POLITECHNIKA ŁÓDZKA

WYDZIAŁ FIZYKI TECHNICZNEJ, INFORMATYKI I MATEMATYKI STOSOWANEJ

Kierunek: Matematyka
Specjalność: Matematyczne Metody Analizy Danych Biznesowych

WYBRANE ZASTOSOWANIE STATYSTYCZNYCH METOD PORZĄDKOWANIA DANYCH WIELOWYMIAROWYCH

Kamila Choja
Nr albumu: 204052

Praca licencjacka
napisana w Instytucie Matematyki Politechniki Łódzkiej

Promotor: dr, mgr inż. Piotr Kowalski

ŁÓDŹ, xxx 2018

Spis treści

Rozdział 1

Wstęp

Wielowymiarowa analiza danych jest istotnym pojęciem we współczesnej analizie. Jedną z jej części są metody porządkowania. Aby móc w pełni zrozumieć to zagadnienie, należy przyjrzeć się podłożu matematycznemu tego tematu. Dane które chcemy poddać analizie, są niczym innym jak próbą losową, składającą są z niezależnych zmiennych losowych. W związku z tym w sekcji 2.3 zostały przytoczone podstawowe pojęcia rachunku prawdopodobieństwa oraz statystyki. Dzięki temu w dalszej części pracy łatwiej zrozumieć zbiór danych jako zbiór obiektów, opisywanych przez zmienne. Mając opracowane te zagadnienia, możemy przejść do matematycznej definicji porządku jako relacji. Dzięki temu relację tę możemy podzielić na porządek liniowy - czyli tę relację, która umożliwia nam porównywanie każdych dwóch obiektów między sobą, a także której wynik może być przedstawiony jako linia, oraz na porządek częściowy, którego to wynik jest reprezentowany przez graf. Ze względu na powyższe zostały opracowane sekcje 2.6 oraz 2.5, w których umieściliśmy wszystkie niezbędne definicje z zakresu teorii matematycznego porządku oraz grafów, z których korzystamy w pracy.

Podstawowym zadaniem analizy danych, które zostało postawione w tej pracy jest odkrywanie porządków z wykorzystaniem metod statystycznych. Problematyce tej został poświęcony rozdział 3, który to opracowano na podstawie [9]. Na potrzeby pracy użyliśmy przyjętych statystycznych nazw, podziału relacji porządku tj. metody porządkowania liniowego oraz metody porządkowania nieliniowego, będących odpowiednikiem relacji porządku częściowego. Wobec powyższego, na początku rozdziału 3 omówiono własności porządkowania liniowego, które to w dalszej kolejności zostały sformalizowane, a następnie samodzielnie udowodnione. Kolejny rozdział pracy - rozdział 4 stanowi praktyczną część. Omówiliśmy tam stworzony zbiór danych, na którym to w dalszej części pracy przetestowano stworzone implementacje wybranych metod porządkowania. Algorytmy te zostały napisane w języku R. Na koniec wyniki uporządkowań zostały poddane analizie.

Rozdział 2

Preliminaria

2.1 Notacja

Poniżej znajduje się lista pojęć powszechnie używanych w pracy wraz z symbolami, które im się przypisuje.

- $\mathbb{R}$ - zbiór liczb rzeczywistych
- $\mathbb{N}$ - zbiór liczb naturalnych
- K - oznaczenie dowolnego ciała zbioru
- $O = \{O_1, O_2, \ldots, O_n\}$ - zbiór obiektów przestrzennych, tj. opisywanych przez wiele atrybutów, $n \in \mathbb{N}$
- $X = [x_{ij}]$ - macierz surowych danych, gdzie x_{ij}-oznacza wartość j-tej zmiennej dla i-tego obiektu, gdzie $i = 1, \ldots, n$, $j = 1, \ldots, m,$, $n, m \in \mathbb{N}$. W rozdziale drugim, przez X najczęściej będziemy oznaczać dowolny zbiór.
- $N = [n_{ij}]$ - macierz znormalizowanych danych, gdzie n_{ij} oznacza wartość j-tej cechy i-tego obiektu
- x^S - oznaczenie zmiennej mającej charakter stymulanty
- x^D - oznaczenie zmiennej mającej charakter destymulanty
- x^N - oznaczenie zmiennej mającej charakter nominanty
- $D = [d_{ik}]$ - oznaczenie macierzy odległości, gdzie d_{ik} oznacza odległość między i-tym i k-tym obiektem
- s_i - oznaczenie zmiennej syntetycznej i-tego obiektu
- $P_0 = [n_{0j}]$ - oznaczenie obiektu wzorcowego, gdzie n_{0j}- znormalizowana j-ta współrzędna obiektu wzorcowego
- Y - w rozdziale drugim używana jest najczęściej do oznaczenie zmiennej losowej
- Ω - oznaczenie dowolnej przestrzeni zdarzeń elementarnych ω, z rodziną podzbiorów $\mathcal{F}$
- $\mathcal{B}$ - rodzina zbiorów borelowskich
- $\mathfrak{B}$ - rodzina wszystkich zbiorów otwartych

- $\leqslant$ - oznaczenie relacji częściowego porządku, przy dołożeniu warunku spójności oznaczać będzie relację liniowego porządku
- $\overline{\overline{\{\cdot\}}}$ - oznaczenie mocy zbioru
- G - oznaczenie ogólnego grafu prostego dla którego $V(G)$ jest zbiorem wierzchołków grafu, a $E(G)$ zbiorem jego krawędzi

2.2 Słownik użytych pojęć

W pracy zostały wykorzystane następujące pojęcia.

- Statystyka matematyczna [6, w oparciu o rozdział 1] Statystyka matematyczna jest nauką zajmującą się opisywaniem i analizą zjawisk przy użyciu metod rachunku prawdopodobieństwa.
- Cecha statystyczna [6, Rozdział 1]

  Cecha statystyczna jest to właściwość wspólna dla danego zbioru obserwacji. Jej wartości pozwalają rozróżnić elementy zbioru między sobą. cechy statystyczne można podzielić na te mierzalne, tj. ilościowe (np. długość, ciężar), oraz niemierzalne tj. jakościowe(np. kolor, płeć, zawód, województwo).

W celu prezentacji dużych ilości danych, w analizie danych korzysta się z pojęcia macierzy. Poniżej zostanie przedstawiona formalna definicja macierzy, oraz definicja macierzy obserwacji, czyli zbiorowi obiektów opisywanych przez zmienne.

Definicja 2.2.1. *Macierz [1, Rozdział 1] Niech K będzie ciałem i $m, n \in \mathbb{N}$. Macierzą o m-wierszach, n-kolumnach i o wyrazach z K nazywamy każdą funkcję postaci $A : \{1, \ldots, m\} \times \{1, \ldots, n\} \to K$*

Przykład 1. *Macierz A o m-wierszach i n-kolumnach najczęściej zapisuje się postaci $A = [a_{ij}]_{i \leqslant m, j \leqslant n}$, tj.*

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

Uwaga 1. *W statystyce koncepcja matematyczna macierzy jest rozszerzana, gdyż niektóre kolumny mają wartości z poza ciała zbioru liczb $\mathbb{R}$ (mogą być np. tekstem).*

Definicja 2.2.2. *Macierz obserwacji [8, Rozdział 2] Niech $m > 1$ oraz $n > 1$ będą liczbami naturalnymi. Macierzą obserwacji nazywamy macierz rozmiaru $n \times m$ postaci*

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}$$

gdzie: x_{ij} - zaobserwowana wartość j-tej cechy dla i-tego obiektu.

Definicja 2.2.3. *Macierz odległości zmiennych [9, Rozdział 1.6] Macierzą odległości cech zmiennych nazywamy macierz, której elementami są odległości między parami badanych obiektów:*

$$D = [d_{ik}].$$

gdzie:
d_{ik}*-odległość między i-tym a k-tym obiektem, dla* $i, k = 1, 2, \ldots, n$

Uwaga 2. *Zauważmy, że macierz obserwacji nie musi być symetryczna, z kolei w przypadku macierzy odległości jest to wymagane.*

W statystyce posługujemy się pojęciami skal do opisu różnych typów danych, które przyjmowane przez nas mogą podlegać analizie. I tak zdefiniujmy następujące rodzaje skal:

- Skala porządkowa [9, Rozdział 1.2] Zmienna opisana jest na skali porządkowej jeśli jej zbiór wartości jest zbiorem, w którym wprowadzony jest porządek np. klasyfikacja zawodników na podium. Nie zawsze porządek ten jest ustalony w sposób matematyczny. W przypadku rozpatrywania zmiennych jakościowych porządek ustalamy na podstawie opinii ekspertów lub ogólnie przyjętych poglądów np. poziom wykształcenia, oceny w systemie szkolnym.
- Skala przedziałowa [9, Rozdział 1.2] Zmienna jest opisana na skali przedziałowej gdy podobnie tak jak na skali porządkowej jej zbiór wartości jest zbiorem, w którym wprowadzony jest porządek, ale dodatkowo możliwe jest obliczenie odległości między dwoma zmiennymi. Dodatkowo na skali tej możliwe jest wyznaczenie umownego punktu - punktu zerowego. Przykład zmiennych przedstawianych na skali przedziałowej: temperatura, czas.
- Skala ilorazowa [9, Rozdział 1.2] Zmienna jest opisana na skali ilorazowej, jeśli jej zbiór wartości jest zbiorem postaci $[0, \infty)$ będący podzbiorem zbioru liczb $\mathbb{R}$ lub też taki zbiór wartości który można utożsamić z podzbiorem liczb $\mathbb{R}$. Przykłady zmiennych opisywanych na skali ilorazowej: napięcie elektryczne, bezrobocie.

Uwaga 3. *Jeżeli zmiennej opisującej dany obiekt nie da się odnieść do żadnej z powyższych skal, to zmienna ta nazywana jest nominalną.*

Ze względu na to, że zmienne opisujące obiekty mogą mieć różny charakter, poniżej zostały wprowadzone definicje trzech różnych typów zmiennych.

Definicja 2.2.4. *Stymulanta [9, Rozdział 1.5] Stymulantami nazywane są te zmienne(cechy), dla których pożądane są wysokie wysokie wartości w badanych obiektach, ze względu na rozpatrywane zjawisko.*

Definicja 2.2.5. *Destymulanta [9, Rozdział 1.5] Destymulantami nazywane są te zmienne, dla których niepożądane są wysokie wysokie wartości w badanych obiektach, ze względu na rozpatrywane zjawisko.*

Definicja 2.2.6. *Nominanta [9, Rozdział 1.5] Nominantami nazywane są te zmienne, które mają określoną najkorzystniejszą wartość. Odchylenia od tej wartości są niepożądane, ze względu na rozpatrywane zjawisko.*

2.3 Podstawowe pojęcia rachunku prawdopodobieństwa oraz statystyki

Na potrzeby pracy, zostały wykorzystane pojęcia rachunku prawdopodobieństwa oraz statystyki, konieczne do zrozumienia danych jako próby losowej. W tym celu niezbędne było wprowadzenie definicji prawdopodobieństwa, zmiennej losowej, a także pojęć powiązanych z tymi definicjami tj. ciała zbiorów, σ-ciała zbiorów, przestrzeni zdarzeń elementarnych, zdarzenia losowego.

Definicja 2.3.1. *Ciało zbiorów [10, Rozdział 8.1] Rodzinę $\mathcal{F}$ podzbiorów, niepustego zbioru Y nazywamy ciałem zbiorów, jeżeli spełnia ona następujące warunku:*

1. $\emptyset \in \mathcal{F}$,
2. *jeżeli* $A \in \mathcal{F}$, *to* $Y \setminus A \in \mathcal{F}$,
3. *jeżeli* $A \in \mathcal{F}$, *to* $A \cup B \in \mathcal{F}$.

Definicja 2.3.2. *σ-algebra/ciało zbiorów[10, Rozdział 8.1] Ciało zbiorów $\mathcal{F}$ nazywamy σ-ciałem zbiorów, jeżeli spełnia ona warunek dla dowolnych zbiorów $A_n \in \mathcal{F}, n \in \mathbb{N}$, mamy $\bigcup_{i=1}^{\infty} A_n \in \mathcal{F}$.*

Najważniejszym σ-ciałem zbiorów w matematyce są σ-ciała zbiorów Borelowskich, dlatego też wprowadzimy definicję zbiorów borelowskich.

Definicja 2.3.3. *Zbiory borelowskie [4, w opraciu o rozdział 2] Zbiorami borelowskimi względem danej przestrzeni Y, nazywamy zbiory należące do σ-ciała Y generowanego przez rodzinę $\mathfrak{B}(Y)$ - wszystkich zbiorów otwartych w Y. Rodzinę wszystkich zbiorów borelowskich względem Y, oznaczamy $\mathcal{B}(Y)$.*

σ-ciała zbiorów często mogą być po prostu utożsamiane ze zbiorami, które można zmierzyć, w związku z tym wprowadzimy definicję miary zbioru.

Definicja 2.3.4. *Miara zbioru [4, Rozdział 2.10] Funkcję μ określoną na ciele $\mathcal{F}$ podzbiorów zbioru Ω nazywamy miarą, jeśli spełnia następujące warunki:*

1. $\mu(A) \in [0, \infty]$ *dla każdego zbioru* $A \in \mathcal{F}$,
2. $\mu(\emptyset) = 0$,
3. *jeśli* $A_1, A_2, ...$ *jest ciągiem rozłącznych zbiorów $\mathcal{F}$-mierzalnych takich, że* $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$, *to*

$$\mu\Big(\bigcup_{k=1}^{\infty} A_k\Big) = \sum_{k=1}^{\infty} \mu(A_k)$$

Definicja 2.3.5. *Przestrzeń mierzalna [4, Rozdział 2.10] Przestrzenią mierzalną nazywamy parę $(Y, \mathcal{F})$, gdzie $\mathcal{F}$ jest σ-ciałem podzbiorów zbioru Y*

Definicja 2.3.6. *Funkcja mierzalna [10, w oparciu o rozdział 8.2] Niech Y będzie niepustym zbiorem, $\mathcal{F}$ σ-ciałem na Y i $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Funkcję $f : Y \to \overline{\mathbb{R}}$ nazywamy mierzalną, jeżeli zbiór $\{y \in Y : f(y) > a\}$*
jest mierzalny przy dowolnym $a \in \mathbb{R}$.

