PRZEGLĄD STATYSTYCZNY
R. LXI – ZESZYT 2 – 2014

JERZY KORZENIEWSKI

# INDEKS WYBORU LICZBY SKUPIEŃ W ZBIORZE DANYCH

## 1. WSTĘP

Wyznaczenie liczby skupień (klas), na które należy podzielić zbiór obiektów jest jednym z etapów analizy skupień warunkującym jakość podziału. Większość indeksów wyznaczających liczbę skupień ma charakter optymalizacyjny, tj. dla ustalonej metody grupowania obiektów zbioru wyznaczana jest najlepsza (spośród liczb od 1 do np. 15) liczba skupień. Wśród najczęściej stosowanych wymienić należy indeksy: Bakera-Huberta, Calińskiego–Harabasza, Dunna, Daviesa-Bouldina, Hartigana, Huberta-Levine'a, Krzanowskiego-Lai, indeks sylwetkowy, indeks Gap. Odrębną grupę tworzą indeksy opracowane tylko pod kątem metod aglomeracyjnych np. Mojeny (1977). Dla metod aglomeracyjnych, Sokołowski (1992) wyróżnia aż pięć różnych grup indeksów liczby skupień. Efektywność wymienionych i innych indeksów badana była przez wielu autorów m.in. Milligan, Cooper (1985), Migdał-Najman, Najman (2005, 2006), Korzeniewski (2005). Wybór właściwych indeksów służących do oceny liczby skupień oraz samej liczby skupień nie jest łatwy. W badaniu Milligana, Cooper (1985) najlepszym indeksem okazała się miara Calińskiego-Harabasza (1974). Badanie Milligana miało jednak miejsce prawie 30 lat temu. Od tego czasu zaproponowano kilka nowych miar, które, zdaniem ich twórców, nie ustępują temu indeksowi. Przykładem później skonstruowanego indeksu, który w badaniu autorów okazał się lepszym od m.in. indeksu Calińskiego-Harabasza i Krzanowskiego-Lai, może być indeks Gap (Tibshirani i inni, 2001). Te dwa indeksy tj. Gap oraz Calinski-Harabasz, będą punktem odniesienia, z którym zostanie porównana efektywność nowego indeksu zaproponowanego w dalszej części artykułu. Wartość indeksu Calińskiego-Harabasza dla liczby skupień równej $k$ dana jest wzorem:

$$CH(k)=\frac{tr(\mathbf{B}_k)/(k-1)}{tr(\mathbf{W}_k)/(n-k)}\ , \tag{1}$$

gdzie $\mathbf{B}_k$ – macierz kowariancji międzygrupowych, $\mathbf{W}_k$ – macierz kowariancji wewnątrzgrupowych, $n$ – liczba obiektów w zbiorze. Za wybraną liczbę skupień należy uznać liczbę $k$, która daje maksymalną wartość wyrażenia (1). Jak widać ze

wzoru (1), nie można przy pomocy tego indeksu rozstrzygnąć czy zbiór danych należy w ogóle dzielić na jakieś skupienia, gdyż *k* musi być większe od 1.

Wartość indeksu Gap dla liczby skupień równej *k* jest wyznaczana przy pomocy wyrażenia:

$$Gap(k) = \frac{1}{B}\sum_{b} \log\left(tr\left(\mathbf{W}_{kb}\right)\right) - \log\left(tr\left(\mathbf{W}_{k}\right)\right), \tag{2}$$

gdzie $\mathbf{W}_{kb}$ macierz kowariancji wewnątrzgrupowych wtedy, gdy każda zmienna oryginalna zastąpiona została zmienną wygenerowaną (w *b*-tym powtórzeniu) z rozkładu jednostajnego nad odcinkiem, który jest rozstępem próby zmiennej oryginalnej; *B* – liczba powtórzeń generowania zmiennych jednostajnych. Za wybraną liczbę skupień należy uznać najmniejszą liczbę *k* która spełnia warunek:

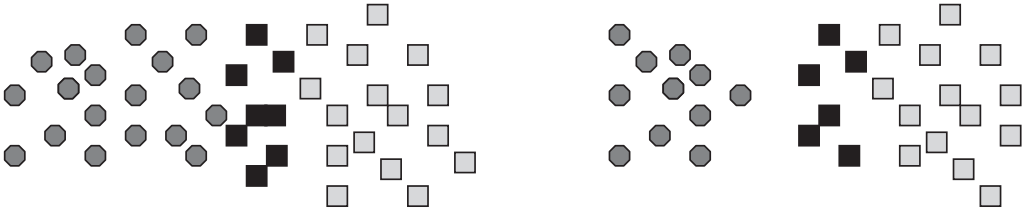
$$Gap(k) \geq Gap(k+1) - s_{k+1}, \tag{3}$$

gdzie $s_k = sd_k\sqrt{1+1/B}$, gdzie $sd_k$ oznacza odchylenie standardowe *B* wartości $\log\left(tr\left(\mathbf{W}_{kb}\right)\right)$. Zamiast rozkładu jednostajnego można zastosować inną technikę szacowania zlogarytmowanego śladu macierzy kowariancji zmiennych nietworzących struktury skupień wykorzystującą kształt rozkładu brzegowego poszczególnych zmiennych oryginalnych. Jak widać z konstrukcji indeksu może on być używany również do zbadania czy zbiór danych należy dzielić na jakiekolwiek skupienia, gdyż *k* może być równe 1. W eksperymencie symulacyjnym przyjęto *B* = 100 powtórzeń.

## 2. FORMUŁA NOWEGO INDEKSU

Nowy indeks oparty jest na wielostopniowym dzieleniu zbioru danych (bądź części zbioru) na dwa skupienia i zachowywaniu tego podziału, gdy jest on wystarczająco dobry tzn. oba skupienia są wyraźnie rozdzielone. Jako metoda podziału na każdym etapie algorytmu stosowana będzie metoda *k*-średnich (*k*=2 lub *k*=3), ze stukrotnym losowym wyborem punktów startowych, ale możliwe jest użycie dowolnej innej metody grupowania danych. Idea nowego indeksu jest prosta. Wyobraźmy sobie, że dzielimy zbiór danych na dwa skupienia, w następnym kroku każde z nich na dalsze dwa itd. Postępując w ten sposób powinniśmy otrzymać poprawną liczbę skupień wyraźnie separowanych miedzy sobą jeśli tylko będziemy dysponowali dobrą miarą jakości podziału na dwa skupienia. Miara taka określi kiedy proces podziału będzie należało przerwać. Konieczne wydaje się również przyjęcie założenia o minimalnej liczebności jednego skupienia.

W celu rozstrzygnięcia czy dwa skupienia są wystarczająco dobrze separowane zdefiniujmy następującą miarę jakości podziału. Załóżmy, że ustalony podzbiór zbioru danych podzielony został w podziale pierwotnym na dwa skupienia $S_1$ oraz $S_2$. Rozważamy teraz nowy podzbiór składający się z mniejszego liczebnościowo z tych dwóch skupień oraz z 1/3 większego skupienia. Nowy podzbiór dzielimy ponownie na dwa skupienia i miarą jakości podziału (pierwotnego) jest skorygowany indeks Randa (por. np. Gatnar, Walesiak, 2004) zgodności podziału pierwotnego z podziałem ponownym. Miarę tę oznaczymy symbolem $R(S_1, S_2)$. Interpretacja graficzna przedstawiona jest na rys. 1. Dla zbioru obiektów po prawej stronie (dwa skupienia) wartość miary powinna być wysoka (bliska 1), zaś dla zbioru niemającego wyraźnej struktury skupień wartość miary powinna być niska (bliska 0), jeśli bowiem nie ma żadnej struktury skupień, to nie ma powodu po temu by podzbiór został podzielony w „tym samym miejscu". Oczywiście, indeks Randa ocenia zgodność podziału pierwotnego z ponownym tylko na podzbiorze składającym się z obiektów mniejszego skupienia i 1/3 większego skupienia. Jeżeli ustalonym podzbiorem, na którym dokonywano podział pierwotny był cały zbiór obiektów, to wartość $R(S_1, S_2)$ jest ostateczną miarą jakości podziału.



Rysunek 1. Ilustracja graficzna miary $R(S_1, S_2)$
Źródło: opracowanie własne.

Dla zbioru obiektów po prawej stronie zgodność podziału podzbioru złożonego z kółek i czarnych kwadratów z podziałem całego zbioru powinna być wysoka.

