PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 6 (1206) 2008
Ekonometria 21

Danuta Strahl

KLASYFIKACJA POZYCYJNA. PODEJŚCIE DYNAMICZNE

1. Wstęp

Literatura przedmiotu przedstawia całe bogactwo metod klasyfikacji. Obszerne opisy tych metod można znaleźć np. w pracach [1; 3; 4; 7; 8; 9]. Jednak niewiele technik klasyfikacyjnych uwzględnia aspekty dynamiczne. Podejście, w którym brane są pod uwagę zmiany w czasie obiektów podlegających klasyfikacji, uwzględnia np.:

sprowadzenie wartości zmiennych diagnostycznych do porównywalności, np. przez standaryzację, w której zarówno średnią wartość, jak i odchylenie standardowe oblicza się dla całego okresu badania,
- klasyfikację obiektów w kolejnych latach i ocenę zgodności otrzymanych wyników klasyfikacji według wybranego współczynnika zgodności klasyfikacji,
- wykorzystanie miar syntetycznych, podobieństwa taksonomicznego, metod porządkowania liniowego ze stałym i zmiennym obiektem-wzorcem,

dynamiczny dobór zmiennych diagnostycznych (por. [7; 10]).

W pracach [5; 6] zaproponowano technikę klasyfikacji obiektów wykorzystującą statystyki pozycyjne. W tym artykule podejście to zostanie rozszerzone o wybrane aspekty dynamiczne.

Zasadniczym celem artykułu jest propozycja klasyfikacji uwzględniająca dane przekrojowo-czasowe oraz oparta na kryterium wartości statystyk pozycyjnych z całego okresu badania.

2. Podstawy formalne klasyfikacji pozycyjnej

Dany jest zbiór obiektów $P = \{P_1, P_2, \ldots, P_K\}$.

Każdy z obiektów P_k $(k = 1, 2, \ldots, K)$ opisany jest zbiorem zmiennych oznaczonych symbolami:

$$X = \{X_1, X_2, \ldots, X_m\}.$$

Wartości zmiennych obserwujemy na obiektach P_k w zadanych momentach $t = 1, 2, \ldots, T$. Zapis obserwacji można ująć macierzą blokową, gdzie:

$$\mathbf{X}_{kj}^{t} = \begin{bmatrix} x_{11}^{1} & \ldots & x_{1m}^{1} \\ \ldots & x_{kj}^{1} & \ldots \\ x_{K1}^{1} & \ldots & x_{Km}^{1} \\ x_{11}^{t} & \ldots & x_{1m}^{t} \\ \ldots & x_{kj}^{t} & \ldots \\ x_{K1}^{t} & \ldots & x_{Km}^{t} \\ x_{11}^{T} & \ldots & x_{1m}^{T} \\ \ldots & x_{kj}^{T} & \ldots \\ x_{K1}^{T} & \ldots & x_{Km}^{T} \end{bmatrix}_{K \times T \times m}, \tag{1}$$

gdzie: x_{kj}^{t} – wartość j-tej cechy $(j = 1, 2, \ldots, m)$ w k-tym obiekcie badania $(k = 1, 2, \ldots, K)$ w t-tym momencie obserwacji $(t = 1, 2, \ldots, T)$.

3. Klasyfikacja pozycyjna w ujęciu dynamicznym

Klasyfikacja przeprowadzona będzie w następujących etapach.

Etap 1

Dla każdej zmiennej X_j $(j = 1, 2, \ldots, m)$ obliczamy wartość wskazanej statystyki pozycyjnej. Dla ustalenia uwagi niech będzie to mediana. Chcąc uwzględnić zmiany zachodzące w badanym okresie i oddziałujące na wartości cech diagnostycznych, możemy wprowadzić jedno z następujących podejść lub też stosować je łącznie:

1) przeprowadzić standaryzację wartości cech, uwzględniając wpływ czasu, stosując odpowiednie wzory dla obliczenia wartości średniej oraz odchylenia standardowego (por. [7]),

2) wprowadzić do procedury klasyfikacyjnej medianę przestrzenną, która jest wektorem wielowymiarowym obliczonym na podstawie obserwacji zmiennych w T momentach czasowych,

3) obliczyć wartość mediany dla każdej zmiennej, biorąc pod uwagę jej wartości w każdym momencie obserwacji.

