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

Kamila Migdał-Najman, Krzysztof Najman
Uniwersytet Gdański

FORMALNA OCENA JAKOŚCI ODWZOROWANIA STRUKTURY GRUPOWEJ NA MAPIE KOHONENA

Streszczenie: Artykuł dotyczy badania możliwości zastosowania siedmiu wskaźników (średni błąd kwantyzacji, błąd topograficzny, błąd dystorsji, współczynnik Kaskiego-Lagus, zlogarytmowany współczynnik Nasha-Sutcliffe'a, indeks Willmotta i liczba martwych neuronów) do oceny jakości odwzorowania struktury grupowej jednostek na mapie Kohonena. W eksperymencie wykorzystano wygenerowane zbiory danych o znanej strukturze przestrzennej jednostek. Poddano analizie związki między wartościami poszczególnych wskaźników ze strukturą badanej sieci i zgodnością grupowania ze wzorcem.

Słowa kluczowe: sieć samoorganizująca się Kohonena (SOM), formalna ocena jakości odwzorowania.

1. Wstęp

Jednym z typów sztucznych sieci neuronowych, znajdujących szerokie zastosowanie w analizie skupień, są sieci samouczące się. Samouczenie się to bezwzorcowy proces odwzorowywania wielowymiarowej przestrzeni wejściowej badanych jednostek w (niskowymiarową) przestrzeń małej liczby jednostek funkcjonalnych, neuronów, z zachowaniem topograficznego podobieństwa tych jednostek. Do tego typu sieci zalicza się między innymi sieć Kohonena (*Self Organising Map* – SOM). W procesie samouczenia się sieci SOM minimalizuje się najczęściej średni błąd kwantyzacji, a więc dąży się do tego, aby neurony minimalizowały odległości (wybraną metrykę) od odwzorowywanych jednostek. Ponieważ proces samouczenia się nie jest deterministyczny, dla sieci o zadanej topografii możliwe jest uzyskanie różnych odwzorowań o podobnym średnim błędzie kwantyzacji, ale znacząco różnych własnościach z punktu widzenia analizy skupień. Wydaje się, że uzyskane odwzorowanie powinno być poddane dalszej wszechstronnej ocenie.

W literaturze przedmiotu można znaleźć szereg miar oceniających różne aspekty uzyskanego odwzorowania jednostek na sieci SOM. Należą do nich: błąd topograficzny, średni błąd kwantyzacji, błąd dystorsji [Sun 2000; Kohonen 2001; Pölzlbauer 2004], współczynnik Kaskiego-Lagus [Kaski, Lagus 1996], zlogaryt-

mowany współczynnik Nasha-Sutcliffe'a [Nash, Sutcliffe 1970] i indeks Willmotta [Willmott 1981; 1982]. Brakuje jednak badań wskazujących na ich przydatność do oceny sieci SOM z punktu widzenia możliwości jej zastosowania w analizie skupień. W prezentowanych badaniach postawiono hipotezę, że poprawne odwzorowanie struktury przestrzennej jednostek na sieci SOM sprzyja poprawności grupowania jednostek dokonanej w oparciu o taką sieć. Jeżeli hipoteza ta jest prawdziwa, to wymienione wyżej współczynniki mogłyby być wykorzystane do oceny potencjału sieci SOM w grupowaniu jednostek. Celem artykułu jest weryfikacja tej hipotezy. Dokonano w nim analizy konstrukcji i własności badanych miar. Przedstawiono także wyniki badań symulacyjnych, w których poddano analizie związki między wartościami poszczególnych miar ze strukturą uzyskanej sieci SOM i zgodnością grupowania uzyskanego dzięki sieci ze znanym wzorcem.

