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

Kamila Migdał-Najman, Krzysztof Najman
Uniwersytet Gdański

ANALITYCZNE METODY USTALANIA LICZBY SKUPIEŃ W ROZMYTYCH ZBIORACH DANYCH

1. Wstęp

Jednym z podstawowych problemów pojawiających się w analizie skupień, jest ustalenie liczby grup, na jaką należy rozdzielić badane obiekty. Jest to zadanie szczególnie trudne, jeżeli klasyfikujemy dziesiątki tysięcy obiektów opisanych kilkudziesięcioma parametrami, co w dobie gwałtownej komputeryzacji wszelkiej ludzkiej działalności nie jest rzadkością. Jedną z wyjątkowo skutecznych metod grupowania takich danych są sztuczne sieci neuronowe typu SOM (Self Organizing Map), por. [Jurkiewicz, Najman 2003]. Szczególną własnością sieci SOM jest możliwość redukcji wymiarowości badanego zagadnienia do dwóch wymiarów (por. [Najman, Najman 2002; 2003]). Obiekty uzyskane na dwuwymiarowej mapie SOM niemal zawsze mają charakter rozmyty. Ostre granice skupień prawie nigdy nie są wyraźnie widoczne. Jest to znaczne utrudnienie w praktyce analizy danych, gdyż aby uzyskać sieć optymalną z punktu widzenia przyjętych kryteriów jej jakości, należy wykonać wiele eksperymentów, „kalibrując" parametry sieci. Niestety mimo bardzo dużej efektywności sieci SOM w grupowaniu obiektów nie posiada ona wewnętrznego kryterium ustalającego liczbę skupień. Niezbędne jest zastosowanie zewnętrznego kryterium wyznaczania liczby skupień.

W niniejszej pracy zaprezentowane zostaną wybrane wskaźniki liczby skupień nazywane też wskaźnikami jakości grupowania (*cluster validity index, cluster separation index*). Wskaźniki te mogą być szczególnie użyteczne w analizie dwuwymiarowych map SOM, a także innych dwuwymiarowych, rozmytych zbiorów danych. Dokonana zostanie ich ocena oparta na wynikach przeprowadzonych badań eksperymentalnych.

2. Wskaźniki jakości grupowania

W literaturze można znaleźć wiele wskaźników jakości grupowania (por. [Milligan, Cooper 1985; Najman, Najman 2005]). W artykule zostaną zaprezentowane jedynie wskaźniki korzystające z rozmytej funkcji przynależności obiektów do skupień[1]. Grupowania obiektów dokonano metodą FCM (*fuzzy c-means*). W prezentowanych badaniach zostaną uwzględnione cztery wskaźniki, wybrane przez autorów jako potencjalnie najsilniejsze spośród proponowanych w literaturze: The Bezdek's partition coefficient (PC), Classification Entropy (CE), Separations Index (S). Do porównania jakości grupowania uzyskiwanej dzięki metodzie FCM z zastosowaniem wybranych wskaźników posłuży indeks Daviesa–Bouldina (DB)[2].

Indeks PC (*The Bezdek's partition coefficient*), zaproponowany przez Bezdeka [Bezdek 1974; Bezdek, Ehrlich, Full 1984] jest definiowany następująco:

$$PC(c)=\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\left(u_{ij}\right)^2, \tag{1}$$

gdzie: $u_{ij}\ (i=1,2,...,c;\ j=1,2,...,N)$ – miara przynależności j-tego obiektu do i-tego skupienia, rozmyta funkcja przynależności uzyskana po grupowaniu metodą FCM.

Konfiguracja skupień, która maksymalizuje wartość *PC*, będzie uznawana za najlepszą. Liczba skupień c, przy której uzyskujemy maksimum *PC*, będzie uznawana za optymalną.

