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

GRAFICZNA OCENA JAKOŚCI ODWZOROWANIA STRUKTURY GRUPOWEJ NA MAPIE KOHONENA

Streszczenie: W eksperymentach wykorzystano wygenerowane zbiory danych metrycznych o znanej strukturze przestrzennej jednostek. Dla każdego zbioru zbudowano sieci SOM (*Self Organizing Map*). Dla każdej sieci dokonano wizualizacji badanych wskaźników jakości odwzorowania i oceniono, która część mapy Kohonena odpowiada za powstałe błędy odwzorowania. W oparciu o tę ocenę wyróżniono jednostki i skupienia poprawnie i błędnie pogrupowane. Ostatecznie dokonano próby lokalnej oceny jakości grupowania, którą można przypisać każdej jednostce zbioru danych.

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

1. Wstęp

Stosując w analizie skupień samouczące się sztuczne sieci neuronowe typu SOM (*Self Organizing Map*), dąży się do budowy takiej sieci, która w najwyższym stopniu odwzorowuje strukturę przestrzenną badanych jednostek. Sieci takie charakteryzują się najwyższą zdolnością do wyróżniania skupień. Sam proces budowy i samouczenia się sieci tego typu nie jest deterministyczny. Zależy on od przyjętej topologii sieci, rozmiaru i liczby neuronów, typu połączeń neuronów, funkcji i zasięgu sąsiedztwa, losowej pozycji neuronów po inicjalizacji sieci i kolejności prezentacji analizowanych jednostek. Powtarzając wielokrotnie ten proces przy identycznych parametrach można uzyskać sieci znacząco różniące się swoimi własnościami. Z tego powodu konieczna jest szczegółowa ocena uzyskanego na sieci odwzorowania badanych jednostek.

W literaturze tematu proponuje się miary oceniające 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], zlogarytmowany współczynnik Nasha-Sutcliffe'a [Nash, Sutcliffe 1970] i indeks Willmotta [Willmott 1981,

1982]. Dodatkową informację o własnościach sieci można uzyskać, obserwując liczbę i udział martwych neuronów[1] w sieci.

Wszystkie powyższe miary spełniają swoje zadanie, a mimo to wydają się niewystarczające. Dostarczają syntetycznej informacji o zbudowanej sieci. Wszystkie są miarami średnimi i jako takie nie zawierają informacji o ważnych szczegółach budowy sieci. Na ich podstawie badacz nie dowie się, które jednostki są lepiej, a które gorzej odwzorowane. Którym neuronom można bardziej zaufać przy definiowaniu skupień, a którym mniej. Informacje tego typu byłyby bardzo przydatne przy ocenie uzyskanej w oparciu o sieć struktury skupień. Wydaje się, że informacje tego typu można uzyskać przez odpowiednią wizualizację uzyskanego w drodze samouczenia się sieci odwzorowania badanych jednostek. Celem prezentowanych badań jest weryfikacja tej hipotezy.

2. Wizualizacja struktury przestrzennej jednostek na mapie Kohonena

Wiele badań naukowych wskazuje, że łatwiej jest zrozumieć złożone zagadnienie, gdy uda się je odpowiednio zwizualizować. Człowiek wykazuje bardzo wysokie zdolności odbierania informacji za pomocą zmysłu wzroku. Z badań empirycznych wynika, że centralny układ nerwowy człowieka za pomocą zmysłu wzroku jest w stanie odebrać aż 87% informacji. Poziom odbieranej informacji przez zmysł słuchu stanowi jedynie 10%, a pozostałych zmysłów zaledwie 3% [Niemann, de Mori, Hanrieder 1994; Migdał-Najman, Najman 2013, s. 169]. Jest to zasadniczą przyczyną stosowania wielu technik wizualizacyjnych do opisu złożonych sztucznych sieci neuronowych. Do częściej stosowanych metod wizualizacji sieci SOM można zaliczyć:

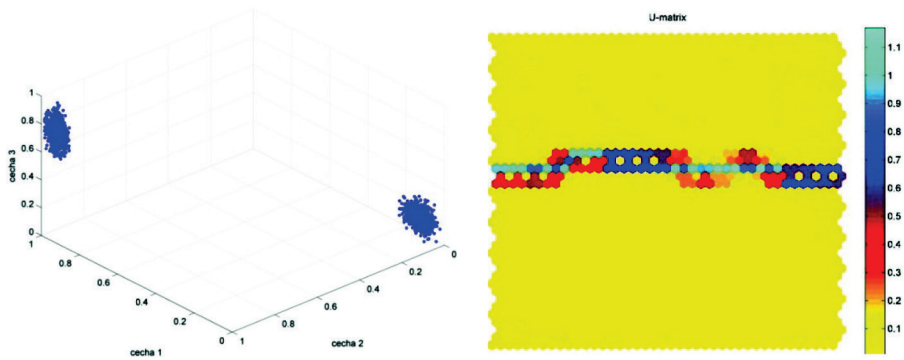
1. macierz ujednoliconych odległości, tzw. macierz U (*unified distance matrix*) i jej składowe,
2. histogram pobudzeń lub diagram pobudzeń (*hit histogram*, *diagram hit's*),
3. kolorowanie podobieństw (*similarity coloring*).

Najważniejszym rodzajem wizualizacji, z punktu widzenia prezentowanych tu badań, jest macierz ujednoliconych odległości [Kohonen 2001; Herbst, Casper 2008; Chattopadhyay, Dan, Mazumdar 2012; Migdał-Najman, Najman 2013]. Jest ona dwuwymiarową mapą, na której prezentowane są odległości (w sensie wybranej w procesie samouczenia się metryki) między neuronami sieci. Odległości te prezentowane są za pomocą skali barw. Uwzględnia się także metodę powiązania neuronów: wiązania kwadratowe lub heksagonalne. Gdy wiązania są kwadratowe, uzyskujemy mapę złożoną z kwadratów, z których każdy swoją barwą opisuje medianę odległości między danym neuronem a jego sąsiadami. Dla wiązań heksago-

[1] Neurony te pozwalają sieci lepiej rozciągnąć się w przestrzeni, jednak nie odpowiadają za odwzorowanie żadnej jednostki. Nie biorą więc udziału w procesie grupowania.

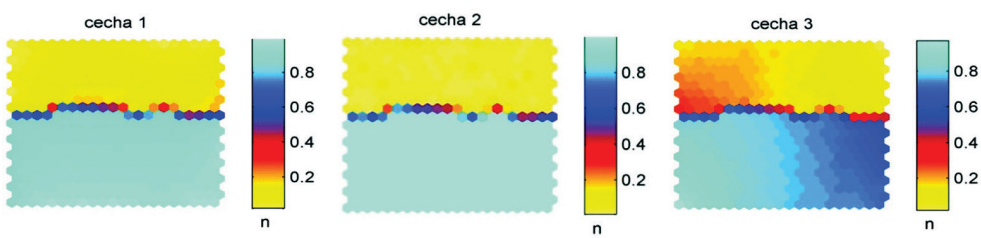
nalnych będą to sześcioboki. Na rysunku 1a przedstawiono przykładowy zbiór 200 jednostek opisanych trzema cechami zmiennymi, znajdujących się w dwóch dobrze separowanych skupieniach. Na rysunku 1b przedstawiono macierz ujednoliconych odległości dla sieci SOM o wymiarze 16 × 16 neuronów, heksagonalnych połączeniach neuronów i gaussowskiej funkcji sąsiedztwa o zasięgu 2, która odwzorowuje te jednostki. Kolor żółty (dół skali kolorów) wskazuje na małe odległości między neuronami, a jasnoniebieski (góra skali) na duże[2]. Na rysunku tym łatwo zauważyć dwa skupienia oddzielone linią dużych odległości między neuronami.

Można również zaprezentować udziały każdej cechy zmiennej (każdego wymiaru) w macierzy ujednoliconych odległości. W ten sposób można ocenić, czy w danym wymiarze skupienia są separowalne czy nie. Na rysunku 2 pokazano udziały poszczególnych cech w macierzy ujednoliconych odległości z rysunku 1b. Łatwo zauważyć, że we wszystkich wymiarach skupienia są separowalne, choć w trzecim nieco mniej niż w pozostałych.



Rys. 1. a) Przykładowy zbiór testowy. b) Macierz ujednoliconych odległości

Źródło: opracowanie własne.