Zatem jeżeli decydujemy się na standaryzację wartości cech diagnostycznych (co nie zawsze jest konieczne), to chcąc uwzględnić zmiany zachodzące w bada-

nym okresie $t = 1, 2, \ldots, T$ oraz ich wpływ na wartości cech, stosować możemy następujące wzory:

$$S_j = \left[\frac{1}{TK} \sum_{k=1}^{K} \sum_{t=1}^{T} \left(x_{kj}^{t} - \bar{x}_j \right)^2 \right]^{0,5}, \tag{2}$$

gdzie:

$$\bar{x}_j = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} x_{kj}^{t}. \tag{3}$$

Rozszerzenie jednowymiarowych parametrów położenia, a więc i mediany, na przypadek wielowymiarowy nie jest tak jednoznaczne jak dla średniej arytmetycznej, która uwzględniając wszystkie zmienne, jest identyczna jak wektor średnich obliczony dla każdej zmiennej z osobna. Mediana wielowymiarowa została precyzyjnie określona w pracach [3; 4]. Szczególnie wiele uwagi poświęcono medianie Webera w pracy [3]. Medianą wielowymiarową będzie taki wektor *m*-wymiarowy, który minimalizuje sumę odległości euklidesowych mediany od każdej obserwacji, czyli:

$$\sum_{k=1}^{K} \sqrt{\sum_{j=1}^{m} \left(x_{kj} - sM_j \right)^2} = \min_{s\mathbf{M}}, \tag{4}$$

gdzie: sM_j – *j*-ta składowa wektora $s\mathbf{M}$,

x_{kj} – wartość *j*-tej zmiennej w *k*-tym obiekcie.

Chcąc ująć parametry położenia dla danych wielowymiarowych, które tworzy określona zmienna obserwowana w *T* momentach czasowych, traktujemy to jako zbiór *T* zmiennych opisujących badane obiekty. W celu obliczenia mediany przestrzennej traktować będziemy obserwacje danej zmiennej w *T* momentach czasowych jako zbiór *T* zmiennych. Stąd wartość mediany wyznaczamy ze wzoru:

$$\sum_{k=1}^{K} \sqrt{\sum_{t=1}^{T} \left(x_{kt} - sM_t \right)^2} = \min_{s\mathbf{M}}, \; t = 1, 2, \ldots, T, \tag{5}$$

gdzie: x_{kt} – wartość *t*-tej zmiennej ($t = 1, 2, \ldots, T$) w *k*-tym obiekcie badania.

Trzeci przypadek będzie określał medianę na podstawie macierzy (1), czyli: wyznaczamy dla każdej zmiennej medianę według jednego ze wzorów (por. [2]):

$$Me\, X_j = \frac{x_{kj}^{ti=(KT:2)} + x_{kj}^{ti=(KT:2)+1}}{2} \tag{6}$$

dla parzystej liczby będącej iloczynem liczby obiektów i okresów badania oraz

$$Me\,X_j = \frac{x_{kj}^{ti=(KT\cdot m):2} + x_{kj}^{ti=(KT\cdot m):2+1}}{2} \tag{7}$$

dla nieparzystej liczby będącej iloczynem liczby obiektów i okresów badania.

Etap 2

Dla każdego obiektu P_k $(k = 1, 2, \ldots, K)$ obliczamy uśrednione wartości cech X_j $(j = 1, 2, \ldots, m)$ według wzoru:

$$\overline{X}_{kj} = \sum_{t=}^{T} X_{jk}^{t}, \quad t = 1, 2, \ldots, T, \; j = 1, 2, \ldots, m, \; k = 1, 2, \ldots, K. \tag{8}$$

Etap 3

Proponowana procedura klasyfikacji uwzględnia dwa przypadki. W przypadku pierwszym algorytm klasyfikacji prowadzi do budowy $m+1$ klas oznaczonych symbolem S_g, gdzie $g = 1, 2, \ldots, G$ $(G = m+1)$, gdy zbiory opisane są za pomocą m zmiennych. W przypadku drugim algorytm klasyfikacji prowadzi do budowy 2^m (czyli $G = 2^m$) klas możliwych kombinacji z m zmiennych.

