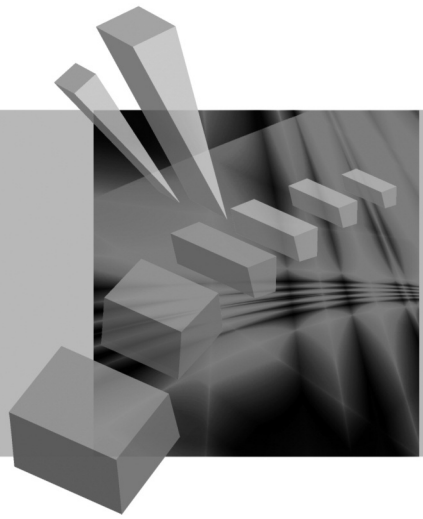
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Andrzej Sokołowski
Uniwersytet Ekonomiczny w Krakowie

Q UNIWERSALNA MIARA ODLEGŁOŚCI

Streszczenie: W zadaniach taksonomicznych coraz częściej mamy do czynienia z danymi zawierającymi zmienne mierzone w różnych skalach. Jedną ze strategii jest wykonanie analizy osobno dla jednorodnej grupy cech, a potem scalenie wyników. Druga strategia to zastosowanie miary odległości, która ze swej natury umożliwia jednoczesne wykorzystanie cech mierzonych w różnych skalach. Najbardziej dojrzałą propozycją jest tu miara Walesiaka GDM. W proponowanym artykule przede wszystkim wykraczamy poza tradycyjny podział cech ze względu na skalę pomiaru – na nominalne, porządkowe, przedziałowe i ilorazowe. Wyróżniono 15 rodzajów cech statystycznych. Następnie zaproponowano takie przekształcenia tych cech (lub odległości), że składowe sumy w odległości typu Manhattan przyjmują wartości z przedziału [0,1] i są niemianowane. To umożliwia policzenie odległości ogólnej.

Słowa kluczowe: miary odległości, skale pomiaru.

W zadaniach taksonomicznych coraz częściej mamy do czynienia z danymi zawierającymi zmienne mierzone w różnych skalach. Jedną z możliwych strategii jest wykonanie analizy osobno dla jednorodnej grupy cech, a potem scalenie wyników. Druga strategia to zastosowanie miary odległości, która ze swej natury umożliwia jednoczesne wykorzystanie cech mierzonych w różnych skalach. Najbardziej dojrzałą propozycją jest miara Marka Walesiaka GDM.

W pracy wykraczamy poza tradycyjny podział cech ze względu na skalę pomiaru – na nominalne, porządkowe, przedziałowe i ilorazowe. Zaproponowano takie przekształcenia tych cech (lub odległości), że składowe sumy w odległości typu Manhattan przyjmują wartości z przedziału [0,1] i są niemianowane. To umożliwia policzenie odległości ogólnej.

Zdarza się, że niektóre cechy mają ograniczenia naturalne. Na przykład wiele cech może przyjmować tylko wartości dodatnie, więc zero jest ograniczeniem z lewej strony. Udziały procentowe przyjmują z kolei wartości z przedziału [0;100]. Te ograniczenia mogą być wykorzystane w procesie normalizacji. Przy braku naturalnych ograniczeń zmienności można wykorzystać ograniczenia empiryczne, czyli zaobserwowane wartości skrajne, ewentualnie przesunięte jeszcze o jakąś umowną stałą.

Wybieramy odległość Manhattan dlatego, że ona daje się dekomponować na poszczególne cechy. W naszych rozważaniach pomijamy wagi, choć wprowadzenie ich do procedury liczenia odległości jest możliwe i proste.

Odległości indywidualne mogą być względne lub bezwzględne. Odległość bezwzględna zależy tylko od współrzędnych tych dwóch punktów, a nie od współrzędnych pozostałych punktów w analizowanym zbiorze. Odległości względne liczone są np. na danych standaryzowanych lub na normalizowanych wykorzystujących empiryczne granice zmienności.

