PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Jerzy Korzeniewski
Uniwersytet Łódzki

OCENA PORÓWNAWCZA KILKU METOD WYZNACZANIA LICZBY SKUPIEŃ W ZBIORZE DANYCH

1. Wstęp

Poniższy artykuł jest kontynuacją pracy [Korzeniewski 2005], która dotyczyła tego samego algorytmu. W niniejszej pracy zaproponowana została nowa wersja algorytmu, w której liczba skupień jest ustalana za pomocą miernika liczbowego opartego na statystyce χ^2; przeprowadzony został też eksperyment oceniający algorytm w porównaniu z innymi wybranymi metodami.

2. Sformułowanie algorytmu

Algorytm wykorzystuje metodę średniego przesunięcia okna/próby, zaproponowaną przez Comaniciu i Meera [2000]. Idea tej metody jest następująca. Niech $\{x_i\}_{i=1,\ldots,n}$, będzie dowolnym zbiorem n punktów z d-wymiarowej przestrzeni euklidesowej. Wielkość

$$M_h(x) = \left(1/n_x \sum_{x_i \in S_h(x)} [x_i - x] \right) = 1/n_x \sum_{x_i \in S_h(x)} x_i - x, \qquad (1)$$

gdzie $S_h(x)$ jest kulą o środku w punkcie x i promieniu h, n_x, jest liczbą punktów zbioru zawartych w tej kuli i nazywa się średnim przesunięciem okna/próby (por. *Korzeniewski* (2005)). Promień kuli h nazywamy szerokością okna/próby. Dla dowolnego punktu ze zbioru danych odpowiadającym mu punktem granicznym przy szerokości okna h nazywamy granicę ciągu środków okien (czyli środków ciężkości próby) występujących po prawej stronie wzoru (1).

Formalne sformułowanie algorytmu jest następujące. Krokiem wstępnym jest przybliżenie mediany odległości pomiędzy dwoma różnymi punktami zbioru danych za pomocą mediany 300 par odległości punktów wylosowanych ze zbioru. Liczba powtórzeń równa 300 daje zawsze bardzo podobne wartości mediany, gdy odległości pomiędzy punktami należącymi do tego samego skupienia są tego samego rzędu co odległości pomiędzy punktami należącymi do różnych skupień. W przypadku gdy te dwa rodzaje odległości znacznie odbiegają od siebie oraz gdy ponadto rozkład odległości par punktów jest symetryczny dwumodalny, to oszacowania mediany takiego rozkładu są mniej stabilne, ale wystarczające na potrzeby naszego algorytmu.

1. Dla $k = 2$ losujemy w sposób zależny 2 punkty ze zbioru danych i dla każdego punktu znajdujemy odpowiadający mu punkt graniczny w procedurze średniego przesunięcia dla ustalonej szerokości okna h.

2. Sprawdzamy, czy wśród wszystkich par utworzonych spośród k punktów granicznych (dla $k = 2$ będzie tylko jedna para) istnieje choć jedna, dla której odległość jest mniejsza od h.

3. Pierwsze dwa kroki powtarzamy 2000 razy w celu ustalenia prawdopodobieństwa spełnienia warunku z kroku drugiego.

4. Pierwsze trzy kroki powtarzamy dla wszystkich możliwych szerokości okna h (ustalając pewien mały przyrost Δh równy np. najmniejszej odległości pomiędzy dwoma różnymi punktami zbioru) z zakresu od zera do największej odległości w badanym zbiorze punktów. Otrzymujemy w ten sposób dla $k = 2$ funkcję o wartościach z przedziału (0;1), zależną od h.

5. Pierwsze cztery kroki powtarzamy kolejno dla $k = 3, 4, 5 \ldots$.

6. Dla każdego k i każdego h obliczamy wartość statystyki χ^2, służącej do weryfikowania hipotezy o równości kilku wskaźników struktury dla 10 kolejnych przyrostów Δh . Uznajemy, że funkcja posiada fazę poziomą, tym samym zbiór ma k skupień, jeżeli wartość statystyki jest mniejsza od 20 dla przynajmniej jednej wartości h. Średnią arytmetyczną 10 wskaźników struktury, dla których obliczona była statystyka χ^2, nazywamy wysokością, na której znajduje się faza pozioma.

