PRZEGLĄD STATYSTYCZNY
R. LVIII – ZESZYT 3-4 – 2011

KAMILA MIGDAŁ-NAJMAN

OCENA JAKOŚCI WYNIKÓW GRUPOWANIA – PRZEGLĄD BIBLIOGRAFII

1. WSTĘP

Jedną z ważnych zdolności człowieka jest umiejętność rozróżniania i rozpoznawania obiektów, zdarzeń czy faktów, a także ich łączenia w grupy. Wydaje się, że czynności te dla człowieka są czymś naturalnym, łatwym i prostym. Można powiedzieć, że wyodrębnianie jednostek podobnych w celu ich pogrupowania jest niemal zdolnością przyrodzoną człowieka. J.A. Hartigan powiedziałby, że jest to sposób myślenia o rzeczach a nie badanie rzeczy samych w sobie, poprzez czerpanie wiedzy z doświadczeń i umiejętności uzyskanych z wielu obszarów życia człowieka (Arabie P., Hubert L.J., Soete G.De [1996]). Ten sposób myślenia stał się podstawą rozwoju wielu dziedzin wiedzy, które zaproponowały w tym zakresie własne bogate terminologie i definicje.

Już 1737 roku Carolus Linnaeus w swoim dziele „*Genera Plantarum*" wskazuje, że cała wiedza, jaką rzeczywiście posiadamy o interesujących nas jednostkach zależy od stosowanych metod, które pozwalają na wyróżnienie podobnych do siebie grup jednostek. Im większe zróżnicowanie badanych jednostek obserwujemy, tym łatwiej nam przy pomocy tych metod pojąć złożoną strukturę badanego zjawiska. Im badane zbiory są bardziej liczne, tym ważniejsze i bardziej konieczne staje się posługiwanie się takimi metodami (Everitt B.S., Landau S., Leese M., Stahl D. [2011]). Ważnym impulsem do rozwoju metod grupowania była publikacja z 1963 roku dwóch biologów P.H. Sneatha i R.R. Sokala zatytułowana „*Principles of numerical taxonomy*". Monografia ta jest często uważana za przełomową i wskazującą nowe kierunki badań nad metodami grupowania.

Grupowanie jednostek jest zadaniem złożonym. Jest wiele czynników wpływających na uzyskane rozwiązanie. Do ważniejszych można zaliczyć: liczbę grupowanych jednostek (stosuje się inne metody grupowania zbiorów o kilkudziesięciu jednostkach a inne o setkach tysięcy), liczbę cech zmiennych opisujących daną jednostkę (tzw. problem wymiarowości), zastosowane skale pomiarowe wszystkich cech (skale mogą być identyczne dla wszystkich cech lub nie), strukturę przestrzenną jednostek (skupienia separowalne lub nie, separowalne liniowo lub nie, posiadające skupienie gęstości obiektów lub rozłożone równomiernie, i inne), istnienie braków danych czy istnienie wartości skrajnych (*outliers*). Każdy z tych czynników powoduje konieczność innego podejścia do problemu grupowania. Różnorodność ta jest także przyczyną istnienia dużej liczby algorytmów grupowania, często opartych na bardzo różnych pomysłach. Ponieważ zbiór jednostek można zwykle pogrupować na bardzo wiele sposobów jedne

z nich wydają się „lepsze" a inne „gorsze". Ponieważ pojęcia te są niejednoznaczne, konieczne stało się wypracowanie obiektywnych kryteriów oceny jakości wyróżnionych skupień.

W artykule podjęto próbę usystematyzowania wiedzy w zakresie istniejących metod oceny jakości struktury grupowej. Przedstawiono propozycje klasyfikacji metod oceny jakości grupowania i wskaźników ustalania liczby skupień.

2. KRYTERIA OCENY JAKOŚCI WYNIKÓW GRUPOWANIA

Wszelkiego rodzaju procedury oceny uzyskanych wyników grupowania znane są w literaturze światowej pod nazwą *cluster validity* (prawdziwość, wiarygodność, jakość, grupowania). Pojęcie to definiuje się jako proces oceny wiarygodności algorytmów grupowania przy zastosowaniu różnych warunków początkowych a także jako kwantyfikowalną ocenę uzyskanych rezultatów grupowania. Ponieważ istnieje bardzo wiele samych metod grupowania, w konsekwencji jest także wiele metod oceny uzyskanej struktury grupowej. Metody oceny struktury grupowej można podzielić na dwie zasadnicze grupy: metody wzorcowe i bezwzorcowe. Metody wzorcowe to te, w których uzyskany podział jednostek porównuje się z innym znanym podziałem. Istnieje więc wzorzec idealnego lub zakładanego podziału, z którym porównuje się podział bieżący. Wzorzec może pochodzić od eksperta, może wynikać z rozważań teoretycznych a może pochodzić z innych badań tej samej zbiorowości, tą samą metodą ale z innymi parametrami lub inną metodą. (Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.R. [2007]). Metody bezwzorcowe to te, które do oceny jakości klasyfikacji wykorzystują jedynie informacje pochodzące bezpośrednio z danych. Bierze się tu pod uwagę takie własności uzyskanych skupień jak ich spójność, separowalność, sferyczność czy gęstość.

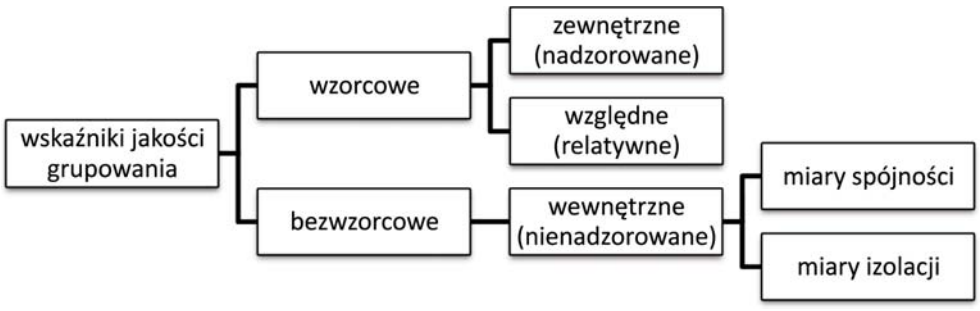
Ta ogólna idea wydaje się być akceptowana przez środowiska naukowe reprezentujące różne dyscypliny. Różnice sprowadzają się głównie do stosowanej nomenklatury lub stopnia szczegółowości opisu. Najbardziej znane propozycje pojęć pochodzą jeszcze z 1967 roku z prac MacQueena. Biorąc pod uwagę źródło pochodzenia informacji o strukturze grupowej, wyróżniono trzy klasy wskaźników (współczynników) jakości grupowania opartych o kryterium: zewnętrzne (*external criteria, external validation*), wewnętrzne (*internal criteria, internal validation*) i względne (*relative criteria, relative validation*) (Theodoridis S., Koutroumbas K. [1999, 2003], Halkidi M., Batistakis Y., Vazirgiannis M. [2001], Maimon O., Rokach L. [2005]).

Ze wskaźnikami biorącymi pod uwagę kryterium zewnętrzne mamy do czynienia, gdy uzyskana struktura grupowa i przynależność jednostek do skupień porównywana jest ze z góry założoną strukturą grupową badanego zbioru jednostek. Zakładana struktura jest odzwierciedleniem naszej wiedzy, wiedzy lub intuicji eksperta o analizowanym zbiorze. Wzorzec grupowania jest, więc zewnętrzny w stosunku do grupowanych jednostek. Ze wskaźnikami biorącymi pod uwagę kryterium wewnętrzne mamy do czynienia, gdy ocena rezultatów uzyskanej struktury grupowej dokonana jest w oparciu o infor-

macje pochodzące z samego analizowanego zbioru jednostek. Informacja o strukturze grupowej pochodzi wyłącznie z danych. Jest więc wewnętrzna w stosunku do nich. Ze wskaźnikami biorącymi pod uwagę kryterium względne mamy do czynienia, gdy ocena rezultatów uzyskanej struktury grupowej porównywana jest w stosunku do innej struktury grupowej uzyskanej w wyniku zastosowania tego samego algorytmu grupowania, ale z założonymi innymi parametrami (np. liczbą skupień). Ocena grupowania jest dokonywana względem innych uzyskanych podziałów populacji (Friedman H.P., Rubin J. [1967], Halkidi M., Batistakis Y., Vazirgiannis M. [2001], Duda R.O., Hart P.E., Stork D.G. [2002], Theodoridis S., Koutroumbas K. [2003], Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.R. [2007], Halkidi M., Gunopulos D., Vazirgiannis M., Kumar N., Domeniconi C. [2008]).