2. Metody oceny odwzorowania badanych jednostek na sieci SOM

Samouczenie się sieci SOM jest procesem iteracyjnym, który może być zrealizowany w oparciu o jeden z wielu algorytmów [Migdał-Najman, Najman 2013, s. 163]. Niezależnie od przyjętego szczegółowego rozwiązania dla każdej badanej jednostki poszukiwany jest neuron do niej najbliższy (neuron zwycięzca, *Best Matching Unit* – BMU), który wraz z sąsiadami podlega uczeniu (zmianie swoich współrzędnych w przestrzeni). Jeżeli współrzędne jednostki *i*-tej (i = 1,2,…, n) nazwiemy x_i, a współrzędne *m*-tego neuronu (m = 1, 2,…,M) – w_m, to neuron zwycięzca w_c spełnia relację:

$$d(x_i, w_c) = \min_{1 \le m \le M} d(x_i, w_m).$$

Odległość $d(x_i, w_c)$ nazywa się błędem kwantyzacji. Proces zaprojektowany jest w ten sposób, że gdyby neuronów było tyle samo co odwzorowywanych jednostek, to w skończonej liczbie iteracji każdy neuron zostałby BMU dla jednej jednostki i uzyskałby identyczne jak ona współrzędne. Ponieważ jednym z celów budowy sieci SOM jest redukcja liczby jednostek i zastąpienie ich niewielką liczbą neuronów, gdy $M < n$, każdy neuron może odwzorowywać wiele jednostek. Oznacza to, że w procesie samouczenia się sieci SOM minimalizowany jest w istocie średni błąd kwantyzacji (*Mean Quantization Error – MQE*). Jego postać można zapisać następująco [Kohonen 1997]:

$$MQE = \frac{\sum_{i=1}^{n} d(x_i, w_c)}{n}.$$

Średni błąd kwantyzacji może być użyty do oceny jakości uczenia i dopasowania sieci SOM do zbioru odwzorowywanych jednostek. *MQE* jest przeciętną odleg-

łością między każdą jednostką i najbliższym jej neuronem i powinien być jak najmniejszy. Należy zauważyć, że zakładając poprawność procesu samouczenia się[1], dla danego zbioru jednostek *MQE* zmniejsza się wraz ze wzrostem liczby neuronów na sieci SOM. Miara ta nie może służyć do porównania sieci o różnym rozmiarze. Niewielki (bliski zeru) średni błąd kwantyzacji oznacza, że proces samouczenia się sieci pozwolił rozciągnąć się sieci w tej części przestrzeni, w której faktycznie znajdują się badane jednostki. Jest to jednak informacja o charakterze ogólnym, ponieważ faktycznie jest to jedynie średnia arytmetyczna z błędów kwantyzacji.

Aby bardziej precyzyjnie ocenić to rozciągnięcie sieci w przestrzeni, można wyznaczyć wartość błędu topograficznego. Błąd topograficzny (*Topographic Error* – *TE*), nazywany również średnim błędem topograficznym lub błędem topologicznym, pozwala na ocenę topograficznego uporządkowania neuronów na sieci i jakości odwzorowania topograficznego. Określa udział jednostek, dla których dwa najbliższe neurony (pierwsze i drugie BMU, tj. neuron zwycięzca i następny po nim, który w stosunku do danej jednostki był najbliżej) nie są neuronami sąsiadującymi na sieci w ogólnej liczbie jednostek. Definiuje się go następująco:

$$TE = \frac{\sum_{i=1}^{n} l(x_i)}{n}.$$

Dla każdej z n jednostek funkcja $l(x_i)$ przyjmuje wartość równą 1, gdy dwa najbliższe wektorowi danych neurony nie sąsiadują ze sobą na sieci SOM. W przeciwnym wypadku przyjmuje wartość 0. Pożądany poziom błędu topograficznego wynosi 0 i oznacza, że każda jednostka jest odwzorowywana przez dwa sąsiednie neurony, które są dla niej pierwszym i drugim BMU. Taka wartość oznaczałaby, że neurony są bardzo dobrze rozłożone w przestrzeni jednostek [Migdał-Najman, Najman 2013, s. 173-175].

Ani średni błąd kwantyzacji, ani błąd topograficzny nie pozwalają jednak ocenić, czy gęstość neuronów jest odpowiednia dla gęstości jednostek w odpowiednich częściach przestrzeni. Pożądaną, z punktu widzenia analizy skupień, własnością sieci SOM byłoby, gdyby w tej części przestrzeni, w której znajduje się wiele jednostek, znajdowało się proporcjonalnie wiele neuronów. Własność tę można opisać, stosując kolejną miarę, którą jest błąd dystorsji (*Distortion Measure* – DM). Można go zdefiniować następująco:

[1] Teoretycznie możliwe jest zbudowanie sieci o dużej liczbie neuronów, z których większość leży w odległej od jednostek części przestrzeni. Gdy liczba iteracji uczących będzie niewystarczająca, *MQE* dla takiej sieci może być znacznie większy niż dla mniejszej sieci, dla której proces samouczenia się pozwolił na przesunięcie neuronów do tej części przestrzeni, w której faktycznie znajdują się badane jednostki. Sytuację taką jest jednak łatwo wykryć dzięki wizualizacji macierzy ujednolicenych odległości [Migdał-Najman, Najman 2013].

$$DM = \sum_{i=1}^{n}\sum_{m=1}^{M} G\left(R, d\left(c,m\right)\right) \times d\left(x_i, w_m\right),$$

gdzie: $d(x_i, w_m)$ oznacza odległość między jednostką x_i a neuronem w_m, $G\left(R, d(c,m)\right)$ jest przyjętą funkcją sąsiedztwa względem neuronu wygrywającego w_c. Błąd dystorsji powinien być jak najmniejszy, ponieważ wtedy neurony będą równomiernie przydzielone poszczególnym jednostkom i w przybliżeniu taka sama liczba jednostek będzie odwzorowywana przez wszystkie neurony.

W literaturze przedmiotu można znaleźć także inne wskaźniki oceny uzyskanego odwzorowania. W ich konstrukcji bierze się pod uwagę nie BMU, lecz drugi najbliższy neuron, lub ocenia odległość między jednostką a neuronem względem wybranej przeciętnej odległości między jednostkami lub neuronami. Należy do nich współczynnik Kaskiego-Lagus (*Kaski-Lagus Measure – KLM*) [Kaski, Lagus 1996] w postaci:

$$KLM = \frac{\sum_{i=1}^{n} d\left(x_i, w_{c'}\right)}{n}.$$

Jest to odpowiednik średniego błędu kwantyzacji, z tą różnicą, że w stosunku do drugiego BMU. Powinien on być bardziej odporny od *MQE* na skrajne dopasowanie i niedopasowanie niektórych neuronów do odwzorowywanych jednostek. Innym wskaźnikiem jest zlogarytmowany współczynnik Nasha-Sutcliffe'a (*logarithmized Nash-Sutcliffe coefficient of efficiency – CEEF*log), który definiuje się następująco:

$$CEEF\log = \frac{\sum_{i=1}^{n}\left(\ln\left(w_c\right) - \ln\left(x_i\right)\right)^2}{\sum_{i=1}^{n}\left(\ln\left(x_i\right) - \ln\left(\overline{x}\right)\right)^2}.$$

Wyraża on przeciętną odległość między każdą jednostką a jej BMU w stosunku do przeciętnej odległości między każdą jednostką a centrum przestrzeni, w której się one znajdują. Im mniejsza wartość wskaźnika, tym lepsze odwzorowanie [Nash, Sutcliffe 1970; Herbst, Casper 2008; Migdał-Najman, Najman 2013]. Na podobnej idei zbudowany jest indeks Willmotta (*Willmott's index of agreement – IAg*), który definiuje się następująco:

$$IAg = 1 - \frac{\sum_{i=1}^{n}\left(w_c - x_i\right)^2}{\sum_{i=1}^{n}\left(\left|w_c - \overline{x}\right| - \left|x_i - \overline{x}\right|\right)^2}.$$

Willmott wskazywał, że błędy kwantyzacji poszczególnych jednostek powinny być oceniane z punktu widzenia odległości nie tylko między jednostkami a centrum przestrzeni, ale także neuronami a tym centrum. Wskaźnik ten przyjmuje wartości z przedziału $0 \leq IAg \leq 1$. Wartości bliskie 1 oznaczają, że sieć jest poprawnie rozciągnięta w przestrzeni i dobrze odwzorowuje obserwowane jednostki [Willmott 1981; 1982].

Dodatkową miarą poprawności struktury uzyskanej sieci może być liczba martwych neuronów (*MN*). Są to neurony, które biorą udział w procesie samo uczenia się, pośredniczą między odległymi neuronami, ale nie odwzorowują żadnych jednostek. Gdy ich liczba stanowi mały ułamek ogólnej liczby neuronów, jest to zjawisko normalne. Gdy mają istotny udział w ogólnej liczbie neuronów, może to świadczyć o niewłaściwej liczbie neuronów w sieci i będą one utrudniały rozpoznanie struktury grupowej badanych jednostek.

