PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Marek Walesiak, Andrzej Dudek

Uniwersytet Ekonomiczny we Wrocławiu

OCENA WYBRANYCH PROCEDUR ANALIZY SKUPIEŃ DLA DANYCH PORZĄDKOWYCH

1. Wstęp

Celem artykułu jest przeprowadzenie oceny przydatności wybranych procedur analizy skupień dla danych porządkowych. Testowanie przydatności wybranych procedur przeprowadzone zostanie na podstawie porządkowych danych symulacyjnych o znanej strukturze klas obiektów wygenerowanych z wykorzystaniem z funkcji `cluster.Gen` pakietu `clusterSim` (zob. [Walesiak, Dudek 2008]).

W teorii pomiaru rozróżnia się cztery podstawowe skale pomiaru, tj. nominalną, porządkową (rangową), przedziałową (interwałową), ilorazową (stosunkową). Skale przedziałową i ilorazową zalicza się do skal metrycznych, natomiast nominalną i porządkową do niemetrycznych. Skale pomiaru są uporządkowane od najsłabszej (nominalna) do najmocniejszej (ilorazowa). Z typem skali wiąże się grupa przekształceń, ze względu na które skala zachowuje swe własności. Na skali porządkowej dozwolonym przekształceniem matematycznym dla obserwacji jest dowolna ściśle monotonicznie rosnąca funkcja, która nie zmienia dopuszczalnych relacji, tj. równości, różności, większości i mniejszości.

Zasób informacji skali porządkowej jest nieporównanie mniejszy niż skal metrycznych. Jedyną dopuszczalną operacją empiryczną na skali porządkowej jest zliczanie zdarzeń (tzn. wyznaczanie liczby relacji większości, mniejszości i równości). Szczegółową charakterystykę skal pomiaru zawierają m.in. prace: [Walesiak 1996, s. 19-24; 2006, s. 12-15].

2. Procedura analizy skupień dla danych porządkowych

Typowa procedura analizy skupień dla danych porządkowych obejmuje (zob. np. [Milligan 1996, s. 342-343; Walesiak 2005]):

1. Wybór obiektów i zmiennych.
2. Wybór miary odległości.
3. Wybór metody klasyfikacji.
4. Ustalenie liczby klas.
5. Ocenę wyników klasyfikacji.
6. Opis i profilowanie klas.

W stosunku do procedury analizy skupień dla danych metrycznych nie występuje tutaj etap normalizacji wartości zmiennych. Normalizacja polega na pozbawieniu wartości zmiennych mian i ujednoliceniu rzędów wielkości w celu doprowadzenia ich do porównywalności. Dla danych porządkowych nie zachodzi potrzeba normalizacji ze względu na to, że są one niemianowane. Ponadto rząd wielkości obserwacji zmiennej porządkowej nie ma znaczenia, gdyż między obserwacjami wyznacza się tylko relacje równości, różności, większości i mniejszości. Miara odległości (zob. etap 2 procedury analizy skupień) dla obiektów opisanych zmiennymi porządkowymi może wykorzystywać w swojej konstrukcji tylko ww. relacje. To ograniczenie powoduje, że musi być ona miarą kontekstową, która wykorzystuje informacje o relacjach, w jakich pozostają porównywane obiekty w stosunku do pozostałych obiektów z badanego zbioru obiektów. Taką miarą odległości dla danych porządkowych jest miara GDM zaproponowana przez Walesiaka [1993, s. 44-45]. W literaturze przedmiotu nie są znane inne miary odległości dla danych porządkowych, które wykorzystywałyby dopuszczalne na tej skali relacje.

Wybrane procedury analizy skupień obejmują w artykule:

1. Miarę odległości GDM dla danych porządkowych (zob. [Walesiak 2006, s. 36-39]). W pakiecie `clusterSim` jest to metoda `GDM2`.