Kolejny indeks, Classification Entropy (*CE*), także zaproponowany przez Bezdeka, jest definiowany następująco:

$$CE(c)=-\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}\log\left(u_{ij}\right), \tag{2}$$

gdzie: $u_{ij}\ (i=1,2,...,c;\ j=1,2,...,N)$ – miara przynależności j-tego obiektu do i-tego skupienia, rozmyta funkcja przynależności uzyskana po grupowaniu metodą FCM.

W formule *CE* zastosowano znaną miarę entropii Shannona. Ideą indeksu *CE* jest minimalizacja entropii obiektów. Im entropia mniejsza, tym badany „układ" jest bardziej uporządkowany. Najwyższe uporządkowanie obiektów (a więc mini-

[1] W przedstawionych tu badaniach zastosowana zostanie jedynie rozmyta funkcja przynależności obiektu do danego skupienia. Nie ma jednak formalnych przeszkód, aby zastosować w prezentowanych wskaźnikach inną miarę przynależności.

[2] Indeks DB należy do najsilniejszych znanych autorom w identyfikacji skupień dla zbiorów separowalnych i nierozmytych. Opis własności teoretycznych indeksu DB, a także wyniki badań empirycznych w podobnej klasie problemów można znaleźć w pracy: [Najman, Najman 2005]. Dane testowe i szczegółowe wyniki tabelaryczne analiz można znaleźć na stronie internetowej: http://panda.bg.univ.gda.pl/~Najman.

malna entropia) powinno jednocześnie gwarantować optymalną liczbę skupień. Miary *PC* i *CE* są ze sobą w związku:

$$0 \le 1 - PC(c) \le CE(c), \tag{3}$$

dla wszystkich możliwych podziałów na c skupień (por. [Kim, Lee, Lee 2004; Sun, Wang, Jiang 2004]).

Trzeci z proponowanych indeksów to Separation Index (*S*) zdefiniowany następująco:

$$S(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^2 \, d\left(x_j - z_i\right)^2}{N \min\limits_{\substack{m,n=1,\ldots,c \\ m \neq n}} \left\{d\left(z_m - z_n\right)\right\}^2} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^2 \, d\left(x_j - z_i\right)^2}{N\left(d_{\min}\right)^2}, \tag{4}$$

gdzie: $u_{ij}\ (i=1,2,\ldots,c;\ j=1,2,\ldots,N)$ – miara przynależności j-tego obiektu do i-tego skupienia, rozmyta funkcja przynależności uzyskana po grupowaniu metodą FCM.

$d_{\min}$ – minimalna odległość euklidesowa między centrami skupień,

$d\left(x_j - z_i\right)^2$ – odległość euklidesowa między j-tym obiektem a centrum i-tego skupienia.

Im $d_{\min}$ jest większe od licznika indeksu *S*, tym grupowanie lepsze, co jest zgodne z zasadą, że im odległość między centrami skupień jest większa niż odległość między obiektami a centrami skupień, tym lepiej. Funkcja u_{ij} jest tu traktowana jako waga wzmacniająca tę zasadę, gdy jej wartość jest wysoka, i osłabiająca, gdy jest niska. Ostatecznie, im mniejsza wartość indeksu *S*, tym grupowanie lepsze. Liczbę skupień c, dla której uzyskujemy minimalną wartość *S*, uznamy za optymalną.

3. Eksperyment badawczy

Do oceny własności i praktycznej przydatności wskaźników zaprojektowano odpowiedni eksperyment. Przygotowano 14 zbiorów danych[3], uporządkowanych od najprostszych do najbardziej złożonych, różniących się:

1) stopniem separowalności, od całkowicie liniowo separowanych po nieseparowalne w przestrzeni dwuwymiarowej,

2) stopniem rozmycia danych,

3) proporcjami między odległościami między obiektami i centrami skupień,

4) rozkładem obiektów w skupieniach, wszystkie skupienia gaussowskie, wszystkie równomierne, mieszane,

[3] Komplet danych źródłowych i rysunki omawianych zbiorów znajdują się na stronie internetowej autorów, http://panda.bg.univ.gda.pl/~Najman.