Przy powyższych definicjach wskaźniki zewnętrzne i względne należą do szerszej grupy wskaźników wzorcowych, a wskaźniki wewnętrzne do bezwzorcowych. Opierając się na tej samej idei, powyższą klasyfikację definiuje się czasem stosując nomenklaturę bardziej techniczną. Metody oparte o kryterium zewnętrzne są nazywane metodami nadzorowanymi (*supervised*), oparte o kryterium wewnętrzne – metodami nienadzorowanymi (*unsupervised*), a oparte o kryterium względne – metodami relatywnymi (*relative*). W ramach metod nienadzorowanych wyróżnia się dodatkowo dwie grupy: miary spójności skupień (*measures of cluster cohesion*), które określają jak blisko powiązane są ze sobą jednostki w skupieniu oraz miary rozdzielenia, izolacji (*measures of cluster separation*), które określają jak oddalone jest dane skupienie od innych (por. schemat 1).

Schemat 1 Wskaźniki jakości grupowania



Źródło: opracowanie własne

Informacja o jakości uzyskanych skupień jest w procesie grupowania danych potrzebna na dwóch etapach badań. Na etapie końcowym, gdy dokonano już grupowania jednostek konieczna jest ocena uzyskanego podziału. Informacja taka jest jednak konieczna także w samym procesie wyróżniania skupień. Wiele metod grupowania opiera się na pomyśle, aby w procesie iteracyjnym optymalizować wybrane kryterium podziału jednostek. Poszukuje się takiego podziału, który jest najlepszy z wybranego punktu widzenia. Ponieważ najlepszy podział to taki, który pozwala uzyskać skupienia o najwyższej jakości, konieczne jest zdefiniowanie kryterium oceny jakości skupień,

które będzie w procesie grupowania optymalizowane. Krytycznym etapem optymalizacji jest wybór liczby skupień, na którą należy zbiór jednostek rozdzielić. Nie można uzyskać dobrej jakości skupień dla błędnie ustalonej ich liczby. Jednocześnie liczba skupień stanowi parametr aprioryczny dla wielu często stosowanych w praktyce metod grupowania danych takich jak metoda k-średnich, k-medoid, rozmyta metoda c-średnich czy DBSCAN. Konieczne jest zastosowanie obiektywnego, ilościowego kryterium wyznaczania liczby skupień. Wskaźniki takie istnieją i nazywane są wskaźnikami jakości grupowania (*cluster validity indices CVIs, cluster separation index, validity measure, cluster validity measures, validity indices, cluster validity methods*). Z powyższych powodów problem ustalania liczby skupień należy także do klasy problemów oceny jakości grupowania. W dalszej części artykułu przyjęto podział wskaźników jakości grupowania oparty o kryterium: zewnętrzne, wewnętrzne i względne.

3. KRYTERIUM ZEWNĘTRZNE

Wskaźniki oceny jakości grupowania oparte o kryterium zewnętrzne mają szczególne zastosowanie w sytuacjach gdy: 1) badacz jest zainteresowany informacją jak bardzo podobne do siebie (zgodne) są podziały uzyskane w drodze zastosowania różnych algorytmów grupowania, 2) należy sprawdzić zgodność grupowań dokonanych tą samą metodą ale z różnymi parametrami, 3) konieczne jest porównanie skupień uzyskanych dla tych samych obiektów obserwowanych w różnych momentach czasu, 4) należy ocenić zgodność uzyskanego grupowania ze znaną, wzorcową strukturą grup.

W literaturze tematu proponuje się wiele współczynników zgodności podziałów, służących do oceny podobieństwa wyników różnych klasyfikacji. Do bardziej znanych i częściej stosowanych należą: współczynnik Jaccarda (*Jaccard Coefficient, Jaccard's Coefficient of Community)* (Jaccard P. [1908]), współczynnik Randa (*Rand Index, Rand Statistic)* (Rand W.M. [1971]), współczynnik Fowlkesa i Mallowsa (*Fowlkes and Mallows Index)* (Fowlkes E.B., Mallows C.L. [1983]), skorygowany współczynnik Randa (*Rand Adjusted Statistic, Modified Rand)* (Hubert L.J., Arabie P. [1985]). Propozycje współczynników zgodności grupowania przedstawiali również: Kulczynski S. [1927], Anderberg M.R. [1973], Arabie P., Boorman S.A. [1973], Baker F.B. [1974], Rohlf F.J. [1974, 1982], Hartigan J.A. [1975], Hubert L.J., Levin J.R. [1976], Szmigiel C. [1976], Goodman L.A., Kruskal W.H. [1979], Wallace D.L. [1983], Nowak E. [1985], Gordon A.D. [1987], Mirkin B. [1996], Ben-Hur A., Elisseeff A., Guyon I. [2002] i inni. Interesujące zastosowanie wskaźników Jaccarda, Randa, skorygowanego Randa oraz Fowlkesa i Mallowsa do oceny podobieństwa wyników grupowania w badaniu preferencji i zachowań komunikacyjnych mieszkańców Gdyni zaprezentowała K.Migdał Najman [2011].

Poza tymi najbardziej znanymi wskaźnikami można również spotkać inne propozycje współczynników zgodności podziałów, jak np. miarę F (*F-Measure, F-Score measure*), wskaźnik *purity*, metrykę Mirina (*Mirkin metric*), entropię (*entropy*), wskaźnik wzajemnej informacji (*mutual information*), miarę NMI (*normalized mutual informa-*

tion), współczynnik zmienności informacji (*variation of information*), współczynnik podziału (*partition coefficient*), miarę V (*V-measure*), współczynnik Minkowskiego (*Minkowski score*), błąd klasyfikacji (*classification error, classification metric, CE*), kryterium van Dongena (*van Dongen criterion*), współczynnik lokalnej precyzji (*micro-average precision*) i inne (por. Celeux G., Soromenho G. [1996], Holliday J.D., Hu C-Y., Willett P. [2002], Zhong S., Ghosh J. [2003], Mirkin B. [1996], Handl J., Knowles J. [2004], Zhao Y., Karypis G. [2004], Rubinov A.M., Soukhorokova N.V., Ugon J. [2006], Meilă M. [2007], Aliguliyev R.M. [2009], Seung-Seok C., Sung-Hyuk C., Tappert C.C. [2010], Albatineh A.N., Niewiadomska-Bugaj M. [2011], Labatut V., Cherifi H. [2011], Pitchandi P., Raju N. [2011], Rendón E., Abundez I.M., Gutierrez C., Diaz Zagal S., Arizmendi A., Quiroz E.M., Arzate E.H. [2011]).

Większość z proponowanych współczynników zgodności przyjmuje wartości z unormowanego przedziału, np. [0,1]. Wartość równą 0 osiągają, gdy dwa porównywane podziały są zupełnie niepodobne a wartość równą 1, gdy dwa porównywane podziały są identyczne.

Szerokie zastosowania wskaźników zgodności podziałów w swoich badaniach przedstawiali: Szultz J.V., Hubert L.J. [1979], Pal N.R., Biswas J. [1997], Bezdek J.C., Pal N.R. [1998], Yeung K.Y., Ruzzo W.L. [2001], Mali K., Mitra S. [2003], Bryan J. [2004], Gatnar E., Walesiak M. [2004], Hoffmann A., Motoda H., Scheffer T. (Eds.) [2005], Maimon O., Rokach L. [2005], Ayala G., Epifanio I., Simó A., Zapater V. [2006], Reulke R., Eckardt U., Flach B., Knauer U., Polthier K. (Eds.) [2006], Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.R. [2007], Falasconi M., Pardo M., Vezzoli M., Sberveglieri G. [2007], Migdał Najman K. [2007], Ding C., Li T., Peng W. [2008], Ming-Tso Chiang M., Mirkin B. [2010], Gurrutxaga I., Muguerza J., Arbelaitz O., Pérez J.M., Martín J.I. [2011], Albatineh A.N., Niewiadomska-Bugaj M. [2011] i inni.

