

TAKSONOMIA 14

Klasyfikacja i analiza danych – teoria i zastosowania

Justyna Wilk

Akademia Ekonomiczna we Wrocławiu

INTERPRETACJA I PROFILOWANIE KLAS W ANALIZIE DANYCH SYMBOLICZNYCH*

1. Wstęp

Metody klasyfikacji są jednymi z najczęściej wykorzystywanych procedur w analizie danych marketingowych [Punj, Stewart 1983; Gatnar, Walesiak 2004]. Rezultatem procesu klasyfikacji jest podział zbioru obiektów na względnie jednorodne klasy. Wyniki klasyfikacji stają się użyteczne, gdy wyodrębnione klasy zostaną scharakteryzowane ze względu na wewnętrzną strukturę, a także relacje międzyklasowe. W badaniach marketingowych istotnymi kwestiami są więc interpretacja i profilowanie klas.

W praktyce wiele zagadnień marketingowych jest charakteryzowanych za pomocą zmiennych symbolicznych (*symbolic variables*). Złożona struktura i nietypowy charakter tych zmiennych powodują, że do opisu klas obiektów opisanych tego rodzaju zmiennymi nie można zastosować dotąd proponowanych procedur.

Artykuł ma na celu przybliżenie sposobu opisu klas obiektów symbolicznych w podstawowym zakresie na przykładach empirycznych.

2. Specyfika danych symbolicznych

Zmienne symboliczne, w przeciwieństwie do zmiennych rozumianych w sensie klasycznym, tj. mierzonych na mocnych bądź słabych skalach pomiaru, mają bar-

* Autorka korzysta ze stypendium współfinansowanego przez Unię Europejską ze środków Europejskiego Funduszu Społecznego oraz przez budżet państwa, realizowanego w ramach ZPORR, priorytet II, działanie 2.6.

dziej złożoną strukturę. Poszczególne obiekty mogą przyjmować więcej niż jeden wariant zmiennej, niekoniecznie w postaci liczby rzeczywistej (np. przedział liczbowy bądź kategoria). Warianty zmiennej nie muszą być rozłączne (np. przedziały liczbowe mogą się nakładać). Wariantom zmiennych dla pojedynczego obiektu mogą być przypisane wagi (np. zmienne opisujące udział, strukturę). Zmienne mogą być powiązane relacjami, zależnościami (np. silnie skorelowane), mogą tworzyć struktury drzewiaste [Billard, Diday 2003].

Zmienne symboliczne są adekwatne do opisu obiektów zarówno prostych (np. produktów, konsumentów), jak i złożonych (np. gospodarstw domowych, województw). W przypadku obiektów opisanych zmiennymi symbolicznymi w miejsce klasycznej macierzy danych (zapis wektorowy) dane prezentuje się w tablicy danych symbolicznych.

