PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Jerzy Korzeniewski
Uniwersytet Łódzki

PROPOZYCJA MODYFIKACJI DOWOLNEGO INDEKSU WYZNACZAJĄCEGO LICZBĘ SKUPIEŃ W ZBIORZE DANYCH

1. Wstęp

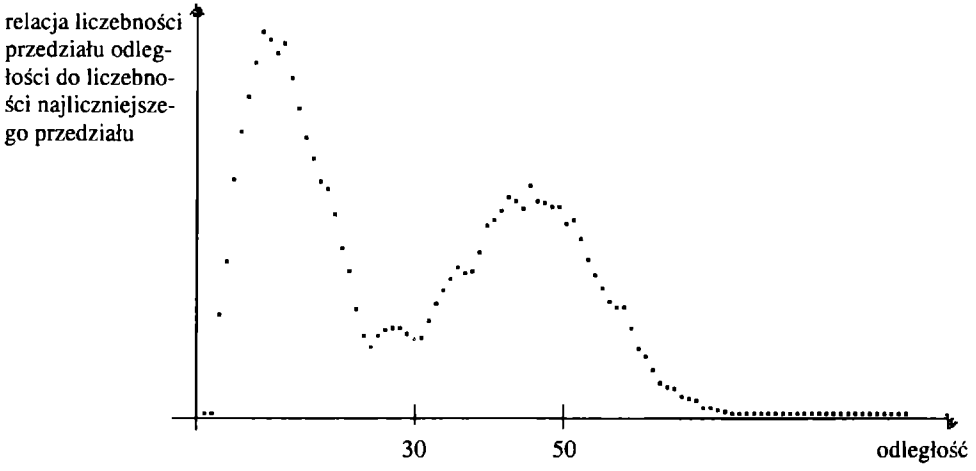
Artykuł zawiera propozycję metody wyznaczenia minimalnej liczby skupień w zbiorze danych. Kiedy znamy taką liczbę, możemy zmodyfikować dowolny indeks, uznając za właściwą liczbę skupień tę, którą wskazuje indeks pod warunkiem, że jest ona nie mniejsza od znalezionej minimalnej liczby skupień. Jeśli jest mniejsza, to za właściwą liczbę skupień uznajemy znalezioną minimalna liczbę skupień. Metoda oparta jest na analizie rozkładu odległości pomiędzy dwoma elementami zbioru danych i wymaga tylko znajomości wszystkich par odległości w danym zbiorze, czyli macierzy odległości. W tekście przedstawiono sformułowanie metody oraz jej ocenę za pomocą eksperymentu, w którym badany jest odsetek poprawionych wskazań dla zbiorów z przestrzeni euklidesowych wygenerowanych z zastosowaniem programu CLUSTGEN.

2. Sformułowanie metody

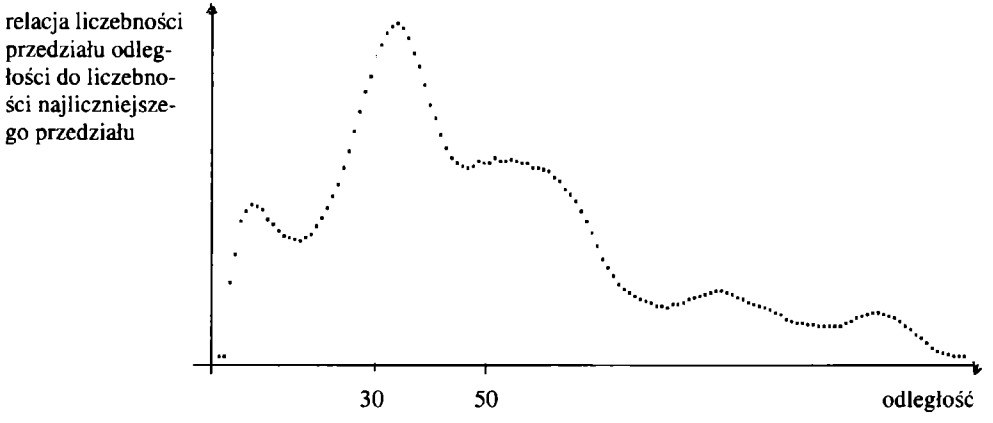
W celu uzasadnienia idei metody rozważmy rozkłady odległości (por. rys. 2 i rys. 3) pomiędzy parami punktów dla dwóch przykładowych zbiorów punktów z przestrzeni R^2 (por. rys. 1). Na osi poziomej zaznaczone są odległości. Każdej odległości odpowiada (na osi pionowej) pewna liczebność względna, która odnosi się do przedziału o szerokości dwóch jednostek. Na przykład punktowi o wartości 50 odpowiada liczba odległości pomiędzy parami punktów, które mają wartość z przedziału od 48 do 50. Na osi pionowej podane są relacje liczebności odległości,