Odległość Q jest sumą odległości indywidualnych. W poniższym wzorze *i* oraz *j* to numery obiektów, *k* – numer cechy, *m* zaś to rozmiar przestrzeni klasyfikacji. Podany wzór można oczywiście różnie przekształcić, choćby zamieniając go na średnią odległość lub wprowadzając wagi.

$$q_{ij} = \sum_{k=1}^{m} d_{ij}^{(k)},$$

$$0 \leq d_{ij}^{(k)} \leq 1.$$

W dalszej części pracy omówiono różne rodzaje cech statystycznych i sposoby sprowadzania ich do przedziału [0;1] (w niektórych przypadkach nie jest to przedział obustronnie domknięty) lub też sposób sprowadzania odległości do przedziału jednostkowego. Opisy są z reguły lakoniczne.

CECHY ILOŚCIOWE CIĄGŁE

To chyba najpopularniejszy rodzaj cech statystycznych. Rozważamy tutaj tylko te cechy, które mogą być zakwalifikowane jako stymulanty lub destymulanty. Nominanty będą przekształcane w inny sposób. We wzorach przyjmujemy konwencję, wedle której x_i to oryginalne wartości cech, a x_i^* to wartości znormalizowane (przekształcone).

Stymulanty

$$x_i^* = \frac{x_i - min_i\{x_i\}}{max_i\{x_i\} - min_i\{x_i\}}.$$

Destymulanty

$$x_i^* = \frac{max_i\{x_i\} - x_i}{max_i\{x_i\} - min_i\{x_i\}}.$$

Odległość Manhattan dla stymulant

$$d_{ij} = \left| \frac{x_i - min_i\{x_i\}}{max_i\{x_i\} - min_i\{x_i\}} - \frac{x_j - min_i\{x_j\}}{max_j\{x_j\} - min_j\{x_j\}} \right|.$$

Minimum i maksimum są liczone po wszystkich obiektach, więc są identyczne. Dla uproszczenia (i dla podkreślenia, że nie muszą to być wartości zaobserwowane) będziemy używali tylko symboli *min* i *max*. Mamy więc

$$d_{ij} = \left| \frac{x_i - min}{max - min} - \frac{x_j - min}{max - min} \right| = \left| \frac{x_i - x_j}{max - min} \right|.$$

Dla destymulant mamy

$$d_{ij} = \left|\frac{x_j - x_i}{max - min}\right|.$$

Obydwa wzory dają więc tę samą odległość.

CECHY ILOŚCIOWE SKOKOWE NIEUJEMNE (*COUNT DATA*)

Przykładem takiej cechy jest liczba dzieci w rodzinie. Ogólnie rzecz biorąc, jeżeli cecha ta przyjmuje dużo wartości, to może być przekształcana tak jak zmienna ciągła. Przy relatywnie małych wartościach najpierw szacujemy wartość przeciętną λ. Załóżmy, że dobrą aproksymacją rozkładu cechy jest rozkład Poissona z dystrybuantą $F(x) = P(X < x)$. Przy takiej wersji definicji dystrybuanty ważna jest jej wartość w punkcie o jeden większym niż x_i, który chcemy przekształcić. Wyjściowa miara odległości ma postać

$$d_{ij} = \left|F(x_i + 1) - F(x_j + 1)\right|.$$

Niedogodnością może być zakres zmienności wartości przekształconych punktów empirycznych. Mamy

$$x_i^* \epsilon\, [F(0); 1).$$

Można te przedziały rozciągnąć, przesuwając początek zakresu zmienności do zera. Przekształcenie ma postać

$$x_i^* = \frac{F(x_i+1) - F(0)}{1 - F(0)}.$$

Wtedy odległość indywidualna to

$$d_{ij} = \frac{\left|F(x_i+1) - F(x_j+1)\right|}{1 - F(0)}.$$

Ta odległość ma zakres zmienności [0;1). Jest odległością względną, ponieważ zależy od λ, szacowanej ze wszystkich obserwacji.

