STUDIA INFORMATICA 2015
Volume 36 Number 1 (119)

Łukasz PAŚKO, Galina SETLAK
Politechnika Rzeszowska, Zakład Informatyki

WPŁYW WYBRANYCH METRYK NA WYNIK BADANIA SKUPISK

Streszczenie. Celem artykułu jest przedstawienie metryk stosowanych do pomiaru odległości pomiędzy obiektami w przestrzeni danych. Analizy wykonano na siedmiu zbiorach danych. W każdym z nich zbadano występowanie skupisk podobnych do siebie obiektów oraz rozproszenie obiektów w skupiskach. Obliczenia przeprowadzono za pomocą czternastu znanych z literatury metryk. Praca zawiera wybrane wyniki obliczeń, ze szczególnym uwzględnieniem różnic wynikających z zastosowania poszczególnych metryk.

Słowa kluczowe: eksploracja danych, metryki, miary jakości skupisk

THE INFLUENCE OF SELECTED METRICS ON THE RESULT OF AN EXAMINATION OF CLUSTERS

Summary. The aim of this paper is to present the metrics used to measure a distance between objects in a feature space. The analyses were performed on seven datasets. For each of them, the occurrence of clusters of similar objects was examined, and the measures of clusters' dispersion were calculated. The calculations were carried out using fourteen metrics known from the literature. The article contains selected results with particular emphasis on the differences arising from the use of various metrics.

Keywords: data mining, metrics, measures of the quality of clusters

1. Wstęp

Analiza danych wymaga niekiedy odpowiedzi na pytanie, jak bardzo różnią się od siebie pewne obiekty. Każdy obiekt posiada zwykle zestaw *n* cech, zwanych też atrybutami, które są istotne z punktu widzenia przeprowadzanego badania. Wartości tych cech, zapisane w zbiorze danych, stanowią podczas analizy tzw. zmienne wejściowe, inaczej zmienne obja-

śniające lub niezależne, które można potraktować jako *n*-wymiarowy wektor cech. Wspomniane zmienne mogą posiadać wartości ilościowe (numeryczne) lub jakościowe (porządkowe i nominalne) [9]. Jeśli dwa obiekty opisane są takim samym zestawem *n* cech ilościowych, wtedy matematyka pozwala na wyznaczenie odległości, jaka dzieli je w przestrzeni danych $\Re^n$ [4]. Tak obliczona odległość stanowi odpowiedź na postawione wyżej pytanie.

Do pomiarów odległości używane są miary nazywane w literaturze metrykami. Nauka spotyka się z tym pojęciem od starożytności, czego przykładem może być metryka euklidesowa – jedna z najpopularniejszych i najczęściej używanych miar. Chociaż przestrzeń zdefiniowana przez Euklidesa służyła do opisu wielkości geometrycznych obiektów, co ograniczało ją do maksymalnie trzech wymiarów (n = 3), to jednak metrykę Euklidesa można łatwo rozszerzyć do przestrzeni o dowolnej liczbie wymiarów. W ciągu kolejnych wieków pojawiło się wiele innych metryk, będących mniejszą lub większą modyfikacją metryki euklidesowej [2, 4]. Ich zróżnicowanie prowadzi do otrzymania odmiennych wyników dla takiej samej pary wektorów, co może powodować problemy we właściwej interpretacji odległości.

Celem artykułu jest przedstawienie wybranych metryk i zastosowanie ich do pomiarów odległości pomiędzy obiektami ze zbioru danych *odkurzacze*, którego analizę opisano w pracach [17, 18]. Rezultaty pomiarów posłużyły do ustalenia, czy w zbiorze danych istnieją naturalne skupiska podobnych do siebie obiektów (test Hopkinsa), a także do wyznaczenia rozproszenia obiektów w sześciu klasach zdefiniowanych w badanym zbiorze (miary σ_1 oraz σ_2). Dla porównania, wyniki tych samych miar wyznaczono dla sześciu innych, znanych z literatury, zbiorów danych. Wyniki obliczeń omówiono, ze szczególnym uwzględnieniem różnic, jakie wywołuje stosowanie badanych metryk.

2. Zastosowane metryki i zbiory danych

W tej sekcji wyszczególniono wszystkie metryki wykorzystywane do pomiarów odległości pomiędzy obiektami (wektorami) w *n*-wymiarowej przestrzeni danych $\Re^n$. Opisano także analizowany zbiór danych oraz przybliżono cechy pozostałych sześciu zbiorów wykorzystanych do porównania wyników badań.

2.1. Metryki

Aby dowolną funkcję $d : \Re^n \times \Re^n \to \Re_+ = [0;+\infty)$ można było nazwać metryką w $\Re^n$, musi ona spełniać każdy z trzech następujących warunków:

1) $\forall x, y \in \Re^n : d(x, y) = 0 \Leftrightarrow x = y$,

2) $\forall x, y \in \Re^n : d(x, y) = d(y, x)$,

3) $\forall x, y, z \in \Re^n : d(x,y) + d(y,z) \geq d(x,z)$

Z warunków 2) i 3) wynika ponadto, że $d(x,y) \geq 0$.

