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

Rozdział 2

Preliminaria

2.1 Notacja

- R^m - przestrzeń liniowa, wektorowa, jej elementy nazywamy zamiennie wektorami lub punktami
- $\mathcal{E}^n$ - przestrzeń euklidesowa n-wymiarowa
- O = $\{O_1, O_2, ..., O_n\}$ - zbiór obiektów przestrzennych
- X = $\{X_1, X_2, ..., X_m\}$ - zbiór zmiennych (cech)
- T = $\{T_1, T_2, ..., T_k\}$ - zbiór okresów (jednostek czasu)
- OX = O $\cdot$ X - zbiór obiekto-zmiennych
- OT = O $\cdot$ T - zbiór obiekto-okresów
- XT = X $\cdot$ T - zbiór zmienno-okresów
- OXT = O $\cdot$ X $\cdot$ T - zbiór obiekto-zmienno-okresów
- Ω - przestrzeń zdarzeń elementarnych
- ω - zdarzenie elementarne
- $\mathcal{F}$ - rodzina podzbiorów zbioru Ω
- med(X_j) - mediana cechy X_j
- ρ - relacja porządkująca
- G - graf prosty
- $V(G)$ - zbiór wierzchołków grafu G
- $E(G)$ - krawędzie grafu G
- D - graf skierowany(digraf)
- $V(D)$ - zbiór wierzchołków digrafu D
- $A(D)$ - rodzina łuków digrafu D

2.2 Słownik użytych pojęć

W pracy zostały wykorzystane następujące pojęcia, których wytłumaczenie znajduje się poniżej.

- Statystyka matematyczna [4, Rozdział 1]
  Statystyka matematyczna zajmuje się metodami wnioskowania o całej zbiorowości statystycznej na podstawie zbadania pewnej jej części zwanej próbką lub próbą.

- Cecha statystyczna [7, Rozdział 1]
  Cecha statystyczna jest to liczbowy opis przedmiotu dociekań tj. konkretnej dziedziny życia społeczno-gospodarczego.Służy ona do scharakteryzowania podmiotu badania.

- Macierz obserwacji ([7, Rozdział 2]
  Niech $m > 1$ oraz $n > 1$ będą liczbami naturalnymi. Macierzą obserwacji nazywamy macierz rozmiaru $n \times m$ postaci

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

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

- Skala porządkowa [8, Rozdział 1.2]
  Skalą porządkową nazywana jest skala, pozwalająca na stwierdzeniu identyczności lub różnic porównywanych obiektów, a także na porównywanie wariantów zmiennych zaobserwowanych w obiektach. Nie pozwala ona określić odległości między obiektami. Umożliwia zliczanie obiektów uporządkowanych(liczby relacji równości, nierówności, większości i mniejszości).

- Skala przedziałowa [8, Rozdział 1.2]
  Skalą przedziałową nazywana jest skala, która w stosunku do skali porządkowej, pozwala obliczyć odległość między obiektami, dokonując pomiaru zmiennych za pomocą liczb rzeczywistych. Dla skali tej możliwe jest, obok operacji arytmetycznych dopuszczalnych dla skal o mniejszej mocy, także dodawanie i odejmowanie. Wartość zerowa na tej skali ma charakter umowny, co prowadzi do zachowania różnic między wartościami cechy, przy zmianie jednostek miary.

- Skala ilorazowa [8, Rozdział 1.2]
  Skalą ilorazową nazywana jest skala, która jest podobna do skali przedziałowej(odwołanie), z tym że występuje w niej zero bezwzględne(zero ogranicza lewostronnie zakres tej skali). Powoduje to, że można na tej skali obok operacji dopuszczalnych na skalach słabszych dokonywać także dzielenia i mnożenia, a tym samym przedstawiać dowolną wartość cechy danego obiektu jako wielokrotność wartości cechy dla innego obiektu.

- Zmienna objaśniająca [3, Rozdział 1.1] Zmienną objaśniającą nazywamy zmienną w modelu statystycznym, która oddziałuje na zmienne objaśniane. Zmiennie objaśniającą oznaczamy jako $X_1, ..., X_k$, z kolei zmienne objaśniane jako Y.

- Stymulanta [8, Rozdział 1.5]
  Stymulantami nazywane są zmienne, których wysokie wartości badany w badanych obiektach są pożądane z punktu widzenia rozpatrywanego zjawiska.

- Destymulanta [8, Rozdział 1.5]
  Destymulantami nazywane są zmienne, których wysokie wartości badany w badanych obiektach są niepożadane z punktu widzenia rozpatrywanego zjawiska.

- Nominanta [8, Rozdział 1.5]
  Nominantami nazywane są zmienne, których odchylenia wartości w badanym obiekcie od wartości (lub przedziału wartości) uznawanych za najkorzystniejsze są niepożądane z punktu widzenia rozpatrywanego zjawiska.

- Transformacja normalizacyjna [8, Rozdział 1.5]
  Transformacją zmiennych diagnostycznych, mających na celu ujednolicenie ich jednostek pomiarowych, przy zastosowaniu zmiennych diagnostycznych, nazywana jest transformacją normalizacyjną. Można ją przeprowadzić na zmiennych, opisujących porównywane obiekty, mierzonych na skali przedziałowej lub ilorazowej.

  Ogólny wzór na przekształcenie normalizacyjne(Borys, 1978; Grabiński i in., 1989):

$$z_{ij} = \left(\frac{x_{ij} - a}{b}\right)^p, i = 1, 2, ..., n; j = 1, 2, ..., m; b \neq 0, \tag{2.1}$$

  gdzie:
  z_{ij} - znormalizowana wartość j-tej zmiennej w i-tym obiekcie,
  a, b, p - parametry normalizacyjne.

- Stopień podobieństwa obiektów [8, Rozdział 1.6]
  Stopień podobieństwa obiektów, jest wielkość mówiąca o podobieństwie obiektów między sobą. Jest on mierzony za pomocą miar odległości lub tez miar bliskości(zgodności).

- Miara odległości [8, Rozdział 1.6]
  Miarą odległości pomiędzy obiektami: i-tym i i'-tym, nazywamy dowolną funkcję rzeczywista d, spełniającą następujące warunki:
  - **dodatniość** (odległość między różnymi obiektami jest zawsze dodatnia): $d_{ii'} > 0$
  - **symetryczność** (odległość i-tego obiektu od i'-tego obiektu jest taka sama, jak odległość i'-tego obiektu od obiektu i-tego: $d_{ii'} = d_{i'i}$
  - **zwrotność** (odległość obiektu od samego siebie jest równa zeru): $d_{ii} = 0$
  - **nierówność trójkąta**: odległość pomiędzy i-tym i i'-tym obiektem będzie nie większa niż odległość pośrednia pomiędzy tymi obiektami definiowana jako suma odległości pomiędzy obiektami i-tym i i''-tym oraz i'-tym i i''-tym): $d_{ii'} \leqslant d_{ii''} + d_{i''i'}$.

Uwaga: Wzrost wartości miary odległości oznacza zmniejszenie stopnia podobieństwa obiektów.

