

Maria Kolenda

PROPOZYCJA NOWYCH MIERNIKÓW HETEROGENICZNOŚCI DO OCENY KLASYFIKACJI

1. Wstęp

W artykule zaproponowano mierniki do oceny jakości skonstruowanej klasyfikacji. Mierniki te oparto na określeniu heterogeniczności poszczególnych grup, w tym również grupy zawierającej jeden obiekt, czyli punktu izolowanego. Zdefiniowano trzy stopnie heterogeniczności dla grup oraz dwa stopnie heterogeniczności dla punktów izolowanych. Ponadto zdefiniowano brak heterogeniczności dla grupy oraz punktu izolowanego. Dla metod hierarchicznych zaproponowano prosty sposób wyboru podziału najlepszego.

W literaturze istnieje bardzo dużo różnych mierników heterogeniczności i homogeniczności, które oceniają zarówno pojedynczą grupę, jak i wszystkie grupy łącznie. Przykładowo w pracy [1] na s. 136 opisano takich mierników ponad 60, ale ocena klasyfikacji za ich pomocą wcale nie jest prosta – trudno bowiem poruszać się w takim gąszczu mierników, zwłaszcza że każdy z nich prezentowany jest za pomocą liczby. Dlatego autorka podjęła się próby skonstruowania (na bazie już istniejących mierników) takich nowych mierników, które byłyby miernikami syntetycznymi, uwzględniały heterogeniczność oraz homogeniczność pojedynczej grupy, miały pewną hierarchię i były przedstawione w postaci symbolu. Proponowane mierniki mają ułatwić oraz uprościć analizę heterogeniczności (a w pewnym sensie i homogeniczności) poszczególnych grup, a także umożliwić ocenę całego podziału.

W literaturze często do oceny klasyfikacji używa się miernika SC^1 oraz średniej wariancji wewnątrzgrupowej. Są to mierniki oceniające poprawność grupowania – syntetycznie oceniają wszystkie grupy i nie umożliwiają analizy heterogeniczności lub homogeniczności w poszczególnych pojedynczych grupach. W artykule oba te mierniki szczegółowo omówiono i porównano z miernikami propono-

¹ Miernik ten nazywany jest *Silhouette coefficient* (zob. [3. s. 320]).

wanymi przez autorkę. Do porównywania mierników wykorzystano empiryczny zbiór obiektów, dla którego za pomocą metody Warda utworzono hierarchiczną klasyfikację tych obiektów.

2. Mierniki typu h – mierniki heterogeniczności grupy

Do oceny stopnia heterogeniczności grupy autorka proponuje trzy następujące mierniki, oznaczone symbolami h^{**} , h^* , h , gdzie h^{**} oznacza *bardzo dobrą heterogeniczność* (odrębność) grupy, h^* – *dobłą heterogeniczność* grupy, h – *słabą, lecz dopuszczalną heterogeniczność* grupy.

Mierniki te są uporządkowane i mają określoną hierarchię. Największa heterogeniczność będzie oznaczona symbolem h^{**} , a najśłabsza symbolem h . Jeśli grupa ma heterogeniczność typu h^{**} , to ma również heterogeniczność typu h^* oraz h . Jeśli natomiast grupa ma heterogeniczność typu h , to nie ma heterogeniczności typu h^* ani typu h^{**} .

Dodatkowo definiuje się miernik h' , który głównie służy do oceny słabszej heterogeniczności punktu izolowanego.

Heterogeniczność punktu izolowanego proponuje się oceniać za pomocą mierników h^{**} oraz h' .

Mierniki te są nazywane w skrócie miernikami h . Inspiracją do zdefiniowania mierników h był znany w literaturze miernik L^{*2} .

Definicje mierników heterogeniczności

Podstawą proponowanych przez autorkę mierników heterogeniczności są znane z literatury następujące mierniki³:

Sep_C – miernik określający odrębność (separację) grupy (*separation of a cluster*) C , gdzie:

$$Sep_C = \min_{t \in C, h \notin C} d_{t,h}. \quad (1)$$

Miernik Sep_C oznacza minimalną odległość grupy C od innej grupy (minimalna odległość punktów grupy C od punktów nie należących do grupy C);

D_C – średnica taksonometryczna (*diameter of a cluster*) grupy C , gdzie:

$$D_C = \max_{i,j \in C} d_{i,j}. \quad (2)$$

² Zob. [3], s. 319, miernik L^* – drugi typ izolacji (*isolated group*). Proponowany miernik h^* zdefiniowany jest tak samo jak miernik L^* . Zmianę oznaczenia L^* na h^* wprowadzono dla porządku.

³ Zob. np. [3; 1].

Miernik D_C oznacza maksymalną odległość między dwoma punktami należącymi do grupy C .

D_l – średnica taksonometryczna l -tej grupy w danej klasyfikacji, gdzie $l = 1, 2, \dots, k$ (k – liczba grup) – skonstruowany jak D_C .

1. Definicja miernika h^{**}

Grupa C ma heterogeniczność typu h^{**} – ma bardzo dobrą heterogeniczność, jeśli:

$$Sep_C = \max_l (D_l), \quad (3)$$

gdzie D_l oznacza średnicę taksonometryczną l -tej grupy.

Grupa C ma bardzo dobrą heterogeniczność, jeśli odrębność grupy C (wyznaczona za pomocą miernika Sep_C) jest większa od maksymalnej średnicy taksonometrycznej grup. Według tego miernika można określać również odrębność punktu izolowanego.

2. Definicja miernika h^*

Grupa C ma heterogeniczność typu h^* – ma dobrą heterogeniczność, jeśli:

$$Sep_C > D_C. \quad (4)$$

Grupa C ma dobrą heterogeniczność jeśli odrębność grupy C jest większa od średnicy tej grupy. Według tego miernika nie można określać odrębności punktu izolowanego.

3. Definicja miernika h

Grupa C ma heterogeniczność typu h – ma słabą, lecz dopuszczalną heterogeniczność, jeśli:

$$Sep_C > \max_i \left(d_{i, R} \right), \text{ gdzie: } R \text{ reprezentant grupy } C. \quad (5)$$

Grupa C ma słabą, lecz dopuszczalną heterogeniczność, jeśli odrębność grupy C jest większa od maksymalnej odległości obiektów od reprezentanta R w grupie C (reprezentantem może być np. obiekt o najmniejszej średniej odległości w grupie C). Według tego miernika nie można określać odrębności punktu izolowanego.

4. Definicja miernika h'

Grupa C jest heterogeniczna według miernika h' , jeśli:

$$Sep_C > \frac{1}{k-p} \sum_{l=1}^{k-p} D_l, \text{ gdzie: } k - \text{liczba grup, } p - \text{liczba obiektów izolowanych.} \quad (6)$$

Wartość Sep_C (odrębność grupy C) jest większa od średniej wartości średnicy grup bez punktów izolowanych.

