PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Justyna Wilk
Akademia Ekonomiczna we Wrocławiu

HIERARCHICZNE METODY KLASYFIKACJI W ANALIZIE DANYCH SYMBOLICZNYCH

1. Wstęp

Klasyfikacja ma na celu podział zbioru obiektów na klasy (skupienia) zawierające obiekty podobne ze względu na wartości zmiennych. Badacz musi więc na wstępie określić takie elementy procesu klasyfikacji, jak [Michalski, Strepp, Diday 1981, s. 35; Walesiak 2004, s. 53]:

1. Zbiór obiektów podlegających analizie.
2. Zbiór zmiennych charakteryzujących obiekty.
3. Miara odległości (podobieństwa) obiektów.
4. Metoda klasyfikacji.

Różnorodność zmiennych, ich złożony (kompleksowy) charakter i nietypowa struktura powodują utrudnienia w procesie klasyfikacji, dotyczące m.in. wyboru miary odległości i metody analizy. Bock i Diday tego rodzaju zmienne określili jako zmienne symboliczne [*Analysis of...* 2000, s. 2].

Jak pokazuje praktyka, wiele problemów marketingowych opisuje się za pomocą tego typu zmiennych. W artykule przedstawione zostaną zmienne symboliczne na przykładzie typowych pytań dotyczących preferencji nabywców samochodów osobowych.

2. Kategorie zmiennych

Dane w ujęciu klasycznym są rozpatrywane w dwóch podstawowych kategoriach: zmiennych mierzonych na mocnych skalach pomiaru, tj. ilorazowej i prze-

działowej (zmienne ilościowe) oraz na słabych skalach pomiaru, tj. porządkowej i nominalnej (zmienne opisowe).

Zmienne ilościowe przyjmują pojedynczą wartość liczbową. Mogą mieć charakter ciągły (np. wzrost = 1,75 cm) lub punktowy (np. wiek = 30 lat). Zmienne opisowe przyjmują jeden wariant cechy (np. wzrost – niski, kolor włosów – blond).

W klasyfikacji obiektów na cele badań marketingowych badacz spotyka się z kolei z wieloma zmiennymi symbolicznymi o bardziej rozbudowanej i skomplikowanej strukturze. W artykule sformułowano typowy zestaw pytań ankietowych na temat preferencji nabywców samochodów osobowych, aby na przykładzie zaprezentować rodzaje i specyfikę zmiennych symbolicznych.

Diday wymienia następujące typy zmiennych symbolicznych [Billard, Diday 2003, s. 471- 478; *Analysis of...* 2000, s. 2, 32- 37; Diday 2002, s. 7]:

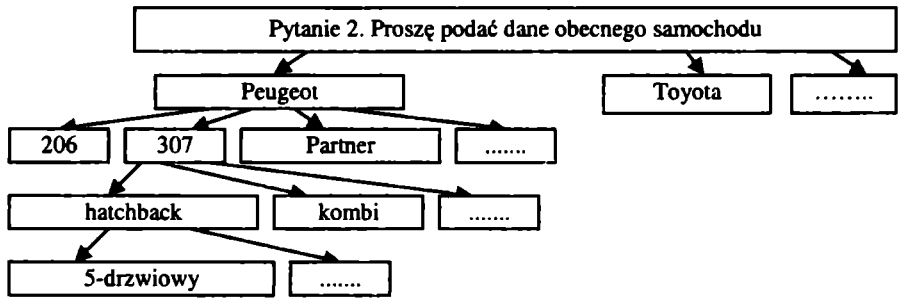
1. Przedziały klasowe (*interval-valued variable*) rozłączne lub spójne (nakładające się). Szczególnym przypadkiem jest tu zmienna ilościowa, gdzie w analizie klasycznej wartością (wariantem) zmiennej jest $x = a$, a w analizie symbolicznej ma ona postać przedziału klasowego $\xi = [a, a]$. Przedział klasowy spójny to zmienna z pytania 10: „Proszę podać przedział kwotowy na zakup samochodu", np. {35-45 tys. zł} i {40-60 tys. zł}. Przedziały klasowe rozłączne to zmienne z pytania o miejsce zamieszkania respondenta, np. {miasto 50-100 tys. mieszkańców} i {miasto 100-150 tys.} oraz pytania o dochód na osobę, np. {1000-1500 zł} i {1500-2000 zł}.