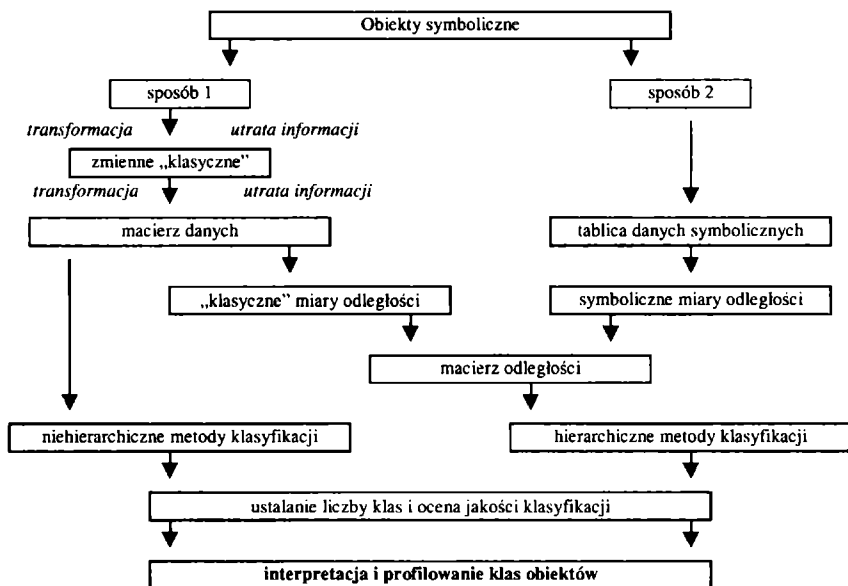
Jeśli przyjmiemy, że Y_k to zmienna symboliczna (k – numer zmiennej, $k = 1, \dots, p$) ze zbioru Y (gdzie V_k zbiór realizacji zmiennej Y_k) o wariantach y_{fk} ($f = 1, \dots, z$), to v_{ik} jest wartością zmiennej Y_k dla obiektu o_i . O jest zbiorem obiektów symbolicznych $O = \{o_1, \dots, o_n\}$, gdzie i to numer obiektu, $i = 1, \dots, n$.

Podstawowymi rodzajami zmiennych symbolicznych są [Billard, Diday 2003, s. 471-478]:

1. Przedziały liczbowe (*interval-valued variable*), gdzie $v_{ik} = [a_{fk}; b_{fk}]$.
2. Listy kategorii, wartości (*multivalued variable*). To zmienne dopuszczające występowanie wielu kategorii lub wartości tej samej zmiennej dla jednego obiektu, gdzie $v_{ik} = \{y_{fk}, \dots, y_{zk}\}$.
3. Listy kategorii (wartości) z wagami, prawdopodobieństwami czy częstościami (*multivalued modal variable*). To podobne zmienne do listy kategorii lub wartości, z tym, że wariantom przyporządkowane są stopnie ważności. Zmienna symboliczna z wagami jest odwzorowaniem $Y_k = (y_{fk}, p(y_{fk}))$, $p(y_{fk})$ to waga f -tego wariantu zmiennej Y_k dla obiektu o_i , gdzie $p(y_{fk}) \in [0; 1]$ i $\sum_{f=1}^z p(y_{fk}) = 1$, zatem $v_{ik} = \{y_{1k}(p(y_{1k})), \dots, y_{zk}(p(y_{zk}))\}$.
4. Zmienne strukturalne (*dependent variables*) w postaci drzewa wariantów. Warianty mogą mieć charakter zmiennych typu 1, 2 lub 3.

3. Proces klasyfikacji obiektów symbolicznych

Klasyfikację obiektów symbolicznych można przeprowadzić na dwa sposoby (zob. rys. 1).



Rys. 1. Proces klasyfikacji obiektów symbolicznych – dwie ścieżki
Źródło: opracowanie własne.

Sposób pierwszy różni się od drugiego tym, że na etapie przygotowania danych do analizy należy przekształcić zmienne symboliczne na zmienne mierzone na mocnych bądź słabych skalach pomiaru. Następnie postępuje się jak w typowej procedurze klasyfikacyjnej (zob. [Walesiak 2004; Gordon 1999]).

Alternatywnym sposobem, korzystniejszym z merytorycznego punktu widzenia, jest wybór hierarchicznych metod klasyfikacji i wykorzystanie symbolicznych miar odległości skonstruowanych do analizy danych symbolicznych (zob. [Bock, Diday 2000, s. 153-185]). Pozwala on analizować dane symboliczne w ich pierwotnej postaci, tj. nie tracąc informacji przy przekształcaniu zmiennych.

