PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 426

Taksonomia 26
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

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 stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

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

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp 9

Jacek Batóg: Identyfikacja obserwacji odstających w analizie skupień / Influence of outliers on results of cluster analysis 13

Andrzej Bąk: Porządkowanie liniowe obiektów metodą Hellwiga i TOPSIS – analiza porównawcza / Linear ordering of objects using Hellwig and TOPSIS methods – a comparative analysis..... 22

Grażyna Dehnel: *MM*-estymacja w badaniu średnich przedsiębiorstw w Polsce / *MM*-estimation in the medium-sized enterprises survey in Poland..... 32

Andrzej Dudek: *Social network analysis* jako gałąź wielowymiarowej analizy statystycznej / Social network analysis as a branch of multidimensional statistical analysis..... 42

Iwona Foryś: Analiza dyskryminacyjna w wyborze obiektów podobnych w procesie szacowania nieruchomości / The discriminant analysis in selection of similar objects in the real estate valuation process 51

Gregory Kersten, Ewa Roszkowska, Tomasz Wachowicz: Ocena zgodności porządkowej systemu oceny ofert negocjatora z informacją preferencyjną / Analyzing the ordinal concordance of preferential information and resulting scoring system in negotiations..... 60

Iwona Konarzewska: Rankingi wielokryteriowe a współzależność liniowa kryteriów / Multi-criteria rankings and linear relationships among criteria 69

Anna Król, Marta Targaszewska: Zastosowanie klasyfikacji do wyodrębniania homogenicznych grup dóbr w modelowaniu hedonicznym / The application of classification in distinguishing homogeneous groups of goods for hedonic modelling..... 80

Marek Lubicz: Problemy doboru zmiennych objaśniających w klasyfikacji danych medycznych / Feature selection and its impact on classifier effectiveness – case study for medical data..... 89

Aleksandra Łuczak: Wpływ różnych sposobów agregacji opinii ekspertów w FAHP na oceny priorytetowych czynników rozwoju / Influence of different methods of the expert judgments aggregation on assessment of priorities for evaluation of development factors in FAHP..... 99

Iwona Markowicz: Tablice trwania firm w województwie zachodniopomorskim według rodzaju działalności / Companies duration tables in Zachodniopomorskie voivodship by the type of activity 108

Małgorzata Markowska, Danuta Strahl: Filary inteligentnego rozwoju a wrażliwość unijnych regionów szczebla NUTS 2 na kryzys ekonomiczny – analiza wielowymiarowa / Smart development pillars and NUTS 2 European regions vulnerability to economic crisis – a multidimensional analysis 118
Kamila Migdał-Najman, Krzysztof Najman: Hierarchiczne deglomeracyjne sieci SOM w analizie skupień / The hierarchical divisive SOM in the cluster analysis 130
Kamila Migdał-Najman, Krzysztof Najman: Hierarchiczne aglomeracyjne sieci SOM w analizie skupień / The hierarchical agglomerative SOM in the cluster analysis 139
Barbara Pawełek, Józef Pociecha, Jadwiga Kostrzewska, Mateusz Baryła, Artur Lipieta: Problem wartości odstających w prognozowaniu zagrożenia upadłością przedsiębiorstw (na przykładzie przetwórstwa przemysłowego w Polsce) / Problem of outliers in corporate bankruptcy prediction (case of manufacturing companies in Poland) 148
Wojciech Roszka: Syntetyczne źródła danych w analizie przestrzennego zróżnicowania ubóstwa / Synthetic data sources in spatial poverty analysis..... 157
Małgorzata Rószkiewicz: Czynniki różnicujące efektywność pracy ankietera w wywiadach *face-to-face* w środowisku polskich gospodarstw domowych / Factors affecting the efficiency of face-to-face interviews with Polish households 166
Adam Sagan, Marcin Pełka: Analiza wielopoziomowa z wykorzystaniem danych symbolicznych / Multilevel analysis with application of symbolic data 174
Marcin Salamaga: Zastosowanie drzew dyskryminacyjnych w identyfikacji czynników wspomagających wybór kraju alokacji bezpośrednich inwestycji zagranicznych na przykładzie polskich firm / The use of classification trees in the identification of factors supporting the choice of FDI destination on the example of Polish companies..... 185
Agnieszka Stanimir: Pomiar wykluczenia cyfrowego – zagrożenia dla Pokolenia Y / Measurement of the digital divide – risks for Generation Y ... 194
Mirosława Sztemberg-Lewandowska: Grupowanie danych funkcjonalnych w analizie poziomu wiedzy maturzystów / Functional data clustering methods in the analysis of high school graduates' knowledge 206
Tadeusz Trzaskalik: Modelowanie preferencji w wielokryterialnych dyskretnych problemach decyzyjnych – przegląd bibliografii / Preference modeling in multi-criteria discrete decision making problems – review of literature 214

