PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

ROZMYTA KLASYFIKACJA *K*-ŚREDNICH DLA DANYCH SYMBOLICZNYCH INTERWAŁOWYCH

Streszczenie: Artykuł przedstawia adaptacyjną i nieadaptacyjną klasyfikację *k*-średnich dla danych symbolicznych. Obydwie te metody znajdują zastosowanie wyłącznie dla interwałowych zmiennych symbolicznych. W artykule przedstawiono także typy zmiennych symbolicznych.

W części empirycznej zastosowano nieadaptacyjną klasyfikację *k*-średnich dla przykładowych danych symbolicznych.

1. Wstęp

W analizie danych symbolicznych zaproponowano wiele różnorodnych metod klasyfikacji, które można podzielić na dwie główne grupy metod.

Pierwszą z nich są metody **sekwencyjne (iteracyjne)**, które optymalizują (wykorzystując w tym celu pewną funkcję – kryterium) początkowy podział obiektów zgodnie z pewnym algorytmem. Wśród tych metod wyróżnia się: **metody tworzące skupienia rozłączne** (np. klasyfikacja dynamiczna, metoda COBWEB Michalskiego i inne) oraz **metody tworzące skupienia nierozłączne** (rozmyta klasyfikacja *k*-średnich dla danych symbolicznych, adaptacyjna rozmyta klasyfikacja *k*-średnich dla danych symbolicznych). Wśród metod tworzących skupienia nierozłączne ważne miejsce **zajmują metody klasyfikacji rozmytej**.

Drugą grupą metod są metody **hierarchiczne** – wśród tych metod wyróżnia się: **metody aglomeracyjne** (np. metodę Brito, metodę klasyfikacji Gowdy-Didaya i inne), **metody deglomeracyjne** (metodę podziałową opartą na kryteriach Chavent, metodę EPAM Simona oraz Feigenbauma i inne), **metody tworzące skupienia nierozłączne** (metodę piramid Brito).

Przegląd różnorodnych metod klasyfikacji danych symbolicznych prezentują m.in. prace: [De Carvalho 2007; Verde 2004; Pełka 2009].

Obiekty symboliczne ze względu na zmienne, które je opisują, oraz fakt, że obiekty symboliczne drugiego rzędu są agregatami (grupami, złożeniami) obiektów pierwszego rzędu (zob. [Bock, Diday 2000]), w wielu przypadkach nie należą tylko i wyłącznie do jednej klasy, lecz do wielu klas, tworząc skupienia nierozłączne.

Wynika z tego potrzeba tworzenia metod klasyfikacji nierozłącznej, w tym metod klasyfikacji rozmytej.

Celem artykułu jest zaprezentowanie metody nieadaptacyjnej rozmytej klasyfikacji *k*-średnich dla danych interwałowych, którą zaproponował De Carvalho [2007]. Celem dodatkowym jest próba oceny wpływu wielkości parametru rozmycia na homogeniczność otrzymanych klas.

W części empirycznej przedstawiono przykład ilustracyjny zastosowania nieadaptacyjnej rozmytej klasyfikacji *k*-średnich na przykładowych danych symbolicznych. Wykorzystano tu dane dostępne w pakiecie SODAS oraz dane o znanej strukturze klas wygenerowane z wykorzystaniem procedury `cluster.Gen` z pakietu `clusterSim` dla programu `R`.

2. Typy zmiennych w analizie danych symbolicznych

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

1) ilorazowe, przedziałowe, porządkowe, nominalne;

2) kategorie, np. biały, zielony;

3) interwałowe, czyli przedziały liczbowe, rozłączne lub nierozłączne, np. ilość spalanej benzyny na 100 km w pewnym samochodzie (6 litrów; 13 litrów);

4) wielowariantowe, przykładem może być typ nadwozia samochodu pewnej marki: sedan, hatchback, minivan, coupé, co oznacza, że dostępny jest on w czterech różnych wariantach nadwozia;

5) wielowariantowe z wagami (prawdopodobieństwami), gdzie oprócz listy kategorii występują wagi (prawdopodobieństwa), z jakimi obiekt ma wybraną kategorię, np. jeżeli wybrać zmienną wybrane kolory nadwozia dla pewnej marki i modelu samochodu: biały (0,45), zielony (0,30), czarny (0,15), to oznacza to, że możemy kupić samochód w kolorze białym i zielonym, natomiast kolor czarny jest o wiele mniej spotykany. Sytuacja taka może wynikać m.in. z polityki producenta czy popularności kolorów;

6) zmienne strukturalne [Bock, Diday 2000, s. 2-3; 33-37] – w literaturze przedmiotu wyróżnia się oprócz wyżej wymienionych typów zmiennych także zmienne strukturalne:

a) zmienne o zależności funkcyjnej lub logicznej pomiędzy poszczególnymi 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) zmienne taksonomiczne, w których *a priori* ustalono systematykę, według której przyjmuje ona swoje realizacje.