5) liczbą skupień (od 2 do 5),

6) liczbą obiektów w skupieniu (równa lub różna: od 50 do 500 obiektów w skupieniu),

7) stopniem „owalności" skupień (od dokładnie okrągłych po bardzo wysmukłe elipsy).

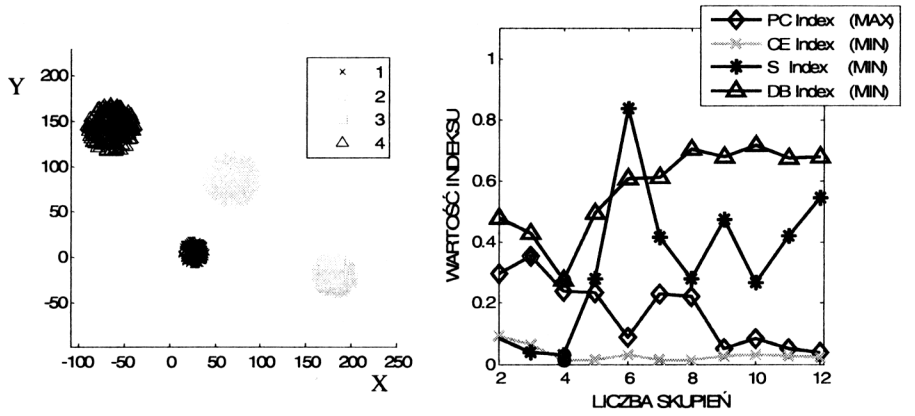
Wszystkie zbiory były klasyfikowane metodą FCM, z zadaną liczbą skupień od 2 do 12. Dla każdego zadanego podziału wyznaczano wartości wszystkich badanych indeksów. Uzyskane wyniki prezentowano graficznie i tabelarycznie.

4. Wyniki badań

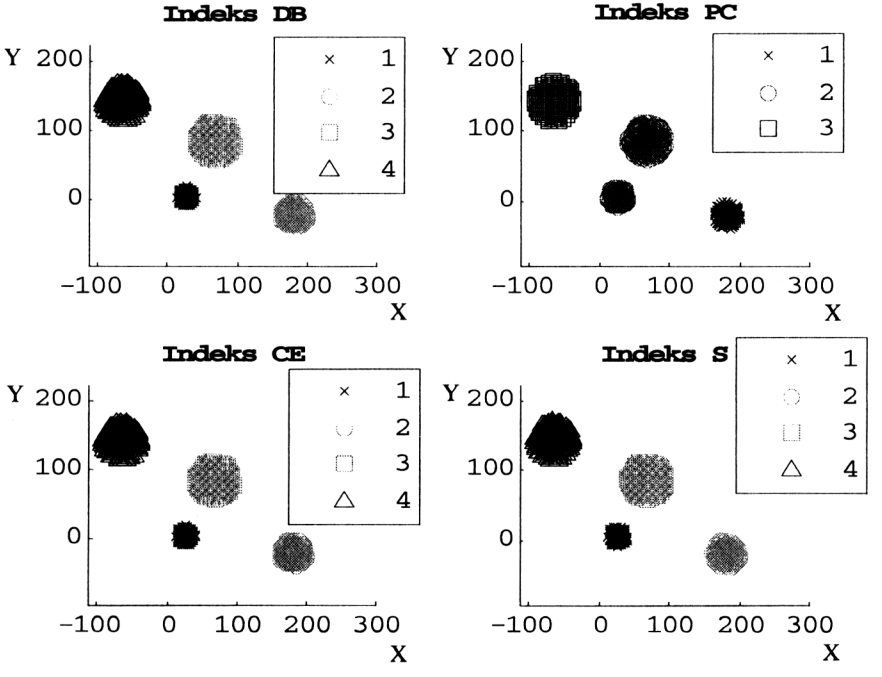
W przypadku zbioru danych nr 1 mamy 4 skupienia w pełni separowalne liniowo, nierozmyte, owalne. Rozkład punktów w każdym ze skupień jest gaussowski, w każdym skupieniu jest 250 obiektów. Problem grupowania wydaje się wyjątkowo łatwy. Indeksy *CE* i *S* a także *DB* prawidłowo wskazały na optymalne grupowanie składające się z 4 skupień. Indeks *PC* wskazał, że optymalny podział to podział na 3 skupienia. Graficzna prezentacja analizy zbioru danych nr 1 jest pokazana na rys. 1 i 2. Ze względu na znaczną objętość uzyskanych wyników w dalszej części pracy zostaną przedstawione syntetyczne wnioski oparte na analizie wszystkich 14 zbiorów danych. Szczegółowe tablice wynikowe można znaleźć na stronie internetowej autorów.