Rys. 2. Udziały poszczególnych cech zmiennych w macierzy ujednoliconych odległości

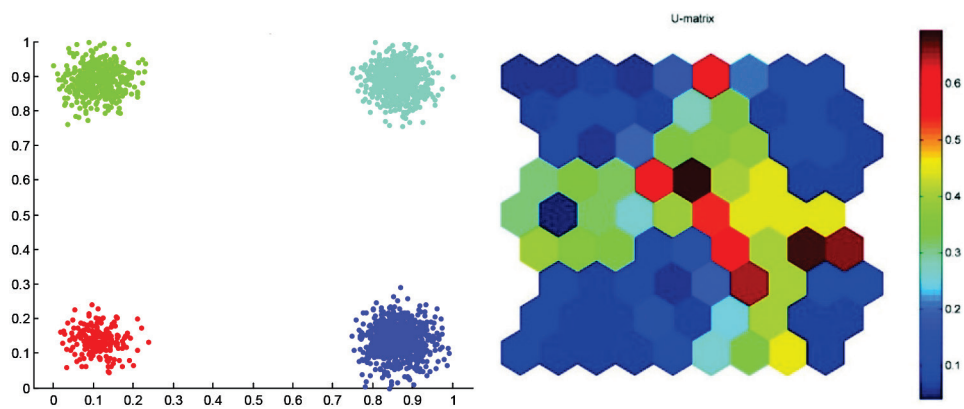
Źródło: opracowanie własne.

[2] Skala kolorów jest umowna i subiektywnie dobrana. Aby poprawnie porównywać różne odległości, zwykle skaluje się je na przedział [0,1].

Jeżeli odpowiednią barwą możliwe jest zaznaczenie na takiej mapie odległości każdego neuronu do każdego jego sąsiada, to możliwe jest także oznaczenie kolorem indywidualnych ocen jakości odwzorowania. Każdy neuron ma bowiem swój udział w średnim błędzie kwantyzacji, topograficznym czy dystorsji. Wizualizacja taka mogłaby być bardzo użyteczna, ponieważ pozwoliłaby zaobserwować, jak w przestrzeni rozkładają się błędy w odwzorowaniu badanych jednostek. Część jednostek jest zwykle lepiej odwzorowana niż pozostałe. To słabsze odwzorowanie może skutkować mniejszą zdolnością sieci do grupowania jednostek znajdujących się w gorzej odwzorowanej części przestrzeni.

3. Wizualizacja błędów odwzorowania na mapie Kohonena

Proces wizualizacji błędów odwzorowania zostanie przedstawiony na prostym przykładzie. Niech zbiór danych liczy 2000 jednostek, opisanych dwiema umownymi cechami zmiennymi (X i Y), które skupiają się w czterech sferycznych i separowalnych skupieniach (por. rys. 3a). Dla tego zbioru zbudowano sieć SOM o wymiarach 5 × 5 neuronów, z heksagonalną strukturą połączeń neuronów, gaussowską funkcją sąsiedztwa o zasięgu 2. Macierz ujednoliconych odległości dla tej sieci jest zaprezentowana na rysunku 3b.



Rys. 3. a) Zbiór testowy 2000 jednostek. b) Macierz ujednoliconych odległości o wymiarze 5 × 5 neuronów

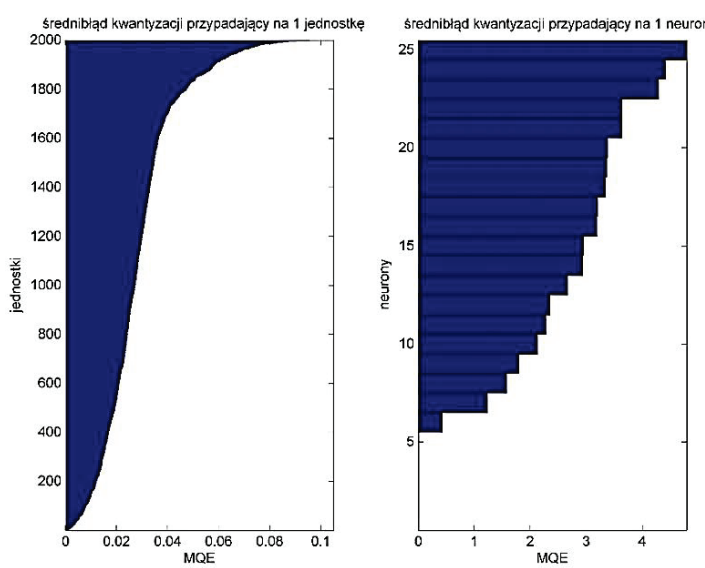
Źródło: opracowanie własne.