4. KRYTERIUM WEWNĘTRZNE

Celem stosowania wskaźników jakości grupowania opartych na kryterium wewnętrznym jest poszukiwanie odpowiedzi na pytanie: na ile uzyskana struktura grupowa otrzymana w wyniku zastosowana danej techniki grupowania stanowi dobre podsumowanie informacji zawartej w danych? Wiele z metod grupowania np. hierarchiczne metody aglomeracyjne (Lance G.N., Williams W.T. [1966a, 1966b], Johnson S.C. [1967], Lance G.N., Williams W.T. [1967a, 1967b], Anderberg M.R. [1973], Gordon A.D. [1987]) oparte są na pomiarze stopnia podobieństwa (*similarity*) lub zróżnicowania (*dissimilarity*) jednostek w przestrzeni cech. To między innymi od właściwego zdefiniowania miary oceniającej stopień podobieństwa (*similarity measures*) lub zróżnicowania (odległości, *dissimilarity measures*) jednostek będzie zależał uzyskany wynik grupowania. (Sneath P.H.A., Sokal R.R. [1973], Anderberg M.R. [1973], Everitt B.S. [1993], Mahalanobis P.C. [1936], Spath H. [1980], Tanimoto T. [1958]). Zastosowanie jednej z metod aglomeracyjnego grupowania hierarchicznego do zdefiniowanej macie-

rzy podobieństwa pozwala na prezentację wyników klasyfikacji w formie graficznej w postaci tzw. dendrogramu lub drzewa połączeń (*dendrogram, tree diagram, hierarchical tree diagram*). Wykres taki prezentuje hierarchię łączenia jednostek w grupy oraz poziomy łączenia (*splitting, fusion levels*), na których jednostki te połączyły się po raz pierwszy (Hartigan J.A. [1967], Gordon A.D. [1987, 1999]). Dendrogram jest efektem zastosowania konkretnej strategii grupowania hierarchicznego i miary podobieństwa lub odległości (Florek K., Łukasiewicz J., Perkal J., Steinhaus H., Zubrzycki S. [1951], Sneath P.H.A. [1957], Sokal R.R., Michener C.D. [1958], McQuitty L.L. [1960, 1966, 1967], Sokal R.R., Sneath P.H.A. [1963], Gower J.C. [1967], Hartigan J.A. [1967], Podani J. [1989], Ward J.H. [1963], Wishart D. [1969]). W połowie lat 60-tych G.N. Lance i W.T. Williams [1966a, 1966b, 1967a, 1967b] opracowali ogólny schemat obliczania odległości między skupieniami biorący pod uwagę wszystkie znane metody aglomeracyjnego grupowania hierarchicznego[1]. W 1978 roku M. Jambu zaproponował poszerzenie tego schematu i umożliwił włączenie do niego jeszcze innych strategii (Jambu M. [1978]). Analizując własności metod aglomeracyjnego grupowania hierarchicznego i możliwość zastosowania w nim różnych miar odległości lub podobieństwa wynika, że możemy uzyskać wiele różnych hierarchii. W literaturze przedmiotu prezentowane są różne metody pozwalające na zmierzenie dopasowania dendrogramu wyznaczonego dla zadanej metody aglomeracyjnego grupowania hierarchicznego do macierzy odległości lub podobieństwa. Ocenę jakości grupowania można oprzeć na porównaniu uzyskanych wyników, prezentowanych np. w formie wyjściowej macierzy odległości do macierzy odległości uzyskanej dla danej strategii grupowania, która prezentuje poziomy łączenia, na których pary jednostek pojawiły się po raz pierwszy w tym samym skupieniu. Najbardziej znanym współczynnikiem pozwalającym na ocenę stopnia dopasowania między macierzą odległości **D** a macierzą dendrogramu $\mathbf{C}_{dendr}$ jest współczynnik korelacji kofenetycznej (*cophenetic correlation coefficient CPCC*), zaproponowany w 1962 przez R.R Sokala i F.J. Rohlfa. (Sokal R.R., Rohlf F.J. [1962]) Jego zastosowania można znaleźć w pracach: Rohlf F.J., Fisher D.R. [1968], Farris J.S. [1969], Kruskal J.B., Carroll J.D. [1969], Rohlf F.J. [1970], Holgersson M. [1978], Trakhtenbrot A., Kadmon R. [2006], Kuramae E.E., Robert V., Echavarri-Erasun C., Boekhout T. [2007], Mérigot B., Durbec J.P., Gaertner J.C. [2010]. Innym współczynnikiem pozwalającym na ocenę stopnia dopasowania dendrogramu do macierzy odległości (podobieństwa) jest współczynnik Goodmana-Kruskala (*Goodman-Kruskal gamma coefficient, gamma index*). Został on zaproponowany w 1954 roku (Goodman L.A., Kruskal W.H. [1954]) do oceny zgodności uporządkowań cech wyrażonych na skali porządkowej. W bogatej literaturze tematu proponuje się także inne współczynniki zgodności dla cech wyrażonych na skali porządkowej. Ich propozycje i szersze dyskusje przedstawiali: Kendall M.G. [1945], Somers R.H. [1962a,b], Costner H.L. [1965], Davis J.A. [1967], Hawkes R.K. [1971], Kim J. [1971], Cunningham K.M.,

[1] Metody objęte schematem przez Lancea i Williamsa: metoda najbliższego sąsiada, metoda najdalszego sąsiada, metoda średniej grupowej, metoda centroidalna, metoda mediany i metoda Warda.

Ogilvie J.C. [1972], Baker F.B. [1974], Hubert L.J. [1974], Rohlf F.J. [1974], Baker F.B., Hubert L.J. [1975], Günter S., Bunke H. [2003], Rousson [2007].

Wielu autorów proponuje także wskaźniki oparte na różnicach odległości (podobieństwa) w dwóch porównywanych macierzach: Kruskal J.B. [1964], Gower J.C. [1966, 1967, 1970], Guttman L. [1968], Hartigan J.A. [1967], Jardine C.J., Jardine N., Sibson R. [1967], Jardine N., Sibson R. [1968], Kruskal J.B., Carroll J.D. [1969], Sammon J.W. [1969], Anderson A.J.B. [1971], Sneath P.H.A., Sokal R.R. [1973], Everitt B. [1978], Balicki A. [2009], Kalinowski S.T. [2009].

Do oceny rezultatów uzyskanej klasyfikacji w oparciu o informacje pochodzące z samego analizowanego zbioru może również posłużyć tzw. wskaźnik sylwetkowy (*silhouette index - SI, silhouette coefficient, SIL index*) zaproponowany przez P.J. Rousseeuw w 1987 roku (Rousseeuw P.J. [1987]). Wartość wskaźnika SI można zinterpretować jako wskaźnik jakości otrzymanej struktury grupowej. Interesujące wykorzystanie wskaźnika SI można znaleźć w artykule D. Waneka [2003] do oceny rezultatów uzyskanej klasyfikacji klientów fińskiego przedsiębiorstwa produkującego domy z prefabrykatów.

Zastosowania indeksu SI w swoich badaniach przedstawiali: Kaufman L., Rousseeuw P.J. [1990], Raymond T.N., Jiawei H. [1994], Tibshirani R., Walther G., Hastie T. [2001], Bolshakova N., Azuaje F. [2003], Mali K., Mitra S. [2003], Sugar C.A., James B.M. [2003], Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.R. [2007], Saitta S., Raphael B., Smith I.F.C. [2007], Hennig C. [2008], Schepers J., Ceulemans E., Van Mechelen I. [2008], Lago-Fernández L.F., Corbacho F. [2010], Ming-Tso Chiang M., Mirkin B. [2010].

5. KRYTERIUM WZGLĘDNE

Biorąc pod uwagę względne kryterium oceny jakości grupowania ocena uzyskanej struktury grupowej dokonywana jest poprzez porównanie otrzymanego podziału z innymi wynikami grupowania. Podejście to wykorzystuje się także w samym procesie grupowania do ustalania liczby skupień. Przyjmuje się a priori pewną liczbę skupień (zwykle dwa), dokonuje grupowania i ocenia jakość uzyskanej struktury grupowej. Następnie przyjmuje się większą o jeden liczbę skupień i powtarza procedurę tak długo, aż uzyska się optymalną wartość miary jakości grupowania. Ten podział, który optymalizuje wartość przyjętego do oceny grupowania wskaźnika będzie uznany za najlepszy.

