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

Krzysztof Najman
Uniwersytet Gdański

METODY USTALANIA LICZBY SKUPIEŃ W ZBIORACH DANYCH BINARNYCH

1. Wstęp

Jednym z podstawowych problemów, przed którym staje analityk w procesie grupowania obiektów w wielowymiarowej przestrzeni cech, jest liczba grup, na które przestrzeń tę należy podzielić. Ustalenie liczby skupień w wielowymiarowym zbiorze danych jest zadaniem trudnym. Jednocześnie liczba skupień stanowi element apriroryczny dla wielu często stosowanych w praktyce metod grupowania danych, takich jak k-średnich, k-medoid, rozmyta c-średnich czy DBSCAN. Szczególną klasą problemów klasyfikacyjnych, w których ustalenie liczby skupień ma wyjątkowy charakter, jest grupowanie obiektów opisanych wieloma cechami binarnymi. Z takimi zbiorami danych mamy coraz częściej do czynienia w badaniach marketingowych i rynkowych[1], psychologicznych i socjologicznych, medycznych czy web miningowych. W badaniach marketingowych respondentom często zadaje się serię pytań z dychotomicznymi wariantami odpowiedzi typu: „tak/nie", „posiada/nie posiada", „stosuję/nie stosuję". W badaniach psychologicznych lub medycznych często stan pacjenta i historię jego kuracji opisuje się wieloma cechami binarnymi typu: „wykonano/nie wykonano badania krwi", „wykonano/nie wykonano prześwietlenia", „zareagował/nie zareagował na lek". W badaniach web czy text miningowych jeden dokument czy jedna strona WWW mogą być opisane przez nawet 1000 cech, z których większość jest binarna typu: „słowo X występuje/nie występuje", „podsieć należy/nie należy do domeny Y". Grupowanie obiektów opisanych w ten sposób jest problemem statystycznym samym w sobie i względnie mało znanym. Dodatkowo wiele metod grupowania przystosowanych do analizy tego typu danych nie ma własnego, wewnętrznego kryterium ustalania liczby skupień.

[1] Autor zna wiele badań marketingowych prowadzonych w Polsce, w których jeden obiekt jest opisywany nawet 900 cechami, w tym kilkudziesięcioma cechami binarnymi.

Celami przedstawionych badań są prezentacja i krytyczna ocena wskaźników jakości grupowania (liczby skupień) obiektów opisanych cechami binarnymi. Ocena będzie oparta na analizie teoretycznych własności metod grupowania i opisanych wskaźników, a także na badaniach symulacyjnych. Ponieważ w szybkim tempie rośnie liczba badań, w których wymaga się grupowania danych binarnych, rozwój metod ich analizy i ustalenie liczby grup wydaje się zadaniem ważnym i aktualnym.

2. Wskaźniki liczby skupień

Wskaźniki liczby skupień są konstruowane na podstawie różnych kryteriów wyróżniania skupień. Można je pogrupować następująco: 1. Wskaźniki uwzględniające jedynie odległości między obiektami w skupieniu i między skupieniami (między obiektami z różnych skupień, między centrami skupień). Należą do nich m.in.: indeks Dunna, indeks *silhouette*, indeks Huberta-Lewina i *cluster separation index* (CSI). 2. Wskaźniki uwzględniające rozproszenie obiektów wewnątrz skupień i między skupieniami (między obiektami z różnych skupień, między centrami skupień), np. indeks Daviesa-Bouldina. 3. Wskaźniki związane z macierzą kowariancji całego zbioru danych i macierzami kowariancji dla skupień. Należą do nich: indeks Calinskiego-Harabasza, indeks Balla-Halla, indeks Hartigana. 4. Wskaźniki związane z macierzą rozrzutu (*scatter matrix*) całego zbioru danych i macierzami rozrzutu obiektów w skupieniach, indeks Scotta-Symonsa, indeks Marriota, i indeksy Friedmana-Rubina: indeks TraceCovW, TraceW, Trace$W^{-1}B$. 5. Wskaźniki mieszane i inne (np. oparte na rozmytej funkcji przynależności do skupień). Należą do nich: indeks GAP, indeks Bezdeka, entropia klasyfikacyjna i inne.

