STUDIA INFORMATICA 2016
Volume 37 Number 1 (123)

Łukasz PAŚKO, Galina SETLAK
Politechnika Rzeszowska, Zakład Informatyki

ZNACZENIE DOBORU METRYK W BADANIU SEPARACJI MIĘDZY KLASTRAMI

Streszczenie. Celem artykułu jest zbadanie znaczenia doboru metryki podczas analizy separacji między skupiskami obiektów w przestrzeni danych. Do analizy wybrano czternaście znanych z literatury metryk, służących do pomiaru odległości. Analizie poddano siedem zbiorów danych, różniących się liczbą obiektów, cech i skupisk. Dla każdego z nich wyznaczano cztery miary separacji klastrów. Praca zawiera wybrane wyniki obliczeń, skupiając się w szczególny sposób na różnicach, wynikających z zastosowania każdej z metryk.

Słowa kluczowe: separowalność klastrów, metryki, jakość grupowania

THE IMPORTANCE OF SELECTION OF METRICS IN THE ANALYSIS OF SEPARATION BETWEEN CLUSTERS

Summary. The aim of this paper is to examine the importance of selection of metric during the analysis of separation between clusters of objects in the feature space. Fourteen metrics known from the literature were selected for the calculations. Seven datasets that differ in the number of objects, attributes, and clusters were examined. For each of them, the four cluster separation measures were calculated. The article contains selected results with particular emphasis on the differences arising from the use of various metrics.

Keywords: separation of clusters, metrics, measures of the quality of clustering

1. Wstęp

Jednym z podstawowych zadań analizy danych jest grupowanie, zwane klasteryzacją. Polega ono na poszukiwaniu naturalnych skupisk podobnych do siebie obiektów, występujących w badanym zbiorze danych. Wynikiem klasteryzacji jest przypisanie każdemu obiekto-

wi jednego z klastrów, nazywanego także skupiskiem lub grupą. Po zrealizowaniu klasteryzacji ważnym problem jest odpowiedź na pytanie, czy algorytm grupujący prawidłowo zlokalizował skupiska oraz czy poprawnie przydzielił obiekty do skupisk. W odpowiedzi pomocne mogą być miary oceny jakości klasteryzacji. W idealnym przypadku każdy z klastrów zawiera bardzo podobne do siebie obiekty, a jednocześnie poszczególne klastry są od siebie wyraźnie odseparowane [7, 11, 17].

Podczas badania jakości klasteryzacji kluczową kwestią jest odpowiednie zdefiniowanie podobieństwa między obiektami. Przyjmując, że każdy obiekt w zbiorze danych opisany jest takim samym zestawem *n* cech (inaczej atrybutów lub zmiennych objaśniających), wówczas obiekty można potraktować jako *n*-wymiarowe wektory cech. W takiej sytuacji podobieństwo między dowolnymi dwoma obiektami można wyrazić za pomocą odległości, jaka je dzieli w przestrzeni danych $\Re^n$ [5]. Problemem może być tylko występowanie cech jakościowych (porządkowych lub nominalnych), jednak można go łatwo rozwiązać, konwertując wcześniej takie zmienne do wartości numerycznych (ilościowych) [10].

Do pomiarów odległości w przestrzeni danych służą funkcje, zwane metrykami. Najczęściej używana jest metryka Euklidesa znana od starożytności. Jednak literatura prezentuje także inne miary, które równie dobrze mogą być zastosowane jako miara odległości. Warunki konieczne do nazwania dowolnej funkcji $d : \Re^n \times \Re^n \to \Re_+ = [0;+\infty)$ metryką w $\Re^n$ są następujące:

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

Poza tym $d(x,y) \geq 0$, co wynika ze wzorów 2) i 3). W wymienionych warunkach *x* i *y* są punktami należącymi do $\Re^n$, które w niniejszych badaniach odpowiadają wektorom cech, czyli obiektom ze zbioru danych. Natomiast obliczona wartość $d(x,y)$ jest szukaną odległością między *x* i *y*. Różnice pomiędzy opisywanymi w literaturze metrykami powodują, że wyniki pomiarów odległości dla danej pary obiektów często znacznie od siebie odbiegają, co może być problemem we właściwej interpretacji otrzymywanych rezultatów. Fakt ten był inspiracją do przeprowadzenia analizy dotyczącej wpływu metryk na ocenę jakości klasteryzacji.

2. Opis badań

Za jakość klasteryzacji odpowiada rozproszenie skupisk i separacja pomiędzy nimi. Pierwsza część badań dotyczyła wpływu wybranych metryk na wyniki miar rozproszenia

skupisk. Zostało to opisane w pracy [19]. Celem niniejszego artykułu jest przedstawienie części drugiej, gdzie skoncentrowano się na separacji między klastrami.