które znalazły się w danym przedziale, do liczebności przedziału najliczniejszego (w największym maksimum lokalnym relacja ta ma wtedy wartość 1). Na rysunku 2 widoczne są wyraźne dwa maksima lokalne występujące w rozkładzie odległości par punktów, na rys. 3 zaś – cztery maksima lub pięć maksimów. Skąd biorą się te maksima? Jeśli weźmiemy pod uwagę typowy punkt zbioru, tzn. należący do któregoś ze skupień, to pierwsze maksimum lokalne tworzą odległości do innych punktów z tego samego skupienia. Następnie wraz ze zwiększaniem się odległości na osi poziomej częstość występowania takich odległości maleje, do momentu napotkania większej grupy punktów z innego skupienia. Gdyby odległości pomiędzy poszczególnymi wyraźnie oddzielonymi skupieniami były parami różne, to na wykresie liczba maksimów lokalnych byłaby równa liczbie skupień. Tak oczywiście być nie musi, jak widać na przykładzie zbioru nr 2. Składa się on z ośmiu skupień, a na wykresie częstości odległości są co najwyżej cztery (może pięć) wyraźne maksima lokalne. Powodem jest to, że odległości pomiędzy niektórymi parami skupień są bardzo podobne. Oczywiste wydaje się jednak to, że liczba skupień w zbiorze danych nie może być większa od liczby maksimów lokalnych na wykresie rozkładu częstości odległości pomiędzy parami punktów. Modyfikacja dowolnego indeksu wyznaczającego liczbę skupień w zbiorze danych, jaką chcemy zaproponować w tym artykule, jest następująca: liczba skupień w zbiorze danych powinna być taka, na jaką wskazuje indeks, przy czym nie mniejsza od liczby maksimów lokalnych rozkładu częstości odległości pomiędzy parami punktów.

Rys. 1. Dwa przykładowe zbiory punktów z przestrzeni dwuwymiarowej: zbiór nr 1 (po lewej) z dwoma skupieniami, zbiór nr 2 (po prawej) z ośmioma skupieniami
Źródło: opracowanie własne.



Rys. 2. Rozkład częstości odległości pomiędzy parami punktów dla zbioru nr 1
Źródło: opracowanie własne.



Rys. 3. Rozkład częstości odległości pomiędzy parami punktów dla zbioru nr 2
Źródło: opracowanie własne.

Przy takiej modyfikacji problemem zasadniczym staje się zdefiniowanie maksimum lokalnego badanej krzywej. Jeżeli przyjmiemy zbyt mały skok odległości (rozpiętość przedziału) na osi poziomej, to zazwyczaj otrzymamy zbyt dużo maksimów lokalnych. Jeżeli zaś skok odległości będzie zbyt duży, to możemy otrzymać zbyt mało maksimów lokalnych i nie uzyskamy żadnej poprawy efektywności indeksów. Postępując podobnie jak w przypadku różnych reguł doboru wartości parametru wygładzania, możemy posługiwać się takimi parametrami, jak rozstęp kwartylowy lub mediana rozkładu odległości pomiędzy parami punktów zbioru. Podobnych problemów może nastręczać „wyrazistość" maksimum lokalnego,

tzn. czy np. dwa maksima lokalne znajdujące się blisko siebie traktować jako jedno maksimum czy jako dwa maksima.

W celu ustalenia konkretnej liczby maksimów lokalnych metodą prób i błędów dobrana została następująca procedura.

