PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 384

Taksonomia 24
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronie internetowej Wydawnictwa
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 384 • 2015
Taksonomia 24 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Beata Basiura, Anna Czapkiewicz
AGH Akademia Górniczo-Hutnicza
e-mail: bbasiura@zarz.agh.edu.pl

SYMULACYJNE BADANIE WYKORZYSTANIA ENTROPII DO BADANIA JAKOŚCI KLASYFIKACJI

Streszczenie: Celem prezentowanej pracy było zweryfikowanie, czy proponowany wskaźnik jakości klasyfikacji wykorzystujący entropię Renyi'ego może być stosowany do określenia liczby podgrup. Badanie przeprowadzono na danych sztucznie wygenerowanych. Zastosowano algorytm aglomeracji Warda, klasyfikacji *k*-średnich oraz klasyfikacji spektralnej. Na podstawie wybranych czterech indeksów zweryfikowano poprawność wykrywania struktury grupowej. Badano tylko liczbę grup, a nie przynależność poszczególnych elementów do grupy. Otrzymane wyniki potwierdzają skuteczność proponowanego wskaźnika w problemie dotyczącym weryfikacji liczby grup.

Słowa kluczowe: wskaźnik klasyfikacji, entropia Renyi'ego, klasyfikacja.

DOI: 10.15611/pn.2015.384.02

1. Wstęp

W literaturze dotyczącej metod klasyfikacji można znaleźć wiele różnych wskaźników miary jakości klasyfikacji. Wyróżnia się podział na trzy klasy wskaźników oceny jakości grupowania (por. [Halikidi i in. 2001; Baarsch, Celebi 2012; Rendón i in. 2011; Walesiak, Dudek 2012]): wskaźniki oparte na kryteriach zewnętrznych (*external criteria*, *external validation*), wskaźniki oparte na kryteriach wewnętrznych (*internal criteria*, *internal validation*) oraz wskaźniki oparte na kryteriach względnych (*relative criteria*, *relative validation*).

W konstrukcji wskaźników wykorzystuje się miarę odległości (podobieństwa) pomiędzy obiektami. Przykładem takich indeksów są między innymi indeks Calińskiego i Harabasza, indeks Davies-Bouldina czy Silhouette indeks. Wykorzystanie miary zwartości i separowalności zazwyczaj wpływa na lepszą ocenę grup danych o rozkładach eliptycznych.

W prezentowanej pracy rozważono wskaźnik miary klasyfikacji danych, do konstrukcji którego wykorzystano własności entropii Renyie'go. Pewne własności

tego wskaźnika były prezentowane w pracy Basiury i Czapkiewicz [2014]. W pracy tej między innymi zaprezentowano wstępne badanie symulacyjne, które pokazało, że dla wielowymiarowych rozkładów normalnych wskaźnik ten poprawnie wykrywa strukturę grupową danych.

W niniejszej pracy dokonano dalszej analizy własności wskaźnika opartego na entropii Reny'ego. Celem pracy było porównanie tego wskaźnika z wybranymi wskaźnikami opisanymi w literaturze przy założonej strukturze danych [Milligan Glenn 1981; Halkidi i in. 2010; Rendón i in. 2011; Walesiak, Gatnar 2009; Walesiak 2013]. Podstawą przeprowadzonych badań były symulacyjnie wygenerowane przykłady. Podziału na grupy dokonano przy zastosowaniu algorytmu aglomeracji Warda, podstawowej klasyfikacji *k*-średnich oraz klasyfikacji spektralnej [Walesiak, Dudek 2012]. Otrzymane wyniki klasyfikacji zweryfikowano, stosując wybrane cztery wskaźniki określające poprawność klasyfikacji.

2. Wybrane wskaźniki klasyfikacji

Do analizy porównawczej wybrano wskaźniki wewnętrzne. Do nich należy między innymi: indeks Calińskiego i Harabasza [1974], indeks Davies-Bouldina [Davies, Bouldin 1979] oraz Silhouette indeks [Rousseeuw 1987].

