

STUDIA METODOLOGICZNE

Marcin SALAMAGA

Miernik podobieństwa grupowania obiektów

Istotą grupowania obiektów jest wyodrębnienie jak najbardziej jednorodnych grup obiektów ze względu na podobieństwo wewnętrznej struktury charakteryzujących je zmiennych. Od procedury grupowania wymaga się spełnienia kryteriów homogeniczności oraz heterogeniczności (Dobosz, 2004). Oznacza to, że obiekty należące do tej samej grupy powinny być do siebie jak najbardziej podobne, natomiast obiekty należące do różnych grup powinny być jak najmniej podobne. Wyodrębnianie skupień obiektów powinno się odbywać m.in. zgodnie z zasadami zupełności i rozłączności.

W literaturze przedmiotu istnieje wiele różnych metod grupowania obiektów, które nierzadko prowadzą do rozbieżnych wyników w zakresie liczby tworzonych skupień oraz ich składu. Autorzy niektórych opracowań naukowych, w których stosowano co najmniej dwie metody grupowania obiektów, porównując otrzymane rezultaty nie zawsze dokonują wyczerpującej oceny stopnia podobieństwa wyników grupowania. Tymczasem zastosowanie właściwych mierników umożliwia precyzyjny pomiar podobieństwa podziałów. To pozwala odpowiedzieć na pytanie, które metody dostarczyły najbardziej podobnych wyników grupowania obiektów oraz w przypadku których metod wyniki grupowania były najmniej podobne? Taka ocena ma szczególne znaczenie, jeśli liczba skupień oraz liczba grupowanych obiektów jest znaczna. Problematyka pomiaru podobieństwa podziałów obiektów uzyskanych różnymi metodami jest od dawna podejmowana, np. w pracach: P. Arabie i S. A. Boormana (1973), E. B. Fowlkesa i C. L. Mallowsa (1983), L. J. Huberta i P. Arabie (1985), P. Jaccarda

(1908), E. Nowaka (1985), W. M. Randa (1971) czy Cz. Szmięła (1976). Nasuwają się przy tym następujące pytania:

1. Która z metod grupowania obiektów daje rezultaty najbardziej zbliżone do optymalnego rozwiązania w sensie określonego kryterium?
2. W jakim stopniu wyniki dyskryminacji obiektów uzyskane różnymi metodami są zbliżone (lub rozbieżne)?
3. W jakim stopniu podziały obiektów uzyskane tą samą metodą zmieniają się na skutek zmian relacji pomiędzy porównywanymi obiektami?

Istnieje wiele propozycji dotyczących pomiaru podobieństwa klasyfikacji obiektów. Należy tu wymienić wskaźniki zaproponowane np. przez P. Jaccarda (1908), W. M. Randa (1971), P. Arabie i S. A. Boormana (1973), E. B. Fowlkesa i C. L. Mallowsa (1983), E. Nowaka (1985), A. Sokołowskiego (1995) i Cz. Szmięła (1976).

Celem artykułu jest zaproponowanie kolejnego wskaźnika podobieństwa wyników grupowania obiektów. Omówiono również jego własności oraz przedstawiono zastosowanie na kilku przykładach.

KONSTRUKCJA WSKAŹNIKA PODOBIEŃSTWA GRUPOWAŃ

Podstawa konstrukcji proponowanego wskaźnika podobieństwa podziałów zbioru obiektów nawiązuje do budowy wskaźników zgodności podziałów Cz. Szmięła oraz E. Nowaka. Podobnie jak w wymienionych miernikach, wykorzystano sumę oraz iloczyn mnogościowy zbiorów.

Załóżmy, że znane są rezultaty dwóch grupowań tych samych obiektów. W wyniku grupowania G_1 uzyskano $1, 2, \dots, k$ grup obiektów, a w wyniku grupowania G_2 uzyskano $1, 2, \dots, l$ grup obiektów. Powstaje pytanie, na ile wyniki obu grupowań są podobne? Odpowiedź można znaleźć obliczając wartość wskaźnika podobieństwa wyników grupowania S .