Joanna Trzęsiok: Metody nieparametryczne w badaniu zaufania do instytucji finansowych / Nonparametric methods in the study of confidence in financial institutions 226
Hanna Wdowicka: Analiza sytuacji na lokalnych rynkach pracy w Polsce / Local labour market analysis in Poland..... 235
Artur Zaborski: Zastosowanie skalowania dynamicznego oraz metody wektorów dryfu do badania zmian w preferencjach / The use of dynamic scaling and the drift vector method for studying changes in the preferences..... 245

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia" nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii" nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pociecha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pociecha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia" nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym" (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pociecha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes".

W konferencji „European Conference on Data Analysis" (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKl) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies", prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II" oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance", prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice", prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II".

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015", Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis" odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017". Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019" organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak", 1 głos na „nie", 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 426 • 2016

Taksonomia 26 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Kamila Migdał-Najman, Krzysztof Najman
Uniwersytet Gdański
e-mails: {kamila.migdal-najman; krzysztof.najman}@ug.edu.pl

HIERARCHICZNE AGLOMERACYJNE SIECI SOM W ANALIZIE SKUPIEŃ

THE HIERARCHICAL AGGLOMERATIVE SOM IN THE CLUSTER ANALYSIS

DOI: 10.15611/pn.2016.426.14

Streszczenie: Samouczące się sztuczne sieci neuronowe typu SOM należą do jednych z bardziej efektywnych narzędzi *data mining*, które są stosowane w grupowaniu i klasyfikacji danych wielowymiarowych. Spadek efektywności sieci SOM w grupowaniu i klasyfikacji danych często wynika z przyjętej nadmiarowej struktury sieci i znacznego przyrostu martwych neuronów w sieci. Proces samouczenia takiej sieci staje się niepotrzebnie długi. Jedną z możliwości rozwiązań tego problemu jest budowa hierarchicznych aglomeracyjnych sieci SOM (*Hierarchical agglomerative* SOM, H_aSOM). W sieciach tych wyróżnia się dwa podejścia: tematyczne i oparte na skupieniach. Celem prezentowanych badań jest analiza własności aglomeracyjnych sieci H_aSOM w analizie skupień danych o hierarchicznej strukturze.

Słowa kluczowe: analiza skupień, nienadzorowane sieci neuronowe, sieć SOM, aglomeracyjne sieci H_aSOM.

Summary: Self-learning artificial neural networks type of SOM are one of the most effective data mining tools which are used in grouping and classification of multidimensional data. The decrease in network efficiency SOM clustering and classification of data often results from the assumed redundant network structure and a significant increase of dead neurons in the network. The process of self-learning of the network becomes unnecessarily long. One possibility of solving this problem is to build a hierarchical agglomerative SOM network. In these networks, there are two approaches: thematic and based on clusters. The aim of this paper is to analyze the properties of agglomerative H_aSOM network in the cluster analysis.

Keywords: cluster analysis, unsupervised neural networks, SOM, agglomerative hierarchical SOM (H_aSOM).

1. Wstęp

Jedną z bardziej efektywnych metod analizy skupień są nienadzorowane sieci neuronowe samoorganizujące się (*Self Organizing Map*, SOM) [Kohonen 1995]. Do ich najważniejszych zalet należą: ich nieparametryczność, niewrażliwość na występowanie wartości skrajnych i szumu, odporność na braki danych, a także brak apriorycznej konieczności ustalenia dokładnej struktury sieci [Migdał-Najman, Najman 2013]. Wadą, szczególnie uciążliwą w przypadku analizy zbioru danych o wysokim wymiarze, jest konieczność budowy sieci o znacznych rozmiarach, czego konsekwencją jest długi czas uczenia się sieci i spadająca wraz ze wzrostem rozmiaru sieci efektywność grupowania. Jednym z powodów spadku efektywności jest to, że w dużej sieci SOM wiele neuronów nie bierze udziału w rozpoznawaniu obiektów (tzw. efekt martwych neuronów). W konsekwencji struktura sieci staje się nadmiarowa, a proces samouczenia się staje się niepotrzebnie długi.