Mając powyższe definicje, wprowadźmy możemy wprowadzić poszukiwane definicje rachunku prawdopodobieństwa.

Definicja 2.3.7. *Przestrzeń zdarzeń elementarnych [6, w oparciu o rozdział 1.1] Zbiór wszyskich możliwych wyników doświadczenia losowego nazywamy przestrzenią zdarzeń elementarnych i oznaczamy przez Ω. Elementy zbioru Ω nazywamy zdarzeniami elementarnymi i oznaczamy ω.*

Definicja 2.3.8. *Zdarzenie losowe [6, w oparciu o rozdział 1.1] Zdarzeniem losowym (zdarzeniem) nazywamy każdy podzbiór A zbioru Ω, taki że $A \in \mathcal{F}$, gdzie $\mathcal{F}$ jest rodziną podzbiorów Ω spełniającą następujące warunki:*

1. $\Omega \in \mathcal{F}$;

2. *Jeśli* $A \in \mathcal{F}$, *to* $A' \in \mathcal{F}$, *gdzie* $A' = \Omega \setminus A$ *jest zdarzeniem przeciwnym do zdarzenia* A;

3. *Jeśli* $A_i \in \mathcal{F}$, $i = 1, 2, ...,$*to* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Rodzinę $\mathcal{F}$ *spełniającą warunki 1 - 3 nazywamy* σ*-ciałem podzbiorów zbioru* Ω

Definicja 2.3.9. *Prawdopodobieństwo [6, w oparciu o rozdział 1.1] Prawdopodobieństwem nazywamy dowolną funkcję* P *o wartościach rzeczywistych, określoną na* σ*-ciele zdarzeń* $\mathcal{F} \subset 2^{\Omega}$, *spełniającą warunki:*

1. $\forall A \in \mathcal{F} P(A) \geqslant 0$

2. $P(\Omega) = 1$

3. *Jeśli* $A_i \in \mathcal{F}$, $i = 1, 2, ...$ *oraz* $A_i \cap A_j$ *dla* $i \neq j$, *to*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i)$$

Definicja 2.3.10. *Przestrzeń probabilistyczna [6, w oparciu o rozdział 1.2] Przestrzenią probabilistyczną nazywamy uporządkowaną trójkę* $(\Omega, \mathcal{F}, P)$, *gdzie* Ω *jest zbiorem zdarzeń elementarnych,* $\mathcal{F}$ *jest* σ*-ciałem podzbiorów* Ω, *zaś* P *jest prawdopodobieństwem określonym na* $\mathcal{F}$.

Dla tak podanej definicji prawdopodobieństwa, definiujemy:

Definicja 2.3.11. *Zmienna losowa [6, Rozdział 2.1] Niech* $(\Omega, \mathcal{F}, P)$ *będzie dowolną przestrzenią probabilistyczną. Dowolną funkcję* $Y : \Omega \to \mathbb{R}$ *nazywamy zmienną losową jednowymiarową, jeśli dla dowolnej liczby rzeczywistej* y *zbiór zdarzeń elementarnych* ω, *dla których spełniona jest nierówność* $Y(\omega) < y$ *jest zdarzeniem, czyli*
$\{\omega : Y(\omega) < y\} \in \mathcal{F}$ *dla każdego* $y \in \mathbb{R}$

Definicja 2.3.12. *Wektor losowy [5, Rozdział 5.1] Wektorem losowym nazywamy odwzorowanie* $Y : \Omega \to \mathbb{R}^n$, *spełniające następujący warunek: dla każdego układu liczb* $t_1, t_2, \ldots, t_n \in \mathbb{R}^n$ *zbiór* $Y^{-1}((-\infty, t_1] \times \ldots \times (-\infty, t_n])$ *należy do* $\mathcal{F}$.

Definicja 2.3.13. *Rozkład prawdopodobieństwa zmiennej losowej* Y *[5, Rozdział 5.1] Rozkładem prawdopodobieństwa zmiennej losowej* Y *o wartościach w* $\mathbb{R}$ *nazywamy funkcję* μ_Y *określoną na* $\mathcal{B}(\mathbb{R})$ *zależnością*
$\mu_Y(B) = P_Y(B) = P(Y^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R})$

Definicja 2.3.14. *Rozkład dyskretny [5, Rozdział 5.1] Mówimy, że zmienna losowa jednowymiarowa* Y *ma rozkład dyskretny, jeśli istnieje przeliczany zbiór* $S \subset \mathbb{R}$, *taki że* $\mu_Y(S) = 1$.

Definicja 2.3.15. *Gęstość i rozkład ciągły [5, Rozdział 5.1] Jeśli* μ *jest rozkładem prawdopodobieństwa na* $\mathbb{R}$ *i istnieje całkowalna funkcja* $f : \mathbb{R} \to \mathbb{R}$ *taka, że*

$$\mu(A) = \int_A f(y)dy, A \in \mathcal{B}(\mathbb{R})$$

to funkcję f *nazywamy gęstością rozkładu* μ. *Rozkład który ma gęstość, nazywamy rozkładem ciągłym.*

Definicja 2.3.16. *Wartość oczekiwana [6, Rozdział 2.6] Niech X będzie zmienną losową typu dyskretnego lub ciągłego. Wartością oczekiwaną zmiennej losowej Y nazywamy*

$$\mathrm{E}(Y) = \begin{cases} \sum_{i=1}^{n} y_i p_i & \text{, jeśli zmienna ma rozkład dyskretny i przyjmuje dokładnie } n \text{ wartości} \\ \sum_{i=1}^{\infty} y_i p_i & \text{, jeśli zmienna przyjmuje nieskończenie ale przeliczalnie wiele wartości} \\ \int_{-\infty}^{\infty} y f(y) dy & \text{, jeśli zmienna ma rozkład ciągły} \end{cases}$$

Definicja 2.3.17. *Wariancja [5, Rozdział 5.6] Niech Y będzie zmienną losową. Jeśli* $\mathrm{E}(Y - \mathrm{E}Y)^2 < \infty$*, to liczbę tę nazywamy wariancją zmiennej losowej Y o o wartościach rzeczywistych i oznaczamy:*

$\operatorname{Var} Y = \mathcal{D}^2 Y = \mathrm{E}(Y - \mathrm{E}Y)^2$.

Definicja 2.3.18. *Odchylenie standardowe [5, Rozdział 5.6] Niech Y będzie zmienną losową. Odchyleniem standardowym zmiennej losowej Y nazywamy pierwiastek z wariancji.*

$\sigma_Y = \mathcal{D}(Y) = \sqrt{\mathcal{D}^2 Y}$

Definicja 2.3.19. *Próba losowa n-elementowa [2, Rozdział 2] Niech $(\Omega, \mathcal{F}, P)$ będzie przestrzenią probabilistyczną, zaś Y będzie wektorem losowym. Próbą losowa n-elementową nazywamy n-elementowy ciąg niezależnych zmiennych losowych o jednakowym prawdopodobieństwie, tzn. ciąg postaci $Y = (Y_1, Y_2, \ldots, Y_n)$.*

Uwaga 4. *Średnią z próby oznaczamy $\overline{Y}$ i wyliczamy za pomocą wzoru: $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.*

Definicja 2.3.20. *Rozkład normalny (Gaussa) [5, Rozdział 5.10] Jeśli zmienna losowa Y ma gęstość postaci*

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\mu_Y)^2}{2\sigma^2}}$$

dla $y \in \mathbb{R}$ i pewnych $\mu_Y \in \mathbb{R}$ i $\sigma^2 > 0$. To mówimy, że zmienna losowa ma rozkład normalny z parametrami μ i σ^2, co zapisujemy $\mathcal{N}(\mu, \sigma^2)$.

W przypadku, gdy $\mu = 0$ i $\sigma^2 = 1$, to rozkład ten nazywamy standardowym rozkładem normalnym i oznaczamy $\mathcal{N}(0, 1)$, a gęstość jest postaci

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(y)^2}{2}}.$$

2.4 Wybrane operacje statystyczne dla zmiennych

Normalizacja zmiennych

Bardzo ważnym krokiem przed rozpoczęciem pracy na zbiorze danych jest ujednolicenie ich charakteru, tj. przekształcenie zmiennych (mierzonych na skali przedziałowej lub ilorazowej) opisujących obiekty w zbiorze, w celu pozbycia się dysproporcji między nimi czy też dominacji jednych zmiennych nad drugimi. W tym celu stosuje się transformację normalizacyjną. Można wyróżnić trzy podstawowe typy przekształceń normalizacyjnych:

- standaryzacja,
- unitaryzacja,
- przekształcenie ilorazowe,

- rangowanie zmiennych.

W dalszej części pracy j-ta zmienna znormalizowana, i-tego obiektu jest oznaczona jako n_{ij}. Dodatkowo przyjmujemy oznaczenia: $\overline{x}$ - średnia z próby, $\sigma(x)$ - odchylenie standardowe z próby.

1. W wyniku standaryzacji zmienne uzyskują odchylenie standardowe równe 1 i wartości oczekiwanej równą 0. W tym celu dla każdej zmiennej, będącej cechą obiektu oblicza się odchylenie standardowe na podstawie wartości tej zmiennej dla wszystkich obiektów, a także wartość oczekiwaną na tej samej zasadzie. W kolejnym kroku dla każdego obiektu liczymy jego znormalizowana wartość tj. od wartości zmiennej odejmujemy średnią wartość danej cechy, a otrzymaną różnicę dzielimy przez odchylenie standardowe dla tej cechy, co można to zapisać w postaci

$$n_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma(x_j)}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

2. W pracy korzystamy również z unitaryzacji, stosowanej w celu uzyskania zmiennych o ujednoliconym zakresie zmienności, najczęściej jest to $[0, 1]$. W tym celu od wartości zmiennej w obiekcie, odejmowana jest minimalna wartość występująca dla tej cechy, a następnie różnica ta dzielona jest przez różnicę między maksymalną a minimalną wartością zmiennej, która została poddana normalizacji. Znormalizowana zmienna jest postaci

$$n_{ij} = \frac{x_{ij} - \min\limits_i\{x_{ij}\}}{\max\limits_i\{x_{ij}\} - \min\limits_i\{x_{ij}\}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

3. Kolejna metoda normalizacyjna, wykorzystana w pracy to przekształcenie ilorazowe. Stosuje się je aby odnieść wartości zmiennej do ustalonej wartości - może to być wartość oczekiwana danej zmiennej na tle analizowanych obiektów, wartość minimalna lub maksymalna tej cechy. W pracy za tę wartość przyjęliśmy średnią wartość zmiennej. Każda wartość zmiennej dla danego obiektu jest dzielona przez wartość oczekiwaną tej zmiennej, a postać znormalizowanej zmiennej to

$$n_{ij} = \frac{x_{ij}}{\overline{x_j}}, \quad \overline{x_j} \neq 0 \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

4. Przy zastosowaniu metody rang, wykorzystuje się normalizację rangową. Przekształcenie to, najczęściej stosowane jest, gdy zmienne opisujące obiekty są wyrażone na skali porządkowej. W pierwszym kroku wartości zmiennych opisujących obiekty zostają uporządkowane ze względu na ich wartości po procesie normalizacji. W kolejnym kroku wartościom zmiennej przyporządkowywane są rangi - czyli wartości liczbowe, będące najczęściej numerami miejsc zajmowanych przez obiekty w uporządkowanym zbiorze. Postać zmiennej znormalizowanej rangowo

$$n_{ij} = r, \quad dla \quad x_{hj} = x_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

   gdzie r-ranga nadana i-temu obiektowi znajdującemu się na r-tym miejscu w uporządkowanym zbiorze, ze względu na wartość j-tej zmiennej.

Stymulacja zmiennych

W celu ujednolicenia charakteru zmiennych należy poddać je pewnym przekształceniom, polegającym na zamianie destymulant i nominant na stymulanty. Tego typu transformacje nazywamy

stymulacjom. Można wyróżnić dwie najczęściej stosowane metody, tj. przekształcenie ilorazowe oraz przekształcenie różnicowe. W zależności od skali na której mierzone są zmienne, należy stosować odpowiednie przekształcenie stymulacyjne.

W przypadku przekształcenia ilorazowego można go stosować tylko dla zmiennych mierzonych na skali ilorazowej. Stymulacja destymulant wygląda następująco:

$$x_{ij}^S = [x_{ij}^D]^{-1}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

dla nominant:

$$x_{ij}^S = \frac{\min\{x_j^N, x_{ij}^N\}}{\max\{x_j^N, x_{ij}^N\}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

gdzie:

x_j^N - oznacza pożądaną wartość j-tej zmiennej

x_{ij}^N - oznacza wartość j-tej zmiennej w i-tym obiekcie

Dla zmiennych mierzonych na skali ilorazowej czy też przedziałowej stosuje się przekształcenie różnicowe. Stymulacja dla destymulant prezentuje się następująco

$$x_{ij}^S = \max_i\{x_{ij}^D\} - x_{ij}^D, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m,$$

a dla nominant

$$x_{ij}^S = -|x_{ij}^N - x_j^N|, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

2.5 Podstawowe pojęcia teorii grafów

W pracy zostaną opisane zarówno metody porządkowania liniowego jak i nieliniowego. W tym celu należy wprowadzić definicje związane z teorią grafów, niezbędne przy opisywaniu metod porządkowania nieliniowego.

W celu wprowadzenia kluczowych definicji, należy wcześniej podać podstawowe pojęcia dotyczące grafów. Zacznijmy od wprowadzenia definicji pary uporządkowanej oraz nieuporządkowanej, gdyż pojęcia te zostały wykorzystane w definicji grafu.

Definicja 2.5.1. *Para uporządkowana [7, w oparciu o rozdział 3] Niech dane będą dwa elementy a i b. Parą uporządkowaną nazywamy parę postaci $< a, b >$, gdzie element a jest poprzednikiem, zaś element b jest następnikiem.*

$< a, b >= \{\{a\}, \{a, b\}\}$.