Rozważmy zatem **przypadek pierwszy**:

1° Do klasy S_1 wchodzą obiekty ze zbioru P, których uśrednione według wzoru (8) wartości wszystkich zmiennych X_j, czyli m zmiennych, są wyższe od zadanej statystyki pozycyjnej lub jej równe. Dla ustalenia uwagi przyjmiemy, że statystyką tą będzie mediana ($\overline{Me}$). Stąd:

$$\bigwedge_j : \bar{x}_{kj} \geq \overline{Me}\,X_j \,. \tag{9}$$

gdzie: $k = 1, 2, \ldots, K;\; j = 1, 2, \ldots, m;\; t = 1, 2, \ldots, T.$

$\overline{Me}\,X_j$ obliczono z wykorzystaniem jednego z podejść uwzględniających odpowiednio wzory od (2) do (7).

2° Do klasy S_2 wchodzą obiekty ze zbioru P (oprócz obiektów wyłonionych w punkcie 1º), których uśrednione wartości tylko $m-1$ zmiennych spełniają warunek:

$$\bar{x}_{kj} \geq \overline{Me}\,X_j \text{ dla } P_k \notin S_1. \tag{10}$$

m° Do klasy S_g ($g = m$) wchodzą obiekty ze zbioru P, których uśredniona wartość tylko jednej zmiennej X_j ze zbioru X spełnia warunek (9).

$(m + 1)^o$ Do klasy S_{g+1} ($g = m + 1$) wchodzą obiekty P_k, których uśredniona wartość $\bar{x}_{kj}$ żadnej ze zmiennych X_j nie spełnia warunku (9).

Przypadek drugi:

1° Klasę S_1 tworzą te obiekty P_k, których uśrednione według wzoru (8) wartości wszystkich *m* zmiennych X_j spełniają warunek:

$$\bigwedge_j \bar{x}_{kj} \geq \overline{Me}\, X_j, \tag{11}$$

gdzie: $j = 1, 2, \ldots, m$.

2° Klasę S_2 tworzą te obiekty P_k, których uśrednione wartości jedynie $(m - 1)$ zmiennych tworzących jedną z kombinacji $\binom{m}{m-1}$ zmiennych spełniają warunek (11).

3° Klasę trzecią S_3 tworzą te obiekty P_k, których uśrednione wartości zmiennych kolejnej kombinacji $(m - 1)$-elementowej spełniają warunek (11).
Po wyczerpaniu kombinacji $(m - 1)$-elementowych tworzymy klasy dla kombinacji $(m - 2)$-elementowych i stawiamy warunek (11).

2^m Klasę S_g $(g = 2^m)$ tworzymy z obiektów P_k, dla których uśrednione wartości $\bar{x}_{kj}$ wszystkich zmiennych X_j nie spełniają warunku (11).

Jak widać, oba przypadki mają wyraźnie odmienne założenia klasyfikacyjne. W przypadku pierwszym przypisujemy identyczne znaczenie wszystkim zmiennym, rozróżniając jedynie klasy obiektów przez liczbę zmiennych spełniających zadane warunki. Natomiast w drugim przypadku rozróżniamy grupy obiektów poprzez identyfikację specyfikacji zmiennych spełniających zadane warunki klasyfikacji. W obu przypadkach wartość statystyki pozycyjnej obliczana jest z uwzględnieniem wartości cech z całego okresu badania, a więc dla $t = 1, 2, \ldots, T$.