Inną konsekwencją budowy dużych sieci SOM jest utrata zdolności do obserwacji struktur danych na różnych poziomach ogólności. Analogicznie jak w badaniach reprezentacyjnych użytkownik potrzebuje czasami ocen parametrów dla warstw i ocen globalnych. Sieć o setkach neuronów może ukazywać jednak tak dużą liczbę lokalnych zależności przestrzennych, że trudno z nich wyłowić związki ogólniejsze, które mogą być dla użytkownika równie istotne.

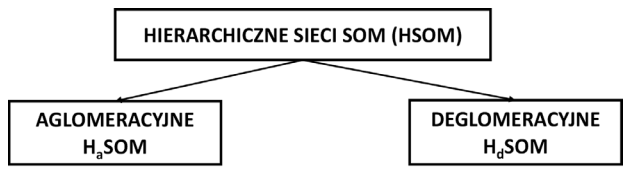
Jednym z możliwych rozwiązań powyższych problemów jest budowa hierarchicznych aglomeracyjnych sieci SOM (*Hierarchical agglomerative* SOM, H_aSOM). Celem prezentowanych badań jest analiza własności aglomeracyjnych sieci H_aSOM w analizie skupień danych o hierarchicznej strukturze.

2. Hierarchiczne aglomeracyjne sieci SOM

Pierwsze próby budowy hierarchicznych sieci SOM podjął w 1989 r. S.P. Luttrell [Luttrell 1989]. Przedstawił wyniki badań nad podejściem hierarchicznym w wektorowej kwantyzacji. W pracy tej autor podkreślił różnicę między podejściem standardowym a hierarchicznym. Co ważniejsze, wykazał, że stosowanie sieci SOM na kolejnych poziomach aglomeracji tylko w minimalny sposób zniekształca dane wejściowe. Do podobnych wniosków doszli P. Koikkalainen, J. Lampinen i E. Oja, którzy analizowali hierarchiczne sieci SOM jako narzędzie grupowania [Koikkalainen, Oja 1990; Lampinen, Oja 1992]. W 1990 r. P. Koikkalainen i E. Oja na międzynarodowej konferencji poświęconej sztucznym sieciom neuronowym (International Joint Conference on Neural Networks, IJCNN) w Waszyngtonie zaprezentowali samoorganizującą się hierarchiczną mapę cech. Podobne badania prowadziła Kamila Migdał-Najman, prezentując wysoką zgodność wyników klasycznych metod aglomeracyjnych opartych na macierzy odległości z wynikami uzyskanymi dzięki sieci SOM, przy jednoczesnej redukcji ilości koniecznych obliczeń i wymagań co do zasobów

komputera [Migdał-Najman 2007]. Sieć SOM jest obecnie jednym z bardziej popularnych i efektywnych narzędzi *data mining*, które znajdują zastosowanie w zagadnieniach klasyfikacji [Corridoni, Bimbo, Landi 1996; Papadimitriou i in. 2002; Ye, Lo 2000] i grupowania [Changchien, Lu 2001; Deboeck 1999; Gómez-Carracedo i in. 2010; Ha, Park 1998; Hui, Jha 2000; Kiang, Hu, Fisher 2006; Kruk i in. 2007; Migdał-Najman, Najman 2003; Vesanto, Alhoniemi 2000].

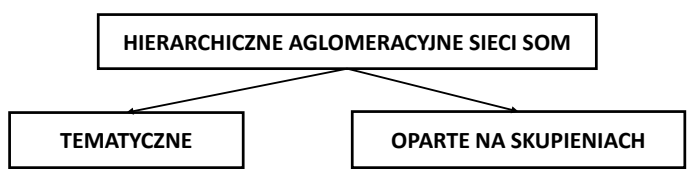
Obecnie, mówiąc o hierarchicznych sieciach SOM, mamy na myśli całą rodzinę różnych podejść do budowy sieci SOM. Wyróżnić należy podejście aglomeracyjne i deglomeracyjne (por. rys. 1).