Indeks Calińskiego i Harabasza (CH) wykorzystuje iloraz zmienności międzygrupowej (Between Groups – BG) oraz zmienności wewnątrzgrupowej (Within Groups – WG). Zmienność międzygrupowa jest ważoną sumą kwadratów odległości pomiędzy środkiem każdej klasy a środkiem całego zbioru. Wagami są wielkości analizowanych klas. Natomiast zmienność wewnątrzgrupowa wyznaczana jest jako suma kwadratów odległości każdego elementu podzbioru od środka klasy.

W indeksie Davies-Bouldina (DB) dla każdego skupienia wyznacza się średnią odległość pomiędzy każdym punktem grupy a jej centrum (oznaczmy je jako δ_k i $\delta_{k'}$) oraz odległość pomiędzy środkami skupienia k i skupienia k' (oznaczmy jako $\Delta_{kk'}$). Następnie dla każdego podzbioru wyznacza się maksymalną wartość zu: $\frac{\delta_k+\delta_{k\prime}}{\Delta_{kk\prime}}$. Indeks DB to średnia wartość z ilorazów po wszystkich podzbiorach.

Silhouette indeks zaproponowany został przez Rousseeuwa (1987). Indeks ten pozwala oceniać prawidłowość zaklasyfikowania poszczególnych obiektów do wyodrębnionych klas na podstawie następującej reguły. Niech:

$$S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}. \tag{1}$$

Rozważmy dwie klasy: C_k oraz $C_{k'}$. Wielkość $a(i)$ określa średnią odległość obiektu i od pozostałych obiektów należących do klasy C_k, natomiast wielkość $b(i)$ określa minimalną odległość i od obiektów należących do klasy $C_{k'}$. Indeks $S(P_r)$ określający prawidłowość wyodrębnienia klasy jest średnią z wartości $S(i)$ dla po-

szczególnych elementów klasy, natomiast średnia z $S(P_r)$ po wszystkich klasach jest indeksem sylwetkowym.

O lepszej jakości klasyfikacji mówią wyższe wartości indeksu Calińskiego i Harabasza, wyższe wartości indeksu Silhouette oraz niższe indeksy Davies-Bouldina.

3. Wskaźnik klasyfikacji na podstawie miary entropii

Pojęcie entropii wprowadził Shannon w 1948 r., następnie w drugiej połowie ubiegłego wieku pojawiło się wiele uogólnień probabilistycznej miary tej entropii. Węgierski matematyk Alfred Rényi [Rényi 1961] zaproponował następujące uogólnienie pojęcia entropii:

$$H(x) = \frac{1}{1-\alpha}\log\left(\int f^{\alpha}(x)dx\right),\ \alpha > 0, \alpha \neq 1. \tag{2}$$

W szczególności dla $\alpha = 2$ otrzymuje się:

$$H(x) = -\log\left(\int f^{2}(x)\right)dx. \tag{3}$$

Niech $\{x_1, \ldots, x_N\}$, gdzie x_i jest d-wymiarowym obiektem, będzie zbiorem danych niezależnych o tym samym rozkładzie $f(x)$. Jeśli nie znamy rozkładu danej funkcji, to do jej estymacji można zastosować metodę nieparametryczną na podstawie estymacje jądrową [Liang i in. 2011; Jensen i in. 2003]. Niech:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} W_{\sigma^2}(x, x_i).$$

Przy czym $W_{\sigma}^{2}(x, x_i)$ jest oknem Parzena, natomiast σ^2 określa szerokość okna w zależności od rozmiaru danych. W naszych badaniach została wykorzystana funkcja jądrowa Gaussa, określona wzorem (4), w którym parametr σ oznacza optymalnej wielkości okno:

$$W_{\sigma^2}(x, x_i) = \frac{1}{(2\pi\sigma^2)^{d/2}}\exp\left(-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right). \tag{4}$$

Można pokazać, że entropię całego układu można wyznaczyć jako:

$$H = -\log\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} W_{2\sigma^2}\left(x_j, x_i\right).$$