W 1953 roku R.L. Thorndike'a w czasopiśmie *Psychometrika* przedstawił subiektywną propozycję ustalania liczby klas[2]. Podejście Thorndike'a polegało na porównywaniu średnich wewnątrzklasowych przy ustalonej, różnej liczbie skupień. Au-

[2] Jest on twórcą pomysłu na iteracyjny algorytm optymalizacyjny podziału zbioru jednostek, nazwanego później przez McQueena metodą k-średnich. Niesłusznie uważa się, że to McQueen jest jego twórcą, gdy w rzeczywistości prowadził badania nad ulepszeniem już znanego pomysłu (ulepszanego do dziś) i nadał mu nazwę.

tor sugerował, że wraz ze wzrostem liczby klas następuje spadek wartości średniej wewnątrzklasowej. Proponował prezentowanie tej zależności w formie graficznej. Na wykresie, na osi odciętej prezentowane były kolejne podziały (zaczynając od dwóch klas, *number of clusters*) a na osi rzędnej średnia wewnątrzklasowa dla proponowanych różnych podziałów (*average within-cluster distance for different numbers of cluters*). Odczytywanie sugerowanej liczby klas następowało poprzez poszukiwanie na liniowym wykresie nagłego „spłaszczenia" wykresu (bądź posłużenie się kryterium „łokcia"), które następowało wraz ze wzrostem liczby klas (Thorndike R.L. [1953]). Podobne podejście proponowane było również przez innych autorów, między innymi J.C. Gowera w 1975 roku. E.M.L. Beale w artykule z 1969 roku pt. „*Euclidean cluster analysis*" prezentowanym w *Bulletin of the International Statistical Institute*, sugerował, że formalną metodą pozwalającą na ustalanie optymalnej liczby skupień może być test F (Fishera-Snedecora, równości wariancji). Rozważał, czy test F może zostać wykorzystany do sprawdzenia, że podział na większą liczbą klas jest „lepszy" niż podział na mniejszą liczbę klas. Prowadzone badania z zastosowaniem testu F wykazały skuteczność proponowanego podejścia, ale tylko w przypadku skupień dość dobrze separowalnych i hipersferycznych (Everitt B.S. [1979]). W 1969 roku L. Engelman i J.A. Hartigan testowali podejście oparte na maksymalizacji ilorazu międzygrupowej sumy kwadratów w stosunku do wewnątrzgrupowej sumy kwadratów dla zbioru jednostek, który został podzielony na dwa skupienia (Engelman L., Hartigan J.A. [1969]). W 1971 roku F.H.C. Marriott przeprowadził interesującą dyskusję nad wykorzystaniem jako kryterium ustalania liczby klas wyznacznika z macierzy **W** (macierz rozrzutu, dyspersji wewnątrzklasowej). Konfiguracja skupień, która minimalizowałaby iloczyn liczby klas podniesiony do potęgi drugiej i wartość wyznacznika macierzy **W** ($k^2 \cdot |\mathbf{W}|$) zostałaby uznana za optymalną liczbę klas (Marriott F.H.C. [1971]).

Od czasu badań Thorndike'a zaproponowano przynajmniej kilkadziesiąt wskaźników jakości grupowania stosowanych do ustalania liczby skupień. W 1985 roku ukazał się artykuł G.W. Milligana i M.C. Cooper, (Milligan G.W., Cooper M.C. [1985]) który do dzisiaj jest cytowany, jako podstawowe źródło wiedzy w tej dziedzinie. Ta wielość utrudnia zapoznanie się z nimi i wybór właściwego z punktu widzenia bieżących celów badawczych. Aby zadanie to ułatwić proponuje się ich klasyfikację uwzględniającą źródła i rodzaj informacji brany pod uwagę przy ich konstrukcji. Można wyróżnić tu 5 kryteriów (por. schemat 2):

1. Wskaźniki uwzględniające jedynie odległości między obiektami w skupieniu i między skupieniami, (między obiektami z różnych skupień, między centrami skupień). Należą do nich między innymi: wskaźnik Dunna (*Dunn's index*) (Dunn J.C. [1974]), wskaźnik sylwetkowy (*Global silhouette index, silhouette index-SI, silhouette coefficient, SIL index*) (Rousseeuw P.J. [1987]), wskaźnik C (*C-index*) (Hubert L.J., Levin J.R. [1976]). Interesujące zastosowanie wskaźnika SI zaprezentowano w artykule K.Migdał Najman, K.Najman [2006a] w klasyfikacji 329 miast Stanów Zjednoczonych ze względu na cechy sprzyjające inwestycjom mieszkaniowym.

2. Wskaźniki uwzględniające rozproszenie obiektów wewnątrz skupień i odległości między skupieniami, (między centrami skupień), np. wskaźnik geometryczny (*Geometrical index*) (Lam B.S.Y., Yan H. [2005, 2007]), wskaźnik Daviesa-Bouldina (*Davies-Bouldin index*) (Davies D.L., Bouldin D.W. [1979]). Interesujące zastosowanie wskaźnika Daviesa-Bouldina zaprezentowano w artykule K.Migdał Najman [2010] w klasyfikacji zwyczajów i prawidłowości zakupowych klientów.

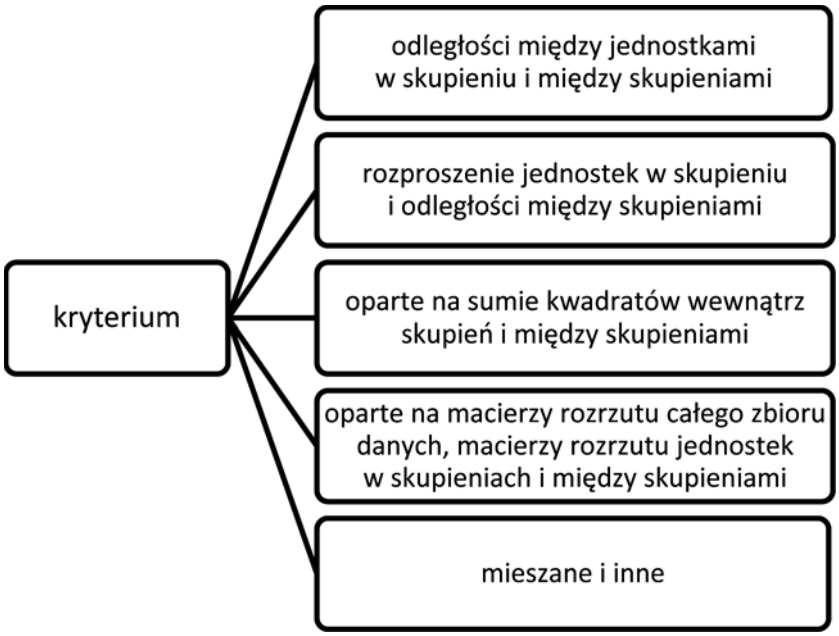
3. Wskaźniki oparte na sumie kwadratów wewnątrz skupień (*within-group sum of squares*) i między skupieniami (*between-group sum of squares*). Należą do nich: wskaźnik Balla-Halla (*Ball-Hall index*) (Ball G.H., Hall D.J. [1965]), wskaźnik Calińskiego-Harabasza (*Calinski-Harabasz index*) (Caliński T., Harabasz J.S. [1974]), wskaźnik Hartigana (*Hartigan index*) (Hartigan J.A. [1975]), wskaźnik Ratkowsky'ego-Lance'a (*Ratkowsky-Lance index*) (Ratkowsky D.A., Lance G.N. [1978]), wskaźnik Krzanowskiego-Lai (*Krzanowski-Lai index*) (Krzanowski W., Lai Y. [1988]), wskaźnik Xu (*Xu index*) (Xu L. [1997]). Interesujące zastosowanie wskaźnika Ratkowsky'ego-Lance'a zaprezentowano w artykule S. Dolnicara i F. Leischa [2004] w segmentacji turystów odwiedzających Austrię. Wyróżnione segmenty nazwano: turystyka aktywna, turystyka zdrowotna i turystyka relaksacyjna. Wskaźnik Krzanowskiego-Lai wykorzystano w artykule J.S. Larsona, E.T. Bradlowa i P.S. Fadera [2005] w klasyfikacji ścieżek zakupowych (*shopper travel path*) klientów pewnego supermarketu w zachodniej części Stanów Zjednoczonych.

4. Wskaźniki oparte na macierzy rozrzutu (*scatter matrix*) całego zbioru danych (**T),** macierzy rozrzutu jednostek w skupieniach (wewnątrzklasowej) (**W**) i między skupieniami (międzyklasowej) (**B**). Należą do nich: wskaźniki Friedmana-Rubina (*Trace*$\mathbf{W}^{(-1)}\mathbf{B}$**,** $|\mathbf{T}| / |\mathbf{W}|$) (Friedman H.P., Rubin J. [1967]), wskaźnik Scotta-Symonsa (*Scott-Symons index*) (Scott A.J., Symons M.J. [1971]), wskaźnik Marriotta (*Marriott index*) (Marriott F.H.C. [1971]), wskaźnik TraceCov**W** (Milligan G.W., Cooper M.C. [1985]).

5. Wskaźniki mieszane i inne (np. oparte na rozmytej funkcji przynależności do skupień). Należą do nich: wskaźnik PC (*Bezdek's partition coefficient PC*) (Bezdek J.C. [1974a]), wskaźnik PE (*Partition entropy PE)* (Bezdek J.C. [1974b]), wskaźnik CE (*Bezdek's classification entropy*) (Bezdek J.C. [1981]), wskaźnik Xie-Beni (*Xie-Beni's separation measure*) (Xie X.L., Beni G. [1991]) i inne.

