PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Marcin Pełka
Akademia Ekonomiczna we Wrocławiu

MODYFIKACJA METODY KLASYFIKACJI PODZIAŁOWEJ OPARTEJ NA KRYTERIACH

1. Wstęp

Metody klasyfikacji, należące do grupy metod statystycznej analizy wielowymiarowej, mają szerokie zastosowanie w badaniach marketingowych, a zwłaszcza w segmentacji rynku, pozycjonowaniu produktu (przedsiębiorstwa) na rynku, identyfikacji rynków testowych oraz określaniu struktury rynku [Walesiak 1996, s. 117, 120, 151].

Celem niniejszego artykułu jest zaprezentowanie idei symbolicznej klasyfikacji opartej na kryteriach (*criterion-based divisive clustering*). Metoda ta została zaproponowana przez M. Chavent (por. [Chavent 1998; Chavent i in. 1999; Chavent, Stephan 1999]). W swych założeniach metoda pozwala albo na klasyfikację obiektów, które opisane są zmiennymi w postaci przedziału liczbowego czy innych „silnych skal" pomiaru, albo na klasyfikację obiektów opisanych zmiennymi w postaci listy kategorii czy listy kategorii z wagami. Natomiast z powodu zastosowanej w niej miary jakości klasyfikacji nie pozwala na klasyfikację obiektów opisanych różnymi typami zmiennych łącznie. Dlatego też dodatkowym celem niniejszego opracowania jest pokazanie modyfikacji tej metody, która pozwala na klasyfikację obiektów opisanych przez zmienne dowolnego typu. W części empirycznej porównano wyniki klasyfikacji uzyskane metodą niezmodyfikowaną i zmodyfikowaną na przykładzie rynku samochodowego.

2. Typy zmiennych w symbolicznej analizie danych

W przypadku obiektów symbolicznych możemy mieć do czynienia z następującymi rodzajami zmiennych [Bock, Diday 2000, s. 2-3]:

1) ilorazowe, przedziałowe, porządkowe, nominalne;

2) kategorie, czyli dane tekstowe, np. biały, zielony;

3) przedziały liczbowe, które np. określają dochód konsumenta (1000 zł, ..., 2500 zł) czy ilość spalanej benzyny na 100 km (5 litrów, ..., 11 litrów), co w przypadku dochodu oznacza, że może on wynieść od 1000 zł do 2500 zł (np. praca akordowa);

4) lista kategorii; tu przykładem może być standard wyposażenia pewnego samochodu (obiektu) określony jako: niski, średni, wysoki, co oznacza, że oferowany jest w trzech rodzajach standardów wyposażenia;

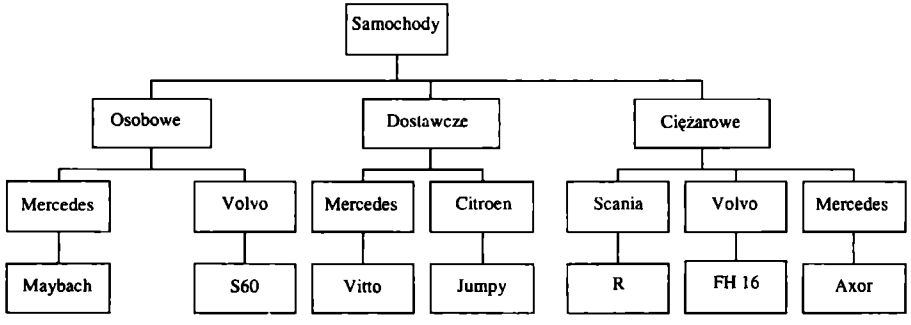
5) lista kategorii z wagami (prawdopodobieństwami), gdzie oprócz listy kategorii występują wagi, z jakimi obiekt posiada wybraną kategorię, np. jeżeli wybrać zmienną „typ nadwozia samochodu" w postaci listy: sedan (0,45), coupé (0,20), pickup (0,25), kabriolet (0,10), inne (0,00) dla pewnej grupy osób, to oznacza to, że 45% samochodów posiadanych przez te osoby to samochody o nadwoziu typu sedan, 10% zaś to kabriolety. Osoba, która w ankiecie zaznaczyła, że jej samochód to kabriolet, należy właśnie do tej grupy. Wagi w zmiennych w postaci list kategorii z wagami najczęściej odpowiadają częstościom występowania danej kategorii zmiennej, w niektórych przypadkach wagi te mogą wynikać też z innych badań czy ocen ekspertów;