Jeśli dowolne $x, y \in \Re^n$, to wtedy liczbę $d(x,y)$ można nazywać odległością, jaka dzieli punkty *x* i *y*. W przeprowadzonych analizach mierzono odległości pomiędzy obiektami wyrażonymi za pomocą wektorów cech. Dlatego w dalszej części pracy używany jest termin *wektor* $\boldsymbol{x}$ zamiast terminu *punkt x*.

Metryki wykorzystane w niniejszych badaniach przedstawiono w tabeli 1. Wymienione metryki prezentowane są m.in. w pracach [2, 3, 4, 7, 8, 12, 13, 15].

Tabela 1

Metryki wykorzystane w przeprowadzonych analizach

Metryka	**Wzór**	**Metryka**	**Wzór**
Euklides	$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^2}$	*Jaccard*	$d(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i y_i}$
Manhattan	$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$	*Dice*	$d(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}$
Czebyszew	$d(\boldsymbol{x},\boldsymbol{y}) = \max_i \lvert x_i - y_i \rvert$	*Canberra*	$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \frac{\lvert x_i - y_i \rvert}{x_i + y_i}$
Lorentzian	$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \ln(1 + \lvert x_i - y_i \rvert)$	*Wave Hedges*	$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \left(1 - \frac{\min(x_i, y_i)}{\max(x_i, y_i)}\right)$
Squared-chord	$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \left(\sqrt{x_i} - \sqrt{y_i}\right)^2$	*Squared* χ^2	$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}$
Sorensen	$d(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} (x_i + y_i)}$	*Dywergencja*	$d(\boldsymbol{x},\boldsymbol{y}) = 2\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{(x_i + y_i)^2}$
Soergel	$d(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} \max(x_i, y_i)}$	*Clark*	$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} \left(\frac{\lvert x_i - y_i \rvert}{x_i + y_i}\right)^2}$

Wektory $\boldsymbol{x}$ i $\boldsymbol{y}$ odpowiadają obiektom należącym do badanych zbiorów, zaś x_i oraz y_i są elementami tych wektorów (cechami obiektów). Wartość n jest liczbą cech i może być różna dla każdego ze zbiorów danych.

2.2. Zbiory danych

Podstawowy zbiór danych poddany analizie nosi nazwę *odkurzacze*. Opracowano go na podstawie badań rynku artykułów gospodarstwa domowego w latach 2003-2005. Zbiór zawiera dane na temat 194 odkurzaczy dostępnych wtedy na rynku. Każdy produkt scharakteryzowany jest przez dwanaście cech, które traktowane są jako zmienne niezależne. Dwie z nich mają charakter numeryczny, zaś dziewięć jest typu binarnego. Ostatnia zmienna ma typ wyliczeniowy i przyjmuje wartości {*niski*, *średni*, *wysoki*}.

Właściwą analizę poprzedzała operacja przekodowania zmiennych. Wartości atrybutów numerycznych przeskalowano do zakresu [0; 1]. Zmienne binarne nie zostały poddane modyfikacji, zachowując wartości {0; 1}. Natomiast atrybut wyliczeniowy zakodowano za pomocą trzech zmiennych binarnych.

Opisywany zbiór danych służył do zrealizowania zadania segmentacji rynku. Stąd, oprócz zmiennych niezależnych, zdefiniowano w nim także zmienne zależne, które dzieliły rynek odkurzaczy na kilka segmentów. Segmenty rynku zostały ustalone za pomocą sieci Kohonena, co opisano w artykule [18]. Segmenty są odpowiednikiem klastrów, składających się z podobnych do siebie obiektów. W niniejszej pracy termin *klaster* będzie używany wymiennie z terminami *skupisko* lub *klasa*.

W pracach [17, 18] brano pod uwagę wersję zbioru danych z czterema klastrami. Jednak obliczone w pracy [17] wyniki błędu kwantyzacji wektorowej wskazują, że optymalna liczba segmentów rynku wynosi sześć, dlatego w niniejszych badaniach przyjęto występowanie w zbiorze sześciu klastrów.

Aby porównać otrzymane wyniki analiz z innymi zestawami danych, wybrano sześć znanych z literatury zbiorów i poddano je takim samym badaniom. Ich charakterystykę, wraz ze zbiorem *odkurzacze*, przedstawia tabela 2.

Tabela 2

Wybrane cechy analizowanych zbiorów danych

Oznaczenie	*Nazwa*	*Liczba obiektów*	*Liczba cech obiektów*	*Liczba klas*
O	*odkurzacze*	194	12	6
I	*balance*	625	4	3
II	*cleveland*	297	13	5
III	*hayes-roth*	160	4	3
IV	*iris*	150	4	3
V	*newthyroid*	215	5	3
VI	*tae*	151	5	3

Każdy ze zborów od **I** do **VI** pochodzi z internetowego repozytorium danych, opisanego w pracy [1]. Dostęp do zbiorów można uzyskać w [19], gdzie znajduje się też opis ich pochodzenia i przeznaczenia. Przed rozpoczęciem analiz zmienne wejściowe wszystkich zbiorów zostały przekodowane podobnie jak w przypadku zbioru *odkurzacze*.