Syntetyczne wyniki zaprezentowane są w tab. 1 i 2. Żaden ze wskaźników nie rozpoznał prawidłowo właściwego podziału dla wszystkich zbiorów. Uzyskano odpowiednio 7,1, 21,4 i 42,9% prawidłowych wskazań dla indeksów *PC*, *CE* i *S*. Indeks *PC* wskazał prawidłową liczbę skupień jedynie w jednym przypadku. Systematycznie wskazywał podział danych na dwa skupienia. Poza zbiorem testowym nr 1, w każdym przypadku wartości indeksu *PC* malały monotonicznie wraz z liczbą klas, na jaką rozdzielany był zbiór danych. Sytuacja taka miała miejsce niezależnie od własności testowanego zbioru danych. Z tego powodu uznajemy ten wskaźnik za nieprzydatny w badaniach prezentowanego typu.

Nieco lepsze wyniki osiągnięto dzięki zastosowaniu indeksu *CE*. Trzykrotnie wskazano prawidłową liczbę skupień, a raz błąd był na poziomie 1 skupienia. Indeks ten dawał lepsze wyniki dla danych charakteryzujących się małym rozmyciem. W trudniejszych przypadkach, gdy dane są znacznie rozmyte, indeks ten wskazuje na maksymalną liczbę skupień spośród testowanych (liczba skupień równa 12). Często wartości indeksu *CE* monotonicznie maleją wraz z wzrostem liczby testowanych klas. Indeks ten może być pomocniczym wskaźnikiem jedynie dla danych dobrze separowalnych, charakteryzujących się minimalnym rozmyciem. Najlepszym wskaźnikiem okazał się indeks *S*, który niemal w połowie przypadków wskazał prawidłową liczbę skupień. W większości pozostałych przypadków



Rys. 1. Zbiór danych nr 1 i wartości wskaźników jakości grupowania
Źródło: opracowanie własne.



Rys. 2. Podział zbioru danych nr 1 wykonany metodą FCM, oparty na ekstremalnych wartościach wskaźników jakości grupowania
Źródło: opracowanie własne.

błąd wynosił 1-2 skupienia. Indeks S wskazywał prawidłową liczbę skupień niezależnie od stopnia separowalności skupień stopnia ich rozmycia. Jego wartości, prze-

ciwnie niż poprzednie indeksy, charakteryzują się znacznym zróżnicowaniem w zależności od liczby testowanych skupień. Wyjaśnienia wymaga jedynie wynik dla zbioru drugiego. W zbiorze drugim znajdują się jedynie dwa skupienia, w niewielkim stopniu rozmyte, owalne, o równomiernym rozkładzie obiektów w skupieniach. W skupieniach tych znajduje się stosunkowo niewiele obiektów, szeroko rozrzuconych. W wyniku grupowania skupienia te zostały rozbite na wiele mniejszych.