W opracowaniu uwzględnione będą indeksy należące do pierwszych 4 grup. Indeksy należące do grupy 5 zostały pominięte, ponieważ w zbiorach binarnych trudno mówić o rozmyciu granic między skupieniami (por.: [Najman, Najman 2005]). Indeksy Dunna, *silhouette*, CSI, Daviesa-Bouldina były od strony formalnej szczegółowo omówione w pracach [Najman, Najman 2005; 2006]. Formalny opis tych indeksów zostanie pominięty.

Indeks Huberta-Levina (znany w literaturze także pod nazwą *C-Index*), zaproponowany przez autorów w roku 1976 [Hubert, Levin 1976], przyjmuje postać:

$$HL = \frac{d_w - \min(d_w)}{\max(d_w) - \min(d_w)}, \tag{1}$$

gdzie: d_w jest sumą wewnętrznych odległości w skupieniach. Podział obiektów minimalizujący wartość HL jest uznawany za optymalny.

Indeks Calinskiego-Harabasza zaproponowany został w roku 1974 [Calinski, Harabasz 1974] i jest definiowany następująco:

$$CH = \frac{trace(\mathrm{cov}(\mathbf{X}))}{k-1} \Bigg/ \frac{\sum_{i=1}^{k} trace(\mathrm{cov}(X_i))}{n-k}, \tag{2}$$

gdzie: $\mathbf{X}$ – macierz obserwacji, X_i – obiekty należące do i-tego skupienia, k – liczba skupień, n – liczba obserwacji[2].

Indeks Hartigana został zaproponowany w roku 1975 [Hartigan 1975] i definiowany jest następująco:

$$HA = \log\left(trace(\mathrm{cov}(\mathbf{X})) \Big/ \sum_{i=1}^{k} trace(\mathrm{cov}(X_i)) \right), \tag{3}$$

gdzie: $\mathbf{X}$ – macierz obserwacji, X_i – obiekty należące do i-tego skupienia.

Indeks Balla-Halla został zaproponowany w roku 1965 [Ball, Hall 1965] i jest definiowany następująco:

$$BH = \sum_{i=1}^{k} trace(\mathrm{cov}(X_i)) \Big/ k, \tag{4}$$

gdzie: X_i – obiekty należące do i-tego skupienia, k – liczba skupień.

Indeks Scotta-Symonsa należy do 4 grupy wskaźników. Zaproponowany został w 1971 r. [Scott, Symons 1971] i definiowany jest następująco:

$$SS = n \log(\det(T)/\det(W)), \tag{5}$$

gdzie: $W = \sum_{k} W_k$ – wewnętrzna macierz rozrzutu (*pooled-within groups scatter matrix*), $W_k = \sum_{l=1}^{n_k} (X_{lk} - C_k)^T (X_{lk} - C_k)$ – macierz rozrzutu dla obiektów ze skupienia k, (*scatter matrix*), $T = W + B$ – łączna macierz rozrzutu (*total scatter matrix*), $B = \sum_{k} n_k C_k^T C_k$ – międzygrupowa macierz rozrzutu (*between groups scatter matrix*).

Indeks Marriota zaproponowano także w roku 1971 [Mariot 1971] i definiowany jest następująco:

$$MA = k^2 \det(W). \tag{6}$$

---

[2] Indeks ten jest często błędnie zapisywany w uproszczonej formie: $CH = \frac{trace(B)}{k-1} \Big/ \frac{trace(W)}{n-k}$, gdzie B to zewnętrzna suma kwadratów, a W to wewnętrzna suma kwadratów. W rzeczywistości w mianowniku powinna być suma po wszystkich skupieniach. Co więcej, poza oryginalnym artykułem trudno znaleźć odpowiedź na pytanie: suma kwadratów „czego" jest wyznaczana do B i W.

Cztery indeksy Friedmana-Rubina należą do 4 grupy indeksów i były proponowane w latach 1965, 1967, 1971 [Edwards, Cavalli-Sforza 1965; Friedman, Rubin 1967]. Nie mają nazw własnych od nazwisk twórców, ponieważ nad ich rozwojem pracowało wielu autorów. Wydaje się, że największy wkład w tym zakresie mieli właśnie Friedman i Rubin. Nazwy indeksów pochodzą wprost od ich formalnego zapisu i definiuje się je następująco:

$$\text{TraceCovW} = trace\left(\text{cov}(W)\right), \tag{7}$$

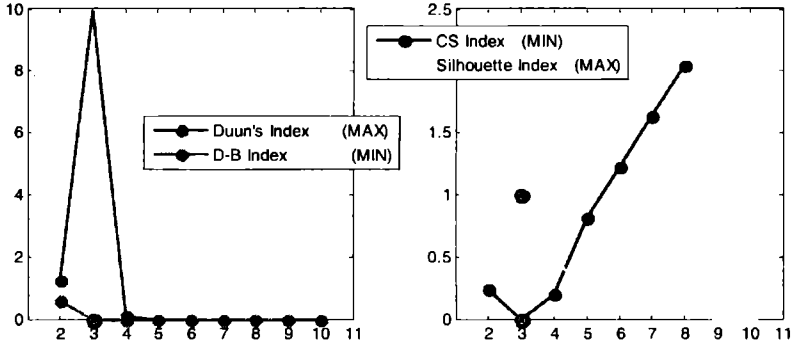
$$\text{TraceW} = trace(W), \tag{8}$$

$$\text{TraceW}^{-1}\text{B} = trace\left(W^{-1}B\right), \tag{9}$$

$$FR2 = \det(T)/\det(W). \tag{10}$$

3. Ustalanie optymalnej liczby skupień na podstawie badanych indeksów

Ustalanie wartości wskaźników wskazujących na optymalną liczbę skupień najczęściej nie jest jednoznaczne. W zasadzie nie ma większych kontrowersji w odniesieniu do indeksów związanych z odległościami i rozproszeniem obiektów. W większości badań teoretycznych i empirycznych przyjmuje się następujące kryteria: optymalna jest liczba skupień, gdy indeks przyjmuje wartość: maksymalną dla indeksu Dunna, maksymalną dla indeksu *silhouette*, minimalną dla indeksu *cluster separation*, minimalną dla indeksu Huberta-Lewina oraz minimalną dla indeksu Daviesa-Bouldina. Łatwość interpretacji wynika z tego, że przebieg wartości tych indeksów dla kolejnych podziałów zbioru obiektów jest monotoniczny w stosunku do wyróżnionego ekstremum. Na rysunku 1 przedstawiony jest typowy przebieg wartości tych indeksów dla przykładowego zbioru danych grupowanego na od 2 do 10 skupień metodą k-średnich.



Rys. 1. Przykładowy przebieg wartości wskaźników jakości grupowania
Źródło: opracowanie własne.

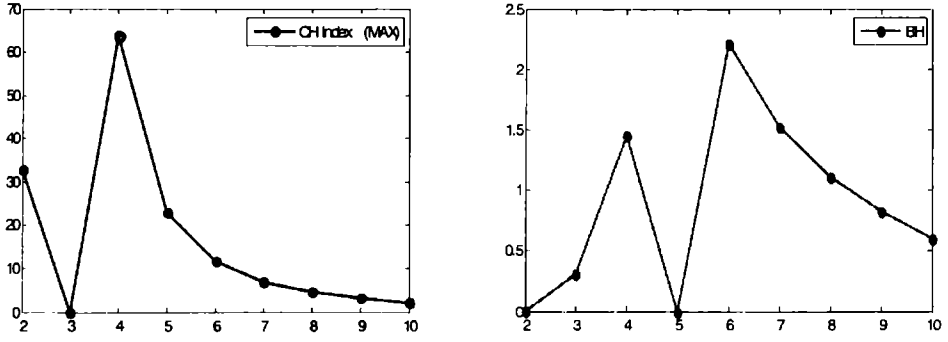
Niestety w przypadku indeksów z pozostałych grup nie ma tak jednoznacznych kryteriów. Spowodowane to jest charakterem przebiegu wartości tych indeksów dla testowanych podziałów zbiorowości obiektów. Zwykle nie mają one tak monotonicznego przebiegu jak poprzednie. Często pojawiają się naprzemiennie występujące niskie i wysokie wartości wskaźników. Przebieg wskaźników Calinskiego-Harabasza i Balla-Halla dla tego samego zbioru przykładowych danych przedstawia rys. 2.

Zwykłe kryterium maksymalizacji czy minimalizacji wartości wskaźnika w tym wypadku nie wystarczy. Proponuje się inne kryteria bazujące na największej zmianie wartości wskaźnika. Do częściej stosowanych kryteriów można zaliczyć następujące kryteria:

a) największej/najmniejszej lewostronnej zmiany wartości wskaźnika, $X_n - X_{n-1}$,

b) największej/najmniejszej prawostronnej zmiany wartości wskaźnika, $X_n - X_{n+1}$,

c) największej/najmniejszej obustronnej zmiany wartości wskaźnika, $(X_{n+1} - X_n) - -(X_n - X_{n-1})$.