6) w literaturze oprócz wyżej wymienionych typów zmiennych symbolicznych, wyróżnia się także zmienne strukturalne [Bock, Diday 2000, s. 33-37]:

a) zmienne o zależności funkcyjnej lub logicznej pomiędzy zmiennymi, gdzie *a priori* ustalono reguły funkcyjne lub logiczne decydujące o tym, jaką wartość przyjmie dana zmienna,

b) zmienne hierarchiczne, w których *a priori* ustalono warunki, od których zależy, czy zmienna dotyczy danego obiektu, czy też nie,

c) zmienną taksonomiczną, dla której *a priori* ustalono systematykę, według której przyjmują one swoje realizacje. Zmienną tego typu przedstawia rys 1.



Rys. 1. Zmienna hierarchiczna – samochody
Źródło: opracowanie własne.

3. Algorytm klasyfikacji podziałowej opartej na kryteriach

Algorytm klasyfikacji podziałowej opartej na kryteriach w każdym kroku poszukuje optymalnego podziału zbioru obiektów (na podstawie wszystkich zmiennych. Podział ten pozwala z jednej klasy C_i uzyskać dwie klasy C_i^1, C_i^2, do których następnie przydzielamy obiekty zgodnie z kryteriami podziału. Kryteria (reguły) przydzielania obiektów do klas zależą od typu zmiennej symbolicznej, z jaką mamy do czynienia, gdy:

1) dla zmiennych w postaci przedziałów liczbowych punktem podziału (*cut value*) jest środek przedziału. Jeżeli środek przedziału dla zmiennej opisującej obiekt jest mniejszy lub równy wartości punktu podziału, to obiekt przydzielamy do klasy C_i^1. W przeciwnym przypadku przydzielamy go do klasy C_i^2.

2) dla zmiennych w postaci list kategorii lub dla zmiennych w postaci list kategorii z wagami punktem podziału jest prawdopodobieństwo (waga) związane z daną kategorią. W metodzie przyjęto, że punktem podziału jest wartość prawdopodobieństwa równa ½. Jeżeli kategoria danej zmiennej ma przypisaną wagę większą niż ½, to obiekt przydzielamy do klasy C_i^1. W przeciwnym przypadku przydzielamy go do klasy C_i^2.

Do wyboru optymalnego podziału w każdym kroku algorytmu wykorzystywana jest funkcja-kryterium w postaci:

$$\Delta(C_i) := I(C_i) - I(C_i^1) - I(C_i^2), \tag{1}$$

gdzie: odpowiednie elementy oblicza się ze wzoru [Chavent, Stephan 1999, s. 115]:

$$I(C_i) := \frac{1}{2\mu_i} \cdot \sum_{k \in C_i} \sum_{l \in C_i} \omega_k \omega_l \cdot d_{kl}^2, \tag{2}$$

gdzie: $I(C_i)$ oznacza oceniany podział na podstawie i-tej zmiennej, μ_i to suma wag w postaci $\mu_i := \sum_{k \in C_i} \omega_k$, ω_k, ω_l to wagi dla obiektów k i l, d_{kl} to miara odległości pomiędzy obiektami k i l, które należą do klasy C_i.