Jeśli grupa ma heterogeniczność typu h^{**} , to ma też heterogeniczność typu h' .

Miernik h' wprowadzono głównie do oceny słabszej heterogeniczności punktu izolowanego. Może on być również wyznaczany dla grupy zawierającej więcej niż jeden punkt, wtedy jednak miernik h' będzie mierzył inną heterogeniczność niż h^* lub h i będzie on od tych mierników niezależny.

5. Definicja braku heterogeniczności grupy

Grupa nie jest heterogeniczna (nie ma odrębności), jeśli nie ma heterogeniczności przynajmniej typu h , a dla obiektu izolowanego przynajmniej typu h' .

3. Mierniki typu H oceniające jakość klasyfikacji

Jakość dokonanej klasyfikacji proponuje się oceniać na podstawie oceny stopnia heterogeniczności wszystkich poszczególnych grup i punktów izolowanych w danej klasyfikacji. O jakości całej klasyfikacji decyduje najsłabsza heterogeniczność grupy lub punktu izolowanego w tej klasyfikacji. Do oceny klasyfikacji zdefiniowano trzy mierniki i na ich podstawie wyróżniono cztery rodzaje klasyfikacji. Oto definicje mierników H^{**} , H^* , H i poszczególnych rodzajów klasyfikacji:

H^{**} – miernik oznacza *bardzo dobrą klasyfikację*, jeśli wszystkie grupy w tej klasyfikacji mają heterogeniczność typu h^{**} ,

H^* – miernik oznacza *dobłą klasyfikację*, jeśli wszystkie grupy w tej klasyfikacji mają heterogeniczność przynajmniej typu h^* , a punkty izolowane przynajmniej typu h' i istnieje przynajmniej jedna grupa, która ma heterogeniczność typu h^* lub punkt izolowany typu h' ,

H – miernik oznacza *słabą, lecz dopuszczalną klasyfikację*, jeśli wszystkie grupy w tej klasyfikacji mają heterogeniczność przynajmniej typu h , a punkt izolowany przynajmniej typu h' i istnieje przynajmniej jedna grupa o heterogeniczności typu h lub punkt izolowany o heterogeniczności typu h' .

Definicja złej klasyfikacji

Klasyfikację uznaje się za *złą*, jeśli przynajmniej jedna z grup w danej klasyfikacji lub punkt izolowany *nie będą miały heterogeniczności*. Brak heterogeniczności grupy zdefiniowano w punkcie 1 definicji 5.

Powyższe mierniki w skrócie będą nazywane miernikami H .

4. Wybór „najlepszego” podziału w metodach hierarchicznych za pomocą mierników H oraz h

Określenie „najlepszy” podział zbioru obiektów (punktów w m -wymiarowej przestrzeni cech) zawsze budzi wątpliwości. Ogólnie wiadomo, że podział powi-

nien być taki, aby obiekty należące do wyodrębnionej grupy były podobne (punkty leżały blisko siebie), a obiekty należące do dwóch różnych grup były niepodobne (punkty leżały daleko od siebie). Do tego celu utworzono dość duży zbiór mierników, które oceniają zwartość grupy – homogeniczność – oraz odrębność grupy – heterogeniczność. Duża różnorodność związana jest ze sposobem pomiaru. Przykładowo: homogeniczność grupy można mierzyć za pomocą średniej odległości między punktami, średniej odległości od reprezentanta albo średniej odległości od środka ciężkości, można ją mierzyć również na podstawie wariancji odległości między obiektami w grupie.

Dobroć skonstruowanej klasyfikacji w literaturze ocenia się, porównując np. średnią homogeniczność wszystkich grup ze średnią heterogenicznością tych grup czy wariancję wewnątrzgrupową z wariancją ogólną. Problem w tym, że mierników jest dużo, a wartość ilorazu dodatkowo będzie zależała od liczby grup w danej klasyfikacji. Następną trudność stanowią punkty izolowane – mają one wprawdzie największą homogeniczność (zwartość), ale nie wiadomo, w jaki sposób oceniać ich odrębność. Poza tym konstruowanie klasyfikacji zależy od sformułowania celu, np. czy to ma być „naturalna” klasyfikacja, wynikająca z układu punktów w przestrzeni wielowymiarowej, czy klasyfikacja „na z góry określoną liczbę grup”.

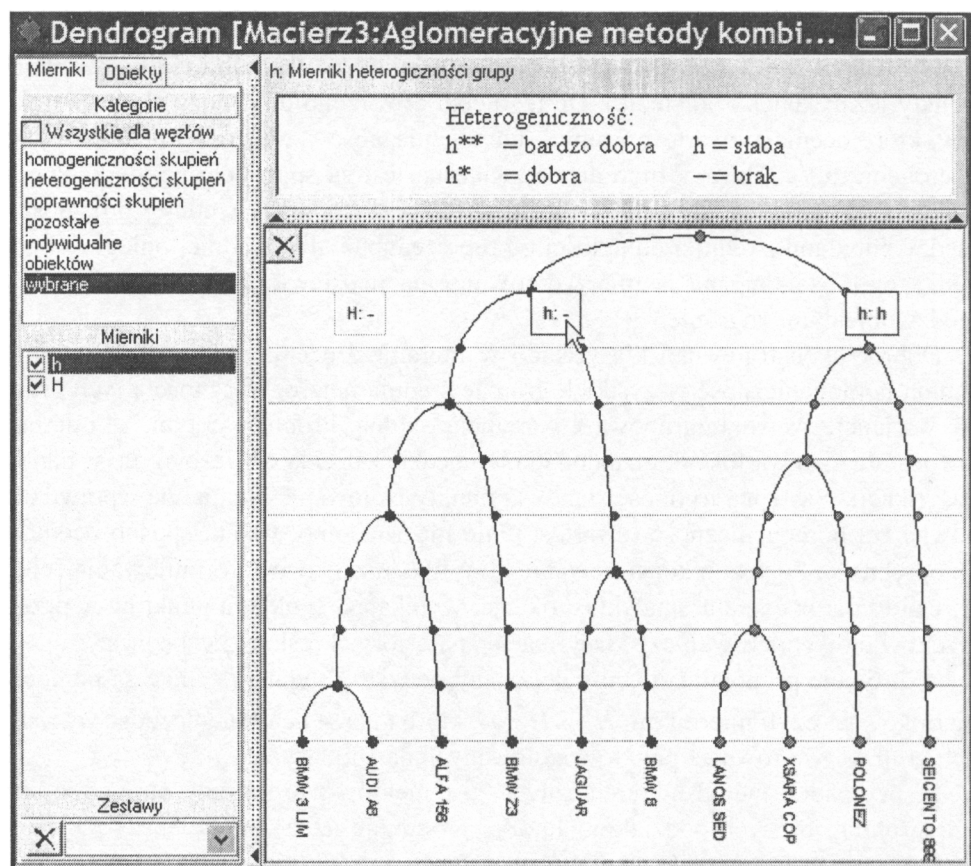
W artykule proponuje się, aby dokonany podział wstępnie oceniać za pomocą mierników H , czyli mierników H^{**} , H^* , H , które mierzą heterogeniczność wszystkich grup, w tym również punktów izolowanych danej klasyfikacji.