3. Wyniki analiz

Przeprowadzone badania są kontynuacją analiz opisanych w pracy [17]. Ich zadaniem była ocena jakości segmentacji rynku. Podział rynku na segmenty uzyskano za pomocą grupowania danych siecią Kohonena. Otrzymane w ten sposób wyniki, w postaci produktów podzielonych na klastry (segmenty rynku), zapisano w zbiorze danych *odkurzacze*. Do oceny segmentacji użyto znanych z literatury miar badających jakość klastrów (skupisk) znalezionych za pomocą metod grupowania danych [6, 11, 16]. W niniejszym artykule, spośród miar rozpatrywanych w pracy [17], wybrano do ponownego przeanalizowania trzy: test Hopkinsa i dwie miary rozproszenia (σ_1 oraz σ_2).

Każda z wymienionych miar bazuje na pojęciu odległości pomiędzy obiektami (wektorami) w przestrzeni danych [14]. W pracy [17] skorzystano przy pomiarze odległości jedynie z metryki euklidesowej. W niniejszym artykule ponownie obliczono wspomniane miary, stosując tym razem czternaście metryk wyszczególnionych w sekcji 2.1. Wybrane rezultaty obliczeń wraz z wnioskami omówiono w kolejnych sekcjach.

3.1. Test Hopkinsa

Test Hopkinsa wykorzystywany jest do ustalenia, czy w badanym zbiorze danych istnieją naturalne skupiska podobnych do siebie obiektów [16]. Wyznaczenie tej miary poprzedza wyselekcjonowanie z badanego zbioru danych pewnej liczby p obiektów. Przyjęto, że dla każdego z siedmiu analizowanych zbiorów liczba ta będzie wynosić $p = 40$. Wybrany podzbiór obiektów, oznaczony jako T, jest usuwany ze zbioru oryginalnego. W kolejnym kroku utworzony zostaje zbiór L. Składa się on z wygenerowanych przypadków o rozkładzie losowym. Jego liczność jest równa liczności podzbioru T.

Mając tak przygotowane zbiory T i L oraz zbiór oryginalny, pomniejszony o przypadki wchodzące w skład T, wykonywana jest właściwa część testu Hopkinsa. Polega ona na znalezieniu najbliższego sąsiada w zbiorze oryginalnym dla wszystkich przypadków ze zbiorów T i L. Po zidentyfikowaniu najbliższego sąsiada ustalona zostaje odległość od niego. W tym miejscu obliczone zostają dwie wartości:

- u_i – odległość i-tego wektora ze zbioru L od najbliższego sąsiada ze zbioru oryginalnego,
- w_i – odległość i-tego wektora ze zbioru T od najbliższego sąsiada ze zbioru oryginalnego,

przy czym $i = 1, 2, \ldots, p$. Przyjmując powyższe oznaczenia, test Hopkinsa ma postać:

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}. \tag{1}$$

Rezultat tej miary mieści się zawsze w zakresie [0; 1]. Interpretacja wyniku jest następująca: wartość zbliżona do 0 lub 1 to wskazanie, że w zbiorze występują skupiska podobnych do siebie obiektów. Natomiast rezultat bliski 0,5 oznacza, że zbiór oryginalnych wektorów T nie różni się zasadniczo od zbioru losowego L, stąd wniosek o braku skupisk w zbiorze.

Aby łatwiej zinterpretować uzyskane wyniki, zaproponowano przekształcenie oznaczone jako H_2, opisane wzorem (2):

$$H_2 = \left|H - 0{,}5\right|. \tag{2}$$

Zastosowanie wzoru H_2 prowadzi do przeskalowania wyników testu Hopkinsa do przedziału [0; 0,5]. W takim ujęciu, wartość zbliżona do 0 wskazuje brak występowania skupisk, natomiast wynik bliski 0,5 świadczy o istnieniu wyraźnych grup podobnych do siebie obiektów w zbiorze danych. Wyznaczone rezultaty tej miary dla badanych zbiorów danych przedstawiono w tabeli 3.

Tabela 3

Wyniki miary H_2

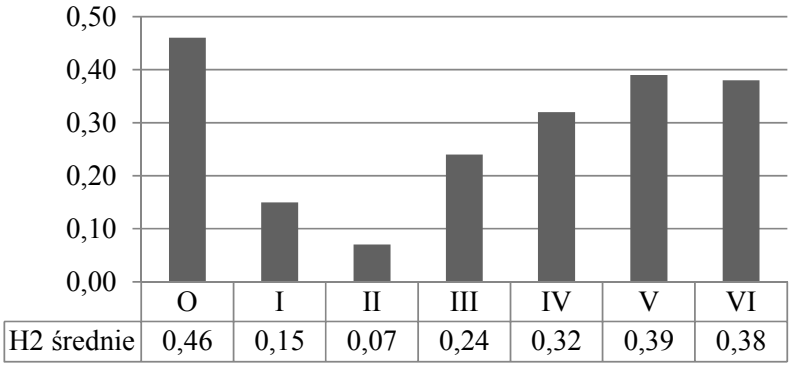
	zbiory danych						
metryki	**O**	**I**	**II**	**III**	**IV**	**V**	**VI**
Euklides	0,46	0,13	0,08	0,17	0,32	0,40	0,34
Manhattan	0,47	0,02	0,08	0,26	0,33	0,40	0,35
Czebyszew	0,44	0,20	0,08	0,12	0,31	0,40	0,34
Lorentzian	0,47	0,03	0,07	0,28	0,31	0,39	0,34
Squared-chord	0,47	0,11	0,12	0,29	0,42	0,48	0,43
Sorensen	0,45	0,04	0,04	0,21	0,28	0,35	0,40
Soergel	0,45	0,04	0,03	0,11	0,27	0,33	0,38
Jaccard	0,45	0,16	0,10	0,08	0,36	0,47	0,45
Dice	0,45	0,16	0,11	0,08	0,34	0,47	0,46
Canberra	0,46	0,29	0,04	0,37	0,29	0,31	0,35
Wave Hedges	0,46	0,26	0,03	0,36	0,26	0,26	0,33
Squared χ^2	0,47	0,05	0,12	0,26	0,43	0,47	0,43
Dywergencja	0,47	0,42	0,05	0,42	0,35	0,43	0,39
Clark	0,45	0,23	0,03	0,33	0,27	0,30	0,33