Do obliczania odległości pomiędzy obiektami w prezentowanym algorytmie nie są stosowane miary odległości dla obiektów symbolicznych, lecz odpowiednio miary odległości dla obiektów opisanych zmiennymi w postaci przedziału liczbowego czy też „mocnych skal" pomiaru i miary odległości dla obiektów opisanych zmiennymi w postaci listy kategorii czy listy kategorii z wagami. Przy czym, jak wspomniano we wprowadzeniu, algorytm nie zakłada możliwości klasyfikacji pomiaru obiektów, w których znajdują się zmienne dowolnego typu. Dla zmiennych w postaci przedziału liczbowego, która przybiera postać (α_k^j, β_k^j) dla obiektu k oraz (α_l^j, β_l^j) dla obiektu l, wykorzystuje się miarę odległości w postaci:

$$d(k,l)=\sqrt{\sum_{j=1}^{p}\max\{\left|\alpha_k^j-\alpha_l^j\right|,\left|\beta_k^j-\beta_l^j\right|\}^2}. \tag{3}$$

Dla zmiennych w postaci listy kategorii czy listy kategorii z wagami należy wpierw zbudować macierz częstości, aby na jej podstawie obliczyć miarę odległości w postaci (por. [Bock, Diday 2000, s. 303]):

$$d^2(k,l)=\sum_{j=1}^{t}\frac{p_{..}}{p_{\cdot j}}\left(\frac{p_{kj}}{p_{k.}}-\frac{p_{lj}}{p_{l.}}\right)^2, \tag{4}$$

gdzie: odpowiednie elementy tej funkcji obliczamy na podstawie tablicy częstości:

$$p_{kj}:=\frac{f_{kj}}{np}, \tag{5}$$

$$p_{k.}:=\sum_{j=1}^{t}p_{kj}, \tag{6}$$

$$p_{\cdot j}:=\sum_{k=1}^{n}p_{kj}, \tag{7}$$

$$p_{..}:=\sum_{k=1}^{n}\sum_{j=1}^{t}p_{kj}=1. \tag{8}$$

Sam algorytm klasyfikacji podziałowej opartej na kryteriach prezentuje rys. 2 (por. [Bock, Diday 2000, s. 299-311]).

4. Modyfikacja algorytmu klasyfikacji podziałowej opartej na kryteriach

Proponowana modyfikacja algorytmu klasyfikacji podziałowej opartej na kryteriach pozwala na klasyfikowanie obiektów, które zawierają zmienne dowolnego typu. Wyjątkiem jest w tym przypadku brak możliwości klasyfikacji zmiennych zawierających wagi.

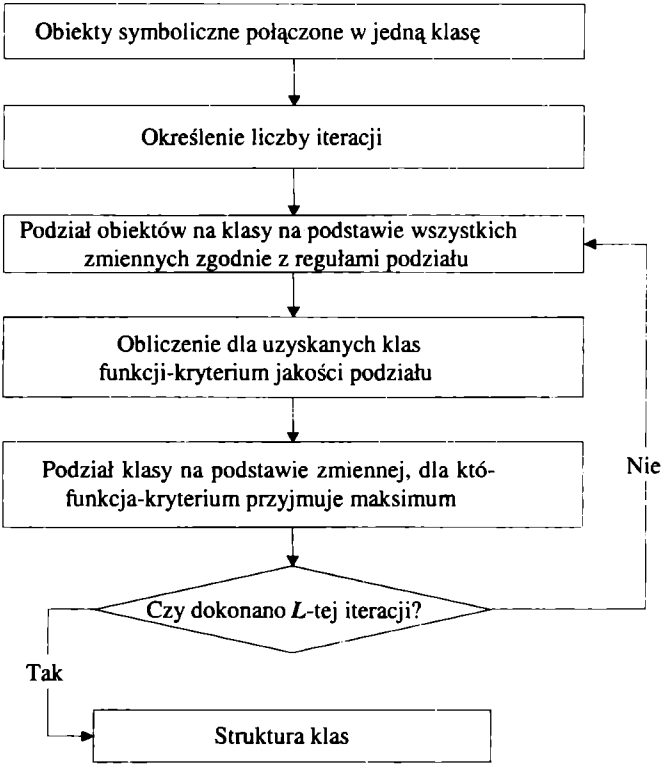
Aby możliwe było klasyfikowanie dowolnych obiektów symbolicznych do funkcji kryterium, zamiast przedstawionych miar odległości proponuje się zastosować normalizowaną miarę Ichino-Yaguchiego. Znormalizowana miara Ichino-Yaguchiego wyraża się wzorem (zob. [Bock, Diday 2000, s. 172]):

