PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1100 — 2006
Ekonometria 16

Marcin Pełka

KLASYFIKACJA OBIEKTÓW SYMBOLICZNYCH OPARTA NA KRYTERIACH

1. Wstęp

Dokonanie klasyfikacji, czyli podziału na jednorodne podzbiory, danego zbioru obiektów jest od dawna ważnym elementem badań ekonomicznych i znajduje się w centrum zainteresowania wielu różnych prac naukowych.

Celem niniejszego artykułu jest zaprezentowanie idei klasyfikacji obiektów symbolicznych opartej na kryteriach[1] (DIV); ideę tę prezentuje Diday [2], dotyczy ona klasyfikacji obiektów symbolicznych, w których mamy do czynienia ze zmiennymi mierzonymi na „silnych" skalach pomiarowych, czyli przedziałowymi lub ilorazowymi, lub, gdy mamy do czynienia ze zmiennymi mierzonymi na „słabych" skalach pomiarowych, porządkowych czy też nominalnych. W części empirycznej przedstawiona zostanie metoda zastosowana do zaklasyfikowania samochodów osobowych ze względu na komfort jazdy.

2. Obiekty i zmienne symboliczne

Zanim zostanie przedstawiona idea klasyfikacji opartej na kryteriach, konieczne jest krótkie przybliżenie samych obiektów symbolicznych wraz z rodzajami zmiennych, jakie mogą opisywać obiekty symboliczne.

Według E. Gatnara [5, s. 62] obiekt symboliczny opisywany jest przez koniunkcję wartości poszczególnych jego cech, przykładowo obiekt symboliczny w postaci:

[marka pojazdu = Škoda Fabia]∧[pojemność silnika = 1800 cm^3]∧[komfort jazdy = = średni]

[1] We wspomnianej publikacji metodę tę nazywa się *Criterion-Based Divisive Clustering* (DIV).

obrazuje samochód marki Škoda Fabia o pojemności silnika 1800 cm^3, którego komfort jazdy został oceniony jako średni, przeciętny. W ramach symbolicznych metod analizy danych taką koniunkcję wartości cech obiektów nazywamy kompleksami (por. [5, s. 67-69]).

W obiekcie symbolicznym oprócz zmiennych w postaci tradycyjnej, tj. zmiennych mierzonych na skalach ilorazowych, przedziałowych, porządkowych czy nominalnych, mogą występować zmienne charakterystyczne dla metod symbolicznych. Najważniejsze typy tych zmiennych przedstawiono w tab. 1.

Tabela 1. Zmienne mogące opisywać obiekt symboliczny

Nazwa zmiennych	Przykład
Liczbowe (nazywane numerycznymi)	wzrost = 175
Kategorie (jakościowe)	kolor = zielony
Przedziały wartości	zarobki = (1560, ..., 2500)
Lista wartości	wybrane programy = {TVP1, TVN, TVP2}
Lista wartości z wagami (prawdopodobieństwami)	miejsce zamieszkania ={miasto 65%, wieś 35%}
Zmienne strukturalne	• *zależność logiczna (funkcyjna)* dla zmiennej ustalono *a priori* warunki logiczne lub funkcyjne decydujące, o tym, jakie wartości przyjmie dana zmienna, • *zależność taksonomiczna* – „miejsce zamieszkania" dla zmiennej ustalono *a priori* systematykę, wg której przyjmuje ona realizacje, • *zależność hierarchiczna* zdefiniowane są warunki, od których zależy czy zmienna dotyczy danego obiektu, czy nie.

Źródło: opracowanie własne na podstawie [2, s. 2; 5].

3. Klasyfikacja podziałowa oparta na kryteriach

Sama idea klasyfikacji podziałowej opartej na kryteriach wiąże się ściśle w swych początkach z algorytmem CART [2, s. 298]), lecz dotyczy obiektów i zmiennych symbolicznych, w związku z czym ma pewne modyfikacje i różnice.

Początkowo w metodzie tej rozważa się zbiór $\Omega = \{1, 2, \ldots, n\}$ obiektów symbolicznych, które z kolei są scharakteryzowane p zmiennymi symbolicznymi, tj. $Y_1, Y_2, \ldots, Y_p$.

$Y_j: \Omega \rightarrow B_j$, a także $k \rightarrow Y_j(k)$. Zakres B_j zależy od dziedziny $\mathcal{Y}_j$ oraz typu zmiennej, z jaką mamy do czynienia.

