PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

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

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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Beata Basiura, Anna Czapkiewicz
AGH Kraków

BADANIE JAKOŚCI KLASYFIKACJI SZEREGÓW CZASOWYCH

Streszczenie: Celem niniejszej pracy było zaprezentowanie wskaźnika jakości klasyfikacji z zastosowaniem entropii Renyiego na tle znanych wskaźników jakości grupowania wielowymiarowych szeregów czasowych. Punktem wyjścia były dane empiryczne. Podziału na grupy dokonano przy zastosowaniu algorytmu aglomeracji Warda, klasyfikacji k-średnich oraz klasyfikacji spektralnej. Otrzymane wyniki klasyfikacji zweryfikowano stosując wybrane indeksy określające poprawność klasyfikacji. Zaproponowany wskaźnik wydaje się obiecujący, ale wymaga zweryfikowania dla różnych rozkładów badanych szeregów.

Słowa kluczowe: wskaźnik klasyfikacji, entropia Renyiego, klasyfikacja szeregów czasowych.

1. Wstęp

Empiryczne badanie jakości struktury grupowej danych jest zadaniem niezwykle trudnym. W literaturze przedmiotu można znaleźć wiele różnych wskaźników miary jakości klasyfikacji. Wyróżnia się metody wzorcowe, w których uzyskany podział na grupy porównuje się z pewnym podziałem wzorcowym, oraz metody bezwzorcowe, w których ocena jakości klasyfikacji wykorzystuje jedynie informacje zawarte w danych. Większość wskaźników opiera się na miarach zwartości klas takich, jak na przykład wariancja wewnątrzgrupowa, oraz na miarach separowalności poszczególnych podgrup określających zmienność międzygrupową.

Istnieją trzy klasy wskaźników oceny jakości grupowania (por. Halkidi 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*). Większość z nich wykorzystuje miarę odległości pomiędzy obiektami i zazwyczaj lepiej ocenia grupy „eliptyczne", natomiast w przypadku nietypowych podzbiorów danych wydaje się, że nie dają one wiarygodnych wyników.

Wszelkie decyzje podejmowane przez badaczy, związane z inwestycjami, rozwojem, wielkością sprzedaży i ceną, mogą być oparte na wartości szeregu czaso-

wego. Ocena wielkości zmian rynkowych i aspekty, do której grupy zaklasyfikować badane szeregi, zależy nie tylko od wyboru miary podobieństwa czy metody grupowania. Ostateczna decyzja dotycząca wyników grupowania powinna być podejmowana na podstawie oceny jakości uzyskanej klasyfikacji. W niniejszej pracy zaproponowano ocenę jakości klasyfikacji krótkich szeregów czasowych, wykorzystującą entropię Rényiego [Rényi 1961; Wędrowska 2010].

Prezentowana praca ma na celu porównanie wybranych wskaźników jakości grupowania wielowymiarowych szeregów czasowych. Punktem wyjścia były dane empiryczne. Podziału na grupy dokonano przy zastosowaniu algorytmu aglomeracji Warda, klasyfikacji *k*-średnich oraz klasyfikacji spektralnej. Otrzymane wyniki klasyfikacji zweryfikowano stosując wybrane wskaźniki określające poprawność klasyfikacji. Zaproponowany wskaźnik, wykorzystujący funkcję entropii, został przedstawiony na tle istniejących już w literaturze wskaźników klasyfikacji. Dla zbadania jakości prezentowanego wskaźnika wykonano krótkie badanie symulacyjne.

Celem pracy było pokazanie wskaźnika skonstruowanego na podstawie entropii Rényiego w stosunku do innych opisanych w literaturze i stosowanych wskaźników jakości klasyfikacji przy założeniu nieznajomości rozkładu badanych szeregów [Milligan, Glenn 1981; Halkidi i in. 2010; Rendón i in. 2011; Walesiak, Gatnar (red.) 2009].

2. Wybrane wskaźniki klasyfikacji

Jak już wspomniano, wyróżnia się trzy klasy wskaźników jakości klasyfikacji. Wskaźniki zewnętrzne, w których uzyskana struktura porównywana jest z założoną, znaną z góry – ekspercką – strukturą danych, wskaźniki wewnętrzne, wykorzystujące jedynie informacje z analizowanego zbioru danych, oraz wskaźniki względne, gdy ocena struktury porównywana jest z grupowaniem uzyskanym za pomocą tego samego algorytmu, ale z założonymi innymi parametrami (np. inna liczba grup). Większość wskaźników opiera się na miarach spójności (zwartości) skupień: elementy każdego skupienia powinny odpowiednio blisko siebie (*custer cohesion*, *cluster compactness*) oraz na miarach rozdzielenia skupień: klastry powinny być odpowiednio od siebie oddalone (*cluster separation*). Popularną miarą zwartości jest wariancja.