2. Listy kategorii lub wartości (*multivalued variable*) to zmienne dopuszczające występowanie wielu kategorii lub wartości tej samej zmiennej dla jednego obiektu. Obiekt może przyjmować wybrane lub wszystkie warianty zmiennej. Warianty zmiennej są najczęściej rangami lub kategoriami. Charakter listy wartości ma zmienna z pytania 8: „Proszę ocenić w skali 1-5 cechy samochodu", np. {5, 3, 4, 2, 5, 3} i {4, 3, 3, 2, 4, 5}. Charakter listy kategorii mają z kolei zmienne z pytania 12. „Czym kieruje się Pan/Pani przy wyborze samochodu?", np. {cena, marka, osiągi, technologia, opinia innych} i {koszty eksploatacji, marka} oraz z pytania 13: „Jakie akcesoria są dla Pana/Pani niezbędne?", np. {ABS, klimatyzacja, wspomaganie kierownicy} i {poduszka powietrzna}.

3. Listy kategorii lub wartości z wagami, prawdopodobieństwami czy częstościami (*multivalued modal variable*), to zmienna podobna do zmiennej listy kategorii lub wartości, z tym że wariantom są przyporządkowane stopnie ważności. Suma wag dla obiektu jest równa 1. Szczególnym przypadkiem tej zmiennej jest zmienna opisowa, w której obiekt przyjmuje wyłączne jeden wariant zmiennej (tzn. waga jest równa 1). Może to być też klasyczna zmienna binarna wzbogacona o wagi. Charakter listy kategorii z wagami ma zmienna z pytania 11: „Jakie, w kolejności, marki rozpatruje Pan/Pani?", np. {Toyota Yaris (0,5), Skoda Fabia (0,3), Volkswagen Polo (0,2)} i {Nissan Micra (0,6), Fiat Uno (0,4)} i {Peugeot 307 (1,0)}.

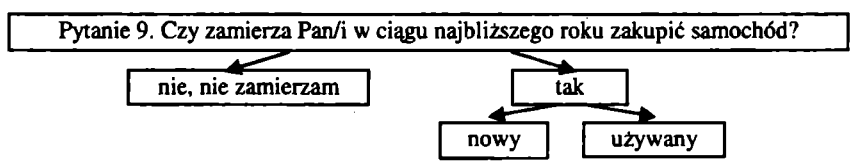
4. Zmienne strukturalne (o strukturze taksonomicznej, hierarchicznej, logicznej lub stochastycznej) przybierają postać drzewa wariantów, tzn. dla wariantów zmiennej można określić podwarianty, tj.:

- zmienne o strukturze taksonomicznej (*taxonomic dependent*), w których wariant zmiennej z najniższego poziomu nie determinuje wariantu z poziomu wyższego (brak wewnętrznej zależności logicznej lub stochastycznej); przykładem tej zmiennej jest zmienna z pytania 2: „Proszę podać dane obecnego samochodu", np. {Peugeot 307, hatchback 5-drzwiowy} (zob. rys. 1),



Rys. 1. Zmienna o strukturze taksonomicznej
Źródło: opracowanie własne.

- w zmiennej o strukturze hierarchicznej (*hierarchical dependent, Mother-Doughter*) dla co najmniej jednego z wariantów zmiennej nie można określić podwariantu (brakujący podwariant w analizie symbolicznej ma wartość *NA – non-applicable*); przykładem tej zmiennej jest zmienna z pytania 9: „Czy zamierza Pan/ Pani w ciągu najbliższego roku zakupić samochód?", np. {tak, nowy} (zob. rys. 2).



