PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 449 Metody statystyczne 1988
wielowymiarowej analizy porównawczej

Marek W a l e s i a k
Akademia Ekonomiczna we Wrocławiu

SPOSOBY WYZNACZANIA OPTYMALNEJ LICZBY KLAS W ZAGADNIENIU KLASYFIKACJI HIERARCHICZNEJ

1. Wstęp

Niech dany będzie n-elementowy zbiór obiektów $A = \{A_1,\ldots,A_n\}$. Każdy z obiektów opisany jest za pomocą wartości m zmiennych: X_1, ..., X_m. Załóżmy, że chcemy dokonać podziału zbioru obiektów $A = \{A_1$, ..., $A_n\}$, na względnie jednorodne (homogeniczne) klasy $P_1,\ldots,P_H$, spełniającego następujące warunki:

1° zupełności $\bigcup_{h=1}^{H} P_h = A$,

2° rozłączności $\bigwedge_{\substack{h,h' = 1,\ldots,H \\ h \neq h'}} P_h \cap P_{h'} \neq \emptyset$,

3° niepustości $\bigwedge_{h} P_h \neq \emptyset$.

W literaturze różnie definiuje się pojęcie klasy jednorodnej (homogenicznej) [4; 15]:

a) klasa jest taką zbiorowością obiektów w przestrzeni, w której podobieństwo pomiędzy dowolną parą obiektów jest większe niż podobieństwo pomiędzy jakimkolwiek obiektem należącym do klasy, a dowolnym obiektem nie należącym do niej;

b) klasy tworzą obiekty, których podobieństwo z najbardziej im podobnym obiektem jest większe niż podobieństwo najbardziej podobnych obiektów należących do różnych klas;

c) klasy tworzą obiekty, jak najbardziej podobne, natomiast w różnych klasach znajdują się obiekty jak najmniej podobne;

d) klasami są takie obszary w przestrzeni, które charakteryzują się większą gęstością i są oddzielone obszarami o mniejszej gęstości;

e) klasy tworzą obiekty, których podobieństwo z najbardziej im podobnym obiektem jest większe niż wszystkie podobieństwa międzyklasowe (te z kolei mogą być zdefiniowane w różny sposób).

Definicja c) - z uwagi na jej ogólny charakter (definicje a), b) i e) są jej szczególnymi przypadkami) - ma w praktyce największe znaczenie, bowiem dopuszcza stosowanie różnych rozwiązań w zakresie metod klasyfikacji.

Po tym krótkim wstępie możemy zdefiniować co będziemy rozumieli przez zagadnienie klasyfikacji hierarchicznej.

Metody hierarchiczne dzielą się na metody aglomeracyjne i deglomeracyjne. W wyniku przeprowadzenia klasyfikacji hierarchicznej otrzymuje się ciąg podziałów (przy czym dla każdego z nich zachodzą warunki 1°-3°), w którym spełniony jest taki warunek, że podział w danym kroku klasyfikacji powstaje z połączenia (rozpadu) dwóch (jednej) lub więcej klas w kroku poprzednim w przypadku klasyfikacji aglomeracyjnej (deglomeracyjnej).

Hierarchiczne metody aglomeracyjne charakteryzują się (w ujęciu klasycznym)[1] następującymi cechami:

a) punktem wyjścia jest N(A) klas jednoelementowych (jest tyle klas ile jest obiektów); N(A) - liczba obiektów w zbiorze A;

b) po każdym kroku klasyfikacji liczba klas zmniejsza się o jeden, przy czym zmniejszenie liczby klas następuje przez połączenie dwóch istniejących;

c) istnieje N(A) - 1 kroków klasyfikacji; po N(A) - 1 krokach otrzymuje się jedną klasę zawierającą wszystkie obiekty A_r (r=1,...,n);

d) proces klasyfikacji można przedstawić graficznie przy pomocy dendrogramu (drzewka połączeń) wskazującego na kolejność połączeń między klasami.