$$d_q(O_1,O_2)=\left(\sqrt[q]{\sum_{i=1}^{p}\psi(A_i,B_i)^q}\right), \tag{9}$$

gdzie: $\psi(A,B)=\dfrac{\varphi(A,B)}{|Y|}$ to miara niepodobieństwa dla zmiennych,

natomiast $\phi(A,B)$ wyraża się wzorem:

$$\phi(A,B) = |A \oplus B| - |A \otimes B| + \gamma(2.|A \oplus B| - |A| - |B|). \quad (10)$$



Rys. 2. Algorytm klasyfikacji podziałowej opartej na kryteriach
Źródło: opracowanie własne.

Miara odległości Ichino-Yaguchiego jest oparta na pojęciach operatorów połączenia (*Cartesian join*) i przekroju (*Cartesian meet*), będącymi rozszerzeniami operatorów $\cup$ i $\cap$ na wszystkie typy danych przechowywanych w obiektach symbolicznych (por. [Dudek 2004]).

Tak zmodyfikowane kryterium oceny jakości dokonanego podziału pozwoli na klasyfikację obiektów zawierających zmienne dowolnego typu, poza zmiennymi zawierającymi wagi. Takie zmienne należałoby zastąpić zmiennymi bez wag, co wiąże się z pewną utratą informacji. Sama procedura algorytmu pozostałaby niezmieniona.

5. Przykład empiryczny

W przykładzie empirycznym wykorzystane zostały wyniki badania ankietowego, dotyczącego preferencji użytkowników samochodów osobowych w razie zmia-

ny samochodu. Badanie przeprowadzano od lipca do sierpnia 2004 r. Uczestniczyło w nim 267 respondentów. Respondenci zostali dobrani do badania techniką nielosową – dobór przypadkowy (por. [Szreder 2004, s. 48-53]).

Ostatecznie do analizy zostało przyjętych 160 prawidłowo wypełnionych ankiet. Na ich podstawie utworzono 25 obiektów symbolicznych drugiego rzędu (powstałych na podstawie ankiet respondentów, którzy aktualnie posiadają samochód danej marki). Obiekty te są opisywane przez 8 zmiennych:

1) preferowane modele danej marki (X_1) – taksonomiczna,
2) wybrana pojemność silnika (X_2) – przedział liczbowy,
3) wybrany wiek używanego samochodu (X_3) – przedział liczbowy,
4) preferowany typ nadwozia (X_4) – lista kategorii,
5) kwota, jaką przeznaczyłby ankietowany na zakup auta (X_5) – przedział liczbowy,
6) paliwo, jakim powinien być zasilany wybrany samochód (X_6) – lista kategorii,
7) wybrane wyposażenie dodatkowe (X_7) – lista kategorii,
8) czynniki, którymi kieruje się ankietowany przy wyborze nowego samochodu (X_8) – lista kategorii.