2. Wybrane metody klasyfikacji (zob. [Walesiak 2008a]):

– metodę *k*-medoidów (`pam`), w której każda klasa jest reprezentowana przez jeden z jej obiektów będący gwiazdą klasy (*medoid, star*);
– siedem metod klasyfikacji hierarchicznej: pojedynczego połączenia (`single`), kompletnego połączenia (`complete`), średniej klasowej (`average`), ważonej średniej klasowej (`mcquitty`), powiększonej sumy kwadratów odległości (`ward`), środka ciężkości (`centroid`), medianową (`median`). Metody Warda, centroidalna i medianowa przyjmują założenie, że odległości między obiektami zostały wyznaczone za pomocą kwadratu odległości euklidesowej (mają one wtedy interpretację geometryczną, zgodną z nazwami tych metod). Metody te mogą być stosowane (por. [Anderberg 1973, s. 141]), gdy macierz odległości jest liczona na podstawie innych miar odległości, lecz interpretacja tak otrzymanych wyników (w sensie odległości międzyklasowej) nie jest zgodna z nazwami tych metod,
– hierarchiczną metodę deglomeracyjną Macnaughtona-Smitha i in. [1964] – w pakiecie **R** nosi ona nazwę `diana`.

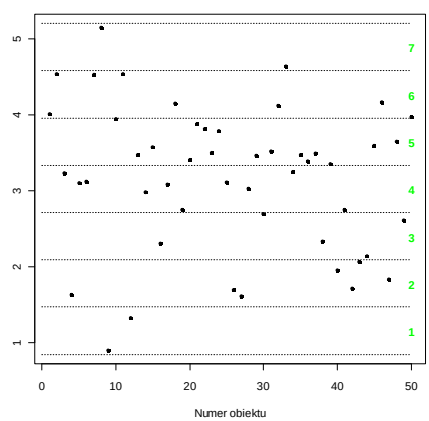
3. Wybrane indeksy służące ustaleniu liczby klas (Daviesa-Bouldina – `index.DB`, Calińskiego i Harabasza – `index.G1`, Bakera i Huberta – `index.G2`, Huberta i Levine – `index.G3`, gap – `index.Gap`, Hartigana – `index.H`, Krzanowskiego i Lai – `index.KL`, Silhouette – `index.S`). Formuły prezentowanych indeksów zawiera praca Walesiaka [2008a].

Indeksy Calińskiego i Harabasza, Krzanowskiego i Lai, Daviesa-Bouldina, Hartigana i gap w swojej konstrukcji wykorzystują środek ciężkości klasy o współrzędnych będących średnimi arytmetycznymi z wartości zmiennych opisujących obiekty danej klasy. Dla danych porządkowych nie jest dopuszczalne obliczanie średnich arytmetycznych. W związku z tym przy obliczaniu tych indeksów zamiast środka ciężkości klasy zastosowano współrzędne obiektu usytuowanego centralnie w klasie (tj. obiektu, dla którego suma odległości od pozostałych obiektów w klasie jest najmniejsza).

Testowanie przydatności wybranych procedur przeprowadzono na podstawie porządkowych danych symulacyjnych o znanej strukturze klas obiektów.

3. Procedura generowania danych porządkowych z wykorzystaniem pakietu `clusterSim`

Generowanie obserwacji porządkowych w pakiecie `clusterSim` (funkcja `clusterGen`) przebiega w sposób następujący (zob. [Walesiak 2008b]). Najpierw generowane są losowo dane metryczne o znanej strukturze klas z wielowymiarowego rozkładu normalnego, w którym położenie i jednorodność skupień zadaje się za pomocą wektorów wartości przeciętnych (środki ciężkości skupień) i macierzy kowariancji (rozproszenie obiektów). Obserwacje na zmiennych zakłócających strukturę klas (*noisy variables*) generowane są niezależnie z rozkładu jednostajnego. Przedział zmienności zmiennych zakłócających jest podobny do zmiennych wyznaczających strukturę klas. Wygenerowane obserwacje mają charakter ciągły (dane metryczne).



Rys. 1. Przykład dyskretyzacji wartości *j*-tej zmiennej
Źródło: opracowanie własne.