Jeżeli jednak oprócz skupień $S_1$ oraz $S_2$ były jeszcze jakieś inne skupienia (podzbiory) całego zbioru danych, to wartość $R(S_1, S_2)$ nie może być miarą ostateczną, bo może zdarzyć się, że jedno ze skupień $S_1$ lub $S_2$ (lub oba) jest „blisko" pozostałej części zbioru danych (lub któregoś z pozostałych skupień). Liczbę $R(S_1, S_2)$ można przyjąć jako miarę jakości podziału dopiero gdy przynajmniej jedno ze skupień $S_1$ lub $S_2$ ma taką samą lub wyższą miarę jakości podziału na dwa skupienia ze wszystkimi pozostałymi skupieniami. Wobec tego ostateczną miarą jakości podziału ustalonego podzbioru zbioru obiektów na 2 skupienia, przy założeniu, że oprócz ustalonego podzbioru istnieje jeszcze $k$-2 innych skupień (o numerach 3, 4, …, $k$), niech będzie wartość:

$$R(S_1,S_2,k)=\min\left\{ R(S_1,S_2), \max_{i=1,2}\left\{ \min_{\substack{j=1,\ldots,k\\ j\neq 1}} R(S_1,S_j), \min_{\substack{j=1,\ldots,k\\ j\neq 2}} R(S_2,S_j) \right\} \right\}. \qquad (4)$$

Przykładowo, pierwsze minimum w wewnętrznym nawiasie klamrowym ma wyłonić najsilniejszy związek skupienia $S_1$ z jakimś spośród skupień $S_3$ $S_4$,…,$S_k$. Analogiczne jest znaczenie drugiego minimum w wewnętrznym nawiasie klamrowym. Po wybraniu z tych dwóch wartości miar związków z pozostałymi skupieniami wartości większej (i co za tym idzie skupienia $S_1$ lub $S_2$ ), porównujemy tę wartość z wartością miary $R(S_1, S_2)$ wybierając mniejszą z tych dwóch wartości. W uzupełnieniu wzoru (4) należy dodać, że jeżeli $k$=2 czyli, gdy w pierwszym kroku dzielimy cały zbiór obiektów, to $R(S_1, S_2,k) = R(S_1, S_2)$ .

Miarę podobną do miary (4) zastosowano w pracy Korzeniewski (2012) do identyfikacji zmiennych tworzących strukturę skupień. Miara ta spisuje się najsłabiej gdy w zbiorze danych istnieją tylko dwa skupienia. W takich przypadkach efektywność miary (4) nieco poprawia wielokrotne jej obliczenie dla zbiorów danych „podobnych" do oryginalnego zbioru, na przykład, zbiorów, które powstają poprzez zastąpienie każdego oryginalnego obiektu obiektem leżącym „blisko" oryginalnego. Takie zbiory można generować, na przykład, według wzoru:

$$x'_v = x_v + 0{,}2 \cdot r\,, \qquad (5)$$

gdzie $v$ – numer zmiennej, $x'_v$ – nowa wartość (na zmiennej $v$) zastępująca oryginalną wartość $x_v$, $r$ – liczba wylosowana z odcinka [0;1] (niezależnie dla każdej zmiennej $v$ i każdego obiektu). Po unitaryzacji zerowanej (por. opis eksperymentu) wszystkie wartości każdej zmiennej $v$ leżą na odcinku [0;1] więc formułę (5) można uznać za generowanie losowe obserwacji leżącej „blisko" obserwacji oryginalnej. Z badań symulacyjnych opisanych w pracy Korzeniewski (2012) wynika, że wielokrotne generowanie zbiorów podobnych do oryginalnego niekiedy poprawia efektywność miary (4) dzięki temu, że niweluje niekorzystny rozkład wartości cech w obszarach granicznych pomiędzy dwoma skupieniami. Działanie algorytmu nie będzie silnie uzależnione od tego jak zmieni się zbiór danych pod wpływem transformacji (5), ponieważ transformacja ta zostanie powtórzona wielokrotnie i z wielu powtórzeń będzie wybierany wynik dominujący. Dokładniej rzecz ujmując, algorytm zostanie zastosowany do każdego z 20 zbiorów danych po transformacji (5) i za ostateczną liczbę skupień uznamy liczbę występującą najczęściej w ciągu otrzymanych 20 kandydatek. Ten sposób postępowania wprowadza pewną losowość do algorytmu, ale zauważmy, że losowość istnieje w samej metodzie $k$-średnich, gdy obiekty startowe wybierane są losowo.