4. Przykład ilustrujący klasyfikację pozycyjną w ujęciu dynamicznym

Dany jest zbiór obiektów, którymi są regiony szczebla NUTS-2 państw Unii Europejskiej (w badaniu z 268 obecnie wydzielonych regionów UE-27 poziomu NUTS-2, analizowano 240[1] regionów), $K = 240$, $k = 1, 2, \ldots, 240$. Każdy z obiektów opisany jest 3 cechami diagnostycznymi:

X_1 – wartość PKB *per capita* w *k*-tym regionie,

X_2 – dynamika wzrostu PKB *per capita* w regionie w momencie *t* + 1 do *t*,

X_3 – stopa aktywności zawodowej w regionie.

[1] Regiony, których nie uwzględniono w badaniu, podano pod tab. 1.

Wartości cech obserwowane są w pięciu momentach czasowych, od roku 2000 do roku 2004, czyli $t = 1, 2, 3, 4, 5$. Stąd macierz obserwacji ma postać macierzy blokowej:

$$X_{kj}^{t} = \begin{pmatrix} x_{1\,1}^{1} & x_{2\,1}^{1} & x_{3\,1}^{1} \\ \vdots & \vdots & \vdots \\ x_{1\,240}^{1} & x_{2\,240}^{1} & x_{3\,240}^{1} \\ x_{1\,1}^{2} & x_{2\,1}^{2} & x_{3\,1}^{2} \\ \vdots & \vdots & \vdots \\ x_{1\,240}^{2} & x_{2\,240}^{2} & x_{3\,240}^{2} \\ x_{1\,1}^{3} & x_{2\,1}^{3} & x_{3\,1}^{3} \\ \vdots & \vdots & \vdots \\ x_{1\,240}^{3} & x_{2\,240}^{3} & x_{3\,240}^{3} \\ x_{1\,1}^{4} & x_{2\,1}^{4} & x_{3\,1}^{4} \\ \vdots & \vdots & \vdots \\ x_{1\,240}^{4} & x_{2\,240}^{4} & x3_{1\,240}^{4} \\ x_{1\,1}^{5} & x_{2\,1}^{5} & x_{3\,1}^{5} \\ \vdots & \vdots & \vdots \\ x_{1\,240}^{5} & x2_{1\,240}^{5} & x_{3\,240}^{5} \end{pmatrix}. \tag{12}$$

Wartości median $\overline{Me}\, X_j^t$, dla $j = 1, 2, 3$ i $t = 1, 2, 3, 4, 5$, każdej zmiennej z uwzględnieniem każdego momentu obserwacji wynoszą:
– dla cechy $X_1 = 19\,960{,}5$,
– dla cechy $X_2 = 104{,}64$,
– dla cechy $X_3 = 57{,}1$.

Klasę I tworzyć będą obiekty P_k – regiony UE, których wartości (obliczone według wzoru (8)) trzech zmiennych w badanym okresie 2000-2004 były wyższe od mediany określonej z uwzględnieniem pełnego okresu badań.

Klasę II tworzą regiony państw UE, których uśrednione według wzoru (8) wartości dwóch zmiennych były wyższe od mediany określonej z uwzględnieniem pełnego okresu badań.

Klasę III tworzą regiony, których wartości (obliczone według wzoru (8)) jednej zmiennej są wyższe od mediany określonej z uwzględnieniem pełnego okresu badań.

Klasę IV tworzą regiony, których uśrednione według wzoru (8) wartości wszystkich zmiennych są niższe od mediany określonej z uwzględnieniem pełnego okresu badań.

Wyniki klasyfikacji regionów UE szczebla NUTS 2 z przyporządkowaniem ich do państw członkowskich podano w tab. 1.