Rys. 2. Zmienna hierarchiczna
Źródło: opracowanie własne.

- zmienna o strukturze logicznej (*logical dependence*), występuje między zmiennymi *Y* i *Z*, gdy wartość drugiej zmiennej *Z* zależy logicznie lub funkcyjnie od wartości pierwszej zmiennej *Y*; przykładem jest zmienna „Rok produkcji samochodu i jego przebieg", np. jeśli samochód jest użytkowany 4 lata można domniemywać, że jego przebieg wynosi nie mniej niż 40 tys. km (zakładając, że przeciętny kierowca pokonuje w ciągu roku 10-15 tys. km).
- w zmiennej o strukturze stochastycznej (*stochastic dependence*) poszczególne szczeble hierarchii są ze sobą skorelowane (współczynnik korelacji i kowariancji jest różny od zera), np. jeśli respondent użytkuje samochód klasy średniej, to planowany samochód będzie podobnej lub wyższej klasy.

W praktyce występują również kombinacje zmiennych symbolicznych, m.in. (zob. [Billard, Diday 2003, s. 475, 477-478]:

- listy wartości lub kategorii (z wagami), gdzie kategorie mają charakter przedziału klasowego (*interval-valued modal varibale*), np. waga respondentów (kobiet) w różnych grupach wiekowych (10% kobiet w wieku 20-29 lat waży 45-50 kg) (zob. tab. 1),

Tabela 1. Zmienna – lista wartości, w której wariantem jest przedział klasowy spójny

Lp.	Wiek (w latach)	Waga (w kg)
1	20 – 29	{[45 – 50), 10; [50 – 55), 24; [55 – 60), 35; [60 – 65), 25; [65 – 70], (6)}
2	30 – 39	{[48 – 52), 12; [52 – 59), 15; [59 – 66), 40; [66 – 75), 10; [75 – 80], (3)}
3	40 – 49	{[51 – 57), 18; [57 – 63), 21; [63 – 70), 52; [70 – 78), (5); [78 – 85], (4)}
n	⋮	⋮

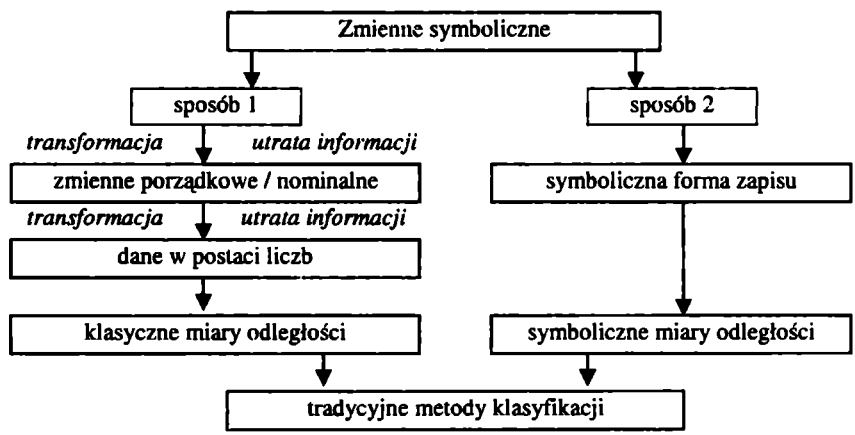
Źródło: opracowanie na podstawie [Billard, Diday 2003, s. 478]

- zmienne strukturalne z podwariantami o charakterze listy kategorii lub wartości (z wagami),
- zmienne strukturalne złożone (*multiple dependencies variable*) o różnego typu wewnętrznych zależnościach.

3. Formy zapisu danych

Metody klasyczne wymagają macierzowego zapisu danych, tj. każdemu obiektowi może być przypisany wyłącznie jeden wariant zmiennej w postaci liczbowej lub kategorii.

Analiza danych symbolicznych tradycyjną metodą klasyfikacji może przebiegać na dwa sposoby (zob. rys. 3).