Badania opisane w niniejszej pracy oraz w [19] są kontynuacją wcześniejszych analiz, dotyczących oceny jakości segmentacji rynku. Rozpatrywana segmentacja została zapisana w zbiorze danych o nazwie *odkurzacze*. Zbiór ten powstał na podstawie badań rynku sprzętu gospodarstwa domowego, które realizowano w latach 2003-2005. Znajdują się w nim dane na temat 194 odkurzaczy. Każdy z nich opisany jest za pomocą dwunastu cech charakterystycznych. Z punktu widzenia analizy danych cechy produktów traktowane są jako zmienne niezależne.

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*	$d(\boldsymbol{x},\boldsymbol{y})=\frac{\sum_{i=1}^{n}(x_i-y_i)^2}{\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}x_iy_i}$
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*	$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{i=1}^{n}\left(\frac{\lvert x_i-y_i\rvert}{x_i+y_i}\right)^2}$
Oznaczenia: $\boldsymbol{x}$, $\boldsymbol{y}$ – wektory odpowiadające badanym obiektom; x_i, y_i – elementy wektorów (cechy obiektów); n – liczba cech obiektów.			

Segmentacja rynku została uzyskana za pomocą grupowania danych siecią Kohonena, co opisano w pracy [20]. Stąd w badanym zbiorze znajduje się także zmienna zależna dzieląca produkty na sześć segmentów rynku. W analizie danych segmenty są odpowiednikiem skupisk, ponieważ powinny zawierać produkty o dużym podobieństwie.

Tak przygotowany zbiór danych poddano ocenie jakości segmentacji. Jej rezultaty opisano w pracy [18]. Jakość znalezionych segmentów rynku oceniano za pomocą miar stosowanych w literaturze do badania jakości skupisk będących wynikiem grupowania danych. Wykorzystane miary jakości są opisane m.in. w [7, 12]. Każda z tych miar bierze pod uwagę odległości pomiędzy grupowanymi obiektami. Jednak w pracy [18] przy wyznaczaniu tych odległości ograniczono się jedynie do użycia metryki euklidesowej.

Niniejsza praca oraz artykuł [19] rozszerzają przeprowadzoną do tej pory ocenę segmentacji przez zastosowanie czternastu metryk, wyszczególnionych w tabeli 1. Dokładny opis i pochodzenie każdej z nich prezentowane jest m.in. w pracach [3, 4, 5, 8, 9, 14, 13, 16].

W uzyskanych rezultatach zwracano szczególną uwagę na różnice wynikające z zastosowania poszczególnych metryk. Otrzymywane wyniki mogą być uzależnione nie tylko od stosowanej metryki, ale także od parametrów badanego zbioru danych, takich jak liczność czy liczba klastrów. Aby to sprawdzić, wybrano sześć innych zbiorów danych często przedstawianych w literaturze, traktując je jako zbiory porównawcze. Każdy z nich pochodzi z internetowego repozytorium danych opisanego w pracy [1], do którego dostęp można uzyskać w [21]. Podstawowe parametry wszystkich analizowanych zbiorów 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

3. Wyniki analiz

Klastry zbiorów porównawczych poddano takiej samej ocenie jak segmenty rynku zbioru podstawowego. W badaniach wykorzystano następujące miary separacji klastrów:

- odległość pary najbliższych wektorów pochodzących z dwóch klastrów,
- odległość między centrami klastrów,

- separowalność między klastrami wyznaczana na podstawie rozkładu wektorów tworzących te klastry,
- separowalność międzyklastrowa w całej przestrzeni danych.

Wszystkie te miary należą do grupy bezwzorcowych metod oceny jakości klasteryzacji. Metody takie, zgodnie z ich założeniem, do oceny skupisk nie wykorzystują zewnętrznych informacji na temat grupowanych obiektów, a tylko biorą pod uwagę informacje zawarte w analizowanym zbiorze. Dlatego często miary te bywają nazywane wskaźnikami wewnętrznymi (ang. *internal validation*) [2, 6, 11, 15].

W kolejnych sekcjach zaprezentowano wybrane wyniki powyższych miar separacji, skupiając się na tych rezultatach, które pokazują rozbieżności wynikające z zastosowania różnych metryk.

