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

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 DEGLOMERACYJNE SIECI SOM W ANALIZIE SKUPIEŃ

THE HIERARCHICAL DIVISIVE SOM IN THE CLUSTER ANALYSIS

DOI: 10.15611/pn.2016.426.13

Streszczenie: W badaniach empirycznych może pojawić się problem hierarchicznej struktury obserwowanych jednostek i skupień. Jednym z możliwych rozwiązań tego problemu jest budowa hierarchicznych, deglomeracyjnych sieci SOM (*Hierarchical divisive* SOM, HdSOM). Można tu wyróżnić dwa podejścia: statyczne (*static divisive* HSOM) i dynamiczne (*dynamic divisive* HSOM). Konstrukcja takich sieci jest hierarchiczna, gdyż fragmenty sieci jednej warstwy stają się zarodkiem sieci w kolejnej warstwie. W konsekwencji taka sieć może uczyć się znacznie szybciej, zredukowana zostanie liczba martwych neuronów i możliwe będzie znacznie bardziej szczegółowe rozpoznanie struktury grupowej. Celem prezentowanych badań jest analiza własności deglomeracyjnych sieci HdSOM w analizie skupień.

Słowa kluczowe: analiza skupień, nienadzorowane sieci neuronowe, sieć SOM, deglomeracyjne sieci HdSOM.

Summary: In the empirical studies hierarchical structure of units and clusters can be a problem. One solution to this problem is to build a hierarchical divisive SOM network. Two approaches can be distinguished: static and dynamic. The construction of such a network is hierarchical, as parts of the network layer of one network are the nucleus in the next layer. As a result, the network can learn much faster, the number of dead neurons will be reduced and much more detailed identification of the group structure will be possible. The aim of this paper is to analyse the properties of divisive HdSOM network in the cluster analysis.

Keywords: cluster analysis, unsupervised neural networks, SOM, divisive hierarchical SOM (HdSOM).

1. Wstęp

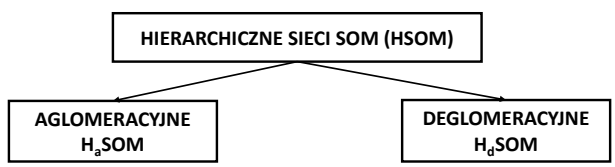
W wielu badaniach empirycznych badacz staje przed problemem hierarchicznej struktury obserwowanych jednostek i skupień. Niejednokrotnie pierwotnie wyróżnione skupienia mają dalszą, głębszą strukturę grupową, jednak na innym poziomie

agregacji. Przykładem mogą być badania przestrzenne, które ujawniają makrostruktury na poziomie państw, jeszcze inne na poziomie regionów, a mikrostruktury na poziomie powiatów. Badacz może być, z różnych powodów, zainteresowany zarówno makro- jak i mikrostrukturami. Aby je wykryć i obserwować, należy zastosować specjalne metody analizy skupień. Jednym z możliwych rozwiązań tego problemu jest budowa hierarchicznych, deglomeracyjnych sieci SOM (*Hierarchical divisive* SOM, H_dSOM). Celem prezentowanych rozważań jest analiza własności tego typu sieci w analizie zbiorów danych o hierarchicznej strukturze skupień.

2. Hierarchiczne deglomeracyjne sieci SOM

Mówiąc o hierarchicznych sieciach SOM, mamy na myśli całą rodzinę sieci o różnych konstrukcjach. Badania nad ich budową i metodami uczenia prowadzili m.in.: S.P. Luttrell [1989], P. Koikkalainen, J. Lampinen i E. Oja, którzy analizowali hierarchiczne sieci SOM jako narzędzie grupowania [Koikkalainen, Oja 1990; Lampinen, Oja 1992]. Sieć SOM [Kohonen 1995] jest obecnie jednym z bardziej efektywnych narzędzi eksploracji danych, które mają zastosowanie w zagadnieniach klasyfikacji [Corridoni, Bimbo, Landi 1996; Ye, Lo 2000; Papadimitriou i in. 2002] 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 2007, 2008;Migdał-Najman, Najman 2003; Vesanto, Alhoniemi 2000].