W celu wykorzystania miary (4) do ustalania liczby skupień należy zaproponować algorytm wielostopniowego dzielenia zbioru danych na dwa skupienia, gdyż

kolejność dzielenia poszczególnych skupień na ewentualne dwa lub trzy mniejsze (sposób przeszukiwania zbioru) ma znaczenie dla ostatecznego wyniku. Progiem dla miary (4), od którego akceptujemy podział jednego skupienia na dwa skupienia będzie wartość 0,4. Ta liczba została ustalona metodą poszukiwań empirycznych w pracy Korzeniewski (2012) w celu identyfikacji zmiennych tworzących strukturę skupień. Wartość progowa 0,4 dała dobre wyniki dla struktur skupień generowanych z mieszanin rozkładów normalnych. Zaproponujmy następujący algorytm wielostopniowego dzielenia zbioru danych na dwa skupienia.

*Krok 1.* Połóż $K$=1 tj. cały zbiór obiektów potraktuj jako jedno skupienie.

*Krok 2.* Każde skupienie o numerze $k$, $k$=1,2,…,$K$ podziel na dwa skupienia i oznacz symbolem $m_k$ wartość miary (4) dla tego podziału jeżeli liczebności obu mniejszych skupień są równe co najmniej 5% liczebności zbioru obiektów. Jeżeli choć jedno ze skupień ma mniejszą liczebność, to przyjmij $m_k = 0$ .

*Krok 3.* Spośród liczb $m_k$ wybierz największą odpowiadającą skupieniu $k_0$ i jeśli jest ona większa od 0,4 to zwiększ liczbę $K$ skupień o 1 zastępując skupienie $k_0$ dwoma mniejszymi, na które zostało ono podzielone. Idź do kroku 2. Jeśli największa spośród liczb $m_k$ jest mniejsza od 0,4 to idź do kroku 4.

*Krok 4.* Pojedynczo, dla każdego ze skupień o numerze $k$, $k$=1,2,…,$K$ przeprowadź następującą procedurę. Podziel skupienie na dwa skupienia $A_k$ oraz $B_k$ (podział wstępny, po którym będzie $K$+1 skupień), a następnie, każde z tych dwóch skupień na kolejne dwa skupienia (podział drugi, po którym będzie $K$+3 skupień). Jeżeli liczebności obu skupień z podziału drugiego są równe co najmniej 5% liczebności zbioru obiektów, to znajdź wartość miary (4) dla podziałów drugich obu skupień $A_k$ oraz $B_k$. Jeżeli liczebność choćby jednego z dwóch skupień drugiego podziału jest niższa od 5% liczebności zbioru obiektów, to połóż wartość miary (4) równą 0.

Spośród dwóch wartości miary (4) (jednej dla skupienia $A_k$ drugiej dla $B_k$) wybierz większą wyodrębniającą skupienie $k'$ ($k$) (z drugiego podziału) i oznacz ją symbolem $m_k$. Spośród liczb $m_k$ dla $k$=1,2,…,$K$, wybierz największą odpowiadającą skupieniu $k_0$.

Jeśli wartość ta jest większa od 0,4 to zmień podział obu skupień $A_{k_0}$ oraz $B_{k_0}$ dzieląc je na skupienie $k'$ ($k_0$) oraz skupienie składające się z drugiego skupienia z podziału pierwszego (tego spośród $A_{k_0}$, $B_{k_0}$ które miało mniejszą wartość miary (4)) i drugiego, różnego od $k'$ ($k_0$), skupienia z podziału drugiego odnośnego skupienia. Połóż $K$+2 jako liczbę skupień. Idź do kroku 2.

Jeśli największa spośród liczb $m_k$ jest mniejsza od 0,4 to idź do kroku 5 zachowując taki podział całego zbioru obiektów na $K$ skupień jaki był na początku kroku 4.

*Krok 5.* Pojedynczo, dla każdego ze skupień o numerze $k$, $k$=1,2,…,$K$ przeprowadź analogiczną procedurę do tej z kroku 4 dzieląc każde skupienie w podziale wstępnym na trzy skupienia. Jeśli któreś z tych trzech skupień da się podzielić na dwa skupienia spełniając te same kryteria co w kroku 4, to idź do kroku 2. Jeśli nie, to idź do kroku 6 zachowując taki podział całego zbioru obiektów na $K$ skupień jaki był na początku kroku 5.