W przypadku metod hierarchicznych (np. metody Warda), gdy otrzymuje się całą strukturę różnych podziałów, autorka proponuje uznać za najlepszy taki podział, w którym jest mała liczba grup i w którym wszystkie grupy są heterogeniczne najlepiej typu H^{**} lub H^* , może być też typu H . Zatem w przypadku metod hierarchicznych należy sprawdzać, czy kolejne podziały – najpierw na dwie grupy, potem na trzy itd. – są heterogeniczne. Mierniki H^{**} , H^* , H będą się bowiem znacznie częściej pojawiały w przypadku większej liczby grup wyłonionych w danym podziale (w danym rzędzie dendrogramu).

Prezentowane rysunki dendrogramów wraz z wartościami mierników są zrzutami ekranowymi z programu komputerowego TAKSONOMIA⁴.

Na rys. 1 na pierwszym rzędzie od góry dendrogramu sprawdzono za pomocą mierników H podział zbioru samochodów na dwie grupy. Podziału dokonano w programie „Taksonomia numeryczna”. Jeśli przy literze „H” znajduje się symbol „-”, oznacza to, że podział jest zły.

⁴ Program komputerowy TAKSONOMIA jest dołączony i opisany w książce M. Kolendy *Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych. Metodologia + program komputerowy* (program komputerowy opracowany przez Karola Kolendę), książka obecnie jest w druku w Wydawnictwie Akademii Ekonomicznej we Wrocławiu.



Rys. 1. Ocena na dendrogramie podziału zbioru obiektów na dwie grupy za pomocą mierników H oraz h
 Źródło: opracowanie własne.

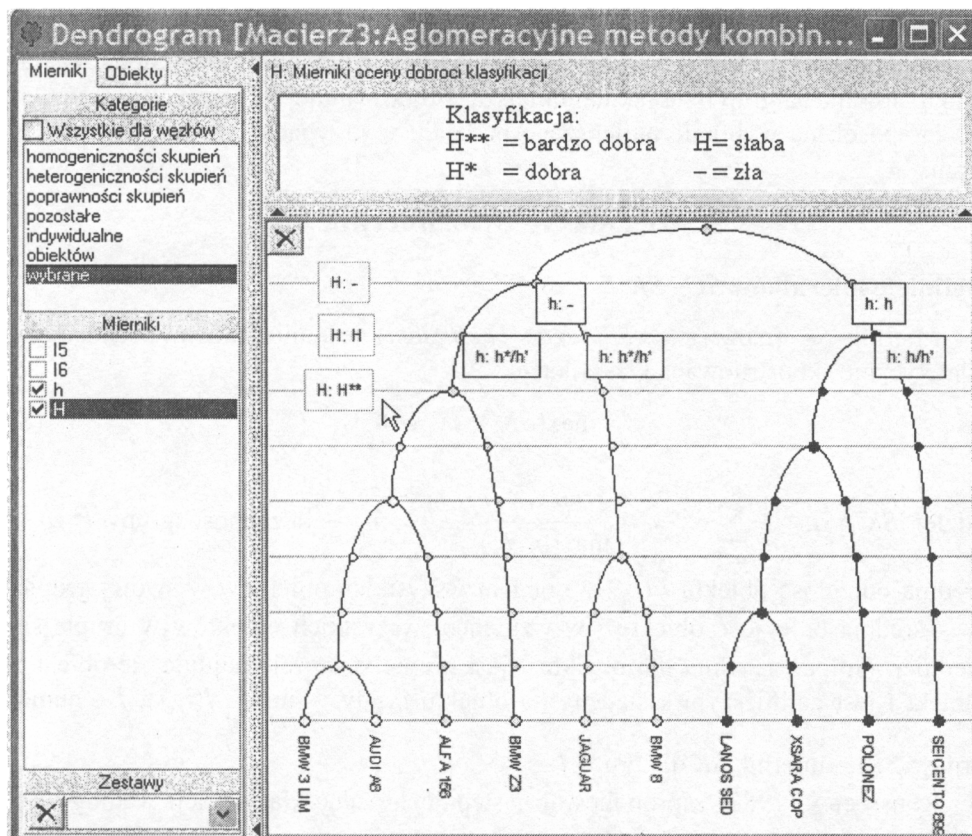
Podział na dwie grupy jest zły, ponieważ pierwsza grupa samochodów (licząc od lewej strony) nie ma heterogeniczności – przy mierniku „ h ” znajduje się „-”, co oznacza brak heterogeniczności (zob. strzałka wskaźnika myszy). Druga grupa samochodów (LANOS SED, XSARA, POLONEZ, SEICENTO) jest heterogeniczna. Wprowadzie heterogeniczność jest słaba (typu h), można jednak uznać, że samochody te należą do jednej grupy.

Na rys. 2 przedstawiono przykład wybierania najlepszego podziału na dendrogramie dla metod hierarchicznych za pomocą mierników H oraz h .

Opis rys. 2.

– W pierwszym rzędzie od góry dendrogramu widoczny jest symbol $H:-$, co oznacza, że klasyfikacja na dwie grupy jest zła (por. rys. 1).

– W drugim rzędzie od góry widnieje symbol $H: H$, co oznacza, że klasyfikacja na trzy grupy jest słaba, choć dopuszczalna. W tym podziale słaba jest grupa



Rys. 2. Wybieranie najlepszego podziału za pomocą mierników „ H ” na dendrogramie w programie komputerowym „Taksonomia numeryczna”

Źródło: opracowanie własne.

{LANOS, XSARA, POLONEZ, SEICENTO} – ma ona heterogeniczność typu h oraz h' (widać, że h' jest miernikiem niezależnym od h – należy porównać tę grupę w rzędzie drugim).

– W trzecim rzędzie od góry klasyfikacja jest typu H^{**} , co oznacza bardzo dobrą klasyfikację. Miernik H^{**} oznacza, że wszystkie grupy, w tym obiekt izolowany, mają heterogeniczność typu h^{**} . Zatem za najlepszą klasyfikacją dla tego zbioru należy uznać podział na cztery grupy, a dokładniej na trzy grupy i jeden obiekt izolowany. Obiektem izolowanym jest SEICENTO. Wyłonione grupy to:

- gr. 1 { BMW 3 LMN; AUDI A8; ALFA 156; BMW Z3},
- gr. 2 { JAGUAR; BMW 8},
- gr. 3 {LANOS SED; XSARA COP; POLONEZ},
- gr. 4 {SEICENTO 886}.