7. Poszukiwanie faz poziomych, czyli tym samym skupień, kończymy na tym k, dla którego wysokość, na której znajduje się faza pozioma jest większa od liczby $w(k)$.

Wartość Δh w poniżej przeprowadzonym eksperymencie porównawczym przyjęto równą 1/100 obliczonej mediany zbioru odległości par punktów. Natomiast wartości $w(k)$, decydujące o zakończeniu poszukiwania większej od dotychczasowej liczby skupień, dobrane zostały tak, by liczebność każdego spośród ewentualnych k skupień była nie mniejsza od przyjętej liczebności minimalnej skupienia. Liczebność minimalną skupienia przyjęto równą 5% liczebności całego zbioru. Warunek ten pozwala wyliczyć wartości $w(k)$ przy dodatkowym założeniu, że istnieje jedno skupienie stanowiące 5% zbioru, liczebności zaś pozostałych $k - 1$ skupień są mniej więcej jednakowe. Dla zbiorów o liczebności równej 100 $w(k)$ jest równe 0,92, 0,93, 0,96, 0,98, 0,99 dla k równego odpowiednio 2, 3, 4, 5, 6. Ze

względu na prawdopodobieństwa tego rzędu liczbę powtórzeń w kroku trzecim przyjęto równą 2000, ponieważ przy mniejszej porównywanie odsetków bliskich 1 jest niemożliwe.

3. Eksperyment porównawczy

Założenia eksperymentu. W celu porównania algorytmu z innymi metodami wyznaczania liczby skupień w zbiorze danych wybrano następujące metody (por. [Sugar, James 2003]). W poniższych wzorach B oznacza wariancję pomiędzy skupieniami (międzygrupową), W zaś wariancję wewnątrzgrupową, d – wymiar przestrzeni euklidesowej, w której znajdują się punkty zbioru, K – badaną liczbę skupień, n – liczebność zbioru danych. Pierwsza metoda to indeks Calińskiego–Harabasza. Według niego, należy wybrać taką liczbę skupień K, która maksymalizuje wielkość

$$CH(K)=\frac{B(K)/(K-1)}{W(K)/(n-K)}. \tag{2}$$

Druga metoda to indeks Krzanowskiego–Lai

$$KL(K)=\left|\frac{DIFF(K)}{DIFF(K+1)}\right|, \tag{3}$$

gdzie

$$DIFF(K)=(K-1)^{2/d}W(K-1)-K^{2/d}W(K), \tag{4}$$

przy czym znów wybrać należy K, które maksymalizuje wielkość daną wzorem (3).

Trzecia metoda to indeks Hartigana

$$H(K)=(n-K-1)\left(\frac{W(K)}{W(K+1)}-1\right), \tag{5}$$

przy którym wybieramy najmniejsze K, dla którego wielkość dana wzorem (5) jest mniejsza lub równa 10.

Czwarta metoda to indeks sylwetkowy Rousseeuwa, dany wzorem

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \tag{6}$$

gdzie $a(i)$ jest średnią odległością pomiędzy i-tym elementem a wszystkimi innymi punktami należącymi do tego samego skupienia, $b(i)$ zaś jest średnią odległością do najbliższego skupienia. Wybieramy K, które maksymalizuje średnią dla całego zbioru wielkość daną wzorem (6).

Powyższe metody różnią się w sposób zasadniczy od zaproponowanego algorytmu, wymagają bowiem uprzedniego pogrupowania zbioru danych w skupienia.

W związku z tym metody te zostały zastosowane do dwóch różnych w swej idei metod grupowania, tj. do metody *k* średnich (z losowym wyborem punktów startowych) oraz do algorytmu aglomeracyjnego średniego łączenia. Ta ostatnia metoda polega na tym, że w każdym kroku aglomeracji łączymy dwa skupienia (lub punkty), których środki ciężkości są najbliższe.