Wyniki miary H_2 dla podstawowego zbioru danych *odkurzacze* (kolumna **O**) pokazują, że w zbiorze istnieją wyraźne skupiska obiektów. Każda z użytych metryk daje tutaj bardzo zbliżony rezultat, mieszczący się w granicach 0,45÷0,47. Przeciwną sytuację, a więc brak występowania skupisk, można zaobserwować na zbiorze **II**, dla którego wszystkie metryki

przyjmują wartość nie większą niż 0,12. W tym przypadku rezultaty również mieszczą się w niewielkim zakresie kilku setnych.

Przykłady zbiorów **O** i **II** pokazują, że zastosowanie każdej z badanych czternastu metryk skutkuje otrzymaniem zbliżonych wyników miary H_2, co jednoznacznie świadczy o istnieniu bądź nieistnieniu grup podobnych obiektów. Jednak porównując wyniki poszczególnych metryk obliczone dla pozostałych zbiorów danych, można zauważyć, że znacznie odbiegały one od siebie. Największe różnice odnotowano dla zbioru **I**. Najwyższa wartość H_2 dla tego zbioru wyznaczona została za pomocą miary dywergencji i wynosiła 0,42, co wskazywałoby istnienie naturalnych skupisk. Natomiast wartość minimalna była równa 0,02 i została obliczona metryką Manhattan. Taka wielkość świadczyłaby z kolei o braku występowania skupisk w zbiorze.

Aby wziąć pod uwagę rezultaty wszystkich metryk, wyznaczono ich wartość średnią dla każdego z badanych zbiorów danych. Wyniki tych obliczeń przedstawia rysunek 1.



	O	I	II	III	IV	V	VI
H2 średnie	0,46	0,15	0,07	0,24	0,32	0,39	0,38

Rys. 1. Średnia wartość miary H_2 dla analizowanych zbiorów danych
Fig. 1. The average of H_2 measure for the analyzed datasets

Otrzymane wartości średnie miary H_2 wskazują, że skupiska istnieją najprawdopodobniej, oprócz zbioru **O**, również w zbiorach **IV**, **V** i **VI**. Odmienna sytuacja zachodzi w przypadku zbioru **I**, gdzie obliczona średnia sugeruje brak wyraźnych grup, podobnie jak w zbiorze **II**. Występowanie skupisk jest najtrudniejsze do określenia w zbiorze **III**, dla którego kilka metryk przyjmowało stosunkowo wysokie wartości, przy czym metryki Jaccarda i Dice dawały wynik równy 0,08. Średnie H_2 nie pozwala jednoznacznie rozstrzygnąć o istnieniu grup podobnych do siebie obiektów w tym zbiorze danych.

Rezultaty testu Hopkinsa porównano w dalszej części pracy z rozproszeniem klas zdefiniowanych w zbiorach danych.

3.2. Miary rozproszenia

Wszystkie metody, pozwalające ocenić jakość struktury klastrów (klas, skupisk) w zbiorach danych, dzieli się w literaturze na wzorcowe i bezwzorcowe [10]. Rozpatrywane w tej

pracy miary rozproszenia klas należą do grupy metod bezwzorcowych, gdyż pozwalają ocenić istniejącą w zbiorze strukturę klas, nie korzystając przy tym z wzorca prawidłowej (optymalnej) struktury. Rezultaty miar rozproszenia wskazują, jak bardzo różnią się od siebie obiekty przypisane do danej klasy. Inaczej mówiąc, testuje się tutaj, jak bardzo oddalone od siebie są wektory obiektów w przestrzeni danych, mierząc ich odległości za pomocą wybranej metryki [6, 11].

Porównując dwie struktury klastrów, za optymalną należy uważać tę, której klastry charakteryzują się mniejszym rozproszeniem. Oznacza to, że obiekty wchodzące w skład takich klastrów są bardziej jednorodne.

Często miarę rozproszenia klastra definiuje się odwrotnie, mówiąc o spójności klastra. Przy takim podejściu większa spójność skupiska oznacza większą jednorodność obiektów do niego przypisanych [6, 10, 11, 16].

Najprostszą miarą rozproszenia jest średnica klastra, równa odległości pomiędzy najbardziej różniącymi się od siebie obiektami do niego przypisanymi [16]. Jednak wnioskowanie na temat rozproszenia całego skupiska, uwzględniając tylko jego dwa skrajne wektory, obarczone może być dużym błędem, jeśli w tym skupisku występują obserwacje odstające (ang. *outliers*). Dlatego najbardziej wiarygodnymi miarami rozproszenia mogą być opisane poniżej wskaźniki σ_1 oraz σ_2.