Wizualizacja pozwala zauważyć, że w zbiorze danych istnieją cztery skupienia. Jednak ich granice można wyznaczyć tylko w przybliżeniu. Wyznaczone wartości miar jakości odwzorowania jednostek na sieci kształtowały się następująco: średni błąd kwantyzacji MQE = 0,0034, błąd topograficzny TE = 0,1045, błąd dystorsji DM = 0,0005, współczynnik Kaskiego-Lagus KLM = 0,0094, współczynnik Nasha-

-Sutcliffe'a *CEEF*log = 0,0356, indeks Willmotta *IAg* = 0,9999. Liczba martwych neuronów jest równa 5, co stanowi 20% ogólnej liczby neuronów. Na rysunku 3b nie można zaobserwować czy są martwe neurony , a jeżeli tak, to gdzie się znajdują. Nie można także zaobserwować lokalnej jakości odwzorowania.

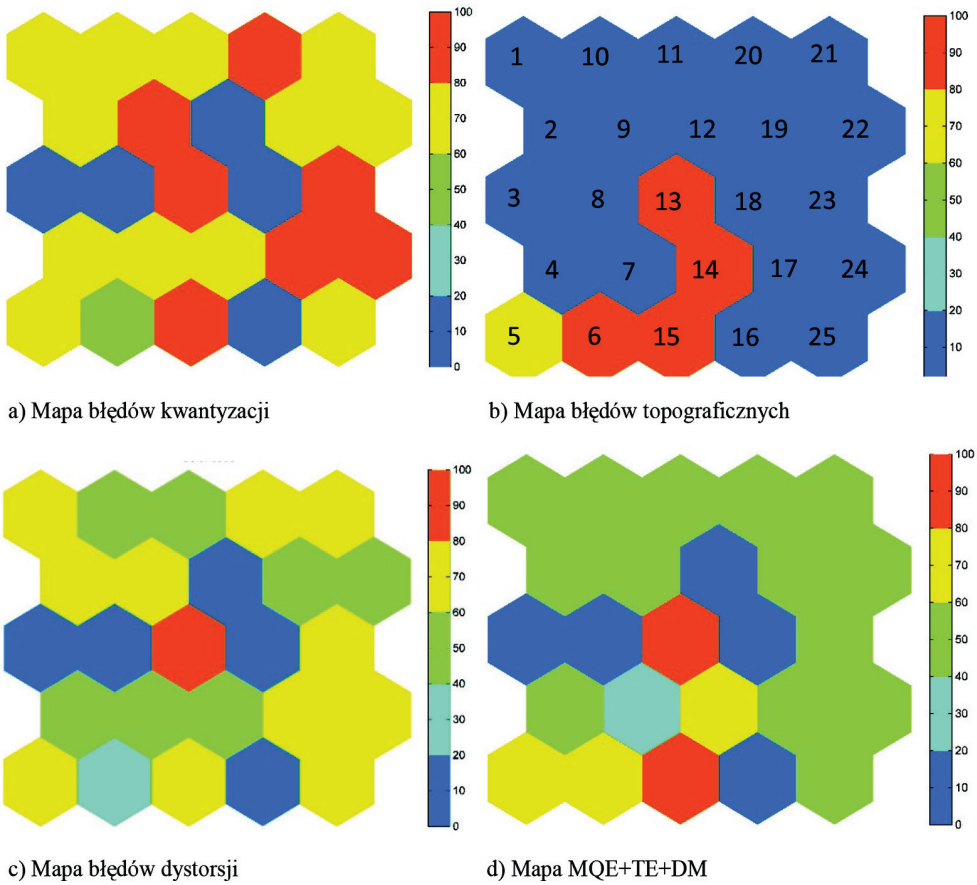
Wizualizacji, która pozwoli znaleźć odpowiedzi na te pytania, można dokonać w oparciu o: 1) mapy i diagramy błędów odwzorowania dla neuronów, 2) mapy i diagramy błędów odwzorowania dla jednostek. Diagramy błędów odwzorowania dla neuronów i jednostek można łatwo zbudować, ponieważ elementy, z których się składają, stanowią składowe sum miar jakości odwzorowania. W ten sposób można np. pokazać, jaki jest udział poszczególnych neuronów w średnim błędzie kwantyzacji. Analogicznie, identyfikując neuron odpowiedzialny za odwzorowanie danej jednostki, można pokazać błąd kwantyzacji związany z każdą z nich. Dla powyższego zbioru danych diagramy takie pokazano na rysunku 4.