- Odległość euklidesowa dla znormalizowanych zmiennych [8, Rozdział 1.6]
  Wzór odległości euklidesowej, dla znormalizowanych zmiennych, jest następujący:

$$d_{ii'} = \Big[\sum_{j=1}^{m} |z_{ij} - z_{i'j}|^2\Big]^{\frac{1}{2}}$$

- Odległość miejska(Manhattan) dla znormalizowanych zmiennych [8, Rozdział 1.6]
  Wzór na odległość miejską(Manhattan) dla znormalizowanych zmiennych, jest następujący:

$$d_{ii'} = \sum_{j=1}^{m} |z_{ij} - z_{i'j}|.$$

- Miara bliskości [8, Rozdział 1.6]
  Miarą bliskości pomiędzy obiektami, nazywamy funkcję p spełniającą następujące warunki:
  - **dodatniość**: $p_{ii'} > 0$,
  - **symetryczność**: $p_{ii'} = p_{i'i}$
  - **zwrotność**: $p_{ii} = 1$.
  Uwaga: Wzrost wartości miary bliskości oznacza zwiększenie stopnia podobieństwa badanych obiektów.

- Macierz odległości [8, Rozdział 1.6]
  Macierzą odległości, nazywamy macierz unormowanych danych wejściowych, tj. macierz, której elementami są odległości między parami badanych obiektów. Macierz odległości jest postaci:

$$D = [d_{ii'}], i, i' = 1, 2, ..., n.$$

- Średnia arytmetyczna z próby [7, Rozdział 2.2]
  Średnią arytmetyczną wartości cechy X_j nazywamy wartość

$$\overline{x_j} = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

- Odchylenie standardowe z próby [7, Rozdział 2.2]
  Odchyleniem standardowym cechy X_j nazywamy wartość

$$s_j = \sqrt{s_j^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}$$

- Mediana [7, Rozdział 2.2]
  Medianę cechy X_j nazywamy wartość

$$\text{med}(X_j) = y = \begin{cases} \frac{1}{2}\left(x_{(\frac{n}{2})j} + x_{(\frac{n}{2}+1)j}\right) & \text{jeśli n jest parzyste} \\ \\ x_{(\frac{n+1}{2})j} & \text{jeśli n jest nieparzyste} \end{cases}$$

- Przestrzeń euklidesowa n-wymiarowa $\mathcal{E}^n$ [6, Rozdział 9]
  Przestrzeń euklidesowa n-wymiarowa, jest przestrzenią metryczną przy zwykłej definicji odległości punktu $x = (x_1, x_2, ..., x_n)$ od punktu $y = (y_1, y_2, ..., y_n)$, danej wzorem Pitagorasa

$$|x - y| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

  gdzie, x i y są ciągami złożonymi z n liczb rzeczywistych.

2.3 Podstawowe pojęcia rachunku prawdopodobieństwa oraz statystyki

Na potrzeby pracy, zostały wykorzystane pojęcia rachunku prawdopodobieństwa oraz statystyki, konieczne to 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. Poniższej zostały one wypisane.

Definicja 2.3.1. *Ciało zbiorów [9, Rozdział 8.1]*
Rodzinę A podzbiorów zbioru X nazywamy ciałem zbiorów, jeżeli spełnia ona następujące warunku:

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

Definicja 2.3.2. *σ-algebra/ciało zbiorów/ zbiorów mierzalnych[9, Rozdział 8.1]*
Ciało zbiorów $\mathcal{A}$ nazywamy σ-ciałem zbiorów, jeżeli spełnia ona warunek dla dowolnych zbiorów $A_n \in \mathcal{A}, n \in \mathbb{N}$, mamy

$$\bigcup_{i=1}^{\infty} A_n \in \mathcal{A}.$$

Elementy σ-ciała $\mathcal{A}$ nazywamy zbiorami mierzalnymi.

Definicja 2.3.3. *Przestrzeń zdarzeń elementarnych [5, 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 2.3.4. *Zdarzenie losowe [5, w oparciu o rozdział 1.1]*
Zdarzeniem losowym (zdarzeniem) nazywamy każdy zbiór $A \in \mathcal{F}$, gdzie $\mathcal{F}$ jest rodziną podzbiorów Ω spełniającą następujące warunki:

1. $\Omega \in F$;
2. *Jeśli* $A \in \mathcal{F}$, *to* $A' \in F$, *gdzie* $A' = \Omega \backslash A$ *jest zdarzeniem przeciwnym do zdarzenia* A;
3. *Jeśli* $A_i \in F$, $i = 1, 2, ...,$*to* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

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

Definicja 2.3.5. *Przestrzeń probabilistyczna [5, 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 F.

Definicja 2.3.6. *Przestrzeń mierzalna w n-wymiarowej przestrzeni euklidesowej [1, Rozdział 1]*
Niech (Ω, F, P) oznacza przestrzeń propabilistyczną. Przestrzenią mierzalną w n-wymiarowej przestrzeni euklidesowej R^n, nazywamy uporządkowaną dwójkę (R^n, B^n), gdzie B^n jest σ-ciałem podzbiorów borelowskich tej przestrzeni, $n \geqslant 1$.

Definicja 2.3.7. *Prawdopodobieństwo [5, 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. $P(A) \geqslant 0 \quad \forall A \in \mathcal{F}$
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 2.3.8. *Zmienna losowa [5, Rozdział 2.1]*
Niech $(\Omega, \mathcal{F}, P)$ będzie dowolną przestrzenią propabilistyczną. Dowolną funkcję $X : \Omega \to \mathbb{R}$ nazywamy zmienną losową jednowymiarową, jeśli dla dowolnej liczby rzeczywistej x zbiór zdarzeń elementarnych ω, dla których spełniona jest nierówność $X(\omega) < x$ jest zdarzeniem, czyli

$$\{\omega : X(\omega) < x\} \in F \text{ dla każdego } x \in \mathbb{R}$$

Definicja 2.3.9. *Funkcja mierzalna [9, w oparciu o rozdział 8.2]*
Niech X będzie niepustym zbiorem, A σ-ciałem na X i $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Funkcję $f : X \to \overline{\mathbb{R}}$ nazywamy mierzalną, jeżeli zbiór

$$\{x \in X : f(x) > a\}$$

jest mierzalny przy dowolnym $a \in \mathbb{R}$.

Definicja 2.3.10. *Wektor losowy n-wymiarowy [1, Rozdział 1]*
Wektorem losowym n-wymiarowym nazywamy funkcję $X : \Omega \to \mathbb{R}^n$ mierzalną względem σ-ciała $\mathcal{F}$ ($\mathcal{F}$-mierzalną), tzn. taką, że $X^{-1}(B) \in \mathcal{F}$ dla każdego $B \in \mathcal{F}$.

Definicja 2.3.11. *Wartość oczekiwana [5, Rozdział 2.6]*
Niech X będzie zmienną losową typu dyskretnego lub ciągłego. Wartością oczekiwaną zmiennej losowej X nazywamy

$$E(X) = \begin{cases} \sum_{i=1}^{n} x_i p_i & \text{jeśli zmienna jest typu dyskretnego} \\ \\ \int_{-\infty}^{\infty} x f(x) dx & \text{jeśli zmienna jest typu ciągłego} \end{cases}$$

Definicja 2.3.12. *Wartość oczekiwana macierzy losowej X [1, Rozdział 1.3]*
Wartością oczekiwaną macierzy losowej X nazywamy macierz postaci

$$E(X) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & ... & E(X_{1r}) \\ E(X_{21}) & E(X_{22}) & ... & E(X_{2r}) \\ ... & ... & ... & ... \\ E(X_{n1}) & E(X_{n2}) & ... & E(X_{nr}) \end{bmatrix}$$

przy założeniu, że wszystkie wartości oczekiwane $E(X_{ij})$, $i = 1, 2, ..., n, j = 1, 2, ..., r$, *istnieją.*