*Krok 6.* Aktualną bieżącą liczbę *K* skupień uznaj za ostateczną. Wygeneruj nowy zbiór danych zgodnie z formułą (5) i idź do kroku 1.

*Krok 7.* Powtórz 20 razy kroki 1–6. Spośród 20 liczb kandydatek na liczbę skupień wybierz liczbę dominującą. Jeśli nie ma liczby dominującej, to jeszcze raz wygeneruj zbiór danych zgodnie z formułą (5), powtórz kroki 1–6 otrzymując kandydatkę numer 21 i wybierz tę spośród kilku najczęściej powtarzających się wartości w ciągu 20 liczb, która jest najbliższa kandydatce numer 21.

Należy zwrócić uwagę na to, że występujące w kroku 4 i kroku 5 podziały pierwsze na dwa lub trzy skupienia są podziałami tymczasowymi i nie oceniamy ich przy pomocy miary (4). Natomiast każde ze skupień powstałych w wyniku podziału tymczasowego próbujemy dzielić dalej na dwa skupienia i dopiero te podziały są oceniane i z nich wybierany jest najlepszy. Podziały tymczasowe są z powrotem scalane, za wyjątkiem skupienia $k'$ ($k_0$) (jeżeli takie zostało wyodrębnione). Wprowadzenie podziałów tymczasowych na dwa lub trzy skupienia jest konieczne, gdyż wielostopniowe dzielenie na tylko dwa skupienia nie da dobrych efektów w przypadku struktury składającej się z kilku skupień (por. rys. 2). Podział na dwa skupienia, z których każde będzie składało się z kilku mniejszych, może nie dać wartości miary (4) przekraczającej próg 0,4. Jeżeli zaś wprowadzimy wstępny, tymczasowy podział na dwa lub trzy skupienia i każde z nich będziemy próbowali dzielić na dwa, to jest większa szansa na to, że przy pomocy miary (4) wyodrębnimy jakieś skupienie wystarczająco dobrze separowane od wszystkich pozostałych.

Rysunek 2. Przykład zbioru obiektów składającego się z siedmiu skupień, który przy podziale na dwa skupienia może dać niską wartość miary $R(S_1, S_2)$

Źródło: opracowanie własne.

## 3. EKSPERYMENT BADAWCZY

Wszystkie zbiory danych składały się z obiektów opisanych przez zmienne ciągłe, generowanych przy pomocy algorytmu OCLUS (Steinley i Henson, 2005) – jednego z najnowszych ze znanych z literatury sposobów. W algorytmie tym separowalność skupień jest typu "łańcuchowego" tzn. na każdym wymiarze jest $k$-1 par skupień zachodzących na siebie w takim samym stopniu równym $(overlap)^{1/T}$ ($k$ – liczba skupień, $T$ – liczba wymiarów). Ogólny stopień pokrywania się skupień jest iloczynem stopni pokrywania się skupień na poszczególnych wymiarach, czyli

$$overlap = \prod_{v=1}^{T} overlap_v \,. \tag{6}$$

Poszczególne skupienia są generowane na każdej współrzędnej z rozkładów normalnych o jednostkowej wariancji i wartościach średnich zdeterminowanych przez stopień zachodzenia na siebie rozkładów generujących skupienia oraz przez wartość średnią rozkładu generującego pierwsze skupienie. Rozkład normalny generowany był według algorytmu Marsaglia-Bray (por. Wieczorkowski, Zieliński, 1997) z generatorem liczb pseudolosowych z języka Delphi4. Następnie, po wygenerowaniu wszystkich skupień (na każdej współrzędnej oddzielnie), zostały one w sposób losowy ponumerowane w sposób niezależny na każdej współrzędnej. W tabeli 1 podane są odległości pomiędzy wartościami średnimi kolejnych skupień w zależności od stopnia zachodzenia na siebie rozkładów generujących skupienia oraz liczby zmiennych. Dla $overlap$=0 przyjęto, że wartości średnie dwóch sąsiednich skupień są odległe o 6 odchyleń standardowych. Taki sposób generowania struktury skupień powoduje to, że otrzymujemy zbiory, których struktura jest efektem „równego" udziału wszystkich zmiennych. Tę cechę można uważać za wadę, wśród zbiorów danych empirycznych takiego równouprawnienia cech, na ogół, nie ma.

