PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005

TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Kamila Migdał-Najman, Krzysztof Najman
Uniwersytet Gdański

ANALITYCZNE METODY USTALANIA LICZBY SKUPIEŃ

1. Wstęp

Jednym z podstawowych problemów, który pojawia się w analizie skupień, jest ustalenie liczby grup, na jaką należy rozdzielić badane obiekty. Zwykle, kierując się zdrowym rozsądkiem i znajomością badanego zagadnienia, próbuje się pogrupować obiekty, tak aby obiekty należące do danej klasy miały jak najwięcej wspólnych cech, a jednocześnie jak najmniej wspólnych cech z obiektami spoza tej klasy. W wielu przypadkach empirycznych badane zbiory obiektów wykazują pewne naturalne tendencje do grupowania się we względnie homogeniczne grupy, co znacznie ułatwia ich analizę. Gdy grupy obiektów są wewnętrznie podobne i jednocześnie między grupami są wyraźne różnice, wyodrębnienie tych grup (i ustalenie ich liczby) zwykle nie jest trudne. Jednak pojawia się coraz więcej przykładów badań empirycznych[1], w których należy uwzględnić znaczną liczbę cech dla ogromnej liczby przypadków, w których podziały na grupy obiektów wewnętrznie jednorodnych są trudne i intuicyjnie nieobserwowalne. Problem wyboru liczby grup jednak pozostaje. Wiele metod grupowania, jak np. metoda *k*-średnich, wymaga *a priori* podania liczby klas.

Z podobnym problemem spotkali się autorzy w badaniach nad zastosowaniami sztucznych sieci neuronowych typu SOM (ang. *self organizing map*). Sieci te pozwalają na redukcję wymiarowości badanego zagadnienia do dwóch wymiarów (por. [Najman, Najman 2002a; 2002b; 2003]). W uproszczeniu można powiedzieć, że w wyniku procedury samouczenia sieci SOM uzyskuje się macierz ujednoliconych odległości (ang. *u-matrix*), którą prezentuje się graficznie na tzw. mapie SOM. Następnie, kierując się roz-

[1] Autorzy z własnych doświadczeń znają przykłady badań, w których zleceniodawca życzy sobie, aby uwzględnić jednocześnie kilkaset cech w analizie 5000 przypadków. W ustaleniu liczby klas, na które należy taki zbiór podzielić, nie wystarczy prosta analiza średnich i wariancji.

kładem barw (obiekty podobne mają podobną, jasną barwę i od innych skupień są oddzielone ciemnymi pasami) na mapie SOM, dokonuje się grupowania, „ręcznie" wyznaczając linie podziału. Podejście takie wprowadza jednak element subiektywny, a w analizach opartych na sztucznych sieciach neuronowych jest to bezcelowe ze względu na to, że aby znaleźć sieć o optymalnej strukturze, należy przeanalizować dużą liczbę map SOM.

W takiego typu badaniach niezbędne jest zastosowanie obiektywnego, ilościowego kryterium wyznaczania liczby skupień. Wskaźniki takie istnieją i najczęściej są nazywane wskaźnikami jakości grupowania (ang. *cluster validity index, cluster separation index*). Są to miary wskazujące w sposób ilościowy na optymalny podział obiektów z pewnego punktu widzenia. Niestety literatura dotycząca tego tematu jest nieuporządkowana i opisywane metody nie są zwykle wystarczająco przebadane. Większość autorów tych publikacji nie zna innych opracowań (brak odwołań do literatury i badań) i odkrywa niektóre wskaźniki od nowa. Prezentowane w artykule badania są fragmentem prac nad uporządkowaniem teorii zagadnienia i możliwościami zastosowania wskaźników w empirycznych badaniach ekonomicznych[2].

2. Wskaźniki jakości grupowania

W literaturze można znaleźć kilkadziesiąt wskaźników jakości grupowania. W prezentowanych badaniach pominięto wskaźniki możliwe do zastosowania wyłącznie w hierarchicznych metodach grupowania, ponieważ dla dużych zbiorów danych (i cech) ich użyteczność jest ograniczona (por. [Aldefender, Blashfield 1996; Milligan, Cooper 1985]). Pominięto także możliwe do zastosowania wskaźniki dla cech o zadanych rozkładach wielowymiarowych (por. [Wolfe 1970]) lub możliwe do zastosowania jedynie przy spełnieniu wielu dodatkowych założeń (por. [Sarle 1983]). W artykule nie będą zaprezentowane wskaźniki możliwe do zastosowania w metodzie FCM, oparte na rozmytej funkcji przynależności, gdyż stanowią one podstawę osobnego opracowania.