3. Nieadaptacyjna rozmyta klasyfikacja *k*-średnich

Metodę rozmytej klasyfikacji *k*-średnich dla danych w rozumieniu klasycznym zaproponował Dunn [1973], a następnie jej modyfikację Bezdek [1981].

De Carvalho [2007] zaproponował modyfikację rozmytej klasyfikacji *k*-średnich dla danych klasycznych, która umożliwia klasyfikację obiektów symbolicznych opisywanych wyłącznie zmiennymi interwałowymi (zob. [De Carvalho 2007, s. 424]).

Algorytm nieadaptacyjnej rozmytej klasyfikacji *k*-średnich dla danych interwałowych przedstawia się następująco [De Carvalho 2007, s. 425]:

1. Ustal liczbę klas c, na które zostanie dokonany podział zbioru obiektów.
2. Wybierz wielkość parametru rozmycia m $(m>1)$.
3. Ustal maksymalną liczbę iteracji T oraz kryterium stopu $\varepsilon > 0$.
4. Dla każdego obiektu ustal stopień przynależności do *i*-tej klasy $u_{ik} \geq 0$ $\left(\sum_{i=1}^{c} u_{ik} = 1\right)$, gdzie $i = 1, \ldots, c$ – liczba klas.
5. Dla ustalonych u_{ik} wyznacz współrzędne prototypów klas zgodnie ze wzorami:

$$\alpha_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^m a_{jk}}{\sum_{k=1}^{n} (u_{ik})^m}, \tag{1}$$

$$\beta_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^m b_{jk}}{\sum_{k=1}^{n} (u_{ik})^m}, \tag{2}$$

gdzie: α_{ij} – dolny (β_{ij} – górny) kraniec przedziału *j*-tej zmiennej $(j = 1, \ldots, p)$ w *i*-tej $(i = 1, \ldots, c)$ klasie.

$k = 1, \ldots, n$ – numer obiektu.

6. Dla obliczonych α_{ij}, β_{ij} oblicz stopień przynależności obiektów do klas zgodnie ze wzorem:

$$u_{ik} = \left[\sum_{h=1}^{n} \left(\frac{\sum_{j=1}^{p} \left[\left(a_{jk} - \alpha_{ji}\right)^2 + \left(b_{jk} - \beta_{ji}\right)^2 \right]}{\sum_{j=1}^{p} \left[\left(a_{jk} - \alpha_{jh}\right)^2 + \left(b_{jk} - \beta_{jh}\right)^2 \right]} \right)^{\frac{1}{m-1}} \right]^{-1}, \tag{3}$$

gdzie: a_{jk} – dolny (b_{jk} – górny) kraniec przedziału *j*-tej zmiennej $(j = 1, \ldots, p)$ w *k*-tym obiekcie;

pozostałe oznaczenia jak we wzorach (1) i (2).

7. Oblicz wartość funkcji – kryterium W, wykorzystując wzór:

$$W_t = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\left[\left(a_{jk}-\alpha_{ji}\right)^2+\left(b_{jk}-\beta_{ji}\right)^2\right], \tag{4}$$

gdzie: oznaczenia jak we wzorach (1), (2) i (3).

Jeżeli $\left|W_{t+1}-W_t\right| \le \varepsilon$ lub osiągnięto maksymalną ustaloną liczbę iteracji T, wówczas należy zakończyć działanie algorytmu, w przeciwnym przypadku przejdź do kroku 5, zwiększając liczbę dokonanych iteracji o jeden.

Do oceny jakości rozmytej klasyfikacji *k*-średnich dla danych symbolicznych zaproponowano miary heterogeniczności: R^1 i R^2 (*overall heterogeneity index*). Jednakże ze względu na sposób ich interpretacji oraz przyjmowany zakres wartości tych miar w artykule proponuje się nazwę **miary homogeniczności** R^1 i R^2. Miary te przyjmują wartości z zakresu [0; 1]. Im wyższe wartości tych miar, tym otrzymane klasy są bardziej homogeniczne, a reprezentanci klas w lepszy i pełniejszy sposób odzwierciedlają (reprezentują) obiekty znajdujące się w tych klasach [De Carvalho 2007, s. 428].