3.1. Odległość pary najbliższych wektorów

Pierwsza miara separacji to najkrótsza odległość pomiędzy wektorami klastrów k_i oraz k_j, co wyraża wzór:

$$d\left(k_i, k_j\right) = \min_{\substack{\boldsymbol{x}\in k_i \\ \boldsymbol{y}\in k_j}} \left\{d\left(\boldsymbol{x}, \boldsymbol{y}\right)\right\}. \tag{1}$$

Wyznaczenie tej miary polega na znalezieniu pary obiektów najbardziej zbliżonych do siebie (o najbardziej zbliżonych cechach), ale należących do różnych klastrów. Wtedy separacja pomiędzy tymi klastrami to odległość między znalezionymi obiektami. Wyniki tej miary wyznaczono dla wszystkich siedmiu zbiorów danych, szukając najkrótszych odległości pomiędzy każdą parą klastrów. Po porównaniu otrzymanych rezultatów stwierdzono, że dla każdego zbioru danych wyniki kształtują się w analogiczny sposób. Dlatego w niniejszej pracy zaprezentowano tylko fragment rezultatów, które umieszczono w tabeli 3.

Tabela 3 przedstawia najkrótsze odległości pomiędzy pierwszym klastrem zbioru podstawowego (oznaczenie O1) a kolejnymi pięcioma klastrami (od O2 do O6). Dane pokazują wyraźne różnice pomiędzy poszczególnymi metrykami. Najniższe wartości otrzymano dla metryk Sorensena, Soergela, Jaccarda i Dice. Każda z nich daje rezultaty mniejsze od 0,3. Natomiast wartości największe są osiągane, korzystając z dywergencji – najwyższa z nich to około 65 jednostek zmierzona dla odległości między klastrem O1 i O3. Jednak wszystkie te różnice są zupełnie normalnym zjawiskiem, które wynika z innego sposobu obliczania poszczególnych metryk.

Zdecydowanie większe znaczenie mają proporcje pomiędzy każdą z sześciu odległości między klastrami. Proporcje te powinny być takie same lub zbliżone, niezależnie od stosowanej metryki. W tabeli 3 widać, że taka sytuacja zachodzi dla 13 metryk. Każda z tych metryk wskazuje najmniejszą separację między klastrami O1 i O2 oraz nieco większe odległości dla

O1-O5, O1-O6 i O1-O4. Najlepiej odseparowanymi klastrami okazują się O1 i O3. Tak więc pomimo różnic w wartościach bezwzględnych, wnioski po zastosowaniu badanych metryk będą dla każdej z nich takie same. Wyjątkiem jest tutaj tylko metryka Czebyszewa. Powodem tego jest specyficzne traktowanie odległości przez tę metrykę. Odległość Czebyszewa wyznacza się jako maksymalną różnicę pomiędzy składowymi dwóch rozpatrywanych wektorów. Ponieważ przed rozpoczęciem analiz wszystkie dane zostały przeskalowane do zakresu [0; 1], więc maksymalną odległością, jaką można uzyskać metryką Czebyszewa, jest 1. Stąd taki właśnie wynik dla każdej pary klastrów.

Tabela 3
Minimalny dystans między klastrem O1 a pozostałymi klastrami

	klastry				
metryki	**O1-O2**	**O1-O3**	**O1-O4**	**O1-O5**	**O1-O6**
Euklides	1,00	4,02	3,09	2,00	3,00
Manhattan	1,16	17,41	11,21	4,00	9,03
Czebyszew	1,00	1,00	1,00	1,00	1,00
Lorentzian	0,59	8,63	5,72	1,92	4,34
Squared-chord	1,00	16,09	9,29	4,00	9,00
Sorensen	0,01	0,12	0,07	0,04	0,08
Soergel	0,03	0,27	0,17	0,12	0,19
Jaccard	0,04	0,29	0,17	0,14	0,22
Dice	0,01	0,13	0,07	0,05	0,09
Canberra	1,20	17,78	11,73	4,01	9,04
Wave Hedges	1,39	19,19	13,26	4,01	9,09
Squared χ^2	1,01	16,17	9,58	4,00	9,00
Dywergencja	4,04	64,95	38,41	16,00	36,00
Clark	1,00	4,03	3,10	2,00	3,00

W czasie wykonywania obliczeń zwrócono również uwagę na to, jakie wektory z obu badanych klastrów tworzą najbliższe pary. Okazuje się, że w zależności od zastosowanej metryki najbliższe pary różnią się od siebie. Jednak trudno tutaj zauważyć jakąkolwiek prawidłowość, ponieważ pary najbliższych wektorów zależały w dużym stopniu od badanego zbioru danych. Przykładowo, w przypadku zbioru **I** niektóre metryki wybierały tę samą parę wektorów, ale na zbiorze **O** te same metryki wskazywały inne pary jako najbliższe. Z kolei dla zbioru **VI** wszystkie najbliższe pary wyznaczone każdą z czternastu metryk były identyczne.