W prezentowanych badaniach uwzględniono dwa wskaźniki możliwe do zastosowania przy grupowaniu obiektów z zastosowaniem klasycznych metod typu *k*-średnich i FCM (ang. *fuzzy c-means*). Za pomocą tych metod często analizuje się duże zbiory danych o tysiącach obiektów opisywanych przez wiele cech. Należą do nich: indeks Dunna (ang. *Dunn's separation index, DI*) i indeks Davisa-Bouldina (ang. *Davies-Bouldin's index* (DB)).

Indeks Dunna bazuje na założeniu, że odległości między obiektami w skupieniu powinny być małe w stosunku do odległości między obiektami z różnych skupień. W związku z tym buduje się dwie miary odległości: wewnętrzną (między obiektami w skupieniu, ang. *intercluster distance*) i zewnętrzną (między skupieniami, ang. *intracluster distance*). Maksymalizacja odległości między skupieniami, przy równoczesnej minimali-

[2] Prezentowany w artykule wycinek szerszych badań przedstawia wskaźniki potencjalnie, szczególnie użyteczne w analizie map SOM. Ponieważ mapy te są 2-wymiarowe, będzie uwzględniony jedynie ten aspekt ustalania liczby grup. Prezentowane wskaźniki mogą jednak być uogólnione na większą liczbę wymiarów.

zacji wewnętrznej odległości, prowadzi do optymalnego podziału obiektów. Ogólnie można tę relację przedstawić następująco:

$$DI(c)=\min_{1\le i\le c}\left\{\min_{\substack{1\le j\le c\\ j\neq i}}\left\{\frac{\delta(X_i,X_j)}{\max_{1\le k\le c}\{\Delta(X_k)\}}\right\}\right\}, \tag{1}$$

gdzie: $\delta(X_i,X_j)$ – odległość między skupieniami X_i i X_j,

$\Delta(X_k)$ – odległości wewnętrzne między obiektami skupienia X_k,

c – to liczba skupień, na którą rozdzielamy obiekty.

Odległości między obiektami obliczono z zastosowaniem dobrze znanych metryk: euklidesowej i miejskiej. W literaturze są proponowane różne sposoby definiowana odległości między skupieniami. Do częściej stosowanych sposobów należą: najbliższego sąsiada (ang. *single linkage),* najdalszego sąsiada (ang. *complete linkage),* średnich połączeń (ang. *average linkage),* środka ciężkości (ang. *centroid linkage)* i odległość Hausdorfa (ang. *Hausdorff metrics*). W prezentowanym badaniu zastosowano następujące odległości: najdalszego sąsiada – największa odległość pomiędzy dwoma obiektami należącymi do dwóch różnych skupień[3] – i średnich połączeń –średnia odległość pomiędzy wszystkimi obiektami należącymi do dwóch różnych skupień[4]. Proponuje się również różne sposoby definiowania odległości w obrębie skupienia. Do częściej stosowanych należą odległości: najdalszego sąsiada (ang. *complete diameter*), średnich połączeń (ang. *average diameter*) i środków ciężkości (ang. *centroid diameter*). W prezentowanym badaniu zastosowano odległość: *average diameter*, czyli średnią odległość między wszystkimi obiektami należącymi do tego samego skupienia [Bolshakova, Azuaje 2003, s. 827--828]. Liczba skupień, która maksymalizuje wartość DI, uznawana będzie za optymalną.

Index Daviesa-Bouldina bazuje na założeniu, że optymalny podział obiektów na grupy to taki, który minimalizuje zróżnicowanie obiektów w klasach i jednocześnie maksymalizuje odległości między centrami klas. Jeżeli odległości między centrami skupień są duże w stosunku do zmienności obiektów w skupieniach, takie grupowanie jest uznawane za dobre. Ogólnie zasadę tę można zapisać następująco:

$$DB(c)=\frac{1}{c}\sum_{i=1}^{c}\max_{j,j\neq i}\left\{\frac{S_{i,q}+S_{j,q}}{d_{ij,t}}\right\}, \tag{2}$$

gdzie:

$S_{i,q}=\left(\frac{1}{|C_i|}\sum_{x\in C_i}\left\{\|x-z_i\|_2^q\right\}\right)^{1/q}$ to zróżnicowanie w i-tym skupieniu,

$d_{ij,t}=\left\{\sum_{s=1}^{p}|z_{is}-z_{js}|^t\right\}^{1/t}=\|z_i-z_j\|_t$ to odległość między skupieniami C_i a C_j.