4. Opis klas obiektów symbolicznych

Interpretacja i profilowanie klas obiektów to ostatni etap procesu klasyfikacji. Polega on na wskazaniu cech charakteryzujących poszczególne klasy oraz na wyjaśnieniu, jakimi czynnikami różnią się wyodrębnione klasy. Podstawą interpretacji klas są zmienne, które brały udział w procesie klasyfikacji zbioru obiektów. Profilowanie przeprowadza się natomiast na podstawie zmiennych, które nie brały udziału w procesie klasyfikacji. Sposób opisu klas w analizie danych symbolicznych jest uzależniony od wybranej ścieżki klasyfikacji obiektów.

4.1. Opis klas obiektów symbolicznych po transformacji

Jeśli klasyfikacja obiektów symbolicznych będzie przeprowadzona z wykorzystaniem sposobu pierwszego, to w wyniku transformacji otrzymane zostaną zmienne mierzone na mocnych lub słabych skalach pomiaru. Do opisu klas można wtedy zastosować dotychczas zalecane techniki [Hair i in. 1998; Jain, Murty, Flynn 1999, s. 282; Gordon 1999, s. 209-211].

Dla zmiennych ilościowych (ilorazowe i przedziałowe) wyznacza się zwykle środki ciężkości poszczególnych klas oraz odchylenia standardowe w poszczególnych klasach. Dla zmiennych jakościowych (tj. mierzonych na słabych skalach pomiaru) wyznacza się frakcje i odsetki występowania w danej klasie poszczególnych kategorii zmiennych [Gatnar, Walesiak 2004, s. 66-67].

4.2. Opis klas obiektów symbolicznych w ich pierwotnej postaci

Jeśli do analizy wykorzysta się sposób drugi, to w wyniku otrzymane zostaną skupienia obiektów opisanych zmiennymi symbolicznymi. Specyfika zmiennych symbolicznych uniemożliwia zastosowanie dotychczas zalecanych technik analitycznych. Zaproponowano więc odrębny sposób opisu klas [Brito 2004, s. 209, 211-213; ASSO... 2001, s. 31-33; Verde, Lechevallier, Chavent 2003, s. 3-4]. Taki sposób opisu klas obiektów symbolicznych wyróżnił się z tzw. klasyfikacji pojęciowej [Fisher, Langley 1986; Michalski, Stepp, Diday 1981]. Zagadnienie opisu klas obiektów symbolicznych pojawiło się również w opracowaniach na temat analizy skupień, m.in. w pracach [Gordon 1999, s. 209] oraz [Jain, Murty, Flynn 1999, s. 282-283].

Skupienia obiektów symbolicznych można opisać w wąskim lub szerokim zakresie [Brito 2004, s. 209-211; ASSO... 2001]. Sposoby opisu klas zostaną zaprezentowane na podstawie wybranych danych z badania preferencji użytkowników Internetu (tab. 1). Zostali oni zapytani m.in. o ilość czasu spędzanego tygodniowo w Internecie (przedział liczbowy), miesięczne wydatki na Internet (przedział liczbowy), najczęściej odwiedzane portale (lista kategorii) oraz motywy korzystania z Internetu (lista kategorii z wagami). W wyniku klasyfikacji otrzymano podział obiektów na trzy klasy. Do profilowania wykorzystano m.in. zmienne dotyczące miesięcznego dochodu netto (przedział liczbowy) oraz źródła dochodu (lista kategorii).

4.2.1. Opis klas obiektów symbolicznych w szerokim zakresie

Charakterystyka skupienia w szerokim zakresie (*long description*) polega na tym, że konstruuje się „nowy” obiekt symboliczny obejmujący wszystkie elementy klasy. Jest to opis rozszerzony (uogólniony), tzn. może obejmować swoim zasięgiem również obiekty nienależące do klasy, które mogą dopiero się pojawić, w miarę kolekcjonowania danych. Opis klas można zapisać w formie koniunkcji cech

[Gatnar 1998, s. 69, 75]. Jego konstrukcja przebiega w różny sposób w zależności od rodzaju analizowanych zmiennych.

Dla przedziałów klasowych należy znaleźć najmniejszy przedział wartości, obejmujący wszystkie przedziały wartości zmiennej charakteryzujące obiekty badanego skupienia.