Proponuję następujący sposób obliczenia wskaźnika podobieństwa wyników grupowania:

$$S = \frac{\sum_{i=1}^k \max_j P_{ij} + \sum_{j=1}^l \max_i P_{ij}}{k + l} \quad (1)$$

Wskaźnik P_{ij} określa podobieństwo dwóch zbiorów (skupień): A_i — skupienia utworzonego w grupowaniu G_1 oraz B_j — skupienia utworzonego w grupowaniu G_2 dla $i \in \{1, 2, \dots, k\}, j \in \{1, 2, \dots, l\}$. Jego wartość można obliczyć zgodnie ze wzorem (2).

$$P_{ij} = \frac{\overline{\overline{A_i \cap B_j}}}{\overline{\overline{A_i \cup B_j}}} \quad (2)$$

gdzie:

$\overline{A_i \cap B_j}$ — liczba elementów zbioru będącego iloczynem mnogościowym zbiorów A_i oraz B_j ,

$\overline{A_i \cup B_j}$ — liczba elementów zbioru będącego sumą mnogościową zbiorów A_i oraz B_j .

Jeśli zbiory A_i oraz B_j zawierają wyłącznie identyczne obiekty (czyli $A_i \cap B_j = A_i \cup B_j$), to wskaźnik $P_{ij} = 1$. Jeżeli zbiory są rozłączne (czyli $A_i \cap B_j = \emptyset$), to $P_{ij} = 0$. Konstrukcja wskaźnika podobieństwa grupowania wymaga obliczenia wartości wskaźnika P_{ij} dla wszystkich par skupień obiektów z obu grupowań $G1$ i $G2$, a następnie polega na wyborze maksymalnej wartości P_{ij} (dla pary skupień najbardziej podobnych).

Warto zwrócić uwagę, że wyrażenie $\max_i P_{ij}$ w e wzorze (1) określa stopień podobieństwa wyników grupowania $G1$ w stosunku do wyników grupowania $G2$, natomiast wyrażenie $\max_j P_{ij}$ określa stopień podobieństwa wyników grupowania

$G2$ w stosunku do wyników grupowania $G1$. Wskaźnik S jest więc w pewnym sensie agregatem podobieństwa rezultatów grupowania $G1$ do $G2$ i odwrotnie. Godzi się też zauważyć, że nie musi zachodzić równość: $\max_i P_{ij} = \max_j P_{ij}$.

Wartości wskaźnika S są unormowane w przedziale $[0;1]$. Wskaźnik przyjmuje wartość 1, jeśli wyniki grupowania $G1$ oraz $G2$ są identyczne (taka sama liczba grup oraz identyczne składy odpowiednich skupień). Wartość 0 jest możliwa, gdy żaden z obiektów grupowania $G1$ nie pojawia się w wynikach grupowania $G2$ (i odwrotnie), czyli gdy grupowaniu podlegają obiekty należące do rozłącznych zbiorów. Jeżeli grupowaniu podlegają zbiory, których część wspólna jest niepusta, to wartość wskaźnika S jest zawsze większa od zera. Zatem jeśli ten sam zbiór obiektów jest poddawany grupowaniu różnymi metodami, to wartość wskaźnika S należy do przedziału $(0;1]$.

Im wartość wskaźnika S jest bliższa 1, tym większe jest podobieństwo wyników grupowania obiektów. Z kolei im wartość wskaźnika S jest bliższa 0, tym mniejsze jest podobieństwo wyników dyskryminacji zbioru obiektów. Wartość wskaźnika S zależy m.in. od liczby identycznych obiektów w porównywanych skupieniach w grupowaniu $G1$ i $G2$ oraz od liczby utworzonych skupień w obu dyskryminacjach zbioru. Większe rozproszenie obiektów między skupieniami oraz zwiększająca się liczba skupień w jednym grupowaniu mogą zmniejszać stopień podobieństwa dyskryminowanego zbioru obiektów. Zaletą wskaźnika S jest możliwość zastosowania go do różnej liczby skupień tworzonych w wyniku grupowania $G1$ i $G2$. Wskaźnik ten daje możliwości procentowej oceny zgodności wyników grupowania obiektów.