1. Dla dziedziny $\mathcal{Y}_j \in \mathcal{R}$ należy rozważyć dwa typy zmiennych, z jakimi możemy mieć do czynienia:

• dla tradycyjnej zmiennej liczbowej (ilorazowej lub przedziałowej), gdzie $Y_j(k) \in \mathcal{R}$, wówczas $B_j \in \mathcal{R}$.

• dla zmiennej symbolicznej w postaci przedziału liczbowego, gdzie $Y_j(k) = (\alpha, \beta) \subset \mathcal{R}$, wówczas B_j jest zbiorem wszystkich zbiorów nieotwartych zawierających się w $\mathcal{R}$.

2. Dla dziedziny $\mathcal{Y}_j = \{a, b, c, \ldots, h\}$ która jest skończonym zbiorem zmiennych porządkowych, należy rozważyć takie przypadki, jak:

• tradycyjna zmienna porządkowa, gdzie $Y_j(k) \in \mathcal{Y}_j$, wówczas $B_j = \mathcal{Y}_j$,

• przypadek zmiennej w postaci listy wartości, gdy $Y_j(k) \subset \mathcal{Y}_j$, wówczas $B_j = P(\mathcal{Y}_j)$,

• lista wartości z wagami czy prawdopodobieństwami, częstościami, gdy mamy $Y_j(k) = \pi_k$, które jest właśnie częstością, prawdopodobieństwem wystąpienia danej wartości w $\mathcal{Y}_j$. Wówczas B_j jest zbiorem $M(\mathcal{Y}_j)$ wszystkich prawdopodobieństw wystąpienia wartości w Y_j [2, s. 298-300].

Każdy z obiektów $k \in \Omega$ jest opisany przez wektor zmiennych symbolicznych w postaci $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, co zapisujemy jako $\mathbf{Y}(k) = (Y_1(k), Y_2(k), \ldots, Y_p(k))'$; zapis ten jest równoważny zapisowi $\xi_k = (\xi_{k1}, \xi_{k2}, \ldots, \xi_{kp})'$. Można to zapisać także w postaci macierzy symbolicznej

$$\mathbf{X} = \begin{bmatrix} \xi_1' \\ \vdots \\ \xi_2' \end{bmatrix}_{n \times p} \tag{1}$$

Istotną wadą, na którą zwraca uwagę Diday w swej pracy, jest brak możliwości klasyfikowania obiektów symbolicznych opisywanych przez zmienne różnych typów. Możliwe jest klasyfikowanie tylko zmiennych w postaci przedziału liczbowego lub tradycyjnych zmiennych przedziałowych lub ilorazowych łącznie. Drugą możliwością jest klasyfikacja zmiennych porządkowych, list wartości, list wartości z wagami oraz zmiennych nominalnych [2, s. 300].

W celu dokonania dalszej klasyfikacji obiektów konieczne jest wyliczenie odległości między obiektami symbolicznymi, przy czym twórcy tej metody nie posługują się miarami odległości, jakie są możliwe do zastosowania w przypadku obiektów symbolicznych[2].

[2] W większości metod analizy danych symbolicznych wykorzystywane są miary odległości: Gowdy-Didaya, Ichino-Yaguchiego oraz de Carvalho.

1. Dla zmiennych liczbowych, przedziałowych lub ilorazowych oraz zmiennych w postaci przedziału liczbowego

Niech δ_j będzie miarą odległości dla B_j taką, że δ_j: $B_j \times B_j \rightarrow \mathcal{R}^+$, co oznacza, że możemy to zapisać jako $(\xi_{kj}, \xi_{lj}) \rightarrow \delta_j(\xi_{kj}, \xi_{lj})$. Jeżeli mamy do czynienia z takimi ξ_{kj}, ξ_{lj}, że są to przedziały liczbowe w postaci (α_k^j, β_k^j) oraz (α_l^j, β_l^j) używana jest odległość Hausdorfa, dla szczególnego przypadku przedziałów, w postaci [2, s. 301]:

$$\delta_j(\xi_{kj}, \xi_{lj}) = \max\left\{\left|\alpha_k^j - \alpha_l^j\right|, \left|\beta_k^j - \beta_l^j\right|\right\}. \tag{2}$$

Dla tradycyjnego przypadku, gdy mamy do czynienia z wartościami pojedynczymi, odległość ta sprowadza się do różnic absolutnych pomiędzy wartościami zmiennych.

