla różnych obiektów, powstaje **problem skalowalności** (np. obiekty $(1, 23)$ oraz $(1, 5023)$).
- Stosuje się wówczas **standaryzację** zmiennych (odejmujemy od wartości atrybutów ich średnią arytmetyczną i dzielimy przez odchylenie standardowe).
- Czasem stosuje się także tzw. **unitaryzację**. Przyjmujemy wówczas nowe wartości atrybutu w próbie, dzieląc wartości dotychczasowe przez ich rozstęp (różnicę pomiędzy wartością maksymalną a minimalną).

Miary odległości obiektów (6)

W przypadku **atrybutów binarnych** (przyjmujących wyłącznie wartości 0 lub 1) stosuje się następującą miarę odległości:

$$D(X_i, X_j) = \frac{\text{liczba atrybutów o różnych wartościach w obiektach } X_i \text{ i } X_j}{\text{liczba wszystkich atrybutów}}.$$

Powyższą miarę odległości stosuje się w przypadku atrybutów binarnych, które mają w populacji podobne wagi (np. płeć). Atrybuty takie nazywamy **symetrycznymi**.

W przypadku, gdy wagi są różne (**atrybuty asymetryczne**), stosuje się modyfikację powyższej miary, zdefiniowaną w następujący sposób:

$$D(X_i, X_j) = \frac{\text{liczba atrybutów o różnych wartościach w obiektach } X_i \text{ i } X_j}{\text{liczba atrybutów, które w choć jednym z obiektów mają wartość 1}}.$$

Miary odległości obiektów (8)

W sytuacji, gdy atrybuty dane są w skali nominalnej lub porządkowej, sposób postępowania opisano poniżej.

- Dla atrybutów danych w skali nominalnej stosuje się miarę postaci

$$D(X_i, X_j) = \frac{\text{liczba atrybutów o różnych kategoriach}}{\text{liczba wszystkich atrybutów}}.$$

W sytuacji, gdy atrybuty dane są w skali nominalnej lub porządkowej, sposób postępowania opisano poniżej.

- Dla atrybutów danych w skali nominalnej stosuje się miarę postaci

$$D(X_i, X_j) = \frac{\text{liczba atrybutów o różnych kategoriach}}{\text{liczba wszystkich atrybutów}}.$$

- Dla atrybutów danych w skali porządkowej przyporządkowuje się im wartości równe $\frac{i-1}{M-1}$, gdzie $i = 1, ..., M$, przy czym M oznacza liczbę różnych kategorii danego atrybutu. Następnie stosuje się jedną ze zwykłych miar stosowanych w przypadku atrybutów o charakterze ilościowym.

Zagregowana miara niepodobieństwa (1)

- Jeżeli atrybuty obiektów należą do różnych typów, wówczas do oceny odległości tych obiektów stosuje się tzw. **zagregowaną miarę niepodobieństwa (odległości)**, zdefiniowaną jako

$$D(X_i, X_j) = \sum_{k=1}^{p} w_k d_k(X_i, X_j), \quad \sum_{k=1}^{p} w_k = 1,$$

gdzie p jest liczbą atrybutów, za pomocą których opisywane są obiekty, $d_k(X_i, X_j)$ jest odległością k-tego atrybutu dla obiektów X_i i X_j, natomiast w_k jest **wagą** k-tego atrybutu.

- Jeżeli atrybuty obiektów należą do różnych typów, wówczas do oceny odległości tych obiektów stosuje się tzw. **zagregowaną miarę niepodobieństwa (odległości)**, zdefiniowaną jako

$$D(X_i, X_j) = \sum_{k=1}^{p} w_k d_k(X_i, X_j), \quad \sum_{k=1}^{p} w_k = 1,$$

  gdzie p jest liczbą atrybutów, za pomocą których opisywane są obiekty, $d_k(X_i, X_j)$ jest odległością k-tego atrybutu dla obiektów X_i i X_j, natomiast w_k jest **wagą** k-tego atrybutu.

- Dodatkowo definiuje się także **średnią wartość niepodobieństwa obiektów** z całej badanej zbiorowości (lub próby) n-elementowej jako

$$\overline{D} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} D(X_i, X_j).$$

Zagregowana miara niepodobieństwa (2)

- Wielkość $\overline{D}$ można wyrazić także w następujący sposób:

$$\overline{D} = \sum_{k=1}^{p} w_k \overline{d}_k,$$

gdzie $\overline{d}_k = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_k(X_i, X_j)$.

Zagregowana miara niepodobieństwa (2)

- Wielkość $\overline{D}$ można wyrazić także w następujący sposób:

$$\overline{D} = \sum_{k=1}^{p} w_k \overline{d}_k,$$

gdzie $\overline{d}_k = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_k(X_i, X_j)$.

- **Względny wpływ k-tego atrybutu na średnią wartość niepodobieństwa** wszystkich par obiektów z badanej grupy wynosi $w_k \overline{d}_k$.

Zagregowana miara niepodobieństwa (2)

- Wielkość $\overline{D}$ można wyrazić także w następujący sposób:

$$\overline{D} = \sum_{k=1}^{p} w_k \overline{d}_k,$$

gdzie $\overline{d}_k = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_k(X_i, X_j)$.

- **Względny wpływ k-tego atrybutu na średnią wartość niepodobieństwa** wszystkich par obiektów z badanej grupy wynosi $w_k \overline{d}_k$.
- Jeżeli chcemy, by wszystkie atrybuty miały podobny wpływ na średnią wartość niepodobieństwa, należy przyjąć w_k proporcjonalne do wartości $\frac{1}{\overline{d}_k}$ i do procentowego udziału atrybutu w wartości zagregowanej miary niepodobieństwa.