Program do generowania zbiorów danych oraz program realizujący zaproponowany algorytm napisane zostały samodzielnie w środowisku języka Delphi4.

Wszystkie zbiory składały się z 200 elementów podzielonych na kilka skupień. Poszczególne parametry wygenerowanych zbiorów danych przedstawione są poniżej.

Parametr pierwszy, liczba skupień mogła być równa 2, 3, 4, 6 lub 8.

Parametr drugi, liczebności skupień, miał trzy warianty: (a) równe liczebności wszystkich skupień; (b) 10% obserwacji i (c) 60% obserwacji w jednym skupieniu, a pozostałe skupienia w przybliżeniu równoliczne.

Parametr trzeci, liczba zmiennych, miał trzy warianty: 2, 4 lub 6.

Parametr czwarty, stopień pokrywania się skupień, miał pięć wariantów – 0; 0,1; 0,2; 0,3; 0,4.

Parametr piąty, siła korelacji wewnątrz skupień miała dwa warianty: (a) macierz kowariancji w każdym skupieniu była macierzą jednostkową; (b) w każdym skupieniu

była taka sama macierz kowariancji z jedynkami na przekątnej zaś poza przekątną, liczbą wylosowaną z przedziału [0,3; 0,8].

Razem wszystkie kombinacje wariantów parametrów dały liczbę 450 zbiorów. Ten zestaw powtórzono 10 razy co dało ostateczną liczbę 4500 zbiorów.

Tabela 1.

Odległości pomiędzy wartościami średnimi rozkładów normalnych dwóch sąsiednich skupień, w zależności od stopnia pokrywania się skupień (*overlap*) oraz liczby *T* zmiennych opisujących obiekty

| Overlap | *T*=2 | *T*=4 | *T*=6 |
|---|---|---|---|
| 0 | 6 | 6 | 6 |
| 0,1 | 2 | 1,16 | 0,82 |
| 0,2 | 1,52 | 0,86 | 0,60 |
| 0,3 | 1,2 | 0,66 | 0,46 |
| 0,4 | 0,96 | 0,51 | 0,35 |

Źródło: opracowanie własne.

W eksperymencie symulacyjnym stosowanie jakiejkolwiek normalizacji zmiennych nie jest konieczne, ale normalizacja zazwyczaj występuje w analizie skupień zbiorów empirycznych. Wybrana została formuła w postaci unitaryzacji zerowanej z dzieleniem przez rozstęp cechy, czyli, dla zmiennej o numerze *v*:

$$x_v^{'} = \frac{x_v - \min x_v}{\max x_v - \min x_v}. \tag{7}$$

Ta formuła ma dobrą opinię, ponadto po jej zastosowaniu można łatwo wprowadzić niewielkie, lokalne transformacje zbioru danych, takie jak, na przykład, transformacja dana wzorem (5).

Indeksy Calińskiego-Harabasza i Gap oceniane były dla dopuszczalnego zakresu liczby skupień od 1 (lub 2) do 10.

## 4. OCENA PORÓWNAWCZA EFEKTYWNOŚCI INDEKSU

Wyniki eksperymentu symulacyjnego są przedstawione w tabelach 2 i 3. Nowy indeks spisał się lepiej w przypadku struktur skupień bez korelacji wewnątrzklasowej. Taki wynik jest poprawny, gdyż zastosowana metoda grupowania *k*-średnich zwraca gorsze wyniki w przypadku skupień „wydłużonych". W porównaniu z kon-