3. Eksperyment badawczy

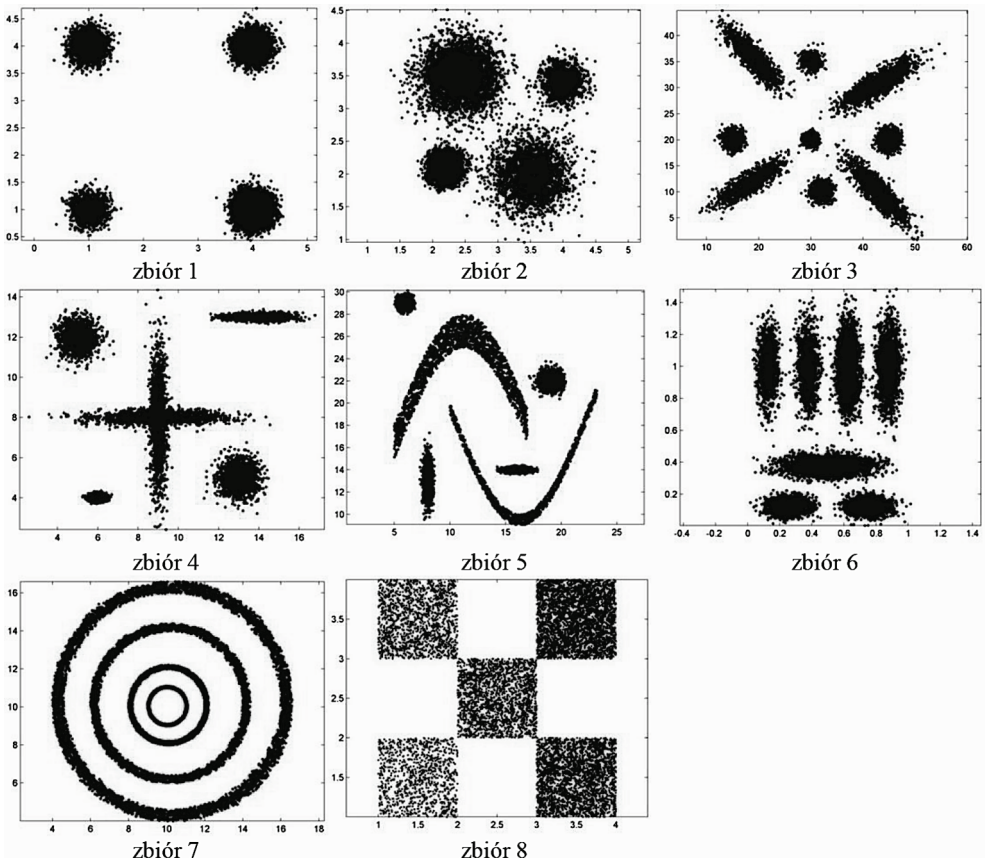
Budowa sieci SOM nie jest procesem deterministycznym, a jej własności zależą od kilku ustalanych *a priori* parametrów. W klasycznym algorytmie Kohonena wstępna konfiguracja neuronów w przestrzeni jest losowa[2]. W kolejnych iteracjach uczących sieć odwzorowuje jednostki prezentowane jej w losowej kolejności. Z tych powodów, powtarzając budowę sieci o przyjętych, stałych parametrach, po przeprowadzeniu procesu samouczenia się można uzyskać sieci o różnych własnościach. Własności te mogą się także istotnie różnić dla sieci o stałym rozmiarze, ale różniących się kształtem, przyjętą funkcją i zasięgiem sąsiedztwa czy typem powiązań neuronów.

Aby zrealizować postawiony we wstępie cel, przeprowadzono eksperyment symulacyjny. Przygotowano osiem zbiorów danych, różniących się istotnie swoją strukturą grupową, liczbą i konfiguracją skupień, liczbą jednostek w skupieniach. Zbiory te zostały zaprezentowane na rysunku 1. Dla każdego z nich budowano sieci SOM o następujących topologiach:

1. rozmiar sieci: od 4×4 do 16×16 neuronów,
2. funkcje sąsiedztwa: gaussowska, ucięta gaussowska, wykładnicza i prostokątna,
3. zasięg sąsiedztwa: od 2 do 4,
4. struktura połączeń neuronów: heksagonalna,
5. sieci uczono procedurą wsadową o liczbie iteracji od 2 do 65.