Definicja 2.3.13. *Macierz kowariancji n-wymiarowego wektora losowego X [1, Rozdział 1]*
Macierzą kowariancji n-wymiarowego wektora losowego X nazywamy macierz

$$\sum = E\{[X - E(X)][X - E(X)]'\}$$

Definicja 2.3.14. *Kowariancja [1, Rozdział 1]*
Niech X_i *i* X_j *będą zmiennymi losowymi,* $\sum$ *będzie macierzą kowariancji n-wymiarowego wektora losowego X. Kowariancją zmiennych losowych* X_i *i* X_j*, nazywamy*

$$cov(X_i, X_j) = \sigma_{ij} = E\{[X_i - E(X_i)][X_j - E(X_j)]\},\ i, j = 1, 2, ..., n$$

gdzie σ_{ij} *jest elementem macierzy kowariancji n-wymiarowego wektora losowego X.*

Definicja 2.3.15. *Współczynnik Pearsona [7, Rozdział 2.2]*
Współczynnik Pearsona oznaczamy:

$$r_{jk} = \frac{cov(X_j, X_k)}{s_j s_k}$$

gdzie:
$cov(X_j, X_k)$ *- kowariancja cech* X_j *i* X_k *.*

Definicja 2.3.16. *Macierz korelacji par zmiennych [7, Rozdział 2.2]*
Macierzą korelacji par zmiennych, nazywamy macierz postaci:

$$R = \begin{bmatrix} 1 & r_{12} & ... & r_{1m} \\ r_{21} & 1 & ... & r_{2m} \\ ... & ... & ... & ... \\ r_{m1} & r_{m2} & ... & 1 \end{bmatrix}$$

gdzie:
r_{jk} *- współczynnik korelacji liniowej Pearsona j-tej i k-tej cechy (czyli* X_j *oraz* X_k*) .*

2.4 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 wprowadzeniu definicji, należy wcześniej podać niezbędne pojęcia dotyczące grafów. Niezmiernie istotne jest podanie pojęć dotyczących grafów, dzięki którym później zostaną wprowadzone definicje.

Poniższe pojęcia zostały opracowane na podstawie [10]

Definicja 2.4.1. *Graf prosty [10, Rozdział 2]*
Niech G będzie grafem prostym, tj. grafem składającym się z niepustego zbioru skończonego $V(G)$, którego elementy nazywamy wierzchołkami (lub węzłami), i skończonego zbioru $E(G)$ różnych par nieuporządkowanych różnych elementów zbioru $V(G)$, które nazywamy krawędziami. Zbiór $V(G)$ nazywamy zbiorem wierzchołków, a zbiór $E(G)$ $\Big(E(G) \subseteq \{\{u,v\} : u,v \in V, u \neq v\}\Big)$ zbiorem krawędzi grafu G.
Mówimy, że krawędź $\{v,w\}$ łączy wierzchołki v i w, i na ogół oznaczamy ją krócej symbolem vw.

Definicja 2.4.2. *Pętle [10, Rozdział 2]*
Pętlami nazywamy krawędzie wielokrotne, łączące wierzchołek z samym sobą.

Definicja 2.4.3. *Graf/graf ogólny [10, Rozdział 2]*
Grafem nazywamy obiekt, w którym występują krawędzie wielokrotne oraz pętle.

Definicja 2.4.4. *Trasa/marszruta [10, Rozdział 3]*
Trasą (lub marszrutą) w danym grafie G nazywamy skończony ciąg krawędzi postaci $v_0v_1, v_1v_2, ..., v_{m-1}v_m$, zapisywany również w postaci $v_0 \to v_1 \to v_2 \to ... \to v_m$, w którym każde dwie kolejne krawędzie są albo sąsiednie, albo identyczne. Taka trasa wyznacza ciąg wierzchołków $v_0, v_1, ..., v_m$. 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 2.4.5. *Ścieżka [10, Rozdział 3]*
Trasą, w której wszystkie krawędzie są różne, nazywamy ścieżką.

Definicja 2.4.6. *Droga [10, Rozdział 3]*
Ścieżkę, w której wierzchołki $v_0, v_1, ..., v_m$ są różne (z wyjątkiem, być może, równości $v_0 = v_m$), nazywamy drogą.

Definicja 2.4.7. *Droga zamknięta/ścieżka zamknięta [10, Rozdział 3]*
Droga lub ścieżka jest zamknięta, jeśli $v_0 = v_m$.

Definicja 2.4.8. *Cykl [10, Rozdział 3]*
Ścieżką zamkniętą zawierającą co najmniej jedną krawędź nazywamy cyklem.

Definicja 2.4.9. *Graf spójny [10, Rozdział 3]*
Graf jest spójny wtedy i tylko wtedy, gdy każda para wierzchołków jest połączona drogą.

Definicja 2.4.10. *Dendryt [8, Rozdział 2.3]*
Graf spójny i otwarty nazywany jest dendrytem.

Definicja 2.4.11. *Drzewo [10, Rozdział 4]*
Drzewem nazywamy graf spójny, nie zawierający cykli.

Definicja 2.4.12. *Graf skierowany(digraf albo graf zorientowany) [10, Rozdział 7]*
Graf skierowany lub digraf D, skała się z niepustego zbioru skończonego $V(D)$ elementów nazywanych wierzchołkami i skończonej rodziny $A(D)$ par uporządkowanych elementów zbioru $V(D)$, nazywanych łukami. Zbiór $V(D)$ nazywamy zbiorem wierzchołków, a rodzinę $A(D)$ rodziną łuków digrafu D. Łuk (v, w) zwykle zapisujemy jako vw. Graf skierowany oznaczamy zwykle w postaci pary uporządkowanej $G =< V, E >$

UWAGA
Każdy graf jednoznacznie wyznacza pewną relację binarną w zbiorze V. Można również powiedzieć odwrotnie, że każda relacja binarna r w zbiorze V, wyznacza jednoznacznie graf zorientowany, którego węzłami są elementy zbioru V, z kolei krawędziami są uporządkowanie pary (v, v'), należące do r.

Definicja 2.4.13. *Graf niezorientowany [8, Rozdział 2.3]*
Grafem niezorientowanym, nazywamy graf $G =< V, E >$, jeżeli relacja binarna tego grafu jest symetryczna, tj. dla dowolnych wierzchołków $v, v' \in V$, $(v, v') \in E$ wttw $(v', v) \in E$.

2.5 Relacja porządkująca

W niniejszej pracy skupiamy się na zagadnieniu porządkowania danych wielowymiarowych. Koniecznym jest zatem przywołanie odpowiednich sformułowań dotyczących matematycznej definicji porządku. Najbardziej podstawowym pojęciem jest relacja porządku, którą teraz definiujemy

Definicja 2.5.1. *Relacja porządkująca [6, Rozdział 1]*
Niech dana relacja ρ, którą oznaczać będziemy przez $\leqslant$, będzie określona dla elementów ustalonego zbioru X. Mówimy, że relacja $\leqslant$ jest relacją porządkującą, jeśli spełnione są warunki:

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

Definicja 2.5.2. *Relacja liniowo porządkująca (liniowy porządek) [2, Rozdział 2]*
Niech dany będzie 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 liniowości:

$$x \leqslant y \text{ lub } y \leqslant x$$

Parę $(X, \leqslant)$ nazywamy zbiorem liniowo uporządkowanym lub łańcuchem.

Definicja 2.5.3. *Dobry porządek [2, 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 2.5.4. *Ograniczenie górne [2, 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 2.5.5. *Ograniczenie dolne [2, 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 $a \leqslant x$ dla każdego $a \in A$.