Miara σ_1 nazywana jest średnim rozproszeniem danego klastra k biorąc pod uwagę odległości między jego wektorami [16]. Wyraża ją następujący wzór:

$$\sigma_1(k) = \frac{1}{m} \sum_{\substack{\boldsymbol{x} \in k \\ \boldsymbol{y} \in k}} d^2(\boldsymbol{x}, \boldsymbol{y}), \tag{3}$$

gdzie $m = \dfrac{n_k(n_k - 1)}{2}$, natomiast n_k jest liczbą wektorów w klastrze k.

Zgodnie ze wzorem (3), obliczenie wartości σ_1 sprowadza się do zbadania odległości pomiędzy każdą parą wektorów należących do danego skupiska k. Rezultaty obliczeń σ_1 dla badanego zbioru danych **O** przedstawiono w tabeli 4.

Patrząc na rozproszenie klastrów zbioru **O**, widać wyraźne różnice w wynikach kilku metryk. Zastosowanie dywergencji skutkowało uzyskaniem najwyższych rezultatów, przekraczających wartość 47 dla skupisk O1 i O4. Z kolei metryki Czebyszewa, Sorensena, Soergela, Jaccarda i Dice dawały dla tych samych klastrów rezultaty nie większe od 1. Identyczne zależności można zaobserwować dla pozostałych zbiorów danych. Wyjątkiem były tylko dwa klastry w zbiorze **IV** i jeden klaster w zbiorze **V**, dla których metryka Wave Hedges osiągnęła wartość wyższą od dywergencji.

Interesujące mogą być również dysproporcje pomiędzy rozproszeniem klastrów z punktu widzenia zastosowania różnych metryk. Aby to przeanalizować, obliczono procentowy udział

rozproszenia każdego klastra w łącznej sumie rozproszenia danego zbioru. Wyniki dla czterech wybranych zbiorów i trzech metryk przedstawiają wykresy z tabeli 5.

Tabela 4

Wyniki miary σ_1 dla klastrów należących do zbioru **O**

	klastry					
metryki	**O1**	**O2**	**O3**	**O4**	**O5**	**O6**
Euklides	2,70	1,50	1,73	3,02	1,63	1,76
Manhattan	10,80	3,56	5,42	13,74	4,57	5,16
Czebyszew	0,92	0,85	0,80	0,92	0,85	0,79
Lorentzian	5,47	1,82	2,75	6,92	2,37	2,61
Squared-chord	9,64	3,04	4,64	11,69	3,75	4,47
Sorensen	0,33	0,02	0,02	0,04	0,04	0,04
Soergel	0,45	0,06	0,06	0,11	0,10	0,09
Jaccard	0,47	0,06	0,06	0,10	0,10	0,09
Dice	0,35	0,02	0,02	0,04	0,04	0,04
Canberra	13,58	3,80	5,37	13,30	5,02	5,26
Wave Hedges	14,84	4,37	5,99	14,56	5,74	5,93
Squared χ^2	9,95	3,10	4,72	11,95	3,88	4,53
Dywergencja	47,93	12,66	18,89	47,43	16,79	18,19
Clark	3,07	1,53	1,73	3,00	1,70	1,77

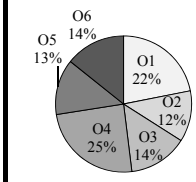
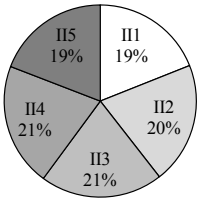
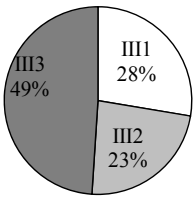
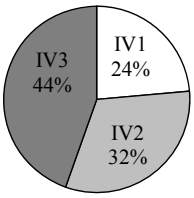
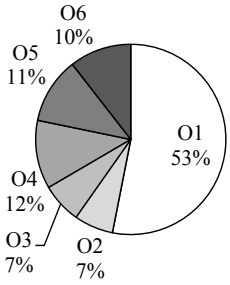
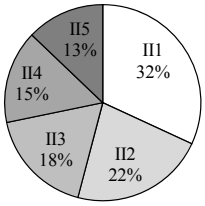
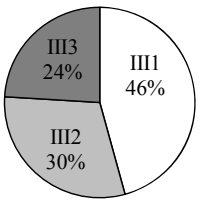
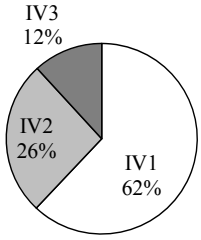
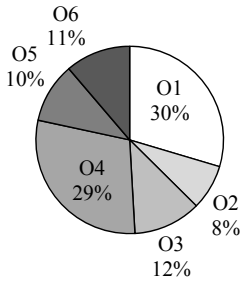
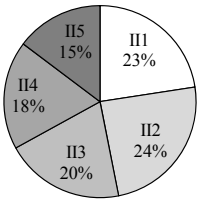
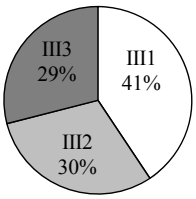
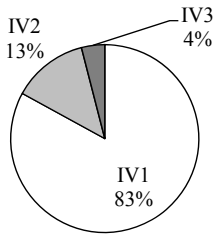
Rezultaty wyznaczone dla zbioru danych **O** pokazują, że miara Euklidesa wskazuje skupiska O1 i O4 jako najbardziej rozproszone. Podobna tendencja zachodzi w przypadku dywergencji, a także większości innych metryk nieujętych w tej tabeli. Odmienna sytuacja występuje dla metryki Jaccarda, gdzie rozproszenie klastra O1 ma ponadpięćdziesięcioprocentowy udział w łącznej sumie rozproszenia zbioru **O**. Zbliżony rezultat do miary Jaccarda uzyskano również dla metryk Sorensena, Soergela oraz Dice.