W oparciu o każdą z uzyskanych sieci dokonano grupowania metodą *k*-średnich [Spath 1985], uzyskując *de facto* grupowanie hybrydowe SOM-KS [Migdał -Najman 2012]. Liczba skupień każdorazowo była ustalana w oparciu

[2] Procedury inicjacji sieci mogą być także inne, np. liniowa. Ich wybór nie ma większego znaczenia dla prezentowanych badań [Migdał-Najman, Najman 2013, s. 163].



Rys. 1. Analizowane zbiory testowe

Źródło: opracowanie własne.

o wskaźnik Daviesa-Bouldina [Davies, Bouldin 1979]. Łącznie wykonano 9828 grupowań (13 × 4 × 3 × 63). Dla każdego z nich wyznaczono wartość wymienionych wyżej siedmiu wskaźników jakości odwzorowania. Ponieważ przynależność każdej jednostki do skupienia we wszystkich zbiorach jest znana, wyznaczono także wartość skorygowanego współczynnika Randa [Rand 1971]. Możliwe było także wyznaczenie wartości współczynnika korelacji Pearsona między badanymi wskaźnikami (zob. tab. 1).

Uzyskane rezultaty są zgodne z oczekiwaniami wynikającymi z analizy konstrukcji badanych wskaźników. Potwierdzają hipotezę, że im lepsze odwzorowanie jednostek na sieci, tym skuteczniejszy może być proces grupowania[3]. Wskazują na to znaki

[3] Jest to prawdą nawet wtedy, gdy grupowanie jest dokonane nieoptymalną metodą. Zbiory 4, 5 i 7 nie mogą być poprawnie pogrupowane metodą *k*-średnich niezależnie od własności sieci SOM.

współczynników korelacji (poza znakami dla *MN*). Wielkość współczynników wskazuje na istotne korelacje, jednak w większości przypadków nie są one wysokie. Przeciętna wartość z modułów współczynnika korelacji wynosi jedynie 0,4375.

Tabela 1. Współczynniki korelacji między wartościami badanych wskaźników a skorygowanym współczynnikiem Randa dla 8 zbiorów testowych

Skorygowany współczynnik Randa	Zbiór	*MQE*	*TE*	*DM*	*KLM*	*CEEF*log	*IAg*	*MN*
	1	–0,5122	–0,0631	–0,5247	–0,4587	–0,4425	0,3780	–0,5310
	2	–0,5784	–0,4886	–0,6848	–0,6071	–0,6590	0,6977	0,1867
	3	–0,5395	–0,2782	–0,4720	–0,6217	–0,5915	0,6056	0,4370
	4	–0,6141	–0,2237	–0,6653	–0,7304	–0,7445	0,7186	0,2928
	5	–0,2855	–0,1851	–0,3723	–0,3988	–0,4268	0,4578	–0,0790
	6	–0,3604	–0,3326	–0,4389	–0,3187	–0,3360	0,3877	0,0136
	7	–0,1977	–0,0859	–0,2098	–0,2170	–0,2878	0,1826	0,2557
	8	–0,7104	–0,1534	–0,7699	–0,7291	–0,7258	0,8007	0,4343

Źródło: opracowanie własne.

Szerszego wyjaśnienia wymagają współczynniki korelacji Pearsona dla udziału martwych neuronów w sieci. Dodatnie ich wartości wydają się zaskakujące. Oznaczałoby to, że im większy udział martwych neuronów w sieci, tym wyższa zgodność grupowania ze wzorcem. W przypadku rozmiarów badanych tu sieci i hybrydowej metody SOM-KS jest tak faktycznie. Wynika to ze znacznej komplikacji struktur grupowych w badanych zbiorach i tego, że metoda *k*-średnich wykorzystana na drugim stopniu nie pozwala na poprawne grupowanie badanych zbiorów (poza zbiorem pierwszym). Sieć potrzebuje znacznej liczby neuronów, w tym wielu martwych, które pełnią funkcję pośredników między neuronami aktywnymi, aby odwzorować istniejącą strukturę przestrzenną jednostek. Gdy jest ich już odpowiednio dużo, dalszy wzrost rozmiaru sieci nie powoduje wzrostu liczby martwych neuronów, a ich względny udział zaczyna maleć. Większość badanych sieci miała rozmiar zbyt mały w stosunku do stopnia komplikacji danych. Gdyby uwzględnić jedynie sieci o rozmiarze 10 × 10 do 16 × 16 wszystkie współczynniki korelacji miałyby ujemny, zgodny z oczekiwaniami, znak. Wydaje się, że jest to wartościowa wskazówka, pozwalająca ustalić właściwy, w stosunku do struktury przestrzennej danych, rozmiar sieci. Dokonując kolejnych symulacji, należy obserwować moment, w którym udział martwych neuronów w sieci przestaje rosnąć. Moment ten będzie wskazywał na osiągnięcie przez sieć koniecznych rozmiarów.