Odległości między zmiennymi, zależne od typu zmiennej, z jaką mamy do czynienia, są następnie elementami następującej miary odległości [2, s. 302]:

$$d: \Omega \times \Omega \rightarrow \mathcal{R}^+,$$

$$d(k,l) = \left(\sum_{j=1}^{p} \delta_j(\xi_{kj}, \xi_{lj})^2\right)^{1/2} \tag{4}$$

Wykorzystując wcześniej wskazaną funkcję odległości Hausdorfa, możemy zapisać powyższy wzór jako:

$$d: \Omega \times \Omega \rightarrow \mathcal{R}^+,$$

$$d(k,l) = \left(\sum_{j=1}^{p} \max\left\{\left|\alpha_k^j - \alpha_l^j\right|, \left|\beta_k^j - \beta_l^j\right|\right\}^2\right)^{1/2} \tag{5}$$

Szczególnym przypadkiem będzie, gdy oba obiekty opisują zmienne przedziałowe lub interwałowe, wówczas ta miara odległości jest równoważna odległości euklidesowej dla tych zmiennych.

2. Dla zmiennych w postaci list wartości, list wartości z wagami

Przyjmuje się, że mamy do czynienia z przypadkiem, gdy wszystkie zmienne to listy wartości lub listy wartości z wagami, takie założenie jest prawdziwe, ponieważ podczas tworzenia obiektów symbolicznych rzadko spotykaną sytuacją jest występowanie zmiennych w postaci zmiennej porządkowej czy nominalnej. W takim przypadku mamy do czynienia z listą wartości dla zmiennych Y_1, Y_2, ..., Y_p

wraz z ich dziedzinami $\mathcal{Y}_1$, $\mathcal{Y}_2$, ..., $\mathcal{Y}_p$. W takim przypadku $Y_j(k)$ jest zbiorem kategorii występujących w $\mathcal{Y}_j$ lub też jest zbiorem częstości, prawdopodobieństw.

Tabelę zawierającą obiekty symboliczne wraz ze zmiennymi w postaci list wartości z wagami można łatwo przekształcić w macierz częstości, gdzie $t = |y_1| + \ldots + |y_p|$ oznacza całkowitą liczbę kategorii, a $j = 1, 2, \ldots, t$ to indeks dla poszczególnych kategorii w macierzy częstości. Przykładową macierz częstości pokazano w tab. 2.

Tabela 2. Budowa macierzy częstości

	1	...	j	...	t	Σ
1	f_{11}	...	f_{1j}	...	f_{1t}	p
⋮	⋮		⋮		⋮	⋮
k	f_{k1}	...	f_{kj}	...	f_{kt}	p
⋮	⋮		⋮		⋮	⋮
n	f_{n1}	...	f_{nj}	...	f_{nt}	p
Σ	$f_{\cdot 1}$	...	$f_{\cdot j}$	...	$f_{\cdot t}$	np

Źródło: [2, s. 303].

Dwa dowolne obiekty k i l możemy porównać, wykorzystując funkcję odległości Φ^2 w postaci [2, s. 302]:

$$d^2(k,l) = \sum_{j=1}^{t} \frac{p_{\cdot\cdot}}{p_{\cdot j}} \left(\frac{p_{kj}}{p_{k\cdot}} - \frac{p_{lj}}{p_{l\cdot}} \right)^2 . \qquad (6)$$

Odpowiednie elementy tej funkcji obliczamy na podstawie tablicy częstości:

$$p_{kj} := \frac{f_{kj}}{np}, \qquad (7)$$

$$p_{k\cdot} := \sum_{j=1}^{t} p_{kj}, \qquad (8)$$

$$p_{\cdot j} := \sum_{k=1}^{n} p_{kj}, \qquad (9)$$

$$p_{\cdot\cdot} := \sum_{k=1}^{n} \sum_{j=1}^{t} p_{kj} = 1. \qquad (10)$$

Kryterium wyboru podziału zbioru na klasy w tym przypadku jest wariancja wewnątrz klasy. Dla macierzy odległości $\mathbf{D}=(d_{kl})$ między obiektami $\Omega=(1,2,\ldots,n)$, z których każdy jest ważony wagą w postaci $\omega_k \geq 0$ ($k = 1, 2, \ldots, n$), możemy zapisać, że przybliżeniem wariancji wewnątrzklasowej jest [2, s. 304]:

$$I(C_i) := \frac{1}{2\mu_i} \cdot \sum_{k \in C_i} \sum_{l \in C_i} \omega_k \omega_l \cdot d_{kl}^2 . \qquad (11)$$

We wzorze tym wagi mają następującą własność: $\mu_i := \sum_{k \in C_i} \omega_k$.