CECHY RANGOWE

Tu wystarczy jeden wzór dla stymulant, bo dla stymulant odwrócenie należy przeprowadzić już na etapie rangowania. Przy stymulantach największa wartość otrzymuje rangę 1, a przy destymulantach tę rangę otrzymuje wartość najmniejsza. Przy takim rangowaniu obiekt *i* jest lepszy od $(n - r_i)$ obiektów. Rangę obiektu o numerze *i* oznaczamy przez r_i. Jeżeli obiekt jest pierwszy, to jest lepszy od $(n - 1)$ obiektów. Sprowadzenie rang do przedziału [0,1] dokonuje się według wzoru

$$x_i^* = \frac{n - r_i}{n - 1}.$$

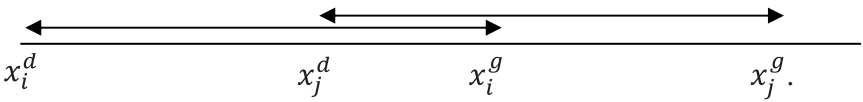
Odległość

$$d_{ij} = \left|\frac{n - r_i}{n - 1} - \frac{n - r_j}{n - 1}\right| = \left|\frac{r_j - r_i}{n - 1}\right|.$$

CECHY PRZEDZIAŁOWE

Nie chodzi tu o cechy mierzone w skali przedziałowej, lecz o takie, dla których zamiast pojedynczej wartości podajemy przedział liczbowy. Przykładami może tu być zakres zmian temperatury czy wydatki na grupę dóbr.

Przedział definiowany jest jako (x_i^d, x_i^g)



Miara odległości

$$d_{ij} = 1 - \frac{x_i^g - x_i^d + x_j^g - x_j^d}{2\left(max\left\{x_i^g, x_j^g\right\} - min\left\{x_i^d, x_j^d\right\}\right)}.$$

Miara jest równa zeru, jeżeli przedziały są identyczne, zaś jeden – jeżeli są zdegenerowane do punktów.

CECHY NOMINALNE

Taka zmienna ma kilka wariantów, które są nieuporządkowane i wzajemnie wykluczające się. Gdy są tylko dwie kategorie, to zmienna staje się cechą binarną. Przykłady to: status na rynku pracy (pełnozatrudniony, część etatu, bezrobotny, nie należy do siły roboczej), kolor oczu, płeć. Przy cechach binarnych są tylko dwie możliwe relacje: albo dwa obiekty mają ten sam wariant cechy, albo różny. W pliku danych każdy wariant ma osobną kolumnę. Jego występowanie oznaczone jest przez 1, a niewystępowanie przez zero.

$$d_{ij} = \frac{\sum_{l=1}^{w} \left| x_{il} - x_{jl} \right|}{2},$$

$$x_{il}, x_{jl} \in \{0, 1\},$$

w – liczba wariantów cechy, l – numer wariantu.

CECHY NOMINALNE Z WIELOKROTNYM POJEDYNCZYM WYBOREM

Przykłady: lokalizacja przerzutów nowotworowych, objawy choroby, wyposażenie gospodarstwa domowego, wskazanie ważnych wartości z listy

$$d_{ij} = \frac{\sum_{l=1}^{w} \left| x_{il} - x_{jl} \right|}{w}.$$

Ta odległość znana jest pod nazwą *niezgodności procentowej*. Dobrze byłoby, gdyby warianty cechy jakościowej były mniej więcej takiej samej ważności (wyjściowo). Problem może pojawić przy wyposażeniu gospodarstwa domowego. Takie

elementy, jak samochód, odtwarzacz DVD, są w odległości traktowane tak samo. Można wprowadzić zmienną pomocniczą: *wartość przedmiotu,* i zamienić cechę na ilościową. Częstym rozwiązaniem jest też zamiana takiej cechy na zestaw cech zero-jedynkowych – dla każdego wariantu tworzy się osobną cechę binarną.