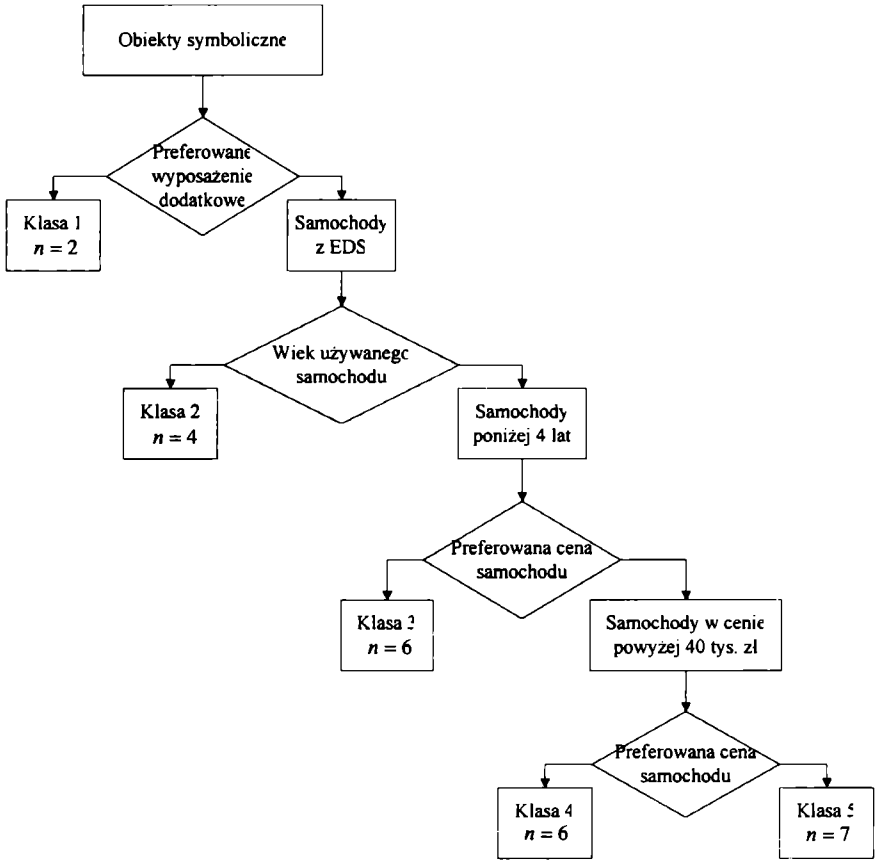
Jeżeli obiekty symboliczne opisywane przez zmienne tego typu miałyby być poddane klasyfikacji z użyciem metody niezmodyfikowanej, to można:

1) usunąć z analizy zmienne jednego z typów, w tym przypadku zmiennych w postaci przedziałów liczbowych, ponieważ jest ich mniej. Z analizy należałoby również wykluczyć zmienną taksonomiczną. Jednakże usuwanie zmiennych opisujących obiekty powoduje znaczną utratę informacji;

2) dokonać odrębnej klasyfikacji dla obiektów opisywanych przez każdy z typów zmiennych odrębnie i porównać wyniki. W tym przypadku należałoby usunąć z analizy zmienną taksonomiczną. Jeżeli wyniki klasyfikacji są w miarę zgodne, to problem można uznać za rozwiązany;

3) na podstawie danych pierwotnych należałoby utworzyć obiekty symboliczne opisywane wyłącznie przez zmienne zawierające wagi. W tym przypadku należałoby także usunąć zmienną taksonomiczną. Problemem może być sytuacja, w której nie mamy dostępu do danych pierwotnych.

W tym przypadku autor na podstawie danych pierwotnych utworzył obiekty symboliczne opisywane wyłącznie przez zmienne zawierające wagi – w tym wypadku wagi odpowiadają częstościom występowania danej kategorii zmiennej. Następnie te obiekty poddano klasyfikacji. Później zbadano jakość klasyfikacji dla wszystkich możliwych podziałów za pomocą wskaźnika *silhouette*; dla struktury trzech klas wartości wskaźnika miały podobny poziom. Ostatecznie przyjęto podział zbioru obiektów na poziomie pięciu klas. Następnie w celu porównania metody niezmodyfikowanej z metodą zmodyfikowaną dokonano klasyfikacji zbioru obiektów opisywanych przez osiem zmiennych. Dla otrzymanych wyników dokonano oceny jakości klasyfikacji na poziomie trzech oraz pięciu klas. W przypadku metody zmodyfikowanej dla zmiennej taksonomicznej przyjęto, że podział będzie dokonywany z *a priori* przyjętą systematyką, według której zmienna przyjmowała realizacje.