Definicja 2.5.2. *Para nieuporządkowana [7, w oparciu o rozdział 3] Niech dane będą dwa elementy a i b. Parą nieuporządkowaną nazywamy zbiór postaci $\{a, b\}$, tj. zawierający elementy a i b i nie zawierający żadnego innego elementu. W przypadku, gdy $a = b$, to para nieuporządkowana $\{a, b\}$, składa się dokładnie z jednego elementu.*

W analogiczny sposób jak parę dwóch punktów można wprowadzić parę dwóch zbiorów. Istotne jest aby podkreślić różnicę pomiędzy parą dwóch wierzchołków, które utworzą krawędź, a parą zbiorów definiującą graf.

Definicja 2.5.3. *Graf [12, w oparciu o rozdział 2] Grafem nazywamy parę $G = (V, E) = (V(G), E(G))$, gdzie V jest niepustym, skończonym zbiorem wierzchołków grafu G, zaś E jest skończonym podzbiorem zbioru nieuporządkowanych par elementów zbioru V.*

Definicja 2.5.4. *Pętle [12, Rozdział 2] Pętlami w grafie nazywamy krawędzie reprezentowane przez $\{a\}$ gdzie a jest pewnym wierzchołkiem, tj. łączące wierzchołek z samym sobą. Innymi słowy jest to para nieuporządkowana składająca się z jednego elementu.*

Definicja 2.5.5. *Wierzchołki sąsiednie [12, Rozdział 2] Mówimy, że dwa wierzchołki v i w grafu G są sąsiednie, jeśli istnieje krawędź vw która je łączy.*

$$v \text{ ——— } w$$

Analogicznie definiuje się krawędzie sąsiednie.

Definicja 2.5.6. *Krawędzie sąsiednie [12, Rozdział 2] Dwie krawędzie e i f grafu G są sąsiednie, jeśli mają wspólny wierzchołek, tj*

$$\forall e, f \in E(G) \quad \exists d \in V(G) \quad d \in e \wedge d \in f$$

$$\text{——}\overset{e}{\text{——}}d\text{——}\overset{f}{\text{——}}$$

Aby dowiedzieć się o połączeniu dwóch wierzchołków w grafie, wprowadzimy pojęcie trasy.

Definicja 2.5.7. *Trasa/marszruta [12, Rozdział 3] Trasą (lub marszrutą) w danym grafie G nazywamy skończony ciąg krawędzi postaci $v_0v_1, v_1v_2, \ldots, v_{m-1}v_m$, zapisywany również w postaci $v_0 \to v_1 \to v_2 \to \ldots \to v_m$, w którym każde dwie kolejne krawędzie są albo sąsiednie, albo identyczne. Taka trasa wyznacza ciąg wierzchołków $v_0, v_1, \ldots, v_m$. Wierzchołek v_0 nazywamy wierzchołkiem początkowym, a wierzchołek v_m wierzchołkiem końcowym trasy, mówimy też wtedy, o trasie od wierzchołka v_0 do wierzchołka v_m. Liczbę krawędzi na trasie nazywamy długością trasy.*

Definicja 2.5.8. *Ścieżka [12, Rozdział 3] Trasą, w której wszystkie krawędzie są różne, nazywamy ścieżką.*

Definicja 2.5.9. *Droga [12, Rozdział 3] Ścieżkę, w której wierzchołki oznaczane kolejno: $v_0, v_1, \ldots, v_m$ są różne (z wyjątkiem, być może, równości $v_0 = v_m$, gdzie v_0 to wierzchołek początkowy a v_m to wierzchołek końcowy), nazywamy drogą.*

Definicja 2.5.10. *Droga zamknięta/ścieżka zamknięta [12, Rozdział 3] Droga jest zamknięta gdy rozpoczyna się i kończy w tym samym punkcie, tj. według przyjętej notacji $v_0 = v_m$.*

Definicja 2.5.11. *Cykl [12, Rozdział 3] Cyklem nazywamy drogę zamkniętą.*

Definicja 2.5.12. *Graf spójny [12, Rozdział 3] Graf jest spójny wtedy i tylko wtedy, gdy każda para wierzchołków jest połączona drogą.*

Definicja 2.5.13. *Drzewo [12, Rozdział 4] Drzewem nazywamy graf spójny, nie zawierający cykli.*

Definicja 2.5.14. *Graf skierowany(digraf albo graf zorientowany) [12, Rozdział 7] Graf skierowany lub digraf D, składa się z niepustego zbioru skończonego $V(D)$ elementów nazywanych wierzchołkami i skończonej rodziny $E(D)$ par uporządkowanych elementów zbioru $V(D)$, nazywanych łukami. Zbiór $V(D)$ nazywamy zbiorem wierzchołków, a rodzinę $E(D)$ rodziną łuków digrafu D (krawędzi grafu skierowanego). Łuk (v, w) zwykle zapisujemy jako vw. Graf skierowany oznaczamy zwykle w postaci pary uporządkowanej $G =< V, E >$*

Uwaga 5. *Każdy graf jednoznacznie wyznacza pewną relację dwuargumentową (binarną) w skończonym zbiorze V. Można również powiedzieć odwrotnie, że każda relacja dwuargumentowa (binarna) r w skończonym zbiorze V, wyznacza jednoznacznie graf zorientowany, którego węzłami są elementy skończonego zbioru V, z kolei krawędziami są uporządkowane pary $< v, v' >$, należące do r.*

Uwaga 6. *Niech dany będzie digraf D składający się z niepustego zbioru skończonego wierzchołków $V(D)$ i skończonej rodziny krawędzi$E(D)$. W momencie gdy $\forall a, b \in V(D) < a, b > \in E(D)$ wówczas $< b, a > \in E(D)$, to taki graf skierowany, może być utożsamiony z grafem niezorientowanym.*

2.6 Wybrane pojęcia z teorii mnogości, topologii i algebry liniowej

2.6.1 Relacja porządkująca

W niniejszej pracy skupiamy się na zagadnieniu porządkowania danych wielowymiarowych. Konieczne jest zatem przywołanie odpowiednich sformułowań dotyczących matematycznej definicji porządku. Najbardziej podstawowym pojęciem jest relacja porządku, która zostanie zdefiniowana poniżej. W sekcji tej zostaną również podane pojęcia równoliczności zbiorów, mocy zbioru, które są potrzebne przy definiowaniu własności porządkowania liniowego zbioru obiektów. Dodatkowo przytoczona została definicja zbioru skończonego, ze względu na zastosowanie metod porządkowania na zbiorze skończonym.

Definicja 2.6.1. *Relacja [7, Rozdział 3] Niech dane będą zbiory X i Y. Relacją (dwuargumentową) między elementami zbiorów X i Y nazywamy dowolny podzbiór $\rho \subset X \times Y$. Jeśli $X = Y$ to mówimy, że ρ jest relacją na zbiorze X.*

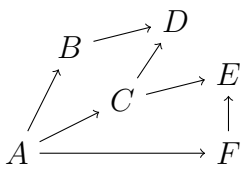
Definicja 2.6.2. *Relacja porządkująca (częściowego porządku) [3, Rozdział 2] Niech dana relacja ρ, którą oznaczać będziemy przez $\leqslant$, będzie określona dla elementów ustalonego zbioru X. Mówimy, że relacja $\leqslant$ jest relacją częściowego porządku, jeśli spełnione są warunki:*

1. *$x \leqslant x$ dla każdego x (zwrotność),*
2. *jeśli $x \leqslant y$ i $y \leqslant x$, to $x = y$ (słaba antysymetryczność),*
3. *jeśli $x \leqslant y$ i $y \leqslant z$, to $x \leqslant z$ (przechodniość).*

Przykład 2. *Częściowego porządku na zbiorze. Wykorzystanie częściowego porządku na płaszczyźnie, obrazuje diagram Hassego, będący grafem skierowanym, którego wierzchołki zostały poddane relacji porządkowania i reprezentują elementy skończonego zbioru X. Aby go skonstruować, należy postępować według poniższych kroków:*

- *Punkty obrazujące elementy zbioru X, umieszcza się na płaszczyźnie.*
- *Punkt $x \in X$ łączony jest odcinkiem z punktem $y \in X$, jeśli x jest następnikiem y, czyli gdy $y < x$ oraz nie istnieje taki punkt $z \in X$, że $y < z < x$.*

Zobrazowanie relacji częściowego porządku, dla punktów $A, B, C, D, E, F \in X$



Definicja 2.6.3. *Relacja liniowo porządkująca (liniowy porządek) [3, Rozdział 2] Niech dany będzie niepusty zbiór X. Relację $\leqslant$ porządkującą zbiór X, nazywamy relacją liniowo porządkującą lub porządkiem liniowym, gdy dla dowolnych $x, y \in X$ spełnia ona następujący warunek spójności tzn. $x \leqslant y$ lub $y \leqslant x$. Parę $(X, \leqslant)$ nazywamy zbiorem liniowo uporządkowanym lub łańcuchem.*

Definicja 2.6.4. *Dobry porządek [3, Rozdział 2] Niech dany będzie zbiór X. Relację $\leqslant$ porządkującą zbiór X, nazywamy dobrym porządkiem na zbiorze X, gdy w każdym niepustym podzbiorze zbioru X istnieje element najmniejszy względem relacji $\leqslant$. Jeśli relacja $\leqslant$ na zbiorze X jest dobrym porządkiem, to mówimy, że para $(X, \leqslant)$ jest zbiorem dobrze uporządkowanym.*

Definicja 2.6.5. *Elementy wyróżnione [3, Rozdział 2] Niech X będzie zbiorem częściowo uporządkowanym przez relacje $\leqslant$ oraz niech $a \in X$. Mówimy, że:*

1. *a jest elementem najmniejszym w X, gdy $\forall x \in X \quad a \leqslant x$*
2. *a jest elementem minimalnym w X, gdy jest jedynym elementem najmniejszym w X*
3. *a jest elementem największym w X, gdy $\forall x \in X \quad x \leqslant a$*
4. *a jest elementem maksymalnym w X, gdy jest jedynym elementem największym w X*

Definicja 2.6.6. *Ograniczenie górne [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Element $x \in X$ nazywamy ograniczeniem górnym zbioru A względem relacji $\leqslant$, gdy $a \leqslant x$ dla każdego $a \in A$.*

Definicja 2.6.7. *Ograniczenie dolne [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Element $y \in X$ nazywamy ograniczeniem dolnym zbioru A względem relacji $\leqslant$, gdy $y \leqslant a$ dla każdego $a \in A$.*

Definicja 2.6.8. *Zbiór ograniczony [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Zbiór nazywamy ograniczonym z góry (ograniczonym z dołu), jeśli ma on ograniczenie górne (dolne).*
Zbiór ograniczony z dołu i z góry nazywamy ograniczonym.

Definicja 2.6.9. *Kres górny [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Jeśli zbiór A jest ograniczony z góry i wśród ograniczeń górnych zbioru A istnienie element najmniejszy x_0, to element ten nazywamy kresem górnym zbioru A i oznaczamy symbolem $supA$. Tak więc $x_0 = supA$, gdy spełnione są następujące warunki:*

1. *$a \leqslant x_0$ dla każdego $a \in A$,*
2. *jeśli $a \leqslant x$ dla każdego $a \in A$, to $x_0 \leqslant x$.*

Definicja 2.6.10. *Kres dolnym [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Jeśli zbiór A jest ograniczony z dołu i wśród ograniczeń dolnych zbioru A istnienie element największy x_0, to element ten nazywamy kresem dolnym zbioru A i oznaczamy symbolem $infA$. Tak więc $x_0 = infA$, gdy spełnione są następujące warunki:*

1. *$y_0 \leqslant a$ dla każdego $a \in A$,*
2. *jeśli $y \leqslant a$ dla każdego $a \in A$, to $y \leqslant y_0$.*

Definicja 2.6.11. *Zbiory równoliczne [3, Rozdział 5] Mówimy, że zbiory A i B są równoliczne (tej samej mocy), gdy istnieje bijekcja, tj. funkcja f różnowartościowa, przekształcająca zbiór A na zbiór B, tzn. $f : A \to B$. Piszemy wtedy: $\overline{\overline{A}} = \overline{\overline{B}}$.*

Definicja 2.6.12. *Zbiór skończony [3, Rozdział 5] Mówimy, że zbiór A jest skończony, gdy jest pusty lub równoliczny ze zbiorem $\{1, \ldots, n\}$, dla pewnego $n \in \mathbb{N}$. Gdy zbiór jest równoliczny ze zbiorem $\{1, \ldots, n\}$, to mówimy że jest on n-elementowy, tj. mocy równej n.*

Definicja 2.6.13. *Zbiór przeliczalny [3, Rozdział 5] Mówimy, że zbiór X jest przeliczalny, gdy jest skończony lub jest równoliczny z $\mathbb{N}$.*

2.6.2 Przestrzenie metryczne, miary odległości

Niezbędnym jest również wprowadzenie podstawowych pojęć z topologii, ze względu na stosowanie funkcji odległości w celu uporządkowania obiektów.

Definicja 2.6.14. *Metryka [7, Rozdzial 9] Niech X będzie niepustym zbiorem, wtedy funkcję $\mathrm{d} : X \times X \to [0, \infty)$, nazywamy metryką jeśli spełnione są warunki:*

1. $\forall x, y \in X \quad \Big(\mathrm{d}(x, y) = 0 \iff x = y\Big)$,
2. $\forall x, y \in X \quad \mathrm{d}(x, y) = \mathrm{d}(y, x)$,
3. $\forall x, y, z \in X \quad \mathrm{d}(x, y) \leqslant \mathrm{d}(x, z) + \mathrm{d}(z, y)$.

Definicja 2.6.15. *Przestrzeń metryczna [7, Rozdział 9] Niech X będzie niepustym zbiorem, d metryką, wówczas parę (X, d) nazywamy przestrzenią metryczną.*

Przykład 3. *Metryka euklidesowa w $\mathbb{R}^2$ Niech $\mathrm{d}_e : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ będzie metryką euklidesową, wówczas*

$$\forall (x_1, x_2), (y_1, y_2) \in \mathbb{R}^2 \quad \mathrm{d}_e((x_1, y_1), (x_2, y_2)) := \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Przykład 4. *Metryka miejska(Manhattan) w $\mathbb{R}^2$ Niech $\mathrm{d}_m : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ będzie metryką miejską, wówczas*

$$\forall (x_1, x_2), (y_1, y_2) \in \mathbb{R}^2 \quad \mathrm{d}_m((x_1, y_1), (x_2, y_2)) := |x_1 - x_2| + |y_1 - y_2|.$$

Przykład 5. *Przestrzeń euklidesowa n-wymiarowa $\mathbb{R}^n$ Niech $\mathrm{d}_e : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ będzie metryką euklidesową, wówczas*

$$\forall (x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n \quad \mathrm{d}_e(x, y) := \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}.$$

Rozdział 3

Metody porządkowania

Rozdział ten został opracowany w oparciu o [9, Rozdział 2]. Omówiono w nim wybrane metody porządkowania zarówno liniowego jak i nieliniowego. W kolejnym rozdziale szczegółowo przyjrzeliśmy się wybranym metodom wraz z przedstawieniem ich algorytmów oraz dokładnych opisów matematycznych.