Jak widać na rys. 2, za pomocą proponowanych mierników dość szybko można ocenić daną klasyfikację oraz wszystkie wyłonione w niej grupy. Jeśli doda się warunek, aby liczba grup była jak najmniejsza⁵, to otrzymuje się prosty, zgodny z intuicją sposób na wybranie najlepszego podziału w przypadku metod hierarchicznych.

5. Miernik SC oraz miernik SK

Definicje mierników SC i SK

Miernik SC (nazywany *Silhouette coefficient*) jest miernikiem oceniającym syntetycznie skonstruowaną klasyfikację.

$$SC = \max_l (SK_l), SC \in (0, 1), \quad (8)$$

gdzie: $SK = \bar{s}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} s_i$; $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$; $i \in l$, n_l – liczebność grupy l , a_i –

średnia odległość obiektu i (o_i) względem wszystkich obiektów w swojej grupie, b_i – średnia odległość obiektu i względem wszystkich obiektów w grupie sąsiedniej. Grupa sąsiednia dla obiektu i jest grupą, w której znajduje się obiekt j . Obiekt j jest najbliższym sąsiadem dla obiektu i , gdy: $\min_{j, i \in l, j \notin l} d(i, j)$, l – numer

grupy; SK_l – miernik SK dla grupy l .

Rousseeuw (1987) zaproponował następującą ocenę klasyfikacji według wartości miernika SC :

0,71-1,00 – bardzo dobry podział na grupy,

0,51-0,70 – dobry podział,

0,26-0,50 – podział słaby (zastosować inną metodę klasyfikacji),

0,25-0,0 – punkty nie tworzą skupień.

Przykład wyznaczania i interpretacji mierników SC , SK , H oraz h

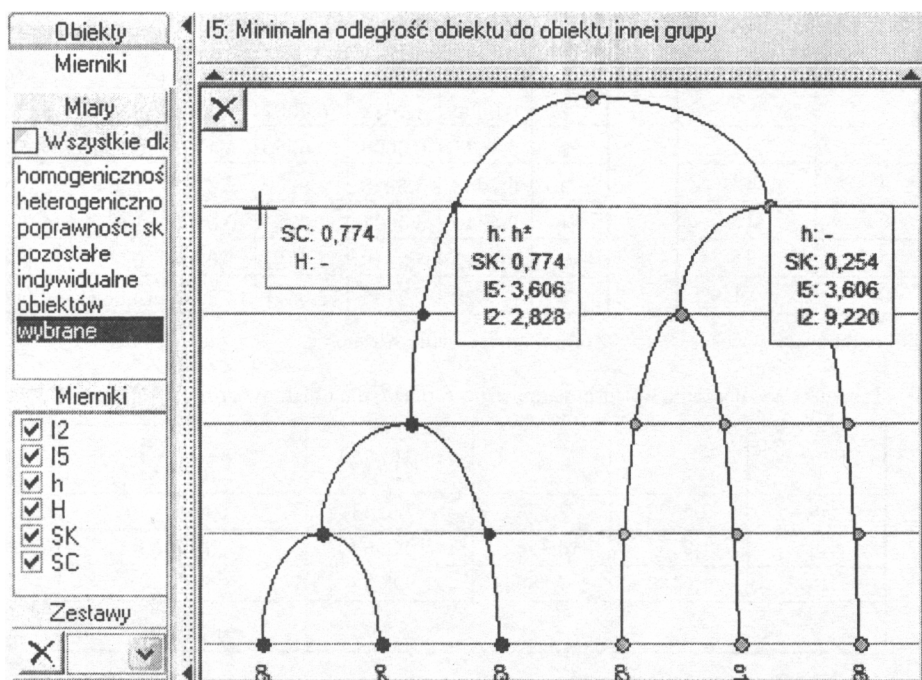
Na przykładowym zbiorze punktów z tab. 1, składającym się z sześciu obiektów, skonstruowano macierz odległości (euklidesową – zob. tab. 2). Następnie za pomocą metody Warda skonstruowano hierarchiczną strukturę podziału i przedstawiono ją na dendrogramie – rys. 3.

Na dendrogramie na rys. 3 w pierwszym rzędzie od góry przedstawiono:

➤ dwa mierniki oceniające cały podział:

– miernik $SC = 0,774$, który oznacza, że podział zbioru na dwie grupy jest bardzo dobry;

⁵ Nie bez znaczenia jest liczebność poszczególnych grup. Na przykład podział na dwie grupy, gdzie jedną grupę będzie stanowił obiekt izolowany, nie będzie dobrym podziałem niezależnie od typu heterogeniczności tych grup.



Rys. 3. Prezentacja dendrogramu oraz mierników

Źródło: opracowanie własne.

– miernik H :- , który oznacza, że podział całego zbioru obiektów na dwie grupy jest zły;

➤ pięć mierników oceniających poszczególne (dwie) grupy:

– mierniki h oceniające heterogeniczność poszczególnych grup: grupa $\{o_3, o_4, o_5\}$ ma heterogeniczność typu h^* – dobrą, natomiast grupa $\{o_2, o_1, o_6\}$ nie jest heterogeniczna;

– miernik SK wyznaczany dla każdej grupy przy obliczaniu miernika SC ;

– miernik $I5$, który przedstawia wartość Sep zdefiniowaną wzorem (3);

– miernik $I2$ przedstawiający średnicę taksonomiczną grupy, czyli wartość DC zdefiniowaną wzorem (2).

Przedstawiony podział całego zbioru obiektów na dwie grupy został przez oba mierniki, tzn. SC i H , oceniony zupełnie inaczej. Miernik $SC = 0,774$ określił tę klasyfikację jako bardzo dobrą ($SC > 0,71$), natomiast miernik H :- ocenił, że dokonana klasyfikacja jest zła.

Według mierników h grupa $\{o_2, o_1, o_6\}$ nie jest heterogeniczna. Grupa ta nie jest heterogeniczna ani wg h^* , ponieważ średnica grupy wynosi 9,220 i jest większa niż $Sep = 3,606$, ani wg h , ponieważ maksymalna odległość obiektu od reprezentanta w grupie wynosi 6,0828 (reprezentantem jest o_1).

Tabela 1. Macierz danych X

X_1	X_2
1	7
2	10
4	13
6	15
5	15
0	1

Źródło: opracowanie własne.