Za pomocą programu CLUSTGEN, autorstwa Milligana (por. [Milligan 1985], źródło http://www.pitt.edu/~csna/Milligan/readme.html), wygenerowano po 160 zbiorów danych składających się ze 100 elementów w każdej z przestrzeni R^4, R^6, R^8, po 40 zbiorów o 2, 3, 4, 5 skupieniach. Następnie wygenerowano po 160 zbiorów danych składających się ze 120 elementów w każdej z przestrzeni R^4, R^6, R^8, po 40 zbiorów o 2, 3, 4, 5 skupieniach, przy czym 100 elementów każdego zbioru było takich samych, jak poprzednio natomiast 20 dodatkowych elementów miało charakter zanieczyszczający.

Tabela 1. Liczby skupień pokazywane przez zaproponowany algorytm oraz częstości ich występowania dla zbiorów 100-elementowych z przestrzeni euklidesowych o różnych wymiarach

Wymiar przestrzeni euklidesowej	Zbiory danych z liczbą skupień 2	Zbiory danych z liczbą skupień 3	Zbiory danych z liczbą skupień 4	Zbiory danych z liczbą skupień 5
R^4	**2 sk.~78** 3 sk.~22	**3 sk.~100**	**4 sk.~75** 3 sk.~25	**5 sk.~44** 4 sk.~56
R^6	**2 sk.~100**	**3 sk.~100**	**4 sk.~48** 3, 5 sk.~52	**5 sk.~66** 4 sk.~34
R^8	**2 sk.~100**	**3 sk.~100**	**4 sk.~100**	**5 sk.~66** 4, 6 sk.~34

Źródło: obliczenia własne.

Tabela 2. Liczby skupień pokazywane przez zaproponowany algorytm oraz częstości ich występowania dla zbiorów 120-elementowych z przestrzeni euklidesowych o różnych wymiarach

Wymiar przestrzeni euklidesowej	Zbiory danych z liczbą skupień 2	Zbiory danych z liczbą skupień 3	Zbiory danych z liczbą skupień 4	Zbiory danych z liczbą skupień 5
R^4	3 sk.~100	4 sk.~100	**4 sk.~22** 5 sk.~78	**5 sk.~12** 3, 4 sk.~56 6 sk.~34
R^6	3 sk.~100	**3 sk.~10** 2 sk.~10 4 sk.~80	**4 sk.~34** 5 sk.~66	**5 sk.~46** 6 sk.~54
R^8	3 sk.~100	**3 sk.~34** 4 sk.~66	**4 sk.~12** 5 sk.~88	**5 sk.~12** 6 sk.~88

Źródło: obliczenia własne.

Wyniki eksperymentu. Wyniki dla zaproponowanego algorytmu przedstawiają tab. 1 i 2 natomiast dla pozostałych czterech metod – tab. 3, 4, 5, 6. W polach tab. 1 oraz tab. 2 podano dla każdego rodzaju zbioru liczby skupień wskazywane przez algorytm, a obok nich wyrażoną w procentach częstość występowania tych liczb.

Tabele 3-6 skonstruowane są nieco inaczej, podano w nich częstości występowania wskazań prawidłowej liczby skupień, liczby różnej od prawidłowej o 1 oraz liczby różnej od prawidłowej o więcej niż 1. W tabelach 3 oraz 4 grupowanie punktów przeprowadzone było metodą aglomeracyjną średniego łączenia, natomiast w tab. 5 oraz 6 metodą *k*-średnich. Ponadto należy zaznaczyć, że w tab. 1, 3, 5 rozważano te same zbiory 100-elementowe, natomiast w tabelach 2, 4, 6 – te same zbiory 120-elementowe.