Metody porządkowania liniowego umożliwiają przeprowadzenie takiego uporządkowania, w wyniku którego zwracana jest uporządkowana lista obiektów na podstawie określonego kryterium (np. wartości zmiennych). Z kolei metody porządkowania nieliniowego zwracają graf połączeń obiektów podobnych ze względu na opisujące je zmienne.

3.1 Metody porządkowania liniowego

W wielowymiarowej przestrzeni zmiennych, porządkowanie liniowe obiektów sprowadza się do rzutowania punktów, które reprezentują obiekty poddane porządkowaniu, na prostą. Taka operacja pozwala na ustaleniu hierarchii obiektów.

Poniżej zostaną przedstawione własności uporządkowania liniowego obiektów, wraz z podaniem ich matematycznej interpretacji.

Obiekty uporządkowane liniowo charakteryzują się tym, że:

- każdy obiekt ma przynajmniej jednego sąsiada i nie więcej niż dwóch sąsiadów,
- jeżeli sąsiadem i-tego obiektu jest k-ty obiekt, to jednocześnie sąsiadem k-tego obiektu jest i-ty obiekt,
- dokładnie dwa obiekty mają tylko jednego sąsiada.

Powyżej wymienione własności są wynikiem posiadania jedynie skończonej ilości obiektów, które poddane są uporządkowaniu. W następnej części chcielibyśmy:

- sformalizować rozumienie powyższych własności,
- udowodnić ich poprawność,
- rozważyć dostateczność tych własności w zbiorach o skończonej ilości obiektów.

Na początku zaczniemy od sprecyzowania takich pojęć jak sąsiad względem relacji.

Definicja 3.1.1. *Sąsiad względem relacji $\leqslant$ Niech X będzie niepustym zbiorem, a x, y będą dwoma różnymi elementami należącymi do tego zbioru. Mówimy, że $y \in X$ jest sąsiadem $x \in X$,co zapisujemy ySx, jeśli*

$(y \leqslant x \vee x \leqslant y) \quad \wedge \quad (\neg\exists_{z \in X} \quad x \neq z \neq y \Rightarrow y \leqslant z \leqslant x \vee x \leqslant z \leqslant y)$.

Twierdzenie 3.1.2. *Własności porządku liniowego w zbiorach skończonych Niech $\leqslant$ będzie relacją porządku liniowego zdefiniowaną w X, gdzie X jest zbiorem ze skończoną liczbą obiektów, złożonym co najmniej z dwóch elementów. Wtedy*

1. $\forall_{x\in X} \quad \overline{\overline{\{y \in X, ySx\}}} \in \{1, 2\}$,
2. $\forall_{x,y\in X} \quad ySx \Rightarrow xSy$,
3. $\overline{\overline{\{x \in X, \quad \overline{\overline{\{y \in X, \quad ySx\}}} = 1\}}} = 2$

gdzie S oznacza sąsiada względem relacji $\leqslant$.

Dowód. Poniżej zostaną udowodnione powyższe własności.

1. Niech $x \in X$. Przypuśćmy na początek, że $\overline{\overline{\{y \in X, ySx\}}} = 0$, tzn. że obiekt x nie posiada sąsiadów w tej relacji. Nasz zbiór X jest jednak co najmniej dwuelementowy, zatem istnieje element $y \in X$ i $x \neq y$. Wobec spójności linowego porządku z Definicji 2.6.3 zachodzi wtedy
$$x \leqslant y \vee y \leqslant x.$$
Jednak wiemy, że y nie może być sąsiadem x gdyż ten nie posiada sąsiadów. Zatem z definicji sąsiada musi istnieć $z \in X$ różny od obu $x \neq z \neq y$, spełniający warunek
$$z \leqslant x \vee x \leqslant z.$$
Powyższe rozumowanie dla y można by dalej zastosować do z, uzyskując kolejne z_1 a później $z_2, z_3, \ldots$ dowolną ilość różnych elementów z których każdy występuje w relacji liniowego porządku z x, ale żaden z nich nie jest sąsiadem. Jednak nasz zbiór X jest zbiorem skończonym, więc nigdy nie uda nam się utworzyć dowolnej ilości różnych elementów ze zbioru X (elementy się wyczerpią). Zatem nasze przypuszczenie, że $\overline{\overline{\{y \in X, ySx\}}} = 0$ jest fałszywe.
Przypuśćmy dalej, że $\overline{\overline{\{y \in X, ySx\}}} \geqslant 3$. Niech a, b, c będą trzema różnymi elementami z X będącymi sąsiadami dla x. Wtedy bez straty ogólności możemy przyjąć, że $a \leqslant x, b \leqslant x$ lub $x \leqslant a, x \leqslant b$. Istotnie mając 3 elementy w relacji wtedy co najmniej dwa muszą znajdować się po zgodnej stronie, a z dokładnością do oznaczeń możemy przyjąć, że będą nimi a oraz b. Ustalmy zatem, że $a \leqslant x, b \leqslant x$. Wobec definicji 2.6.3 wiemy, że $a \leqslant b$ lub $b \leqslant a$. Jeśli $a \leqslant b$ to $a \leqslant b \leqslant x$. Co przeczy temu, że a jest sąsiadem x. Jeśli $b \leqslant a$ to $b \leqslant a \leqslant x$ co przeczy temu, że b jest sąsiadem. Analogicznie postępujemy dla przypadku $x \leqslant a, x \leqslant b$. Uzyskujemy zatem sprzeczność, będącą efektem przypuszczenia, że mogą istnieć takie 3 elementy a, b, c. Zatem ostatecznie $\overline{\overline{\{y \in X, ySx\}}} \in \{1, 2\}$.

2. Niech $x, y \in X$ oraz niech ySx. Korzystając z definicji sąsiada 3.1.1 mamy, że skoro ySx to
$$(y \leqslant x \vee x \leqslant y) \quad \wedge \quad (\neg\exists_{z\in X} \quad x \neq z \neq y \Rightarrow y \leqslant z \leqslant x \vee x \leqslant z \leqslant y),$$
Natomiast xSy oznacza, że
$$(x \leqslant y \vee y \leqslant x) \quad \wedge \quad (\neg\exists_{z\in X} \quad y \neq z \neq x \Rightarrow x \leqslant z \leqslant y \vee y \leqslant z \leqslant x).$$
Wobec powyższego widać, że te dwa zdania znaczą to samo, stąd widać że $ySx \Rightarrow xSy$.

3. Intuicyjnie te dwa elementy posiadające po jednym sąsiedzie są elementami maksymalnym i minimalnym w tym zbiorze. Udowodnimy kolejno:

- Element minimalny w zbiorze ma pojedynczego sąsiada. Załóżmy, że zbiór musi posiadać dokładnie 1 element minimalny, tzn. $x_m \in X$ takie, że

$$\forall x \in X \quad x_m \leqslant x.$$

Istotnie przypuśćmy, że nie istnieje element minimalny. Niech x_1 będzie dowolnym elementem z X. Skoro nie istnieje element minimalny, to istnieje $x_2 \in X$ takie, że $x_2 \leqslant x_1$ i $x_2 \neq x_1$. Dla x_2 z braku elementu minimalnemu, musi istnieć z kolei $x_3 \leqslant x_2$ takie, że $x_2 \neq x_3$. Itd. Co nie jest możliwe, gdyż zbiór X jest przecież skończonym zbiorem. Rozważmy dalej przypuszczenie gdyby były dwa lub więcej takich elementów. Wtedy to, z antysymetryczności, oczywiście musiałyby być sobie równe. Jeśli x_m, y_m są jednocześnie minimalne to

$$\forall x \in X \quad x_m \leqslant x,$$

oraz

$$\forall x \in X \quad y_m \leqslant x.$$

Skąd natychmiast mamy, że $x_m \leqslant y_m$ oraz $y_m \leqslant x_m$. Wobec antysymetryczności z definicji 2.6.2 mamy, że $x_m = y_m$ wbrew naszemu przypuszczeniu, że są od siebie różne. Pozostaje pokazać, że element minimalny ma pojedynczego sąsiada. Przypuśćmy, że $y, z \in X$ są dwoma różnymi sąsiadami dla x_m. Wtedy $x_m \leqslant y \vee y \leqslant x_m$ oraz $x_m \leqslant z \vee z \leqslant x_m$. Skoro x_m jest minimalny to musi to oznaczać, że

$$x_m \leqslant y \wedge x_m \leqslant z.$$

Wobec spójności z definicji 2.6.3 zachodzi $y \leqslant z$ lub $z \leqslant y$. Sprzeczność, gdyż wtedy któryś z nich nie mógłby być sąsiadem dla x_m.

- Element maksymalny x_M w zbiorze ma pojedynczego sąsiada. Analogicznie do powyższego punktu, zbiór musi posiadać dokładnie 1 element maksymalny, tzn. $x_M \in X$ takie, że

$$\forall x \in X \quad x \leqslant x_M.$$

Istotnie przypuśćmy, że nie istnieje element maksymalny. Niech x_1 będzie dowolnym elementem z X. Skoro nie istnieje element maksymalny, to istnieje $x_2 \in X$ takie, że $x_1 \leqslant x_2$ i $x_1 \neq x_2$. Dla x_2 z braku elementu maksymalnego, musi istnieć taki element $x_3 \in X$ i $x_3 \neq x_2$, że $x_2 \leqslant x_3$. Itd. Co nie jest możliwe, gdyż z założenia zbiór X jest skończonym zbiorem. Rozważmy dalej przypuszczenie gdyby były dwa lub więcej takich elementów. Wtedy to z antysymetryczności, musiałyby być sobie równe. Jeśli x_M, y_M są jednocześnie maksymalne, to

$$\forall x \in X \quad x \leqslant x_M,$$

oraz

$$\forall x \in X \quad x \leqslant y_M.$$

Stąd natychmiast mamy, że $x_M \leqslant y_M$ oraz $y_M \leqslant x_M$. Wobec antysymetryczności z definicji 2.6.2, mamy że $x_M = y_M$, co wbrew naszemu przypuszczeniu daje, że elementy te nie są od siebie różne. Pozostaje pokazać, że element maksymalny ma pojedynczego sąsiada. Przypuśćmy, że $y, z \in X$ są dwoma różnymi sąsiadami dla x_M. Wtedy $x_M \leqslant y \vee y \leqslant x_M$ oraz $x_M \leqslant z \vee z \leqslant x_M$. Skoro x_M jest elementem maksymalny to musi to zatem oznaczać

$$y \leqslant x_M \wedge z \leqslant x_M.$$

Wobec spójności z definicji 2.6.3 zachodzi $y \leqslant z$ lub $z \leqslant y$. Sprzeczność, gdyż wtedy któryś z nich nie mógłby być sąsiadem dla x_M.

- Żaden inny element nie może mieć pojedynczego sąsiada. Przypuśćmy, że $x \in X$ nie będąc ani elementem minimalnym ani maksymalnym ma pojedynczego sąsiada. Wobec definicji elementu minimalnego i maksymalnego oraz spójności zachodzi

$$x_m \leqslant x \leqslant x_M.$$

Zatem albo x_m jest sąsiadem x albo istnieje $y_1 \in X$ taki, że $y_1 \leqslant x$. Tworzy to kilka możliwych przypadków. W pierwszym x_m będzie tym jedynym sąsiadem, w drugim x_M nim będzie, w ostatnim natomiast, ani x_m, ani x_M nie będą sąsiadami.

Zajmiemy się najpierw pierwszym z nich, tj. x_m jest sąsiadem x. Zauważmy teraz, że z faktu, iż x_M jest elementem maksymalnym zbioru X, wynika że $x \leqslant x_M$. Nie jest jednak sąsiadem elementu x. Zatem istnieje takie $x_1 \in X$, że $x \leqslant x_1 \leqslant x_M$. Jednak x_1 również nie może być sąsiadem X co powoduje, że istnieje taki element $x_2 \in X$, że $x \leqslant x_2 \leqslant x_1 \leqslant x_M$. Itd. Jednakże, skoro zbiór X jest zbiorem skończonym, to musi istnieć taki element $x_j \in X$, że $x \leqslant x_j$, który będzie sąsiadem z x, zatem xSx_j. Zatem ostatecznie xSx_m i xSx_j, a to przeczy założeniu, że x ma pojedynczego sąsiada.

Przejdźmy teraz do drugiego przypadku, tj. gdy x_M jest sąsiadem dla x. Wtedy x_m nie jest sąsiadem dla x jednak wiedząc, że $x_m \leqslant X$ musi istnieć $y_1 \in X$, taki że $y_1 \leqslant x$. Jednak wiedząc iż y_1 nie jest sąsiadem dla X wnioskujemy, że istnieje taki $y_2 \in X$, że $x_m \leqslant y_1 \leqslant y_2 \leqslant x$. Itd. Jednakże, skoro zbiór X jest zbiorem skończonym, to musi istnieć taki element $y_i \in X$, że $y_i \leqslant x$, który będzie sąsiadem z x. Zatem ostatecznie y_iSx i xSx_M, co przeczy założeniu że x ma pojedynczego sąsiada.