Rys. 1. Klasyfikacja hierarchicznych sieci SOM

Źródło: opracowanie własne.

W **podejściu aglomeracyjnym** na pierwszym poziomie uczenia budowanych jest kilka sieci SOM, każda dla subdomen lub grup zmiennych. Na kolejnym poziomie uzyskane wyniki sieci z poziomu pierwszego łączy się i buduje nową sieć/sieci na poziomie wyższym agregując informacje uzyskane na niższym poziomie. W **podejściu deglomeracyjnym**, na poziomie pierwszym wychodzi się zazwyczaj od jednej dużej sieci SOM, a następnie na kolejnych poziomach rozbija się ją na części i dla każdej z nich buduje się kolejne sieci SOM. Budując sieci H_aSOM można wyróżnić dwa podejścia: tematyczne (*thematic agglomerative* HSOM) i oparte na skupieniach (*based on clusters* HSOM) (por. rys. 2).



Rys. 2. Klasyfikacja hierarchicznych aglomeracyjnych sieci SOM

Źródło: opracowanie własne.

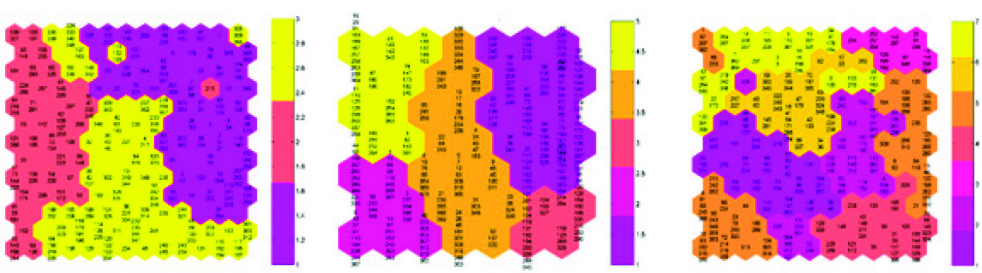
W **podejściu tematycznym** na pierwszym poziomie buduje się serię sieci SOM dla domen lub strukturalnych części zbioru jednostek (np. cech ekonomicznych, społecznych, demograficznych, itp.) a następnie, na drugim poziomie, kolejną sieć na wykrytych mikroskupieniach. Podejście to wymaga, aby domeny czy

grupy zmiennych były zdefiniowane przez badacza *a priori*. Jeżeli takiej wiedzy badacz nie posiada, może zdać się na samą sieć. W **podejściu opartym na skupieniach** na pierwszym poziomie buduje się sieć SOM dla wszystkich posiadanych jednostek. Sieć ta może mieć znaczne rozmiary. Na drugim poziomie neurony sieci z poziomu pierwszego są przekazywane, jako dane wejściowe do sieci na poziomie drugim. Sieć ta jest zazwyczaj mniejsza od sieci budowanej na poziomie pierwszym. Podejście to pozwala na obserwacje klas na różnych poziomach szczegółowości. Sieć na poziomie pierwszym zachowuje informacje szczegółowe, w dużej części o znaczeniu lokalnym. Sieć poziomu drugiego przekazuje z kolei informacje zagregowane o znacznie wyższym poziomie ogólności. W zależności od badanego problemu i posiadanych danych sieć taka może posiadać dwa lub więcej poziomów.

3. Konstrukcja tematycznych sieci H_aSOM

Konstrukcja tematycznej sieci H_aSOM zostanie zaprezentowana na przykładzie badania powiatów w Polsce [Migdał-Najman, Najman 2003].[1] Niech zbiór danych stanowi 307 powiatów (bez miast na prawach powiatów) opisanych 43 cechami, odpowiednio 9 demograficznymi, 14 ekonomicznymi i 20 ekologicznymi. Załóżmy, że w badaniu ważne jest, aby zaobserwować zróżnicowanie i strukturę grupową powiatów w Polsce dla każdej z trzech grup badanych cech, a także strukturę ogólną wynikającą ze wszystkich badanych cech. Na pierwszym poziomie należy zbudować niezależną sieć SOM dla wszystkich badanych jednostek opisanych daną grupą cech. W opisanym przypadku będą to trzy niezależne sieci SOM.