Funkcja `cluster.Gen` zawiera wbudowane modele, z zadanymi wektorami wartości przeciętnych i macierzami kowariancji, różniące się:
- liczbą zmiennych (wymiarów) i liczbą skupień,
- gęstością skupień (tj. liczebnością obiektów w klasach),
- kształtem skupień w wyniku uwzględnienia zróżnicowanych macierzy kowariancji dla poszczególnych skupień.

W celu otrzymania danych porządkowych należy przeprowadzić dla każdej zmiennej proces dyskretyzacji. Liczba kategorii (k_j) zmiennej porządkowej X_j określa szerokość przedziału klasowego $\left[\max_i\{x_{ij}\} - \min_i\{x_{ij}\}\right] \Big/ k_j$. Niezależnie dla każdej zmiennej kolejne przedziały klasowe otrzymują kategorie $1,\ldots,k_j$ i aktualna wartość zmiennej x_{ij} jest zastępowana przez te kategorie. Dla poszczególnych zmiennych liczba kategorii może być inna (np. $k_1 = 7$, $k_2 = 4$, $k_3 = 5$). Przykład dyskretyzacji wartości *j*-tej zmiennej zawiera rys. 1.

4. Charakterystyka eksperymentu symulacyjnego

Dane symulacyjne, o znanej strukturze klas obiektów, składają się z 9 modeli różniących się liczbą zmiennych, liczbą, gęstością i kształtem skupień oraz liczbą zmiennych zakłócających (zob. tab. 1). Następnie dla danych z poszczególnych modeli zastosowano 72 procedury analizy skupień obejmujące 9 metod klasyfikacji, miarę odległości GDM dla danych porządkowych i 8 indeksów jakości klasyfikacji służących ustaleniu liczby klas. Dla każdego modelu przeprowadzono 50 symulacji.

Tabela 1. Charakterystyka modeli w analizie symulacyjnej

Model	*v*	*lk*	*cl*	*lo*	Środki ciężkości klas	Macierz kowariancji Σ	*ks*
1	2	4, 6	3	60, 30, 30	(0; 0), (1,5; 7), (3; 14)	$\sigma_{jj} = 1$, $\sigma_{jl} = -0{,}9$	1
2	3	7	3	45	(1,5; 6, – 3), (3; 12; –6) (4,5; 18; –9)	$\sigma_{jj} = 1$ $(1 \le j \le 3)$, $\sigma_{12} = \sigma_{13} = -0{,}9$, $\sigma_{23} = 0{,}9$	1
3	2	5, 7	5	50, 20, 25, 25, 20	(5; 5), (–3; 3), (3; –3), (0; 0), (–5; –5)	$\sigma_{jj} = 1$, $\sigma_{jl} = 0{,}9$	2
4	3	5, 7, 5	5	25	(5; 5; 5), (–3; 3; –3), (3; –3; 3), (0; 0; 0), (–5; –5; –5)	$\sigma_{jj} = 1$ $(1 \le j \le 3)$, $\sigma_{jl} = 0{,}9$ $(1 \le j \ne l \le 3)$	2
5	2	5	5	20, 45, 15, 25, 35	(0; 0), (0; 10), (5; 5), (10; 0), (10; 10)	$\sigma_{jj} = 1$, $\sigma_{jl} = 0$	3
6	2	3, 5	4	35	(–4; 5), (5; 14), (14; 5), (5; –4)	$\sigma_{jj} = 1$, $\sigma_{jl} = 0$	3
7	3	6	4	25, 25, 40, 30	(–4; 5; –4), (5; 14; 5), (14; 5; 14), (5; –4; 5),	a	4
8	3	7	5	35, 25, 25, 20, 20	(5; 5; 5), (–3; 3; –3), (3; –3; 3), (0; 0; 0), (–5; –5; –5)	b	4
9	2	7	3	40	(0; 4), (4; 8), (8; 12)	c	4