Tabela 1. Wyniki klasyfikacji regionów szczebla NUTS-2[a] w europejskiej przestrzeni regionalnej

Kraj	Liczba regionów kraju	Klasa – liczba regionów w klasie			
		1	2	3	4
Belgia	11		1	6	4
Czechy	8	1	6	1	
Dania	1		1		
Niemcy	41	1	24	13	3
Estonia	1		1		
Grecja[a]	(13) 12		1	7	4
Hiszpania[a]	(19) 17		5	11	1
Francja[a]	(26) 21		3	8	10
Włochy	21		1	12	8
Cypr	1			1	
Łotwa	1			1	
Litwa	1			1	
Luksemburg	1		1		
Węgry	7			6	1
Malta	1				1
Holandia	12		1	11	
Austria	9		6	2	1
Polska	16			2	14
Portugalia[a]	(7) 5		1	2	2
Słowenia	1		1		
Słowacja	4	1	3		
Finlandia	5		4		1
Szwecja	8	00	8		
Wielka Brytania[a]	(37) 35	12	21	1	1
Ogółem[a]	240	15	89	85	51

[a] Bez regionów bułgarskich, rumuńskich i irlandzkich oraz hiszpańskich zamorskich Ciudad Autónoma de Ceuta i Ciudad Autónoma de Melilla, francuskich zamorskich Guadeloupe, Martinique, Guyane, Reunion oraz Corse, portugalskich autonomicznych Região Autónoma dos Açores, Região Autónoma da Madera, greckiego Voreio Aigaio, brytyjskich (szkockich) Eastern Scotland South, Western Scotland.

Źródło: obliczenia własne na podstawie danych Eurostatu.

5. Podsumowanie

Zaproponowana procedura klasyfikacji pozwala na wykorzystanie danych przekrojowo-czasowych oraz obliczanie statystyk pozycyjnych w całym okresie badania. Podejście to uwzględnia dynamikę zmian zachodzących w wartościach zmiennych ilustrujących badane obiekty w całym okresie badania i uzupełnia dotychczasowy dorobek metod klasyfikacji dynamicznej.

Literatura

[1] Jajuga K., *Statystyczna analiza wielowymiarowa*, PWN, Warszawa 1993.
[2] Luszniewicz A., Słaby T., *Statystyka stosowana*, PWE, Warszawa 1998.
[3] Młodak A., *Analiza taksonomiczna w statystyce regionalnej*, Difin, Warszawa 2006.
[4] *Statystyczne metody analizy danych*, red. W. Ostasiewicz, AE, Wrocław 1999.
[5] Strahl D., *Dynamiczno-strukturalna miara rozwoju obiektów hierarchicznych*, [w:] Ekonometria 18, *Zastosowanie metod ilościowych*, red. J. Dziechciarz, AE, Wrocław 2006.
[6] Strahl D., *Klasyfikacja regionów z medianą*, [w:] Ekonometria 10, *Zastosowania metod ilościowych*, red. J. Dziechciarz, AE, Wrocław 2002, s. 11-18.
[7] *Taksonomiczna analiza przestrzennego zróżnicowania poziomu życia w Polsce w ujęciu dynamicznym*, red. A. Zeliaś, AE, Kraków 2000.
[8] Walesiak M., *Metody analizy danych marketingowych*, Wyd. Naukowe PWN, Warszawa 1996.
[9] Ward J.H., *Hierarchical Grouping to Optimize on Objective Functions*, „Journal of the American Statistical Association" 1963 nr 58.
[10] Zeliaś A., *Some Notes the Selection of Normalization of Diagnostic Variables*, „Statistics in Transition" 2002, vol. 5, nr 5, s. 787-802.

POSITIONAL CLASSIFICATION. DYNAMIC APPROACH

Summary

The article presents the proposal of classification referring to a set of objects and performed by means of positional statistics in a dynamic perspective. The article focuses on a median and takes into consideration three aspects of a dynamic approach in the procedure for objects classification.

1) performing normalization of attributes value considering the flow of time and applying due formulas for calculating the mean value and standard deviation covering the whole research period,

2) introducing the spatial median into classification procedure in which space is made up of variables values in particular research periods,

3) calculating the value of median for every variable considering its value at each observation point.

The article is concluded by an example illustrating the proper procedure.

Danuta Strahl – prof. zw. dr hab., kierownik Katedry Gospodarki Regionalnej Uniwersytetu Ekonomicznego we Wrocławiu – Wydział w Jeleniej Górze.