W niniejszej pracy wybrane zostały trzy najczęściej używane, wewnętrzne współczynniki: indeks Calińskiego i Harabasza [Calinski, Harabasz 1974], indeks Daviesa-Bouldina [Davies, Bouldin 1979] oraz Silhouette indeks [Rousseeuw 1987].

Pierwszy z nich oparty jest na zmienności międzygrupowej (*betwen groups* – BG) oraz zmienności wewnątrz grupowej (*within groups* – WG). Zmienność międzygrupowa jest ważoną sumą kwadratów odległości pomiędzy środkiem każdej klasy a środkiem całego zbioru. Wagami są wielkości klastrów.

$$CH = \frac{tr(BG)/(K-1)}{tr(WG)/(N-K)}. \tag{1}$$

Zmienność wewnątrzgrupowa wyznaczana jest jako suma kwadratów odległości każdego elementu podzbioru od środka klasy.

Wybrany drugi indeks także wykorzystuje iloraz zwartości i separowalności klasy. Dla każdego skupienia wyznacza się w nim średnią odległość pomiędzy każdym punktem grupy a jej centrum (oznaczmy je jako δ_k i$\delta_{k\prime}$). Oznaczmy ko $\Delta_{kk\prime}$ odległość pomiędzy środkami skupienia k i skupienia k'. Następnie dla każdego podzbioru wyznacza się maksymalną wartość ilorazu: $\frac{\delta_k+\delta_{k\prime}}{\Delta_{kk\prime}}$. Wtedy indeks DB oblicza się według wzoru (2):

$$DB = \frac{1}{K}\sum_{k=1}^{K} \max_{k\neq k\prime} \frac{\delta_k+\delta_{k\prime}}{\Delta_{kk\prime}}. \tag{2}$$

Trzeci z indeksów, zaproponowany przez Rousseeuwa [1987], pozwala oceniać prawidłowość zaklasyfikowania poszczególnych obiektów do wyodrębnionych klas. Dany jest wzorem (3):

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i);b(i)\}}. \tag{3}$$

Określa się, że $a(i)$ to średnia odległość obiektu i od pozostałych obiektów należących do klasy C_k, $b(i)$ to minimalna ze średnich odległości obiektu i od obiektów należących do klasy $C_{k'}$, co zapisuje się następująco:

$$b(i) = \min_{k\neq k'}\{d_{iC_{k'}}\}, \text{ gdzie } d_{iC_k} = \frac{1}{n_k}\sum_{m\in C_k} d_{im}.$$

Na tej podstawie określona zostaje $S(P_r)$ jako prawidłowość wyodrębnienia poszczególnych klas (oznaczonych przez P_r) oraz wskaźnik $S(P)$, opisujący ogólną jakość klasyfikacji dane równaniami (5).

$$S(P_r) = \frac{1}{n_k}\sum_{i\in P_r} S(i),\ \ S(P) = \frac{1}{k}\sum_r S(P_r). \tag{5}$$

We wszystkich wymienionych indeksach wykorzystano odległość zastosowaną do wyznaczenia klasyfikacji. W obliczeniach niektórych indeksów wykorzystano procedury z pakietu clusterSim programu R [Walesiak, Dudek 2012; R. Development Core Team 2005].

O lepszej jakości klasyfikacji mówią wyższe wartości indeksu Calińskiego i Harabasza, a także niższe indeksy Daviesa-Bouldina. Silhouette indeks z przedziału od 0,5 do 0,7 świadczy o poważnej strukturze klas, natomiast wartości wyższe niż 0,7 charakteryzują silną strukturę klas [Gatnar, Walesiak (red.) 2004].

3. Entropia

Niektóre sposoby określania miary podobieństwa wywodzą się z teorii informacji. Zakłada się, że takie grupowanie, które daje największy przyrost informacji, jest optymalne, gdyż odpowiada to małemu zróżnicowaniu kategorii w podzbiorach. Entropia jako miara zróżnicowania wydaje się funkcją bardzo uniwersalna, niezależną od charakteru zmiennych.