Problem ustalania optymalnej liczby klas prezentowali w pracach: Thorndike R.L. [1953], Ling R.F. [1972], Marriott F.H.C. [1971], Dunn J.C. [1973, 1974], Sneath P.H.A., Sokal R.R. [1973], Bezdek J.C. [1974a,b], Caliński T., Harabasz J.S. [1974], Baker F.B., Hubert L.J. [1975], Davies D.L., Bouldin D.W. [1979], Jain A.K., Dubes R.C [1979, 1988], Hubert L., Arabie P. [1985], Milligan G.W., Cooper M.C. [1985, 1987], Rousseeuw P.J. [1987], Krzanowski W.J., Lai Y.T. [1988], Gath I., Geva A.B. [1989], Kaufman L., Rousseeuw P.J. [1990], Wilson R., Spann M. [1990], Xie X.L., Beni G. [1991], Bensaid A.M., Hall L.O., Bezdek J.C., Clarke L.P., Silbiger M.L., Arrington J.A., Murtagh R.F. [1996], Celeux G., Soromenho G. [1996], Hardy A. [1996], Bezdek J.C., Pal N.R. [1998], Bandyopadhyay S., Maulik U. [2001], Halkidi

Schemat 2 Podstawowe kryteria podziału wskaźników jakości grupowania



Źródło: opracowanie własne

M., Batistakis Y., Vazirgiannis M. [2001], Tibshirani R., Walther G., Hastie T. [2001], Dimitriadou E., Dolničar S., Weingessel A. [2002], Dudoit S., Fridlyand J. [2002], Bolshakova N., Azuaje F. [2003], Mali K., Mitra S. [2003], Sugar C.A., James G.M. [2003], Chou C.H., Su M.C., Lai E. [2004], Sun H., Wang S., Jiang Q. [2004], Lam B.S.Y., Yan H. [2005, 2007], Migdał Najman K., Najman K. [2005, 2006, 2008], Steinley D. [2006], Walesiak M., Dudek A. [2006], Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.R. [2007], Cios K.J., Pedrycz W., Świniarski R.W., Kurgan L.A. [2007], Migdał Najman [2007], Halkidi M., Vazirgiannis M. [2008], Saitta S., Raphael B., Smith I.F.C. [2008], Wang J.S., Chiang J.C. [2008], Aliguliyev R.M. [2009], Muhlenbach F., Lallich S. [2009], Chicco G., Akilimali J.S. [2010], Lago-Fernández L.F., Corbacho F. [2010], Yue S., Wang J.S., Wu T., Wang H. [2010], Ming-Tso Chiang M., Mirkin B. [2010], Mirkin B. [2011] i inni.

B.S. Everitt, S. Landau, M. Leese w książce z 2001 roku „*Cluster analysis*" twierdzą, że nie ma optymalnego kryterium ustalania liczby klas w badanym zbiorze. Na podstawie badań symulacyjnych podobne wnioski zaprezentowali K. Migdał Najman, K. Najman [2005, 2006b, 2008], K. Migdał Najman [2007]. H.H. Bock w 1985 roku zastosował cztery testy do weryfikacji hipotezy zerowej stwierdzającej homogeniczność, jednorodność populacji wobec hipotezy alternatywnej zakładającej heterogeniczność, niejednorodność populacji. We wnioskach użył określenia, że żaden z czterech badanych rodzajów testów nie okazał się „idealny" (*no one of these tests is the „ideal"*)

w poszukiwaniu struktury grupowej. K. Jajuga potwierdził, że „Jest to problem nie rozwiązany dotychczas w sposób zadowalający" (Jajuga K. [1990]).

6. PODSUMOWANIE

Podobnie jak sam problem grupowania danych tak problem oceny jakości uzyskanej struktury grupowej jest bardzo złożony. Składa się na to duża liczba możliwych do uzyskania podziałów, wielość metod grupowania, z których każda posiada niemały czasem zbiór parametrów, wielość kryteriów i punktów widzenia na cel grupowania. Z tego powodu istnieje duża liczba wskaźników oceny jakości uzyskanej struktury grupowej. Tak jak różne są cele stawiane przed grupowaniem jednostek tak różne muszą być metody oceny ich realizacji. Nie istnieje jeden uniwersalny wskaźnik, który można stosować zawsze, niezależnie od rozwiązywanego problemu i zastosowanej metody. Każdy z nich w swojej konstrukcji uwzględnia jedynie część informacji o strukturze grupowej. Aby wybrać właściwy wskaźnik dla danego zastosowania konieczna jest znajomość szerokiego spektrum istniejących wskaźników i podstawowych ich własności.

W artykule dokonano syntetycznego przeglądu literatury tematu począwszy od prac P. Jaccarda z roku 1908 a skończywszy na pracach B. Mirkina z 2011 roku. Dokonano próby klasyfikacji znanych wskaźników jakości grupowania, uwzględniając kryteria pochodzące z różnych dyscyplin naukowych. W szczególności dokonano klasyfikacji wskaźników optymalnej liczby skupień jako podklasy wskaźników jakości grupowania. Wyniki prezentowanych badań powinny być użyteczne dla wszystkich zajmujących się problemami grupowania i klasyfikacji.

LITERATURA

[1] Albatineh A.N., Niewiadomska-Bugaj M. [2011], *MCS: A method for finding the number of clusters*, Journal of classification, 28, 2, 184-209.
[2] Aliguliyev R.M. [2009], *Performance evaluation of density-based clustering methods*, Information Sciences, 179, 20, 3583-3602.
[3] Anderberg M.R. [1973], *Cluster analysis for applications*, Academic Press, New York, San Francisco, London.
[4] Anderson A.J.B. [1971], *Numeric examination of multivariate soil samples*, Mathematical geology, 3, 1, 1-14.
[5] Arabie P., Boorman S.A. [1973], *Multidimensional scaling of measures of distance between partitions*, Journal of Mathematical Psychology, 10, 2, 148-203.
[6] Arabie P., Hubert L.J., Soete G.De (editors) [1996], *Clustering and classification*, World Scientific Publishing Co. Pte. Ltd., Singapore.
[7] Ayala G., Epifanio I., Simó A., Zapater V. [2006], *Clustering of spatial point patterns*, Computational Statistics & Data Analysis 50.
[8] Baker F.B. [1974], *Stability of two hierarchical grouping techniques, case 1: sensitivity to data errors*, Journal of the American Statistical Association, 69, 346, 440-445.
[9] Baker F.B., Hubert L.J. [1975], *Measurig the power of hierarchical cluster analysis*, Journal of the American Statistical Association, 70, 349, 31-38.

[10] Balicki A. [2009], *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.
[11] Ball G.H., Hall D.J. [1965], ISODATA, *A novel method of data analysis and pattern classification*, Technical report, Stanford Research Institute, Menlo Park, CA, (NTIS No. AD 699616).
[12] Bandyopadhyay S., Maulik U. [2001], *Nonparametric genetic clustering: comparison of validity indices*, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 31, 1, 120-125.
[13] Ben-Hur A., Elisseeff A., Guyon I. [2002], *A stability based method for discovering structure in clustered data*, Pacific Symposium on Biocomputing 7, 6-17.
[14] Bensaid A.M., Hall L.O., Bezdek J.C., Clarke L.P., Silbiger M.L., Arrington J.A., Murtagh R.F. [1996], *Validity-guided (re)clustering with applications to image segmentation*, IEEE Transactions on, Fuzzy Systems, 4, 2, 112-123.
[15] Bezdek J.C. [1974a], *Numerical taxonomy with fuzzy sets*, Journal of Mathematical Biology, 1, 1, 57-71.
[16] Bezdek J.C. [1974b], *Cluster validity with fuzzy sets*, Journal of Cybernetics, 3, 58-72.
[17] Bezdek J.C. [1981], *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, NY.
[18] Bezdek J.C., Pal N.R. [1998], *Some new indexes of cluster validity, Systems, Man, and Cybernetics, Part B: Cybernetics*, IEEE Transactions, 28, 3, 301-315.
[19] Bock H.H. [1985], *On some significance tests in cluster analysis*, Journal of classification, 2, 1, 77-108.
[20] Bolshakova N., Azuaje F. [2003], *Cluster validation techniques for genome expression data*, Signal Processing, 83, 4, 825-833.
[21] Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.R. [2007], *Model-based evaluation of clustering validation measures*, Pattern Recognition, 40, 3, 807-824.
[22] Bryan J. [2004], *Problems in gene clustering based on gene expression data*, Journal of multivariate analysis, 90, 1, 44-66.
[23] Caliński T., Harabasz J.S. [1974], *A dendrite method for cluster analysis*, Communications in Statistics – Theory and Methods, 3, 1, 1-27.
[24] Celeux G., Soromenho G. [1996], *An entropy criterion for assessing the number of clusters in a mixture model*, Journal of classification, 13, 2, 195-212.
[25] Chicco G., Akilimali J.S. [2010], *Renyi entropy-based classification of daily electrical load patterns*, IET Generation, Transmission & Distribution, 4, 6, 736-745.
[26] Chou C.H., Su M.C., Lai E. [2004], *A new cluster validity measure and its application to image compression*, Pattern Analysis & Applications, 7, 2, 205-220.
[27] Cios K.J., Pedrycz W., Świniarski R.W., Kurgan L.A. [2007], *Data mining, a knowledge discovery approach*, Springer Science+Business Media, LLC.
[28] Costner H.L. [1965], *Criteria for measures of association*, American Sociological Review, 30, 3, 341-353.
[29] Cunningham K.M., Ogilvie J.C. [1972], *Evaluation of hierarchical grouping techniques: a preliminary study*, The Computer Journal, 15, 3, 209-213.
[30] Davis J.A. [1967], *A partial coefficient for Goodman and Kruskal's gamma*, Journal of the American Statistical Association, 62, 317, 189-193.
[31] Davies D.L., Bouldin D.W. [1979], *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, 2, 224-227.
[32] Dimitriadou E., Dolničar S., Weingessel A. [2002], *An examination of indexes for determining the number of clusters in binary data sets*, Psychometrika, 67, 3, 137-160.
[33] Ding C., Li T., Peng W. [2008], *On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing*, Computational statistics and data analysis, 52, 8, 3913-3927.