3.2. Odległość między centrami klastrów

Drugi sposób wyznaczania separacji bada odległości między centrami klastrów k_i i k_j według wzoru:

$$s_2\left(k_i, k_j\right) = d^2\left(\boldsymbol{c}_{k_i}, \boldsymbol{c}_{k_j}\right). \tag{2}$$

Centra klastrów, zwane wektorami centralnymi lub reprezentantami skupisk, wyznacza się przez uśrednienie wartości cech wszystkich obiektów należących do danego klastra. Można to wyrazić wzorem:

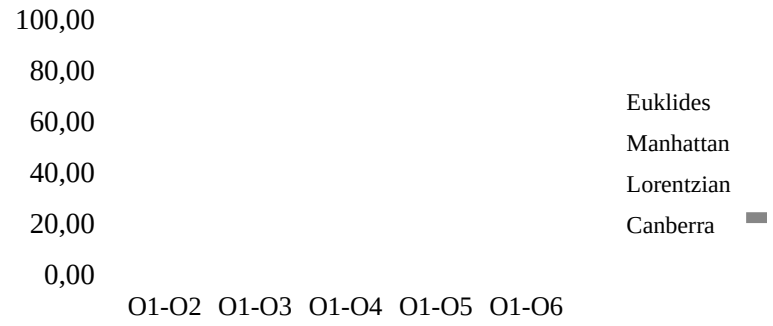
$$\boldsymbol{c}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{x}_i \,, \tag{3}$$

gdzie n_k jest liczbą wektorów w klastrze k.

Miara ta jest mniej wrażliwa na obiekty odstające w porównaniu do miary stosowanej w sekcji 3.1 dzięki temu, że odległość mierzona jest nie między skrajnymi obiektami a między reprezentantami skupisk, czyli wektorami, mającymi odzwierciedlać cechy wszystkich obiektów w danym skupisku.

Wyniki tej miary zaprezentowano na rysunku 1 dla tych samych klastrów, które porównywano w sekcji 3.1. Skupiono się tutaj na metryce Canberra. Wyniki obliczone za pomocą tej metryki odbiegały od pozostałych. Zgodność z innymi metrykami zachodzi tylko dla odległości największej (O1-O4) i najmniejszej (O1-O2). Natomiast w przypadku pozostałych separacji metryka Canberra pokazuje, że odległości O1-O5 i O1-O6 są większe niż odległość O1-O3, co nie znajduje potwierdzenia w pozostałych metrykach.

Inną rozbieżność można zauważyć, porównując wyniki sekcji 3.1 z rezultatami separacji między centrami klastrów. Poprzednia miara wskazywała separację O1-O3 jako największą, natomiast w tym przypadku jest nią O1-O4.



Rys. 1. Separacja między klastrem O1 a pozostałymi klastrami
Fig. 1. Separation between O1 cluster and other clusters

3.3. Analiza rozkładu wektorów tworzących klastry

Separowalność między klastrami wyznaczana na podstawie rozkładu wektorów tworzących te klastry jest najbardziej złożoną obliczeniowo miarą w porównaniu z poprzednimi, gdyż zakłada mierzenie odległości pomiędzy każdą parą wektorów z obu badanych skupisk. Obliczone odległości należy następnie zsumować, a otrzymaną sumę podzielić przez iloczyn liczności obu skupisk, tak by wyeliminować wpływ liczności klastrów na otrzymywany wynik. Dla klastrów k_i i k_j miarę tę można sformułować następująco:

Tabela 4

Wyniki miary s_1 dla klastrów należących do zbiorów **O** i **IV**

	metryki: *Euklides* / Manhattan					
klastry	**O1**	**O2**	**O3**	**O4**	**O5**	**O6**
O1		*4,14*	*6,69*	*7,25*	*5,22*	*5,20*
O2	21,74		*4,22*	*4,34*	*3,81*	*4,79*
O3	51,22	20,94		*3,64*	*2,85*	*3,53*
O4	61,46	23,85	16,83		*4,79*	*4,50*
O5	32,15	16,87	10,31	27,72		*2,92*
O6	31,22	25,61	15,09	25,27	11,07	

	Euklides / Manhattan		
klastry	**IV1**	**IV2**	**IV3**
IV1		*0,65*	*1,37*
IV2	2,23		*0,29*
IV3	4,51	0,99	

	metryki: *Lorentzian* / Squared chord					
klastry	**O1**	**O2**	**O3**	**O4**	**O5**	**O6**
O1		*10,86*	*25,44*	*30,52*	*16,07*	*15,58*
O2	19,82		*10,45*	*11,99*	*8,43*	*12,65*
O3	46,48	18,67		*8,44*	*5,20*	*7,56*
O4	55,14	20,47	14,70		*13,81*	*12,65*
O5	28,95	15,19	8,95	24,74		*5,57*
O6	28,47	23,76	13,31	22,07	9,75	