Rys. 2. Przykładowy przebieg wartości wskaźników jakości grupowania
Źródło: opracowanie własne.

Tabela 1. Ustalanie optymalnej liczby skupień dla indeksów z grupy 1 i 2

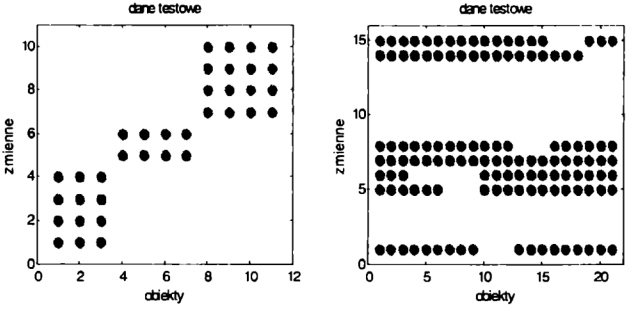
Indeks	Kryterium	Indeks	Kryterium
Calinskiego-Harabasza	c/min.	TraceCovW	a/min.
Hartigana	c/min.	TraceW	c/max
Balla-Halla	c/max	$\text{TraceW}^{-1}\text{B}$	a/max
Scotta-Symonsa	a/max	FR	c/min.
Marriota	c/max	Huberta-Levina	c/max

Źródło: opracowanie własne.

W literaturze brak jest jednoznacznych wskazówek, które kryterium należy zastosować dla danego wskaźnika w danej sytuacji. W konsekwencji są one wybierane subiektywnie na podstawie analizy wykresów przedstawiających ich przebieg. Kryteria przyjęte w prezentowanych badaniach prezentuje tab. 1.

4. Wyniki badań symulacyjnych

W celu weryfikacji zdolności badanych wskaźników do dostarczania prawidłowej informacji o liczbie skupień w zbiorze obiektów zaprojektowano eksperyment badawczy. Wygenerowano 7 zbiorów testowych o założonej strukturze grupowej, zawierających od 2 do 15 skupień. Przykładowe zbiory nr 2 i 5 przedstawia rys. 3. Poszczególne skupienia różnią się od siebie informacją od 1 do 10 cech[3]. Dane dzielono według grupowania hierarchicznego i metodą k-średnich. W grupowaniu hierarchicznym zastosowano metodę środka ciężkości i odległość Jaccarda. W metodzie k-średnich zastosowano odległość Hamminga. W odniesieniu do obu metod grupowania budowano od 2 do 20 skupień. Zgodność uzyskanych wyników grupowania ze znanym wzorcem testowano, opierając się na wskaźnikach Randa, korygowanym Randa, Jaccarda i Fowklesa-Mallowsa.



Rys. 3. Dane testowe, zbiory 2 i 5
Źródło: opracowanie własne.

Tabela 2. Syntetyczne wyniki badań dla grupowania hierarchicznego

	PROPONOWANA LICZBA SKUPIEŃ						
INDEX	zbiór1	zbiór2	zbiór3	zbiór4	zbiór5	zbiór6	zbiór7
silhouette	3	3	5	5	7	15	16
Dunn	3	3	5	5	2	2	2
D-B	3	3	5	5	7	15	16
CS	3	3	5	5	7	15	16
CH	3	3	3	3	3	3	3
HA	3	3	3	3	3	3	3
BH	3	3	4	3	3	3	3
SS	3	3	3	3	3	3	10
MA	2	2	2	2	2	2	4
TCW	3	3	3	3	3	3	10
TW	3	3	4	5	7	15	10
FR	3	3	3	3	3	3	10
FR2	3	3	3	3	3	3	3
HL	2	2	3	3	4	13	9
Prawidłowa liczba skupień	3	3	5	5	7	15	14

Nazwy indeksów uproszczono do pierwszych liter. Indeks $\mathrm{Trace}W^{-1}B$ zapisano jako FR.

Źródło: opracowanie własne.

[3] W zbiorze o niewielkiej liczbie wymiarów, np. 10, binarne skupienia mogą się różnić nawet jednym bitem (informacją o jednej cesze, pozostałe cechy przyjmują w obu skupieniach identyczne wartości). Dla zbiorów o dużej liczbie wymiarów (ponad 100) liczba bitów różniących skupienia powinna być większa. W przeciwnym wypadku wyróżniona liczba skupień będzie zwykle bardzo duża – równa liczbie wariantów cech obiektów.