Charakterystykę klas obiektów opisanych przedziałami liczbowymi prezentuje tab. 2, kolumny nr 1, 2, 5. Internauci należący do klasy 1 spędzają najwięcej czasu w Internecie – od 16 do 30 godzin tygodniowo, wydając miesięcznie na ten cel do 120 zł, przy relatywnie najwyższych dochodach (między 2001 a 3200 zł). Opis klas można zapisać w następujący sposób:

Klasa 1 := [czas $\subseteq$ [16 – 30] $\wedge$ wydatki $\subseteq$ [0 – 120] $\wedge$ dochód $\subseteq$ [2001 – 3200]].

Klasa 2 := [czas $\subseteq$ [5 – 18] $\wedge$ wydatki $\subseteq$ [21 – 80] $\wedge$ dochód $\subseteq$ [0 – 2000]].

Klasa 3 := [czas $\subseteq$ [1 – 5] $\wedge$ wydatki $\subseteq$ [0 – 80] $\wedge$ dochód $\subseteq$ [0 – 1600]].

W przypadku list kategorii (wartości) wyznacza się zbiór wariantów, w skład którego wchodzi wszystkie warianty charakteryzujące poszczególne obiekty skupienia.

Internauci wszystkich klas odwiedzają portal Wirtualna Polska. Internauci klasy drugiej najczęściej odwiedzają portale poświęcone tematyce komputerowej, osoby z klasy trzeciej – portale związane z polityką i telefonią komórkową. Źródłami dochodów w klasie drugiej są m.in. stypendium i kieszonkowe, a klasy pierwszej – głównie praca zarobkowa (tab. 2, kolumny nr 3, 6).

W odniesieniu do listy kategorii (wartości) z wagami opis skupienia można otrzymać na dwa sposoby:

1. Uogólnianie za pomocą kryterium maksimum. Dla wariantu y_f zmiennej Y_k , opisującej obiekty tworzące skupienie, ustala się maksymalną wagę spośród wszystkich. Wyznaczony rozkład wag nie jest rozkładem prawdopodobieństwa i suma wag może być większa niż 1.

Internauci należący do klasy pierwszej ocenili motyw „usługi komercyjne” co najwyżej na 80% ważności. Internauci z klasy drugiej i trzeciej przyznają mu maksimum 30% ważności. Użytkownicy z klasy drugiej zwykle poszukują w Internecie informacji i korzystają z usług edukacyjnych (wysokie oceny). W grupie drugiej motyw „rozrywka” jest istotny (do 50% stopnia ważności), w klasie trzeciej jest on natomiast mało istotny (co najwyżej w 10%), zob. tab. 1, kolumna 4. Opis klas za pomocą kryterium maksimum można wyrazić w następujący sposób:

Klasa 1 := [motywy $\leq$ {usługi komercyjne (80%), rozrywka (30%), wyszukiwanie informacji (20%)}].

Klasa 2 := [motywy $\leq$ {usługi komercyjne (30%), rozrywka (50%), wyszukiwanie informacji (80%), komunikacja (40%), edukacja (60%)}].

Klasa 3 := [motywy $\leq$ {usługi komercyjne (20%), rozrywka (10%), wyszukiwanie informacji (30%), komunikacja (30%)}].