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

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

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

$$H(x) = -\log(\int f^{2}(x))\, dx. \tag{6}$$

Niech $\{x_1, \dots, x_N\}$, 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 w oparciu o estymację jądrową [Liang i in. 2011; Jenssen i in. 2003]. Niech:

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

W naszych badaniach została wykorzystana funkcja jądrowa Gaussa, określona wzorem (7).

$$W_{\sigma^2}(x, x_i) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right). \tag{7}$$

Można pokazać, że entropię układu można wyznaczyć jako:

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

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

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

Wskaźnik postaci:

$$V(C_1, C_2, \dots C_K) = \sum_{i=1}^{K}\frac{N_k}{N}H(C_K) \tag{8}$$

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

Rényiego, otrzymujemy równanie (8). Ponadto zdefiniujmy entropię pomiędzy grupami jako:

$$H\left(C_{1,}C_{2,}\dots C_K\right) = -\log\frac{1}{2\prod_{k=1}^{K}N_k}\sum_{j=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{N}W_{2\sigma^2}\left(x_j,x_i\right). \tag{9}$$

Jeśli klastry są dobrze wybrane, wskaźnik ten powinien być odpowiednio duży [Jenssen i in. 2003]. Jeśli zatem zastosujemy iloraz:

$$V = \frac{H(C_{1,}C_{2,}\dots C_K)}{V(C_{1,}C_{2,}\dots C_K)}, \tag{10}$$

to otrzymamy wskaźnik jakości grupowania, zgodny z ideą tworzenia wskaźnika CH, ale może lepiej działający dla grup z różnych rozkładów, w szczególności dla grup nietypowych.

4. Badanie empiryczne

Własności wskaźnika przedstawionego wzorem (10) zostały zaprezentowane na tle innych wybranych indeksów dla danych empirycznych. W badaniu empirycznym porównane zostały wyniki klasyfikacji dla czterech zestawów danych. Pod uwagę zostały wzięte dane empiryczne w postaci krótkich szeregów czasowych. W przykładzie pierwszym i drugim zastosowano klasyfikację 24 spółek branży informatycznej notowanych na GPW w Warszawie. W przykładzie pierwszym brano pod uwagę zysk netto tych spółek w latach (na zakończenie roku kalendarzowego) 2004-2011. W drugim spółki te klasyfikowano pod kątem wartości ich przychodów w tym samym okresie. Długość szeregu określona została na $d = 8$. W przykładzie trzecim rozważono szeregi czasowe zawierające wartości procentowe miesięcznej inflacji (zmiany miesiąc/miesiąc poprzedni) z 22 wybranych krajów w okresie od stycznia 2010 do czerwca 2013. W tym przypadku wymiar każdego szeregu wyniósł 42 pomiary. Ostatni zbiór danych zawiera liczbę rejestracji samochodów dokonywaną na terenie 16 wybranych państw w poszczególnych miesiącach roku 2011. Rozważano 12 pomiarów dla każdego szeregu. Uzyskane wartości wskaźników jakości klasyfikacji zaprezentowano w tabelach 1, 2, 3 i 4.

Tabela 1. Wskaźniki grupowania zysku netto w latach 2004-2011 spółek branży informatycznej

Liczba grup	Ward				*k*-means				spectral			
	CH	DB	S	E	CH	DB	S	E	CH	DB	S	E
2	17,20	1,05	**0,41**	**1,97**	**17,97**	1,11	0,41	**1,96**	**16,99**	1,13	**0,40**	1,73
3	**17,76**	1,06	0,41	1,83	17,96	**1,05**	**0,42**	1,83	13,35	1,28	0,37	1,90
4	16,85	**0,99**	0,40	1,97	13,05	1,18	0,27	1,62	15,94	**0,99**	0,39	**1,93**

Źródło: opracowanie własne.