Tabela 1. Syntetyczne wyniki dla grupowania metodą FCM

FCM		Liczba skupień			
Numer zbioru testowego	rzeczywista liczba skupień	indeks Daviesa–Bouldin'a	indeks *PC*	indeks *CE*	indeks *S*
1	4	4	3	4	4
2	2	2	2	12	12
3	5	5	2	5	5
4	3	2	2	2	2
5	3	3	2	12	3
6	5	4	2	3	4
7	5	4	2	5	2
8	3	3	2	12	3
9	3	2	2	12	2
10	4	3	2	11	3
11	4	4	2	12	4
12	5	11	2	12	9
13	5	5	2	12	5
14	5	4	2	12	4
Poprawnych wskazań		50,0%	7,1%	21,4%	42,9%

Źródło: opracowanie własne.

Wydaje się, że uzyskane wyniki pozwalają uznać indeks *S* za przydatny w zastosowaniach empirycznych, dla danych typu testowanego w tym badaniu.

Wszystkie wyniki zastosowania wskaźników opartych na rozmytej funkcji przynależności okazały się „gorsze" niż wyniki uzyskane z zastosowaniem indeksu *DB*. Ale nawet indeks *DB* prawidłowo wskazał liczbę skupień jedynie w 50% przypadków. Indeks ten prawidłowo wskazał liczbę skupień dla wszystkich zbiorów separowalnych, a także tych charakteryzujących się niewielkim rozmyciem. Błąd wskazania wynosił zwykle 1 skupienie, poza zbiorem 12. W tym przypadku błąd wyniósł aż 6 skupień. Zbiór ten był wyjątkowo trudny. Składał się z 5 skupień rozłożonych w bardzo smukłych elipsach, wszystkie skupienia są rozmyte, dwa skupienia „wnikają" we wszystkie pozostałe.

W drugiej części badania zmierzono jakość grupowania możliwą do uzyskania metodą FCM. W tym celu wyznaczono udział prawidłowo sklasyfikowanych obiektów w dwóch wariantach:

1) przy założeniu prawidłowego ustalenia liczby skupień,
2) dla takiej liczby skupień, która dawała najlepszą klasyfikację przy nieprawidłowej liczbie skupień (zwykle o jedno skupienie więcej lub mniej niż właściwa).

Tabela 2. Maksymalny odsetek prawidłowo zaklasyfikowanych obiektów dla prawidłowej i nieprawidłowej liczby skupień

Zbiór danych	Najlepsza klasyfikacja		
	prawidłowa liczba skupień (w %)	nieprawidłowa liczba skupień (w %)	różnica
1	100,00	87,63	12,37
2	96,00	71,00	25,00
3	100,00	91,20	8,80
4	97,94	80,96	16,98
5	99,78	76,48	23,30
6	92,10	82,42	9,68
7	86,81	81,74	5,07
8	90,59	82,55	8,04
9	91,02	81,09	9,93
10	97,00	86,75	10,25
11	92,25	84,00	8,25
12	87,31	79,14	8,17
13	95,00	86,56	8,44
14	87,44	78,32	9,12

Źródło: opracowanie własne.

W tabeli 2 przedstawiono szczegółowe wyniki tych analiz. W każdym przypadku, co nie jest zaskoczeniem, najwyższy możliwy do uzyskania procent prawidłowo sklasyfikowanych obiektów uzyskiwano przy znajomości prawidłowej liczby skupień. Różnica w stosunku do grupowania zaproponowanego przez badane indeksy, w przypadku uzyskania błędnej liczby skupień, wynosiła od 5,07 do 25 punktów procentowych. Największe różnice pojawiają się dla najprostszych zbiorów danych. Dane silnie rozmyte, silnie elipsoidalne, o niekorzystnym stosunku odległości między najdalszymi obiektami w skupieniach do odległości między centrami skupień i równomiernym rozkładzie obiektów w skupieniach udawało się stosunkowo dobrze sklasyfikować nawet mimo błędnego ustalenia liczby skupień. Działo się tak być może dlatego, że zbiory te były trudne do prawidłowego sklasyfikowania i nawet przy znajomości właściwej liczby skupień udział poprawnie sklasyfikowanych obiektów był niezbyt wysoki (78,32-86,56%).