1. Szerokość przedziału odległości (skok na osi odciętych) przyjęto równą 1/15 rozstępu kwartylowego rozkładu odległości.
2. Dla każdego przedziału odległości znajdujemy liczbę par punktów, dla których odległość należy do tego przedziału. Otrzymujemy krzywą częstości rozkładu odległości pomiędzy parami punktów.
3. Wygładzamy krzywą częstości za pomocą średniej ruchomej pięciookresowej (np. częstość w przedziale odległości (48,50) jest 1/5 sumy częstości z przedziałów (44,46), (46,48), (48,50), (50,52), (52,54)).
4. Przyjmujemy, że maksimum lokalne istnieje w danym przedziale odległości, jeżeli częstość tego przedziału jest ostro większa od częstości przedziałów sąsiednich i nie mniejsza od częstości przedziałów oddalonych o dwa przedziały.

3. Eksperyment badawczy

Założenia eksperymentu

W celu dokonania oceny zaproponowanej modyfikacji do badania wybrano te same indeksy wyznaczające liczbę skupień w zbiorze danych jak w pracy [Sugar, James 2003]. Poniżej podane są formuły na te indeksy. W tych wzorach B oznacza wariancję pomiędzy skupieniami (międzygrupową), W zaś – wariancję wewnątrzgrupową, d – wymiar przestrzeni euklidesowej, w której znajdują się punkty zbioru, K to badana liczba skupień, n – liczebność zbioru danych. Pierwszy indeks to indeks Calińskiego-Harabasza. Według niego należy wybrać taką liczbę skupień K, która maksymalizuje wielkość:

$$CH(K)=\frac{B(K)/(K-1)}{W(K)/(n-K)}. \tag{1}$$

Drugi indeks to indeks Krzanowskiego-Lai

$$KL(K)=\left|\frac{DIFF(K)}{DIFF(K+1)}\right|, \tag{2}$$

gdzie

$$DIFF(K)=(K-1)^{2/d}W(K-1)-K^{2/d}W(K), \tag{3}$$

przy czym znów wybrać należy K, które maksymalizuje wielkość daną wzorem 2.

Trzeci indeks to indeks Hartigana:

$$H(K)=(n-K-1)\left(\frac{W(K)}{W(K+1)}-1\right), \tag{4}$$

przy którym wybieramy najmniejsze K, dla którego wielkość dana wzorem 4 jest mniejsza lub równa 10.

Czwarty indeks to indeks sylwetkowy Rousseeuwa dany wzorem

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \tag{5}$$

gdzie $a(i)$ jest średnią odległością pomiędzy i-tym elementem a wszystkimi innymi punktami należącymi do tego samego skupienia, $b(i)$ jest zaś średnią odległością do najbliższego skupienia. Wybieramy K, które maksymalizuje średnią dla całego zbioru wielkość daną wzorem 5.

Za pomocą programu CLUSTGEN, autorstwa J. Milligana, (por. [Milligan 1985]), wygenerowanych zostało 216 stuelementowych zbiorów danych. Zbiory były podzielone równo po 72 z każdej z przestrzeni R^4, R^6 oraz R^8. Równy był też podział ze względu na liczbę skupień w każdym zbiorze, tzn. 54 zbiory z 2 skupieniami, 54 z 3 skupieniami, 54 z 4 skupieniami i 54 zbiory z 5 skupieniami. Każdy ze zbiorów został 100 razy pogrupowany w rozłączne klasy metodą k-średnich (k = 2, 3, ..., 11) z losowo wybranymi punktami startowymi. Dla każdego grupowania znalezione zostały badane 4 indeksy i za wskazanie końcowe przyjęto najczęściej występującą (dla danego indeksu) liczbę skupień. Wskazanie to zostało porównane ze wskazaniem wynikającym z zaproponowanej modyfikacji.

Wyniki eksperymentu

Modyfikacja raczej nie wpływa na indeksy Calinskiego-Harabasza i Hartigana. Indeksy te prawie zawsze pokazują za dużo skupień. Wskazania poprawnej liczby skupień tych indeksów były słabe – poniżej 15 % (odsetka poprawnych wskazań indeksów nie ma w tab. 1).