Zajmijmy się teraz trzecim przypadkiem, tj. gdy ani x_m oraz x_M nie są sąsiadami elementu x. Z faktu, iż zbiór X posiada element minimalny x_m, który nie jest sąsiadem elementu x, wynika że istnieje taki element $x_1 \in X$, że $x_m \leqslant x_1 \leqslant x$. Co więcej istnieje taki $x_2 \in X$, że $x_m \leqslant x_1 \leqslant x_2 \leqslant x$. Itd. I znów skoro zbiór X jest skończony, to istnieje taki element $x_j \in X$, że $x_j \leqslant x$ i x_jSx. Z drugiej strony, skoro zbiór X posiada element maksymalny x_M, który nie jest sąsiadem elementu x, wynika że istnieje taki element $y_1 \in X$, że $x \leqslant y_1 \leqslant x_M$. Analogicznie do wcześniejszych kroków, istnieje takie element $y_2 \in X$, że $x \leqslant y_2 \leqslant y_1 \leqslant x_M$. Itd. Zbiór X jest zbiorem skończonym, zatem musi istnieć taki element $y_i \in X$, że $x \leqslant y_i$ i xSy_1. Łącząc te dwa warunki, wynika że x musi mieć dwóch sąsiadów. Co kończy dowód własności.

□

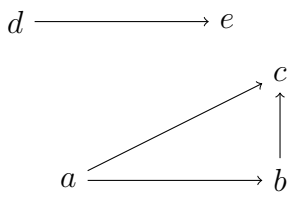
Własności, które wykazaliśmy powyżej są często podawane niemal na równi z definicją takiego uporządkowania. Poniżej zostaną podane przykłady takich relacji, które mimo że posiadają powyższy zestaw własności, to nie opisują relacji będących porządkami liniowymi.

Przykład 6. *Rozważmy zbiór dwuelementowy $X = \{a, b\}$ gdzie $a \leqslant b$ jest jedynym punktem tej relacji. Tak zdefiniowana relacja spełnia wszystkie własności, ale nie spełnia założenia o zwrotności - zatem relacja ta nie jest liniowym porządkiem.*

Diagram Hassego prezentujący tę relację, jest postaci:

Przykład 7. *Rozważmy zbiór $X = \{a, b, c, d, e\}$ oraz relację definiującą następujące sąsiedztwa (wypisaną bez par symetrycznych) aSb, bSc, aSc, dSe. Ponadto dołóżmy warunek zwrotności, tj. $a \leqslant a$, $b \leqslant b$, $c \leqslant c$, $d \leqslant d$, $e \leqslant e$.*

Diagram Hassego prezentujący relację porządku tego zbioru, jest postaci:



Z diagramu widać, że taka relacja spełnia wszystkie omawiane wcześniej własności - jednak nie jest spójna. I tak np. nie możemy porównać elementów a i d, bowiem nie możemy określić czy $d \leqslant a$ *lub* $a \leqslant d$.

W podsumowaniu tej sekcji należy podkreślić, że by uporządkować liniowo obiekty, charakteryzujące je zmienne muszą być mierzone przynajmniej na skali porządkowej. Istotne jest również aby miały jednakowy charakter. Na potrzeby pracy zakładamy, że zmienne opisujące obiekty powinny być stymulantami. Dlatego też gdy nimi nie są, należy poddać je np. stymulacji. Operacja ta umożliwia w dalszym kroku przejście do transformacji normalizacyjnej, która konieczna jest gdy zmienne opisujące obiekty mierzone są na skali przedziałowej lub ilorazowej, a chcemy uzyskać ich porównywalność.

Metody porządkowania liniowego można podzielić na metody diagramowe, procedury oparte na zmiennej syntetycznej oraz procedury iteracyjne bazujące funkcji kryterium dobroci uporządkowania. W kolejnej sekcji zostaną pokrótce przedstawione różne metody, z wyszczególnieniem najważniejszych założeń o każdej z nich.

3.1.1 Metody diagramowe

W metodach diagramowych stosuje się graficzną reprezentację macierzy odległości zwanej diagramem. Macierz konstruowana jest w oparciu o odległości między obiektami, wyznaczone za pomocą dowolnej metryki. Porządkowanie obiektów polega na porządkowaniu diagramu. Tzn. przestawieniu wierszy i odpowiadających im kolumn, aby wzdłuż przekątnej skupiały się najmniejsze odległości zaś im dalej od głównej przekątnej tym większe odległości między zmiennymi opisującymi porządkowane obiekty.

Narzędzie pomocnicze w porządkowaniu danych, może stanowić kryterium postaci:

$$F^1 = \sum_{i=1}^{n} \sum_{k>1}^{n} d_{ik} w_{ik}$$

gdzie:

$D = [d_{ik}]$ - macierz odległości między i-tym i k-tym obiektem

$W = [w_{ik}] \quad i, k = 1, 2, \ldots, n$ - macierz wag elementów macierzy odległości, wymiaru równego wymiarowi macierzy odległości

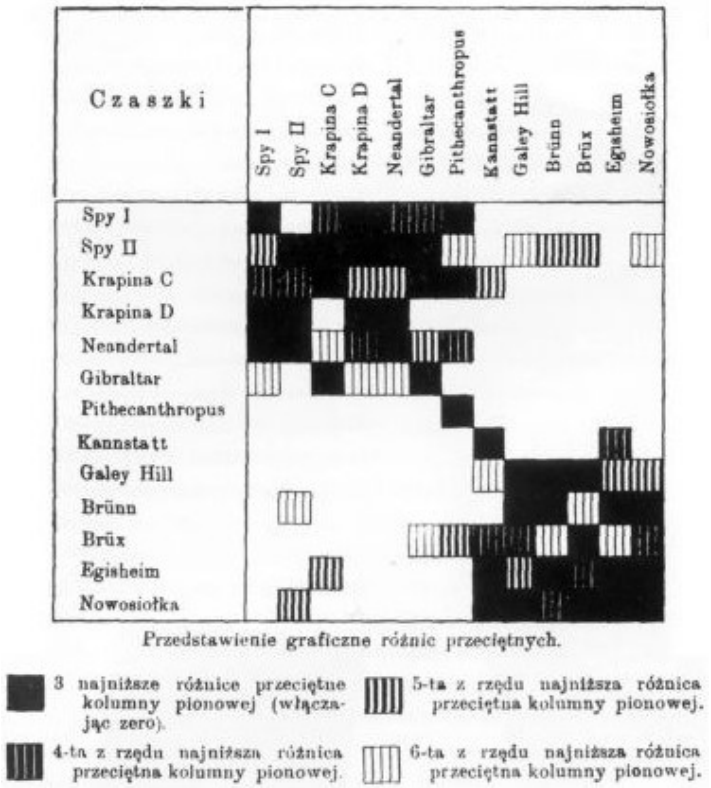
Elementy macierzy wag wyznaczone są za pomocą jednego z poniższych wzorów:

$$w_{ik} = \frac{|i-k|}{n-1},$$

$$w_{ik} = \frac{1}{n(n-1)}[2n|i-k-1| + i + k - (i-i)^2],$$

$$w_{ik} = \frac{1}{n(n-1)}[2n|i-k| + 2 - i - k - (i-i)^2].$$

Zaprezentujmy teraz przykład uporządkowanego diagramu, przedstawiającego wynik badań Jana Czekanowskiego, dotyczących metod badania różnic między kopalnymi czaszkami ludzkimi. Analizując diagram, zauważamy że wzdłuż głównej przekątnej skupiają się najmniejsze odległości między zmiennymi, a im dalej od niej tym odległości zwiększają się. Przykład został znaleziony na stronie [11].



Rysunek 3.1: Diagram opublikowany w podręczniku "Statystyka dla antropologów" Jana Czekanowskiego, 1913r.

Metody diagramowe nie leżą w zakresie zainteresowań tej pracy, z uwagi na mały formalizm matematyczny. Są jednak istotnym narzędziem do wizualizacji porządku.

3.1.2 Metody oparte na zmiennych syntetycznych

W tym podrozdziale zostaną opisane metody porządkowania oparte na zmiennych syntetycznych, tj. funkcji wyznaczonej na podstawie wartości zmiennych opisujących obiekty, której wartości będą służyć do porządkowania zbioru. Metody oparte na zmiennych syntetycznych dzielimy na wzorcowe i bezwzorcowe. Poniżej zostaną one opisane szczegółowo, jednak wcześniej zostaną przedstawione wzory wyznaczające zmienną syntetyczną.

Sposoby wyznaczania zmiennej syntetycznej

W pracy dla zachowania ogólności, przyjmujemy że wszystkie zmienne opisujące obiekty mają jednakowe wagi, w związku z tym wzory służące do wyznaczenia zmiennej syntetycznej są postaci:

1. średniej arytmetycznej:

$$s_i = \frac{1}{m} \sum_{j=1}^{m} n_{ij}, \quad i = 1, 2, \ldots, n,$$

2. średniej geometrycznej:

$$s_i = \prod_{j=1}^{m} (n_{ij})^{\frac{1}{m}}, \quad i = 1, 2, \ldots, n,$$

3. średniej harmonicznej

$$s_i = \Big[\sum_{j=1}^{m} \frac{1}{n_{ij}} \Big]^{-1} \cdot m, \quad i = 1, 2, \ldots, n,$$

gdzie: s_i - wartość zmiennej syntetycznej w i-tym obiekcie,

Metoda wzorcowe

W metodach tych zakłada się istnienie obiektu wzorcowego $P_0 = [n_{0j}], \quad j = 1, 2, \ldots, m$, w którym znormalizowane zmienne wejściowe $n_{0j}, j = 1, 2, \ldots, m$, będące współrzędnymi obiektu wzorcowego, przyjmują optymalne wartości, które to są ustalane na podstawie ogólnie przyjętych norm, subiektywnej opinii dotyczącej obserwowanego obiektu, lub też opinii ekspertów. Wtedy obiekty, porządkowane są na podstawie odległości od obiektu wzorcowego. Poszczególne metody mogą różnić się sposobem wyznaczania obiektu wzorcowego, odległości oraz miary syntetycznej, na podstawie której dokonywane jest porządkowanie.

Metoda Hellwiga

Metoda Hellwiga jest jedną z najstarszych metod wzorcowych. W metodzie tej, obiekt wzorcowy wyznaczony jest na podstawie wystandaryzowanych zmiennych wejściowych. Współrzędnym obiektu wzorcowego przyporządkowuje się maksimum, gdy zmienne wejściowe są stymulantami lub minimum gdy zmienne są destymulantami. Obiekty są uporządkowywane na podstawie odległości od obiektu wzorcowego, przy wykorzystaniu odległości euklidesowej. Do porządkowania używana jest natomiast miara syntetyczna:

$$s_i = 1 - \frac{d_{i0}}{d_0}, \quad i = 1, 2, \ldots, n,$$

gdzie:

d_{i0} - odległość i-tego obiektu, od obiektu wzorcowego

d_0 - wartość wyznaczana dla każdej zmiennej, będąca sumą średniej odległości od obiektu oraz podwojonej wartości odchylenia standardowego dla tej zmiennej

Współrzędne obiektu wzorcowego są obliczane na podstawie wzoru:

$$n_{0j} = \begin{cases} \max\limits_{i}\{n_{ij}\} & \text{dla } n_j^S, \quad j = 1, 2, \ldots, m, \quad i = 1, 2, \ldots, n, \\ \min\limits_{i}\{n_{ij}\} & \text{dla } n_j^D, \quad j = 1, 2, \ldots, m, \quad i = 1, 2, \ldots, n, \end{cases}$$

Etapy wyznaczania miary syntetycznej:

- Dla każdego obiektu, wyznaczana jest odległość od obiektu wzorcowego za pomocą:

$$d_{i0} = \left[\sum_{j=1}^{m} (n_{ij} - n_{0j})^2 \right]^{\frac{1}{2}}$$

- Następnie dla każdej zmiennej wyznaczana jest średnia odległość od obiektu wzorcowego wg. wzoru:

$$\overline{d_0} = \frac{1}{n} \sum_{i=1}^{n} d_{i0}$$

- Wyznaczamy odchylenie standardowe dla każdej zmiennej za pomocą wzoru:

$$\sigma(d_0) = \left[\frac{1}{n}\sum_{i=1}^{n}(d_{i0} - \overline{d_0})^2\right]^{\frac{1}{2}}$$

- Mając powyższe, możemy wyznaczyć wartość d_0 jako sumę średniej odległości oraz podwojonej wartości odchylenia standardowego:

$$d_0 = \overline{d_0} + 2\sigma(d_0)$$

Wartości miary s_i zazwyczaj są z przedziału $[0, 1]$, jednakże mogą zdarzyć się wartości ujemne. Należy tu zaznaczyć, że wartości miary są tym wyższe, im mniej jest oddalony obiekt od obiektu wzorcowego.

Metoda dystansowa

Podobnie jak we wcześniejsze metodach, na początku zmienne należy poddać stymulacji, oraz przekształceniu normalizacyjnemu, wybranemu na podstawie skal do których należą zmienne opisujące obiekty. W kolejnym kroku wyznaczane są współrzędne obiektu wzorcowego, a następnie macierz odległości każdego obiektu od obiektu wzorcowego. Odległość od obiektu wzorca jest wyznaczana przy zastosowaniu dowolnej metryki, np. metryki euklidesowej. Dla metody tej, miara syntetyczna jest wyznaczana za pomocą przekształcenia unitaryzacyjnego postaci:

$$s_i = \left(\frac{d_{i0} - \min\limits_i\{d_{i0}\}}{\max\limits_i\{d_{i0}\} - \min\limits_i\{d_{i0}\}}\right)^p, \quad i = 1, 2, \ldots, n \quad p \in \mathbb{N}$$

Miara syntetyczna uzyskana tą metodą jest unormowana i przyjmuje wartości z przedziału: $[0, 1]$. Czym niższa wartość miary, tym bliżej obiektu wzorcowego leży dany obiekt.

Metody bezwzorcowe

W metodach tych zakładamy, że nie istnieje obiekt wzorcowy. Porządkowanie dokonywane jest na podstawie wartości zmiennej syntetycznej, wyznaczonej dla każdego obiektu. Poniżej zostaną omówione wybrane metody porządkowania bezwzorcowego.