Tabela 2. Macierz odległości dla zbioru punktów w tab. 1

	o_1	o_2	o_3	o_4	o_5	o_6
o_1	0,0000	3,1623	6,7082	9,4340	8,9443	6,0828
o_2	3,1623	0,0000	3,6056	6,4031	5,8310	9,2195
o_3	6,7082	3,6056	0,0000	2,8284	2,2361	12,6491
o_4	9,4340	6,4031	2,8284	0,0000	1,0000	15,2315
o_5	8,9443	5,8310	2,2361	1,0000	0,0000	14,8661
o_6	6,0828	9,2195	12,6491	15,2315	14,8661	0,0000

Źródło: opracowanie własne.

Tabela 3. Wyznaczanie wartości parametrów a oraz b dla obiektów grupy $\{o_3, o_4, o_5\}$

	a_i	b_i	$\max(a_i, b_i)$	$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$
o_3	2,5322	7,6543	7,6543	0,669173
o_4	1,9142	10,3562	10,3562	0,815163
o_5	1,6180	9,8804	9,8804	0,836239
$\bar{s}_k =$				0,773525

Źródło: opracowanie własne.

Przykładowo: $a_i(o_3) = \frac{2,8284 + 2,2361}{2} = 2,5322$; $b_i(o_3) = \frac{6,7082 + 3,6065 + 12,6491}{3} = 7,6543$.

Tabela 4. Wyznaczanie wartości parametrów a oraz b dla obiektów grupy $\{o_1, o_2, o_6\}$

	a_i	b_i	$\max(a_i, b_i)$	$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$
o_1	4,6225	8,3622	8,3622	0,447209
o_2	6,1909	5,2799	6,1909	-0,14716
o_6	7,6512	14,2489	14,2489	0,463036
$\bar{s}_k =$				0,254363

Źródło: opracowanie własne.

Dla grupy $\{o_3, o_4, o_5\}$ wartość $\bar{s}_k = \frac{0,669173 + 0,815163 + 0,836239}{3} = 0,773525$.

Dla grupy $\{o_1, o_2, o_6\}$ wartość $\bar{s}_k = \frac{0,447201 - 0,14716 + 0,463036}{3} = 0,254363$.

Wartość miernika $SC = \max(0,773525; 0,254363) = 0,774$.

Wyznaczone wartości mierników warto porównać z wartościami na dendrogramie, przy czym miernik $\bar{s}_k$ oznaczono na dendrogramie symbolem SK – widać, że wszystkie wartości są zgodne.

6. Wariancja ogólna, wewnątrzgrupowa i międzygrupowa

Wariancję wewnątrzgrupową lub wariancję międzygrupową wyznacza się dla cechy X w przypadku, gdy cały zbiór obiektów jest podzielony na k grup (gdy wartości cechy są podzielone na k rozłącznych grup). Wówczas dla cechy X porównuje się średnią wariancję liczoną w poszczególnych grupach względem wariancji ogólnej cechy X – wariancji wyznaczonej dla wszystkich wartości cechy (dla całego zbioru obiektów). Porównywanie wariancji wewnątrzgrupowej z wariancją ogólną cechy X ma na celu sprawdzenie, w jakim stopniu zmniejszyła się wariancja cechy w wyniku podziału całego zbioru obiektów na k grup.

Definicje

Wariancja ogólna: Wariancję $S^2(X)$ wyznaczoną na podstawie wszystkich wartości cechy X nazywa się wariancją ogólną. Wariancję $S^2(X)$ dla cechy X wyznacza się według wzoru:

$$S^2(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (10)$$

gdzie x_i oznacza elementy cechy X .

Wariancja wewnątrzgrupowa $S_W^2(X)$ dla cechy X jest średnią ważoną wariancji wyznaczonych dla poszczególnych grup. Uwzględnia się przy tym liczebności grup. Wariancję wewnątrzgrupową $S_W^2(X)$ wyznacza się ze wzoru:

$$S_W^2(X) = \frac{1}{n} \sum_{l=1}^k S_l^2(X) \cdot n_l, \quad (11)$$

gdzie k oznacza liczbę grup, $S_l^2(X)$ – wariancję wartości x_i cechy X opisującej obiekty grupy l , czyli dla $x_i \in l$, ($i = 1, 2, \dots, n_l$), n_l – liczebność grupy l , $\sum_{l=1}^k n_l = n$.

Wariancja międzygrupowa $S_B^2(X)$ dla cechy X jest to wariancja średnich arytmetycznych poszczególnych grup, przy czym uwzględnia się liczebności grup. Wariancję międzygrupową $S_B^2(X)$ wyznacza się ze wzoru:

$$S_B^2(X) = \frac{1}{n} \sum_{l=1}^k (\bar{x}_l - \bar{\bar{x}})^2 \cdot n_l, \quad \bar{\bar{x}} = \frac{1}{n} \sum_{l=1}^k \bar{x}_l \cdot n_l, \quad (12)$$

gdzie $\bar{x}_l$ oznacza średnią arytmetyczną dla $x_i \in l$, dla obiektów należących do l -tej grupy.

Równość wariancyjna

$$S^2(X) = S_W^2(X) + S_B^2(X). \quad (13)$$

Średnia wariancja wewnątrzgrupowa $\bar{V}_W$

$$\bar{V}_W = \frac{1}{m} \sum_{j=1}^m S_W^2(X_j), \quad (14)$$

Oceny klasyfikacji dokonuje się za pomocą średniej wariancji wewnątrzgrupowej $\bar{V}_W$, jest ona konstruowana na podstawie wariancji wewnątrzgrupowych wyznaczonych dla wszystkich cech.

Przykład wyznaczania wariancji wewnątrzgrupowej oraz macierzy wariancji i kowariancji wewnątrzgrupowej

Na podstawie macierzy danych $\mathbf{X}$ zawierającej zbiór cech $\{X_j\}$, ($j = 1, 2, \dots, m$) oraz zbiór obiektów $\{o_i\}$, ($i = 1, 2, \dots, n$) i dla określonego podziału obiektów na k grup o liczebnościach n_l ($l = 1, 2, \dots, k$) można wyznaczyć trzy rodzaje wariancji dla każdej cechy X_j . Ze względu na równość wariancyjną (zob. wzór (13)) wystarczy wyznaczyć dla danej cechy X_j tylko dwie wariancje.

Wyznaczanie wariancji wewnątrzgrupowej dla cech zostanie pokazane na przykładowej macierzy danych $\mathbf{X}$, przedstawionej w tab. 1.

Macierz danych $\mathbf{X}$ w tab. 1 zawiera pięć obiektów $\{o_1, o_2, o_3, o_4, o_5\}$ oraz trzy cechy $\{X_1, X_2, X_3\}$. W macierzy $\mathbf{X}$ zbiór obiektów podzielono na dwie grupy $\{o_1, o_2, o_3\}$ i $\{o_4, o_5\}$. Poszczególne grupy obiektów wraz z wartościami cech osobno są przedstawione w tab. 2 i tab. 3.