Niech dane będą podzielone na K klastrów: C_k dla k = 1,2,…K, w których pojawia się N_k obiektów. Entropię w k-tym klastrze można zdefiniować jako:

$$H(C_k) = -\log\frac{1}{N_k^2}\sum_{j=1}^{N_k}\sum_{i=1}^{N_k} W_{2\sigma^2}\left(x_j, x_i\right). \tag{5}$$

Wskaźnik postaci:

$$V(C_1, C_2, \ldots C_K) = \sum_{i=1}^{K}\frac{N_k}{N}H(C_K) \tag{6}$$

mógłby być interpretowany jako wskaźnik entropii wewnątrzgrupowej. W literaturze pojawiła się taka ważona suma entropii w każdym klastrze, ale wyznaczana dla danych dyskretnych [Rendón i in. 2011]. Stosując to rozumowanie dla entropii Reny'ego, otrzymujemy (6). Entropię pomiędzy grupami zdefiniujmy jako:

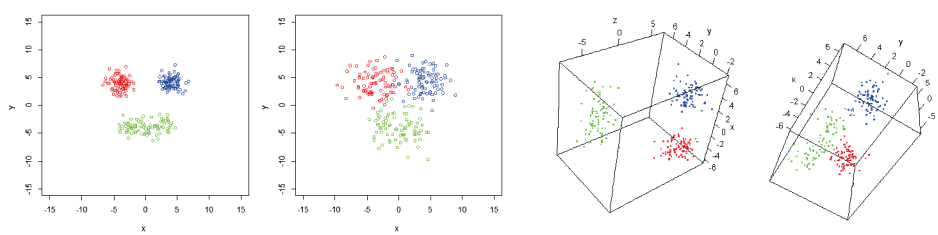
$$H(C_1, C_2, \dots C_K) = -\log \frac{1}{2\prod_{k=1}^{K} N_k} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} W_{2\sigma^2}\left(x_j, x_i\right). \tag{7}$$

Jeśli klasy są dobrze utworzone, wskaźnik ten powinien być duży [Jenssen i in. 2003]. Jako wskaźnik klasyfikacji można rozważyć zatem następujący iloraz:

$$E = \frac{H(C_1, C_2, \dots C_K)}{V(C_1, C_2, \dots C_K)}. \tag{8}$$

4. Badanie empiryczne

W pierwszym eksperymencie symulacyjnym rozważane były dane pochodzące z wielowymiarowego rozkładu normalnego. W przypadku rozkładów dwu- i trójwymiarowych zadeklarowano klasyfikacje na trzy grupy. Liczba elementów w poszczególnych grupach w pierwszym eksperymencie była taka sama, ale zbadano też przypadki różnej liczebności klas. Grupy różniły się przede wszystkim środkami ciężkości. Przy generowaniu danych wielowymiarowych rozważano 20-, 30- i 36- wymiarowe wektory z podziałem na dwa, trzy i cztery skupienia. Badanie przeprowadzono dla różnych liczebności prób (20, 40 i 60 elementów). W każdym eksperymencie tak dobrano środki ciężkości klas i rozrzut elementów, aby rozważane zbiory były mniej lub bardziej separowalne. Opis modeli symulacyjnych prezentuje tab. 1, a przykładowe dane zaprezentowano na rys. 1.



Rys. 1. Przykłady danych symulacyjnych dwuwymiarowych i trójwymiarowych

Źródło: opracowanie własne.

W tabeli 2 porównane zostały wyniki uzyskane przy zastosowaniu klasyfikacji hierarchicznej metodą Warda, klasyfikacji *k*-średnich oraz klasyfikacji spektralnej z wybraną miarą odległości euklidesowej dla eksperymentów o numerach 1, 2, 3 i 4 z tab. 1. Przy stosunkowo słabo separowalnych grupach otrzymuje się niski procent poprawnie określonych podgrup. Na tle badanych wskaźników indeks E wypada korzystnie w grupie 1.