	Lorentz. / Sq. chord		
klastry	**IV1**	**IV2**	**IV3**
IV1		*1,53*	*2,69*
IV2	0,46		*0,73*
IV3	1,17	0,04	

	metryki: *Sorensen* / Canberra					
klastry	**O1**	**O2**	**O3**	**O4**	**O5**	**O6**
O1		*0,29*	*0,47*	*0,45*	*0,47*	*0,43*
O2	24,50		*0,10*	*0,10*	*0,12*	*0,16*
O3	54,28	21,21		*0,06*	*0,06*	*0,08*
O4	64,08	23,90	16,49		*0,13*	*0,11*
O5	35,00	17,54	10,56	27,76		*0,08*
O6	34,39	26,21	15,08	24,95	11,43	

	Sorensen / Canberra		
klastry	**IV1**	**IV2**	**IV3**
IV1		*0,30*	*0,36*
IV2	5,35		*0,05*
IV3	6,04	0,90	

	metryki: *Wave Hedges* / Squared χ^2					
klastry	**O1**	**O2**	**O3**	**O4**	**O5**	**O6**
O1		*26,21*	*57,38*	*67,66*	*37,37*	*36,71*
O2	20,33		*22,96*	*26,16*	*19,03*	*28,00*
O3	47,83	19,02		*17,89*	*11,58*	*16,47*
O4	57,09	21,09	14,95		*29,70*	*27,09*
O5	29,81	15,43	9,15	25,22		*12,58*
O6	29,17	23,97	13,51	22,52	9,91	

	Wave Hedges / Sq. χ^2		
klastry	**IV1**	**IV2**	**IV3**
IV1		*7,55*	*8,11*
IV2	1,06		*1,99*
IV3	2,54	0,12	

	metryki: *Dywergencja* / Clark					
klastry	**O1**	**O2**	**O3**	**O4**	**O5**	**O6**
O1		*89,23*	*200,80*	*237,03*	*127,75*	*125,69*
O2	4,37		*76,59*	*84,62*	*63,19*	*96,56*
O3	6,91	4,25		*59,60*	*37,46*	*54,06*
O4	7,44	4,35	3,62		*101,79*	*89,77*
O5	5,48	3,87	2,88	4,80		*40,50*
O6	5,45	4,82	3,52	4,48	2,96	

	Dywergencja / Clark		
klastry	**IV1**	**IV2**	**IV3**
IV1		*11,07*	*13,97*
IV2	1,61		*0,59*
IV3	1,83	0,27	

$$s_1(k_i, k_j) = \frac{1}{n_{k_i} n_{k_j}} \sum_{\substack{\boldsymbol{x} \in k_i \\ \boldsymbol{y} \in k_j}} d^2(\boldsymbol{x}, \boldsymbol{y}), \tag{4}$$

gdzie n_k jest liczbą wektorów w klastrze k.

Fragment uzyskanych wyników prezentuje tabela 4. Pominięto w niej metryki Soergela, Jaccarda i Dice, gdyż dawały one bardzo podobne rezultaty do metryki Sorensena. Wykluczono także metrykę Czebyszewa ze względu na to, że wszystkie jej wartości były równe 1.

Omawiana miara jest najmniej wrażliwa na obiekty znacznie odbiegające od centrum klastra, dlatego jej wyniki mogą zostać uznane za najbardziej wiarygodne. Otrzymane rezultaty pokazują pełną zgodność wszystkich metryk co do separacji skupisk. Oczywiście poszczególne metryki często znacząco różnią się wartościami, jednak wskazanie najlepiej odseparowanych skupisk jest dla każdej z nich identyczne. Można to zaobserwować na przykładzie odległości między klastrem O1 a pozostałymi skupiskami. Wszystkie metryki wskazują, że największą odległością jest O1-O4, a kolejne separacje to O1-O3, O1-O5, O1-O6 i O1-O2. Prezentowany wynik jest bardzo zbliżony do rezultatu odległości między centrami klastrów – jedyną różnicą jest nieco większa wartość odległości O1-O6 w przypadku miary s_2.

Zgodności między metrykami potwierdzają również pozostałe zbiory porównawcze, a wśród nich zbiór **IV** również zaprezentowany w tabeli 4. W jego przypadku największa odległość dzieli skupiska IV1-IV3 oraz IV1-IV2. Klastry IV2 i IV3 położone są znacznie bliżej siebie.