Tabela 3. Syntetyczne wyniki badań dla grupowania metodą k-średnich

	PROPONOWANA LICZBA SKUPIEŃ						
INDEX	zbiór1	zbiór2	zbiór3	zbiór4	zbiór5	zbiór6	zbiór7
Silhouette	3	3	5	5	7	15	14
Dunn	3	3	5	5	2	2	3
D-B	3	3	5	5	7	15	14
CS	3	3	5	5	7	15	14
CH	3	3	3	3	3	3	3
HA	3	3	3	3	3	3	3
BH	3	3	3	3	7	3	6
SS	3	3	3	3	3	3	7
MA	2	2	2	2	2	2	6
TCW	3	3	4	3	5	3	4
TW	3	3	4	5	5	15	3
FR	3	3	3	3	3	3	7
FR2	3	3	3	3	3	3	3
HL	2	2	3	3	5	13	8
Prawidłowa liczba skupień	3	3	5	5	7	15	14

Nazwy indeksów uproszczono do pierwszych liter. Indeks Trace$W^{-1}B$ zapisano jako FR2.

Źródło: opracowanie własne.

Dla każdego zbioru i dla obu metod grupowania budowano 3 szczegółowe tablice wynikowe. Pierwszą z wartościami wskaźników, drugą z sugerowaną liczbą skupień, trzecią z wartościami wskaźników porównań ze wzorcem. W sumie jest to 21 tablic[4], na podstawie których przeprowadzono wnioskowanie. Syntetyczne przedstawienie uzyskanych wyników znajduje się w tab. 2 i 3. Błędne wskazania wskaźników zaznaczono na szaro.

5. Wnioski

Żaden z badanych wskaźników nie był budowany z myślą o danych binarnych. Fakt ten znalazł odzwierciedlenie w prowadzonych badaniach empirycznych. Pojawiły się bowiem problemy specyficzne jedynie dla tego typu danych. Otóż, wartości żadnego z tych wskaźników nie można wyznaczyć, gdy wśród wyróżnionych skupień będzie choć jedno składające się z tylko jednego elementu. Dla danych wyrażonych w innych skalach taki obiekt zwykle zostanie wykryty i usunięty jako nietypowy. W odniesieniu do danych binarnych nie można tego zrobić, ponieważ taki obiekt może różnić się od innych tylko jednym bitem (wartością tylko jednej cechy) i trudno go uznać za nietypowy. Problem ten jest bardzo wyraźny dla metody k-średnich, która dla danych binarnych ma tendencję do wyróżniania skupień składających się wyłącznie z identycznych obiektów. Wtedy liczba skupień równa się liczbie wariantów cech pojawiających się w zbiorze danych. Problem ten ma dalsze konsekwencje. Jeżeli skupienie składa się z dowolnej liczby identycznych obiektów większej niż 1, to nie można wyznaczyć wartości badanych wskaźników.

[4] Ze względu na ograniczenia objętości publikacji nie zamieszczono tablic szczegółowych. Znajdują się one na stronie WWW autora pod adresem: http://panda.bg.univ.gda.pl/~Najman.

W przypadku wskaźników opartych na macierzy rozrzutu wewnętrzna macierz rozrzutu będzie osobliwa, a wskaźniki przyjmą wartość nieokreśloną[5]. Dla wskaźników opartych na macierzy kowariancji ślad macierzy kowariancji będzie równy 0, a wartość wskaźników będzie także nieokreślona. Dla wskaźników opartych na odległościach suma odległości wewnętrznych skupień jest równa zeru. Nie można wtedy wyznaczyć wartości wskaźników z powodu problemu dzielenia przez zero. Podobnie w odniesieniu do wskaźników opartych na rozproszeniu obiektów w skupieniu. Rozproszenie jest także równe zeru i nie można wyznaczyć wartości wskaźników.

Kolejny problem, także typowy jedynie dla danych binarnych, to możliwość pojawienia się skupień pustych. Stosując klasyczną metodę k-średnich, może się tak zdarzyć. Jeżeli w zbiorze danych są np. dwa skupienia składające się z identycznych elementów, to obiektów nie można podzielić metodą k-średnich na 3 skupienia. Problem ten sumuje się z poprzednimi i uniemożliwia wyznaczenie wartości wskaźników.