Tabela 1. Charakterystyka modeli symulacyjnych danych o wielowymiarowym rozkładzie normalnym

Nr	Liczba klas	Liczba zmiennych	Liczba pomiarów	Środki ciężkości klas	Macierz kowariancji
1	3	2	100	(-6,4) (6, -4) (6, 6)	$\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \begin{bmatrix} 8 & 0.9 \\ 0.9 & 8 \end{bmatrix}$
2	3	2	50	(0, 0), (1.5, 7), (3, 14)	$\begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$
3	3	3	30,50,50	(-4,4,0), (4,4,0), (0,-4,5)	macierz jednostkowa
4	3	3		(1.5,6,-3), (3,12,-6), (4.5,18,-9)	$\begin{bmatrix} 1 & -0.9 & -0.9 \\ -0.9 & 1 & 0.9 \\ -0.9 & 0.9 & 1 \end{bmatrix}$
5	2	20	20, 40	(4,4,…4,0,…,0), (0,…0,4,…,4)	macierz jednostkowa
6	3	30	60	(4,…,4,0,…,0,0,…,0) (0,…,0,4,…,4,0,…,0) (0,…,0,0,…,0,4,…,4)	macierz jednostkowa
7	4	36	36	(4,…,4,0,…,0,0,…,0,0,…,0)), (0,…,0,4,…,4,0,…,0,0,…,0) (0,…,0,0,…,0,4,…,4,0,…,0) (0,…,0, 0,…,0,0,…,0,4,…,4)	macierz jednostkowa

W punkcie 2) wykorzystano model 13 z pakietu clusterSim [Walesiak 2013], natomiast w punkcie 4) model 5 z tego pakietu; dane z punktów 5, 6 i 7 były generowane przy założeniu niezależności.

Źródło: opracowanie własne.

Tabela 2. Liczba poprawnie określonych podgrup (w procentach)

Indeks	Metody klasyfikacji											
	metoda Warda				metoda *k*-średnich				metoda spektralna			
	Nr 1	Nr 2	Nr 3	Nr 4	Nr 1	Nr 2	Nr 3	Nr 4	Nr 1	Nr 2	Nr 3	Nr 4
CH	45	13	98	68	16	15	99	76	13	17	98	65
DB	15	15	95	76	14	25	97	65	52	14	99	62
S	17	10	98	85	5	14	96	82	3	15	95	83
E	45	14	99	69	55	16	86	67	56	16	98	67

CH – indeks Calińskiego i Harabasza; DB – indeks Davies-Bouldina; S – Silhouette indeks; E – indeks entropii; nr to numer modelu z tab. 1.

Źródło: opracowanie własne.

Wyniki eksperymentów o numerach 5, 6 i 7 nie zostały zestawione w tabeli, ponieważ dotyczyły bardzo dobrze separowalnych danych, których grupy są bardzo dobrze wykrywane przez wszystkie wskaźniki. Liczba poprawnie określonych podgrup wahała się od 96% do 100%. Także wskaźnik *E* na równi z pozostałymi trzema we wszystkich przypadkach dawał poprawne wyniki w 97-99%. Zatem można zauważyć, że indeks klasyfikacji konstruowany na podstawie entropii Reny'ego (*E*) ma bardzo dobre własności. Jego skuteczność w wykrywaniu liczby klas jest porównywalna ze skutecznością klasycznych indeksów.

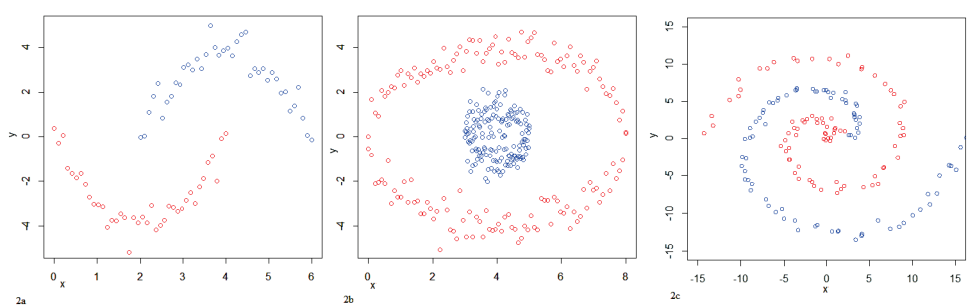
W drugim zestawie modeli symulacyjnych (tab. 3) rozważono niekorelowane dane o rozkładzie skośnym *t*-Studenta. Przyjęto dwa parametry skośności $\eta = 1.7$ i $\eta = 0.6$. W ten sposób osłabione zostało założenie o eliptyczności rozkładów.