Objaśnienia do tab. 1:
v – liczba zmiennych, *lk* – liczba kategorii (jedna liczba oznacza stałą liczbę kategorii); *cl* – liczba klas; lo – liczba obiektów w klasach (jedna liczba oznacza klasy równoliczne); *ks* – kształt skupień (1 – skupienia wydłużone, 2 – skupienia wydłużone i słabo separowalne, 3 – skupienia normalne, 4 – skupienia zróżnicowane dla klas);

a: $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1 & -0{,}9 & -0{,}9 \\ -0{,}9 & 1 & 0{,}9 \\ -0{,}9 & 0{,}9 & 1 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 1 & 0{,}9 & 0{,}9 \\ 0{,}9 & 1 & 0{,}9 \\ 0{,}9 & 0{,}9 & 1 \end{bmatrix}$, $\Sigma_4 = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$;

b: $\Sigma_1 = \begin{bmatrix} 1 & -0{,}9 & -0{,}9 \\ -0{,}9 & 1 & 0{,}9 \\ -0{,}9 & 0{,}9 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0{,}5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 1 & 0{,}9 & 0{,}9 \\ 0{,}9 & 1 & 0{,}9 \\ 0{,}9 & 0{,}9 & 1 \end{bmatrix}$, $\Sigma_4 = \begin{bmatrix} 1 & 0{,}6 & 0{,}6 \\ 0{,}6 & 1 & 0{,}6 \\ 0{,}6 & 0{,}6 & 1 \end{bmatrix}$, $\Sigma_5 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$;

c: $\Sigma_1 = \begin{bmatrix} 1 & -0{,}9 \\ -0{,}9 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1{,}5 & 0 \\ 0 & 1{,}5 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{bmatrix}$.

Źródło: opracowanie własne z wykorzystaniem pakietu `clusterSim` (zob. [Walesiak, Dudek 2008]).

Nie rozpatrywano wszystkich możliwych podziałów zbioru obiektów. W badaniu uwzględniono podziały zbioru obiektów od dwóch do dziesięciu klas.

Ocenę przydatności wybranych procedur analizy skupień dla danych porządkowych przeprowadzono za pomocą skorygowanego indeksu Randa (zob. [Hubert i Arabie 1985]), porównując znaną strukturę klas z wynikami uzyskanymi za pomocą odpowiednich procedur analizy skupień.

5. Dyskusja rezultatów analizy symulacyjnej

Tabela 2 prezentuje uporządkowanie 9 analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 50 symulacji dla 9 modeli i 8 indeksów oceny jakości klasyfikacji.

Tabela 2. Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa

Metoda	Liczba zmiennych zakłócających						Średnia	
	0		2		4			
Average	0,765	1	0,502	1	0,062	1	0,443	1
Mcquitty	0,733	4	0,456	3	0,057	3	0,415	2
Centroid	0,746	2	0,440	4	0,055	4	0,413	3
Ward	0,707	7	0,473	2	0,059	2	0,413	4
Diana	0,738	3	0,430	5	0,053	5	0,407	5
Complete	0,724	5	0,415	7	0,051	7	0,397	6
Pam	0,694	8	0,416	6	0,052	6	0,387	7
Median	0,708	6	0,371	8	0,046	8	0,375	8
Single	0,652	9	0,322	9	0,040	9	0,338	9

Źródło: obliczenia własne.

Na podstawie wyników zawartych w tab. 2 można sformułować następujące wnioski:
- zdecydowanie najlepszą metodą klasyfikacji danych porządkowych (dla 0, 2 i 4 zmiennych zakłócających) jest metoda średniej klasowej (`average`), najgorszą zaś metoda pojedynczego połączenia (`single`),
- metoda Warda (`ward`) w relacji do innych metod jest dość skuteczna w przypadku uwzględnienia zmiennych zakłócających.

Tabela 3 prezentuje uporządkowanie 8 analizowanych indeksów oceny jakości klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 50 symulacji dla 9 modeli i 9 metod klasyfikacji.