N. Jardine i R. Sibson [10, s. 49] definiują dendrogram w sposób następujący. Niech E(A) oznacza zbiór relacji równoważności (ekwiwalentności) określony na zbiorze A. Dendrogram jest to funkcja: $f: (0; \infty) \longrightarrow E(A)$ spełniająca następujące warunki:

1) każda klasa na danym poziomie α' jest połączeniem klas na poziomie α, gdzie

$0 \leqslant \alpha \leqslant \alpha'$ (dla miar odległości),

$1 \geqslant \alpha \geqslant \alpha'$ (dla miar bliskości)[2].

---

[1] W ujęciu nieklasycznym proces klasyfikacji można zakończyć w ciągu mniejszej liczby kroków niż N(A) - 1 (por. [2]).

[2] Wśród miar podobieństwa wyodrębnia się miary odległości oraz bliskości (por. np. [3]). Podobieństwo między obiektami (klasami) jest tym większe, im miara odległości (bliskości) przyjmuje mniejszą (większą) wartość i odwrotnie.

Poziom połączenia klas jest to wartość podobieństwa między najbardziej podobnymi klasami łączonymi w danej iteracji hierarchicznej metody aglomeracyjnej (przy czym podobieństwo międzyklasowe jest różnie definiowane w poszczególnych metodach aglomeracyjnych).

2) dla dostatecznie dużej (małej dla miar bliskości) wartości α wszystkie obiekty znajdują się w jednej klasie.

3) mając daną wartość α istnieje takie $\delta > 0$ ($\delta < 0$ - dla miar bliskości), że klasy na poziomie α i $\alpha + \delta$ są identyczne.

Ta definicja dendrogramu "odrzuca" te metody aglomeracyjne, w których wartości poziomu α związane z łączeniem najbardziej podobnych klas mogą podnosić się i spadać przy przechodzeniu z kroku na krok w klasyfikacji hierarchicznej (metoda środka ciężkości, mediany, średniego połączenia wewnątrz nowej klasy - por. np. [1]).

Omówienie algorytmów hierarchicznych metod aglomeracyjnych zawiera wiele monografii (por. np. [1; 4; 7; 13]).

Z kolei hierarchiczne metody deglomeracyjne charakteryzują się następującymi cechami (w ujęciu klasycznym):

a) punktem wyjścia jest jedna klasa obejmująca wszystkie obiekty badania $A_1, \ldots, A_n$;

b) po każdym kroku klasyfikacji liczba klas zwiększa się o jeden, przy czym zwiększenie liczby klas następuje przez rozdzielenie jednej z istniejących klas;

c) istnieje $N(A) - 1$ kroków klasyfikacji; po $N(A) - 1$ krokach otrzymuje się liczbę klas równą liczbie obiektów badania, tzn. każdy obiekt tworzy jedną klasę.

Omówienie metod deglomeracyjnych zawierają m.in. takie pozycje literatury, jak [7; 14; 15]. Można zaliczyć do nich m.in. metody Huberta, metodę taksonomii wrocławskiej oraz najkrótszą sieć połączeń Prima [5; 7; 14; 15].

2. Przegląd sposobów wyznaczania optymalnej liczby klas

W literaturze [9; 14] znane są liczne sposoby wyboru klasyfikacji optymalnej w przypadku, gdy liczba klas na które należy podzielić zbiór badanych obiektów jest dana. W praktycznych zastosowaniach zazwyczaj brak jest takiej informacji. Dlatego też bardziej interesującym problemem jest przypadek, w którym badacz nie dysponuje informacją o liczbie klas, na które należy podzielić rozpatrywaną zbiorowość obiektów. Taka sytuacja stanowić będzie przedmiot rozważań niniejszego artykułu. W zasadzie wszystkie omawiane sposoby służą do wyznaczenia klasyfikacji optymalnej z ciągu klasyfikacji, ale pośrednio rozwiązują również problem wyznaczania liczby klas. Zatem należy przyjąć założenie,

że najwłaściwszą liczbą klas dla danego zbioru obiektów jest liczba klas wynikająca z klasyfikacji optymalnej [9].