Rys. 3. Postępowanie w przypadku analizy zmiennych symbolicznych metodami klasycznymi
Źródło: opracowanie własne.

Sposób pierwszy polega na przekształceniu zmiennych symbolicznych na zmienne klasyczne (porządkowe lub nominalne), a następnie przekształceniu wa-

riantów zmiennych na wartości liczbowe. W wyniku tego procesu otrzymujemy klasyczną macierz danych (zob. tab. 2).

Transformacja pierwotnych zmiennych prowadzi jednak do utraty informacji i tym samym może zniekształcać wyniki badania, a więc wpływać na użyteczność dokonanej analizy.

Tabela 2. Klasyczna tabela danych – wybrane zmienne

Lp.	Pytanie 1	Pytanie 3 (w tys. km)	Pytanie 4	Pytanie 7	Pytanie 9	Płeć	Wiek	Liczba osób w gospodarstwie
1	tak	300	2000	4	nowy	K	25	4
2	tak	96	2002	3	nowy	M	56	2
3	tak	165	1998	5	nie	M	40	2
4	nie	--	--	--	używany	K	32	6
5	tak	40	2003	2	nie	M	48	1
n	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Źródło: opracowanie własne.

Przekształcone zmienne o charakterze nominalnym lub porządkowym implikują zastosowanie właściwych dla nich klasycznych miar odległości (zob. pierwsze trzy kolumny tab. 3).

Tabela 3. Miary odległości dla zmiennych nominalnych binarnych, nominalnych wielostanowych, porządkowych i symbolicznych

Zmienne nominalne binarne	Zmienne nominalne wielostanowe	Zmienne porządkowe	Zmienne symboliczne (i tradycyjne)
Jaccard	Jukes & Cantor	GDM	Gowda & Diday
Roggers & Tonimoto	Tajima		Ichino & Yacuchi
Sokal & Sneath	Sokal & Michener		De Carvalho
Gower & Legendre			
miary klasyczne			miary symboliczne

Źródło: opracowanie na podstawie: [Walesiak 2002, s. 29, 33; *Analysis of...* 2000, s. 139-183; Malerba 2001, s. 474-476].

Sposób drugi polega na analizowaniu zmiennych w ich pierwotnej postaci, tj. wykorzystaniu w analizie podobieństwa obiektów, symbolicznych miar odległości (zob. ostatnia kolumna tab. 3). Tak przygotowaną macierz odległości można następnie poddać analizie adekwatną tradycyjną metodą klasyfikacji.

W metodach symbolicznych obiekty i ich charakterystyki zapisuje się w formie tzw. kompleksów za pomocą koniunkcji cech $[YRz]$, np. obiekt O_1 ma postać:

$$Y_1 = \{tak\} \wedge Y_2 \in \{Peugeot307, hatchback, 5\text{-}drzwiowy\} \wedge$$

$$\wedge Y_3 = \{300\} \wedge .... \wedge Y_8 \in \{5,3,4,2,5,3\} \wedge Y_9 \in \{Tak, nowy\} \wedge$$

$$\wedge Y_{10} \subseteq \{50-60\} \wedge Y_{11} \in \{ToyotaYaris(0,6) \vee SkodaFabia(0,4)\} \wedge$$
$$\wedge Y_{12} \in \{cena \wedge marka \wedge opinia\} \wedge Y_{13} \in \{ABS \wedge alarm\} \wedge .... \wedge$$
$$\wedge Y_{16} = \{średnie\} \wedge .... \wedge Y_{19} > 2000$$

gdzie: Y_i – zmienna ($i = 1,...,20$),

$\in, \subseteq, =, >$ – operator relacyjny R,

$\{z\}$ – odwzorowanie i-tej zmiennej dla j-otego obiektu (*mapping*).