Szczególnym przypadkiem będzie, gdy mamy do czynienia w obiekcie symbolicznym ze zmiennymi ilorazowymi czy przedziałowymi. Wówczas powyższy wzór można zapisać dla $d_{kl} = \|x_k - x_l\|$ oraz $\omega_k \equiv 1$ jako [2, s. 304]:

$$I(C_i) = \frac{1}{n \times n_i} \cdot \sum_{k,\, l \in C_i} \sum_{l>k} d_{kl}^2 . \tag{12}$$

Warto tu zaznaczyć, że $n_i := |C_i|$ to liczebność obiektów w danej klasie.

Uogólnieniem kryterium wariancji dla podziału na m klas w postaci $C = (C_1, C_2, \ldots, C_m)$ dla danej macierzy odległości jest miara

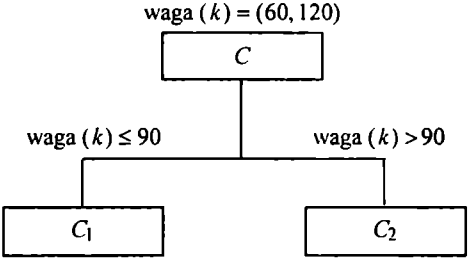
$$W(C) := \sum_{i=1}^{m} I(C_i), \tag{13}$$

w której kryterium jest jedna z wymienionych miar.

Ważnym elementem, z którym mamy do czynienia przy klasyfikacji, jest wybór czy też dobór sposobu podziału na klasy. W tym przypadku poszukujemy takiego binarnego podziału, aby kryterium wariancji wewnątrzklasowej było jak najmniejsze.

W przypadku, gdy mamy do czynienia z obiektem symbolicznym, wybór „punktu podziału" (*cut value*) jest zależny od typu zmiennej, z jaką mamy do czynienia.

1. Dla zmiennych w postaci przedziału, gdzie $Y_j(k) = (\alpha, \beta)$ punktem podziału jest środek tego przedziału. W sposób formalny zapisujemy to jako:
$q_c(k) =$ prawda dla $m_k \le c$,
$q_c(k) =$ fałsz dla $m_k > c$.
Wartość $m_k = \frac{\alpha + \beta}{2}$ jest tutaj środkiem przedziału.



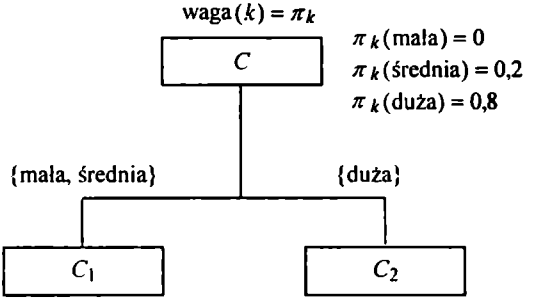
Rys. 1. Punkt podziału dla zmiennej w postaci przedziału liczbowego

Źródło: opracowanie własne.

2. Dla zmiennych w postaci listy wartości lub listy wartości z wagami, punktem podziału jest prawdopodobieństwo (wagi). Zmienne w postaci listy wartości można zapisać w postaci listy wartości z wagami, korzystając z początkowego zbioru obserwacji. Wówczas punkt podziału w sposób formalny definiujemy jako:

$q_c(k) =$ prawda dla $\sum_{x \le c} \pi_k(x) \ge 1/2$,

$q_c(k) =$ fałsz dla $\sum_{x \le c} \pi_k(x) < 1/2$.



Rys. 2. Punkt podziału dla zmiennej w postaci listy wartości z wagami

Źródło: opracowanie własne.

Ważnym elementem, z którym zetknie się każdy użytkownik tej metody, będzie wybór klasy (grupy), którą należy podzielić na dwie nowe klasy. Przyjmuje się, że na samym początku istnieje jedna klasa zawierająca wszystkie obiekty: $P_m = (C_1, C_2, \ldots, C_m)$. Należy wybrać klasę C_i, której podział spowoduje zwiększenie liczby klas o jeden, a otrzymamy dwie nowe klasy C_i^1, C_i^2.