[3] W połączeniu z odległością euklidesową indeks nazwiemy DI1, a z odległością miejską – DI3.

[4] W połączeniu z odległością euklidesową indeks nazwiemy DI2, a z odległością miejską – DI4.

Przy $q = 2$ miarą zmienności jest odchylenie standardowe, a przy $t = 2$ mamy odległość euklidesową między centrami skupień. Konfiguracja klas, która minimalizuje indeks DB, uznawana jest za optymalną [Mali, Mitra 2003, s. 2371].

3. Eksperyment badawczy

Do oceny własności i praktycznej przydatności wskaźników zaprojektowano odpowiedni eksperyment. Przygotowano 15 zbiorów danych[5] różniących się:

- liczbą grup, w badanych zbiorach danych występowało od 4 do 6 grup,
- stopniem separowalności grup,
  - grupy separowalne jedynie w osi X,
  - grupy separowalne jedynie w osi Y,
  - grupy separowalne w obu osiach,
  - grupy nieseparowalne w obu osiach,
- odległościami między centrami grup i obiektami w grupach,
  - odległości wewnątrzgrupowe małe w stosunku do odległości między centrami grup,
  - odległości wewnątrzgrupowe równe odległościom między centrami grup,
  - odległości wewnątrzgrupowe większe od odległości między centrami grup.

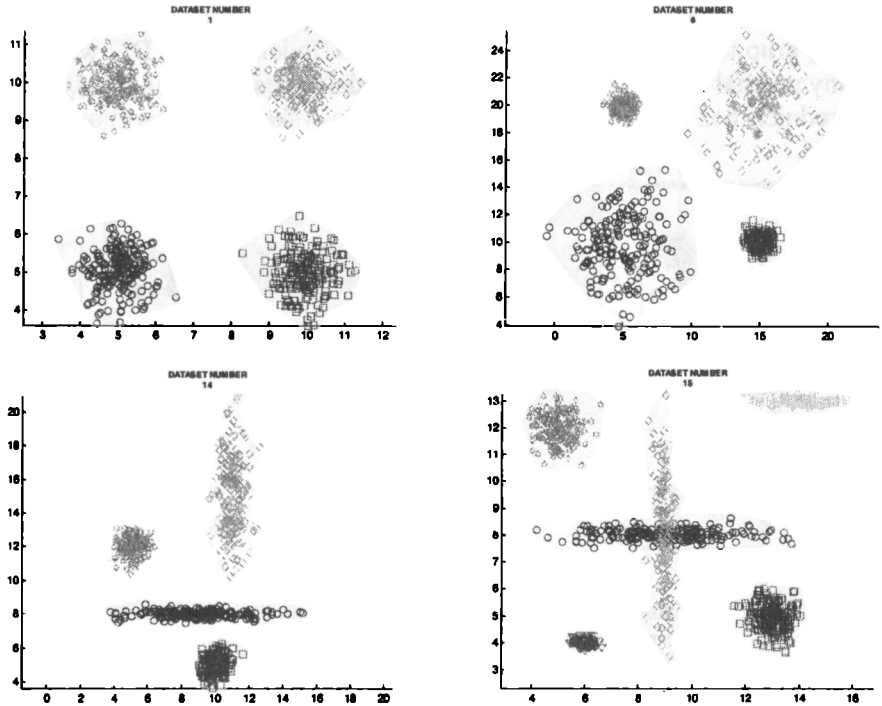
Wszystkie zbiory (poza 15) są obiektywnie separowalne i nie mają charakteru rozmytego. Wszystkie mają grupy 200-elementowe o budowie owalnej lub elipsoidalnej, o gęstości obiektów rosnącej w miarę zbliżania się do centrum grupy. Są to uproszczenia, które wraz z rozwojem badań będą kolejno znoszone. Przykładowe zbiory o numerach 1, 8, 14, 15 przedstawiono na rys. 1. Procedura testowa przebiegała następująco:

1) wszystkie zbiory danych grupowano metodą k-średnich i FCM,
2) zbiór danych grupowano na 2 do 12 grup,
3) dla każdego podziału wyznaczano 6 wskaźników jakości grupowania,
4) uzyskane wyniki zapisywano w postaci tablic i prezentowano graficznie.