Stosunkowo duże różnice w procentowym ujęciu można zauważyć także dla zbiorów **III** i **IV**. W zbiorze **III**, metryki Jaccarda i dywergencja dają porównywalne procentowe wyniki, wskazując największe rozproszenie skupiska III1. Natomiast używając metryki Euklidesa, otrzymuje się wynik sugerujący, że to klaster III3 zawiera najmniej jednorodne obiekty. Porównywalny do metryki euklidesowej rezultat występuje przy zastosowaniu miar Manhattan, Czebyszewa, Lorentzian, a także Squared-chord i Squared χ^2. Dla zbioru **IV** wyniki są jeszcze bardziej zróżnicowane, co widać dla klastra IV1. Udział rozproszenia tego skupiska w łącznej sumie wszystkich skupisk waha się od 24% (metryka Euklidesa) do 83% (dywergencja).

Pozostałe zbiory danych, łącznie z prezentowanym w tabeli 5 zbiorem **II**, cechowały się w przypadku każdej z metryk podobnymi rezultatami w ujęciu procentowym. Jedynie metryki Sorensena, Soergela, Jaccarda oraz Dice nieznacznie odbiegały od pozostałych miar, co objawiało się niewielką przewagą rozproszenia jednego z klastrów nad pozostałymi. W zbiorze **II** widać to zjawisko na przykładzie klastra II1.

Tabela 5

Procentowy udział rozproszenia klastrów w łącznej sumie rozproszenia zbioru

	zbiory danych			
	O	**II**	**III**	**IV**
Euklides				
Jaccard				
Dywergencja				

Wskaźnik σ_2 mierzy rozproszenie skupiska k, badając odległości jego wektorów od centrum $\boldsymbol{c}_k$ tego skupiska [16]. Miarę tę można sformułować następująco:

$$\sigma_2(k) = \frac{1}{n_k} \sum_{\boldsymbol{x} \in k} d^2(\boldsymbol{x}, \boldsymbol{c}_k). \tag{4}$$

Wektor centralny $\boldsymbol{c}_k$ jest średnią wszystkich wektorów wchodzących w skład klastra k, co wyraża wzór:

$$\boldsymbol{c}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{x}_i \,. \tag{5}$$

Miarę σ_2 obliczono dla wszystkich klastrów każdego z badanych zbiorów danych, wykorzystując wszystkie rozpatrywane metryki. W rezultacie otrzymano 364 wartości rozproszenia. Ujawniają one analogiczne zależności, jakie zaobserwowano w przypadku miary σ_1. Aby łatwiej porównać rozproszenie całych struktur klastrów dla poszczególnych zbiorów, wyzna-

czono dla każdego z nich wskaźnik $r(\sigma_2)$. Jego wartość jest sumą rozproszenia wszystkich klastrów znajdujących się w danym zbiorze. Obliczone rezultaty przedstawia tabela 6.

Tabela 6

Wyniki miary $r(\sigma_2)$

	zbiory danych						
metryki	**O**	**I**	**II**	**III**	**IV**	**V**	**VI**
Euklides	5,98	1,31	6,31	1,37	0,16	0,30	1,35
Manhattan	30,58	4,32	49,59	4,32	0,45	0,72	4,64
Czebyszew	2,88	0,65	2,35	0,70	0,10	0,22	0,81
Lorentzian	20,61	3,07	35,14	3,04	0,38	0,57	3,27
Squared-chord	11,07	1,07	12,85	1,40	0,01	0,04	1,18
Sorensen	0,34	0,33	0,37	0,55	0,05	0,08	0,18
Soergel	0,72	0,67	0,88	0,99	0,13	0,23	0,41
Jaccard	0,38	0,36	0,43	0,69	0,03	0,06	0,22
Dice	0,18	0,16	0,16	0,35	0,01	0,02	0,10
Canberra	108,86	8,42	144,86	13,82	1,29	3,33	7,44
Wave Hedges	130,53	12,21	197,42	17,47	2,73	7,38	12,15
Squared χ^2	14,32	1,40	17,16	1,80	0,03	0,11	1,54
Dywergencja	339,14	20,46	370,94	40,09	1,50	3,76	12,89
Clark	20,44	3,04	19,57	4,94	0,51	1,18	2,52

Dane zebrane w tabeli 6 pokazują, że wyniki wyznaczone dla tego samego zbioru mniej lub bardziej różnią się od siebie. Jest to sytuacja normalna, ponieważ każda z użytych metryk w nieco inny sposób traktuje odległość w przestrzeni, dzielącą parę wektorów. Jednak porównując $r(\sigma_2)$ pomiędzy zbiorami, można zaobserwować przypadki, w których obliczone miary wskazują sprzeczne wnioski. Przykładowo, metryki osiągające wyższe wyniki wskazują, że zbiory **O** i **II** posiadają znacznie bardziej rozproszone struktury klastrów od pozostałych zbiorów. Natomiast metryki Sorensena, Soergela, Jaccarda i Dice osiągają najwyższy wynik dla zbioru **III**, wskazując największe rozproszenie właśnie tego zbioru.