Tabela 4. Tablica danych symbolicznych – wybrane zmienne

Lp.	Pytanie 1	Pytanie 2	Pytanie 7	Pytanie 8	Pytanie 9	Pytanie 10 (w tys. zł)	Pytanie 11	Pytanie 12	Dochód (w zł)	Wiek
1	tak	Peugeot 307, hatchback 5-drzw.	4	{5,3,4,2,5,3}	tak, nowy	50-60	{Toyota Yaris (0,6), Skoda Fabia (0,4)}	{cena, marka, opinia}	{pow. 2000}	25
2	tak	Opel Astra, kombi 5-drzw.	3	{4,3,4,3,3,4}	tak, nowy	70-80	{Ford Mondeo (0,5}, Ford Focus (0,3), Opel Omega (0,2)}	{komfort, sylwetka}	{1500--2000}	56
3	tak	Honda Accord 1.8 sedan 3-drzwiowy	5	{5,4,4,4,3,4}	nie	NA	NA	{technologia, osiągi}	{pow. 2000}	40
4	nie	NA	NA	NA	tak, używany	15-20	{Skoda Felicia (1,0)}	{cena}	{1000--1500}	32
5	tak	Opel Zafira van 5-drzwiowy	2	{3,5,4,2,5,5}	nie	NA	NA	{koszty eksploatacji}	{1500--2000}	48
n	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Źródło: opracowanie własne.

Obiekty i ich charakterystyki prezentuje się w tabeli danych symbolicznych (zob. tab. 4).

4. Hierarchiczne metody klasyfikacji

Cechą charakterystyczną hierarchicznych metod klasyfikacji jest możliwość ustalenia hierarchii skupień oraz graficzna prezentacja wyników w postaci drzewa skupień (dendrogramu lub piramidy).

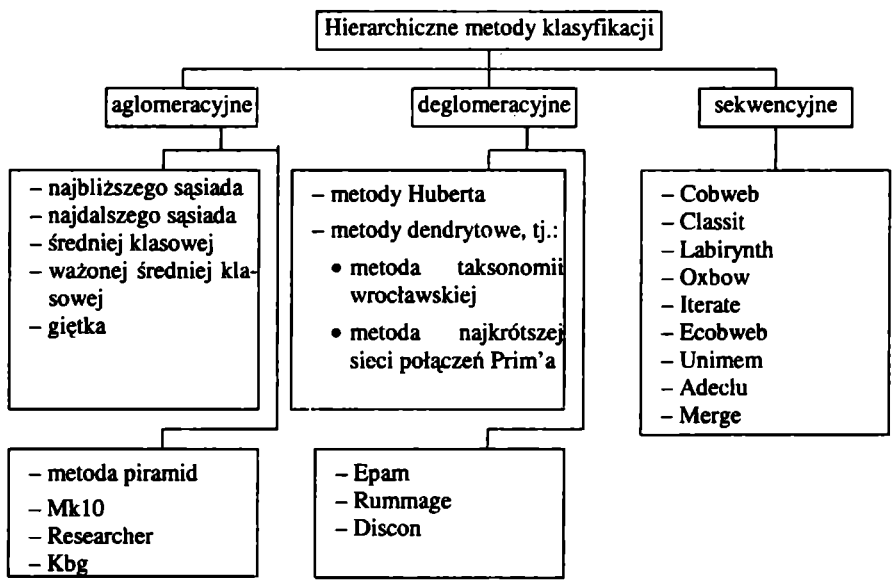
Wśród hierarchicznych metod klasyfikacji wyróżnia się techniki aglomeracyjne, deglomeracyjne oraz sekwencyjne.

W procedurach aglomeracyjnych na początku każdy obiekt tworzy odrębne skupienie. Poprzez łączenie skupień w tzw. grupy wyższego rzędu zmniejsza się liczbę klas do momentu, gdy wszystkie obiekty znajdą się w jednej klasie. Różnice między poszczególnymi wariantami metod aglomeracyjnych polegają na odmiennym pojmowaniu odległości między skupieniami [Ostasiewicz 1999, s. 88-89].

W procedurach deglomeracyjnych (podziałowych) na początku analizowany zbiór obiektów traktuje się jako skupienie n-elementowe. Proces podziału polega na stopniowym zwiększaniu liczby grup, aż do otrzymania skupień jednoelementowych.