Metody klasyfikacji nie dają odpowiedzi na pytanie jak wyznaczyć optymalną liczbę klas na które należy podzielić zbiór badanych obiektów A. Problem ten jest niezwykle istotny, jeżeli chodzi o potencjalnego użytkownika metod klasyfikacji, bowiem nie interesuje go zazwyczaj sama zastosowana metoda, lecz podział zbioru obiektów na klasy otrzymany w wyniku jej zastosowania. W literaturze przedmiotu [2; 5; 6; 9; 11; 12] można spotkać wiele propozycji w tym zakresie (świadczy to o ważności tego problemu), lecz nie ma wśród nich doskonałych. Trwają ciągłe poszukiwania sposobów rozwiązania problemu wyznaczania optymalnej liczby klas.

W pracy [11] R. Mojena podaje dwie propozycje wyznaczania optymalnej liczby klas w hierarchicznych metodach aglomeracyjnych. Sposoby te bazują na rozkładzie wartości poziomu połączenia klas (α). Wartości poziomu α rozłożone są monotonicznie rosnąco w przypadku miar odległości oraz monotonicznie malejąco w przypadku miar bliskości (dla metod, których graficzną prezentacją jest dendrogram spełniający wymogi określenia podanego w punkcie pierwszym). Sposoby te nie znajdują zastosowania dla wyników metod centroidalnych ("mediany" i "środka ciężkości") oraz metody średniego połączenia wewnątrz nowej klasy. Wynika to z faktu, że wartości poziomu α związane z łączeniem klas najbardziej podobnych mogą w tych metodach podnosić się i spadać przy przechodzeniu z kroku na krok w klasyfikacji hierarchicznej.

Sposób 1. W sposobie pierwszym na podstawie n - 1 wartości poziomu połączenia klas wyznacza się średnią arytmetyczną ($\bar{\alpha}$) oraz odchylenie standardowe (S_α). Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok e (e=1,...,n-2) jako pierwszy spełnia nierówność:

$$\alpha_{e+1} > \bar{\alpha} + aS_\alpha \quad \text{(dla miar odległości)} \tag{2.1}$$

gdzie:

α_{e+1} - wartość poziomu połączenia klas w kroku e+1,

a - dowolna liczba rzeczywista ustalona tak, aby otrzymać nietrywialny podział wynikowy (zazwyczaj $a \in \langle -3; 3 \rangle$);

e - numer kroku w hierarchicznej procedurze klasyfikacji.

W przypadku, gdy macierz podobieństwa [c] była ustalona w oparciu o miarę bliskości to znak nierówności (2.1) należy zamienić na przeciwny.

Sposób 2. W sposobie drugim również liczy się średnią arytmetyczną oraz odchylenie standardowe z wartości poziomu połączenia klas (dodat-

kowo jeszcze pewne wielkości korygujące), jednakże wyznacza się je sekwencyjnie po każdym kolejnym e-tym kroku procedury. Z tego tytułu wynika przewaga tego sposobu wyznaczania optymalnej liczby klas nad sposobem pierwszym. Chcąc wyznaczyć optymalną liczbę klas w sposobie pierwszym należy znać wszystkie wartości poziomu połączenia klas (tj. n-2), podczas gdy w sposobie drugim tylko e ($e \leqslant n-2$). Sekwencyjność postępowania w sposobie drugim powoduje, że proces wyznaczania kolejnych podziałów kończy się, gdy spełniona jest relacja (2.2).

Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok e ($e=e', e'+1,\ldots,n-2$) jako pierwszy spełnia nierówność:

$$\alpha_{e+1} > \bar{\alpha}_e + \beta_e + \gamma_e + a\,S_e \quad \text{(dla miar odległości)}, \tag{2.2}$$

gdzie: $\bar{\alpha}_e$ - średnia ruchoma w kroku e (z wartości $\alpha_1,\ldots,\alpha_e$),
S_e - odchylenie standardowe ruchome w kroku e (z wartości $\alpha_1,\ldots,\alpha_e$),

$$\beta_e = \frac{6\left[2\sum_{f=e-e'+1}^{e} w_f\alpha_f - (e'+1)\sum_{f=e-e'+1}^{e}\alpha_f\right]}{e'(e'^2-1)},$$

$$\gamma_e = (e'-1)\,\beta_e : 2\,,$$

$$w_f = w_{f-1} + 1,\ f = e-e'+2,\ldots,e, \quad \text{gdzie: } w_{e-e'+1} = 1.$$

W sytuacji, gdy macierz podobieństw ustalona była w oparciu o miarę bliskości, to znak powyższej nierówności należy zamienić na przeciwny.