3.4. Separowalność międzyklastrowa w całej przestrzeni danych

Miara prezentowana w tej sekcji wyznacza wartość separacji nie dla jednej pary klastrów, jak czynią to miary opisane w sekcjach 3.1-3.3. Tym razem badana jest separowalność między klastrami dla całej przestrzeni danych analizowanego zbioru. Miara ta bierze więc pod uwagę wszystkie klastry jednocześnie, dzięki czemu staje się wygodnym wskaźnikiem pozwalającym ocenić jakość całej struktury klastrów. Wzór opisujący tę miarę jest następujący:

$$s(s_1) = \sum_{\substack{i,j=1 \\ j \neq i}}^{K} \frac{s_1(k_i, k_j)}{\sigma_1(k_i)}, \tag{5}$$

gdzie K jest liczbą klastrów w badanym zbiorze danych.

Wartość σ_1 z mianownika wzoru (5) jest miarą rozproszenia klastra k_i. Rozproszenie wszystkich klastrów każdego z siedmiu zbiorów danych zostało obliczone przy użyciu analizowanych metryk w pracy [19]. Jej celem było badanie wpływu metryk na wynik rozproszenia skupisk. W niniejszej pracy skorzystano z tych rezultatów. Miara σ_1 nazywana jest średnim rozproszeniem klastra. Jej wyznaczenie sprowadza się do zsumowania odległości między

każdą parą wektorów klastra k i podzielenia uzyskanej sumy przez liczbę par m tego klastra, co wyraża wzór:

$$\sigma_1(k) = \frac{1}{m} \sum_{\substack{\boldsymbol{x} \in k \\ \boldsymbol{y} \in k}} d^2(\boldsymbol{x}, \boldsymbol{y}), \tag{6}$$

gdzie $m = \frac{n_k(n_k - 1)}{2}$, natomiast n_k jest licznością klastra k.

Wyniki miary $s(s_1)$ dla wszystkich zbiorów danych i metryk przedstawiają wykresy na rysunku 2. W zależności od zastosowanej metryki rezultaty pokazują kilka dysproporcji mogących powodować problem we właściwej interpretacji rozproszenia.

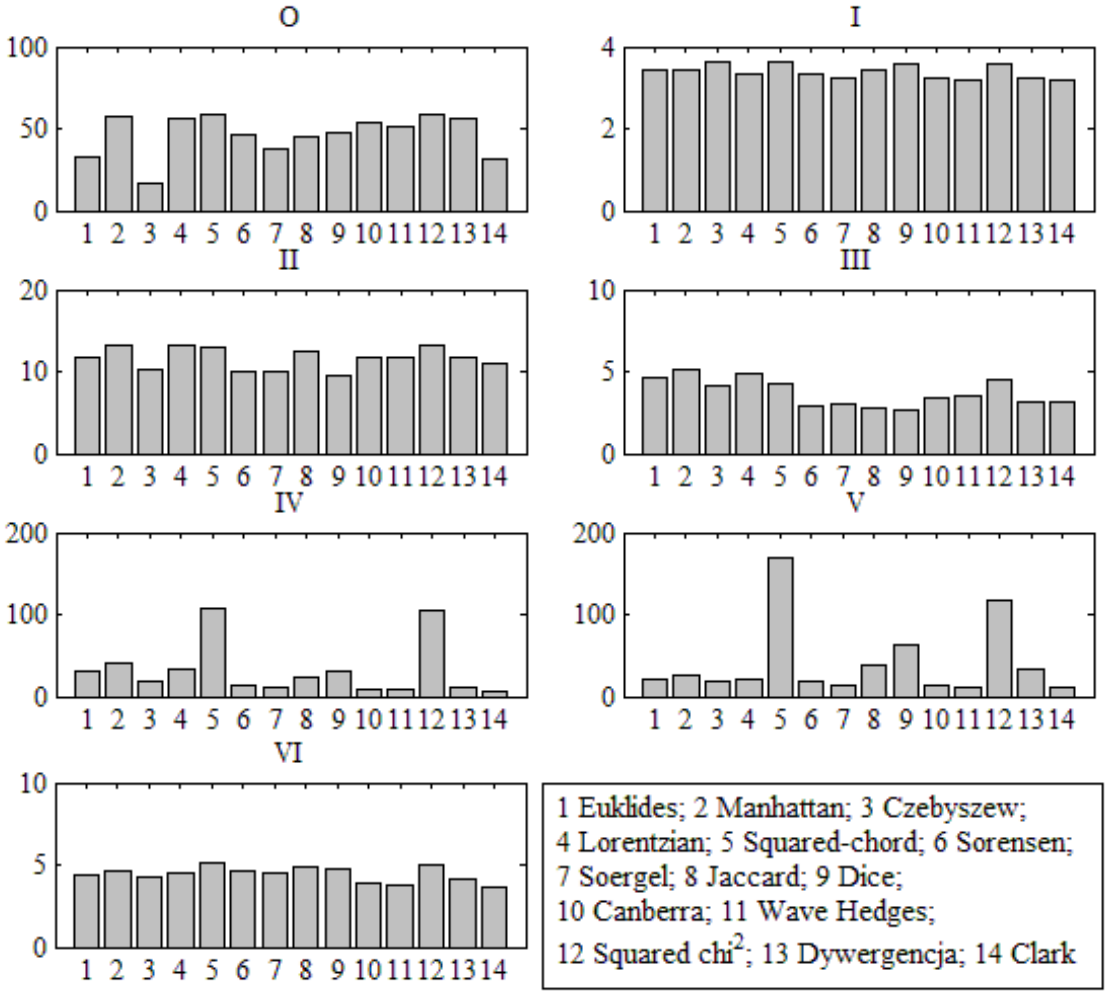
Porównując ze sobą **IV** i **V** zbiór danych, zwracają uwagę dwie najwyższe wartości obliczone za pomocą metryk Squared-chord oraz Squared χ^2. Obie metryki oraz większość pozostałych sugerują, że zbiór **V** zawiera bardziej rozproszoną strukturę klastrów w porównaniu ze zbiorem **IV**. Jednak biorąc pod uwagę metrykę Euklidesa, Manhattan lub Lorentzian, należałoby wskazać klastry zbioru **IV** jako lepiej odseparowane.