Problemem jest wybór takiej klasy, której podział spowoduje polepszenie kryterium minimum wariancji wewnątrzklasowej.

Kryterium to można sprowadzić do zapisu: $\Delta(C_i) := I(C_i) - I(C_i^1) - I(C_i^2)$, kryterium to należy maksymalizować, wówczas wariancja wewnątrzklasowa osiąga minimum wartości [2, s. 306].

Ostatnim ważnym elementem jest zatrzymanie, zakończenie podziału początkowego zbioru klas. Podział kończony jest po dokonaniu L iteracji, których liczbę podaje prowadzący badanie. Wiadomo, że największą liczbą iteracji będzie $L = n$, kiedy to otrzymamy pojedyncze obiekty. Dlatego należy wybrać takie L, żeby $L < n$.

Sam algorytm można by zapisać jako:

1. Połącz wszystkie obiekty w jedną klasę, ważne jest, by zmienne w ramach obiektu miały jednolity charakter:

• zmienne w postaci przedziału liczbowego, a także zmienne mierzone na skalach ilorazowej i przedziałowej,

• zmienne w postaci listy wartości czy listy wartości z wagami oraz zmienne mierzone na skali porządkowej i nominalnej.

2. Dla zmiennych, zgodnie z ich rodzajem, oblicz macierz odległości.

3. Dokonaj podziału na podstawie wszystkich zmiennych zgodnie z regułą, która dotyczy danego typu zmiennych.

4. Dokonaj podziału na klasy na podstawie wszystkich zmiennych zgodnie z regułą, która dotyczy danego typu zmiennej.

5. Dla każdego podziału zbadaj wartość $\Delta(C_i) := I(C_i) - I\left(C_i^1\right) - I\left(C_i^2\right)$.

6. Wybierz taki podział, dla którego $\Delta(C_i)$ wynosi maksimum, na jego podstawie i w oparciu o kryteria podziału, związane z punktem przecięcia, dokonaj podziału obiektów symbolicznych pomiędzy otrzymane klasy. Dla otrzymanych klas zbuduj obiekt symboliczny, który będzie je reprezentować.

7. Powtarzaj kroki 4-6 tak długo, aż dokonasz L iteracji.

4. Klasyfikacja samochodów osobowych pod względem komfortu jazdy

W części empirycznej wykorzystano przedstawioną metodę klasyfikacji do przeanalizowania samochodów osobowych z punktu widzenia komfortu jazdy. Wykorzystano w tym celu dane forum internetowego gazety „Auto Świat"[3], a także informacje gazety „Auto Świat – Testy 2004". Na tej podstawie zbudowano ankietę, którą następnie wypełniło 200 osób oceniających 12 marek samochodów. Do analizy wybrano następujące zmienne:

1) wygoda siedzeń w samochodzie – siedzenia przednie, także ilość miejsca dla pasażera oraz kierowcy – X_1,

2) wygoda siedzeń w samochodzie – siedzenia tylne, także ilość miejsca dla pasażerów – X_2,

3) wykończenie wnętrza pojazdu – czy poszczególne elementy są rozłożone „intuicyjnie", czy nie hałasują w trakcie jazdy i użytkowania, a także, jaka jest ich jakość wykończenia – X_3,

4) nieprzenikanie hałasu pracy silnika w czasie jazdy – X_4,

5) nieprzenikanie odgłosów jazdy do wnętrza, czyli głośność toczenia pojazdu – X_5,

6) komfort pracy układu amortyzacyjnego, zarówno na prostej drodze, jak i na wybojach – X_6.

[3] autoswiat.redakcja.pl/forum.

Wszystkie zmienne były scharakteryzowane na skali porządkowej w postaci ocen od 1 (stan niepożądany, niewygoda, duży hałas itp.) do 7 (stan najbardziej pożądany, wysoki komfort siedzeń, brak hałasu itp.), przy czym każdy ankietowany oceniał dany element tylko raz. Elementy dotyczące wygody siedzeń, czy to przednich, czy to tylnich będą przedstawione jako listy wartości z wagami, pozostałe zmienne zaś jako listy wartości. Ostatecznie tabelę wybranych obiektów i zmiennych je charakteryzujących prezentuje poniższa tabela danych.