Tabela 1. Liczby niezmienionych, lepszych oraz gorszych wskazań liczby skupień indeksów zmodyfikowanych

Wymiar przestrzeni euklidesowej	Względna jakość indeksu zmodyfikowanego	Zbiory danych z liczbą skupień 2				Zbiory danych z liczbą skupień 3				Zbiory danych z liczbą skupień 4				Zbiory danych z liczbą skupień 5			
		S	CH	H	KL	S	CH	H	KL	S	CH	H	KL	S	CH	H	KL
R^4	bez zmiany	12	17	18	14	4	17	16	7	16	18	18	17	18	18	18	18
	wskazanie lepsze	0	0	0	0	13	1	2	9	2	0	0	1	0	0	0	0
	wskazanie gorsze	6	1	0	4	1	0	0	2	0	0	0	0	0	0	0	0
R^6	bez zmiany	13	17	17	14	6	16	17	8	17	18	18	16	18	18	18	18
	wskazanie lepsze	0	0	0	0	12	2	1	9	1	0	0	2	0	0	0	0
	wskazanie gorsze	5	1	1	4	0	0	0	1	0	0	0	0	0	0	0	0
R^8	bez zmiany	14	16	17	15	2	17	17	10	16	18	18	18	18	18	18	18
	wskazanie lepsze	0	0	0	0	14	1	1	8	2	0	0	0	0	0	0	0
	wskazanie gorsze	4	2	1	3	2	0	0	2	0	0	0	0	0	0	0	0

Źródło: obliczenia własne.

Wskazania indeksu Krzanowskiego-Lai (poprawne w ok. 25%) zostały pogorszone w ok. 7% przypadków oraz poprawione w ok. 14% przypadków. Pogorszenia miały miejsce tylko w odniesieniu do zbiorów z dwoma skupieniami. Wtedy zaproponowana modyfikacja wskazywała o jedno skupienie za dużo. Poprawienia wskazań miały miejsce głównie w przypadku zbiorów z trzema skupieniami, wtedy indeks wskazywał dwa skupienia, modyfikacja zaś wskazywała trzy.

Najlepszym indeksem w tym badaniu okazał się indeks sylwetkowy, który wskazywał poprawną liczbę skupień w ok. 32% zbiorów. Wskazania tego indeksu zostały pogorszone w ok. 8% przypadków oraz poprawione w 25% przypadków. Podobnie jak w przypadku indeksu Krzanowskiego-Lai pogorszenia miały miejsce tylko w odniesieniu do zbiorów z dwoma skupieniami (modyfikacja wskazywała o jedno skupienie za dużo), lepsze wskazania zaś – w przypadku zbiorów z trzema skupieniami.

4. Wnioski

Ogólnie zaproponowana modyfikacja dla badanych czterech indeksów liczby skupień w zbiorze danych poprawia wskazania tych indeksów (przynajmniej najlepszych z nich). Tak jak można się było spodziewać, modyfikacja nie wnosi wiele nowego, gdy liczba skupień w zbiorze jest dość duża – wtedy odległości pomiędzy niektórymi skupieniami są podobne i przy badaniu rozkładu odległości pomiędzy parami punktów trudno te skupienia rozróżnić. Trzeba oczywiście pamiętać o tym, że przeprowadzone badanie dotyczyło tylko grupowania elementów zbioru metodą k-średnich. W celu uzyskania pełniejszej oceny należałoby zastosować, być może, inne metody grupowania. Nie należy się łudzić, że samo analizowanie macierzy odległości da ostateczną i poprawną odpowiedź na pytanie o liczbę skupień w zbiorze danych, ale taka analiza (jak pokazuje zaproponowana modyfikacja) może być pomocna. Wydaje się, że analizowanie macierzy odległości (pod względem rozkładu odległości pomiędzy parami punktów) może wspomagać również metody klasyfikacyjne – badanie tego zagadnienia będzie prowadzone w najbliższej przyszłości.

Literatura

Gordon A.D. (1999), *Classification,* Chapman & Hall, London.

Milligan G.W. (1985), *An Algorithm for Generating Artificial Test Clusters*, „Psychometrika" vol. 50, nr 1, s. 123-127.

Sugar C.A., James G.M. (2003), *Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach*, „JASA" vol. 98.

A PROPOSAL OF MODIFICATION OF ARBITRARY DATA SET CLUSTER NUMBER INDEX

Summary

In the paper a proposal of the modification of an arbitrary data set cluster number index is given. The idea of the modification is to obtain a lower bound of the number of clusters in a data set through analysing the distribution of pairwise distances. The modification was tested on 216 data sets from Euclidean spaces with the grouping done by the k-means method. The indices modified were the Rousseeuw silhouette index, the Krzanowski-Lai index, the Caliński-Harabasz index and the Hartigan index.