Konstrukcja wskaźnika S różni się od popularnych mierników podobieństwa podziałów zbioru, takich jak wskaźnik Jaccarda, Randa, Arabie i Boormana, a także Fowlkesa i Mallowsa. Idea budowy tych wskaźników wykorzystuje m.in. obecność identycznych par obiektów znajdujących się w zbiorach utworzonych w dwóch odrębnych grupowaniach czy par obiektów, które nie występują w żadnym grupowaniu. Tymczasem konstrukcja wskaźnika S opiera się na podobieństwie zbiorów w zdefiniowanych miernikiem P_{ij} . Zatem brak jednakowych par obiektów w dwóch grupowaniach wyklucza podobieństwo podziału np. w sensie Jaccarda lub Fowlkesa i Mallowsa, ale niekoniecznie w sensie wskaźnika S .

Jak już wspomniano, konstrukcja wskaźnika S wykorzystuje koncepcję podobieństwa zbiorów zaproponowaną w pracach Szmigla. Z kolei sposób agregacji składowych, określających stopień wzajemnego podobieństwa wyników dwóch grupowań w budowie wskaźnika S , nawiązuje do budowy miernika zaproponowanego w pracy Nowaka. Wymienione mierniki zgodności podziałów dwóch zbiorów są unormowane w przedziale $[0,1]$, a im ich wartości są bliższe 1, tym większa jest zgodność wyników podziałów zbioru obiektów.

Warto zaznaczyć, że wartość wskaźnika (1) jest wrażliwa na nierówne liczby skupień w obu dyskryminacjach zbioru, jak również na stopień niesymetrycznego rozproszenia obiektów między skupieniami w porównywanych grupowaniach. Wskaźnik ten może być też wykorzystany do badania zgodności wyników dyskryminacji dwóch nieidentycznych zbiorów obiektów (czyli zbiorów, których różnica jest zbiorem niepustym). Wówczas jednak wartość wskaźnika S będzie zawsze mniejsza od 1.

Oto przykład zastosowania wskaźnika S do oceny podobieństwa grupowania w trzech równolicznych skupieniach. Grupowaniu podlegają obiekty oznaczone jako: a, b, c, d, e, f, g, h, i. W tabl. 1 podano składy skupień w poszczególnych grupowaniach obiektów, a w tabl. 2 znajdują się obliczenia pomocnicze do wyznaczenia wartości wskaźnika podobieństwa utworzonych zbiorów.

TABL. 1. SKŁAD SKUPIEŃ UTWORZONYCH W GRUPOWANIACH G1—G5

Grupowania	Numery skupienia		
	1 2		3
G1	a, b, c	d, e, f	g, h, i
G2	a, b, c	d, e, f	g, h, i
G3	a, b, d	c, e, f	g, h, i
G4	a, b, d	c, e, g	f, h, i
G5	a, d, g	b, e, h	c, f, i

Źródło: opracowanie własne.

Szczególony sposób obliczania wskaźników P_{ij} podobieństwa grupowania $G1$ względem $G2$ wygląda następująco:

$$P_{11} = \frac{3}{3} = 1, \quad P_{12} = \frac{0}{6} = 0, \quad P_{13} = \frac{0}{6} = 0, \quad \max P_{1j} = \max \{1, 0, 0\} = 1$$

$$P_{21} = \frac{0}{6} = 0, \quad P_{22} = \frac{3}{3} = 1, \quad P_{23} = \frac{0}{6} = 0, \quad \max P_{2j} = \max \{0, 1, 0\} = 1$$

$$P_{31} = \frac{0}{6} = 0, \quad P_{32} = \frac{0}{6} = 0, \quad P_{33} = \frac{3}{3} = 1, \quad \max P_{3j} = \max \{0, 0, 1\} = 1$$

zatem:

$$\sum_{i=1}^3 \max_j P_{ij} = 1 + 1 + 1 = 3$$

Zauważmy również, że wyniki grupowania $G1$ i $G2$ są takie same (tabl. 1). Dlatego odpowiednie obliczenia w wskaźnikach podobieństwa grupowania $G2$ względem $G1$ są identyczne, jak w przypadku grupowania $G1$ względem $G2$ i zostaną dalej pominięte.