W wyniku zastosowania odpowiednich procedur budowy sieci SOM i oceny wyróżnionej struktury grupowej [Migdał-Najman 2007, 2008; Migdał-Najman, Najman 2013] uzyskano sieci zaprezentowane na rys. 3. Analiza sieci pierwszego



a) sieć 1, rozmiar 13 × 13
9 cech demograficznych

b) sieć 2, rozmiar 7 × 7
14 cech ekonomicznych

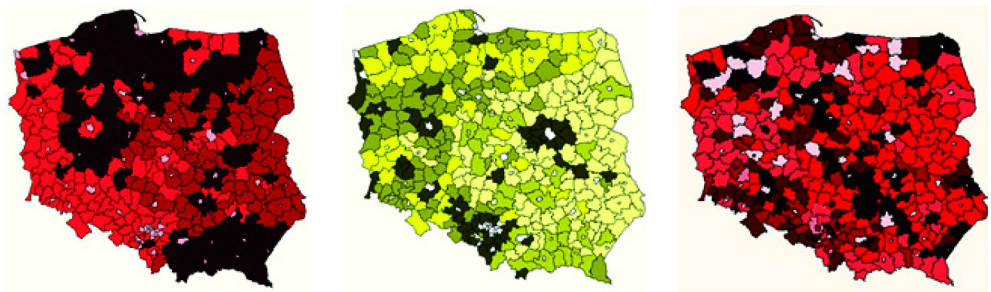
c) sieć 3, rozmiar 11 × 11
20 cech ekologicznych

Rys. 3. Trzy niezależne tematyczne sieci SOM

Źródło: opracowanie własne.

[1] Prezentowane badanie jest ilustracją procedury budowy sieci. Z tego powodu pominięto szczegółowe rozważania nad badanymi cechami i konstrukcją każdej z prezentowanych sieci.

poziomu pozwoliła wyróżnić trzy skupienia powiatów ze względu na cechy demograficzne (por. rys. 3a i 4a), pięć skupień ze względu na cechy ekonomiczne (por. rys. 3b i 4b) i siedem skupień ze względu na cechy ekologiczne (por. rys. 3c i 4c). Dekodując informacje z sieci, można dokonać wizualizacji uzyskanych struktur na mapie Polski (por. rys. 4).

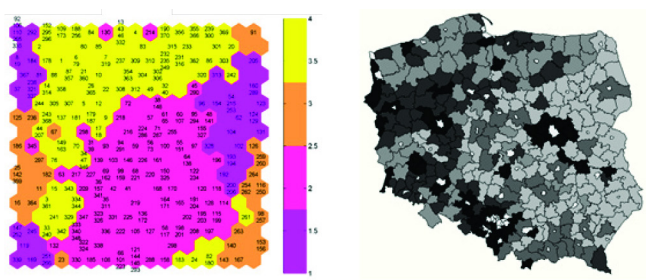


a) 3 skupienia cechy demograficzne | b) 5 skupień cechy ekonomiczne | c) 7 skupień cechy ekologiczne

Rys. 4. Podział powiatów ze względu na cechy demograficzne, ekonomiczne i ekologiczne

Źródło: opracowanie własne.

Na drugim poziomie budowy sieci każdy powiat jest odwzorowywany przez trzy neurony – po jednym z każdej z sieci. Struktura przestrzenna tych neuronów (13 + 7 + 11 neuronów) stanowi dane wejściowe do budowy sieci wyższego poziomu. W wyniku takiej agregacji zbudowano sieć o rozmiarze 17 × 17 neuronów, która wskazała na istnienie czterech ogólnych grup powiatów w Polsce (zob. rys. 5a). Po odkodowaniu uzyskanej struktury grupowej można ją pokazać na mapie powiatów Polski (zob. rys. 5b).