Tabela 1. Wycinek tabeli danych pochodzących z badania preferencji użytkowników Internetu

Wyszczególnienie		Czas spędzany w Internecie w tygodniu (liczba godzin)	Miesięczne wydatki na Internet (zł)	Najczęściej odwiedzane portale	Główne motywy korzystania z Internetu (%)	Miesięczny dochód netto (zł)	Źródła dochodów
zmienne	przebieg	przebieg	przebieg	przebieg	przebieg	przebieg	przebieg
nr klasy	nr obiektu	nr	nr	nr	nr	nr	nr
1	1	[16, 20]	[0, 20]	Interia, Wirtualna Polska, Gazeta	usługi komercyjne (70%), rozrywka (30%)	[2001, 2400]	praca
	2	[18, 24]	[41, 60]	Interia, Fotoklub, Futbol	usługi komercyjne (60), rozrywka (30%), wyszukiwanie informacji (10%)	[2401, 2800]	praca
	3	[20, 30]	[101, 120]	Gazeta, Interia	usługi komercyjne (80%), wyszukiwanie informacji (20%)	[2001, 2400]	praca, renta
	4	[20, 25]	[21, 40]	Interia, Wirtualna Polska	usługi komercyjne (70%), rozrywka (30%)	[2801, 3200]	praca
	5	[6, 11]	[41, 60]	Orange, Wirtualna Polska, Chip, inne	usługi komercyjne (20%), komunikacja (20%), edukacja (60%)	[0, 800]	kieszonkowe
	6	[6, 10]	[21, 40]	Era, Komputer Świat, O2, Chip	rozrywka (50%), komunikacja (40%), edukacja (10%)	1201, 1600	kieszonkowe, stypendium
2	7	[8, 12]	[61, 80]	Gazeta, Orange, Polityka, Chip, Komputer Świat	wyszukiwanie informacji (70%), komunikacja (10%), edukacja (20%)	[1201, 1600]	stypendium, renta, kieszonkowe
	8	[5, 10]	[21, 40]	Polityka, Chip, Komputer Świat	usługi komercyjne (30%), wyszukiwanie informacji (60%), edukacja (10%)	[0, 800]	kieszonkowe
	9	[7, 14]	[21, 40]	Chip, Komputer Świat, Komputer World, inne	rozrywka (40%), wyszukiwanie informacji (30%), edukacja (30%)	[1201, 1600]	stypendium, kieszonkowe
	10	[9, 18]	[61, 80]	Chip, Komputer Świat, Orange	wyszukiwanie informacji (80%), edukacja (20%)	[1600, 2000]	kieszonkowe
3	11	[1, 4]	[0, 20]	Onet, Yahoo	rozrywka (100%)	[801, 1200]	emerytura
	12	[2, 5]	[41, 60]	Onet, Wirtualna Polska, Yahoo	usługi komercyjne (10%), komunikacja (30%), rozrywka (60%)	[1201, 1600]	renta
	13	[3, 4]	[21, 40]	Yahoo, O2, Panorama Firm, Plus, Onet	usługi komercyjne (20%), rozrywka (80%)	[1201, 1600]	emerytura
	14	[3, 5]	[41, 60]	Wprost, Polityka, Onet, Komputer Świat	usługi komercyjne (10%), rozrywka (60%), wyszukiwanie informacji (30%)	[1201, 1600]	emerytura
	15	[2, 4]	[61, 80]	Era, Onet	rozrywka (90%), komunikacja (10%)	[0, 800]	zasilek

Źródło: opracowanie własne.

2. Uogólnianie za pomocą kryterium minimum polega na tym, że porównując obiekty względem f -tego wariantu zmiennej Y_k , wybiera się minimalną wagę analizowanego wariantu. W otrzymanym rozkładzie suma wag może być mniejsza niż 1.