Metoda rang

Definicja 3.1.3. *Ranga [9, Rozdział 1.5] Rangą nazywamy zmienną o wartościach ze zbioru $\mathbb{R}_+ \setminus \{0\}$, będącą najczęściej liczbą całkowitą. Wartości reprezentują numery miejsc obiektów po uporządkowaniu.*

Metoda ta opiera się na normalizacji rangowej, w związku z tym zmienne poddane porządkowaniu, powinny być mierzone są na skali porządkowej. Dla każdego obiektu wyznacza się sumę przyporządkowanych mu rang ze względu na wszystkie zmienne. Na końcu obliczana jest wartość zmiennej syntetycznej, jako średniej wartości rang. W oparciu o tę wartość następuje porządkowanie obiektów, tj. im wartość zmiennej syntetycznej jest mniejsza tym wyżej w hierarchii znajduje się uporządkowany obiekt. Wzór na obliczenie wartości zmiennej syntetycznej:

$$s_i = \frac{1}{m}\sum_{j=1}^{m} n_{ij}, \quad i = 1, 2, \ldots, n,$$

gdzie:

n_{ij}-zmienna znormalizowana rangowo, tj. $n_{ij} = r$ dla $x_{rj} = x_{ij}, \quad h, i = 1, 2, \ldots, n.$ gdzie:

r-ranga nadana i-temu obiektowi znajdującemu się na r-tym miejscu w uporządkowanym szeregu obiektów ze względu na j-tą zmienną.

Metoda sum

Metoda ta używana jest w momencie, gdy zmienne mierzone są na skali ilorazowej lub przedziałowej. W związku z tym tuż po stymulacji zmiennych, należy dokonać przekształcenia normalizacyjnego, za pomocą unitaryzacji. W kolejnym kroku, dla każdego obiektu wyznaczana jest zmienna syntetyczna, jako średnia arytmetyczna wartości zmiennych przy przyjęciu jednakowych wag dla każdej zmiennej, jako średnia arytmetyczna wartości zmiennych. Następnie muszą zostać wyeliminowane ujemne wartości zmiennej syntetycznej, do czego służy poniższe przekształcenie:

$$s'_i = s_i - \min_i\{s_i\}, \quad i = 1, 2, \ldots, n.$$

Końcowa postać zmiennej syntetycznej otrzymywana jest, przy wykorzystaniu normalizacji postaci:

$$s''_i = \frac{s'_i}{\max_i\{s'_i\}}, \quad i = 1, 2, \ldots, n.$$

Powyższe przekształcenia ujednolicają zakres miary syntetycznej do przedziału [0, 1]. Im wyższa wartość zmiennej syntetycznej, tym wyżej w hierarchii znajduje się obiekt.

3.1.3 Metody iteracyjne

W metodach tych przyjmowania jest funkcja kryterium dobroci porządkowania, dla której w kolejnych iteracjach poszukiwane jest takie uporządkowanie liniowe obiektów, które optymalizuje wartość funkcji kryterium, aż do osiągnięcia przez nią wartości optymalnej tj. maksymalnej lub minimalnej.

Metoda Szczotki

Metoda ta polega na znalezieniu takiego uporządkowania liniowego obiektów, dla którego funkcja kryterium dobroci uporządkowania osiąga maksimum:

$$F^2 = \sum_{k=1}^{n-1} k \sum_{i=1}^{n-k} d_{ik} \rightarrow \max$$

gdzie:

d_{ik} - odległość euklidesowa miedzy i-tym i k-tym obiektem.

Sposób postępowania:

W pierwszym kroku przeprowadzane jest dowolne liniowe uporządkowanie obiektów. W kolejnym kroku, dla tego uporządkowania obliczana jest wartość funkcji kryterium dobroci uporządkowania, według powyższego wzoru. W kolejnych etapach wyznaczana jest wartość tej funkcji, dla każdej transpozycji pary obiektów. Powyższe kroki wykonywane są do momentu, gdy dowolna transpozycja pary obiektów nie spowoduje zwiększenia wartości funkcji kryterium dobroci uporządkowania.

3.2 Metody porządkowania nieliniowego

Metody porządkowania nieliniowego w odróżnieniu od metod porządkowania liniowego, polegają nie na uporządkowaniu obiektów w sposób hierarchiczny, a na określeniu dla każdego z nich, stopnia podobieństwa z innymi obiektami, na podstawie opisujących je zmiennych.

Aby zastosować metody porządkowania nieliniowego, zmienne opisujące obiekty, powinny być mierzone na skali przedziałowej lub ilorazowej. Gdy zmienne te mierzone są na skali przedziałowej lub ilorazowej, należy dokonać ich normalizacji.

Metody porządkowania nieliniowego dzielimy na metody dendrytowe oraz aglomeracyjne. Metody dendrytowe prowadzą do powstania dendrytu, prezentującego położenie obiektów ze względu na ich podobieństwo między sobą. Z kolei metody aglomeracyjne sprowadzają się do powstania grafu będącym drzewkiem połączeń, prezentującym sposób łączenia obiektów do siebie podobnych.

3.2.1 Metody dendrytowe

Definicja 3.2.1. *Dendryt [9, Rozdział 2.3] Dendrytem nazywamy graf spójny, niezawierający cykli.*

Metody dendrytowe opierają się na pojęciach teorii grafów. Metody te polegają na stworzeniu dendrytu, którego wierzchołki odpowiadają odpowiednim obiektom poddanym porządkowaniu. Krawędzie łączące wierzchołki odpowiadają zaś odległością między obiektami. Poniżej zostały opisane przykłady metod dendrytowych, tj. taksonomia wrocławska oraz metoda Prima.

Taksonomia wrocławska

W metodzie tej obiekty dzielone są na grupy obiektów najbardziej do siebie podobnych, tj. takich, dla których odległość między sobą jest jak najmniejsza. W związku z tym w pierwszej kolejności należy wyznaczyć macierz odległości obiektów D np. przy użyciu metryki euklidesowej, a następnie w każdym wierszu(kolumnie) macierzy, wyznaczamy jest element najmniejszy:

$$d_{ik} = \min_k d_{ik}, \quad i, k = 1, 2, \ldots, n, i \neq k.$$

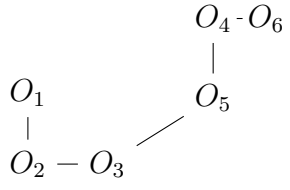
Za pomocą grafu prezentowane są pary obiektów, najbardziej do siebie podobnych. W grafie tym, długość krawędzi łączących wierzchołki(czyli obiekty, poddane porządkowaniu) odpowiadają odległości między obiektami. Jeżeli wśród połączonych par obiektów, pojawią się krawędzie dwukrotne, należy je wyeliminować, ze względu na to że kolejność połączeń w dendrycie nie jest istotna. Obiekty w dendrycie nie mogą się powtarzać, jeżeli natomiast niektóre obiekty w łączeniu wystąpią wielokrotnie, to obiekty te zostaną połączone w zespoły zwane skupieniami. Metoda ta kończy swoje działanie, w momencie uzyskania grafu spójnego.

Zaprezentujemy teraz tę metodę na przykładzie.

Przykład 8. *Niech dana będzie macierz odległości $D = [d_{ik}]_{i\leqslant 6, k\leqslant 6}$ sześciu różnych obiektów $\{O_1, O_2, ..., O_6\}$, w której to w kolejnych wierszach znajdują odległości i-tego obiektu od obiektu k-tego:*

$$D = \begin{bmatrix} 0.00 & \underline{0.35} & 0.70 & 0.95 & 2.36 & 2.99 \\ \underline{0.35} & 0.00 & 0.45 & 1.45 & 2.00 & 0.36 \\ 0.70 & \underline{0.45} & 0.00 & 1.05 & 0.90 & 0.75 \\ 0.95 & 1.45 & 1.05 & 0.00 & 0.50 & \underline{0.16} \\ 2.36 & 2.00 & 0.90 & \underline{0.50} & 0.00 & 0.52 \\ 2.99 & 0.36 & 0.75 & \underline{0.16} & 0.52 & 0.00 \end{bmatrix}$$

Mając wyznaczoną macierz odległości, postępujemy według wyżej wskazanej reguły, która prowadzi do uzyskania dendrytu postaci:



Metoda Prima

W odróżnieniu od taksonomii wrocławskiej, metoda Prima nie wymaga posługiwania się cały czas wyjściową macierzą odległości. W trakcie tworzenia dendrytu, zbiór porządkowanych obiektów jest przyporządkowywany do jednego z dwóch zbiorów, np. niech C i E oznaczają te zbiory. Niech zbiór C będzie pierwszym z nich a zbiór E drugim. Pierwszy z nich zawiera obiekty należące na danym etapie do dendrytu, zaś drugi zawiera obiekty nie należące na tym etapie do dendrytu.

W początkowym etapie, zbiór C jest zbiorem pustym, z kolei do zbioru E należą wszystkie obiekty, należące do zbioru poddanego porządkowaniu. Następnie w zbiorze C zostaje umieszczony dowolny element ze zbioru E, dowolność nie ma wpływu na ostateczną postać dendrytu. W tym momencie zostaje utworzony wektor c w którym przechowywane są odległości tego elementu od pozostałych elementów zbioru E. W kolejnym etapie do zbioru C włączany jest ten obiekt, którego odległość od elementu zbioru C jest jak najmniejsza. Po dołączeniu tego elementu wektor c przechowuje odległości tych dwóch obiektów zbioru C od pozostałych elementów zbioru E, a procedura dołączania kolejnych obiektów polega na tym samym. Cały proces trwa do momentu, aż zbiór E nie będzie zawierał żadnego elementu.

W powstałym dendrycie wierzchołkami są obiekty przechodzące zbioru C, z kolei krawędzie łączące te wierzchołki są współrzędne wektora c, które powstały przez wybór najmniejszej odległości miedzy dołączanymi obiektami, w kolejnych etapach dołączania obiektów do dendrytu.

3.2.2 Metody aglomeracyjne

Przed opisem działania metod aglomeracyjnych, wprowadzimy definicję łańcucha połączeń.

Definicja 3.2.2. *Łańcuch połączeń Niech* $O = \{O_1, O_2, ..., O_n\}, \quad n \in \mathbb{N}$ *oznacza zbiór obiektów. Łańcuchem połączeń nazwiemy skończony ciąg* $(C_i)_{i=1}^{n}$*, taki że:*

1. $C_i = \{O_i\}$ *dla* $i = 1, 2, \cdot, n$
2. $\forall_{i<n} \quad \exists_{j,k\in\mathbb{N}} \quad C_i = C_j \cup C_k$
3. $C_n = O = \{O_1, O_2, ..., O_n\}$
4. $\forall_{i<j,i,j\in\mathbb{N}} \quad C_i \cap C_j \neq \emptyset \Rightarrow C_i \subset C_j$

Uwaga 7. *Graficzną reprezentacje łańcucha połączeń nazywamy dendrogramem.*

Istotą metod aglomeracyjnych jest utworzenie łańcucha połączeń - dendrogramu. W ten sposób zobrazowana jest kolejność łączenia obiektów, na podstawie zmniejszającego się podobieństwa między obiektami włączonymi do dendrogramu, a tymi wcześniej do niego należącymi. Położenie obiektów oraz grup obiektów, które powstały w kolejnych etapach tworzenia drzewka, jest przeprowadzone na podstawie kolejności połączeń tych obiektów i grup. Każda gałąź w drzewku oznacza grupy obiektów podobnych do siebie. Wyjściowym założeniem metod aglomeracyjnych jest to, że każdy obiekt stanowi odrębną, jednoelementową grupę $(\mathrm{G_z}, z = 1, 2, \ldots, n)$.

W kolejnych krokach następuje łączenie ze sobą grup obiektów najbardziej podobnych do siebie ze względu na ich zmienne. Podobieństwo weryfikowane jest na podstawie odległości między grupami.

Na początku odległości między jednoelementowymi grupami $\mathrm{G_1}, \ldots, \mathrm{G_n}$ wyznacza wyjściowa macierz odległości D. W macierzy D poszukiwane są najmniejsze odległości pomiędzy grupami obiektów:

$$d_{zz'} = \min_{ik} d_{ik}, \quad i = 1, 2, \ldots, n_z, \quad k = 1, 2, \ldots, n_{z'}, \quad z, z^{'} = 1, 2, \ldots, n, z \neq z^{'}.$$

gdzie:

$d_{zz'}$ - odległość z-tej od $z^{'}$-tej grupy.

W kolejnym kroku, obiekty o najmniejszej odległości między sobą łączone są w jedną grupę, dzięki czemu liczba grup zmniejsza się o jeden. Zostaje rozpoczęty proces tworzenia drzewka połączeń. Ponownie badane są odległości między nowo stworzoną grupą a pozostałymi grupami. Proces trwa do momentu stworzenia pełnego drzewka połączeń, tj. jednej grupy.

Ogólna postać wzoru służącego do wyznaczenia odległości nowo powstałej grupy $\mathrm{G_{z''}}$, (powstałej dzięki połączeniu grup $\mathrm{G_z}$ i $\mathrm{G_{z'}}$), od grup które zostały $G_z{}'''$ to:

$$d_{z'''z''} = \alpha_z d_{z'''z} + \alpha_{z'} d_{z'''z'} + \beta d_{zz'} + \gamma |d_{z'''z} - d_{z'''z'}|$$

gdzie: $\alpha_z, \alpha_{z'}, \beta, \gamma$ - współczynniki przekształceń, różne dla poszczególnych metod aglomeracyjnych

Możemy wyróżnić sześć różnych metod aglomeracyjnych, różniących się sposobem wyznaczenia odległości między grupami obiektów. Poniżej zostaną podane współczynniki przekształceń dla każdej z nich, a w dalszej części pracy wybrane z nich zostaną szczegółowej omówione.