5. Wnioski

W literaturze proponuje się wiele wskaźników jakości grupowania, wskazujących w sposób ilościowy na optymalny podział obiektów według przyjętego kryterium. W prezentowanym badaniu opisano trzy z nich: The Bezdek's partition coefficient (*PC*), Classification Entropy (*CE*), Separations Index (*S*). Wskaźniki te znajdują zastosowanie przy grupowaniu obiektów metodą FCM. Wykazano, że dla dwuwymiarowych, rozmytych zbiorów danych o opisanych powyżej własnościach indeksy *PC* i *CE* nie mogą być podstawą obiektywnego wyróżniania liczby skupień. Indeks *S* okazał się użytecznym narzędziem, wymaga jednak bacznej kontroli, gdyż może zawyżać lub zaniżać optymalną liczbę skupień (zwykle o 1). Najlepszym indeksem okazał się indeks *DB*, który jednak nie korzysta z rozmytej funkcji przynależności obiektów do skupień i był tu badany jedynie jako punkt odniesienia.

Literatura

Bezdek J.C. (1974), *Cluster Validity with Fuzzy Sets*, „J. Cybernet." nr 3, s. 58-72.

Bezdek J.C., Ehrlich R., Full W. (1984), *FCM: Fuzzy C-Means Algorithm*, Computers and Geoscience.

Jurkiewicz T., Najman K. (2003), *Bootstrapowa analiza własności modyfikowanego estymatora syntetycznego*, [w:] *Metoda reprezentacyjna w badaniach ekonomiczno-społecznych*, red. J. Wywiał, część II, AE, Katowice, s. 15-26.

Kim D.W., Lee K., Lee D. (2004), *On Cluster Validity Indem for Estimation of the Optima Numer of Fuzzy Clusters*, „Pattern Recognition" nr 37, s. 2009-2025.

Milligan G.W., Cooper M.C. (1985), *An Examination of Procedures for Determining the Number of Clusters in Data Set*, „Psychometrika" nr 50(2), s. 159-179.

Najman K., Najman K. (2002), *Zastosowanie sieci neuronowej typu SOM do wyboru najatrakcyjniejszych spółek na WGPW*, Prace Naukowe Akademii Ekonomicznej nr 952, AE, Wrocław, s. 493-499.

Najman K., Najman K. (2002), *Zastosowanie sieci neuronowej typu SOM do wyboru najatrakcyjniejszych spółek na WGPW na bazie wskaźników analizy technicznej*, Rynek Kapitałowy – Skuteczne Inwestowanie, Międzyzdroje, t. II, Wydawnictwo Uniwersytetu Szczecińskiego, s. 403-417.

Najman K., Najman K. (2003), *Próba zastosowania sieci neuronowej typu SOM w badaniu przestrzennego zróżnicowania powiatów w Polsce*, „Wiadomości Statystyczne" nr 4, s. 72-84.

Najman K., Najman K. (2005), *Analityczne metody ustalania liczby skupień, Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 12, AE, Wrocław, s. 265-273.

Sun H., Wang S., Jiang Q. (2004), *FCM-Based Model Selection Algorithms for Determining the Number of Clusters*, „Pattern Recognition" nr 37, s. 2027-2037.

ANALYTICAL PROCEDURES FOR DETERMINING THE NUMBER OF CLUSTERS IN FUZZY DATA SETS

Summary

Several clustering techniques have been proposed for the analysis of fuzzy data sets. Cluster validity indices represent useful tools to support such a task. In this paper three validation indices were applied to fourteen data sets. The resultant optimal clusters have been found to be stable for the different validity indices used, viz. Bezdek's Partition Coefficient (PC), Classification Entropy (CE) and Separation Index (S). It was shown that these methods might support the prediction of the optimal cluster partitioning for those data sets but the determination of the optimal number of clusters is an open problem. Two indices (PC and CE) were called into question their usefulness. Index S was characterized by relatively not large errors and significant effectiveness.