kurencją nowy indeks spisał się bardzo dobrze. Z dwóch konkurencyjnych metod znacznie lepiej wypadł indeks Calińskiego-Harabasza, indeks Gap okazał się wiele słabszy w każdym przypadku. Nowy indeks traci znacznie wolniej efektywność wraz ze wzrostem stopnia rozmycia skupień czyli spadkiem separowalności skupień. W przypadku indeksu Calińskiego-Harabasza jest widoczny ostry wzrost średniego błędu przy wprowadzeniu niewielkiego rozmycia struktury skupień (*overlap*=0,1), w konsekwencji, średni błąd popełniany przez ten indeks jest wysoki. Nowy indeks osiągnął dużo lepszy wynik średni od indeksu Calińskiego-Harabasza. Indeks konkurencyjny jest nieznacznie lepszy tylko w przypadku struktur skupień bardzo wyraźnie separowalnych (*overlap*=0) przy towarzyszącym temu skorelowaniu wewnętrznym skupień. Takie wnioski można sformułować analizując odsetek poprawnie wybranych liczb skupień oraz wartość średnią popełnionego błędu. Przeanalizowane zostało również rozproszenia tych dwóch miar. Rozproszenie mierzone odchyleniem standardowym miary (odsetka poprawnie wybranych liczb skupień i wartości błędu) było dla wszystkich trzech indeksów podobne (w grupach zbiorów z jednakowymi parametrami).

Tabela 2.

Efektywność porównywanych metod w zależności od typu struktury skupień

| Typ zbioru danych | Metoda | Odsetek poprawnie znalezionych liczb skupień | Odsetek błędów równych 1 | Odsetek błędów równych 2 | Średnia arytmetyczna wartość błędu |
|---|---|---|---|---|---|
| Brak korelacji wewnątrz klas | **Nowy indeks** | **0,511** | **0,366** | **0,086** | **0,656** |
| | Indeks Gap | 0,334 | 0,191 | 0,130 | 2,119 |
| | Indeks CH | 0,477 | 0,194 | 0,154 | 1,299 |
| Korelacja wewnątrz klas | **Nowy indeks** | **0,413** | **0,398** | **0,136** | **0,845** |
| | Indeks Gap | 0,330 | 0,196 | 0,131 | 2,120 |
| | Indeks CH | 0,467 | 0,213 | 0,155 | 1,272 |
| Średnio | **Nowy indeks** | **0,462** | **0,382** | **0,111** | **0,751** |
| | Indeks Gap | 0,332 | 0,193 | 0,130 | 2,120 |
| | Indeks CH | 0,472 | 0,203 | 0,154 | 1,285 |

Źródło: obliczenia własne.

Tabela 3.

Średnia arytmetyczna błędu porównywanych metod w zależności od stopnia separowalności struktury skupień

| Typ zbioru danych | Metoda | *Overlap* równy 0 | *Overlap* równy 0,1 | *Overlap* równy 0,2 | *Overlap* równy 0,3 | *Overlap* równy 0,4 |
|---|---|---|---|---|---|---|
| Brak korelacji wewnątrz klas | **Nowy indeks** | **0,132** | **0,579** | **0,746** | **0,849** | **0,972** |
| | Indeks Gap | 1,279 | 2,157 | 2,269 | 2,406 | 2,485 |
| | Indeks CH | 0,211 | 1,143 | 1,387 | 1,771 | 1,983 |
| Korelacja wewnątrz klas | **Nowy indeks** | **0,373** | **0,777** | **0,948** | **0,999** | **1,131** |
| | Indeks Gap | 1,360 | 2,134 | 2,303 | 2,378 | 2,429 |
| | Indeks CH | 0,453 | 1,025 | 1,372 | 1,632 | 1,878 |
| Średnio | **Nowy indeks** | **0,253** | **0,678** | **0,847** | **0,924** | **1,052** |
| | Indeks Gap | 1,319 | 2,145 | 2,286 | 2,392 | 2,457 |
| | Indeks CH | 0,332 | 1,084 | 1,379 | 1,702 | 1,930 |

Źródło: obliczenia własne.

Przedmiotem dalszych badań będzie efektywność nowego indeksu przy zastosowaniu innej metody grupowania obiektów, mniej wrażliwej na nieregularny kształt skupień.

*Uniwersytet Łódzki*

LITERATURA

Caliński R. B., Harabasz J., (1974), A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 3, 1–27.

Gatnar E., Walesiak M., (red.), (2004), *Metody Statystycznej Analizy Wielowymiarowej w Badaniach Marketingowych*, Wydawnictwo AE we Wrocławiu.

Korzeniewski J., (2005), Propozycja nowego algorytmu wyznaczającego liczbę skupień, *Prace Naukowe AE we Wrocławiu nr 1076, Taksonomia 12*, 257–265.

Korzeniewski J., (2012), *Metody selekcji zmiennych w analizie skupień. Nowe procedury*, Wydawnictwo Uniwersytetu Łódzkiego.