Definicja 2.5.6. *Zbiór ograniczony z góry, zbiór ograniczony z dołu, zbiór ograniczony [2, 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 2.5.7. *Kres górny [2, 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 2.5.8. *Kres dolnym [2, 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 $y \leqslant y_0$.*

Przechodząc do dalszej dalszych pojęć związanych z porządkowaniem, przytoczę opis obiektu wzorcowego, a następnie podam jego formalną definicję.

Stwierdzenie w oparciu o [8, Rozdział 2.2]
Obiektem wzorcowym nazywany jest obiekt modelowy o pożądanych wartościach zmiennych wejściowych.

Definicja 2.5.9. *Obiekt wzorcowy [7, Rozdział 2.1]*
Obiektem wzorcowym, nazywamy obiekt powstały na podstawie macierzy wystandaryzowanych zmiennych wejściowych. Współrzędnymi obiektu są:

$$O_0 = [z_{0j}], j = 1, 2, ..., m.$$

gdzie:
Współrzędne obiektu wzorcowego obliczamy na podstawie następującego wzoru:

$$z_{oj} = max_i$$

$$z_{oj} = \begin{cases} \max\limits_i \left\{ z_{ij} \right\} & dla\ z_j^S \\ \\ \min\limits_i \left\{ z_{ij} \right\} & dla\ z_j^D \end{cases} \tag{2.2}$$

gdzie:
$j = 1, 2, ..., m; i = 1, 2, ..., n.$

Definicja 2.5.10. *Funkcja kryterium dobroci uporządkowania [8, Rozdział 2.2]*
Funkcją kryterium dobroci uporządkowania nazywamy funkcję:

$$F^2 = \sum_{i'=1}^{n-1} i' \sum_{i=1}^{n-i'} d_{ii'}$$

gdzie:
$d_{i,i'}$ *- odległość euklidesowa między i-tym i i'-tym obiektem .*

Rozdział 3

Metody porządkowania

Rozdział opracowany w oparciu o [8, Rozdział 2], metody porządkowania zbioru obiektów można podzielić na metody porządkowania liniowego i metody porządkowania nieliniowego. Obie grupy metod mogą stanowić punkt wyjścia do grupowania obiektów.
Metody porządkowania liniowego pozwalają na ustalenie hierarchii obiektów ze względu na określone kryterium. Problematyka związana z grupowaniem obiektów ma tutaj znaczenie drugoplanowe. Natomiast stosowanie metod porządkowania nieliniowego nie pozwala na ustalenie hierarchii obiektów, lecz wyłącznie wskazanie dla każdego z tych obiektów podobnych ze względu na wartości opisujących je zmiennych. Powoduje to, że porządkowanie nieliniowe stanowi przede wszystkim etap wstępny do grupowania obiektów.

3.1 Metody porządkowania liniowego

Porządkowanie liniowe obiektów polega, w ujęciu geometrycznym, na rzutowaniu na prostą punktów reprezentujących obiekty, umieszczonych w wielowymiarowej przestrzeni zmiennych. Takie postępowanie pozwala na ustalenie hierarchii obiektów, czyli uporządkowanie ich od obiektu stojącego najwyżej w tej hierarchii do obiektu znajdującego się najniżej. Poniżej zostaną przedstawione własności uporządkowania liniowego obiektów.

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

Aby uporządkować liniowo obiekty, charakteryzujące je zmienne muszą być mierzone przynajmniej na skali porządkowej. Gdy zmienne te mierzone są na skali przedziałowej lub ilorazowej, należy dokonać ich normalizacji, dla zapewnienia ich porównywalności.
Metody porządkowania liniowego można podzielić na metody diagramowe, procedury oparte na zmiennej syntetycznej oraz procedury iteracyjne bazujące na optymalizacji funkcji kryterium dobroci uporządkowania.

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. W kolejnym etapie następuje dzielenie mierników odległości macierzy, na klasy podobieństwa obiektów. Kolejny krok polega na przyporządkowaniu poszczególnym klasom podobieństwa obiektów odpowiedniego symbolu graficznego. Samo porządkowanie obiektów polega na porządkowaniu diagramu, tj. przestawieniu wierszy i odpowiadających im kolumn diagramu, tak aby symbole graficzne reprezentujące najmniejsze odległości skupiały się wzdłuż głównej przekątnej, zaś w miarę oddalania się od głównej przekątnej znajdowały się symbole graficzne odpowiadające coraz to większym odległością.

Jako narzędzie pomocnicze w porządkowaniu danych, może stanowić kryterium postaci:

$$F^1 = \sum_{i=1}^{n} \sum_{i'>1}^{n} d_{ii'}, w_{ii'}$$

gdzie:
$d_{i,i'}$ - odległość euklidesowa między i-tym i i'-tym obiektem .
$w_{i,i'}$ - wagi elementów macierzy odległości, zdefiniowane w oparciu o jeden z następujących wzorów:

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

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

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

Dodatkowo wagi elementów macierzy odległości tworzą macierz wag postaci:

$$W = [w_{ii'}], i, i' = 1, 2, ..., n.$$

3.1.2 Metody oparte na zmiennych syntetycznych

W tym podrozdziale zostaną opisane metody porządkowania oparte na zmiennych syntetycznych, tj. funkcji których wartości będą służyć do porządkowania danych. 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

1. dla średniej arytmetycznej:

$$s_i = \frac{1}{m}\sum_{j=1}^{m} z_{ij}w_j, i = 1, 2, ..., n,$$

2. dla średniej geometrycznej:

$$s_i = \prod_{j=1}^{m} (z_{ij})^{w_j}, i = 1, 2, ..., n,$$

3. dla średniej harmonicznej

$$s_i = \Big[\sum_{j=1}^{m} \frac{w_j}{z_{ij}}\Big]^{-1}, i = 1, 2, ..., n,$$

gdzie:
s_i - wartość zmiennej syntetycznej w i-tym obiekcie,
w_j - waga j-tej zmiennej.

Metody bewzorcowe

W metodach tych, unormowane wartości podanych zmiennych wejściowych są uśrednianie, przez przypisywanie im odpowiednich wag. Poniżej zostaną omówione wybrane metody porządkowania bezwzorcowe.

Metoda rang

W metodzie ta bazuje na normalizacji rangowej, w związku z tym zmienne mierzone są na skali porządkowej. Na początku dokonywana jest stymulacja zmiennych, w kolejnym kroku dla każdego obiektu wyznacza się sumę przyporządkowanych mu rang ze względu na wszystkie zmienne. W chwili gdy dana wartość zmiennej występuje w jednej niż jednym obiekcie, następuje przyporządkowanie im jednakowej rangi będącej średnią arytmetyczną z przysługujących im rang. Na końcu zostaje obliczona wartość zmiennej syntetycznej, jako średnia wartość rang:

$$s_i = \frac{1}{m}\sum_{j=1}^{m} z_{ij}, i = 1, 2, ..., n,$$

gdzie:
z_{ij}-zmienna znormalizowana rangowo, tj.

$$z_{ij} = h \text{ dla } x_{hj} = x_{ij}, h, i = 1, 2, ..., n.$$

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

Metoda sum

Metoda ta bazuje na konstrukcji zmiennej syntetycznej przy pomiarze zmiennych na skali ilorazowej lub przedziałowej. W pierwszym etapie zostaje dokonana stymulacja zmiennych. Następnie obliczana jest wartość zmiennej syntetycznej dla każdego obiektu, jako średnia arytmetyczna z wartości zmiennych, przy przyjęciu jednakowych wag dla każdej zmiennych. Następnie eliminowane są ujemne wartości zmiennej syntetycznej, poprzez przesuwanie jej skali do punktu zerowego, przy użyciu przekształcenia:

$$s_i' = s_i - min\{s_i\}, i = 1, 2, ..., n$$

.