Tabela 3. obiekty symboliczne i zmienne je charakteryzujące

	X1	X2	X3	X4	X5	X6
Skoda Fabia	4 (0.40), 5 (0.40), 6 (0.20)	3 (0.80), 4 (0.20)	3, 4	4, 5	5, 6, 4	4, 5, 6
Skoda Octavia	6 (0.50), 7 (0.50)	4 (1.00)	4, 5	5, 6	5, 6	4, 5
Fiat 126p	2 (0.40), 1 (0.60)	1 (1.00)	1	1	1	2, 1
Renault Megane	4 (0.50), 5 (0.50)	3 (0.25), 4 (0.75)	3, 4	4, 5	5, 4	3, 4, 5
Citroen C3	4 (0.40), 5 (0.60)	3 (0.60), 4 (0.40)	5, 6	5, 6	5, 4	6, 7
Citroen Saxo	4 (0.50), 3 (0.50)	4 (0.50), 5 (0.50)	4, 5, 6	5, 6	5, 6, 4	4, 5, 6
Opel Agila	4 (0.33), 5 (0.67)	3 (0.67), 4 (0.33)	3, 4	4, 5, 3	5, 3	3, 4
Opel Astra	5 (0.67), 6 (0.33)	4 (0.33), 5 (0.67)	5	5	5, 6	6
Peugot 406	5 (0.50), 6 (0.25), 7 (0.25)	4 (0.50), 5 (0.50)	3, 5	4, 5	5, 4	4, 5
Polonez Caro	4 (0.20), 3 (0.80)	3 (0.60), 4 (0.40)	3, 2	3, 2	3, 2	3, 2, 1
Fiat Punto	4 (0.50), 3 (0.50)	3 (0.50), 4 (0.25), 5 (0.25)	4, 5	4, 3	5, 6, 4	4, 5
Daewoo Tico	4 (0.40), 5 (0.20), 3 (0.40)	3 (0.80), 2 (0.20)	3, 4	4, 3	4, 3	3, 4, 2

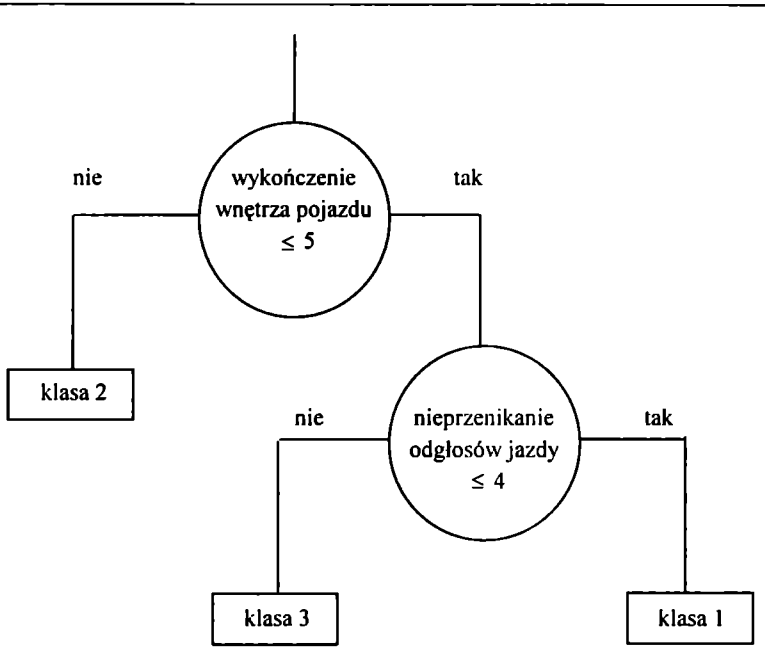
Źródło: opracowanie własne.

Następnie dla tych obiektów w programie SODAS, SODAS wersja 1.20, dokonano klasyfikacji na kolejno 3, 4, 5, 6, 7 klas. Ostatecznie wybrano podział zbioru na 3 klasy, ponieważ wówczas można zinterpretować wyniki jako klasy pojazdów: pojazdy o „niskim, niezadowalającym" komforcie, pojazdy o „średnim, przeciętnym" komforcie oraz pojazdy o „wysokim, dobrym" komforcie jazdy.