Metody klasyczne[1] wymagają, aby zmienne symboliczne zostały zamienione na zmienne porządkowe lub nominalne, dlatego w analizie danych symbolicznych można zastosować tylko te metody, które wykorzystują w swoim algorytmie miary odległości właściwe do analizy zmiennych mierzonych na słabych skalach pomiaru (zob. rys. 4). Do klasyfikacji obiektów opisanych zmiennymi symbolicznymi nie można więc wykorzystać metody Warda, środków ciężkości czy medianowej. W swojej specyfice do pomiaru odległości między obiektami wykorzystują one bowiem metryki (np. odległość euklidesową).

Przegląd metod klasycznych można znaleźć m.in. w pracach: [Walesiak 1996; Gordon 1999; Everitt, Landau, Leese 2001].



Rys. 4. Hierarchiczne metody klasyfikacji
Źródło: opracowanie na podstawie: [Gatnar 1998, s. 100-120, 128-131; Walesiak 1996, s. 107- 112; Gordon 1999, s. 69-109; Everitt, Landau, Leese 2001, s.55- 89; Jain, Murty, Flynn 1999, s. 11- 13].

[1] Termin „klasyczne metody klasyfikacji" dotyczy metod analizy danych klasycznych, tj. pochodzących z mocnych (ilorazowa, przedziałowa) lub słabych (porządkowa, nominalna) skal pomiaru. Obiekty i ich charakterystyki są przedstawione w macierzy danych.

Metody symboliczne wyróżniają się formą zapisu danych, dopuszczalnymi kategoriami zmiennych (które mogą przyjmować więcej niż jeden wariant, przedział liczbowy i być wewnętrznie powiązane zależnościami), możliwością uwzględnienia w analizie zmiennych mierzonych na skalach pomiaru zarówno mocnych, jak i słabych bez konieczności transformacji zmiennych oraz wykorzystaniem symbolicznych miar odległości.

Najpopularniejszymi algorytmami klasyfikacji symbolicznej są: metoda piramid (aglomeracyjna), algorytm Rummage (deglomeracyjny) oraz Cobweb (sekwencyjny).

Metoda piramid tworzy klasy nierozłączne, tzn. obiekt może zostać przyporządkowany do więcej niż jednej klasy. Kryterium jakości podziału obiektów na skupienia jest różnica między zakresami obiektów symbolicznych reprezentujących klasy a charakterystyką obiektów do nich należących. Im mniejsza różnica, tym lepszy podział [zob. Gatnar 1998, s. 128-131; *Analysis of...* 2000, s. 312-323; Gordon 1999, s. 125-127; Everitt, Landau, Leese 2001, s. 152-154].

Rummage jest podstawowym algorytmem deglomeracyjnym; monotetycznym, tzn. dokonuje podziału obiektów ze względu na wartości pojedynczych cech. Jest szybszy niż Epam i Discon, bo przeszukuje znacznie mniejszą przestrzeń podziałów. W pierwszej kolejności wybiera cechę, dającą najlepszy podział. Epam z kolei, dokonując klasyfikacji, przeprowadza test. Jeśli obiekt go spełni, to dołącza się go do klasy, jeśli nie, tworzy się dla niego nową klasę. Discon nie stosuje żadnych kryteriów jakości podziału, więc musi zbudować wszystkie możliwe hierarchie klas i wybrać najlepszą z nich. Przeszukuje zatem całą przestrzeń podziałów.

Symboliczne hierarchiczne metody klasyfikacji rozwinięto o jeszcze jedną grupę technik – o metody sekwencyjne (*incremental methods*). Jest to nowa grupa, dlatego zostanie szerzej omówiona. Wyróżniającą cechą tych metod jest to, że przed rozpoczęciem klasyfikacji nie musi być znany cały zbiór obiektów. W przeciwieństwie do dwóch pierwszych grup metod dokonują one zarówno aglomeracji, jak i deglomeracji obiektów. Drzewo (hierarchia klas) nie jest drzewem binarnym, bo z jednego węzła (zbioru obiektów) może wychodzić kilka podzbiorów [Gatnar 1998, s. 102-103, 118-121].