Końcowa postać zmiennej syntetycznej otrzymywana jest po przeprowadzeniu normalizacji według wzoru:

$$s_i'' = \frac{s_i'}{max\{s_i'\}}, i = 1, 2, ..., n.$$

Powyższe przekształcenia powodują unormowanie miary syntetycznej w przedziale [0,1].

Metoda wzorcowe

W metodach tych zakłada się istnienie obiektu wzorcowego, w którym zmienne wejściowe przyjmują optymalne wartości, które to mogą być ustalane na podstawie ogólnie przyjętych norm, subiektywnej opinii dotyczącej obserwowanego obiektu, lub też opinii ekspertów. Poniżej zostaną omówione wybrane metody porządkowania, wzorcowe.

Metoda Hellwiga

W metodzie tej, w pierwszym etapie wyznacza się obiekt wzorcowy, na podstawie wystandaryzowanych zmiennych wejściowych. Współrzędnym obiektu wzorcowego przyporządkowuje się maksimum, gdy zmienne wejściowe są stymulantami, lub też minimum gdy zmienne wejściowe są destymulantami. Następnie dla każdego obiektu, następuje obliczenie jego odległości od obiektu wzorcowego, w tym celu najczęściej wykorzystywana jest metryka euklidesowa. Miara syntetyczna jest postaci:

$$s_i = 1 - \frac{d_{i0}}{d_0}, i = 1, 2, ..., m,$$

gdzie:
współrzędne obiektu wzorcowego są obliczane na podstawie wzoru:

$$z_{0j} = \begin{cases} \max\limits_i\{z_{ij}\} & \text{dla } z_j^S, j = 1, 2, ..., m; i = 1, 2, ..., n \\ \\ \min\limits_i\{z_{ij}\} & \text{dla } z_j^D, j = 1, 2, ..., m; i = 1, 2, ..., n \end{cases};$$

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

$$d_0 = \overline{d_0} + 2S(d_0);$$

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

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

Wartości miary s_i zazwyczaj są z przedziału $[0; 1]^2$. Należy tu zaznaczyć, że wartości miary są tym wyższe, im mniej jest oddalony obiekt od obiektu wzorcowego.

Metoda Walesiaka

Metoda ta bazuje na konstrukcji zmiennej syntetycznej w oparciu o badanie odległości obiektów od obiektu wzorcowego , przy wykorzystaniu uogólnionej miary odległości. Umożliwia ona porządkowanie obiektów, jeżeli opisujące je charakterystyki są mierzone przynajmniej na skali porządkowej. W takim przypadku, zmienne wejściowe o postaci nominant muszą zostać podaje stymulacji. Z kolei gdy zmienne są mierzone na skali przedziałowej lub ilorazowej, należy je znormalizować. Miara syntetyczna oparta na uogólnionej mierze odległości przyjmuje postać:

$$s_i = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j a_{i0j} b_{0ij} + \sum_{j=1}^{m}\sum_{i''=1}^{n} w_j a_{ii''j} b_{0i''j}}{2\left[\sum_{j=1}^{m}\sum_{i''=1}^{n} w_j a_{ii''j}^2 \cdot \sum_{j=1}^{m}\sum_{i''=1}^{n} w_j b_{0i''j}^2\right]^{\frac{1}{2}}} \tag{3.1}$$

gdzie:

$$w_j \in [0; m]$$

$$\sum_{j=1}^{m} w_j = m$$

Ostateczna postać zmiennej syntetycznej zależy od skali pomiaru zmiennych.

Jeśli zmienne charakteryzujące obiekty mierzone są na skali ilorazowej lub przedziałowej, stosowane jest następujące podstawienie:

$$a_{ii^*j} = z_{ij} - z_{i^*j} \text{ dla } i^* = 0, i^{''} \; b_{0i^*j} = z_{0j} - z_{i^*j} \text{ dla } i^* = i, i^{''}.$$

gdzie:
z_{0j}-wystandaryzowana wartość j-tej zmiennej dla obiektu wzorcowego

Z kolei, gdy zmienne charakteryzujące obiekty mierzone są na skali porządkowej to stosowne jest podstawienie:

$$a_{ii^*j} = \begin{cases} 1 & \text{dla } z_{ij} > z_{i^*j}, \\ 0 & \text{dla } z_{ij} = z_{i^*j}, i^* = 0, i^{'} \\ -1 & \text{dla } z_{ij} < z_{i^*j} \end{cases} \tag{3.2}$$

$$b_{0i^*j} = \begin{cases} 1 & \text{dla } z_{0j} > z_{i^*j}, i^* = i, i^{'} \\ 0 & \text{dla } z_{0j} = z_{i^*j}, i^* = i, i^{'} \\ -1 & \text{dla } z_{0j} < z_{i^*j}, \end{cases} \tag{3.3}$$

Zmienna syntetyczna przyjmuje wartości z przedziału $[0; 1]$. Czym niższa wartość zmiennej syntetycznej, tym bliżej wzorca leży dany obiekt.

Metoda dystansowa

Podobnie jak wcześniej opisane metody, w pierwszym kroku należy wyznaczyć zmienną syntetyczną w oparciu o jej odległość od obiektu wzorca, dla każdego z porównywanych obiektów, przy wykorzystaniu np. metryki euklidesowej. Miara syntetyczna, przy wykorzystaniu przekształcenia unitaryzacyjnego, jest postaci:

$$s_i = \left(\frac{d_{i0} - \min\limits_i \{d_{i0}\}}{\max\limits_i \{d_{i0}\} - \min\limits_i \{d_{i0}\}} \right)^p, i = 1, 2, ..., m \tag{3.4}$$

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

3.1.3 Metody iteracyjne

W metodach tych przyjmowania jest funkcja kryterium dobroci porządkowania i 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

W metodzie tej takie liniowe uporządkowanie obiektów, dla którego funkcja kryterium dobroci uporządkowania osiąga maksimum:

$$F^2 = \sum_{i'=1}^{n-1} i' \sum_{i=1}^{n-i'} d_{ii'} \rightarrow \max \tag{3.5}$$

gdzie:
$d_{ii'}$ - odległość euklidesowa miedzy i-tym i i'-tym obiektem.
Sposób postępowania:
W pierwszym kroku dokonywane jest dowolne liniowe uporządkowanie obiektów, dla którego obliczana jest wartość funkcji kryterium (3.5). W kolejnym etapie obliczana jest wartość funkcji kryterium dla każdej możliwej transpozycji pary obiektów. Jeżeli wartość funkcji kryterium dla każdej z transpozycji par obiektów są mniejsze od wartości tej funkcji dla uporządkowania wyjściowego obiektów, uporządkowanie to uważane jest za najlepsze. W przeciwnym wypadku, dokonywana jest transpozycja tej pary obiektów, dla której wzrost wartości funkcji kryterium jest największy.
Uporządkowanie to stanowi punkt wyjścia do oceny, czy kolejna transpozycja dowolnej pary obiektów pozwoli na wzrost wartości funkcji kryterium. Powyższe postępowanie kontynuowane jest tak długo, aż transpozycja dowolnej pary obiektów nie prowadzi do wzrostu wartości funkcji kryterium.