Tabela 3. Uporządkowanie analizowanych indeksów oceny jakości klasyfikacji według średnich wartości skorygowanego indeksu Randa

Indeks	Liczba zmiennych zakłócających						Średnia	
	0		2		4			
KL	0,804	1	0,473	1	0,052	1	0,443	1
G1	0,721	3	0,463	2	0,051	2	0,412	2
Gap	0,771	2	0,384	7	0,042	7	0,399	3
S	0,691	6	0,451	4	0,050	3	0,397	4
G3	0,667	8	0,453	3	0,050	3	0,390	5
G2	0,686	7	0,417	5	0,046	6	0,383	6
H	0,695	5	0,398	6	0,044	5	0,379	7
DB	0,713	4	0,361	8	0,040	8	0,371	8

Źródło: obliczenia własne.

Na podstawie wyników zawartych w tab. 3 można sformułować następujące wnioski:
- najlepsze indeksy w klasyfikacji danych porządkowych to indeksy Krzanowskiego i Lai (`index.KL`) oraz Calińskiego i Harabasza (`index.G1`),
- o ile indeksy gap (`index.Gap`) i Daviesa-Bouldina (`index.DB`) bez zmiennych zakłócających dość dobrze odkrywały strukturę klas, o tyle ze zmiennymi zakłócającymi ich skuteczność wyraźnie spadała.

Tabela 4. Uporządkowanie analizowanych procedur analizy skupień według średnich wartości skorygowanego indeksu Randa

Lp.	Metoda	Indeks	Liczba zmiennych zakłócających			Średnia	Lp.	Metoda	Indeks	Liczba zmiennych zakłócających			Średnia
			0	2	4					0	2	4	
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Average	KL	0,854	0,554	0,429	0,612	37	Pam	S	0,641	0,455	0,335	0,477
2	Ward	KL	0,843	0,537	0,396	0,592	38	Complete	Gap	0,762	0,385	0,283	0,477
3	Ward	Gap	0,854	0,505	0,362	0,574	39	Centroid	KL	0,830	0,505	0,076	0,470
4	Average	Gap	0,883	0,496	0,342	0,574	40	Mcquitty	G2	0,688	0,405	0,312	0,468
5	Average	H	0,764	0,536	0,417	0,572	41	Complete	DB	0,718	0,383	0,296	0,465
6	Average	G1	0,767	0,537	0,383	0,562	42	Median	G2	0,714	0,461	0,219	0,465
7	Mcquitty	KL	0,802	0,493	0,371	0,555	43	Median	KL	0,782	0,421	0,183	0,462
8	Pam	KL	0,837	0,469	0,340	0,549	44	Pam	DB	0,692	0,387	0,300	0,460