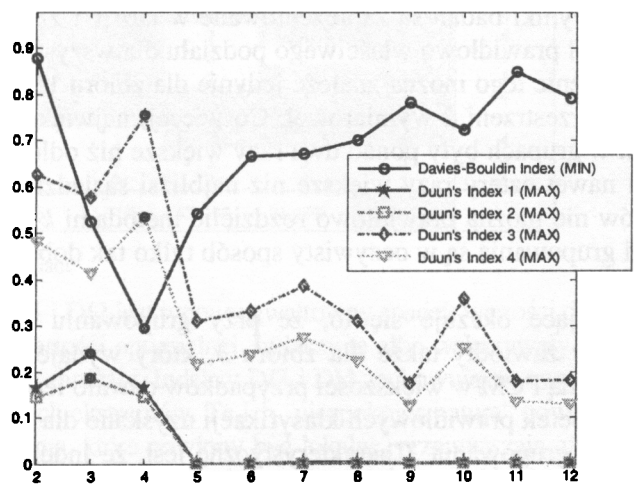
Ze względu na zastosowania empiryczne istotne są jedynie wartości ekstremalne badanych wskaźników. Z naukowego punktu widzenia, ważne jest jednak także ich zachowanie w przypadku błędnych podziałów. Zachowanie to było analizowane na podstawie wyników liczbowych i odpowiednich wykresów. Przykład[6] takiej prezentacji przedstawiono na rys. 2 i 3.

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

[6] Komplet wyników tabelarycznych zajmuje 36 stron A4 i nie umieszczono go w tej pracy. Komplet tablic wynikowych znajduje się na stronie internetowej autorów.

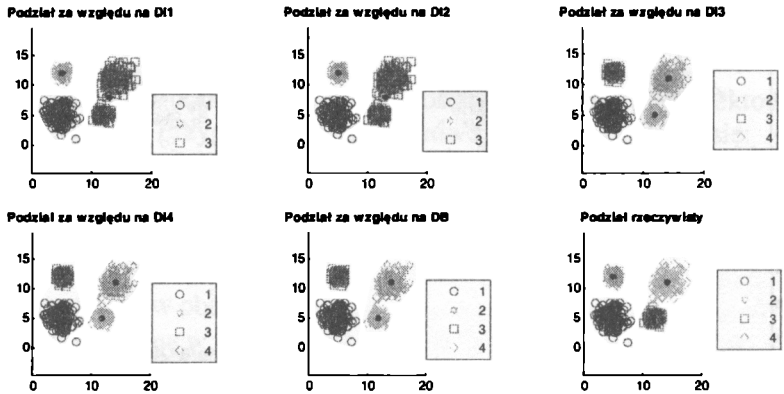


Rys. 1. Przykładowe zbiory danych
Źródło: opracowanie własne.



Rys. 2. Wartości wskaźników jakości grupowania dla 13 zestawu danych
Źródło: opracowanie własne.

W prezentowanym przykładzie zbiór powinien być podzielony na 4 grupy. Na taką liczbę grup wskazują wskaźnik DB, DI3 i DI4. Indeksy DI1 i DI2 wskazują na optymalny podział zbioru na 3 grupy. Sposób podziału jest pokazany na rys. 3. Metodą *k*-średnich bez trudu rozdzielono zbiór, o ile wskaźniki prawidłowo oceniły liczbę grup.



Rys. 3. Podział 13. zbioru danych wykonany metodą *k*-średnich oparty na ekstremalnych wartościach wskaźników jakości grupowania
Źródło: opracowanie własne.

4. Wyniki badań

Syntetyczne wyniki badań są zaprezentowane w tab. 1 i 2. Żaden ze wskaźników nie rozpoznał prawidłowo właściwego podziału dla wszystkich zbiorów. Pewne usprawiedliwienie tego można znaleźć jedynie dla zbioru 15. Zbiór ten był nieseparowalny w przestrzeni 2-wymiarowej. Co więcej, największe odległości między obiektami w grupach były ponad dwa razy większe niż odległości między centrami grup, a nawet cztery razy większe niż najbliżsi sąsiedzi z sąsiednich grup. Takich zbiorów nie można prawidłowo rozdzielić metodami *k*-średnich lub FCM. Miary jakości grupowania są w oczywisty sposób tylko tak dobre jak metoda użyta do grupowania.

Dość zaskakujące okazuje się to, że przy grupowaniu metodą *k*-średnich wszystkie metody zawiodły także dla zbioru 3, który wydaje się bardzo prosty. Grupowanie metodą FCM w większości przypadków dawało lepsze rezultaty.