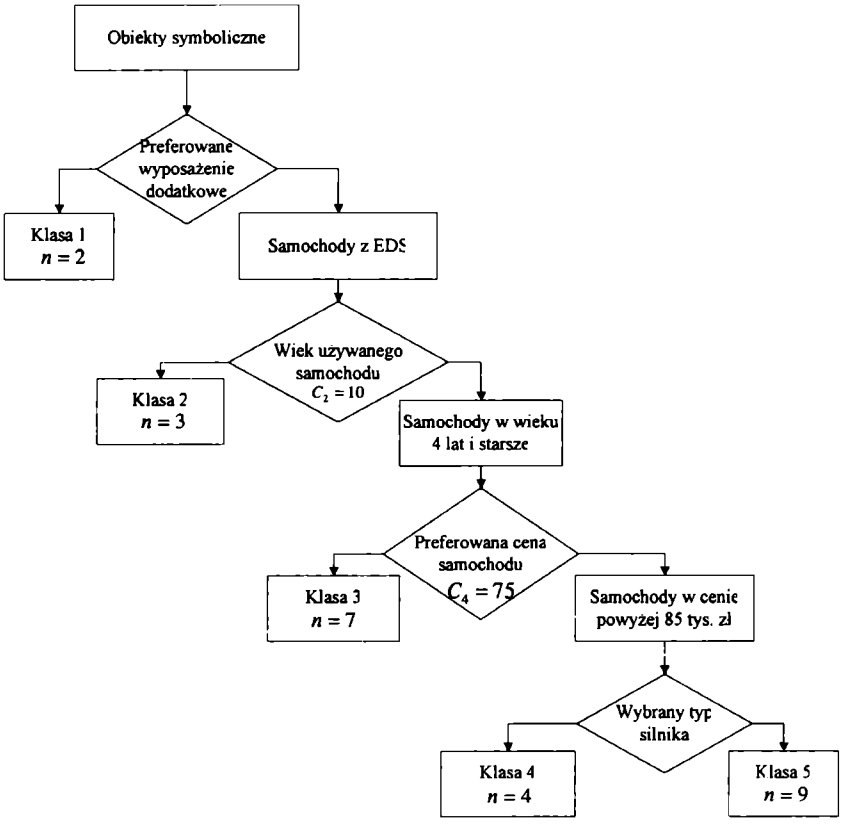


Rys. 3. Wyniki klasyfikacji metodą niezmodyfikowaną
Źródło: opracowanie własne.

W przypadku metody niezmodyfikowanej z początkowego zbioru obiektu należało odrzucić zmienne X_1 – preferowane modele danej marki. Z kolei zmienne X_2 – wybrana pojemność silnika, X_3 – wybrany wiek używanego samochodu, X_5 – kwota, którą ankietowany przeznaczyłby na zakup nowego auta, musiały zostać przekształcone w zmienne w postaci list kategorii z wagami. Usunięcie zmiennych powoduje, po pierwsze, utratę znacznej części informacji o obiektach, a co za tym idzie – wpływa na jakość klasyfikacji. Po drugie, usunięcie części zmiennych powoduje, że zbiór obiektów nie zawsze może być podzielony na *n* klas (gdzie *n* to liczba obiektów w zbiorze początkowym). Przypadek taki będzie miał miejsce, jeżeli:

a) w przypadku zmiennych w postaci list kategorii czy list kategorii z wagami – zmienne mają mało kategorii;

b) w przypadku przedziałów liczbowych – w badaniu pozostawiono niewiele przedziałów liczbowych.



Rys. 4. Wyniki klasyfikacji metodą zmodyfikowaną
Źródło: opracowanie własne.

W wypadku metody zmodyfikowanej nie ma potrzeby usuwania zmiennych, co oznacza, że wykorzystuje się całość informacji o obiektach. Po drugie, wyniki otrzymane przy zastosowaniu metody zmodyfikowanej są „bardziej intuicyjne" niż w przypadku metody niezmodyfikowanej.

Dla pięciu klas w przypadku metody niezmodyfikowanej wartość wskaźnika *silhouette* wyniosła $S = 0{,}54$, a w przypadku metody zmodyfikowanej ten sam wskaźnik wyniósł $S = 0{,}71$.

6. Wnioski

Zaprezentowana modyfikacja algorytmu klasyfikacji podziałowej opartej na kryteriach pozwala na klasyfikowanie obiektów zawierających zmienne różnego typu, z wyjątkiem zmiennych zawierających wagi lub prawdopodobieństwa. Mo-

dyfikacja taka pozwala na zastosowanie tej metody w przypadkach, gdy obiekty opisywane są przez zmienne różnych typów.