Miary homogeniczności R^1 i R^2 są obliczane zgodnie ze wzorami:

$$R^1 = \frac{B^1}{B^1+W^1}, \tag{5}$$

$$R^2 = \frac{B^2}{B^2+W^2}, \tag{6}$$

gdzie:

$$B^1 = \sum_{i=1}^{c} u_i \sum_{j=1}^{p}\left[\left(\alpha_{ji}-\alpha_j\right)^2+\left(\beta_{ji}-\beta_j\right)^2\right], \tag{7}$$

$$B^2 = \sum_{i=1}^{c} u_i \sum_{j=1}^{p}\lambda_{ij}\left[\left(\alpha_{ji}-\alpha_j\right)^2+\left(\beta_{ji}-\beta_j\right)^2\right], \tag{8}$$

$$W^1 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\left[\left(a_{kj}-\alpha_{ij}\right)^2+\left(b_{kj}-\beta_{ij}\right)^2\right], \tag{9}$$

$$W^2 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\lambda_{ij}\left[\left(a_{kj}-\alpha_{ij}\right)^2+\left(b_{kj}-\beta_{ij}\right)^2\right], \tag{10}$$

$$\lambda_{ij} = \frac{\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ik})^m\left[\left(a_{hk}-\alpha_{hi}\right)^2+\left(b_{hk}-\beta_{hi}\right)^2\right]\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n}(u_{ik})^m\left[\left(a_{hk}-\alpha_{hi}\right)^2+\left(b_{hk}-\beta_{hi}\right)^2\right]}, \; j=1,\ldots,p, \tag{11}$$

pozostałe oznaczenia jak we wzorach (1), (2), (3), (6) i (7).

Parametr λ_{ij} to wektor wag związanych z odległościami obiektów od prototypów klas. Parametr ten ma szczególne znaczenie dla **adaptacyjnej rozmytej klasyfikacji *k*-średnich dla danych interwałowych**, gdzie podlega obliczaniu w każdym kroku iteracyjnym, a jego zastosowanie ma na celu otrzymanie bardziej jednorodnych klas (zob. [De Carvalho 2007, s. 426-427]).

4. Przykład empiryczny

Zbiór pierwszy (**model I**) to dane pochodzące z programu SODAS (plik `CAR.SDS`) opisujące 33 marki samochodów 11 zmiennymi różnych typów. Do badania wybrano fragment zbioru danych (20 marek samochodów) oraz zmienne symboliczne interwałowe (m.in. cena w euro, przyspieszenie, długość, wysokość, szerokość, rozstaw osi).

Zbiór drugi (**model II**) to również dane pochodzące z programu SODAS (plik `ABALONE.SDS`) opisujący 24 gatunki ślimaków morskich z rodziny uchowców (*Haliotidae*). Zbiór opisywany jest siedmioma zmiennymi symbolicznymi interwałowymi (m.in. długość, średnica muszli, waga mięczaka).

Zbiór trzeci (**model III**) to 50 obiektów podzielonych na pięć niezbyt dobrze separowanych klas opisywanych przez dwie zmienne symboliczne interwałowe. Zmienne w tym zbiorze są losowane niezależnie z dwuwymiarowego rozkładu normalnego o średnich (5, 5), (–3, 3), (3, –3), (0, 0), (–5, –5) oraz macierzy kowariancji $\sum(\sigma_{jj}=1,\sigma_{jl}=-0,9)$. Zbiór ten wygenerowano z wykorzystaniem funkcji `cluster.Gen` z pakietu `clusterSim`. Model ten nie zawiera zmiennych zakłócających czy obserwacji odstających.

Klasyfikacji dokonano, przyjmując liczbę klas od 2 do 5 przy dwóch parametrach rozmycia 2 i 4[1]. Wyniki klasyfikacji (w sensie miar homogeniczności R^1 i R^2) zawarto w tab. 1 i 2.

Niezależnie od przyjętego w badaniu parametru rozmycia *m* otrzymano takie same wyniki (w sensie homogeniczności poszczególnych klas). Dla zbioru samochodów osobowych najlepszą strukturą jest struktura pięciu klas. W przypadku zbioru ślimaków morskich najlepszym podziałem jest podział na cztery klasy. Dla sztucznie wygenerowanego zbioru danych rozmyta klasyfikacja *k*-średnich dla danych symbolicznych wskazuje na strukturę pięciu klas. W przypadku tego modelu porównano wyniki klasyfikacji rozmytej ze znaną strukturą klas, przyjmując, że obiekt jest przydzielony do klasy o największym stopniu przynależności. Otrzymano w ten sposób trafność klasyfikacji na poziomie 0,74.

[1] Wielkość parametru rozmycia $m=2$ jest jedną z częściej wykorzystywanych w literaturze przedmiotu (por. [De Carvalho 2007; El-Sonbaty, Ismail 1998]). Parametr $m=4$ przyjęto celem sprawdzenia wpływu zmian jego wielkości na homogeniczność klas.