Najwyższy odsetek prawidłowych klasyfikacji uzyskano dla indeksu DB, niezależnie od metody grupowania. Charakterystyczne jest, że indeks DB przebiega w każdym przypadku bardzo gładko, z wyraźnym minimum globalnym. Jego interpretacja jest bardzo jednoznaczna. Odstępstwa od tej reguły pojawiały się w sposób szczególny. Jeżeli zbiór liczył 4 grupy, a wymuszaliśmy podział na 8 grup, gdy każda z 4 podstawowych grup dzieliła się na pół, to na wykresie indeksu DB poja-

wiało się wyraźne minimum lokalne. Jest to silny sygnał potwierdzający właściwy podział.

Tabela 1. Syntetyczne wyniki dla grupowania metodą *k*-średnich

Dane	Dunn's Index 1	Dunn's Index 2	Dunn's Index 3	Dunn's Index 4	Davies--Bouldin Index	Liczba grup
Zbiór 1	4	4	4	4	4	4
Zbiór 2	4	4	2	2	4	4
Zbiór 3	3	3	2	2	3	4
Zbiór 4	4	4	4	4	4	4
Zbiór 5	4	4	4	4	4	4
Zbiór 6	4	4	4	4	4	4
Zbiór 7	4	4	4	4	4	4
Zbiór 8	4	4	2	4	4	4
Zbiór 9	2	4	2	2	4	4
Zbiór 10	4	4	2	4	4	4
Zbiór 11	4	4	4	4	4	4
Zbiór 12	4	4	4	4	4	4
Zbiór 13	3	3	4	4	4	4
Zbiór 14	3	3	3	3	3	4
Zbiór 15	4	4	3	3	5	6
Odsetek prawidłowych	71,43%	78,57%	57,14%	71,43%	85,71%	

Źródło: opracowanie własne[7].

Tabela 2. Syntetyczne wyniki dla grupowania metodą FCM

Dane	Dunn's Index 1	Dunn's Index 2	Dunn's Index 3	Dunn's Index 4	Davies--Bouldin Index	Liczba grup
Zbiór 1	4	4	4	4	4	4
Zbiór 2	4	4	2	2	4	4
Zbiór 3	4	4	2	2	4	4
Zbiór 4	4	4	4	4	4	4
Zbiór 5	4	4	4	4	4	4
Zbiór 6	4	4	4	4	4	4
Zbiór 7	4	4	4	4	4	4
Zbiór 8	4	4	2	4	4	4
Zbiór 9	4	4	2	3	4	4
Zbiór 10	4	4	2	4	4	4
Zbiór 11	4	4	4	4	4	4
Zbiór 12	4	4	4	4	4	4
Zbiór 13	2	3	4	4	4	4
Zbiór 14	3	3	3	3	3	4
Zbiór 15	2	2	3	3	5	6
Odsetek prawidłowych	85,71%	85,71%	57,14%	71,43%	92,86%	

Źródło: opracowanie własne[8].

Dla indeksów DI1 i DI2 jest typowy gwałtowny spadek wartości do 0 dla liczby grup większej o 2-3 od wartości optymalnej. Indeksy te albo wskazywały prawidłową liczbę grup albo ją znacznie zaniżały. Indeksy DI3 i DI4 zachowują się zwykle mniej stabilnie, mając więcej lokalnych ekstremów. Ta ich „niespokojna natura" powoduje w niektórych sytuacjach, że maksima, które powinny być lokalne, przewyższają globalne, prowadząc w konsekwencji do błędnych wniosków. Ta różnica zwykle jest minimalna (rzędu

[7] Ze względu na to, że ze zbiorem 15. nie poradziły sobie same metody grupowania, tego zbioru nie uwzględniono przy ocenie odsetka prawidłowych klasyfikacji.

[8] Jak wyżej.

0,001), ale może być decydująca. Żaden z indeksów w żadnym przypadku nie zawyżył liczby grup. Zaniżenie zwykle nie przekraczało jednej grupy.

Jakość wyników silnie zależała od zastosowanych metod grupowania. Prawidłowością jest, że im grupy bardziej owalne, a odległości między centrami grup są większe w stosunku do najbliższych sąsiadów z sąsiednich grup, tym wskaźniki wyraźniej i bardziej stanowczo wskazywały na prawidłową liczbę grup. Trudno powiedzieć, czy jest to własność wskaźników czy metod grupowania. Z pewnością metoda *k*-średnich „lubi" takie dane, dlatego prawidłowe wskazania wskaźników mogą być efektem metody grupowania. Zbiory, w których grupy miały rozkład zdecydowanie eliptyczny, sprawiały większe problemy samym metodom grupowania, w związku z tym uzyskano słabsze wyniki samych wskaźników[9].