Porównanie zbiorów **O** i **V** także daje nieprecyzyjną informację na temat ogólnej separacji tych zbiorów. Część metryk zwraca większą separację zbioru **O**. Przykładowo, metryka Manhattan przyjmuje wartość 56,87 dla **O** i 24,90 dla **V**. Jednak wspomniane już wcześniej metryki Squared-chord oraz Squared χ^2 prezentują wyraźnie odmienną sytuację. Pierwsza z nich dla zbioru **V** daje wynik 169,67, zaś dla **O** tylko 58,84.

Kolejna niejednoznaczna sytuacja wynika z porównania zbiorów **I** oraz **III**. W przypadku zbioru **I** metryka Euklidesa i Manhattan przyjmują wartości odpowiednio: 3,42 i 3,41. Dla zbioru **III** będą to natomiast wyniki równe 4,66 i 5,18, co sugeruje większą separację zbioru **III**. Jednak patrząc, przykładowo, na metryki Sorensena (3,31 dla **I** i 2,85 dla **III**) lub Soergela (3,24 dla **I** i 2,97 dla **III**), wniosek na temat separowalności tych zbiorów mógłby być odwrotny.

4. Podsumowanie

W niniejszej pracy badano wpływ wybranych metryk na wynik miar separacji między klastrami zbioru danych. Zastosowano cztery miary separacji o różnej złożoności obliczeniowej. Trzy z nich badały separację każdej pary klastrów z osobna, natomiast czwarty wskaźnik mierzył ogólną separowalność całej struktury klastrów. Wszystkie te miary w nieco odmienny sposób traktują odległości pomiędzy skupiskami danych, jednak ich interpretacja jest zawsze taka sama: im większa separowalność dwóch skupisk, tym mniejsze podobieństwo pomiędzy nimi, co w praktyce uważa się za optymalne zjawisko.



Rys. 2. Rezultaty miary $s(s_1)$
Fig. 2. Results of $s(s_1)$ measure

Rezultaty wszystkich czterech miar pokazują często wyraźnie różne wartości obliczone za pomocą poszczególnych metryk. Takie rozbieżności biorą się z innego sposobu wyznaczania każdej metryki, dlatego są oczywistą sytuacją. Poważnym problemem mogą być jednak takie sytuacje, w których dana metryka uzna za najlepiej odseparowaną pewną parę klastrów, natomiast inna metryka dla tego samego zbioru danych wskaże separację zupełnie innych klastrów jako największą.

Pierwsze miary, jakie przeanalizowano to odległość między parą najbliżej położonych od siebie obiektów z dwóch klastrów oraz odległość między centrami klastrów. Obie miary zwracały odbiegające od siebie rezultaty, co pokazano na przykładzie zbioru podstawowego. Wynika to z innego traktowania separacji i nie mają na to bezpośredniego wpływu stosowane metryki. Wyjątkiem była tylko metryka Canberra, wskazująca odmienną separację poszczególnych klastrów podstawowego zbioru danych w porównaniu z innymi metrykami, a także metryka Czebyszewa, która zwraca dla każdej pary skupisk wartość 1.