Dla rezultatów w grupowań $G1$ i $G2$ wartość wskaźnika S zgodnie ze wzorem (1) można wyznaczyć następująco:

$$S = \frac{3+3}{3+3} = 1$$

**TABL. 2. OBLICZENIA POMOCNICZE DO WSKAŹNIKA
PODOBIEŃSTWA GRUPOWAŃ $G1$ — $G5$**

Grupowania	Max P_{ij}	Numery skupienia			Suma	Wskaźnik S
		1 2		3		
$G1$ i $G2$	$\max_i P_{ij}$	1,000 1,	000	1,000 3,	000	1,000
	$\max_j P_{ij}$	1,000 1,000		1,000 3,000		
$G1$ i $G3$	$\max_i P_{ij}$	0,500 0,	500	1,000 2,	000	0,667
	$\max_j P_{ij}$	0,500 0,	500	1,000 2,	000	
$G1$ i $G4$	$\max_i P_{ij}$	0,500 0,	200	0,500 1,	200	0,400
	$\max_j P_{ij}$	0,500 0,	200	0,500 1,	200	
$G1$ i $G5$	$\max_i P_{ij}$	0,200 0,	200	0,200 0,	600	0,200
	$\max_j P_{ij}$	0,200 0,	200	0,200 0,	600	

Źródło: opracowanie własne.

W tabl. 1 podano niektóre warianty konfiguracji obiektów w równolicznych grupach, dokonując stopniowo coraz liczniejszych przesunięć obiektów pomiędzy poszczególnymi skupieniami w kolejnych podziałach zbioru $\{a, b, c, d, e, f, g, h, i\}$. W miarę wzrostu stopnia niezgodności składu grup, wartość wskaźnika S obniża się o d poziomu 1 00% (w przypadku pełnej zgodności wyników aż do poziomu 20% — tabl. 2). Dalsze „pogorszenie” stopnia podobieństwa wyników

grupowania można uzyskać zmieniając liczbę tworzonych grup i „rozpraszając” obiekty pomiędzy takimi skupieniami. W tabl. 3 przedstawiono skrajny przypadek następującej sytuacji: w grupowaniu G_6 mamy jedno skupienie $\{a, b, c, d, e, f, g, h, i\}$, a w grupowaniu G_7 mamy dziewięć jednoelementowych skupień $\{a\}$, $\{b\}$, $\{c\}$, $\{d\}$, $\{e\}$, $\{f\}$, $\{g\}$, $\{h\}$, $\{i\}$. Na podstawie danych z tabl. 3 wartość wskaźnika S można obliczyć następująco:

$$S = \frac{0,111 + 0,999}{1 + 9} = 0,111$$

Otrzymana wartość jest możliwie najniższa w przypadku grupowania tych samych 9 obiektów.

**TABL. 3. OBLICZENIA POMOCNICZE DO WSKAŹNIKA PODOBIEŃSTWA
DYSKRYMINACJI G_6 I G_7**

Numery skupienia w grupowaniach		Podobieństwo grupowania	
G_6	G_7	G_6 i G_7 — $\max_i P_{ij}$	G_7 i G_6 — $\max_j P_{ij}$
1 1		0,111	0,111
x 2		x	0,111
x 3		x	0,111
x 4		x	0,111
x 5		x	0,111
x 6		x	0,111
x 7		x	0,111
x 8		x	0,111
x 9		x	0,111
Suma 0,		111	0,999
Wskaźnik S 0,		111	

Źródło: opracowanie własne.

Niższa wartość wskaźnika S jest możliwa, gdy grupowane są dwa nieidentyczne zbiory obiektów. Zaznaczmy, że wskaźnik S może przyjąć wartość 0 jedynie w przypadku grupowania dwóch zbiorów obiektów, które są rozłączne.