Migdał-Najman K., Najman K. (2005), Analityczne metody ustalania liczby skupień, *Prace Naukowe AE we Wrocławiu nr 1076, Taksonomia 12*, 265–273.

Milligan G. W., Cooper M., (1985), An Examination of Procedures for Determining the Number of Clusters in a Data Set, *Psychometrika*, 2, 159–179.
Mojena R. (1977), Hierarchical Grouping Methods and Stopping Rules: an Evaluation, *Computer Journal*, 20 (4), 359–363.
Najman K., Migdał-Najman K., (2006), Wykorzystanie indeksu Silhouette do ustalania optymalnej liczby skupień, *Wiadomości Statystyczne*, 6, 1–10.
Sokołowski A., (1992), Empiryczne testy istotności w taksonomii, *Zeszyty Naukowe AE w Krakowie*, Seria specjalna: Monografie nr 108.
Steinley D., Henson R., (2005), OCLUS: An Analytic Method for Generating Clusters with Known Overlap, *Journal of Classification*, 22, 221–250.
Tibshrani R., Walther G., Hastie T., (2001), Estimating the Number of Clusters in a Dataset via the Gap Statistic, *Journal of the Royal Statistical Society*, 32, 411–423.
Wieczorkowski R., Zieliński R., (1997), *Komputerowe generatory liczb losowych*, Wydawnictwa Naukowo Techniczne, Warszawa.

## INDEKS WYBORU LICZBY SKUPIEŃ W ZBIORZE DANYCH

### Streszczenie

W artykule zaproponowany jest nowy indeks wyznaczający liczbę skupień w zbiorze danych opisanych przez zmienne ciągłe. Indeks oparty jest na wielostopniowym dzieleniu zbioru danych (lub jego części) na dwa skupienia i sprawdzaniu czy podział taki należy zachować czy pominąć. Kryterium sprawdzającym jest indeks Randa przy pomocy którego oceniana jest zgodność podziału pierwotnego na dwa skupienia z podziałem na dwa skupienia zbioru węższego, składającego się ze skupienia mniejszego z podziału pierwotnego i 1/3 skupienia większego z podziału pierwotnego. Podziały dokonywane są przy pomocy metody *k*-średnich (dla *k*=2) z wielokrotnym losowym wyborem punktów startowych. Efektywność nowego indeksu została zbadana w obszernym eksperymencie na kilku tysiącach zbiorów danych wygenerowanych w postaci struktur skupień o różnej liczbie zmiennych, skupień, względnej liczebności skupień i różnych wariantach skorelowania zmiennych wewnątrz skupień. Ponadto, zmienny był również stopień separowalności skupień – kontrolowany według algorytmu OCLUS. Podstawą oceny efektywności było porównanie z dwoma innymi indeksami liczby skupień, mającymi w literaturze przedmiotu opinię jednych z najlepszych spośród dotychczas opracowanych tj. indeksem Calińskiego-Harabasza oraz indeksem Gap. Efektywność zaproponowanego indeksu jest znacznie wyższa od obu konkurencyjnych indeksów w przypadkach niezbyt wyraźnej struktury skupień.

**Słowa kluczowe:** analiza skupień, liczba skupień w zbiorze danych, indeks Calińskiego-Harabasza, indeks Gap

## INDEX OF THE CHOICE OF THE NUMBER OF CLUSTERS

### Abstract

In the article a new index for determining the number of clusters in a data set is proposed. The index is based on multiple division of the data set (or a part of it) into two clusters and checking if this division should be retained or neglected. The checking criterion is the Rand index by means of which

the extent to which the primary division and the second division of the narrower subset consisting of the smaller cluster from the primary division and 1/3 of the bigger cluster coincide. The divisions are made by means of the classical *k*-means (for $k=2$) with multiple random choice of starting points. The efficiency of the new index was examined in a broad experiment on a couple of thousands of data sets generated to possess cluster structures with different number of variables, clusters, cluster densities and different variants of within cluster correlation. Moreover, the cluster overlap controlled according to the OCLUS algorithm was also varied. A basis for efficiency assessment was the comparison with two other leading indices i.e. Caliński-Harabasz index and the Gap index. The efficiency of the new index proposed is higher than that of the competition when the cluster structure is not very distinct.

**Keywords:** cluster analysis, number of clusters in a data set, Caliński-Harabasz index, Gap index