- metoda najbliższego sąsiedztwa (metoda pojedynczego wiązania)
  parametry przekształceń $\alpha_z = 0,5 \quad \alpha_{z'} = 0,5 \quad \beta = 0 \quad \gamma = 0,5.$
- metoda najdalszego sąsiedztwa (metoda pełnego wiązania)
  parametry przekształceń $\alpha_z = 0,5 \quad \alpha_{z'} = 0,5 \quad \beta = 0 \quad \gamma = -0,5.$
- metoda średniej międzygrupowej (metoda średnich połączeń)
  parametry przekształceń $\alpha_z = \frac{n_z}{n_z+n_{z'}} \quad \alpha_{z'} = \frac{n_{z'}}{n_z+n_{z'}} \quad \beta = 0 \quad \gamma = 0).$
- metoda mediany
  parametry przekształceń $\alpha_z = 0,5 \quad \alpha_{z'} = 0,5 \quad \beta = -025 \quad \gamma = 0.$
- metoda środka ciężkości
  parametry przekształceń $\alpha_z = \frac{n_z}{n_z+n_{z'}}; \alpha_{z'} = \frac{n_{z'}}{n_z+n_{z'}} \quad \beta = \frac{-n_z n_{z'}}{(n_z+n_{z'})^2} \quad \gamma = 0.$
- metoda Warda
  parametry przekształceń $\alpha_z = \frac{n_z+n_{z'''}}{n_z+n_{z'}+n_{z'''}} \quad \alpha_{z'} = \frac{n_{z'}+n_{z'''}}{n_z+n_{z'}+n_{z'''}} \quad \beta = \frac{-n_{z'''}}{n_z+n_{z'}+n_{z'''}} \quad \gamma = 0.$

Metoda najbliższego sąsiedztwa

W metodzie tej odległość między dwoma grupami obiektów jest równa odległości pomiędzy najbliższymi obiektami(sąsiadami), które należą do dwóch różnych grup. Odległość ta opisana jest wzorem:

$$d_{zz'} = \min_{ik} d_{ik}(\mathbf{O_i} \in \mathbf{G_z}, \mathbf{O_k} \in \mathbf{G_{z'}}),$$

$$i = 1, 2, \ldots, n_z, \quad k = 1, 2, \ldots, n_{z'}, \quad z, z^{'} = 1, 2, \ldots, n, \quad z \neq z^{'},$$

gdzie:

$$\mathbf{O_i} = [n_{ij}], \quad j = 1, 2, \ldots, m.$$

Metoda najdalszego sąsiedztwa

W metodzie tej odległość między dwoma grupami obiektów jest równa odległości pomiędzy najdalszymi obiektami(sąsiadami), które należą do dwóch różnych grup. Odległość ta opisana jest wzorem:

$$d_{zz'} = \max_{ik} d_{ik}(\mathbf{O_i} \in \mathbf{G_z}, \mathbf{O_k} \in \mathbf{G_{z'}}),$$

$$i = 1, 2, \ldots, n_z, \quad k = 1, 2, \ldots, n_{z'}, \quad z, z^{'} = 1, 2, \ldots, n, \quad z \neq z^{'},$$

Metoda średniej międzygrupowej

W metodzie tej odległość między dwoma grupami obiektów równa jest średniej arytmetycznej odległości między wszystkimi parami obiektów należących do dwóch różnych grup. Odległość ta opisana jest wzorem:

$$d_{zz'} = \frac{1}{n_z n_{z'}} \sum_{k=1}^{n_{z'}} \sum_{i=1}^{n_z} d_{ik}(\mathbf{O_i} \in \mathbf{G_z}, \mathbf{O_k} \in \mathbf{G_{z'}})$$

$$z, z^{'} = 1, 2, \ldots, n, \quad z \neq z^{'}.$$

Metoda mediany

W metodzie tej odległość między grupami obiektów jest równa medianie odległości pomiędzy wszystkimi parami obiektów należących do dwóch grup. Odległość ta opisana jest wzorem:

$$d_{zz'} = \mathbf{med}_{i,k}\{d_{ik}(\mathbf{O_i} \in \mathbf{G_z}, \mathbf{O_k} \in \mathbf{G_{z'}})\},$$

$$i = 1, 2, \ldots, n_z, \quad k = 1, 2, \ldots, n_{z'}, \quad z, z^{'} = 1, 2, \ldots, n, \quad z \neq z^{'}.$$

Rozdział 4

Zastosowanie wybranych metod porządkowania danych wielowymiarowych

4.1 Opis zbioru

Zbiór danych jest opracowaniem własnym, na podstawie ofert sprzedaży samochodów osobowych, zamieszczonych na portalu *www.otomoto.pl* w okresie listopad - grudzień 2017 roku. Zebrane dane dotyczą szczegółowych informacji odnośnie samochodu, tj. jego marki, modelu, wersji, typu, koloru lakieru, pojemności silnika, roku produkcji, przebiegu, liczby drzwi, rodzaju skrzyni biegu, rodzaju paliwa, rodzaju napędu, wyposażenia w: ABS, komputer pokładowy, ESP, klimatyzację. Oprócz danych ściśle związanych z budową i wyposażeniem samochodu, pojawiły się również atrybuty, tj. cechy umieszczone w kolumnach, związane z informacją o tym czy auto jest uszkodzone oraz bezwypadkowe,czy jest sprowadzane, jaki jest kraj aktualnej rejestracji, czy było serwisowane, czy sprzedający jest pierwszym właścicielem. Dodatkowo oprócz powyższych, został dodany atrybut najbardziej interesujący kupującego - czyli cena oraz województwo tj. miejsce skąd wystawiana została oferta. Zbiór został dołączony do pracy na płycie.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	MARKA	MODEL	WERSJA	TYP	WOJEWC	CENA.NET	CENA.BRU	MOC[km]	POJEMNC	ROK.PRO	PRZEBIEC	KOLOR	L.DZRZWI	RODZAJ.F	SKRZYNIA	NAPED	KRAJ.AKT
2	Hyundai	i20	II	kompakt	malopolskie		46500	90	1396	2016	18300	bialy	3	diesel	manualna	na przedni	Polska
3	Hyundai	i20	I	kompakt	mazowieckie		22900	86	1248	2013	319000	niebieski	5	benzyna+L	manualna	na przedni	Polska
4	Subaru	Legacy	V	kombi	mazowieckie		36900	150	1998	2010	149000	srebrny-m	5	benzyna	manualna	4x4(stały)	Polska
5	Ford	Mondeo	Mk4	sedan	dolnoslaskie		30000	146	1999	2008	166290	czarny	5	benzyna+L	manualna	na przedni	Polska
6	Opel	Astra	G	kompakt	slaskie		3990	136	1998	1998	230000	czarny	5	benzyna	manualna	na przedni	Polska
7	Mazda	Premacy		minivan	dolnoslaskie		7750	100	1998	2004	210563	srebry	5	diesel	manualna	na przedni	Niemcy
8	Seat	Leon	II	kompakt	dolnoslaskie		19900	200	1984	2006	198000	czarny	5	benzyna	manualna	na przedni	Szwajcaria
9	Volkswage	Passat	B6	sedan	lodzkie		13999	105	1900	2005	202749	czarny	5	diesel	manualna	na przedni	Polska
10	Opel	Zafira	A	minivan	lodzkie		7900	125	1800	2001	196000	srebrny	5	benzyna	manualna	4x4(stały)	Niemcy
11	Mercedes-	Klasa A	W168	kompakt	lodzkie		6500	102	1598	2002	200000	niebieski	5	diesel	manualna	na przedni	Polska
12	Skoda	Octavia	II	kombi	malopolskie		19500	140	1986	2008	220000	czarny	5	diesel	manualna	na przedni	Polska
13	Seat	Toledo	II	kompakt	wielkopolskie		11300	105	1598	2003	174600	szary	4	benzyna	manualna	na przedni	Niemcy
14	Peugeot	508		kombi	slaskie		42500	115	1560	2014	156800	bialy	5	diesel	manualna	na przedni	Polska
15	Skoda	Octavia	III	kombi	wielkopolskie		50900	105	1598	2015	91800	bialy	5	diesel	manualna	na przedni	Polska
16	Peugeot	Partner	I	kombi	lodzkie		7990	90	2000	2004	275763	zloty	5	diesel	manualna	na przedni	Polska
17	Porsche	Cayman		coupe	wielkopolsl	95900	117957	300	3400	2006	83000	srebrny	3	benzyna	automatyc	na tylne ko	Polska
18	Toyota	Auris	II	kompakt	mazowieckie		46000	132	1598	2014	46000	bialy-metal	5	benzyna	manualna	na przedni	Polska
19	Toyota	Auris	II	kompakt	dolnoslaskie		30300	99	1329	2014	151783	bialy	5	benzyna	manualna	na przedni	Polska
20	Mazda	3	II	kompakt	zachodniopomorskie		25200	105	1598	2009	135800	czarny	5	benzyna	manualna	na przedni	Austria
21	Mazda	3	II	kombi	malopolskie		30600	150	1999	2009	116000	brazowy	5	benzyna	manualna	na przedni	Niemcy
22	Mazda	6	I	kombi	dolnoslaskie		15700	146	2000	2007	190000	czarny	5	diesel	manualna	na przedni	Polska
23	Mazda	6	I	kompakt	lodzkie		17900	147	2000	2006	238482	szary	5	benzyna+L	manualna	na przedni	Polska
24	Mazda	6	I	kombi	kujawsko-pomorskie		11000	121	1998	2005	204000	srebrny	5	diesel	manualna	na przedni	Polska
25	Citroen	C4	II	kompakt	wielkopolskie		36200	92	1560	2014	86561	bialy	5	diesel	manualna	na przedni	Polska
26	Citroen	C4	II	kompakt	malopolskie		36997	90	1397	2014	33000	bialy	5	benzyna	manualna	na przedni	Polska
27	Citroen	C4	II	kompakt	lodzkie		16900	90	1397	2014	35000	szary	5	benzyna	manualna	na przedni	Francja

Rysunek 4.1: Podgląd stworzonego zbioru

Użyte zmienne

W stworzonym zbiorze danych znajduje się 29 atrybutów, opisujących 61 różnych rekordów, tj. obiektów reprezentowanych przez wiersze, którym przypisano pewne wartości atrybutów. Wśród zebranych danych można wyróżnić zarówno zmienne jakościowe, jak i ilościowe.

Zmiennymi jakościowymi są atrybuty:

- marka,
- model,
- typ,
- wojewodztwo,
- kolor,
- rodzaj.paliwa,
- skrzynia.biegow,
- naped,
- kraj.aktualnej.rejestracji,
- kraj.pochodzenia,
- stan,
- ABS,
- uszkodzony,
- pierwszy.wlasciciel,
- kto.sprzedaje,
- serwisowany,
- komputer.pokladowy,
- ESP,
- klimatyzacja,
- bezwypadkowy,
- status.pojazdu.sprowadzanego

Zmiennymi ilościowymi są atrybuty:

- cena.netto,
- cena.brutto,
- moc,
- pojemnosc.skokowa,
- rok.produkcji,
- przebieg,
- l.drzwi.

4.2 Użyte programy

Zbiór danych został umieszczony w programie Excel, z kolei implementacje zostały stworzone w języku R, który to ma zastosowanie w statystyce i ekonometrii.

4.3 Metoda ...

4.4 Wstęp

Zostanie tutaj zaprezentowane zastosowanie nieliniowego porządkowania danych przy pomocy istniejących funkcji biblioteki cluster, w której to funkcja agnes umożliwiająca uporządkowanie zbioru po wyborze odpowiedniej metody aglomeracyjnej. Mamy tu do wyboru metody: single - metoda najbliższego sąsiedztwa, complete - metoda najdalszego sąsiedztwa, ward - metoda Warda, average - metoda średniej między grupowej. Poniżej zostanie zaprezentowane zastosowanie metod single oraz complete wraz z porównaniem wyników porządkowania.

4.4.1 Import danych

Na początku należy zaimportować dane, które chcemy poddać porządkowaniu. W tym celu należy zaimportować bibloteke readxl - gdyż dane pobieramy z excela, a w kolejnym kroku wywołujemy plik, podając naszą ścieżkę pliku z rozszerzeniem xlsx. My użyjemy tutaj zbioru zawierającego 8 obiektów, będących ofertami sprzedaży aut.

```
library(readxl)
zbior_danych <- read_excel("datasets/zbior_danych.xlsx",
                           sheet = "DANE_INNA_WERSJA")
```

Podgląd danych:

```
head(zbior_danych)
```

```
## # A tibble: 6 x 30
##      Nr MARKA   MODEL   WERSJA TYP     WOJEWODZTWO  `CENA.NETTO_[pln]`
##   <dbl> <chr>   <chr>   <chr>  <chr>   <chr>                     <dbl>
## 1  1.00 Hyundai i20     II     kompakt malopolskie                  NA
## 2  2.00 Hyundai i20     I      kompakt mazowieckie                  NA
## 3  3.00 Subaru  Legacy  V      kombi   mazowieckie                  NA
## 4  4.00 Ford    Mondeo  Mk4    sedan   dolnoslaskie                 NA
## 5  5.00 Opel    Astra   G      kompakt slaskie                      NA
## 6  6.00 Mazda   Premacy <NA>   minivan dolnoslaskie                 NA
## # ... with 23 more variables: `CENA.BRUTTO_[pln]` <dbl>, `MOC_[km]` <dbl>,
## #   `POJEMNOSC.SKOKOWA_[cm3]` <dbl>, ROK.PRODUKCJI <dbl>, `PRZEBIEG_[km]`
## #   <dbl>, KOLOR <chr>, L.DZRZWI <dbl>, RODZAJ.PALIWA <chr>,
## #   SKRZYNIA.BIEGOW <chr>, NAPED <chr>, KRAJ.AKTUALNEJ.REJESTRACJI <chr>,
## #   KRAJ.POCHODZENIA <chr>, STATUS.POJAZDU.SPROWADZONEGO <chr>,
## #   PIERWSZY.WLASCICIEL <dbl>, KTO.SPRZEDAJE <chr>, STAN <chr>,
## #   SERWISOWANY <dbl>, ABS <dbl>, KOMPUTER.POKLADOWY <dbl>, ESP <dbl>,
## #   KLIMATYZAJCA <dbl>, BEZWYPADKOWY <dbl>, USZKODZONY <dbl>
```

4.4.2 Podzbiór danych

W kolejnym, kroku po przyjrzeniu się zbiorowi danych, użytkownik musi zadecydować na których danych ilościowych chce pracować - ważna jest znajomość danych. Dodatkowo pierwszą kolumną musi być kolumna zawierająca numery indeksów obiektów, ze względu na to, że w wyniku zastosowania funkcji odpowiedzialnej za porządkowanie, zostaną zwrócone w kolejności malejącej numery indeksów, mówiące o kolejności uporządkowania. W związku z tym, za pomocą poniższej procedury użytkownik tworzy podzbiór zaimportowanego zbioru, gdzie w miejsce "" wpisuje nazwy kolumn zawierających zmienne ilościowe, wybrane do porządkowania(przyjmijmy założenie, że podzbiór będzie nazywał się dane_porzadkowanie - będzie to pomocne w dalszej części programu). U mnie wybranymi kolumnami są: cena, moc, pojemność, rok produkcji, przebieg.

```
dane_porzadkowanie<-zbior_danych[c("Nr","CENA.BRUTTO_[pln]","MOC_[km]",
                                   "POJEMNOSC.SKOKOWA_[cm3]",
                                  "ROK.PRODUKCJI","PRZEBIEG_[km]")]
```

4.4.3 Transformacje danych

Przed samym porządkowaniem, wymaganym jest aby zmienne miały charakter stymulant oraz by zostały poddane transformacji normalizacyjnej. Aby funkcja dokonująca porządkowania dawała poprawny wynik, użytkownik musi zająć się tranformacją przed jej zastosowaniem. Poniżej podałam tego przykład. Dla zmiennych które stymulantami nie są, należy dokonać stymulacji.