Zestawiając zbiory **O** i **II**, które są najbardziej podobne pod względem liczności i liczby klas, widać inny przykład konfliktu wyników. Przeważająca część rezultatów sugeruje większe rozproszenie klastrów **II**. Jednak występują też miary (Czebyszewa, Dice, Clark) wskazujące odmienną sytuację. Analogiczne różnice można zauważyć także w innych przypadkach. Przykładowo, dla zbiorów **II** i **III** metryki Euklidesa i Sorensena dają zupełnie odmienne rezultaty. Pierwsza z nich wskazuje ponadczterokrotnie większe rozproszenie zbioru **II**, natomiast druga sugeruje, że to zbiór **III** posiada bardziej niejednorodną strukturę klastrów.

3.3. Wnioski

W przypadku każdego zbioru danych zastosowanie metryki dywergencji skutkowało otrzymaniem wartości największych w porównaniu do innych metryk. Rezultaty znacznie wyższe od pozostałych uzyskiwały także metryki Wave Hedges i Canberra. Wartości naj-

mniejsze były najczęściej rezultatem stosowania metryk Euklidesa, Czebyszewa, Sorensena, Soergela, Jaccarda i Dice.

Wyniki testu Hopkinsa obliczone za pomocą każdej z metryk były do siebie zbliżone tylko w przypadku zbiorów **O** i **II**, wskazując istnienie naturalnych skupisk w zbiorze **O**, a brak ich występowania w zbiorze **II**. Dla pozostałych zbiorów trudniej jednoznacznie określić obecność skupisk jednorodnych obiektów z powodu znacznych różnic w wartościach testu Hopkinsa.

Miary rozproszenia wyznaczane za pomocą różnych metryk mogą dawać odmienne rezultaty nie tylko w wielkościach bezwzględnych, ale także w procentowym ujęciu rozproszenia każdego klastra w łącznej sumie rozproszenia danego zbioru.

Mniejszym wartościom miary H_2 danego zbioru towarzyszą większe wartości rozproszenia jego klastrów. Zależność taka zachodzi dla wszystkich wykorzystywanych porównawczo zbiorów. W przypadku badanego zbioru **O** sytuacja jest odmienna – miara H_2 wykazuje istnienie wyraźnych skupisk, przy jednoczesnym występowaniu wysokich wartości rozproszenia w porównaniu do pozostałych zbiorów danych. Jest to spowodowane specyfiką tego zbioru. Szczegółowa analiza obiektów wchodzących w skład wszystkich klastrów zbioru **O** ujawniła, że każdy z nich składa się z kilku lub kilkunastu mniejszych skupisk bardzo podobnych do siebie obiektów. Taki układ skupisk spowodował otrzymanie prawie maksymalnych wyników miary H_2. Mimo to same klastry O1, …, O6 zdefiniowane w zbiorze danych charakteryzują się jednymi z najwyższych wartości rozproszenia, porównywalnymi do zbioru **II**.

4. Podsumowanie

Badane miary rozproszenia i test Hopkinsa wyznaczane są podczas oceny jakości klastrów istniejących w zbiorze lub uzyskanych za pomocą odpowiednich metod grupowania. Ta część analizy danych stanowi ważny jej fragment, decydujący często o optymalności badanych klastrów lub o poprawności zrealizowanego grupowania. Dlatego istotne jest, aby nie tylko poznać metody oceny skupisk, ale również sprawdzić, jaki wpływ na ich wynik mogą wywierać wykorzystywane metryki.

Jest rzeczą oczywistą, że wyniki badanych miar, wyznaczone dla pewnego zbioru klas za pomocą różnych metryk, będą się od siebie mniej lub bardziej różnić w wartościach bezwzględnych. Jednak przeprowadzone analizy pokazują, że metryki mogą wpływać również na względny wynik rozproszenia poszczególnych klastrów. Co za tym idzie, wynikiem zastosowania pewnej metryki może być większe rozproszenie klastra X w porównaniu do skupiska Y. Natomiast używając innej metryki dla tego samego zbioru danych, rezultat rozproszenia klastra Y może okazać się większy od rezultatu skupiska X. Sytuację taką wykazano dla

kilku zbiorów danych i w przypadku kilku różnych metryk. Może to powodować błędne interpretacje wyników otrzymywanych podczas oceny jakości skupisk.