Tabela 5. Macierz danych $\mathbf{X}$ – cały zbiór

	Obiekty	X_1	X_2	X_3
Grupa 1	o_1	1	9	10
	o_2	3	4	14
	o_3	2	5	18
Grupa 2	o_4	8	12	30
	o_5	11	10	25
	$\bar{x}$	5	8	19,4
	S^2	14,8	9,2	52,6

Tabela 6. Dane – grupa 1

Obiekty	X_1	X_2	X_3
o_1	1	9	10
o_2	3	4	14
o_3	2	5	18
$\bar{x}$	2	6	14
S^2	0,67	4,6667	10,667

Źródło: opracowanie własne.

Tabela 7. Dane – grupa 2

obiekty	X_1	X_2	X_3
o_4	8	12	30
o_5	11	10	25
$\bar{x}$	9,5	11	27,5
S^2	2,25	1	6,25

Źródło: opracowanie własne.

Źródło: opracowanie własne.

Dwa ostatnie wiersze w każdej tabeli prezentują wartości średnich arytmetycznych ($\bar{x}$) oraz wariancje S^2 wyznaczone dla każdej cechy (dla wartości elementów w kolumnie).

Wariancje wewnątrzgrupowe S_W^2 dla poszczególnych cech są następujące (podstawiono wartości S^2 z tab. 2 i tab. 3 do wzoru (11)):

$$S_W^2(X_1) = \frac{3 \cdot 0,6667 + 2 \cdot 2,25}{5} = 1,3; \quad S_W^2(X_2) = \frac{3 \cdot 4,6667 + 2 \cdot 1}{5} = 3,2;$$

$$S_W^2(X_3) = \frac{3 \cdot 10,6667 + 2 \cdot 6,25}{5} = 8,9.$$

Poszczególne rodzaje wariancji można łatwo otrzymać na podstawie poniższych trzech macierzy (zob. wartości wariancji S^2 na dole każdej macierzy).

Tabela 8. Macierz **X** do wyznaczania wariancji i kowariancji ogólnej

X =	1	9	10
	3	4	14
	2	5	18
	8	12	30
	11	10	25
S^2	14,8	9,2	52,6

Tabela 9. Macierz **W** do wyznaczania wariancji i kowariancji wewnątrzgrupowej

B =	2	6	14
	2	6	14
	2	6	14
	9,5	11	27,5
	9,5	11	27,5
S^2	13,5	6	43,74

Tabela 10. Macierz **W** do wyznaczania wariancji i kowariancji wewnątrzgrupowej

W =	-1	3	-4
	1	-2	0
	0	-1	4
	-1,5	1	2,5
	1,5	-1	-2,5
S^2	1,3	3,2	8,9

Macierz **B** jest macierzą składającą się ze środków ciężkości (średnich arytmetycznych grup). Każdy obiekt w grupie zastępowany jest wartościami środka ciężkości swojej grupy (zob. wartości $\bar{x}$ w tab. 6 i tab. 7). Macierz **W** jest macierzą danych **X** unormowaną (scentrowaną) wg środków ciężkości dla poszczególnych grup (odejmuje się średnie arytmetyczne grup).

$$\text{Średnia wariancja wewnątrzgrupowa } \bar{V}_W = \frac{1,3 + 3,2 + 8,9}{3} = 4,4666.$$

Na podstawie macierzy **X**, **B**, **W** można otrzymać trzy macierze wariancji i kowariancji:

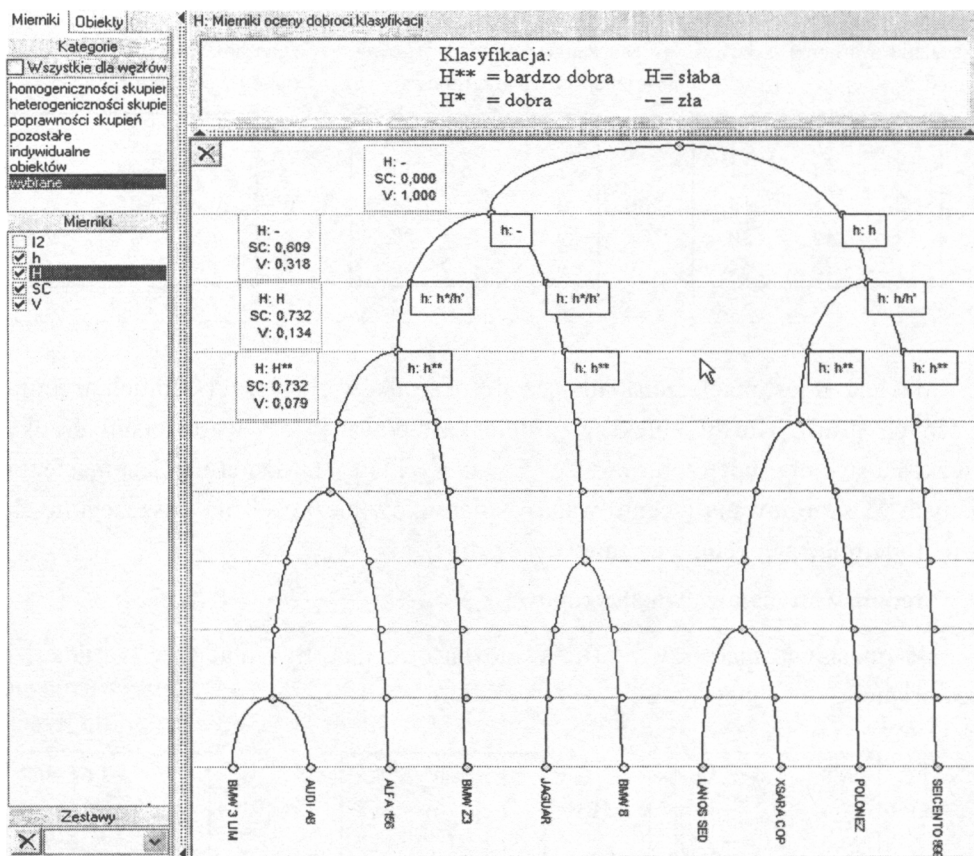
$$\text{Cov}(\mathbf{X}) = \begin{bmatrix} 14,8 & 7,4 & 23,6 \\ 7,4 & 9,2 & 14 \\ 23,6 & 14 & 52,64 \end{bmatrix} \quad \text{Cov}(\mathbf{B}) = \begin{bmatrix} 13,5 & 9 & 24,3 \\ 9 & 6 & 16,2 \\ 24,3 & 16,2 & 43,74 \end{bmatrix} \quad \text{Cov}(\mathbf{W}) = \begin{bmatrix} 1,3 & -1,6 & -0,7 \\ -1,6 & 3,2 & -2,2 \\ -0,7 & -2,2 & 8,9 \end{bmatrix}$$

dla których zachodzi następująca równość: $\text{Cov}(\mathbf{X}) = \text{Cov}(\mathbf{B}) + \text{Cov}(\mathbf{W})$.