Tabela 2. Opis klas obiektów symbolicznych w szerokim zakresie

Zmienne		Czas spędzany w Internecie w tygodniu (liczba godzin)	Miesięczne wydatki na Internet (zł)	Najczęściej odwiedzane portale	Główne motywy korzystania z Internetu (%)	Miesięczny dochód netto (zł)	Źródła dochodów
Nr klasy	% obiektów	1	2	3	4	5	6
1	26	[16, 30]	[0, 120]	Interia, Wirtualna Polska, Gazeta, Fotoklub, Futbol	usługi komercyjne (60-80), rozrywka (30), wyszukiwanie informacji (10-20)	[2001, 3200]	praca, renta
2	40	[5, 18]	[21, 80]	Orange, Wirtualna Polska, Era, Komputer Świat, O2, Gazeta, Polityka, Chip, Computer Word, inne	usługi komercyjne (20-30), rozrywka (50), komunikacja (10-40), edukacja (10-60), wyszukiwanie informacji (30-80)	[0, 2000]	stypendium, kieszonkowe, renta
3	34	[1, 5]	[0, 80]	Wirtualna Polska, O2, Panorama Firm, Plus, Wprost, Polityka, Komputer Świat, Era, Onet, Yahoo	usługi komercyjne (10-20), rozrywka (60-100), wyszukiwanie informacji (30), komunikacja (10-30)	[0, 1600]	emerytura, renta, zasiłek

Źródło: opracowanie własne z wykorzystaniem modułu *Clint* programu *Sodas v 2.0*.

Internauci należący do klasy pierwszej najczęściej wykorzystują Internet do usług komercyjnych (minimum 60% ważności). Internauci z klas drugiej i trzeciej stosunkowo rzadko korzystają z tego typu usług. Użytkownicy należący do klasy drugiej zwykle poszukują w Internecie informacji. Użytkownicy z klasy trzeciej oceniają na jednakowym poziomie motywy „rozrywka” i „wyszukiwanie informacji” (co najmniej 30% poziom ważności).

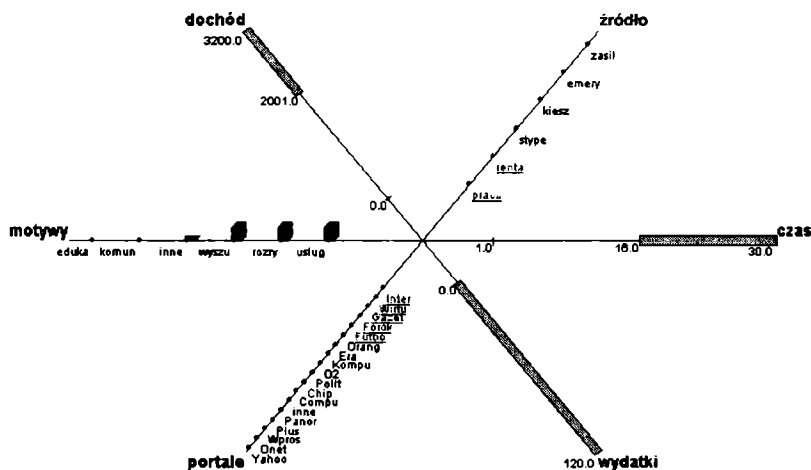
Klasa 1: = [motywy $\leq$ {usługi komercyjne (60%), rozrywka (30%), wyszukiwanie informacji (10%)}].

Klasa 2: = [motywy $\leq$ {usługi komercyjne (20%), rozrywka (50%), wyszukiwanie informacji (30%), komunikacja (10%), edukacja (10%)}].

Klasa 3: = [motywy $\leq$ {usługi komercyjne (10%), rozrywka (60%), wyszukiwanie informacji (30%), komunikacja (10%)}].