Wsród moich zmiennych poddanych porządkowaniu, do stymulant nie należy zmienna zmienna: przebieg - jest destymulantą, w związku z tym, została przekształcona na stymulantę, za pomocą przekształcenia ilorazowego.

```
stymulacja_przeksztalcenie_ilorazowe<-function(x,y){
  for (i in 1:nrow(x)){
    x[i,which(colnames(x)==y)]=1/x[i,which(colnames(x)==y)]
  }
  return(x)
}
```

Gdy użytkownik chce skorzystać z tej funkcji, w miejsce x musi wpisać nazwę zbioru, a w miejsce y nazwę kolumny w "", którą chce poddać stymulacji. UWAGA - kolumny wymagające stymulacji, muszą zostać osobno poddane działaniu poniższej funkcji, dodatkowo po każdym zastosowaniu funkcji, należy nadpisać zbiór by zmiany zostały zapisane.

```
dane_porzadkowanie<-stymulacja_przeksztalcenie_ilorazowe(dane_porzadkowanie,"PRZEBIEG_[k
```

W celu uzyskania porównywalności między zmiennymi, zostały one poddane transformacji normalizacyjnej - unitaryzacja

```
unitaryzacja<-function(x){
  maksi=0
  minim=0
  for (j in 2:ncol(x)){
    maksi[j]=max(x[j])
    minim[j]=min(x[j])
    for (i in 1:nrow(x)){
      x[i,j]=(x[i,j]-minim[j])/(maksi[j]-minim[j])
    }
  }
  return(x)
}

dane_porzadkowanie<-unitaryzacja(dane_porzadkowanie)
```

4.5 Nieliniowe porządkowanie przy pomocy metod aglomeracyjnych

Chcąc zastosować funkcję agnes, należy w pierwszej kolejności wyznaczyć macierz odległości pomiędzy wszystkimi parami obiektów. Do wyznaczenia odległości zostanie użyta metryka euklidesowa - w tym celu zostanie wykorzystania funkcja dist(x, method="") - w miejsce x należy wpisać nazwę tabeli zawierających dane do uporządkowania, a w nawiasie[,] na miejscu drugiej współrzędnej należy podać wektor kolumn, na podstawie którego wartości zostanie wyznaczona macierz odległości. W miejsce argumentu method należy wpisać nazwę metryki na podstawie której zostana obliczona odległość - w naszym przypadku będzie to euclidean - euklidesowa.

```
odleglosci <- dist(dane_porzadkowanie[,c("CENA.BRUTTO_[pln]","MOC_[km]",
            "POJEMNOSC.SKOKOWA_[cm3]","ROK.PRODUKCJI","PRZEBIEG_[km]")],
```

Następnie należy zaimportować bibliotekę cluster, by móc skorzystać z metod aglomerycyjnych. Jak już zostało wspomnanie we wstępnie, wywołanie metod aglomeracyjnych odbywa się dzięki funkcji agnes(x, method=“”) . W miejsce argumentu x - zostanie podana wyznaczona macierz odległości z kolei, w kolejnym argumencie - method zostanie podana reguła wyznaczania odległości pomiędzy nową grupą a pozostałymi obiektów. Regułą tą może być metoda najbliższego sąsiedztwa, najdalszego, Warda lub średniej między grupowej. My wykorzytsamy metodę najbliższego sąsiedztwa oraz najdalszego.

```
library(cluster)
metoda_najblizszego <- agnes(odleglosci, method = "single")
metoda_najdalszego<- agnes(odleglosci, method = "complete")
```

Ostatni etap to graficzne zaprezentowanie wyniku w postaci dendrogramu. W tym celu należy zaimportować bibliotekę factoextra, w której to jest funkcja fviz_dend(x, main = “”). W miejsce argumentu x należy wpisać nazwę obiektu powstałego przy pomocy funkcji agnes, main to tytuł wykresu. Dodatkowo na osi pionowej zaprezentowane są odległości między obiektami, z kolei na osi poziomej znajdują się numery indeksów obiektów.

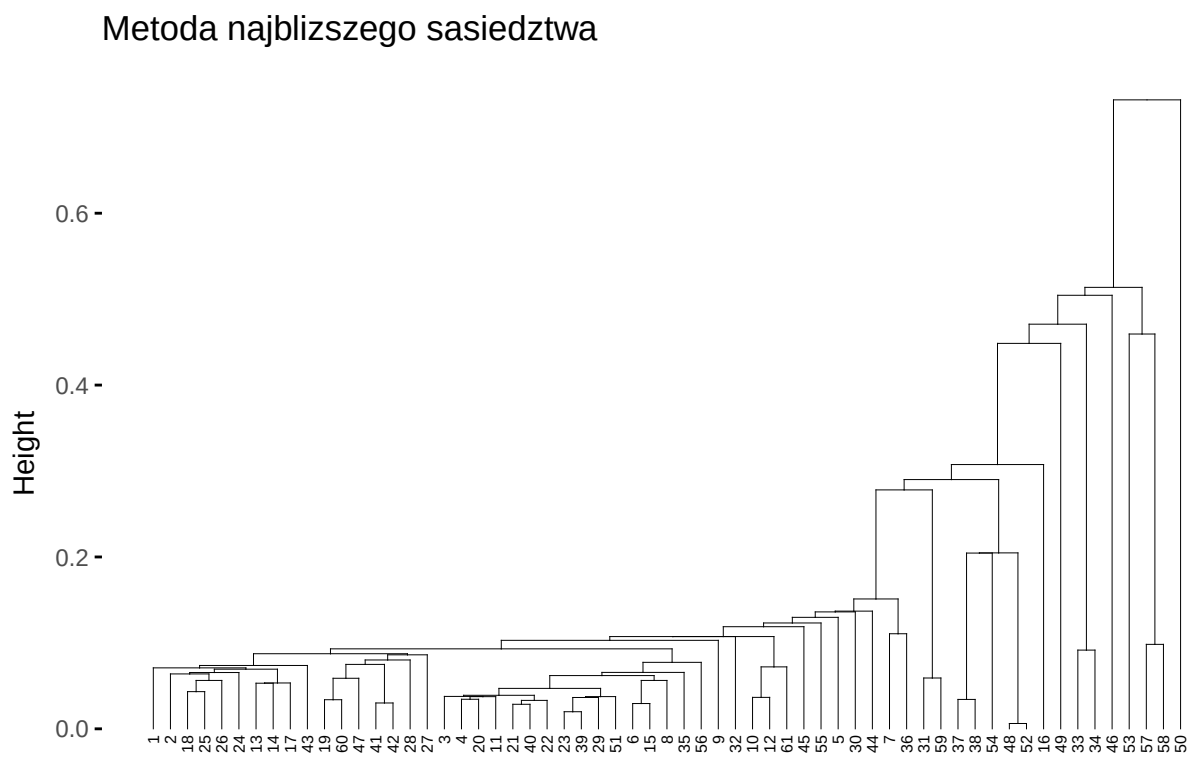
Dendrogram dla metody najbliższego sąsiada:

```
library(factoextra)
```

```
## Loading required package: ggplot2
```

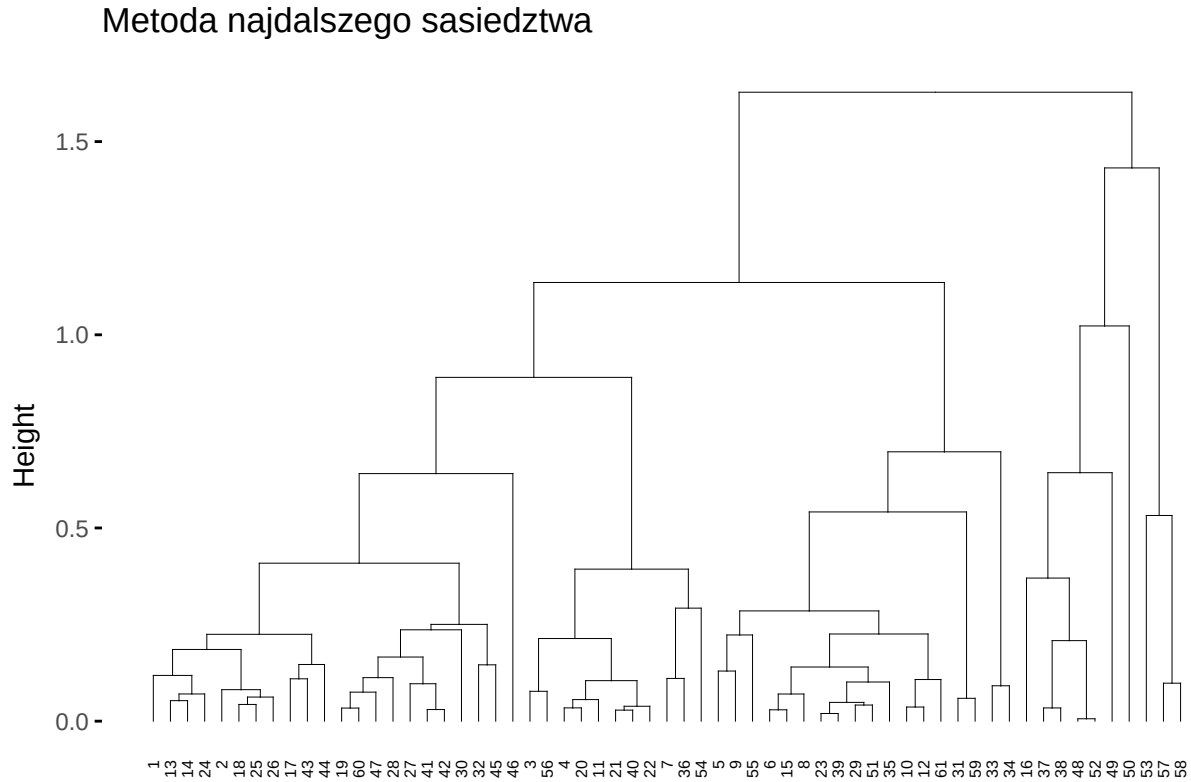
```
## Welcome! Related Books: ‘Practical Guide To Cluster Analysis in R‘ at https://goo.gl/
```

```
fviz_dend(metoda_najblizszego, lwd=0.1, cex=0.45, main = "Metoda najbliższego sąsiedztwa
```



Dendrogram dla metody najdalszego sąsiada:

```
fviz_dend(metoda_najdalszego, lwd=0.1, cex=0.45, main = "Metoda najdalszego sąsiedztwa")
```



4.5.1 Porównanie wyników uporządkowania

W przypadku dendrogramu uzyskanego metodą najdalszego sąsiedztwa można zauważyć większą odległość wiązań niż dla metody najbliższego sąsiedztwa, które to obrazują odległość między grupami. Dodatkowo obiekty charakteryzujące się najbardziej korzystnym wartości zmiennych, zostały pogrupowane w jedną grupę (tu mowa o obiektach o numerach indeksów 49, 50, 53, 57, 58), natomiast w przypadku metody najbliższego sąsiedztwa obiekty te zostały rozdzielone w pojedyncze grupy. Można również zauważyć, że w przypadku uporządkowania metodą najdalszego sąsiedztwa obiekty tworzą pewnego rodzaju pogrupy, skupiska, z kolei w przypadku uporządkowania metodą najbliższego sąsiada, wynik porządkowania można porównać do wyglądu lawiny, lub góry, u której podnóża znajdują się obiekty o najniższych wartościach opisywanych je zmiennych, które to łączą się i dochodzą do szczytu, który stanowią obiekty, o najwyższych wartościach opisywanych je zmiennych.

Bibliografia

[1] Grzegorz Banaszak and Wojciech Gajda. *Elementy algebry liniowej (część 1)*. Wydawnictwo Naukowo-Techniczne, Warszawa, 2002.

[2] Jarosław Bartoszewicz. *Wykłady ze statystyki matematycznej*. Państwowe Wydawnictwo Naukowe, Warszawa, 1996.

[3] Aleksander Błaszczyk and Sławomir Turek. *Teoria mnogości*. Państwowe Wydawnictwo Naukowe, Warszawa, 2007.

[4] Patric Billingsley. *Prawdopodobieństwo i miara*. Państwowe Wydawnictwo Naukowe, Warszawa, 1987.

[5] Jacek Jakubowski and Rafał Sztencel. *Wstęp do teorii prawdopodobieństwa*. SCRIPT, Warszawa, 2004.

[6] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, and M. Wasilewski. *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach: część I. rachunek pradopodobieństwa*. Państwowe Wydawnictwo Naukowe, Warszawa, 1999.

[7] Kazimierz Kuratowski. *Wstęp do teorii mnogości i topologii*. Państwowe Wydawnictwo Naukowe, Warszawa, 2004.

[8] Andrzej Młodak. *Analiza taksonomiczna w statystyce regionalnej*. Centrum Doradztwa i Informacji Difin, Warszawa, 2006.

[9] Tomasz Panek and Jan Karol Zwierzchowski. *Statystyczne metody wielowymiarowej analizy porównawczej: teoria i zastosowania*. Oficyna Wydawnicza, Szkoła Główna Handlowa, Warszawa, 2013.

[10] Ryszard Rudnicki. *Wykłady z analizy matematycznej*. Państwowe Wydawnictwo Naukowe, Warszawa, 2006.

[11] Arkadiusz Sołtysiak and Piotr Jaskulski. *http://www.antropologia.uw.edu.pl/MaCzek/maczek.html*.

[12] Robin J. Wilson. *Wprowadzenie do teorii grafów*. Państwowe Wydawnictwo Naukowe, Warszawa, 2008.