Dzięki diagramowi błędów kwantyzacji dla jednostek wiemy, że ponad 1600 jednostek jest odwzorowanych na poziomie mniejszym niż 0,04. Błędy odwzorowania pozostałych 400 szybko rosną do poziomu 0,08. Rozkład błędów jest więc wyraźnie asymetryczny. Na diagramie błędów odwzorowania dla neuronów można zaobserwować, że nie ma neuronów o znacząco różnych wartościach błędów kwantyzacji. Ich rozkład jest względnie symetryczny. Można także zaobserwować pięć neuronów, które mają zerowy udział w średnim błędzie kwantyzacji. Są to martwe neurony.



Rys. 4. Średni błąd kwantyzacji przypadający na 1 jednostkę i średni błąd kwantyzacji przypadający na 1 neuron

Źródło: opracowanie własne.



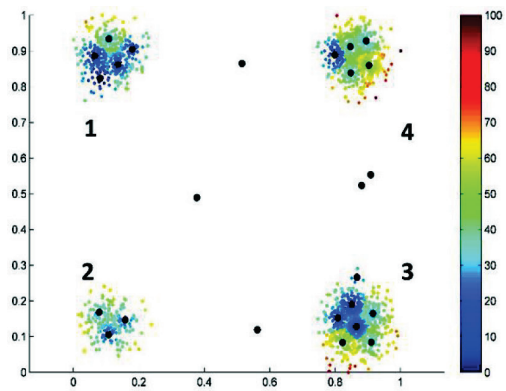
Rys. 5. Mapy błędów sieci Kohonena

Źródło: opracowanie własne.

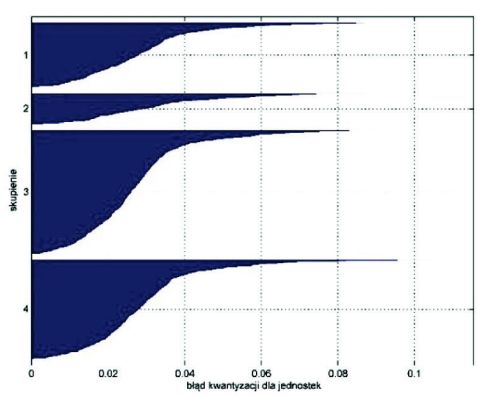
Aby zidentyfikować rozkład błędów odwzorowania między poszczególnymi neuronami, można dokonać ich wizualizacji, stosując mapy błędów. Ich konstrukcja jest podobna do wykresów macierzy ujednoliconych odległości. Zamiast median odległości między sąsiednimi neuronami na wykres naniesione zostają indywidualne błędy odwzorowania dla każdego neuronu. Dla większej czytelności wizualizacji błędy indywidualne można przeskalować na przedział (0;1) lub (0;100). Aby z kolei zidentyfikować obszary neuronów o podobnym poziomie błędów indywidualnych, przyjęty obszar zmienności można podzielić na pewną liczbę przedziałów. W badaniu przyjęto pięć przedziałów oznaczonych na skali obok map. Na rysunku 5a przedstawiono mapę błędów kwantyzacji, 5b – mapę błędów topograficznych, 5c – mapę błędów dystorsji przypadających na dany neuron. Sumy tych błędów przeskalowane na przedział (0;100) pokazano na rysunku 5d. Liczby od

1 do 25 przedstawione na rysunku 5b odpowiadają kolejnym 25 neuronom sieci SOM o wymiarze 5 × 5. Na rysunkach 5a, 5c i 5d można zauważyć pięć neuronów o bardzo niskich wartościach błędów odwzorowania (neurony: 3, 8, 12, 16 i 18). Są to jednak wyłącznie martwe neurony. Mają one zerowy poziom błędu kwantyzacji i dystorsji, ale nie topograficznego. Najmniejszy błąd kwantyzacji i dystorsji ma neuron 6, ale jego błąd topograficzny należy do największych. Ze względu na trzy wymienione miary jakości odwzorowania względnie najmniejszym błędem charakteryzuje się neuron 7, a największym neurony 15 i 13.