W hierarchicznych procedurach aglomeracyjnych klasy tworzy się w ten sposób, aby minimalizować stratę informacji towarzyszącą wzrostowi uogólnienia w poszczególnych krokach klasyfikacji. B.J.L. Berry [2] zaproponował procedurę pozwalającą z ciągu klasyfikacji wybrać optymalną. Jako miernik straty informacji Berry zastosował odległość wewnątrzklasową (jest to suma wszystkich odległości między obiektami wewnątrz klas w danym kroku klasyfikacji). Celem uzyskania właściwych podziałów konieczne jest przerwanie procesu klasyfikacji na jakimś poziomie straty informacji odpowiadającym określonej iteracji. Problem ten jak dotychczas nie został w pełni rozwiązany. Proponuje się przerwać proces klasyfikacji po tej iteracji, po której następuje wyraźny skok w utracie informacji szczegółowej. Jeśli jest jeden taki skok to wybór jest oczywisty, natomiast przestaje on być oczywisty, gdy nie ma wyraźnego skoku w ogóle lub jest ich kilka.

Innego typu funkcję służącą do ustalenia optymalnej liczby klas

zaproponowali Fortier i Solomon [6]. Z ciągu klasyfikacji (otrzymanych hierarchicznymi metodami aglomeracyjnymi lub deglomeracyjnymi) proponują oni wybrać tę, dla której następująca funkcja:

$$\sum_{P_h^e \in P^e} \Big(\sum_{\substack{A_r, A_s \in P_h^e \\ r<s}} d(A_r, A_s) - \frac{1}{2} N(P_h^e)[N(P_h^e) - 1] d^x \Big), \tag{2.3}$$

$$\sum_{P_h^e \in P^e} \Big(\frac{1}{2} N(P_h^e)[N(P_h^e) - 1] g^x - \sum_{\substack{A_r, A_s \in P_h^e \\ r<s}} g(A_r, A_s) \Big), \tag{2.4}$$

gdzie:

$d(A_r, A_s)$, $g(A_r, A_s)$ - wartość miary odpowiednio odległości i bliskości między obiektami A_r i A_s ($r, s = 1, \ldots, n$);

P^e - zbiór klas w e-tym kroku procedury hierarchicznej;

P_h^e - h-ta klasa w e-tym kroku procedury hierarchicznej;

$N(P_h^e)$ - liczebność klasy P_h w e-tym kroku procedury hierarchicznej; $e = 1, \ldots, n-2$; d^x, g^x - krytyczna wartość miary odpowiednio odległości i bliskości;

osiąga minimum.

Interesujący sposób wyznaczania klasyfikacji optymalnej (a więc i liczby klas) zaproponował G.N. Žitkov (por. [9]). Rozpatruje się w nim nie tylko odległości między obiektami znajdującymi się w jednej klasie (por. sposób Fortiera i Solomona), ale również odległości między obiektami należącymi do różnych klas.

W sposobie tym ogół wartości miar odległości między obiektami A_r i A_s dzieli się w następujący sposób:

$$S_{rs} = \begin{cases} 1, & \text{jeśli } d(A_r, A_s) \leqslant d^x \quad \text{odległości małe,} \\ 0, & \text{jeśli } d(A_r, A_s) > d^x \quad \text{odległości duże.} \end{cases}$$

Z ciągu klasyfikacji wybiera się tę, która osiąga wartość maksymalną dla funkcji:

$$M_e = \sum_{P_h^e \in P^e} V_h : N(P^e), \tag{2.5}$$

gdzie:

$$V_h = I_{hh} - \Big(\sum_{P_{h'}^e \in P^e} I_{hh'} \Big) : [N(P^e) - 1]$$

a

$$I_{hh'} = \Big(\sum_{A_r \in P_h^e} \sum_{A_s \in P_{h'}^e} s_{rs} \Big) : N(P_h^e)\, N(P_{h'}^e),$$

$N(P^e)$ - liczba klas w e-tym kroku klasyfikacji.