W wyniku przeprowadzenia klasyfikacji uzyskano następujący wynik, który ukazano na rys. 3.

Uzyskane punkty przecięcia mogą początkowo wydawać się mało intuicyjne, ale po dłuższym zastanowieniu okazuje się, że to właśnie wnętrze pojazdu, a następnie hałas podczas jazdy często decydują o wyborze tego czy innego samochodu. Wiadomo, że są to jedne z pierwszych elementów, które podlegają ocenie przez klienta i jeżeli ich ocena wypadnie negatywnie, to klient często nie zdecyduje się na wybór. Klasy przedstawione na rys. 3 to:

klasa 1 – samochody o wysokiej ocenie wnętrza, nie mniejszej niż „5" w skali siedmiostopniowej, dodatkowo ocenione powyżej przeciętnej ze względu na ciszę



Rys. 3. Wyniki klasyfikacji pojazdów pod względem komfortu, gdy przyjęto trzy klasy

Źródło: opracowanie własne na podstawie wyników.

podczas jazdy, nie mniej niż „4" w 7-stopniowej skali; są to samochody: Škoda Fabia, Škoda Octavia, Renault Mégane, Citroën C3, Citroën Saxo, Opel Agila, Opel Astra, Peugeot 406, Fiat Punto;

klasa 3 – jest to klasa jednoelementowa, stanowi ją samochód Polonez Caro, oceniony nieco powyżej przeciętnej za wnętrze, lecz poniżej „4" w 7-stopniowej skali za hałas wewnątrz pojazdu w czasie jazdy;

klasa 2 – uzyskana po pierwszej iteracji, podziale, jest to również klasa jednoelementowa, stanowi ją samochód Fiat 126p, oceniony najniżej we wszystkich kategoriach przez większość respondentów.

5. Wnioski

Metoda klasyfikacji opartej na kryteriach stanowi ciekawą adaptację tradycyjnej metody CART dla obiektów symbolicznych. Do jej największych wad, na które zwrócili uwagę jej twórcy, zalicza się brak możliwości klasyfikowania obiektów zawierających różnorodne zmienne. Ważne jest zbadanie, czy metoda ta nie może być stosowana dla macierzy odległości, która powstanie na podstawie miar Ichino-Yaguchiego czy de Carvalho, ponieważ umożliwiłoby to klasyfikację bardziej złożonych obiektów symbolicznych.

Uzyskane wyniki sugerują, że ważnymi elementami w ocenie komfortu jazdy samochodem, a tym samym w ocenie jego kupna, są często: jego wnętrze, zarówno pod względem stylistyki, jak i funkcjonalności, a następnie hałas podczas jazdy samochodem, czyli tzw. hałas toczenia. Opisy klas z kolei sugerują, że polskie produkty, tj. Fiat 126p i Polonez nie przystają do nowoczesnych samochodów pod względem komfortu, jednakże mogą z nimi konkurować pod względem ceny.

Literatura

[1] „Auto Świat – Testy 2004", Axel Springer Polska 2004.

[2] Bock H.H., Diday E., *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Heidelberg 2000.

[3] Dudek A., *Tworzenie obiektów symbolicznych z baz danych*, Ekonometria XIV, Prace Naukowe Akademii Ekonomicznej nr 1021, AE, Wrocław 2004.

[4] Dudek A., *Miary podobieństwa obiektów symbolicznych. Odległość Ichino-Yaguchiego*, Ekonometria XIV, Prace Naukowe Akademii Ekonomicznej nr 1021, AE, Wrocław 2004.

[5] Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998.

[6] Hair E., Anderson E., Tatham L., Black C., *Multivariate Data Analysis*, Prentice Hall, Englewood Cliffs 1998.

[7] Wójcik T., *Zarys teorii klasyfikacji*, Wydawnictwo Naukowe PWN, Warszawa 1965.

SYMBOLIC CRITERION-BASED DIVISIVE CLUSTERING

Summary

The article presents symbolic criterion-based divisive clustering proposed by Diday [2]. The aim of the article is to present all the parts of this method and to characterize it. The article presents also suggestions of Polish translations for words used by Diday. The usage possibility was presented on the basis of car marked clustering. At the end of the article some weaknesses and advantages of this idea were presented.

Mgr Marcin Pelka jest pracownikiem Katedry Ekonometrii i Informatyki w Akademii Ekonomicznej we Wrocławiu.