1	2	3	4	5	6	7	8	9	10	11	12	13	14
9	Average	S	0,715	0,517	0,391	0,541	45	Pam	H	0,631	0,402	0,344	0,459
10	Diana	KL	0,805	0,456	0,360	0,540	46	Diana	G2	0,719	0,373	0,285	0,459
11	Mcquitty	H	0,739	0,481	0,363	0,528	47	Centroid	G1	0,757	0,491	0,116	0,454
12	Ward	G1	0,687	0,518	0,378	0,528	48	Pam	G3	0,624	0,420	0,315	0,453
13	Diana	H	0,743	0,447	0,391	0,527	49	Pam	G1	0,637	0,424	0,296	0,452
14	Average	DB	0,771	0,457	0,352	0,527	50	Median	G1	0,725	0,430	0,202	0,452
15	Diana	G1	0,759	0,447	0,374	0,527	51	Median	G3	0,676	0,439	0,224	0,447
16	Mcquitty	G1	0,738	0,487	0,343	0,522	52	Centroid	G2	0,690	0,532	0,114	0,445
17	Average	G3	0,684	0,493	0,389	0,522	53	Ward	G2	0,646	0,386	0,294	0,442
18	Diana	S	0,735	0,462	0,357	0,518	54	Complete	G2	0,692	0,366	0,268	0,442
19	Complete	KL	0,785	0,438	0,325	0,516	55	Centroid	G3	0,675	0,523	0,121	0,439
20	Mcquitty	S	0,696	0,492	0,355	0,514	56	Pam	G2	0,654	0,362	0,270	0,429
21	Pam	Gap	0,834	0,406	0,297	0,513	57	Centroid	S	0,710	0,473	0,007	0,397
22	Ward	S	0,653	0,503	0,375	0,510	58	Median	S	0,697	0,410	0,082	0,396
23	Diana	G3	0,715	0,443	0,370	0,509	59	Single	G2	0,684	0,437	0,052	0,391
24	Mcquitty	Gap	0,788	0,426	0,311	0,508	60	Centroid	Gap	0,819	0,351	0,002	0,391
25	Ward	DB	0,729	0,428	0,343	0,500	61	Single	G1	0,697	0,394	0,061	0,384
26	Diana	Gap	0,709	0,419	0,360	0,496	62	Single	G3	0,631	0,431	0,068	0,376
27	Ward	H	0,619	0,458	0,409	0,495	63	Single	KL	0,697	0,382	0,043	0,374
28	Mcquitty	G3	0,685	0,445	0,344	0,491	64	Centroid	H	0,754	0,345	0,002	0,367
29	Complete	G1	0,726	0,440	0,307	0,491	65	Median	H	0,702	0,288	0,053	0,348
30	Mcquitty	DB	0,730	0,416	0,320	0,489	66	Centroid	DB	0,732	0,296	0,005	0,344
31	Complete	S	0,703	0,451	0,311	0,488	67	Median	Gap	0,693	0,275	0,064	0,344
32	Complete	H	0,716	0,424	0,316	0,485	68	Single	S	0,673	0,301	0,008	0,327
33	Average	G2	0,685	0,429	0,341	0,485	69	Median	DB	0,679	0,246	0,054	0,326
34	Diana	DB	0,718	0,397	0,332	0,483	70	Single	DB	0,647	0,239	0,009	0,298
35	Ward	G3	0,628	0,450	0,357	0,478	71	Single	Gap	0,601	0,190	0,008	0,266
36	Complete	G3	0,687	0,433	0,312	0,477	72	Single	H	0,583	0,202	0,006	0,264

Źródło: obliczenia własne.

Tabela 4 prezentuje uporządkowanie procedur analizy skupień (miara GDM dla danych porządkowych, 9 metod klasyfikacji, 8 indeksów jakości klasyfikacji) według średnich wartości skorygowanego indeksu Randa policzonego z 50 symulacji dla 9 modeli.

Na podstawie wyników zawartych w tab. 4 można sformułować następujące wnioski:

- najskuteczniejsza w sensie przeprowadzonego eksperymentu symulacyjnego jest procedura analizy skupień obejmująca metodę średniej klasowej (`average`) oraz indeks oceny jakości klasyfikacji Krzanowskiego i Lai (`index.KL`). Metoda ta z indeksami odpowiednio gap (`index.Gap`), Hartigana (`index.H`) oraz Calińskiego i Harabasza (`index.G1`) zajęła wysokie pozycje, tj. czwartą, piątą i szóstą,
- drugą i trzecią pozycję zajęła metoda Warda (`ward`) z indeksami odpowiednio Krzanowskiego i Lai (`index.KL`) oraz gap (`index.Gap`),

– najmniej skuteczna w klasyfikacji danych porządkowych jest metoda pojedynczego połączenia (`single`) z indeksami Hartigana (`index.H`), gap (`index.Gap`) i Daviesa-Bouldina (`index.DB`).

6. Wnioski końcowe

Autorzy zdają sobie sprawę, że na otrzymane rezultaty w pewnym stopniu ma wpływ wybór modeli i sposób generowania danych o znanej strukturze klas. W analizie uwzględniono losowe generowanie zbiorów danych z wielowymiarowego rozkładu normalnego, w którym położenie i jednorodność skupień zadaje się za pomocą wektorów wartości przeciętnych (środki ciężkości skupień) i macierzy kowariancji (rozproszenie obiektów). Takie podejście jest typowe w wielu analizach symulacyjnych prezentowanych m.in. w pracach: [Tibshirani, Walther, Hastie 2001; Dudoit, Fridlyand 2002; Soffritti 2003; Tibshirani, Walther 2005].