Tabela 3. Charakterystyka modeli symulacyjnych danych o rozkładach brzegowych skośnych

Nr	Liczba klas	Liczba zmiennych	Liczba pomiarów	Środki ciężkości klas	Macierz kowariancji i przyjęte stopnie swobody i parametry skośności
8	2	20	20, 40	(1,1,…1,0,…,0), (0,…0,1,…,1)	macierz jednostkowa, *df* = 5, η = 1.7
9	2	20	20, 40	(4,4,…4,0,…,0), (0,…0,4,…,4)	macierz jednostkowa, *df* = 5, η = 0.6
10	3	30	60	(1,…,1,0,…,0,0,…,0) (0,…,0,1,…,1,0,…,0) (0,…,0,0,…,0,1,…,1)	macierz jednostkowa, *df* = 5, η = 1.7
11	3	30	60	(4,…,4,0,…,0,0,…,0) (0,…,0,4,…,4,0,…,0) (0,…,0,0,…,0,4,…,4)	macierz jednostkowa, *df* = 5, η = 0.6
12	4	36	36	(1,…,1,0,…,0,0,…,0,0,…,0)), (0,…,0,1,…,1,0,…,0,0,…,0) (0,…,0,0,…,0,1,…,1,0,…,0) (0,…,0, 0,…,0,0,…,0,1,…,1)	macierz jednostkowa, *df* = 5, η = 1.7
13	4	36	36	(4,…,4,0,…,0,0,…,0,0,…,0)), (0,…,0,4,…,4,0,…,0,0,…,0) (0,…,0,0,…,0,4,…,4,0,…,0) (0,…,0, 0,…,0,0,…,0,4,…,4)	macierz jednostkowa, *df* = 5, η = 0.6

Dane generowane z założeniem niezależności i uwzględnieniem skośności ($\eta = 0.6$, $\eta = 1.7$) w rozkładzie brzegowym *t*-Studenta.

Źródło: opracowanie własne.



Rys. 2. Przykładowe wygenerowane dane nieklasyczne: 2a – dane typu Worms, 2b – dane typu koła, 2c – dane typu spirala

Źródło: opracowanie własne.

W wyniku eksperymentów od 8 do 13 liczba poprawnie określonych podgrup wahała się od 90% do 100% dla wszystkich wskaźników. W eksperymencie nr 8 i 10 najlepszym wynikiem było 96% (dla *E* i *S* w metodzie Warda), przy czym pozostałe wartości były niewiele niższe. W pozostałych grupach eksperymentów symulacyjnych (9,11,12,13) wszystkie wskaźniki, także wskaźnik *E,* dawały poprawne wyniki w 97-99%.

W trzecim eksperymencie weryfikowano wartości wskaźników na podstawie nieklasycznych zbiorów danych. Dane te były losowo zaburzane w taki sposób, aby nie utraciły swojej struktury grupowej. Przykładowe wygenerowane zbiory danych przedstawia rys. 2.

Niestety w tych eksperymentach symulacyjnych wszystkie cztery badane wskaźniki wypadają słabo. Jedynie dane typu „Worms" są dobrze oceniane, bowiem w tym przypadku liczba poprawnie określonych podgrup wynosiła od 78% do 98%. Najlepiej wypadła metoda aglomeracyjna Warda i metoda spektralna. Wyniki poprawnie określonej liczby podgrup dla danych typu dwa koła wynosiły od 18% do 65%. Najlepiej w tym przypadku wypadła metoda spektralna: wskaźnik *CH* – 65%, a wskaźnik *S* – 56%. Badany wskaźnik *E* tylko w ok. 18% poprawnie wykrywała podział na dwie grupy przy wszystkich rozważanych metodach klasyfikacji.

Wybór dwóch klas przy danych typu „spirala" potwierdzał się tylko w 10% dla wszystkich badanych indeksów.