Kłopotem wskazywanym wcześniej jest także subiektywność ustalania kryterium optymalności wskaźników. Kryteria przyjęte w prezentowanej pracy są tylko jedną z możliwych opcji, a brak jest teoretycznych podstaw do dokonania obiektywnego wyboru.

Z przedstawionych powodów w analizie danych binarnych żadnego ze wskaźników nie można uznać za dobry. Duże rozbieżności między prawidłową liczbą skupień a liczbą wyznaczoną na podstawie wskaźników wynikają głównie z przyczyn numerycznych wskazanych w artykule. Przeprowadzone badanie symulacyjne wskazuje, że względnie najlepsze są wskaźniki: indeks *silhouette*, CSI i indeks Daviesa-Bouldina, które uzyskały identyczne wyniki.

Prezentowany w pracy problem z całą pewnością nie został rozwiązany. Żaden z indeksów omawianych w artykule nie może być rekomendowany przez autora do zastosowań empirycznych. Niezbędne wydaje się skonstruowanie specjalnego wskaźnika uwzględniającego specyfikę danych binarnych.

Literatura

Ball G., Hall D.J. (1965), *ISODATA, A Novel Method of Data Analysis and Pattern Classification*, Stanford Research Institute, Menlo Park.

Calinski R.B., Harabasz J. (1974), *A Dendrite Method for Cluster Analysis*, „Communications in Statistics" 3, s. 1-27.

Davies D.L., Bouldin D.W. (1979), *A Cluster Separation Measure*, „IEEE Transactions on Pattern Analysis and Machine Intelligence" 1, s. 224-227.

[5] Stwarza to dodatkowe problemy informatyczne. Ten sam algorytm dla różnych języków programowania będzie dawał różne wyniki. W jednych będzie to „inf", w innych „NaN". Czasami pojawi się komunikat błędu dzielenia przez zero.

Dolnicar S., Leisch F., Weingessel A., Buchta C., Dimitradou E. (1998), *A Comparison of Several Cluster Algorithms on Artificial Binary Data Scenarios from Tourism Marketing*, Working Paper 7, SFB, Adaptive Information Systems and Modeling in Economics and Management Science.

Edwards A.W.F., Cavalli-Sforza L. (1965), *A Method for Clustering Analysis*, „Biometrics" 21, s. 362-375.

Friedman H.P., Rubin J. (1967), *On Some Invariant Criteria for Grouping Data*, „Journal of the American Statistical Association" 62, s. 1159-1178.

Hartigan J.A. (1975), *Clustering Algorithms*, Wiley, New York.

Hubert L.J, Levin J.R. (1976), *A General Statistical Framework for Assessing Categorical Clustering in Free Recall*, „Psychological Bulletin" 83, s. 1072-1080.

Marriot F.H.C. (1971), *Practical Problems in a Method of Cluster Analysis*, „Biometrics" 27, s. 501-514.

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

Najman K., Najman K. (2005), *Analityczne metody ustalania liczby skupień*, Prace Naukowe AE we Wrocławiu, *Taksonomia 12, Klasyfikacja i analiza danych – teoria i zastosowania*, AE, Wrocław.

Najman K., Najman K. (2006), *Analityczne metody ustalania liczby skupień w rozmytych zbiorach danych*, Prace Naukowe AE we Wrocławiu, *Taksonomia 13, Klasyfikacja i analiza danych – teoria i zastosowania*, AE, Wrocław.

Scott A.J., Symons M.J. (1971), *Clustering Methods Based on Likelihood Ratio Criteria*, „Biometrics" 27, 387-397.

DETERMINING THE NUMBER OF CLUSTERS IN BINARY DATA SETS

Summary

In this paper the performance of fourteen indexes for determining the number of clusters in a binary data set is analyzed. To ensure that the right number of clusters is known, only artificial sets, designed to simulate data, are used. The resultant optimal clusters have been found to be stable for the different validity indices used, e.g.: Global Silhouette Index, Hubert-Lewin Index, Calinski-Harabasz Index, Ball-Hall Index, Hartigan Index and others. For the evaluation of the performance of the indexes, k-means and hierarchical algorithms are applied. The selection of the number of clusters based on the indexes values for the different number of clusters is done in an automatic way. It was shown that these indexes mightn't support the prediction of the optimal cluster partitioning for those binary data sets.