Podstawowym problemem związanym z generowaniem danych o znanej strukturze klas jest to, że istnieje nieskończenie wiele kształtów skupień dla dowolnej liczby wymiarów (zob. [Carmone, Kara i Maxwell 1999, s. 508]). Celowe byłoby uwzględnienie innych rozkładów oraz tzw. funkcji połączenia (*copula*) do generowania zbiorów danych o niestandardowych kształtach skupień. Nie jest to zadanie łatwe, szczególnie w przypadku danych porządkowych.

Literatura

Anderberg M.R. (1973), *Cluster analysis for applications*, Academic Press, New York, San Francisco, London.

Carmone F.J., Kara A., Maxwell S. (1999), *HINoV: a new method to improve market segment definition by identifying noisy variables*, „Journal of Marketing Research", November, vol. 36, s. 501-509.

Dudoit S., Fridlyand J. (2002), *A prediction-based resampling method for estimating the number of clusters in a dataset*, „Genome Biology", 3(7).

Hubert L.J., Arabie P. (1985), *Comparing partitions*, „Journal of Classification" no 1, s. 193-218.

Macnaughton-Smith P., Williams W.T., Dale M.B., Mockett L.G. (1964), *Dissimilarity analysis: a new technique of hierarchical sub-division*, „Nature", 202, s. 1034-1035.

Milligan G.W. (1996), *Clustering validation: results and implications for applied analyses*, [w:] *Clustering and classification*, red. P. Arabie, L.J. Hubert, G. de Soete, World Scientific, Singapore, s. 341-375.

R Development Core Team (2008), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org.

Soffritti G. (2003), *Identifying multiple cluster structures in a data matrix*, „Communications in Statistics. Simulation and Computation", vol. 32, no 4, s. 1151–1177.

Tibshirani R., Walther G., Hastie T. (2001), *Estimating the number of clusters in a data set via the gap statistic*, „Journal of the Royal Statistical Society", ser. B, vol. 63, part 2, s. 411-423.

Tibshirani R., Walther G. (2005), *Cluster validation by predicting strength*, „Journal of Computational and Graphical Statistics", vol. 14, no 3.

Walesiak M. (1993), *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654, Seria: Monografie i Opracowania nr 101, AE, Wrocław.

Walesiak M. (1996), *Metody analizy danych marketingowych*, PWN, Warszawa.
Walesiak M. (2005), *Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów*, [w:] *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, red. A. Zeliaś, AE, Kraków, s. 185-203.
Walesiak M. (2006), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydanie drugie rozszerzone, AE, Wrocław.
Walesiak M. (2008a), *Analiza skupień*, [w:] *Statystyczna analiza danych z wykorzystaniem programu* `R`, red. M. Walesiak, E. Gatnar, Wydawnictwo Naukowe PWN, Warszawa (w druku).
Walesiak M. (2008b), *Losowe generowanie danych o znanej strukturze klas w pakiecie* `clusterSim`, [w:] *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk społeczno-ekonomicznych*, red. J. Pociecha Wydawnictwo UE, Kraków (w redakcji).
Walesiak M., Dudek A. (2008), `clusterSim` *package*, URL http://www.R-project.org.

FINDING GROUPS IN ORDINAL DATA – AN EXAMINATION OF SOME CLUSTERING PROCEDURES

Summary

The major steps in a cluster analysis procedure for ordinal data contain: selection of objects and variables, selection of a distance measure, selection of clustering method, determining the number of clusters, cluster validation, describing and profiling clusters.

In the article, based on data simulated with `cluster.Gen` function of `clusterSim` package working in **R** environment, some cluster analysis procedures containing GDM distance for ordinal data, nine clustering methods and eight internal cluster quality indices are evaluated.