CECHY NOMINALNE Z WIELOKROTNYM WYBOREM CZĘSTOŚCI

Tu dobrym przykładem jest wyposażenie gospodarstwa domowego. Można mieć trzy telewizory, dwa samochody itd. Przez wartości trzeba przejść na sumaryczną zmienną ilościową lub stworzyć osobną cechę dla każdego wariantu. Wtedy będą to cechy *ilościowe punktowe*.

KATEGORIE UPORZĄDKOWANE LINIOWO

Wykluczające się, wyczerpujące kategorie uporządkowane. Nie są zdefiniowane odległości pomiędzy kategoriami. Przykłady to: zadowolenie z pracy, rating agencji, odpowiedzi w skali Likerta, NYHA, indeks oftalmopatii.

Wydaje się, że są trzy możliwości:

1. Porangowanie kategorii i postępowanie takie jak w przypadku zmiennych rangowych – jest to raczej „wyjście rozpaczy".
2. Wykorzystanie zmiennej pomocniczej – np. przy wykształceniu jest nią średnia płaca osób z danym wykształceniem. Wtedy klasy: co najwyżej podstawowe, zasadnicze zawodowe, średnie zawodowe, średnie ogólne, wyższe nie są kodowane liczbami naturalnymi od 1 do 5.
3. Opinie ekspertów, które pozwolą ustalić odległości między wariantami, biorąc pod uwagę ogólne kryterium grupowania.

KATEGORIE UPORZĄDKOWANE NIELINIOWO

Przykład: klasyfikacja TNM – stopnia zaawansowania nowotworu. Klasa zależy od rozmiaru guza (*Tumor*), zajęcia węzłów chłonnych (*Nodes*), przerzutów odległych (*Metastasis*). Przykład ocen w tej skali to IIA, IIB, IIIA. Tutaj odpada pierwsza – wspomniana powyżej – możliwość prostego numerowania, natomiast praktycznie łączy się wykorzystanie zmiennej pomocniczej z opiniami ekspertów. Praktycznie dla wszystkich nowotworów znane są średnie przeżycia (dla ustalonego okresu t^*) i to może być wykorzystane do kodowania wariantów cechy. Przy oznaczeniu funkcji przeżycia przez $S(t)$ mamy następujący wzór na odległość

$$d_{ij} = \left|S_i(t^*) - S_j(t^*)\right|.$$

UDZIAŁY i PROPORCJE

Wartości tego typu cech mieszczą się z natury w przedziale [0,1]. Wobec tego odległość to po prostu

$$d_{ij} = \left|x_i - x_j\right|.$$

CECHY NIEUJEMNE Z DUŻĄ LICZBĄ ZER

Klasyczny przykład rozważany przez Tobina [1958]: wydatki na pewien rodzaj dóbr w gospodarstwie domowym, w określonym czasie. Cecha nieujemna z dużą liczbą zer jest moim zdaniem cechą jakościowo-ilościową. Tu dobitnym przykładem jest liczba zajętych węzłów chłonnych (inny przykład: tzw. indeks wiązania w cytologii, w którym inną wartością jakościową jest 1, a inną liczby większe od 1). Inny dobry przykład to liczba miesięcy odsiedzianych w więzieniu. Wzory podajemy dla zmiennej z dużą liczbą zer. Najpierw należy przekształcić zmienną oryginalną według wzoru

$$x_i^* = \begin{cases} 0 & \text{dla } x_i = 0 \\ \frac{x_i + max\{x_i\}}{2max\{x_i\}} & \text{dla } x_i > 0 \end{cases}.$$

Dalej odległość indywidualna to po prostu

$$d_{ij} = \left| x_i^* - x_j^* \right|.$$

PRZEŻYCIE (CZAS TRWANIA ZJAWISKA)