Wykorzystywana w metodzie zmodyfikowanej znormalizowana miara odległości Ichino-Yaguchiego pozwala na weryfikację wyników klasyfikacji za pomocą innych metod klasyfikacji hierarchicznej oraz za pomocą wskaźników oceny jakości klasyfikacji, które oparte są na macierzy odległości.

Niezmodyfikowana metoda klasyfikacji oparta na kryteriach wymaga znacznie większej ilości informacji niż metoda zmodyfikowana, przez co jest bardziej czasochłonna i pracochłonna. Gdy nie dysponujemy danymi pierwotnymi, metoda niezmodyfikowana nie pozwala na klasyfikowanie obiektów opisywanych przez zmienne różnych typów łącznie.

Odrębnym zagadnieniem, które wymaga dalszych prac, jest opracowanie reguł podziału dla zmiennych strukturalnych. W przypadku zmiennych taksonomicznych autor sugeruje dokonywanie podziału zgodnie z przyjętą systematyką danej zmiennej.

Uzyskane w przykładzie empirycznym wyniki klasyfikacji pozwalają stwierdzić, że metoda zmodyfikowana jest, po pierwsze, bardziej efektywna (pozwala skrócić proces klasyfikacji oraz badać obiekty opisywane przez zmienne różnych typów), a po drugie – uzyskane wyniki klasyfikacji mają lepszą jakość (ocenianą za pomocą miernika *silhouette*).

Literatura

Bock H.-H., Diday E. (red.) (2000), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg.

Chavent M. (1998), *A Monothetic Clustering Method*, „Pattern Recognition Letters" vol. 19, s. 989-996.

Chavent M., Guinot C., Lechevallier Y., Tenenhaus M. (1999), *Méthodes divisives de classification et segmentation non supervisee: recherchere d'une typologie de la peau humaine saine*, „Reveue de Statisiques Appliquées" XLVII (4), s. 87-99.

Chavent M., Stephan V. (1999), *From Generalization to Clustering in Relational Data Base Context*, [w:] *Studies and Research: Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis* (KESD '98), Office for Official Publications of the European Communities, Luxemburg, s. 105-117.

Chavent M. (2004), *An Hausdorff Distance between Hyper-rectangles for Clustering Interval Data*, [w:] D. Banks, L. House, R. McMorris, P. Arabie, W. Gaul, (red.), *Classification, Clustering and Data Mining Applications*, Springer-Verlag, New York, s. 333-340.

Dudek A. (2004), *Miary podobieństwa obiektów symbolicznych. Odległość Ichino-Yaguchiego*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Ekonometria 15, (w druku).
Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa.
Hair J.E., Anderson R.E., Tatham R.L., Black W.C. (1998), *Multivariate Data Analysis*, Prentice Hall, New Jersey.
Ichino M., Yaguchi H. (1994), *Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis*, „IEEE Transactions on Systems, Man, and Cybernetics" vol. 24, no. 4, s. 698-707.
Malerba D., Espozito F., Giovalle V., Tamma V. (2001), *Comparing Dissimilarity Measures for Symbolic Data Analysis,* materiały konferencyjne „New Techniques and Technologies for Statistcs" and „Exchange of Technology and Know-how" (ETK-NTTS'01), s. 473-481.
Malerba D., Esposito F., Monopoli M. (2002), *Comparing Dissimilarity Measures for Probabilistic Symbolic Objects*, [w:] A. Zanasi, C.A. Brebbia, N.E.F. Ebecken, P. Meli (Eds.), WIT Press Southampton UK, „Data Mining III, Series Management Information Systems" vol. 6, s. 31-40.
Walesiak M. (1996), *Metody analizy danych marketingowych*, PWN, Warszawa.

MODIFICATION OF CRITERION-BASED DIVISIVE CLUSTERING METHOD

Summary

The aim of this article is to present one of symbolic clustering methods, the criterion-based divisive clustering method, besides that article presents a modification of this method that will allow to clustering symbolic objects with different variable types. The article will compare in the empirical part the clustering of a car market with modified and unmodified method.