Tabela 3. Podane w procentach prawidłowo (OK) oraz błędnie (Bł.=1, Bł.>1) wskazane częstości występowania liczb skupień pokazywanych przez cztery oceniane metody dla zbiorów 100-elementowych

Wymiar przestrzeni euklidesowej	Wielkość błędu	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	OK.	**65**	**66**	**100**	**0**	**0**	**34**	**50**	**2**	**2**	**34**	**0**	**0**	**0**	**0**	**16**	**0**
	Bł.=1	2	16	0	0	50	32	50	28	2	16	16	50	2	18	4	50
	Bł.>1	33	18	0	100	50	34	0	70	96	50	84	50	98	82	80	50
R^6	OK.	**34**	**50**	**100**	**18**	**18**	**32**	**34**	**30**	**0**	**50**	**15**	**0**	**0**	**0**	**0**	**80**
	Bł.=1	0	0	0	18	34	50	66	56	0	18	0	68	0	34	16	4
	Bł.>1	64	50	0	64	50	18	0	14	100	32	85	32	100	66	84	16
R^8	OK.	**16**	**50**	**68**	**16**	**16**	**16**	**34**	**16**	**0**	**0**	**0**	**18**	**0**	**32**	**18**	**32**
	Bł.=1	18	16	32	50	50	50	66	34	2	50	50	0	0	18	0	32
	Bł.>1	66	34	0	34	34	34	0	50	98	50	50	82	100	50	82	36

Źródło: obliczenia własne.

Tabela 4. Podane w procentach prawidłowo (OK) oraz błędnie (Bł.=1, Bł.>1) wskazane częstości występowania liczb skupień pokazywanych przez cztery oceniane metody dla zbiorów 120-elementowych

Wymiar przestrzeni euklidesowej	Wielkość błędu	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	OK.	**64**	**67**	**100**	**18**	**33**	**32**	**32**	**15**	**0**	**0**	**32**	**35**	**0**	**15**	**0**	**38**
	Bł.=1	2	17	0	0	35	50	68	35	5	68	18	15	0	35	0	12
	Bł.>1	34	17	0	82	32	18	0	50	95	32	50	50	100	50	100	50
R^6	OK.	**16**	**50**	**95**	**0**	**0**	**17**	**18**	**0**	**0**	**0**	**0**	**35**	**0**	**0**	**0**	**12**
	Bł.=1	2	0	5	32	68	67	82	50	0	0	0	50	0	35	32	38
	Bł.>1	82	50	0	68	32	17	0	50	100	100	100	15	100	65	68	50
R^8	OK.	**18**	**32**	**82**	**67**	**0**	**17**	**32**	**17**	**0**	**18**	**0**	**0**	**0**	**15**	**0**	**0**
	Bł.=1	16	18	18	15	68	17	68	50	0	0	0	18	5	35	0	50
	Bł.>1	66	50	0	18	32	67	0	33	100	82	100	82	95	50	100	50

Źródło: obliczenia własne.

Tabela 5. Podane w procentach prawidłowo (OK) oraz błędnie (Bł.=1, Bł.>1) wskazane częstości występowania liczb skupień pokazywanych przez cztery oceniane metody dla zbiorów 100-elementowych

Wymiar przestrzeni euklidesowej	Wielkość błędu	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	OK.	**65**	**22**	**0**	**10**	**56**	**55**	**0**	**34**	**34**	**20**	**0**	**32**	**22**	**22**	**0**	**34**
	Bł.=1	20	0	0	22	34	10	2	22	10	36	22	34	32	78	0	22
	Bł.>1	15	78	100	68	10	35	98	44	56	44	78	34	46	0	100	44
R^6	OK.	**88**	**44**	**12**	**12**	**0**	**12**	**10**	**12**	**34**	**22**	**2**	**10**	**10**	**10**	**0**	**22**
	Bł.=1	0	12	22	10	100	44	12	56	44	22	0	56	68	65	10	22
	Bł.>1	12	44	68	78	0	44	78	34	22	56	98	34	22	25	90	56
R^8	OK.	**78**	**56**	**12**	**22**	**32**	**34**	**0**	**0**	**22**	**22**	**0**	**22**	**12**	**32**	**10**	**12**
	Bł.=1	0	0	0	22	66	22	22	44	44	44	10	46	44	68	10	44
	Bł.>1	22	44	88	56	2	44	78	56	34	34	90	32	44	0	80	44

Źródło: obliczenia własne.