Analizując różne podejścia do konstrukcji hierarchicznych sieci SOM, wyróżnić należy dwa podstawowe podejścia: aglomeracyjne i deglomeracyjne (rys. 1).



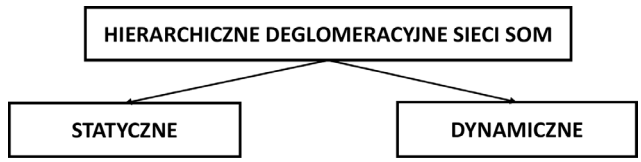
Rys. 1. Klasyfikacja hierarchicznych sieci SOM

Źródło: opracowanie własne.

W **podejściu deglomeracyjnym (**H_dSOM), na poziomie pierwszym buduje się zazwyczaj jedną sieć SOM o stosunkowo dużej liczbie neuronów, a następnie, na kolejnych poziomach rozbija się ją na części i dla każdej z nich buduje się kolejne sieci SOM (podsieci), zwykle znacznie mniejsze niż na pierwszym stopniu.

Budując sieci H_dSOM, można wyróżnić dwa podejścia: statyczne (*static divisive* HSOM) i dynamiczne (*dynamic divisive* HSOM) (rys. 2).

W **podejściu statycznym** stosuje się jeden z dwóch wariantów budowy sieci. W **wariancie pierwszym** na pierwszym poziomie budowana jest pojedyncza sieć



Rys. 2. Klasyfikacja hierarchicznych deglomeracyjnych sieci SOM

Źródło: opracowanie własne.

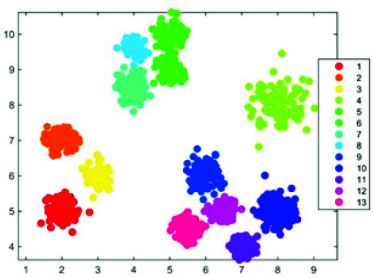
SOM o niewielkiej liczbie neuronów. Jej celem jest wyróżnienie jedynie najważniejszych grup badanych jednostek. Na poziomie drugim dla każdego neuronu sieci SOM z poziomu pierwszego budowana jest nowa sieć SOM. Zwykle każda z sieci drugiego poziomu jest budowana dla jednostek, za które odpowiadał dany neuron na pierwszym poziomie. Sieci drugiego poziomu odpowiadają za odwzorowanie mikrostruktur grupowych znajdujących się w skupieniach wyższego rzędu. W **wariancie drugim** na pierwszym poziomie budowana jest jedna większa sieć SOM. Na sieci wyróżniane są skupienia neuronów. Na poziomie drugim, dla każdego wyróżnionego skupienia neuronów na sieci SOM z poziomu pierwszego, budowana jest nowa sieć SOM. Każda z sieci drugiego poziomu jest budowana dla jednostek, za które odpowiadał dany zespół neuronów na pierwszym poziomie.

W **podejściu dynamicznym** na poziomie pierwszym buduje się pojedynczą sieć SOM, pozwalając ewoluować jej rozmiarowi. Ewolucję tę realizuje się, stosując algorytm Growing SOM [Fritzke 1991a, b, 1996], rozpoczynając od sieci o rozmiarze 2 × 2 neurony i zakładając jedynie maksymalny rozmiar sieci. Na sieci pierwszego poziomu ponownie dokonywane jest grupowanie i dla każdego wykrytego skupienia buduje się kolejne sieci. W procesie tym dopuszcza się dwa modele wzrostu sieci: **w poziomie** i w **pionie**. Pierwszy dotyczy wzrostu liczby neuronów każdej sieci SOM, drugi dotyczy liczby poziomów w H_dSOM. Rozmiar sieci SOM na każdym poziomie i liczba poziomów określane są w trakcie uczenia i zależą od przyjętego kryterium, np. błędu kwantyzacji.