W tabeli 1 porównane zostały wyniki uzyskane przy zastosowaniu klasyfikacji hierarchicznej metodą Warda, klasyfikacji *k*-średnich oraz klasyfikacji spektralnej z wybraną miarą odległości euklidesowej. Najlepsze wskaźniki zostały wytłuszczone. Należy zwrócić uwagę na zgodność współczynnika konstruowanego w oparciu o entropię ze wskaźnikiem CH, ale tylko w przypadku metody *k*-średnich. W przypadku metody Warda i klasyfikacji spektralnej wnioski są rozbieżne. Dla tej ostatniej wskaźnik entropii jest zgodny ze wskaźnikiem DB, a dla metody aglomeracyjnej Warda daje wynik taki sam jak indeks S. Wybierając najlepszą klasyfikację na podstawie wskaźnika CH, otrzymujemy podział na dwie grupy metodą *k*-średnich. Indeks DB wskazuje na klasyfikację metodą Warda i klasyfikację spektralną z podziałem na cztery grupy. Wskaźnik S jako najlepszą określa metodę *k*-średnich z liczbą czterech grup. Natomiast według proponowanego indeksu E najlepsze byłaby grupowanie metodą Warda z podziałem na dwie grupy.

Tabela 2 zawiera wyniki uzyskane dla zbioru danych zawierającego wartości przychodów wybranych spółek branży informatycznej. We wszystkich grupowaniach wartości wskaźników wskazały na wybór podziału zbioru spółek na dwa zbiory. Należy zwrócić uwagę, że indeks entropii wskazał na grupowanie analogicznie jak pozostałe wskaźniki jakości klasyfikacji.

Tabela. 2. Wskaźniki grupowania przychodów w latach 2004-2011 spółek branży informatycznej

Liczba grup	Ward				*k*-means				spectral			
	CH	DB	S	E	CH	DB	S	E	CH	DB	S	E
2	**21,00**	**0,66**	**0,57**	**2,24**	**21,00**	**0,66**	**0,57**	**2,24**	**21,00**	**0,66**	**0,57**	**2,24**
3	18,46	1,13	0,31	1,42	18,46	1,13	0,31	1,42	3,55	2,96	0,10	1,36
4	14,95	1,28	0,24	1,47	15,34	1,14	0,26	1,46	13,82	1,03	0,22	1,42

Źródło: opracowanie własne.

Tabela 3. Wskaźniki grupowania inflacji z 22 krajów w okresie od stycznia 2010 do czerwca 2013

Liczba grup	Ward				*k*-means				spectral			
	CH	DB	S	E	CH	DB	S	E	CH	DB	S	E
2	**7,57**	1,54	**0,24**	**1,97**	**7,57**	1,54	**0,24**	1,97	**5,91**	**1,80**	**0,16**	1,66
3	5,63	1,70	0,14	1,72	5,63	**1,29**	0,24	**1,99**	3,59	2,25	0,11	**1,70**
4	5,07	**1,52**	0,16	1,74	4,69	1,55	0,14	1,60	2,29	2,37	0,05	1,58

Źródło: opracowanie własne.

Przy klasyfikacji wskaźnika inflacji proponowany indeks wykorzystujący miarę entropii dla algorytmu Warda ustala liczbę klas podobnie jak indeks CH. Przy zastosowaniu metody *k*-średnich wskaźnik E pozwala na określenie, że najlepszą klasyfikację otrzyma się przy podziale na trzy grupy, zgodnie z indeksem DB. Niestety, pozostałe dwa wskaźniki proponują podział zbioru szeregów na dwie grupy. W przypadku grupowania metodą spektralną wskaźnik entropii jako najlepszą kla-

syfikację proponuje podział na trzy grupy, w przeciwieństwie do pozostałych wskaźników, które określają podział 22 państw na dwie grupy jako lepszy.

Tabela 4. Wskaźniki grupowania dla miesięcznej liczby rejestrowanych samochodów w 16 krajach w roku 2011

Liczba grup	Ward				*k*-means				spectral			
	CH	DB	S	E	CH	DB	S	E	CH	DB	S	E
2	5,76	1,66	0,23	1,51	6,01	1,44	**0,23**	1,75	3,89	1,88	**0,15**	1,73
3	5,16	1,37	0,20	1,61	**6,03**	**1,24**	0,22	**1,79**	**4,02**	1,69	0,14	**1,74**
4	**6,10**	**1,20**	**0,23**	**1,72**	4,87	1,34	0,17	1,77	3,29	**1,44**	0,11	1,66

Źródło: opracowanie własne.