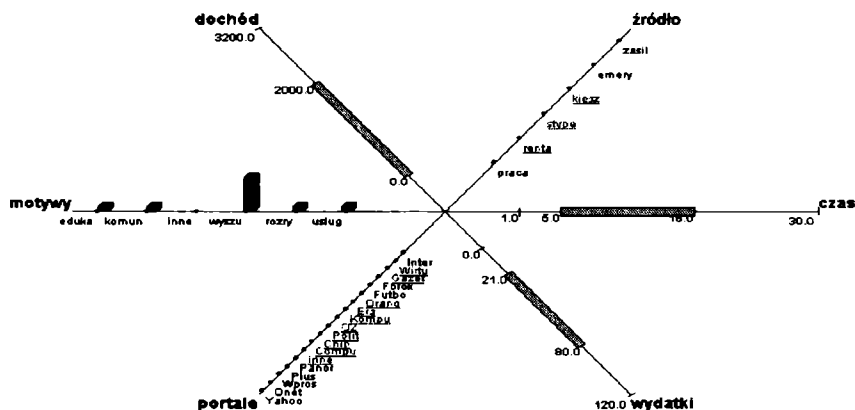
Opis klas obiektów, charakteryzowanych zmiennymi listami kategorii z wagami, można uzyskać, stosując jedno z omówionych kryteriów bądź z połączenia obu możliwości (wyznaczając przedziały liczbowe wag dla poszczególnych wariantów), zob. tab. 2, kolumna 4.

Na rysunkach 2 i 3 przedstawione zostały charakterystyki klas pierwszej i drugiej w szerokim zakresie na podstawie zmiennych z tab. 1. Osie prezentują realizacje zmiennych dla poszczególnej klasy. Słupki oznaczają wagi wariantów zmiennej listy kategorii z wagami w odniesieniu do danego obiektu (kryterium maksimum). Dla przedziałów liczbowych zaznaczono obszary realizacji zmiennej w klasie, podkreślono realizacje list kategorii dla poszczególnej klasy.



Rys. 2. Charakterystyka klasy pierwszej w szerokim zakresie

Źródło: opracowanie na podstawie danych z tab. 2, w module *Clint* programu *Sodas v 2.0*.



Rys. 3. Charakterystyka klasy drugiej w szerokim zakresie

Źródło: opracowanie na podstawie danych z tab. 2, w module *Clint* programu *Sodas v 2.0*.

4.2.2. Opis klas obiektów symbolicznych w wąskim zakresie

Zmienne strukturalne przyjmują warianty w postaci podobnej przedziałów liczbowych lub list kategorii (z wagami), opisuje się je więc podobnie jak pozostałe zmienne symboliczne.

Opis w wąskim zakresie (*short description*) polega na tym, że poszukuje się takiego wariantu zmiennej (wartości, kategorii lub przedziału liczbowego) lub stopnia ważności wariantów, który nie występuje w pozostałych klasach, a występuje dla każdego obiektu należącego do danej klasy. Jest to zawężony (uproszczony) opis klas, w formie tzw. pojęć [Gatnar 1998, s. 75]. Dla zmiennej „dochód” nie udało się wyznaczyć takich wariantów (tab. 3).

Do klasy pierwszej trafili np. internauci, których czas spędzany w Internecie pokrywa się z liczbą 20 godzin tygodniowo, a wydatki pokrywają się z przedziałem 81-120 zł miesięcznie i są one inne niż w pozostałych klasach. Użytkownicy należący do tej klasy odwiedzają portal Interia, a głównym motywem korzystania z Internetu są „usługi komercyjne” (co najmniej 60% ważności). Źródłem utrzymania jest przede wszystkim praca zarobkowa.

Tabela 3. Opis klas obiektów symbolicznych w wąskim zakresie

Zmienne	Czas spędzany w Internecie w tygodniu (liczba godzin)	Miesięczne wydatki na Internet (zł)	Najczęściej odwiedzane portale	Główne motywy korzystania z Internetu (%)	Miesięczny dochód netto (zł)	Źródła dochodów
Nr klasy	1	2	3	4	5	6
1	[20, 20]	[81, 120]	Interia	usługi komercyjne (60)	-	praca
2	[9, 10]	-	Chip	edukacja (10)	-	kieszonkowe
3	[4, 4]	-	Onet	rozrywka (60)	-	-

Źródło: opracowanie własne z wykorzystaniem modułu *Clint* programu *Sodas v 2.0*.