[34] Dolnicar S., Leisch F. [2004], *Segmenting markets by bagged clustering*, Australasian marketing journal, 12, 1, 51-65.
[35] Dubes R.C., Jain A.K. [1979], *Validity studies in clustering methodologies*, Pattern Recognition, 11, 4, 235-254.
[36] Duda R.O., Hart P.E., Stork D.G. [2002], *Pattern classification*, Wiley, New York.
[37] Dudoit S., Fridlyand J. [2002], *A prediction-based resampling method for estimating the number of clusters in a dataset*, Genome Biology, 3, 7, research 0036-research 0036.21.
[38] Dunn J.C. [1973], *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters*, Journal of Cybernetics, 3, 3, 32-57.
[39] Dunn J.C. [1974], *Well separated clusters and optimal fuzzy partitions*, Journal of Cybernetics, 4, 1, 95-104.
[40] Engelman L., Hartigan J.A. [1969], *Percentage points of a test for clusters*, Journal of the American Statistical Association, 64, 328, 1647-1648.
[41] Everitt B. [1978], *Graphical techniques for multivariate data*, Heinemann Educational Books Ltd. London.
[42] Everitt B.S. [1979], *Unresolved problems in cluster analysis*, Biometrics, 35, 1, Perspectives in Biometry, 169-181.
[43] Everitt B.S. [1993], *Cluster Analysis*, Edward Arnold, London.
[44] Everitt B.S., Landau S., Leese M., Stahl D. [2011], *Cluster analysis*, 5th edition, John Wiley & Sons, Ltd., Chichester.
[45] Falasconi M., Pardo M., Vezzoli M., Sberveglieri G. [2007], *Cluster validation for electronic nose data*, Sensors and Actuators B: Chemical, 125, 2, 596-606.
[46] Farris J.S. [1969], *On the cophenetic correlation coefficient*, Systematic Biology, 18, 3, 279-285.
[47] Florek K., Łukasiewicz J., Perkal J., Steinhaus H., Zubrzycki S. [1951], *Taksonomia wrocławska*, Przegląd Antropologiczny, 17, 193-210.
[48] Fowlkes E.B., Mallows C.L. [1983], *A Method for Comparing two hierarchical clusterings*, Journal of the American Statistical Association, 78, 383, 553-569.
[49] Friedman H.P., Rubin J. [1967], *On some invariant criteria for grouping data*, Journal of the American Statistical Association, 62, 320, 1159-1178.
[50] Gath I., Geva A.B. [1989], *Unsupervised optimal fuzzy clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11, 7, 773-781.
[51] Gatnar E., Walesiak M. [2004], *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE Wrocław, 333-336.
[52] Goodman L.A., Kruskal W.H. [1954], *Measures of association for cross classifications*, Journal of the American Statistical Association, 49, 268, 732-764.
[53] Goodman L.A., Kruskal W.H. [1979], *Measures of association for cross classifications*, Sprinter-Verlag, New York, Heidelberg.
[54] Gordon A.D. [1987], *A review of hierarchical classification*, Journal of the Royal Statistical Society, Series A (General), 150, 2, 119-137.
[55] Gordon A.D. [1999], *Classification*, Chapman&Hall/CRC, Boca Raton.
[56] Gower J.C. [1966], *Some distance properties of latent root and vector methods used in multivariate analysis*, Biometrika, 53, 3/4, 325-338.
[57] Gower J.C. [1967], *A comparison of some methods of cluster analysis*, Biometrics, 23, 4, 623-638.
[58] Gower J.C. [1970], *Classification and geology*, Review of the International Statistical Institute, 38, 1, 35-41.
[59] Gurrutxaga I., Muguerza J., Arbelaitz O., Pérez J.M., Martín J.I. [2011], *Towards a standard methodology to evaluate internal cluster validity indices*, Pattern Recognition Letters, 32, 3, 505-515.
[60] Guttman L. [1968], *A general nonmetric technique for finding the smallest coordinate space for a configuration of points*, Psychometrika, 33, 2, 469-506.

[61] Günter S., Bunke H. [2003], *Validation indices for graph clustering*, Pattern Recognition Letters, 24, 8, 1107-1113.
[62] Halkidi M., Batistakis Y., Vazirgiannis M. [2001], *On clustering validation techniques*, Journal of Intelligent Information Systems, 17, 2-3, 107-145.
[63] Halkidi M., Gunopulos D., Vazirgiannis M., Kumar N., Domeniconi C. [2008], *A clustering framework based on subjective and objective validity criteria*, ACM Transactions on Knowledge Discovery from Data, 1, 4, 18, 18:1-18:25.
[64] Halkidi M., Vazirgiannis M. [2008], *NPCIu: an approach for clustering spatially extended objects*, Intelligent Data Analysis, 12, 6, 587-606.
[65] Handl J., Knowles J. [2004], *Multiobjective clustering with automatic determination of the number of clusters*, Technical Report TR- COMPSYSBIO-2004-02. UMIST, Manchester, UK.
[66] Hardy A. [1996], *On the number of clusters*, Computational Statistics & Data Analysis, 23, 1, 15, 83-96.
[67] Hartigan J.A. [1967], *Representation of similarity matrices by tree*, Journal of the American Statistical Association, 62, 320, 1140-1158.
[68] Hartigan J.A. [1975], *Clustering Algorithms*, New York: John Wiley.
[69] Hawkes R.K. [1971], *The multivariate analysis of ordinal measures*, The American Journal of Sociology, 76, 5, 908-926.
[70] Hennig C. [2008], *Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods*, Journal of multivariate analysis, 99, 6, 1154-1176.
[71] Hoffmann A., Motoda H., Scheffer T. (Eds.) [2005], *Discovery Science*, 8th International Conference, DS 2005, Singapore, October, Proceedings, Springer-Verlag Berlin Heidelberg, 302.
[72] Holgersson M. [1978], *The limited value of cophenetic correlation as a clustering criterion*, Pattern Recognition, 10, 4, 287-295.
[73] Holliday J.D., Hu C-Y., Willett P. [2002], *Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit-strings*, Combinatorial Chemistry & High Throughput Screening, 5, 2, 155-166.
[74] Hubert L.J. [1974], *Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures*, Journal of the American Statistical Association, 69, 347, 698-704.
[75] Hubert L.J., Arabie P. [1985], *Comparing partitions*, Journal of Classification, 2, 1, 193-218.
[76] Hubert L.J., Levin J.R. [1976], *Evaluating object set partitions: Free-sort analysis and some generalizations*, Journal of Verbal Learning and Verbal Behavior, 15, 4, 459-470.
[77] Jaccard, P. [1908] *Nouvelles recherchés, sur la distribution florale.* Bulletin de la Société vaudoise des Sciences Naturelles, 44, 223-270.
[78] Jain A.K., Dubes R.C. [1988], *Algorithms for clustering data*, Englewood Cliffs NJ: Prentice Hall, Chapter 4.
[79] Jajuga K. [1990], *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa.
[80] Jambu M. [1978], *Classification automatiqe pour l'analyse des donnees*, tom I, Paris Dunod.
[81] Jardine C.J., Jardine N., Sibson R. [1967], *The structure and construction of taxonomic hierarchies*, Mathematical Biosciences, 1, 2, 173-179.
[82] Jardine N., Sibson R. [1968], *The construction of hierarchic and non-hierarchic classification*, The computer journal, 11, 2, 177-184.
[83] Johnson S.C. [1967], *Hierarchical clustering schemes*, Psychometrika, 32, 3, 241-254.
[84] Kalinowski S.T. [2009], *How well to evolutionary trees describe genetic relationships among populations?*, Heredity, 102, 5, 506-513.
[85] Kaufman L., Rousseeuw P.J. [1990], *Finding groups in data: a introduction to cluster analysis*, Wiley, New York.
[86] Kendall M.G. [1945], *The treatment of ties in ranking problems*, Biometrika, 33, 3, 239-251.
[87] Kim J. [1971], *Predictive measures of ordinal association*, The American Journal of Sociology, 76, 5, 891-907.