Do wad prezentowanych miar należy także zaliczyć to, że są to miary *ex-post*. Należy najpierw dokonać podziałów, a dopiero wtedy ocenić, który z nich jest optymalny. Dla dużych zbiorów danych jest to podejście kłopotliwe. Kolejnym problemem jest ogromna pracochłonność przy wyznaczaniu wartości wskaźników. Często muszą być policzone odległości między wszystkimi obiektami ze wszystkich grup, a także między nimi a centrami grup, między każdym obiektem a wszystkimi pozostałymi obiektami w grupie we wszystkich grupach. Policzenie 5 wskaźników dla 1000 obiektów przy 12 grupach na komputerze (2Gh) może zająć kilka minut. Czas obliczeń rośnie proporcjonalnie do liczby obiektów i niemal w kwadracie do liczby grup.

5. Podsumowanie

W literaturze proponuje się kilkadziesiąt wskaźników jakości grupowania, wskazujących w sposób ilościowy na optymalny podział obiektów z pewnego punktu widzenia. W prezentowanym badaniu opisano dwa wskaźniki: indeks Dunna i indeks Davisa-Bouldina, możliwe do zastosowania przy grupowaniu obiektów z zastosowaniem klasycznych metod typu *k*-średnich i FCM (ang. *fuzzy c-means*). Podsumowując przeprowadzone badanie, można stwierdzić, że indeks Davisa-Bouldina jest silnym narzędziem wspomagającym podejmowanie decyzji o liczbie grup, na którą należy pogrupować zbiór danych. Przy zastosowaniu właściwej metody pogrupowania obiektów wskazania uzyskane dzięki temu indeksowi są prawie zawsze prawidłowe. Indeks Dunna, ze względu na znaczną liczbę kombinacji odległości wewnętrznych i zewnętrznych między obiektami oraz ze względu na jego wrażliwość na dobór metod pomiaru tych odległości, musi być jeszcze szczegółowo przebadany.

[9] Potwierdzeniem jest zastosowanie do grupowania sieci neuronowej typu SOM, która bez problemu w każdym przypadku podzieliła badane zbiory we właściwy sposób, spowodowało poprawienie wyników wszystkich wskaźników. Ich przebiegi były znacznie gładsze, ekstrema były zaś wyraźniejsze. Badania te nie zostały jeszcze zakończone.

Literatura

Aldenderfer M.S., Blashfield R.K., *Cluster Analysis*, SAGE, 1996.

Bolshakova N., Azuaje F., *Cluster Validation Techniques for Genome Expression Data*, „Signal Processing" 2003, 83, s. 827-828.

Günter S., Bunke H., *Validation Indices for Graph Clustering*, „Pattern Recognition Letters" 2003, 24, s. 1109-1110.

Mali K., Mitra S., *Clustering and its Validation in a Symbolic Framework*, „Pattern Recognition Letters" 2003, 24, s. 2371.

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

Najman K., Najman K., *Zastosowanie sieci neuronowej typu SOM do wyboru najatrakcyjniejszych spółek na WGPW*, Zeszyty Naukowe AE we Wrocławiu, AE, Wrocław 2002a, s. 493-499.

Najman K., Najman K., *Zastosowanie sieci neuronowej typu SOM do wyboru najatrakcyjniejszych spółek na WGPW na bazie wskaźników analizy technicznej*, [w:] *Rynek kapitałowy – skuteczne inwestowanie*, tom II, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Międzyzdroje 2002b, s. 403-417.

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

Sarle W.S., *Cubic Clustering Criterion*, „Technical Report" A-108, SAS Institute, Cary, N.C. 1983.

Wolfe J.H., *Pattern Clustering by Multivariate Mixture Analysis*, „Multivariate Behavioral Research" 1970, 5, s. 329-350.

ANALYTICAL PROCEDURES FOR DETERMINING THE NUMBER OF CLUSTERS

Summary

Several clustering techniques have been proposed for the analysis of data sets. Cluster validity indices represent useful tools to support such a task. They are particularly relevant in applications in which there is not a priori indication of the actual number of clusters. In this paper two validation indices were applied to fifteen data sets, using different intracluster and intercluster distances. The resultant optimal clusters have been found to be stable for the different validity indices used, viz. Davies-Bouldin Index and Dunn's Index. 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.