$Klasa\ 1 := [czas \subseteq [20 - 20] \wedge wydatki \subseteq [81 - 120] \wedge portale \{Interia\} \wedge$
 $\wedge motywy \{usługi\ komercyjne\ (60\%)\} \wedge źródła\ dochodów \{praca\}].$

5. Podsumowanie

Sposób charakteryzowania klas obiektów symbolicznych jest uzależniony od wyboru ścieżki klasyfikacji obiektów. Symboliczny sposób charakteryzowania skupień daje pełniejszy opis klas, pozwala przewidzieć, do której klasy trafi „nowy” obiekt, bez konieczności powtarzania analizy. Opis w szerokim zakresie daje uogólnioną charakterystykę klasy, natomiast opis w wąskim zakresie – charakterystykę zawężoną. Opis w wąskim zakresie może być stosowany jedynie wówczas, gdy można ustalić wariant zmiennej (kategorię lub przedział liczbowy) bądź stopień ważności wariantu, który nie występuje w pozostałych klasach, a występuje dla każdego obiektu należącego do danej klasy. Charakterystyki poszczególnych

klas można przedstawić w postaci opisowej, tabelarycznej i graficznej. Symboliczny opis klas może być również uzupełnieniem wyników klasyfikacji obiektów opisanych zmiennymi „klasycznymi”.

Literatura

- ASSO Developers Partners (2001), *Scientific-technical Report for WP 6. Unsupervised Classification, Validation and Cluster Representation*.
- Billard L., Diday E. (2003), *From the Statistics of Data to the Statistic of Knowledge: Symbolic Data Analysis*, „Journal of American Statistical Association” June, vol. 98, nr 470-487.
- Bock H.H., Diday E. (red.) (2000), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg.
- Brito P. (2004), *Clustering Interpretation. Interpreting Clusters by Using the Module CLINT*, [w:] ASSO Developers Partners, *User Manual for SODAS 2 Software*, s. 207-214.
- Fisher D., Langley P. (1996), *Conceptual Clustering and Its Relationship to Numerical Taxonomy*, [w:] W.A. Gale (red.), *Artificial Intelligence and Statistics*, Addison-Wesley, Reading.
- Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.
- Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław.
- Gordon A.D. (1999), *Classification*, „Monographs on Statistics and Applied Probability” nr 82, Chapman & Hall/CRC.
- Hair J.F., Anderson R.E., Tatham R.L., Black W.C. (1998), *Multivariate Data Analysis*, Pearson Prentice Hall, Englewood Cliffs.
- Jain A.K., Murty M.N., Flynn P.J. (1999), *Data Clustering: A Review*, „ACM Computing Survey” vol. 31, nr 3, s. 264-322.
- Michalski R.S., Stepp E., Diday E. (1981), *A Recent Advanced in Data Analysis: Clustering Objects Into Classes Characterized by Conjunctive Concepts*, [w:] L. Kanal, A. Rosenfeld (red.), „Progress in Pattern Recognition” vol. 1.
- Punj G., Stewart D.W. (1983), *Cluster Analysis in Marketing Research: Review and Suggestions for Application*, „Journal of Marketing Research” vol. 10, s. 134-148.
- Verde R., Lechevallier Y., Chavent M. (2003), *Symbolic Clustering Interpretation and Visualization*, „Electronic Journal of Symbolic Data Analysis” vol. 1, nr 1.
- Walesiak M. (2004), *Problemy decyzyjne w procesie klasyfikacji zbioru obiektów*, [w:] J. Dziechciarz (red.), *Ekometria 13*, Prace Naukowe AE we Wrocławiu nr 1010, AE, Wrocław, s. 52-71.

CLUSTER INTERPRETATION IN SYMBOLIC DATA ANALYSIS

Summary

There was presented the way of cluster interpretation in symbolic data analysis in the paper. Firstly the idea and the types of symbolic variables were defined. Then two ways of classification procedure for symbolic objects were suggested. After that the ways of cluster interpretation of symbolic objects were characterized according to classification procedure, with numerical examples.