Najważniejszym algorytmem sekwencyjnym jest Cobweb. Stał się on bazą do powstania kolejnych ulepszonych algorytmów (Classit, Labirynth itd.). Realizuje podział zbioru obiektów, tak żeby maksymalizować przydatność struktury skupień do przewidywania klasy, do której trafi nowy obiekt. Miarą jakości podziału jest tzw. użyteczność kategorii (zob. [Gatnar 1998, s. 119]). Identyfikuje ona własności obiektów należące do klasy. Pozwala też uzyskać dużą ilość informacji o obiekcie na podstawie znajomości klasy. Cobweb, szukając najodpowiedniejszej klasy, przydziela obiekt kolejno do każdego z istniejących skupień i ocenia jakość dokonanej klasyfikacji na podstawie użyteczności kategorii. Przydziela obiekt tam, gdzie jest największa użyteczność. Podczas procesu klasyfikacji cały czas oblicza wartość użyteczności. Łączy klasy i dzieli je, sprawdzając, czy nie da to lepszego

podziału, działa więc w dwie strony [Gatnar 1998, s. 108-109; Li, Biskas 1998, s. 4; Gordon 1999, s. 135].

Algorytm Iterate niweluje wrażliwość algorytmu Cobweb na kolejność pojawiania się obiektów. Algorytm Classit wzbogacono o możliwość analizowania zmiennych ilościowych. Algorytm Ecobweb dopuszcza występowanie w zbiorze zmiennych o charakterze listy wartości lub kategorii. Labirynth pozwala zaś na klasyfikację obiektów charakteryzowanych przez zmienne strukturalne (podobnie jak Merge).

5. Podsumowanie

W artykule zaprezentowano różne typy zmiennych symbolicznych, ich specyfikę oraz hierarchiczne metody klasyfikacji używane do ich analizy.

Klasyfikacja symboliczna, w odróżnieniu od klasyfikacji tradycyjnej, wykorzystuje głównie pojęcia naturalne, a zmienne mogą przyjmować więcej niż jedną wartość czy kategorię lub przedział wartości, uwzględnia wzajemne relacje w wariantach zmiennej. Ponadto dopuszcza występowanie w zbiorze zmiennych mierzonych zarówno na skalach pomiaru mocnych, jak i słabych, bez konieczności ich transformacji. Proces klasyfikacji nie wiąże się więc z utratą informacji na etapie przygotowywania danych do analizy.

Literatura

Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data, red. H.H. Bock, E. Diday, Springer Verlag 2000.

Billard L., Diday E., *From the Statistics of Data to the Statistic of Knowledge: Symbolic Data Analysis*, „Journal of the American Statistical Association", czerwiec 2003, vol. 98, nr 462, s. 456-469.

Diday E., *An Introduction to Symbolic Data Analysis and The Sodas Software*, „The Electronic Journal of Symbolic Data Analysis", lipiec 2002.

Everitt B.S., Landau S., Leese M., *Cluster Analysis*, Arnold Publishers, London 2001.

Gatnar E., *Symboliczne metody klasyfikacji danych*, PWN, Warszawa 1998.

Gordon A.D., *Classification*, Chapman&Hall/CRC, London 1999.

Jain A.K., Murty M.N., Flynn P.J., *Data Clustering: a Review*, „ACM Computing Survey", wrzesień 1999, vol. 31, nr 3.

Li C., Biswas G., *Conceptual Clustering with Numeric-and-Nominal Mixed Data – a New Similarity Based System*, „IEEE Transcript on KCE", 1998.