Zakładamy możliwość wykorzystania danych cenzurowanych. Zazwyczaj są one jeszcze prawostronnie ucinane. Na przykład jeżeli interesują nas przeżycia 5-letnie, to wszystkie dłuższe przeżycia skracamy do 60 miesięcy. Oznaczmy tę wartość graniczną jako x_g. Przeżycie to specyficzna zmienna, która „składa się" z dwóch kolumn. Jedna to czas przeżycia, a druga to zmienna wskaźnikowa (dwustanowa) informująca, czy na końcu okresu zapisanego w pierwszej kolumnie zaszło badane zdarzenie (zgon, upadek przedsiębiorstwa, awaria) czy też nie. Są tu możliwe dwa podejścia.

1. Na podstawie danych szacujemy funkcję przeżycia $S(t)$. Jeżeli przez x_i oznaczymy przeżycie i-tego obiektu, to przekształcenie ma postać:

$$x_i^* = 1 - S(x_i).$$

Ta metoda nie bierze pod uwagę faktu, czy przeżycie jest kompletne czy nie. Naprawdę wszystkie przeżycia traktowane są jako cenzurowane. Jest to dość karkołomna interpretacja, że sekundy przed zgonem ktoś jednak żył i wtedy nastąpiło cenzurowanie.

2. W tym podejściu zakłada się, że istnieje jednak fundamentalna różnica pomiędzy tymi, co zmarli, a tymi, co jeszcze żyją. Odcinek [0,1] zostaje podzielony na trzy równe części. Pierwsza 1/3 to ci, którzy żyją, druga 1/3 to strefa buforowa, która oddziela żyjących od zmarłych, i wreszcie ostatnia 1/3, czyli zmarli.

Przekształcenie odbywa się według wzoru

$$x_i^* = \begin{cases} \frac{x_i}{3x_g} & \text{dla przeżyć kompletnych} \\ \frac{2x_i}{3x_g} & \text{dla przeżyć cenzurowanych} \end{cases}.$$

WYBÓR Z PORZĄDKOWANIEM

Na przykład z listy 15 wartości trzeba wybrać pięć i uporządkować je od najważniejszej. Załóżmy, że to porządkowanie odbywa się przez przyznawanie punktów. Widać dwa sposoby:

1. Jeżeli należy wybrać *s* wartości, to ta uznana za najważniejszą otrzymuje *s* punktów, następna (*s* – 1) itd., a wartości niewybrane „otrzymują" po 0 punktów. Oznaczmy przez x_l liczbę punktów przyznanych wariantowi o numerze *l*. Odległość jest wtedy dana wzorem

$$d_{ij} = 1 - \frac{2\sum_{l=1}^{w} min\{x_{il}, x_{jl}\}}{s(s+1)}.$$

2. Zliczamy punkty z wszystkich i tworzymy hierarchię wartości. Punkty można przeliczyć na wartości sumujące się do jedności. Każdej wartości wybranej przyporządkowana jest wartość według tych punktów. Maksimum punktów to wybranie *s* pierwszych w klasyfikacji, a minimum to *s* ostatnich.

$$max = \sum_{(l)=1}^{s} x_{(l)}$$

$$min = \sum_{(l)=s-1}^{w} x_{(l)} .$$

WYNIKI SPOTKAŃ SPORTOWYCH

Jest to typowa zmienna ilościowo-jakościowa, gdyż w przykładowym wyniku meczu piłkarskiego 3:2 mamy zawartą informację o tym, że mecz wygrała drużyna gospodarzy, a także informację o liczbie bramek strzelonych przez poszczególne drużyny. Propozycję odległości sformułował Sokołowski [2007]. Odległość liczona tu jest odległością względną. Najpierw wyniki są przekształcane według następujących wzorów

$$\text{jeżeli } x_H > x_A \rightarrow \begin{cases} x_H^* = x_H + 2 \cdot max \\ x_A^* = x_A \end{cases}$$

$$\text{jeżeli } x_H = x_A \rightarrow \begin{cases} x_H^* = x_H + max \\ x_A^* = x_A + max \end{cases}$$