7. Porównywanie mierników H , SC oraz $\bar{V}_W$ na przykładzie empirycznym

Zastosowanie mierników H , SC oraz $\bar{V}_W$ – średniej wariancji wewnątrzgrupowej do oceny skonstruowanej klasyfikacji – zaprezentowano i skomentowano na przykładzie empirycznym. Ponadto pokazano sposób wyboru „najlepszej” klasyfikacji (podziału) na dendrogramie prezentującym hierarchiczną strukturę podziału (grupowania) otrzymaną za pomocą metody *Warda* – jednej z hierarchicznych metod klasyfikacyjnych.

Przykład empiryczny dotyczy zbioru 10 samochodów, które scharakteryzowano za pomocą pięciu cech (tab. 13). Macierz odległości (euklidesową, tab. 14) wyznaczono na podstawie standaryzowanych cech, a następnie metodą Warda wyznaczono hierarchiczną strukturę podziału (rys. 4).



Rys. 4. Dendrogram zbioru samochodów oraz prezentacja mierników H , SC oraz V (gdzie: $V = \bar{V}_W$)
oceniających różne podziały zbioru

Dendrogram na rys. 4 przedstawia hierarchiczną strukturę skonstruowaną za pomocą metody Warda oraz mierniki oceniające trzy rodzaje klasyfikacji: na dwie, na trzy oraz na cztery grupy (pierwszy, drugi oraz trzeci od góry rząd dendrogramu).

Ocena klasyfikacji według miernika SC oraz mierników H

Mierniki SC oraz H zostaną porównane i ocenione na przykładzie trzech różnych klasyfikacji prezentowanych na trzech poziomach dendrogramu na rys. 4:

- na poziomie pierwszym od góry, gdzie cały zbiór podzielono na dwie grupy,
- drugim, gdzie zbiór samochodów podzielono na trzy grupy,
- trzecim, gdzie zbiór podzielono na cztery grupy.

Klasyfikacje na trzy oraz na cztery grupy wraz z wartościami mierników SC i H oraz ich interpretacją przedstawiono w tab. 13 oraz 14.

Tabela 11. Podział zbioru na 4 grupy i ocena podziału według mierników SC oraz H

Ocena klasyfikacji według mierników SC oraz H	Nr grupy	Heterogeniczność grupy	Nazwy obiektów
$SC = 0,732$ (bardzo dobra klasyfikacja) $H: H^{**}$ (bardzo dobra klasyfikacja)	1.	h^{**} bardzo dobra	BMW3 AUDI A8 ALFA 156 BMW Z3
	2.	h^{**} bardzo dobra	JAGUAR BMW 8
	3.	h^{**} bardzo dobra	LANOS SED XSARA CCP POLONEZ
	4.	h^{**} bardzo dobra	SEICENTO

Tabela 12. Podział zbioru na 3 grupy i ocena podziału według mierników SC oraz H

Ocena klasyfikacji według mierników SC oraz H	Nr grupy	Heterogeniczność grupy	Nazwy obiektów
$SC = 0,732$ (bardzo dobra klasyfikacja) $H: H$ (słaba chociaż dopuszczalna klasyfikacja)	1.	h^* dobra	BMW3 AUDI A8 ALFA 156 BMW Z3
	2.	h^* dobra	JAGUAR BMW 8
	3.	h oraz h' słaba	LANOS SED XSARA CCP POLONEZ SEICENTO

Źródło: opracowanie własne.

Źródło: opracowanie własne.

Jak widać na dendrogramie (rys. 4) oraz w tab. 11 i 12, oba mierniki różnią się w ocenie tych klasyfikacji. Oto szczegółowy opis mierników SC oraz H dla trzech rodzajów klasyfikacji.

1. Podział zbioru na dwie grupy (zob. dendrogram na rys. 4):

- miernik SC ocenił tę klasyfikację jako dobrą,
- według mierników H klasyfikacja jest zła (ponieważ pierwsza grupa na dendrogramie nie jest heterogeniczna). Wartości DC oraz Sep dla grupy 1 są następujące:

DC (średnica) = $d(\text{BMW3}, \text{BMW8}) = 3,126$ (d – odległość euklidesowa),

Sep (odrębność) = $d(\text{BMW3}, \text{LANOS}) = 1,877$,

co oznacza, że BMW3 dużo bardziej różni się od BMW8 (który jest w tej samej grupie) niż od LANOSA, który jest w innej grupie.

2. Podział zbioru na trzy grupy (zob. tab. 12):

– miernik SC ocenił podział jako bardzo dobry i niczym się nie różniący od podziału wcześniejszego,

– miernik typu H przypisany dla tej klasyfikacji oznacza, że jest to klasyfikacja słaba, chociaż dopuszczalna.

Według mierników H wejście SEICENTO do grupy 3 osłabiło heterogeniczność tej grupy, co wpłynęło niekorzystnie na całą klasyfikację (por. mierniki h dla poszczególnych grup w tab. 11 i 12). Wartości DC oraz Sep dla grupy 3 są następujące:

DC (średnica) = $d(\text{LANOS}, \text{SEICENTO}) = 2,281$,

Sep (odrębność) = $d(\text{LANOS}, \text{BMW8}) = 1,877$.

Z przeprowadzonej analizy widać, że miernik SC nie zareagował na dołączenie SEICENTO do grupy 3 (na poziomie trzecim od góry dendrogramu), chociaż powinien, ponieważ w grupie tej po dołączeniu tego obiektu znacznie wzrosła średnica grupy i zmniejszyła się heterogeniczność grupy.

3. Podział zbioru na cztery grupy oba wskaźniki oceniły tak samo jako bardzo dobrą klasyfikację – zob. tab. 11.

Porównanie i ocena mierników

1. Według mierników H : najlepszy podział zbioru obiektów to podział na cztery klasy (grupy). Do przyjęcia jest również podział na trzy grupy. Natomiast nie powinno się dzielić tego zbioru obiektów na dwie grupy.

2. Według miernika SC : ocena jest mniej wyrazista. Miernik ocenia podział na dwie grupy jako dobry, natomiast podziały na trzy grupy oraz na cztery grupy uznaje za bardzo dobre i w ogóle ich nie różnicuje.

3. Średnia wariancja wewnątrzgrupowa (na dendrogramie symbol V) nie umożliwia w zasadzie wyboru „najlepszego podziału”, największy bowiem spadek wartości średniej wariancji wewnątrzgrupowej będzie przy podziale całego zbioru na dwie grupy, a następnie na trzy. Przy większej liczbie grup średnia wariancja wewnątrzgrupowa będzie coraz słabiej zależna od liczby grup. Tę prawidłowość zaobserwowano na wielu dendrogramach skonstruowanych za pomocą różnych metod klasyfikacyjnych i dla różnych zbiorów obiektów. Metody klasyfikacyjne działają bowiem na algorytmach, które w różny sposób minimalizują wariancję. Otrzymany dendrogram jest już tak skonstruowany, że największe zmniejszenie wariancji wewnątrzgrupowej będzie przy podziale zbioru na dwie części, następnie na trzy i na cztery.