a) sieć 4, rozmiar 17 × 17 | b) mapa powiatów, 43 cechy

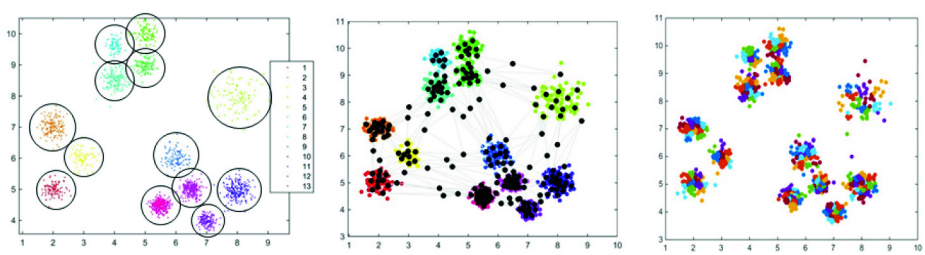
Rys. 5. Hierarchiczna aglomeracyjna tematyczna sieć SOM

Źródło: opracowanie własne.

Zalety sieci H_aSOM wydają się tu dość oczywiste. Podejście takie pozwala na obserwację struktur przestrzennych niezależnie dla domen i dla całego zbioru cech. Każda z domen może być analizowana niezależnie, można dokonać profilowania klas, oceny ich struktury, podobieństwa wewnętrznego czy zróżnicowania zewnętrznego. Jednocześnie po zagregowaniu uzyskujemy ogólniejszą strukturę, obejmującą wszystkie badane cechy. Co istotne agregacja następowała nie po jednostkach, a po neuronach sieci pierwszego stopnia, co pozwala zachować strukturę mikroskupień i maksymalnie uogólnić informacje zawarte w danych.

4. Konstrukcja sieci H_aSOM opartej na skupieniach

Gdy badacz nie może wyróżnić *a priori* domen czy subpopulacji, a nadal zależy mu na wielopoziomowej analizie, może zbudować aglomeracyjną, hierarchiczną sieć SOM opartą na skupieniach. Badacz zdaje się tu na samą sieć i jej zdolność do wyróżniania skupień. Budowa takiej sieci jest przynajmniej dwupoziomowa. Niech ilustracją tego procesu będzie analiza zbioru złożonego z 2000 jednostek, o strukturze przestrzennej pokazanej na rys. 6a. Na pierwszym poziomie budowana jest sieć o znacznym rozmiarze w celu uchwycenia możliwie dużej liczby szczegółów badanej struktury. Na rysunku 6b pokazano neurony sieci o rozmiarze 15×15 neuronów i heksagonalnej strukturze ich połączeń, naniesione na zbiór danych. Każdy neuron odpowiada za pewną liczbę jednostek, które są do siebie najbardziej podobne. Uzyskuje się w ten sposób pierwszy wgląd w strukturę badanych jednostek. Jest to spojrzenie bardzo precyzyjne i szczegółowe (por. rys. 6c).

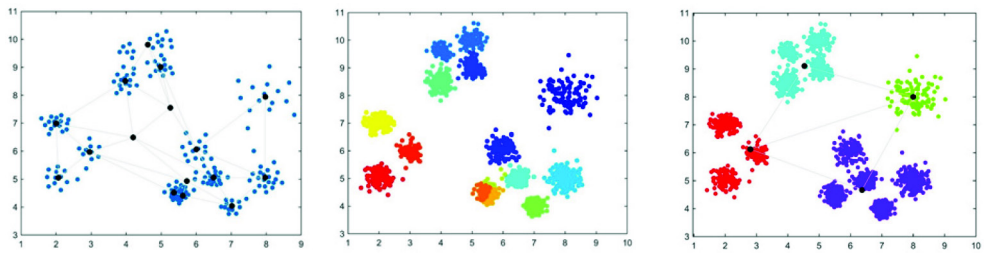


a) b) c)

Rys. 6. Hierarchiczna aglomeracyjna sieć SOM oparta na skupieniach – I poziom

Źródło: opracowanie własne.