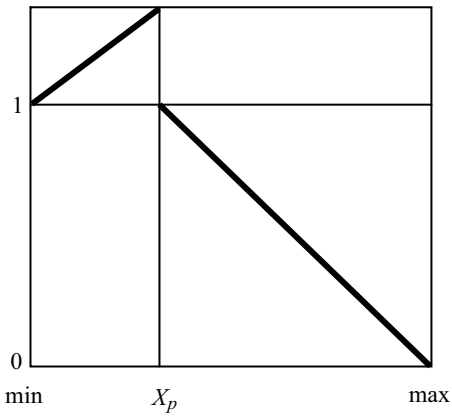
$$\text{jeżeli } x_H < x_A \rightarrow \begin{cases} x_H^* = x_H \\ x_A^* = x_A + 2 \cdot max \end{cases}.$$

Subskrypt H oznacza drużynę gospodarzy, A – drużynę gości, a *max* jest największą liczbą zdobytych bramek (punktów) przez jedną drużynę w zbiorze wyników. Znormalizowana odległość między dwoma wynikami to

$$d_{ij} = \left| \frac{x_{Hi}^* - x_{Hj}^* + x_{Ai}^* - x_{Aj}^*}{6 \cdot max} \right|.$$

CECHA CIĄGŁA Z PUNKTEM PRZEŁAMANIA

Jest to coś w rodzaju nominanty, ale po przekroczeniu tzw. punktu przełamania wartości cechy są oceniane jako coraz gorsze, i to gorsze niż jakakolwiek wartość mniejsza od punktu przełamania. Z taką cechą miałem do czynienia przy ocenie gospodarstw rolnych, gdzie istniała optymalna dawka nawozu na hektar, a jej przekroczenie było oceniane gorzej, niż gdy stosowano mniej niż dawkę optymalną, a nawet gorzej, niż gdy nawozu w ogóle nie stosowano.



Schematycznie taką sytuację przedstawiono powyżej. Wartość x_i przyjmującą wartości z przedziału [*min, max*] należy przekształcić na $0 \leq x_i^* \leq 1$. Można to zrobić według następujących wzorów

$$x_i^* = \frac{max - min + x_i - x_p}{max - min} \qquad \text{dla } x_i < x_p$$

$$x_i^* = \frac{max - x_i}{max - min} \qquad \text{dla } x_i \geq x_p.$$

Podsumowując, podkreślmy, że ideą uniwersalnej miary odległości jest wykorzystanie dystansu typu Manhattan, w którym odległości „po poszczególnych wymiarach" mogą być liczone dla różnego rodzaju cech. Wartości tych cech są wstępnie sprowadzane do przedziału [0;1] bądź do tego przedziału sprowadzana jest sama odległość cząstkowa.

Literatura

Sokołowski A., *The Football Distance*, [w:] Taksonomia 14, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Akademii Ekonomicznej im. Oskara Langego we Wrocławiu nr 1169, Wydawnictwo AE, Wrocław 2007.

Tobin J., *Estimation of relationship for limited dependent variables*, „Econometrica" 1958, no 26.

Walesiak M., *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wydawnictwo UE, Wrocław 2011.

Q UNIVERSAL DISTANCE MEASURE

Summary: In clustering tasks we usually have to use variables which are measured in different scales. One of possible strategies is to cluster objects separately for each scale and then try to combine them into final partition. The second strategy is to use the distance measure which allows for calculation of distances in the classification space formed by attributes measured in different scales. This second approach is possible while applying Walesiak GDM distance. In the paper we go beyond the classical four scales – nominal, order, interval and rational. Fifteen different types of variables have been defined. Then we transfer them into [0.1] interval, or such a transformation is applied to the individual (univariate) distance. The overall distance is calculated through Manhattan approach which makes possible to agglomerate distances on individual axes.

Keywords: distance measures, measurement scales.