W przypadku podziałów na większą liczbę grup spadek wariancji nie zależy tak silnie od liczby grup wyłonionych w danym podziale. Może się zdarzyć, że wariancja wewnątrzgrupowa przy podziale np. na 7 grup będzie mniejsza niż przy podziale na 8 grup, ale to dopiero przy większej liczbie grup, natomiast podziałowi na dwie, trzy lub cztery grupy (części) towarzyszy bardzo silny spadek wartości średniej wariancji wewnątrzgrupowej.

Dla punktu izolowanego wariancja wewnątrzgrupowa będzie wynosiła zero. Zatem dla podziału zawierającego tyle grup, ile jest punktów, średnia wariancja wewnątrzgrupowa będzie wynosiła zero, natomiast dla całego zbioru (jednej grupy) średnia wariancja wewnątrzgrupowa wynosi jeden (dla cech unormowanych za pomocą standaryzacji).

Średnia wariancja wewnątrzgrupowa może być pomocna tylko w przypadku, gdy będą rozpatrywane różne podziały, ale o tej samej liczbie grup, i w przypadku, kiedy mierniki H nie różnicują jakości tych klasyfikacji (dadzą takie same wskazania o jakości tych klasyfikacji).

Tabela 13. Macierz danych

Obiekty	Moc maksymalna [KM]	Maksymalny moment obrotowy [Nm]	Przyspieszenie 0-100 km/h [s]	Średnie zużycie paliwa [l]	Masa własna [kg]
ALFA 156	190	222	7,3	11,4	1320
AUDI A8	193	280	8,4	9,9	1510
BMW 3 LIM	193	280	7	9,1	1445
BMW Z3	321	350	5,4	11,1	1350
BMW 8	326	490	6,3	14	1880
XSARA COP	90	135	13,6	7,9	1068
LANOS SED	106	145	11,5	8,2	1116
POLONEZ	76	121	18	8,9	1135
SEICENTO 899	39	65	18	5,9	690
JAGUAR	370	505	5,6	12,7	1825

Źródło: opracowanie własne.

8. Podsumowanie

1. Proponowane w niniejszym artykule mierniki H oraz h służą do oceny dobroci (jakości) skonstruowanej klasyfikacji oraz wszystkich wyłonionych w niej grup. Ocenę jakości klasyfikacji oparto na badaniu heterogeniczności (odrębności) wszystkich występujących w tym podziale grup, w tym punktów izolowanych.

2. Zdefiniowano trzy stopnie heterogeniczności grupy, przy czym dla punktu izolowanego zdefiniowano dwa stopnie heterogeniczności (odrębności). Ponadto zdefiniowano *brak heterogeniczności* dla grupy i punktu izolowanego.

3. W przypadku metod hierarchicznych zaproponowano prosty sposób wybierania podziału najlepszego. Autorka proponuje uznać za najlepszy taki podział, w którym jest najmniejsza liczba grup i w którym wszystkie grupy są heterogeniczne.

Tabela 14. Macierz odległości skonstruowana na podstawie standaryzowanych wartości z tab. 11.

	ALFA 156	AUDI A8	BMW 3 LIM	BMW Z3	BMW 8	XSARA COP	LANOS SED	POLONEZ	SEICENTO 899	JAGUAR
ALFA 156	0,0000	0,9774	1,1408	1,5470	2,9995	2,4272	1,9949	2,8874	4,1852	3,0271
AUDI A8	0,9774	0,0000	0,4983	1,5809	2,8525	2,3621	1,9572	2,8300	4,1566	2,7770
BMW 3 LIM	1,1408	0,4983	0,0000	1,5943	3,1257	2,3235	1,8769	2,9625	4,0787	2,9664
BMW Z3	1,5470	1,5809	1,5943	0,0000	2,2354	3,5150	3,1008	4,0159	5,1656	1,9453
BMW 8	2,9995	2,8525	3,1257	2,2354	0,0000	5,0825	4,7276	5,2654	6,8141	0,7344
XSARA COP	2,4272	2,3621	2,3235	3,5150	5,0825	0,0000	0,5167	1,0720	1,8270	5,0296
LANOS SED	1,9949	1,9572	1,8769	3,1008	4,7276	0,5167	0,0000	1,4671	2,2809	4,6628
POLONEZ	2,8874	2,8300	2,9625	4,0159	5,2654	1,0720	1,4671	0,0000	1,9180	5,3059
SEICENTO 899	4,1852	4,1566	4,0787	5,1656	6,8141	1,8270	2,2809	1,9180	0,0000	6,7313
JAGUAR	3,0271	2,7770	2,9664	1,9453	0,7344	5,0296	4,6628	5,3059	6,7313	0,0000

Źródło: opracowanie własne.

4. Przedstawiono i zaprezentowano na przykładach miernik SC oraz porównano jego skuteczność i wrażliwość z proponowanymi miernikami H – porównanie wyszło na korzyść mierników H .

5. Przedstawiono wariacje wewnątrzgrupową oraz wskazano na jej małą przydatność w wyborze „najlepszego” podziału w przypadku hierarchicznych metod klasyfikacji, zwłaszcza gdy liczba grup jest mała. Natomiast zaleca się ją stosować wówczas, gdy należy wybrać jedną z dwóch otrzymanych różnych klasyfikacji, ale o tej samej liczbie grup, oraz w przypadku, gdy mierniki H nie będą w stanie zróżnicować jakości tych podziałów.

Literatura

- [1] Grabiński T., Wydymus S., Zeliaś A., *Metody taksonometryczne w modelowaniu zjawisk społeczno-gospodarczych*, PWN, Warszawa 1989.
- [2] Kaufman L., Rousseeuw P.J., *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York 1990.
- [3] IDAMS – *Internationally Developed Data Analysis and Management*, UNESCO 2001-2002.
- [4] Rousseeuw P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, „Journal of Computational and Applied Mathematics” 1987, 20.

COEFFICIENTS OF CLASSIFICATION SCORING

Summary

The paper describes techniques of classification scoring. The basis of a score is determined by a heterogeneity of every cluster and every isolated object in a given classification. A few coefficients have been proposed, which determine various degrees of heterogeneity of clusters and isolated objects. On the basis of these coefficients, four types of classifications have been defined: very good, good, weak but sufficient, and bad classification.

The paper also describes how to easily find „the best” classification for hierarchical methods. The proposed coefficients have been commented and compared with different coefficients, such as SC (Silhouette Coefficient) and average innercluster variation.