Na drugim poziomie budowy sieci H_aSOM jednostkami badania stają się neurony sieci z pierwszego poziomu. W ten sposób pierwotne dane zostają zastąpione abstrakcyjnymi jednostkami, neuronami, które reprezentują uzyskane na pierwszym poziomie informacje. Na rysunku 7a pokazano neurony sieci SOM o rozmiarze 4 × 4 drugiego poziomu naniesione na neurony poziomu pierwszego (po usu-



a) b) c)

Rys. 7. Hierarchiczna aglomeracyjna sieć SOM oparta na skupieniach – II i III poziom

Źródło: opracowanie własne.

nięciu martwych neuronów). Po przypisaniu do neuronów drugiego poziomu wyjściowych jednostek ze zbioru danych uzyskujemy strukturę skupień pokazaną na rys. 7b. Informacje uzyskane po drugim poziomie są już bardziej ogólne, opisują struktury szersze, złożone z setek jednostek. Możliwe są także dalsze kroki aglomeracji. Do poziomu trzeciego, jako dane wejściowe należy wziąć neurony uzyskane na drugim poziomie. W ten sposób uzyskana zostanie struktura na największym poziomie ogólności. Wynik grupowania po trzecim poziomie dla sieci SOM o rozmiarze 2 × 2 przedstawiono na rys. 7c).

5. Zakończenie

Jak pokazano na powyższych przykładach, hierarchiczne, aglomeracyjne sieci SOM mogą być interesującym narzędziem analizy skupień. W przeciwieństwie do klasycznych sieci SOM pokazują hierarchię podobieństwa jednostek i skupień. Pozwala to zwykle na pogłębioną analizę badanego zbioru danych, uwzględniającą różne poziomy agregacji. W analizie wykorzystuje się całą serię sieci SOM, jednak są to zwykle sieci o niewielkim rozmiarze i znacznie mniejsze niż pojedyncza sieć wykorzystywana w podejściu niehierarchicznym. Ponieważ szybkość uczenia sieci SOM zależy w decydującym stopniu od liczby neuronów, a ta dla sieci kwadratowych i heksagonalnych rośnie w kwadracie rozmiaru sieci, proces uczenia serii niewielkich sieci jest krótszy niż jednej dużej. Znacząco spadają także wymagania co do zasobów komputera, które przy budowie dużej sieci łatwo przekraczają możliwości typowych komputerów. Budując serię niewielkich sieci, każdorazowo potrzebny jest tylko niewielki, w porównaniu do jednej dużej sieci, ich ułamek. Inne własności sieci H_aSOM wynikają wprost z własności podstawowej sieci SOM. Sieć taka będzie wykazywała zdolność do analizy skupień separowalnych, częściowo także zaszumionych i nieseparowalnych [Migdał-Najman, Najman 2013, s. 219]. Sieci H_aSOM stają się więc interesującym narzędziem eksploracyjnej analizy danych. W tabeli 1 zaprezentowano podstawowe własności sieci H_aSOM.

Tabela 1. Własności sieci H_aSOM

Własności	H_aSOM
Struktura sieci	zmienna/zależna od modelu
Liczba krytycznych parametrów sterujących	6+/zależna od modelu
Szybkość uczenia	szybsza niż SOM/zależna od modelu
Wymagania pojemności pamięci	mniejsze niż SOM/zależne od modelu
Dowolna konfiguracja skupień	tak
Martwe neurony	znacznie mniej niż w SOM
Rozmycie skupień	dopuszczalne (małe błędy)
Skupienia nieseparowalne	dopuszczalne (małe błędy)
Wizualizacja danych wielowymiarowych	tak
Wizualizacja sieci	tak
Eksploracja danych	bardzo wysoka
Hierarchia podobieństwa jednostek / cech	tak

Źródło: opracowanie własne.

Niestety poza wskazanymi pozytywnymi cechami sieci tego typu pojawiają się problemy, trudne do obiektywnego rozwiązania. Budując sieć SOM, trzeba zdefiniować szereg parametrów konstrukcji sieci i procesu samouczenia się. Są to przynajmniej: rozmiar sieci, struktura powiązań neuronów, funkcja i zasięg sąsiedztwa. Jak wynika z badań teoretycznych i empirycznych [Migdał-Najman, Najman 2013] obiektywne i optymalne ustalenie tych parametrów dla danego problemu może nie być możliwe. W sieciach hierarchicznych, gdy budowanych jest wiele sieci, parametry te trzeba ustalać wielokrotnie, niezależnie dla każdej z sieci, co dodatkowo utrudnia prawidłowe wnioskowanie. Dodatkowo własności zbudowanej sieci muszą zostać ocenione przynajmniej z punktu widzenia błędów kwantyzacji, topograficznego i dystorsji. W sieciach H_aSOM, gdy podsieci jest wiele, ocenie podlegać musi każda z nich. Jednocześnie brakuje ogólnej, całościowej oceny.