[88] Kruskal J.B. [1964], *Nonmetric multidimensional scaling: a numerical method*, Psychometrika **29**, 2, 115-129.
[89] Kruskal J.B., Carroll J.D. [1969], *Geometrical models and badness-of-fit functions*. In Multivariate analysis (P.R.Krishnaiah, ed.), vol. II, (Proceedings of the 2. International Symposium on Multivariate Analysis held at Wright State Universaity, Dayton, Ohio, June 17-22, 1968), New York, Academic Press, 639-671.
[90] Krzanowski W.J., Lai Y.T. [1988], *A criterion for determining the number of groups in a data set using sum-of-squares clustering*, Biometrics, 44, 1, 23-34.
[91] Kulczynski S. [1927], *Die Pflanzenassociationen der Pienenen*, Bulletin International de L'Académie Polonaise des Sciences et des lettres, Classe des sciences mathématiques et naturelles, Série B, Supplément II, 2, 57-203.
[92] Kuramae E.E., Robert V., Echavarri-Erasun C., Boekhout T. [2007], *Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom*, BMC Evolutionary Biology, 7, 134.
[93] Labatut V., Cherifi H. [2011], *Accuracy Measures for the comparision of classifiers*, ICIT, The 5^{th} International Conference on Information Technology (artykuł wygłoszony na konferencji ICIT).
[94] Lago-Fernández L.F., Corbacho F. [2010], *Normality-based validation for crisp clustering*, Pattern Recognition, 43, 3, 782-795.
[95] Lam B.S.Y., Yan H. [2005], *A new cluster validity index for data with merged clusters and different densities*, IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC), 1, 798-803.
[96] Lam B.S.Y., Yan H. [2007], *Assessment of microarray data clustering results based on a new geometrical index for cluster validity*, Soft Computing, 11, 4, 341-348.
[97] Lance G.N., Williams W.T. [1966a], *Computer programs for hierarchical polythetic classification ("Similarity analysis")*, The Computer Journal, 9, 1, 60-64.
[98] Lance G.N., Williams W.T. [1966b], *A generalized sorting strategy for computer classifications*, Nature 212, 218 (8 October), Letters to Nature.
[99] Lance G.N., Williams W.T. [1967a], *A general theory of classificatory sorting strategies*, I. Hierarchical systems, The Computer Journal, 9, 4, 373-380.
[100] Lance G.N., Williams W.T. [1967b], *A general theory of classificatory sorting strategies: II. Clustering systems*, The Computer Journal, 10, 3, 271-277.
[101] Larson J.S., Bradlow E.T., Fader P.S. [2005], *An exploratory look at supermarket shopping paths*, International Journal of Research in Marketing, 22, 4, 395-414.
[102] Ling R.F. [1972], *On the theory and consucion of k-clusters*, The computer journal, 15, 4, 326-332.
[103] Mahalanobis P.C. [1936], *On the generalised distance in statistics*, Proceedings of the National Institute of Sciences of India, 2, 1, 49-55.
[104] Maimon O., Rokach L. (editors) [2005], *The data mining and knowledge discovery handbook*, Halkidi M., Vazirgiannis M.: Quality assessment approaches in data mining, Springer.
[105] Mali K., Mitra S. [2003], *Clustering and its validation in a symbolic framework*, Pattern Recognition Letters, 24, 14, 2367-2376.
[106] Marriott F.H.C. [1971], *Practical problems in a method of cluster analysis*, Biometrics, 27, 3, 501-514.
[107] McQuitty L.L. [1960], *Hierarchical linkage analysis for the isolation of types*, Educational and Psychological Measurement, 20, 1, 55-67.
[108] McQuitty L.L. [1966], *Similarity analysis by reciprocal pairs for discrete and continuous data*, Educational and Psychological Measurement, 26, 4, 825-831.
[109] McQuitty L.L. [1967], *Expansion of similarity analysis by reciprocal pairs for discrete and continuous data*, Educational and Psychological Measurement, 27, 2, 253-255.
[110] Meilă M. [2007], *Comparing clusterings – an information based distance*, Journal of multivariate analysis, 98, 5, 873-895.

[111] Mérigot B., Durbec J.P., Gaertner J.C. [2010], Ecology, 91, 6, 1850-1859.
[112] Migdał Najman K. [2007], *Analiza podobieństwa wyników grupowania uzyskanych w oparciu o metodę k-średnich dla wybranych metod ustalania optymalnej liczby skupień*, Prace i Materiały WZ UG, 5, 601-610.
[113] Migdał Najman K. [2010], *Zastosowanie samouczącej się sieci neuronowej typu SOM w analizie koszykowej*, Taksonomia 17, Prace Naukowe UE we Wrocławiu, 305-315.
[114] Migdał Najman K. [2011], *Propozycja hybrydowej metody grupowania opartej na sieciach samouczących*, Referat wygłoszony na konferencji: SKAD 2011, Wągrowiec.
[115] Migdał Najman K., Najman K. [2005], *Analityczne metody ustalania liczby skupień*, Taksonomia 12, Prace Naukowe AE we Wrocławiu, 1076, 265-273.
[116] Migdał Najman K., Najman K. [2006a], *Wykorzystanie indeksu silhouette do ustalania optymalnej liczby skupień*, Wiadomości Statystyczne, 6, 1-10.
[117] Migdał Najman K., Najman K. [2006b], *Analityczne metody ustalania skupień w rozmytych zbiorach danych*, Taksonomia 13, Prace Naukowe AE we Wrocławiu, 1126, 159-167.
[118] Migdał Najman K., Najman K. [2007], *Charakterystyka mierników oceny podobieństwa wyników podziałów*, Prace i Materiały WZ UG, 3, 191-201.
[119] Migdał Najman K., Najman K. [2008], *Wykorzystanie wskaźnika Dunn'a do ustalania optymalnej liczby skupień*, Wiadomości Statystyczne, 11, 26-34.
[120] Milligan G.W., Cooper M.C. [1985], *An examination of procedures for determining the number of clusters in a data set*, Psychometrika, 50, 2, 159-179.
[121] Milligan G.W., Cooper M.C. [1987], *Methodology review: clustering methods*, Applied psychological measurement, 11, 4, 329-354.
[122] Ming-Tso Chiang M., Mirkin B. [2010], *Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads*, Journal of classification, 27, 1, 3-40.
[123] Mirkin B. [1996], *Mathematical classification and clustering*, Nonconvex optimization and its applications, Kluwer Academic Publishers, Dordrecht.
[124] Mirkin B. [2011], *Choosing the number of clusters*, WIREs Data Mining and Knowledge Discovery, vol. 1, May/June, John Wiley&Sons, Inc., 252-259.
[125] Muhlenbach F., Lallich S. [2009], *A new clustering algorithm based on regions of influence with self-detection of the best number of clusters*, Data Mining, 2009, ICDM'09, Ninth IEEE International Conference on Data Mining, Miami, Florida, 884-889.
[126] Nowak E. [1985], *Wskaźnik podobieństwa wyników podziału*, Przegląd Statystyczny, 1, 41-48.
[127] Pal N.R., Biswas J. [1997], *Cluster validation using graph theoretic concepts*, Pattern Recognition, 30, 6, 848-849.
[128] Pitchandi P., Raju N. [2011], *Improving the performance of multivariate Bernoulli model based documents clustering algorithms using transformation techniques*, Journal of Computer Science, 7, 5, 762-769.
[129] Podani J. [1989], *New combinatorial clustering methods*, Vegetatio (Plant ecology), 81, 61-77.
[130] Rand W.M. [1971], *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association, 66, 336, 846-850.
[131] Ratkowsky D.A., Lance G.N. [1978], *A criterion for determining the number of groups in a classification*, Australian Computer Journal, 10, 3, 115-117.
[132] Raymond T.N., Jiawei H. [1994], *Efficient and effective clustering methods for spatial data mining*, Tech. Rep. TR-93-13, University of British Columbia, Vancouver, B.C., Canada.
[133] Rendón E., Abundez I.M., Gutierrez C., Diaz Zagal S., Arizmendi A., Quiroz E.M., Arzate E.H. [2011], *A comparison of internal and external cluster validation indexes*, in Applications of mathematics & computer engineering, Zemliak A., Mastorakis N. (editors), American Conference on Applied Mathematics (AMERICAN-MATH'11), 5^{TH} WSEAS (World Scientific and Engine-