5. Zakończenie

W pracy przedstawiono badanie przydatności wskaźnika jakości klasyfikacji opartego na własnościach entropii Reny'ego. Uzyskane wyniki dla wybranych danych symulacyjnych wskazują na dużą przydatność tego wskaźnika do określenia liczby klas. Indeks *E* generalnie jest podobny do innych klasycznych wskaźników. Badanie symulacyjne pokazało, że dla dobrze separowalnych wskaźnik *E*, na równi z innymi indeksami, poprawnie wykrywa strukturę grupową danych. Jednakże w przypadku danych o bardzo rozbudowanej strukturze indeks ten, na równi z pozostałymi indeksami, nie wybiera właściwie liczby klas.

Wydaje się zatem, że wskaźnik *E* mógłby być wykorzystany jako miara jakości klasyfikacji na równi z pozostałymi klasycznym wskaźnikami. Może służyć jako informacja wspomagająca wybór właściwej decyzji.

Literatura

Baarsch J., Celebi M.C., 2012, *Investigation of Internal Validity Measures for K-Means Clustering*, IMECS 2012, Hong Kong.

Basiura B., Czapkiewicz A., 2014, *Badanie jakości klasyfikacji szeregów czasowych*, PN 327 Taksonomia 22, *Klasyfikacja i analiza danych – teoria i zastosowania*, Jajuga K., Walesiak M. (red.), Uniwersytet Ekonomiczny we Wrocławiu.

Davies D., Bouldin D., 1979, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence 1(2): 224-227.

Calinski R.B., Harabasz J., 1974, *A dendrite method for cluster analysis*, Communications in Statistics – Theory and Methods 3(1): 1-27.

Halkidi M., Yannis B., Vazirgiannis M., 2001, *On clustering validation techniques*, Journal of Intelligent Information Systems, 17:2/3, 107-145.

Jenssen R., Hild K.E., Erdogmus D., Principe J.C., Eltoft T., 2003, *Clusterin using renyi's entropy*, Neural Networks, Proceedings of the International Joint Conference on (Volume:1).

Liang J., Zhao X., Li D., Cao F., Dang C., 2011, *Determining the number of clusters using information entropy for Mixed Data*, Patter Recognition, v. 45, 2251-2265.

Milligan G., Glenn W., 1981, *A Monte Carlo study of thirty internal criterion measures for cluster analysis*, Psychometrika 46(2): 187-199.

Rendón E., Abundez I., Arizmendi A., Quiroz E.M., 2011, *Internal versus external cluster validation indexes,* Intenational Journal of Computers and Communications, no. 1, vol. 5.

Rényi A., 1961, *On measures of information and entropy*, Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960, pp. 547-561.

Rousseeuw P.J., 1987 Silhouettes: A , *Graphic aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics 20(1): 53-65.

Walesiak M., Gatnar E., 2009, *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa.

Walesiak M., 2013, *Zagadnienie doboru liczby klas w klasyfikacji spektralnej, Taksonomia 20 – Klasyfikacja i analiza danych – teoria i zastosowania*, UE, Wrocław.

Walesiak M., Dudek M., 2012, *Package 'clusterSim' in R project*, http://keii.ue.wroc.pl/clusterSim/index.html.

Wędrowska E., 2011, *Wykorzystanie entropii Shanona i jej uogólnień do badania rozkładu prawdopodobieństwa zmiennej losowej dyskretnej*, Przegląd Statystyczny, RLVII, zeszyt 4.

R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.12.2 (2011-02-25) R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Kovács F., Legány C., & Babos A., 2005, *Cluster validity measurement techniques. Proceedings of the 6th Iternational Symposium of Hungarian Researchers on Computational Intelligence*, Budapest, Nov. 2005, 18-19.

SIMULATION STUDY OF THE USE OF ENTROPY TO VALIDATION OF CLUSTERING

Summary: The aim of this paper is to present a quality index classification using Renyi entropy against known quality indicators grouping of multidimensional time series. The study was conducted on artificially generated data and empirical data. The division into groups was made by using Ward's agglomeration algorithm, *k*-means method's and spectral clustering. The results were verified using the selected indices of clustering validation.

Keywords: clustering validation, Renyi's entropy, clustering.