4. Wnioski

Wyniki prowadzonych badań wskazują na prawdziwość postawionej hipotezy badawczej. Wraz ze wzrostem jakości odwzorowania struktury badanych jednostek na sieci SOM zgodność grupowania ze wzorcem rośnie. Żaden ze wskaźników nie

może być jednak uznany za decydujący. Najwyższą przeciętną wartość współczynnika korelacji z skorygowanym współczynnikiem Randa dla badanych zbiorów uzyskano dla współczynnika *CEEF*log i wynosiła ona 0,5268. Wszystkie badane wskaźniki opisują na różne sposoby różne własności sieci SOM. Wszystkie powinny być brane pod uwagę w ocenie potencjału sieci w procesie grupowania danych.

Literatura

Davies D.L., Bouldin D.W. (1979), *A cluster separation measure*, „Pattern Analysis and Machine Intelligence", IEEE Transactions on, PAMI-1, 2, s. 224-227.

Herbst M., Casper M.C. (2008), *Towards model evaluation and identification using self-organizing maps*, „Hydrology and Earth System Science", 12, 2, s. 657-667.

Kaski S., Lagus K. (1996), *Comparing self-organizing maps*, Proceedings of the 1996 International Conference on Artificial Neural Networks, Springer-Verlag, Berlin, s. 809-814.

Kohonen T. (2001 [1997]), *Self-Organizing Maps*, Springer-Verlag, Berlin – Heidelberg.

Migdał-Najman K. (2012), *Propozycja hybrydowej metody grupowania opartej na sieciach samouczących*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 19, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 242, Wyd. UE, Wrocław, s. 342-351.

Migdał-Najman K., Najman K. (2013), *Samouczące się sztuczne sieci neuronowe w grupowaniu i klasyfikacji danych. Teoria i zastosowania w ekonomii*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

Nash J.E., Sutcliffe J.V. (1970), *River flow forecasting through conceptual models part I – A discussion of principles*, „Journal of Hydrology", 10, 3, s. 282-290.

Pölzlbauer G. (2004), *Survey and comparison of quality measures for self-organizing maps*, Proceedings of the Fifth Workshop on Data Analysis WDA'04, Elfa Academic Press, Slovakia, s. 67-82.

Rand W.M. (1971), *Objective criteria for the evaluation of clustering methods*, „Journal of the American Statistical Association", 66, 336, s. 846-850.

Spath H. (1985), *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples*, Halsted Press, New York.

Sun Y. (2000), *On quantization error of self-organizing map network*, „Neurocomputing", 34, 1-4, s. 169-193.

Willmott C.J. (1981), *On the validation of models*, „Physical Geography", 2, s. 184-194.

Willmott C.J. (1982), *Some comments on the evaluation of model performance*, „Bulletin of the American Meteorological Society", 63, 11, s. 1309-1313.

FORMAL QUALITY ASSESSMENT OF GROUP STRUCTURE MAPPING ON THE KOHONEN'S MAP

Summary: In the article the authors studied seven coefficients: mean quantization error, topographic error, distortion measure, Kaski-Lagus measure, logarithmized Nash-Sutcliffe coefficient of efficiency, Willmott's index of agreement and the number of "dead" neurons, to assess the quality of the mapping of the group structure on the Kohonen's map. In the experiment the authors used generated data sets with known spatial structure of units. The authors analyzed the relationship between the values of the coefficients and the structure of the test network. The authors analyzed the similarity between data clustering and pattern.

Keywords: Self Organizing Map (SOM), formal assessment of the quality mapping.