ering Academy and Society) International Conference on Computer Engineering and Applications (CEA'11), Puerto Morelos, Mexico, Published by WSEAS Press, 158-163.
[134] Reulke R., Eckardt U., Flach B., Knauer U., Polthier K. (Eds.) [2006], *Combinatorial Image Analysis*, 11^{th} International Workshop, IWCIA, Berlin, Germany, June, Springer-Verlag, 108-110.
[135] Rohlf F.J. [1970], *Adaptive hierarchical clustering schemes*, Systematic Biology, 19, 1, 58-82.
[136] Rohlf F.J. [1974], *Methods of comparing classifications*, Annual Review of Ecology and Systematics, 5, 1, 101-113.
[137] Rohlf F.J. [1982], *Consensus Indices for Comparing Classifications*, Mathematical Biosciences, 59.
[138] Rohlf F.J., Fisher D.R. [1968], *Tests for hierarchical structure in random data sets*, Systematic Biology, 17, 4, 407-412.
[139] Rousseeuw P.J. [1987], *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, Journal of of computational and applied mathematics, 20, 1, 53-65.
[140] Rousson V. [2007], *The gamma coefficient revisited*, Statistics & Probability Letters, 77, 17, 1696-1704.
[141] Rubinov A.M., Soukhorokova N.V., Ugon J. [2006], *Classes and clusters in data analysis*, European Journal of Operational Research, 173, 3, 849-865.
[142] Saitta S., Raphael B., Smith I.F.C. [2007], *A bounded index for cluster validity*, Machine Learning and Data Mining in Pattern Recognition, Heidelberg, Germany, Springer, 174-187.
[143] Saitta S., Raphael B., Smith I.F.C. [2008], *A comprehensive validity index for clustering*, Intelligent Data Analysis, 12, 6, 529-548.
[144] Sammon J.W. [1969], *A nonlinear mapping for data structure analysis*, IEEE Transactions on computers, C-18, 5, 401-409.
[145] Schepers J., Ceulemans E., Van Mechelen I. [2008], *Selecting among multi-mode partitioning models of different complexities: a comparison of four model selection criteria*, Journal of Classification, 25, 1, 67-85.
[146] Schultz J.V., Hubert L.J. [1979], *A nonparametric test for the correspondence between two proximity matrices*, Journal of Educational Statistics, 1, 1, 59-67.
[147] Scott A.J., Symons M.J. [1971], *Clustering methods based on likelihood ratio criteria*, Biometrics, 27, 2, 387-397.
[148] Seung-Seok C., Sung-Hyuk C., Tappert C.C. [2010], *A survey of binary similarity and distance measures*, Journal of Systemics, Cybernetics & Informatics, 8, 1, 43-48.
[149] Sneath P.H.A. [1957], *The application of computers to taxonomy*, Journal of General Microbiology, 17, 201-226.
[150] Sneath P.H.A., Sokal R.R. [1973], *Numerical Taxonomy*. The principles and practice of numerical classification, W.H. Freeman & Company, San Francisco, CA.
[151] Sokal R.R., Michener C.D. [1958], *A statistical method for evaluating systematic relationships*, The University of Kansas Scientific Bulletin 38, 1409-1438.
[152] Sokal R.R., Rohlf F.J. [1962], *The comparison of dendrograms by objective methods*, Taxon, 11, 2, 33-40.
[153] Sokal R.R., Sneath P.H.A. [1963], *Principles of numerical taxonomy*, W.H. Freeman & Company, San Francisco, CA.
[154] Somers R.H. [1962a], *A new asymmetric measure of association for ordinal variables*, American Sociological Review, 27, 6, 799-811.
[155] Somers R.H. [1962b], *A similarity between Goodman and Kruskal's tau and Kendall's tau, with a partial interpretation of the latter*, Journal of the American Statistical Association, 57, 300, 804-812.
[156] Spath H. [1980], *Cluster Analysis Algorithms*, Ellis Horwood.
[157] Steinley D. [2006], *K-means clustering: a half-century synthesis*, British Journal of Mathematical and Statistical Psychology, 59, 1, 1-34.

[158] Sugar C.A., James G.M. [2003], *Finding the number of clusters in a dataset: an information theoretic approach*, Journal of the American Statistical Association, 98, 463, 750-763.
[159] Sun H., Wang S., Jiang Q. [2004], *FCM-based model selection algorithms for determining the number of clusters*, Pattern Recognition, 37, 10, 2027-2037.
[160] Szmigiel C. [1976], *Wskaźnik zgodności kryteriów podziału*, Przegląd Statystyczny, 4.
[161] Tanimoto T. [1958], *An elementary mathematical theory of classification and prediction*, Internal Report, IBM Corp.
[162] Theodoridis S., Koutroumbas K. [1999], *Pattern recognition*, Elsevier, Academic Press.
[163] Theodoridis S., Koutroumbas K. [2003], *Pattern recognition*, second edition, Elsevier, Academic Press.
[164] Thorndike R.L. [1953], *Who belongs in the family?*, Psychometrika, 18, 4, 267-276.
[165] Tibshirani R., Walther G., Hastie T. [2001], *Estimating the number of clusters in a data set via the gap statistic*, Journal of the Royal Statistical Society; Series B (Statistical Methodology), 63, 2, 411-423.
[166] Trakhtenbrot A., Kadmon R. [2006], *Effectiveness of environmental cluster analysis in representing regional species diversity*, Conservation Biology, 20, 4, 1087-1098.
[167] Walesiak M., Dudek A. [2006], *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – charakterystyka problemu*, Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 450, Prace Katedry Ekonometrii i Statystyki nr 17, 635-646.
[168] Wallace D.L. [1983], *A method for comparing two hierarchical clustering: comment*, Journal of the American Statistical Association, 78, 383, 569-576.
[169] Wanek D. [2003], *Fuzzy spatial analysis techniques in a business GIS environment*, ERSA 2003 Congress, University of Jyväskylä, Finland.
[170] Ward J.H. [1963], *Hierarchical grouping to optimize an objective function*, Journal of the American Statistical Association, 58, 301, 236-244.
[171] Wilson R., Spann M. [1990], *A new approach to clustering*, Pattern Recognition, 23, 12, 1413-1425.
[172] Wishart D. [1969], *An algorithm for hierarchical classifications*, Biometrics, 25, 1, 165-170.
[173] Xie X.L., Beni G. [1991], *A validity measure for fuzzy clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13, 8, 841-847.
[174] Xu L. [1997], *Bayesian Ying-Yang machine, clustering and number of clusters*, Pattern Recognition Letters, 18, 11-13, 1167-1178.
[175] Yeung K.Y., Ruzzo W.L. [2001], *Details of the adjusted Rand index and clustering algorithms*, suplement do artykułu: An empirical study on principal component analysis for clustering gene expression data, Bioinformatics, 17, 9, 763-774.
[176] Yue S., Wang J.S., Wu T., Wang H. [2010], *A new separation measure for improving the effectiveness of validity indices*, Information Sciences, 180, 5, 748-764.
[177] Zhao Y., Karypis G. [2004], *Empirical and theoretical comparisons of selected criterion functions for document clustering*, Machine Learning, 55, 3, 311-331.
[178] Zhong S., Ghosh J. [2003], *A comparative study of generative models for document clustering*, The University of Texas at Austin, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.4583&rep=rep1&type=pdf.

OCENA JAKOŚCI WYNIKÓW GRUPOWANIA – PRZEGLĄD BIBLIOGRAFII

S t r e s z c z e n i e

W artykule dokonano syntetycznego przeglądu literatury tematu począwszy od prac P. Jaccarda z roku 1908 a skończywszy na pracach B. Mirkina z 2011 roku. Dokonano próby klasyfikacji znanych wskaźników jakości grupowania, uwzględniając kryteria pochodzące z różnych dyscyplin naukowych. W szczególności dokonano klasyfikacji wskaźników optymalnej liczby skupień jako podklasy wskaźników jakości grupowania. Wyniki prezentowanych badań powinny być użyteczne dla wszystkich zajmujących się problemami grupowania i klasyfikacji.

Słowa kluczowe: analiza skupień, wskaźniki jakości grupowania

CLUSTER VALIDITY MEASUREMENT – A BIBLIOGRAPHY REVIEW

S u m m a r y

In the article are presented the synthetic review of the literature from P. Jaccard in 1908 to B. Mirkin, 2011. In this paper, the concept and classification of cluster validity indices are proposed. There are presented classification of validity indices to find the optimal number of clusters. The results of this study should be useful for all concerned with the problems of classification.

Keywords: cluster analysis, cluster validity indices