W przypadku najbardziej precyzyjnej miary, która analizuje rozkład wszystkich wektorów tworzących klastry, stosowane metryki dawały zgodne rezultaty jednoznacznie świadczące o separacji klastrów. Natomiast ostatnia miara, która tym razem odnosi się do całej przestrzeni danych, bierze pod uwagę również wyniki rozproszenia klastrów. Uzyskiwane dla niej rezultaty powodują najwięcej problemów we właściwej interpretacji. Poszczególne metryki wyraźnie wpływają na wynik separacji, co uniemożliwia porównanie zbiorów danych i udzielenie odpowiedzi na pytanie, w którym zbiorze znajduje się optymalna struktura klastrów.

Podsumowując, najlepszym sposobem obliczania separacji spośród przeanalizowanych miar jest analiza rozkładu wszystkich wektorów badanych klastrów. Stosowane tutaj metryki nie wpływały na zmianę wyników separacji. Jednak gdy do analizy zostanie włączone rozproszenie klastrów, wówczas metryki zaczynają wyraźnie wpływać na uzyskiwane rezultaty, utrudniając interpretację separowalności całego zbioru danych.

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. Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.R.: Model-based evaluation of clustering validation measures. Pattern Recognition, Vol. 40, No. 3, Elsevier, 2007, s. 807÷824.
3. 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.
4. Cox T.F., Cox M.A.A: Multidimensional Scaling, 2nd edition. Chapman & Hall/CRC Press, 2000.
5. Deza M.M., Deza E.: Encyclopedia of distances. Springer-Verlag, Berlin, Heidelberg 2009.
6. 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, No. 2, 2003, s. 5÷12.
7. Everitt B.S., Landau S., Leese M.: Cluster analysis. Wiley Publishing, Nowy Jork 2009.

8. 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.
9. Gordon A.D.: Classification, 2nd edition. Chapman & Hall/CRC Press, 1999.
10. Hand D., Mannila H., Smyth P.: Eksploracja danych. WNT, Warszawa 2005.
11. Jain A.K., Dubes R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey 1988.
12. Jain A.K., Murty M.N., Flynn P.J.: Data clustering: a review. ACM Computing Surveys, Vol. 31, No. 3, 1999, s. 264÷323.
13. Krause E.F.: Taxicab Geometry: An Adventure in Non-Euclidean Geometry. Dover, New York 1986.
14. 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.
15. Meila M.: Comparing clusterings – an information based distance. Journal of Multivariate Analysis, Vol. 98, No. 5, 2007, s. 873÷895.
16. Monev V.: Introduction to similarity searching in chemistry. MATCH Communications in Mathematical and in Computer Chemistry, Vol. 51, 2004, s. 7÷38.
17. Osowski S.: Metody i narzędzia eksploracji danych. Wydawnictwo BTC, Legionowo 2013.
18. Paśko Ł., Setlak G.: Ocena segmentacji rynku za pomocą miar jakości grupowania danych. Zeszyty Naukowe Politechniki Śląskiej, Seria Informatyka, Vol. 35, No. 2(116), Gliwice 2014, s. 157÷173.
19. Paśko Ł., Setlak G.: Wpływ wybranych metryk na wynik badania skupisk. Zeszyty Naukowe Politechniki Śląskiej, Seria Informatyka, Vol. 36, No. 1(119), Gliwice 2015, s. 31÷45.
20. Setlak G., Paśko Ł.: Zastosowanie metod eksploracji danych do segmentacji rynków. Zeszyty Naukowe Politechniki Śląskiej, Seria Informatyka, Vol. 34, No. 2A(111), Gliwice 2013, s. 311÷323.
21. http://sci2s.ugr.es/keel/datasets.php – wykorzystane w czasie badań zbiory danych od I do VI wraz z ich opisem – ostatni dostęp 7.02.2016 r.

Abstract

This paper explores the influence of selected metrics on the results of separation between clusters in a dataset. Four indices of cluster separation were analyzed: the distance between

a pair of nearest objects from each of two clusters, the distance between the centers of the clusters, the distribution of all vectors that forming clusters, and the measure of overall separation of data space. The separation measures are calculated during the validation of quality of clusters.

The first section introduces theoretical foundation on clustering, cluster validation, and distances in feature space. In the second section, fourteen metrics used during the analysis are presented (table 1), and seven analyzed datasets are introduced. Next section is divided into four subsections, which contains the results of the measures. Table 3 shows the outcomes of the first separation measure given by equation (1). Figure 1 presents the results of second indicator for one of the datasets. Table 4 contains the results of the most accurate measure of separation expressed by formula (4). Figure 2 consists of seven bar graphs that show the overall separation of the dataset, depicted for all metrics and all datasets.

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.