Tabela 1. Wartości miar homogeniczności w zależności od liczby klas ($m = 2$)

Lp.	Numer modelu	Liczba klas			
		2	*3*	*4*	*5*
R^1					
1	I	0,34	0,45	0,65	0,79
2	II	0,13	0,61	0,89	0,67
3	III	0,37	0,53	0,78	0,98
R^2					
1	I	0,37	0,50	0,71	0,83
2	II	0,16	0,66	0,88	0,72
3	III	0,43	0,57	0,81	0,99

Źródło: obliczenia własne z wykorzystaniem programu Excel.

Tabela 2. Wartości miar homogeniczności w zależności od liczby klas ($m = 4$)

Lp.	Numer modelu	Liczba klas			
		2	*3*	*4*	*5*
R^1					
1	I	0,22	0,31	0,44	0,64
2	II	0,06	0,20	0,63	0,43
3	III	0,12	0,26	0,40	0,75
R^2					
1	I	0,30	0,38	0,52	0,69
2	II	0,09	0,26	0,70	0,48
3	III	0,20	0,33	0,51	0,85

Źródło: obliczenia własne z wykorzystaniem programu Excel.

Miary oceny homogeniczności klas R^1 oraz R^2 wskazują w przypadku tych zbiorów danych oraz przyjętych parametrów m na podobną homogeniczność struktur klas.

5. Podsumowanie

Istotnym ograniczeniem rozmytej klasyfikacji *k*-średnich dla danych interwałowych jest fakt, że pozwala na analizowanie obiektów symbolicznych opisywanych wyłącznie zmiennymi interwałowymi. Pewne rozwiązanie tego ograniczenia jest proponowane w artykule autorów, takich jak Yang, Hwang, Chen [2004]. Drugim z ograniczeń rozmytej klasyfikacji *k*-średnich dla danych interwałowych jest fakt, że wykorzystuje w obliczeniach odległość euklidesową, lepszym rozwiązaniem w przypadku danych symbolicznych jest wykorzystanie miar odległości adekwatnych dla tego typu danych (np. De Carvalho, Hausdorffa czy Ichino i Yaguchiego).

Z przeprowadzonych badań wynika, że zwiększanie liczby klas oraz wielkości parametru rozmycia *m* prowadzi do spadku homogeniczności klas. Wyniki w podobnym brzmieniu są formułowane dla rozmytej klasyfikacji *k*-średnich dla danych klasycznych (por. [Lasek 2002, s. 146]).

Kierunkiem dalszych prac powinno stać się porównanie rozmytej klasyfikacji *k*-średnich dla danych symbolicznych oraz adaptacyjnej rozmytej klasyfikacji *k*-średnich dla danych symbolicznych z innymi metodami klasyfikacji rozmytej dla danych symbolicznych (np. metodą piramid).

Innym obszarem dalszych badań powinno stać się zbadanie skuteczności rozmytej klasyfikacji *k*-średnich dla danych symbolicznych w przypadku, gdy w zbiorze zmiennych znajdują się zmienne zakłócające, a w zbiorze danych obserwacje odstające.

Literatura

Bezdek J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York 1981.

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

De Carvalho F.A.T., *Fuzzy c-means clustering methods for symbolic interval data*, „Pattern Recognition Letters" 2007 vol. 28, Issue 4, s. 423-437.

Dunn J.C., *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters*, „Journal of Cybernetics" 1973 no 3, s. 32-57.

El-Sonbaty Y., Ismail M.A., *Fuzzy clustering for symbolic data*, „IEEE Transactions on Fuzzy Systems" 1998 vol. 6, no 2, s. 195-204.

Lasek M., *Data mining. Zastosowanie w analizach i ocenach klientów bankowych*, Biblioteka Menedżera i Bankowca, Warszawa 2002.

Milligan G.W., *Clustering Validation: Results and Implications for Applied Analyses*, [w:] *Clustering and Classification*, P. Arabie, L.J. Hubert, G. de Soete (red.), World Scientific, Singapore 1996, s. 341-375.

Pełka M., *Porównanie strategii klasyfikacji danych symbolicznych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu (w druku), 2009.

Verde R., *Clustering Methods in Symbolic Data Analysis*, [w:] *Classification, Clustering and Data Mining Applications*, D. Banks, L. House, E.R. McMorris, P. Arabie, W. Gaul (red.), Springer-Verlag, Heidelberg 2004, s. 299-317.

Yang M.S., Hwang P.Y., Chen D.H., *Fuzzy clustering algorithms for mixed feature variables*, „Fuzzy Sets and Systems" 2004 vol. 141, Issue 2, s. 301-317.

FUZZY *C*-MEANS CLUSTERING FOR SYMBOLIC DATA

Summary: This paper introduces adaptive and non-adaptive fuzzy c-means clustering methods for symbolic data. Both methods are suitable only for interval-valued symbolic data. The article also presents types of symbolic variables.

In the empirical part of the paper, non-adaptive fuzzy c-means clustering method was applied to exemplary symbolic data.