Malerba D., Esposito F., Gioviale V., Tamma V., *Comparing Dissimilarity Measures in Symbolic Data Analysis*, „Proceedings of the Joint Conferences on: New

Techniques and Technologies for Statistics and Exchange of Technology and Know-How (ETK-NTTS'01)" 2001, s. 473-481.
Michalski R.S., Strepp R., Diday E., *A Recent Advance in Data Analysis: Clustering Objects into Classes Characterized by Conjunctive Concepts*, „Progress in Pattern Recognition", North Holland Publishing Company, vol. 1, Holland 1981.
Ostasiewicz W., *Statystyczne metody analizy danych*, AE, Wrocław 1999.
Walesiak M., *Problemy decyzyjne w procesie klasyfikacji zbioru obiektów*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1010, AE, Wrocław 2004.
Walesiak M., *Metody analizy danych marketingowych*, PWN, Warszawa 1996.
Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, AE, Wrocław 2002.

Załącznik
kwestionariusz ankiety

1. Czy posiada Pani/i własny samochód?	● tak ○ nie
2. Proszę podać następujące dane obecnie posiadanego samochodu:	
Marka:	PEUGEOT
Model:	307
Typ nadwozia:	HATCHBACK 5- DRZWIOWE
3. Przebieg (km):	50.000
4. Rok produkcji:	2001
5. Pojemność silnika:	1.6
6. Rodzaj silnika:	BENZYNOWY
7. Proszę ocenić w skali od 1 do 5 zadowolenie z samochodu: 1- niezadowolony, 5 - b. zadowolony	○ 1 ○ 2 ○ 3 ● 4 ○ 5
8. Proszę ocenić w skali 1- 5 cechy obecnego samochodu: (1- ocena najniższa, 5- ocena najwyższa)	
Zużycie paliwa	○ 1 ○ 2 ○ 3 ● 4 ○ 5
Wygoda wnętrza	○ 1 ○ 2 ● 3 ○ 4 ○ 5
Pokonywanie nierówności	○ 1 ○ 2 ○ 3 ○ 4 ● 5
Widoczność	○ 1 ● 2 ○ 3 ○ 4 ○ 5
Wentylacja i ogrzewanie	○ 1 ○ 2 ○ 3 ● 4 ○ 5
Manewrowanie	● 1 ○ 2 ○ 3 ○ 4 ○ 5
9. Czy zamierza Pani/i w ciągu najbliższego roku zakupić samochód?	TAK, NOWY
10. Proszę podać przedział kwotowy, jaki zamierza Pani/i przeznaczyć na zakup samochodu:	od ..35.. do..45.. (tys. zł)

11. Jakie, w kolejności, marki rozpatruje Pani?	Marka:	TOYOTA	SKODA	VW
	Model:	YARIS	FABIA	POLO

12. Czym kieruje się Pani przy wyborze samochodu? (max 5 odpowiedzi)	☑ ceną zakupu ☐ dostępnością części ☐ komfortem jazdy ☐ kosztami eksploatacji ☐ kosztami ubezpieczenie ☐ estetyką pojazdu ☑ marką ☐ technologią ☑ opinią innych ☑ osiągami ☐ sylwetką pojazdu
13. Jakie akcesoria są dla Pani niezbędne? (max 3 odpowiedzi)	☐ ABS ☐ Szyby elektryczne ☑ Klimaryzacja ☐ Poduszki powietrzne ☐ Alarm ☑ Wspomaganie kierownicy

Charakterystyka respondenta:	
14. Płeć	KOBIETA
15. Wiek	..30.. (lat)
16. Wykształcenie	WYŻSZE
17. Miejsce zamieszkania	50 - 100 tys. MIESZK
18. Liczba osób w gospodarstwie domowym	3 OSOBY
19. Miesięczny dochód na osobę w Panaśi gospodarstwie (zł)	1000 - 1500zł
20. Status społeczny	PRACUJĄCY

HIERARCHICAL CLUSTERING METHODS IN SYMBOLIC DATA ANALYSIS

Summary

The article looks at the types and concepts of symbolic data and then attempts to review the available hierarchical clustering methods (traditional and symbolic) to analyze. Symbolic data were defined and contrasted with classical data and the examples were presented.