PRZYKŁAD EMPIRYCZNY

Wskaźnik podobieństwa grupowania obiektów został wykorzystany do oceny stopnia zgodności wyników grupowania wybranych funduszy inwestycyjnych akcji, zrównoważonego wzrostu i stabilnego wzrostu działających w Polsce w latach 1999—2005 przy użyciu trzech metod: algorytmu eliminacji wektorów, metody Warda oraz metody k -średnich. Celowo zastosowano metody różniące się właściwościami i celami dyskryminacji, aby uzyskać skupienia możliwie różnych składach. Grupowanie funduszy przeprowadzono uwzględniając prze-

ciężny udział walorów z poszczególnych sektorów polskiej gospodarki w portfelach inwestycyjnych w kolejnych półroczach lat 1999—2005.

ZESTAWIENIE. WYBRANE DZIAŁY I SEKTORY GOSPODARKI

Działy gospodarki	Sektory
Przemysł	spożywczy lekki drzewny chemiczny materiałów budowlanych budownictwo elektromaszynowy metalowy inne
Finanse	bankowy ubezpieczeniowy inne
Usługi	handel informatyka telekomunikacja media inne
Inne	x

Źródło: opracowanie własne.

Do klasyfikacji funduszy inwestycyjnych wykorzystano najpierw algorytm eliminacji wektorów opracowany przez S. Chomątowskiego i A. Sokołowskiego (1978). Stosując kolejne kroki algorytmu obliczono elementy macierzy braku podobieństwa struktur portfeli inwestycyjnych poszczególnych funduszy. Na tej podstawie wyodrębniono cztery jednorodne grupy funduszy inwestycyjnych (za krytyczny poziom braku podobieństwa przyjęto $\alpha = 0,1$).

Wyniki grupowania funduszy inwestycyjnych przy zastosowaniu algorytmu eliminacji wektorów są następujące (Salamaga, 2008):

- **grupa 1:** DWS (akcji +), ING Akcji typu A i B, PKO/CS (akcji), SEB 3 (akcji), Skarbiec Akcja, UniK ORONA (akcji), ING (z równoważony), PKO/CS (zrównoważony), SEB 1 (zrównoważony),
- **grupa 2:** Pioneer AI FIO, Citi Akcji FIO, Citi Zrównoważony FIO (A),
- **grupa 3:** DWS (akcji), DWS (zrównoważony),
- **grupa 4:** ARKA 3 Zrównoważony FIO, Skarbiec WAGA, UniKORONA (zrównoważony), PKO/CS (stabilnego wzrostu).

W celu osiągnięcia większej przejrzystości prowadzonej analizy, wybranymi parametrami pozostałych metod sterowano tak, aby uzyskać taką samą liczbę skupień, jak przy użyciu metody eliminacji wektorów (jedynakowa liczba skupień oczywiście nie jest warunkiem koniecznym do obliczenia wartości wskaźnika S).

Wyniki grupowania funduszy inwestycyjnych metodą Warda (z odległością euklidesową) przedstawiono na wykresie (Salamaga, 2008).

„Przycinając” dendrogram według przyjętego kryterium jednakowej liczby skupień na wysokości wiązania równej 2,75 można wyodrębnić grupy funduszy uzyskane metodą Warda:

- **grupa 1:** ARKA 3 Zrównoważony FIO,
- **grupa 2:** DWS (akcji), DWS (zrównoważony),
- **grupa 3:** Citi Akcji FIO, Citi Zrównoważony FIO (A), Skarbiec WAGA, Skarbiec Akcja, DWS (akcji +), PKO/CS (stabilnego wzrostu), ING Akcji typu A i B, ING (zrównoważony), PKO/CS (zrównoważony), PKO/CS (akcji),
- **grupa 4:** Pioneer AI FIO, UniKORONA (akcji), UniKORONA (zrównoważony), SEB 3 (akcji), SEB 1 (zrównoważony).