3.1.4 Metody gradientowe

W metodach gradientowych dąży się do takiego liniowego uporządkowania obiektów, które jak najmniej zniekształca relacje strukturalne porządkowanego zbioru obiektów. Od strony geometrycznej oznacza to, że odległości pomiędzy punktami reprezentującymi obiekty w przestrzeni jednowymiarowej, określonej przez zmienną syntetyczną, w jak najmniejszym stopniu zniekształcają odległości pomiędzy tymi punktami w przestrzeni wielowymiarowej, określonej przez zmienne wejściowe. Metody gradientowe poszukują takich współrzędnych punktów reprezentujących obiekty w przestrzeni jednowymiarowej, dla których funkcja dobroci uporządkowania osiąga minimum, co można przedstawić za pomocą wariantów:

$$F^3 = \frac{\sum_{\substack{i,i'=1 \\ i \neq i'}}^{n} (d^s_{ii'} - d_{ii'})^2}{\sum_{\substack{i,i'=1 \\ i<i'}}^{n} d_{ii'},} \rightarrow \min \tag{3.6}$$

$$F^4 = \sum_{\substack{i,i'=1 \\ i<i'}}^{n} \left(\frac{d^s_{ii'} - d_{ii'}}{d_{ii'}} \right)^2 \rightarrow \min \tag{3.7}$$

$$F^5 = \frac{1}{\sum_{\substack{i,i'=1 \\ i \neq i'}}^{n} d_{ii'}} \sum_{\substack{i,i'=1 \\ i<i'}}^{n} \frac{\left(d^s_{ii'} - d_{ii'} \right)^2}{d_{ii'}} \rightarrow \min \tag{3.8}$$

gdzie:
$d_{ii'}$ - odległość euklidesowa miedzy i-tym i i'-tym obiektem.

Punkt wyjścia procedury, polega na dowolnym liniowym uporządkowaniu obiektów, następnie

w trakcie kolejnych iteracji poszukiwane jest ekstremum funkcji wielu zmiennych, zapewniające jak największy spadek wartości funkcji kryterium. Poniżej zostały szczegółowo omówione kroki postępowania:

Na początku wyznaczana jest wartość funkcji kryterium dla wyjściowego, liniowego uporządkowania obiektów (wyjściowych wartości zmiennych syntetycznych w tych obiektach), przy czym wartość funkcji jest przyjmowana jako wynik iteracji dla $t = 0$:

$$F^5 = \frac{1}{c} \sum_{\substack{i,i^{'}=1 \\ i<i^{'}}}^{n} \frac{\left(d^s_{ii^{'},t} - d_{ii^{'}}\right)^2}{d_{ii^{'}}} \tag{3.9}$$

gdzie:

$$c = \frac{1}{\sum_{\substack{i,i^{'}=1 \\ i<i^{i}}}^{n} d_{ii^{'}}}, \tag{3.10}$$

Przy czym wartości zmiennych oryginalnych oraz wyjściowych wartości zmiennych syntetycznych, zostały znormalizowane na przedziale $[0; 1]$.

Współrzędne zmiennych syntetycznych dla obiektów w kolejnej iteracji $t + 1$ wyznacza się na podstawie wzoru:

$$s_{i,t+1} = s_{i,t} - W\Delta_i(t), \tag{3.11}$$

gdzie:

$$\Delta_i(t) = \frac{\delta F^5_t}{\delta s_{i,t}} : \frac{\delta F^5_t}{(\delta s_{i,t})^2}, \tag{3.12}$$

przy czym:

$$\frac{\delta F^5}{\delta s_i} = -\frac{2}{c} \sum_{\substack{i^{'}=1 \\ i\neq i^{'}}}^{n} \left(\frac{d_{ii^{'}} - d^s_{ii^{'}}}{d_{ii^{'}} + d^s_{ii^{'}}}\right)(s_i - s_{i^{'}}), \tag{3.13}$$

$$\frac{\delta^2 F^5}{(\delta s_i)^2} = -\frac{2}{c} \sum_{\substack{i\neq 1 \\ i\neq i^{'}}}^{n} \frac{1}{d_{ii^{'}} d^s_{ii^{'}}} \left[(d_{ii^{'}} - d^s_{ii^{'}}) - \frac{(s_i - s_{i^{'}})^2}{d^s_{ii^{'}}} \left(1 + \frac{d_{ii^{'}} - d^s_{ii^{'}}}{d^s_{ii^{'}}}\right)\right]. \tag{3.14}$$

W - parametr.

Na wstępie zakłada się maksymalną oraz minimalną wartość parametru W, wskaźnik skali zmian wartości tego parametru pomiędzy iteracjami W_{t+1}/W_t oraz maksymalną liczbę iteracji. Procedurę iteracyjną rozpoczynamy od przyjętej maksymalnej wartości parametru W. Postępowanie iteracyjne jest kontynuowane do momentu gdy nastąpi wzrost wartości wzrost wartości funkcji kryterium. Po tym następuje powrót do wartości zmiennej syntetycznej z poprzedniej iteracji, przy jednoczesnym zmniejszeniu wartości parametru W o przyjęty wskaźnik jego zmian. Procedura kontynuowana jest do momentu, aż wartość parametru W, nie spadnie poniżej założonej wartości minimalnej lub gdy osiągnie z góry założoną liczbę iteracji.

3.2 Metody porządkowania nieliniowego

Porządkowanie nieliniowe polega, od strony geometrycznej, na rzutowaniu na płaszczyznę obiektów umieszczonych w wielowymiarowej przestrzeni zmiennych. Ta metoda porządkowania nie pozwala na ustaleniu hierarchii obiektów, lecz na określeniu dla każdego z nich, stopnia podobieństwa do innych obiektów, ze względu na ich charakterystyki.

Aby uporządkować nieliniowo obiekty, charakteryzujące je zmienne powinny być mierzone przynajmniej na skali przedziałowej lub ilorazowej. Gdy zmienne te mierzone są na skali przedziałowej lub ilorazowej, należy dokonać ich normalizacji, dla zapewnienia ich porównywalności.

Metody porządkowania nieliniowego można podzielić na metody dendrytowe i metody aglomeracyjne. Metody dendrytowe prowadzą do powstania dendrytu, będącego ilustracją graficzną położenia względem siebie obiektów ze względu na ich podobieństwo. Z kolei metody aglomeracyjne prowadzą do utworzenia drzewka połączeń, będącego graficzną ilustracją hierarchii łączenia obiektów, ze względu na ich podobieństwo.

3.2.1 Metody dendrytowe

Metody dendrytowe opierają się na regułach i pojęciach teorii grafów. Porządkowanie dendrytowe polega na przyporządkowaniu obiektom poszczególnych wierzchołków dendrytu, w tym celu budowany jest dendryt. Poniżej zostaną opisane przykłady metod dendrytowych, tj. taksonomia wrocławska oraz metoda Prima.

Taksonomia wrocławska

W pierwszym etapie tej metody, dla każdego obiektu O_i poszukiwany jest obiekt $O_{i'}$, który jest najbardziej do niego podobny. W tym celu w każdym wierszu(kolumnie) macierzy odległości D, wyznaczamy jest element najmniejszy:

$$d_i i^{'} = \max_{i^{'}} d_{ii'}, i, i^{'} = 1, 2, ..., n; i \neq i^{'}. \tag{3.15}$$

Następnie otrzymane pary najbardziej podobnych do siebie obiektów, przedstawiane są w postaci grafu niezorientowanego, Długość krawędzi łączących wierzchołki grafu, są proporcjonalne do odległości między obiektami. Może się zdarzyć, że wśród wyznaczonych par połączeń, pojawią się połączenia występujące dwukrotnie, jedno z nich zostanie wyeliminowane, ponieważ kolejność połączeń w dendrycie nie jest istotne. Warto również zwrócić uwagę, na fakt iż w dendrycie danych obiekt może występować tylko jeden raz, w związku z tym jeżeli w łączeniu występują wielokrotnie te same obiekty, to zostaną one połączone w zespoły zwane skupieniami.