Operowanie liczebnościami odległości małych i dużych może spowodować zniekształcenie rezultatów obliczeń. Dlatego też poprawniejsze wydaje się być podzielenie ogółu odległości między obiektami A_r i A_s w nieco inny sposób:

$$s_{rs} = \begin{cases} d(A_r,A_s), & \text{jeśli } d(A_r,A_s) \leqslant d^x \text{ odległości małe,} \\ 0 \quad\quad\;\; , & \text{jeśli } d(A_r,A_s) > d^x \text{ odległości duże.} \end{cases}$$

Wtedy wzór na $I_{hh'}$ przyjmie następującą postać:

$$I_{hh'} = \Big(\sum_{A_r \in P_h^e} \sum_{A_s \in P_{h'}^e} s_{rs} \Big) : \Big(\sum_{A_r \in P_h^e} \sum_{A_s \in P_{h'}^e} d(A_r, A_s) \Big),$$

gdzie:

I_{hh} - miara gęstości klasy P_h (jest to udział sumy odległości małych w sumie wszystkich odległości między obiektami w klasie P_h),

$I_{hh'}$ - udział sumy odległości małych w sumie wszystkich odległości między obiektami klas P_h i $P_{h'}$.

Zaletą powyższych dwóch sposobów (a więc Fortiera i Solomona oraz Žitkova) jest to, że mogą być zastosowane do wyników klasyfikacji otrzymanych każdą metodą hierarchiczną (aglomeracyjną czy też deglomeracyjną).

Metoda taksonomii wrocławskiej (a co za tym idzie i metoda Prima [15]) zawiera pewne sposoby pozwalające z ciągu klasyfikacji wybrać optymalną. Mając ustawione w dendrycie podobieństwa między sąsiadującymi obiektami (graficznie reprezentowane przez krawędzie) w porządku wzrastającego podobieństwa (malejących wartości odległości d_e lub wzrastających wartości bliskości g_e) oblicza się ilorazy sąsiednich wyrazów [5]:

$$w_e = \frac{d_e}{d_{e+1}}, \quad (e,e' = 1,\ldots,n-2), \tag{2.6}$$

$$w_e = \frac{g_{e+1}}{g_e}. \tag{2.7}$$

Z otrzymanych podziałów na $N(P^e)$ i $N(P^{e+1})$ klas wybiera się ten dla którego zachodzi (otrzymuje się wtedy podział naturalny):

$$w_e < w_{e+1}. \tag{2.8}$$

Z dwóch podziałów naturalnych na $N(P^e)$ i $N(P^{e'})$ klas ten jest lepszy, który ma mniejszą wartość w_e.

Inny sposób wyznaczania optymalnej liczby klas prezentuje Z. Hellwig w pracy [8]. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok e (e = 1,..., n-2) jako ostatni spełnia nierówność:

$$d_e > \overline{d} + a\, S_d, \tag{2.9}$$

$$g_e < \overline{g} + a\, S_g, \tag{2.10}$$

gdzie:

a - jak we wzorze (2.1); w pracy [8] a = 2;

d_e, g_e - wartość miary odległości (odpowiednio bliskości) między obiektami sąsiadującymi w dendrycie w kroku e-tym;

$$\overline{d} = \frac{1}{n} \sum_{s} \min d(A_r, A_s);$$

$$S_d = \left[\frac{1}{n} \sum_{s} (\min d(A_r, A_s) - \overline{d})^2\right]^{0,5};$$

analogiczne wzory otrzymuje się dla miary bliskości wstawiając w miejsce symbolu d symbol g.

Przedstawiony przegląd z pewnością nie wyczerpuje zbioru istniejących sposobów wyznaczania optymalnej liczby klas. Inne sposoby zawarte są m.in. w pracach [9; 12].