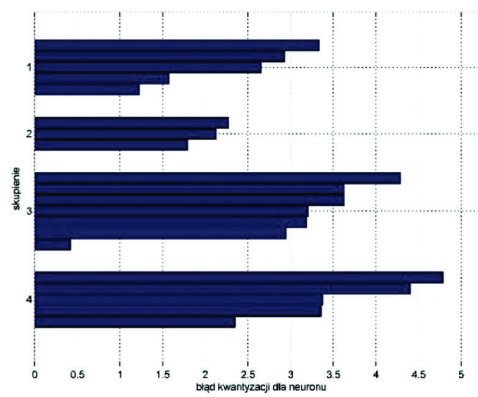
W dalszej kolejności możliwe jest zidentyfikowanie jednostek odwzorowanych przez poszczególne neurony. Przypisując każdej z nich sumę błędów odwzorowujących je neuronów, można dokonać kolejnej wizualizacji. Na rysunku 6a pokaza-



a) odwzorowywane jednostki przez neurony sieci SOM, kolor oznacza sumę błędów odwzorowania (kwantyzacji, topologiczny i dystorsji) przypadających na każdą z nich



b) błędy kwantyzacji jednostek w skupieniach



c) błędy kwantyzacji neuronów w skupieniach

Rys. 6. Odwzorowane jednostki i wyróżnione skupienia

Źródło: opracowanie własne.

no odwzorowywane jednostki, oznaczając kolorem sumę błędów odwzorowania (kwantyzacji, topologiczny i dystorsji) przypadających na każdą z nich. Na rysunku 6a czarnymi kropkami zaznaczono także pozycje neuronów sieci SOM. Łatwo zauważyć, że w sieci faktycznie znajduje się 5 martwych neuronów. Można także zaobserwować, że im dalej znajdują się jednostki od najbliższego neuronu, tym większy dotyczy ich błąd odwzorowania. Kolor pozwala obserwować jego natężenie.

Ostatecznie, dokonując grupowania neuronów sieci SOM, w tym wypadku metodą *k*-średnich [Spath 1985] z liczbą skupień ustaloną w oparciu o indeks Daviesa-Bouldina [Davies, Bouldin 1979], można dokonać wizualizacji poszczególnych błędów odwzorowania lub ich sum dla jednostek i neuronów według skupień (por. rysunek 6b i 6c).

W analizowanym przykładzie neurony znajdujące się w drugim skupieniu charakteryzują się najmniejszymi błędami kwantyzacji (por. rys. 6c), a znajdujące się w czwartym – największymi. Im słabsze indywidualne oceny jakości odwzorowania dla jednostek, tym mniejsze zaufanie do jakości ich grupowania.

4. Wnioski

Jakość odwzorowania jest zjawiskiem nie tylko globalnym dla całej sieci, ale także lokalnym – dla każdego neuronu i każdej jednostki. Znane analityczne miary jakości odwzorowania jednostek na mapie Kohonena pozwalają ocenić efekt samouczenia się sieci jako całości. Nie dostarczają jednak informacji o ich zróżnicowaniu, rozkładzie na sieci ani wśród jednostek. Z tego powodu obserwacja lokalnych błędów odwzorowania wydaje się ważna. Informacje takie można uzyskać i dokonać ich wizualizacji. Pozwala ona lepiej rozumieć własności uzyskanej sieci SOM, co może prowadzić do bardziej precyzyjnego grupowania i zwiększać zaufanie badacza do uzyskanych wyników. Bezpośredni wpływ tych ocen będzie stanowił podstawę dalszych badań.

Literatura

Chattopadhyay M., Dan P.K., Mazumdar S. (2012), *Application of visualclusteringproperties of self-organizing map in machine-part cellformation*, „Applied Soft Computing" 12, 2, s. 600-610.

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., 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 fore casting through conceptual models part I – A discussion of principles*, „Journal of Hydrology", 10, 3, s. 282-290.
Niemann H., de Mori R., Hanrieder G. (1994), *Progress and Prospects of Speech Research Technology*, Proceedings in Artificial Intelligence CRIM/FORWISS Workshop, München, September.
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.
Sun Y. (2000), *On quantization error of self-organizing map network*, „Neurocomputing", 34, 1-4, s. 169-193.
Spath H. (1985), *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples*, Halsted Press, New York.
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.

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

Summary: In the article the authors used generated data sets with known spatial structure of units. For each of the data set there was built the neural network type of SOM (*Self Organizing Map*). For each neural network there was made a visualization of coefficients of quality mapping. This approach allows to decide which part of Kohonen's map is responsible for mapping errors as well as to recognize those units and clusters which are correctly and incorrectly grouped. The authors conducted a local assessment of the quality of clustering for each unit.

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