W kolejnym kroku, sprawdza się, czy utworzony graf jest spójny. Jeżeli tak będzie, to algorytm zostaje zakończony. W przeciwnym wypadku, poszczególne składowe dendrytu łączy się w większe zespoły. Odpowiednie skupienia łączone są ze sobą w miejscach, określonych przez minimalną odległość między nimi. Tworzone są w ten sposób skupienia 2-go rzędu. W tym celu znajdowana jest najmniejsza odległość każdego obiektu jednego skupienia, od obiektów należących do pozostałych skupień. Z uzyskanych odległości wybierana jest odległość najmniejsza, która zostaje wiązadłem łączącym skupienia.

Powyższy proces przeprowadzany jest do momentu, aż nie powstanie graf spójny, w ten sposób tworzone są skupienia wyższego rzędu

Metoda Prima

W odróżnieniu od taksonomii wrocławskiej, metoda Prima nie wymaga operowania cały czas pełną, wyjściową macierzą odległości. W trakcie tworzenia dendrytu, na każdym etapie zbiór porządkowanych obiektów jest klasyfikowany do jednego z dwóch podzbiorów, np. A i B. Niech zbiór A będzie pierwszym z nich a zbiór B drugim. Pierwszy z nich zawiera obiekty należące na danym etapie do dendrytu, zaś drugi zawiera obiekty nie należące na tym etapie do dendrytu.

Sposób postępowania:

Na początku procedury, zbiór A jest zbiorem pustym, z kolei zbiór B zawiera wszystkie obiekty.

W pierwszym kroku do zbioru A zostaje włączony dowolny obiekt, nie ma to wpływu na ostateczną postać dendrytu. Następnie do zbioru A zostają włączone te obiekty zbioru B, najbardziej podobne do obiektów należących już do zbioru A. Proces ten trwa do momentu, aż zbiór B nie będzie pusty. W tym celu w pierwszym kroku algorytmu zostaje stworzony wektor d, zawierający odległości wybranego obiektu zbioru A, od pozostałych obiektów zbioru B. Po utworzeniu wektora, sprawdzane jest dla którego elementu odległość od elementu ze zbioru A, jest najmniejsza. Po znalezieniu tego elementu, zostaje on włączony do zbioru A i usunięty ze zbioru B. Po tym etapie, sprawdzane jest czy zbiór B jest pusty, jeżeli tak jest to algorytm kończy działanie, zaś w przeciwnym wypadku, zostaje ponownie tworzony wektor d, którego elementami są najmniejsze z odległości każdego z obiektów pozostających jeszcze w zbiorze B od obiektów, które należą do zbioru A. Ponownie wybierany jest z wektora d najmniejszy element i włączany do zbioru A przy jednoczesnym usunięciu ze zbioru B.

W powstałym dendrycie wierzchołkami są obiekty przechodzące kolejno do zbioru A, z kolei wiązadłami łączącymi wierzchołki są minimalne wartości elementów wektora d, otrzymanego w kolejnych krokach przyłączania obiektów do dendrytu.

3.2.2 Metody aglomeracyjne

Istotą metod aglomeracyjnych jest utworzenie drzewka połączeń - dendrogramu. W ten sposób zobrazowana jest hierarchia łączenia obiektów, na podstawie zmniejszającego się podobieństwa między obiektami włączonymi do dendrogramu, w kolejnych etapach a obiektami należącymi już do dendrogramu. Hierarchia połączeń określa wzajemnie położenie względem siebie obiektów oraz grup obiektów powstających w kolejnych etapach tworzenia drzewka. Grupy podobnych do siebie obiektów tworzą oddzielne gałęzie. Punktem wyjściem metod aglomeracyjnych stanowi założenie, że każdy obiekt stanowi odrębną, jednoelementową grupę $(\mathrm{G_r}, r = 1, 2, ..., z)$.. W kolejnych krokach następuje łączenie ze sobą grupy obiektów najbardziej do siebie podobnych ze względu na wartości opisujących je zmiennych. Podobieństwo weryfikowane jest na podstawie odległości między grupami obiektów. Poniżej zostały szczegółowo omówione kroki postępowania:

Na początku odległości między jednoelementowymi grupami obiektów $\mathrm{G_1}, ..., \mathrm{G_z}$ są elementami wyjściowej macierzy odległości **D**. W macierzy **D** poszukiwane są najmniejsze odległości pomiędzy grupami obiektów:

$$d_{rr'} = \min_{ii'} d_{ii'}, i = 1, 2, ..., n_r; i^{'} = 1, 2, ..., n_{r'}; r, r^{'} = 1, 2, ..., z; r \neq r^{'}. \tag{3.16}$$

gdzie:
$d_{rr'}$ - odległość r-tej od $r^{'}$-tej grupy. W kolejnym kroku następuje łączenie do siebie obiektów podobnych w jedną grupę, w wyniku czego wyjściowa liczba grup zmniejszona jest o jeden, oraz rozpoczęta jest budowa drzewka połączeń. Następnie wyznacza się ponownie odległość nowo utworzonej grupy obiektów od wszystkich pozostałych grup obiektów. Odległości te umieszczone zostają w macierzy odległości D - w miejscu wierszy i kolumn odpowiadających obiektom(grupom obiektów) połączonych w jedną grupę. Po każdym etapie grupowania ponownie określana jest odległość między nowo powstałą grupa a pozostałymi grupami. Warto również dodać, że odległości te tworzą nową, aktualną na danym etapie grupowania, macierz odległości o co raz mniejszym wymiarze $(n-u)(n-u)$, gdzie u jest u-tym etapem łączenia grup obiektów. Procedura łączenia grup obiektów powtarzana jest tak długo, aż nie zostanie utworzona jedna grupa, tj. zostanie utworzone pełne drzewko połączeń.

Ogólny wzór wyznaczania odległości nowo powstałej grupy $\mathrm{G_{r''}}$, powstałej w wyniku połączenia grup $\mathrm{G_r}$ i $\mathrm{G_{r'}}$, od pozostałych grup $G_r{}'''$, przy tworzeniu drzewka połączeń ma postać:

$$d_{r'''r''} = \alpha_r d_{r'''r} + \alpha_{r'} d_{r'''r'} + \beta d_{rr'} + \gamma |d_{r'''r} - d_{r'''r'}| \tag{3.17}$$

gdzie:
$\alpha_r, \alpha_{r'}, \beta, \gamma$ - współczynniki przekształceń odmienne dla różnych metod aglomeracyjnych

Poszczególne metody aglomeracyjne, różnią się między sobą sposobami wyznaczania odległości między obiektami. Poniżej zostały wymienione najczęściej stosowane metody aglomeracyjne, które będą dokładniej omówione w kolejnym podrozdziale.