Do pogrupowania funduszy inwestycyjnych według struktury branżowej portfeli inwestycyjnych zaproponowano w dalszej kolejności metodę k -średnich. Przyjęto, że $k = 4$ (aby uzyskać porównywalność grupowania co do liczby skupień z pozostałymi metodami). Wyniki grupowania funduszy inwestycyjnych metodą k -średnich są następujące:

- **grupa 1:** ARKA 3 Zrównoważony FIO,
- **grupa 2:** DWS (akcji), DWS (zrównoważony),

- **grupa 3:** Pioneer AI FIO, ING Akcji typu A i B, Citi Akcji FIO, PKO/CS (akcji), SEB 3 (akcji), UniK ORONA (akcji), Citi Zrów noważony FIO (A), ING (zrównoważony), PKO/CS (zrównoważony), UniKORONA (zrównoważony),
- **grupa 4:** DWS (akcji +), Skarbiec Akcji a, SEB 1 (zrównoważony), Skarbiec WAGA, PKO/CS (stabilnego wzrostu).

Aby zbadać, które metody grupowania funduszy inwestycyjnych dostarczyły najbardziej podobnych wyników, zastosowano wskaźnik podobieństwa grupowania S .

W tabl. 4 przedstawiono szczegółowe etapy obliczenia wskaźnika (1) dla poszczególnych par metod.

TABL. 4. OBLICZENIA POMOCNICZE DO WSKAŹNIKA PODOBIEŃSTWA GRUPOWANIA FUNDUSZY INWESTYCYJNYCH PRZY UŻYCIU METODY WARDA (W), ALGORYTMU ELIMINACJI WEKTORÓW (A) ORAZ METODY k -ŚREDNICH (K)

Metody grupowania	Max P_{ij}	Numery skupienia				Suma	Wskaźnik S
		1	2	3	4		
Metoda Warda i algorytm eliminacji wektorów	$\max_i P_{ij}$	0,250	1,000	0,462	0,143	1,855	0,469
	$\max_j P_{ij}$	0,462	0,182	1,000	0,250	1,894	
Algorytm eliminacji wektorów i metoda k -średnich	$\max_i P_{ij}$	1,000	1,000	0,429	0,364	2,793	0,698
	$\max_j P_{ij}$	1,000	1,000	0,429	0,364	2,793	
Metoda Warda i metoda k -średnich	$\max_i P_{ij}$	0,462	1,000	0,300	0,250	2,012	0,501
	$\max_j P_{ij}$	0,250	1,000	0,462	0,286	1,998	

Źródło: opracowanie własne.

Z obliczeń przedstawionych w tabl. 4 wynika, że najbardziej podobne wyniki grupowania uzyskano w przypadku metody k -średnich oraz algorytmu eliminacji wektorów (zgodność w 69,8%). Najniższy stopień podobieństwa wyników grupowania otrzymano natomiast stosując metodę Warda i algorytm eliminacji wektorów (zgodność w 46,9%). Naturalnie nie należy wnioskować, że stosując metodę eliminacji wektorów zawsze osiągniemy bardziej podobne wyniki grupowania do metody k -średnich. Trzeba wziąć pod uwagę, że w wymienionych metodach dyskryminacji zbioru występują parametry, których wartości są ustalane przez użytkownika w sposób arbitralny (poziom wskaźnik braku podobieństwa struktur w algorytmie eliminacji wektorów). W efekcie, w zależności od przyjętych wartości odpowiednich parametrów, możliwe jest uzyskanie mniej lub bardziej podobnych wyników grupowania w obu metodach.

Podsumowanie

Zastosowaniu kilku kryteriów dyskryminacji zbioru obiektów na rozłączne i niepuste podzbiory towarzyszy często pytanie: które z metod podziału zbioru dają najbardziej podobne wyniki grupowania obiektów, a które dają wyniki najmniej podobne? Odpowiedź można uzyskać obliczając wartość z aprezentowanego wskaźnika podobieństwa podziału zbiorów S . Wskaźnik S można również wykorzystać do oceny podobieństwa wyników dyskryminacji przeprowadzanej tą samą metodą, ale stosowaną w różnych okresach obserwacji wartości cech użytych do podziału zbioru. Zmieniające się wartości cech grupowanych obiektów mogą bowiem tworzyć inny skład powstających skupień obiektów w różnych momentach. Obliczając wartości wskaźnika (2) można określić, w których okresach (momentach) otrzymane wyniki grupowania były najbardziej podobne (lub niepodobne). Zaletą tego miernika jest również możliwość zastosowania go do badania podobieństwa podziału różnych zbiorów (kórnich różnica nie jest zbiorem pustym).