Tabela 6. Podane w procentach prawidłowo (OK) oraz błędnie (Bł.=1, Bł.>1) wskazane częstości występowania liczb skupień pokazywanych przez cztery oceniane metody dla zbiorów 120- elementowych

Wymiar przestrzeni euklidesowej	Wielkość błędu	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	OK.	**90**	**12**	**0**	**0**	**22**	**0**	**2**	**12**	**10**	**44**	**0**	**34**	**0**	**32**	**2**	**22**
	Bł.=1	2	10	2	32	68	22	2	44	34	22	5	44	10	56	0	22
	Bł.>1	8	78	98	68	10	78	96	44	56	34	95	22	90	12	98	56
R^6	OK.	**56**	**22**	**2**	**12**	**22**	**22**	**0**	**34**	**22**	**12**	**0**	**10**	**22**	**22**	**10**	**44**
	Bł.=1	32	0	0	10	56	34	5	44	34	56	10	56	12	44	0	22
	Bł.>1	12	78	98	78	22	44	95	22	44	32	90	34	68	34	90	34
R^8	OK.	**44**	**30**	**10**	**22**	**0**	**10**	**0**	**10**	**22**	**12**	**0**	**22**	**22**	**34**	**10**	**12**
	Bł.=1	22	5	0	22	78	22	10	34	34	32	12	44	44	44	22	44
	Bł.>1	34	65	90	56	22	68	90	56	44	56	88	34	34	22	68	44

Źródło: obliczenia własne.

4. Wnioski końcowe

Zaproponowany algorytm należy ocenić jako dobry. W przypadku zbiorów niezanieczyszczonych algorytm na ogół poprawnie pokazuje liczbę skupień – gdy pokazuje błędną liczbę skupień, to błąd nie przekracza 1. W przypadku zbiorów zanieczyszczonych algorytm pokazuje na ogół o jedno skupienie za dużo. To może wynikać stąd, że dodatkowo generowane punkty zanieczyszczające wokół istniejących skupień (tak działa program Milligana) mogą formować nowe, niewielkie skupienia i algorytm je odnajduje. Dużo słabiej spisały się pozostałe metody – wybrane działają

dobrze w nielicznych przypadkach. Fakt ten wynika zapewne z niedoskonałości zastosowanych metod grupowania. Aglomeracja, jak wiadomo, posiada wadę łańcucha, a metoda *k*-średnich z losowymi punktami startu też nie spisuje się dobrze (np. w sensie odsetka punktów z dodatnimi indeksami *silhouette*). Jednak wadą, którą potwierdzają również inne badania, jest wysoka niestabilność tych indeksów względem zastosowanej metody grupowania danych w skupienia. Bardzo często oceniane indeksy wykazują przy ustalaniu liczby skupień błąd większy od 1. Wydaje się, że poszukiwanie metod wyznaczania liczby skupień w zbiorze danych niezależnie od metody grupowania elementów zbioru ma sens.

Literatura

Comaniciu D., Meer P. (1999), *Mean Shift Analysis and Applications*, IEEE Int. Conf. Computer Vision (ICCV'99), Kerkyra, Greece, s. 1197-1203.

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

Korzeniewski J. (2005), *Propozycja nowego algorytmu wyznaczającego liczbę skupień, Klasyfikacja i analiza danych – teoria i zastosowani*, Taksonomia 12, AE, Wrocław.

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.

PERFORMANCE COMPARISON OF SOME METHODS OF DETERMINING THE NUMBER OF CLUSTERS IN A DATA SET

Summary

This paper is an attempt to compare the performance of an algorithm for determining the number of clusters in a data set proposed by the author with other methods of determining the number o clusters. The other methods include the Rousseeuw silhouette index, the Krzanowski-Lai index, the Caliński-Harabasz index and the Hartigan index. The comparison was carried out on the Euclidean space sets generated by the Milligan's CLUSTGEN program in 4,6 and 8 dimensions. The algorithm proposed does not depend on the grouping of set's point therefore the other methods were applied to two different in nature grouping methods i.e. the *k*-means method and the average linkage agglomeration method. The results seem to prove the new algorithm's usefulness.