Literatura

Changchien S.W., Lu T.C., 2001, *Mining association rules procedure to support on-line recommendation by customers and products fragmentation*, Expert Systems with Applications, vol. 20, no. 4, s. 325–335.

Corridoni J.M., Bimbo A., Landi L., 1996, *3D object classification using multi-object Kohonen networks*, Pattern Recognition, vol. 29, no. 6, s. 919–935.

Deboeck G.J., 1999, *Value maps: Finding value in markets that are expensive*, [w:] E. Oja, S. Kaski (red.), *Kohonen Maps*, Elsevier Science, Amsterdam, s. 15–32.

Gómez-Carracedo M.P., Andrade J.M., Carrera G.V.S.M., Aires-de-Sousa J., Carlosena A., Prada D., 2010, *Combining Kohonen neural networks and variable selection by classification trees to cluster road soil samples*, Chemometrics and Intelligent Laboratory Systems, vol. 102, no. 1, s. 20–34.

Ha S., Park S., 1998, *Application of data mining tools to hotel data mart on the Internet for database marketing*, Expert Systems with Applications, vol. 15, no. 1, s. 1–31.

Hui S.C., Jha G., 2000, *Data mining for customer service support*, Information & Management, vol. 38, no. 1, s. 1–13.

Kiang M.Y., Hu M.Y., Fisher D.M., 2006, *An extended self-organizing map network for market segmentation – a telecommunication example*, Decision Support Systems, vol. 42, no. 1, s. 36–47.

Kohonen T., 1995, *Self-organizing Maps*, Springer, Berlin.

Koikkalainen P., Oja E., 1990, *Self-organizing hierarchical feature maps*, [w:] *Proceedings of the International Joint Conference on Neural Networks (IJCNN'90)*, Washington, DC, vol. 2, s. 279–284.

Kruk A., Lek S., Park Y.S., Penczak T., 2007, *Fish assemblages in the large lowland Narew River system (Poland): Application of the self-organizing map algorithm*, Ecological Modelling, vol. 203, no. 1-2, s. 45–61.

Lampinen J., Oja E., 1992, *Clustering properties of hierarchical self-organizing maps*, Journal of Mathematical Imaging and Vision, vol. 2, no. 2-3, s. 261–272.

Luttrell S.P., 1989, *Hierarchical vector quantisation*, Communications, Speech and Vision, IEE Proceedings I, vol. 136, no. 6, s. 405–413.

Migdał-Najman K., 2007, *Propozycja hybrydowej metody grupowania dużych zbiorów danych wykorzystującej sieć Kohonena i taksonomiczne metody grupowania*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, nr 1169, Taksonomia 14: *Klasyfikacja i analiza danych – teoria i zastosowania*, s. 305–313.

Migdał-Najman K., 2008, *Analiza porównawcza własności nienadzorowanych sieci neuronowych typu Self Organizing Map i Growing Neural Gas w analizie skupień*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 47, Taksonomia 16: *Klasyfikacja i analiza danych – teoria i zastosowania*, s. 205–213.

Migdał-Najman K., Najman K., 2003, *Zastosowanie sieci neuronowej typu SOM w badaniu przestrzennego zróżnicowania powiatów*, Wiadomości Statystyczne, nr 4, s. 72–85.

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.

Papadimitriou S., Mavroudi S., Vladutu L., Pavlides G., Bezerianos A., 2002, *The supervised network self-organizing map for classification of large data sets*, Applied Intelligence, vol. 16, no. 3, s. 185–203.

Vesanto J., Alhoniemi E., 2000, *Clustering of the self-organizing map*, IEEE Transactions on Neural Networks, vol. 11, no. 3, s. 586–600.

Ye H., Lo B.W.N., 2000, *A visualised software library: Nested self-organizing maps for retrieving and browsing reusable software assets*, Neural Computing and Applications, vol. 9, no. 4, s. 266–279.