W celu pełnej oceny efektywności wskaźnika S w zakresie zdolności wykrywania podobieństwa (bądź braku podobieństwa) wyników dyskryminacji, wydaje się uzasadnione przeprowadzenie w kolejnych badaniach analizy porównawczej z innymi wskaźnikami podobieństwa podziału zbioru. Ciekawe wydaje się również porównanie stopnia skomplikowania procedury obliczania wartości mierników. Nie jest to bez znaczenia dla użytkowników metod, tym bardziej że wymienione wskaźniki na ogół nie doczekały się oprogramowania w popularnych pakietach komputerowych.

dr Marcin Salamaga — Uniwersytet Ekonomiczny w Krakowie

LITERATURA

- Arabie P., Boorman S. A. (1973), *Multidimensional scaling of measures of distance between partitions*, „Journal of Mathematical Psychology”, No. 10
- Chomątowski S., Sokołowski A. (1978), *Taksonomia struktur*, „Przegląd Statystyczny”, nr 2
- Dobosz M. (2004), *Wspomagana komputerowo statystyczna analiza wyników badań*, Akademicka Oficyna Wydawnicza EXIT, Warszawa
- Fowlkes E. B., Mallows C. L. (1983), *A Method for Comparing Two Hierarchical Clusterings*, „Journal of the American Statistical Association”, No. 78
- Hubert L. J., Arabie P. (1985), *Comparing partitions*, „Journal of Classification”, No. 1
- Jaccard P. (1908), *Nouvelles recherches sur la distribution florale*, „Vaudoise des Sciences Naturelles”, No. 44
- Nowak E., (1985), *Wskaźnik podobieństwa wyników podziału*, „Przegląd Statystyczny”, z. 1

- Pal N. R., Biswas J. (1997), *Cluster validation using graph theoretic concepts*, Pattern Recognition, vol. 30, No. 6
- Rand W. M. (1971), *Objective Criteria for the Evaluation of Clustering Methods*, „Journal of the American Statistical Association”, No. 66
- Salamaga M. (2008), *Badanie podobieństwa struktur portfeli funduszy inwestycyjnych w Polsce w latach 1999—2000. Modelowanie preferencji a ryzyko '07*, praca zbiorowa pod red. T. Trzaskalika, Wydawnictwo AE w Katowicach
- Sokołowski A. (1995), *Percentage Points of the Similarity Measure for Partitions*, „Statistics in Transition”
- Szmigiel Cz. (1976), *Wskaźnik zgodności kryteriów podziału*, „Przegląd Statystyczny”, nr 4

SUMMARY

There are many similarity measures used in cluster analysis, like Rand, Jaccard, Szmigiel or Sokołowski coefficients. In this paper the Author presented another measure of classification partition. The value of the proposed S-indicator belongs to $[0,1]$ interval. This depends on number of identical objects in comparative clusters as well as on number of clusters. This indicator can be used for comparison of different partitions, for example by different clustering methods, or by the same classification method if observations are from different periods S-indicator can be used to calculate similarity partition in two different sets.

РЕЗЮМЕ

Существуют многие показатели сходства группировки объектов, такие как показатели Жаккарда, Ранда, Шмигеля, или показатель Соколовского. Статья представляет очередное предложение показателя сходства дискриминации множества объектов. Значение показателя S нормализованное в интервале $[0,1]$. Оно зависит, между прочим, от числа тождественных объектов в сопоставляемых кластерах, а также от числа образованных кластеров. Предлагаемый показатель может использоваться для оценки соответствия результатов дискриминации проведенной этим же методом, но использованной в разные периоды наблюдения (обследования). Преимуществом показателя S является также возможность его использования для обследования сходства разделения разных множеств.