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

Metoda najbliższego sąsiedztwa

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

$$d_{rr'} = \min_{ii'} d_{ii'}(\mathbf{O_i} \in \mathbf{G_r}, \mathbf{O_{i'}} \in \mathbf{G_{r'}}), \tag{3.18}$$
$$i = 1, 2, ..., n_r; i^{'} = 1, 2, ..., n_{r'}; r, r^{'} = 1, 2, ..., z; r \neq r^{'},$$

gdzie:

$$\mathbf{O_i} = [z_{ij}], j = 1, 2, ..., m. \tag{3.19}$$

Metoda najdalszego sąsiedztwa

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

$$d_{rr'} = \max_{ii'} d_{ii'}(\mathbf{O_i} \in \mathbf{G_r}, \mathbf{O_{i'}} \in \mathbf{G_{r'}}), \tag{3.20}$$
$$i = 1, 2, ..., n_r; i^{'} = 1, 2, ..., n_{r'}; r, r^{'} = 1, 2, ..., z; r \neq r^{'},$$

Metoda średniej międzygrupowej

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

$$d_{rr'} = \frac{1}{n_r n_{r'}} \sum_{i'=1}^{n_{r'}} \sum_{i=1}^{n_r} d_{ii'}(\mathbf{O_i} \in \mathbf{G_r}, \mathbf{O_{i'}} \in \mathbf{G_{r'}}) \tag{3.21}$$
$$r, r^{'} = 1, 2, ..., z; r \neq r^{'}.$$

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:

$$d_{rr'} = \mathbf{med}_{i,i'}\{d_{ii'}(\mathbf{O_i} \in \mathbf{G_r}, \mathbf{O_{i'}} \in \mathbf{G_{r'}})\}, \tag{3.22}$$

$$i = 1, 2, ..., n_r; i^{'} = 1, 2, ..., n_{r'}; r, r^{'} = 1, 2, ..., z; r \neq r^{'}.$$

Metoda środków ciężkości

W metodzie tej odległość między dwoma grupami jest równa odległości między środkami ciężkości tych grup. Odległość ta opisana jest wzorem:

$$d_{rr'} = d_{i^c i'^c}(\mathbf{O_i^c} = \overline{\mathbf{O}}_\mathbf{r} \in \mathbf{G_r}, \mathbf{O_{i'^c}} = \overline{\mathbf{O}}_\mathbf{r} \in \mathbf{G_{r'}}), \tag{3.23}$$

$$i = 1, 2, ..., n_r; i^{'} = 1, 2, ..., n_{r'}; r, r^{'} = 1, 2, ..., z; r \neq r^{'}.$$

gdzie:
$d_{i^c i'^c}$ - odległość środka ciężkości r-tej grupy od środka ciężkości r'-tej grupy,
$\overline{\mathbf{O}}_{\mathbf{i^c}}, \overline{\mathbf{O}}_{\mathbf{i'^c}}$ - środki ciężkości odpowiednio r-tej i $r^{'}$-tej grupy obiektów. przy czym:

$$\mathbf{O}_{i^c} = \overline{\mathbf{O}}_r = \frac{1}{n_r}\sum_{\mathbf{i=1}}^{\mathbf{n_r}}\mathbf{O}_i \tag{3.24}$$

$$\mathbf{O}_{i'^c} = \overline{\mathbf{O}}_{r'} = \frac{1}{n_{r'}}\sum_{\mathbf{i'=1}}^{\mathbf{n_{r'}}}\mathbf{O}_{i'}. \tag{3.25}$$

Metoda Warda

W metodzie tej odległości między dwoma grupami obiektów nie można przedstawić wprost za pomocą odległości między obiektami należącymi do tych grup. Dwie grupy obiektów podczas tworzenia drzewka połączeń, na dowolnym etapie są łączone w jedną grupę,w celu zminimalizowania sumy kwadratów odchyleń wszystkich obiektów z tych dwóch grup od środka ciężkości nowej grupy, powstałej w wyniku połączenia tych dwóch grup. Proces ten oznacza, że na każdym etapie łączenia grup obiektów, w jedną grupę łączy się te grupy, które charakteryzują się najmniejszym zróżnicowaniem ze względu na opisujące je zmienne. Zróżnicowanie badania się przy pomocy kryterium $ESS(ErrosSumofSquares)$ sformułowanego przez J.H. Warda, które jest postaci:

$$ESS = \sum_{\mathbf{i''=1}}^{\mathbf{n_r''}} d^2_{i''i''^c}(\mathbf{O}_{i''} \in \mathbf{G}_{r''}, \mathbf{O}_{i''^c} = \overline{\mathbf{O}}_{r''} \in \mathbf{G}_{r''}), \tag{3.26}$$

gdzie: $d_{i''i''^c}$ - odległość $i^{''}$-tego obiektu, należącego do nowo powstałej $r^{''}$-tej grupy od środka ciężkości tej grupy,

$$\mathbf{O}_{i''^c} = \overline{\mathbf{O}}_{r''} = \frac{1}{n_{r''}}\sum_{\mathbf{i''=1}}^{\mathbf{n_r''}}\mathbf{O}_{i''}. \tag{3.27}$$

Rozdział 4

Zbiór danych

4.1 Opis zbioru

Zbiór danych jest opracowaniem własnym, na podstawie ofert sprzedaży samochodów osobowych, zamieszczonych na portalu *www.otomoto.pl*. 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 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 jest oferta.

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

Rysunek 4.1: Podgląd stworzonego zbioru

Użyte zmienne

W stworzonym zbiorze danych znajduje się 29 atrybutów, opisujących 61 różnych rekordów. Wśród zebranych danych można wyróżnić zarówno zmienne jakościowe, jak i ilościowe.

Zmiennymi jakościowymi są atrybuty: MARKA, MODEL, WERSJA, TYP, WOJEWODZTWO, KOLOR, RODZAJ.PALIWA, SKRZYNIA.BIEGOW, NAPED, KRAJ.AKTUALNEJ.REJESTRACJI, KRAJ.POCHODZENIA, STAN, ABS, STATUS.POJAZDU.SPROWADZONEGO, PIERWSZY.WLASCICIEL, KTO.SPRZEDAJE, SERWISOWANY, KOMPUTER.POKLADOWY, ESP, KLIMATYZACJA, BEZWYPADKOWY, USZKODZONY.

Wśród zmiennych jakościowych można wyróżnić zmienne porządkowe, nominalne oraz binarne. W stworzonym zbiorze danych zmiennymi binarnymi są atrybuty: PIERWSZY.WLASCICIEL, SERWISOWANY, ABS, KOMPUTER.POKLADOWY, ESP, BEZWYPADKOWY, USZKODZONY. Pozostałe atrybuty są zmiennymi nominalnymi.

Zmiennymi ilościowymi są atrybuty: CENA.[PLN].NETTO, CENA.[PLN].BRUTTO, MOC, POJEMNOSC.SKOKOWA[cm3], ROK.PRODUKCJI, PRZEBIEG[km], L.DRZWI.
Wśród zmiennych ilościowych można wyróżnić zmienne skokowe oraz dyskretne. W stworzonym zbiorze danych, zmiennymi skokowymi są: MOC, POJEMNOŚĆ.SKOKOWA[cm3], ROK.PRODUKCJI, PRZEBIEG, L.DRZWI. Z kolei atrybuty: CENA.[PLN].NETTO, CENA.[PLN].BRUTTO są zmiennymi ciągłymi.

Bibliografia

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

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

[3] Tadeusz Grabiński, Stanisław Wydmus, and Aleksander Zelias. *Metody doboru zmiennych w modelach ekonometrycznych.* Państwowe Wydawnictwo Naukowe, Warszawa, 1982.

[4] Jerzy Greń. *Statystyka matematyczyna: modele i zadania.* Państwowe Wydawnictwo Naukowe, Warszawa, 1984.

[5] 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.

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

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

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

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

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