Nie trudno zauważyć, że podstawowym mankamentem powyższych sposobów wyznaczania optymalnej liczby klas w zagadnieniu klasyfikacji hierarchicznej jest arbitralne dobieranie pewnych wielkości krytycznych. Nikt się jeszcze nie wypowiedział na temat ustalania tych wielkości. Problem ten nie jest zresztą bardzo istotny, jeśli badacz nie wychodzi poza rozważania teoretyczne. W sytuacji, gdy rezultaty badań mają służyć celom decyzyjnym wszelkie niejasności w trakcie obliczeń mogą mieć poważne konsekwencje.

W związku z powyższym od badacza wykorzystującego w badaniach metody klasyfikacji wymaga się dwojakiego rodzaju umiejętności, tzn. w zakresie wybranej dyscypliny badawczej (merytoryczna znajomość zagadnienia) oraz w zakresie opanowania metod statystyczno-ekonometrycznych (metodologiczna znajomość zagadnienia), wyrażające naukową swobodę badacza. Otrzymane przy pomocy metod klasyfikacji rezultaty powinny odpowiadać logiczno-intuicyjnej interpretacji obserwowanych zjawisk.

LITERATURA

[1] A n d e r b e r g M.R.: Cluster analysis for applications. New York, San Francisco, London: Academic Press 1973.

[2] C h o j n i c k i Z., C z y ż T.: Metody taksonomii numerycznej w regionalizacji geograficznej. Warszawa: PWN 1973.

[3] D ą b r o w s k i M., L a u s - M ą c z y ń s k a K.: Metody wyszukiwania i klasyfikacji informacji. Warszawa: WNT 1978.

[4] E v e r i t t B.S.: Cluster analysis. London: Heinemann Educational Books Ltd 1977.

[5] F l o r e k K., Ł u k a s z e w i c z J., P e r k a l J., S t e i n h a u s H., Z u b r z y c k i S.: Taksonomia Wrocławska. "Przegląd Antropologiczny". Poznań: 1951. T. 17.

[6] F o r t i e r J.J., S o l o m o n H.: Clustering procedures. W: Multivariate analysis. Red. P.R. Krishnaiah. New York, London: Academic Press 1966.

[7] G r a b i ń s k i T., W y d y m u s S., Z e l i a ś A.: Metody doboru zmiennych w modelach ekonometrycznych. Warszawa: PWN 1982.

[8] H e l l w i g Z.: Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr. "Przegląd Statystyczny", 1968 z. 4.

[9] J a j u g a K.: O sposobach określania ilości klas w zagadnieniu klasyfikacji i klasyfikacji rozmytej. W: Metody taksonomiczne i ich zastosowanie w badaniach ekonomicznych. Wrocław: AE 1984. Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 262.

[10] J a r d i n e N., S i b s o n R.: Mathematical taxonomy. London, New York, Sydney, Toronto: John Wiley & Sons Ltd 1971.

[11] M o j e n a R.: Hierarchical grouping methods and stopping rules: an evaluation. "The Computer Journal", 1977 No 4 Vol. 20.

[12] P a m u ł a J., S o k o ł o w s k i A.: Propozycja wyznaczania podziału wynikowego aglomeracyjnych algorytmów taksonomicznych. W: Prace z zakresu statystyki, ekonometrii, programowania matematycznego i matematyki. Kraków: AE 1980. Zeszyty Naukowe Akademii Ekonomicznej w Krakowie nr 127.

[13] S n e a t h P.H.A., S o k a l R.R.: Numerical taxonomy. The principles and practice of numerical classification. San Francisco: W.H. Freeman 1973.

[14] S z c z o t k a F.A.: Podstawy taksonomii numerycznej. Opracowanie wykonane w ramach problemu węzłowego 11.2.1. Grupa tematyczna 03. Temat A1. Warszawa: IGiPZ PAN 1975 (maszynopis powielony).

[15] W a l e s i a k M.: Metody klasyfikacji w badaniach strukturalnych. Rozprawa doktorska. Wrocław: AE 1985 (maszynopis).