Optymalność struktury klastrów ma duże znaczenie w problemie segmentacji rynku. Otrzymany rezultat segmentacji może mieć wpływ na strategię przedsiębiorstwa, a przez to również na jego wynik finansowy, stąd ważne jest skupienie się na ocenie uzyskanych segmentów rynku. Literatura z zakresu segmentacji rynku potwierdza istotność badania wpływu stosowanych metryk na wynik segmentacji. W pracy [5], która jest analizą literatury dotyczącej zagadnień segmentacji, stwierdzono, że w przypadku siedemdziesięciu trzech procent empirycznych badań opisanych w literaturze, autorzy nie wspominają o metrykach wykorzystanych do pomiaru podobieństwa między obiektami. W pozostałych badaniach, które precyzowały, jaką miarę zastosowano, autorzy głównie używali miary euklidesowej (96% opisanych zastosowań segmentacji rynku wykorzystuje tę metrykę). Praca [5] stwierdza, że analizowanie zasadności użycia odległości euklidesowej oraz rozpatrywanie również innych miar podobieństwa może wpłynąć na poprawę rezultatów badań nad segmentacją rynku.

Ocena segmentacji rynku to nie tylko analiza rozproszenia poszczególnych segmentów. Aby wykonać pełną ocenę jakości struktury klastrów, należy także wziąć pod uwagę miary separacji pomiędzy klastrami. Kontynuując przeprowadzone badania, możliwe jest porównanie wyników wspomnianych miar separacji, stosując do ich obliczenia różne metryki, tak by poznać całkowity wpływ używanych metryk na ocenianą jakość klastrów.

BIBLIOGRAFIA

1. Alcalá-Fdez J., Fernandez A., Luengo J., Derrac J., García S., Sánchez L., Herrera F.: KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing, Vol. 17, No. 2÷3, 2011, s. 255÷287.
2. Cha S.: Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 1, No. 4, 2007, s. 300÷307.
3. Cox T. F., Cox M. A. A: Multidimensional Scaling, 2nd edition, Chapman & Hall/CRC Press, 2000.
4. Deza M. M., Deza E.: Encyclopedia of distances. Springer-Verlag, Berlin, Heidelberg 2009.
5. Dolnicar S.: Using cluster analysis for market segmentation – typical misconceptions, established methodological weaknesses and some recommendations for improvement. Australasian Journal of Market Research, Vol. 11 (2), 2003, s. 5÷12.

6. Everitt B. S., Landau S., Leese M.: Cluster analysis, Wiley Publishing, Nowy Jork 2009.
7. Gavin D. G., Oswald W. W., Wahl E. R., Williams J. W.: A statistical approach to evaluating distance metrics and analog assignments for pollen records. Quaternary Research, Vol. 60, 2003, s. 356÷367.
8. Gordon A. D.: Classification, 2nd edition, Chapman & Hall/CRC Press, 1999.
9. Hand D., Mannila H., Smyth P.: Eksploracja danych. WNT, Warszawa 2005.
10. Jain A. K., Dubes R. C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey 1988.
11. Jain A. K., Murty M. N., Flynn P. J.: Data clustering: a review. ACM Computing Surveys, Vol. 31, No. 3, 1999, s. 264÷323.
12. Krivulin N.: An algebraic approach to multidimensional minimax location problems with Chebyshev distance. WSEAS Transaction on Mathematics, Vol. 10, No. 6, 2011, s. 191÷200.
13. Krause E. F.: Taxicab Geometry: An Adventure in Non-Euclidean Geometry. Dover, New York 1986.
14. Meila M.: Comparing clusterings – an information based distance. Journal of Multivariate Analysis, Vol. 98, No. 5, 2007, s. 873÷895.
15. Monev V.: Introduction to similarity searching in chemistry. MATCH Communications in Mathematical and in Computer Chemistry, Vol. 51, 2004, s. 7÷38.
16. Osowski S.: Metody i narzędzia eksploracji danych. Wyd. BTC, Legionowo 2013.
17. Paśko Ł., Setlak G.: Ocena segmentacji rynku za pomocą miar jakości grupowania danych. Studia Informatica, Vol. 35, No. 2 (116), Gliwice 2014, s. 157÷173.
18. Setlak G., Paśko Ł.: Zastosowanie metod eksploracji danych do segmentacji rynków. Studia Informatica, Vol. 34, No. 2A (111), Gliwice 2013, s. 311÷323.
19. http://sci2s.ugr.es/keel/datasets.php.

Abstract

The article is an analysis of the influence of uses various metrics on the results of two dispersion measures and Hopkins statistic. The measures are calculated during the assessment of quality of clusters, which is an important part of data mining. The first section introduces theoretical foundation on metrics and feature space. In the second section, the metrics used during the analysis are listed (table 1). Moreover, in table 2, the analyzed datasets are presented. Next section contains the results of the measures. Table 3 shows the outcomes of H_2 measure given by equation (2), which was determined based on Hopkins statistics. The fol-

lowing tables illustrate the dispersion measures, called σ_1 and σ_2. The paper ends with conclusions that reveal the differences between results calculated using various metrics.

Adresy

Łukasz PAŚKO: Politechnika Rzeszowska, Zakład Informatyki, al. Powstańców Warszawy 8, 35-959 Rzeszów, Polska, lpasko@prz.edu.pl.
Galina SETLAK: Politechnika Rzeszowska, Zakład Informatyki, al. Powstańców Warszawy 8, 35-959 Rzeszów, Polska, gsetlak@prz.edu.pl.