W klasyfikacji Warda wszystkie indeksy wybrały jako najlepszą liczbę czterech podgrup. Klasyfikacja metodą *k*-średnich jest mniej jednoznaczna, ale indeks E daje wynik zgodny z wskaźnikami CH i DB. W klasyfikacji spektralnej indeks oparty na entropii jest zgodny jedynie z indeksem CH. Indeksy CH i DB określają jako najlepszą klasyfikację metodą aglomeracji Warda z podziałem na cztery grupy, natomiast według wskaźnika S najlepszy wynik daje metoda *k*-średnich z podziałem na dwie grupy. Indeks E proponuje natomiast wybrać metodę *k*-średnich, ale z podziałem na trzy podzbiory.

Tabela 5. Wskaźniki grupowania dla wygenerowanych modelowych danych

Liczba grup	Ward				*k*-means				spectral			
	CH	DB	S	E	CH	DB	S	E	CH	DB	S	E
2	13,29	1,09	0,29	2,89	12,81	1,16	0,28	2,75	11,56	1,71	0,26	2,78
3	19,91	1,02	0,41	2,95	19,97	1,02	0,42	3,10	17,42	1,11	0,39	2,95
4	**40,96**	**0,67**	**0,55**	**3,12**	**40,96**	**0,67**	**0,55**	**3,12**	**40,96**	**0,67**	**0,55**	**3,12**
5	33,2	0,87	0,47	2,05	31,47	1,07	0,42	1,66	33,21	0,87	0,47	2,05

Źródło: opracowanie własne.

Aby zbadać przydatność proponowanego indeksu do określenia jakości klastrowania, wykonano krótkie badanie symulacyjne, które jest wstępem do dalszych badań. Wygenerowano zestawy danych pochodzących z populacji wielowymiarowych rozkładów normalnych o określonej liczbie klas, a następnie wspomnianymi metodami dokonano ich klasyfikacji i wyznaczono omawiane wskaźniki. Zazwyczaj wyniki indeksów dobrze wykrywały strukturę grupową wygenerowanych danych. W tabeli 5zawarto wyniki dla wygenerowanych 36 wektorów o długości 20 pomiarów. Dane pochodziły z populacji podzielonej na cztery klasy.

5. Podsumowanie

W pracy przedstawiono próbę określenia jakości klasyfikacji za pomocą wskaźnika opartego na entropii Rényiego. Uzyskane wyniki dla wybranych danych empirycznych wskazują na podobieństwo do istniejących wskaźników klasyfikacji. Proponowany indeks E najczęściej dawał wyniki zgodne ze wskaźnikami CH i DB, ale były przypadki, w których zachowywał się inaczej. Krótkie badanie symulacyjne pokazało, że dla wielowymiarowych rozkładów normalnych wskaźnik E, na równi z innymi wskaźnikami, poprawnie wykrywa strukturę grupową danych. Zaproponowany wskaźnik mógłby być wykorzystany jako miara jakości klasyfikacji w przypadku nieznajomości rozkładów szeregów czasowych i jest zachętą do dalszych badań.

Literatura

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

Davies D., Bouldin D. (1979), *A Cluster Separation Measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence 1(2), s. 224-227.

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

Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo AE, Wrocław.

Halkidi M., Yannis B., Vazirgiannis M. (2001), *On Clustering Validation Techniques*, „Journal of Intelligent Information Systems", 17, 2/3, s. 107-145.

Jenssen R., Hild K.E., Erdogmus D., Principe J.C., Eltoft T. (2003), *Clusterin Using Renyi's Entropy*, Proceedings of the International Joint Conference on Neural Networks, Vol. 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, Vol. 45, s. 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*, „International 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, s. 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), s. 53-65.

Walesiak M., Gatnar E. (red.) (2009), *Statystyczna analiza danych z wykorzystaniem programu R*, Wyd. Naukowe PWN, Warszawa.

Walesiak M., Dudek M. (2012), *Package 'clusterSim' in R project*, http://keii.ue.wroc.pl/clusterSim/index.html (30.08.2013).

Wędrowska E. (2010), *Wykorzystanie entropii Shanona i jej uogólnień do badania rozkładu prawdopodobieństwa zmiennej losowej dyskretnej*, „Przegląd Statystyczny", LVII, 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, http://www.R-project.org (30.08.2013).

VALIDATION OF TIME SERIES CLUSTERING

Summary: The aim of this paper is to present a quality index classification using Renyi's entropy against known quality indicators grouping of multidimensional time series. The starting point was the 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. The proposed index seems to be promising but it needs to be verified for various distributions of time series.

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