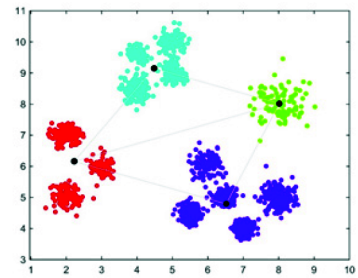
3. Konstrukcja statycznej sieci H_dSOM

Konstrukcja statycznej sieci H_dSOM zostanie zaprezentowana na przykładzie abstrakcyjnego dwuwymiarowego zbioru składającego się z 2000 jednostek (rys. 3a). Dane zostały wygenerowane w taki sposób, aby jednostki tworzyły wyraźne sferyczne skupienia o gęstości wzrastającej w kierunku centrum skupienia, a jednocześnie same skupienia tworzyły grupy (rys. 3a). Dla prezentowanych jednostek zbudowano na pierwszym poziomie sieć SOM o rozmiarze 2 × 2 neurony, uzyskując podstawową makrostrukturę skupień. Łatwo zauważyć, że wyróżnione skupienia mają głębszą strukturę, gdyż trzy z nich złożone są z kilku skupień drugiego rzędu (rys. 3b). Na drugim poziomie budowy statycznej sieci H_dSOM, dla jednostek wyróżnionych przez każdy z czterech neuronów buduje się niezależną sieć. Neuron pierwszy

(1) odpowiada za cztery skupienia, neuron drugi (2) za jedno skupienie, neuron trzeci (3) za pięć skupień, a neuron czwarty (4) za trzy skupienia (rys. 4).



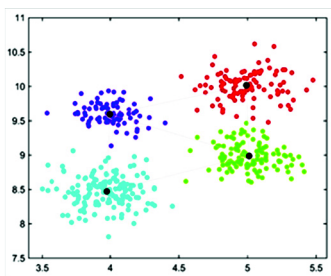
a) struktura grupowa 2000 jednostek



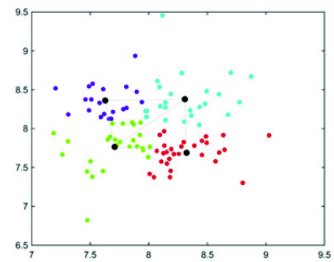
b) sieć SOM o rozmiarze 2 × 2 neurony i makrostruktura skupień

Rys. 3. Zbiór 2000 jednostek w przestrzeni dwuwymiarowej i sieć SOM

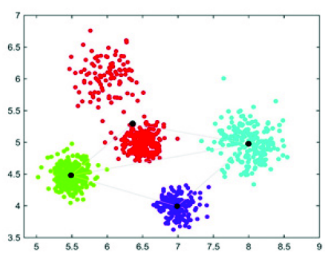
Źródło: opracowanie własne.



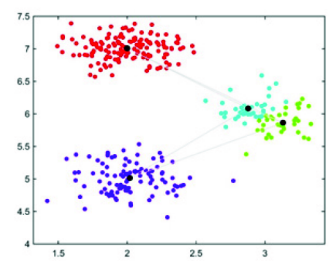
a) sieć 2 × 2 zbudowana dla jednostek przypisanych do 1 neuronu z I poziomu uczenia się



b) sieć 2 × 2 zbudowana dla jednostek przypisanych do 2 neuronu z I poziomu uczenia się



c) sieć 2 × 2 zbudowana dla jednostek przypisanych do 3 neuronu z I poziomu uczenia się



d) sieć 2 × 2 zbudowana dla jednostek przypisanych do 4 neuronu z I poziomu uczenia się

Rys. 4. Struktura grupowa uzyskana na drugim poziomie budowy statycznej sieci H_dSOM

Źródło: opracowanie własne.

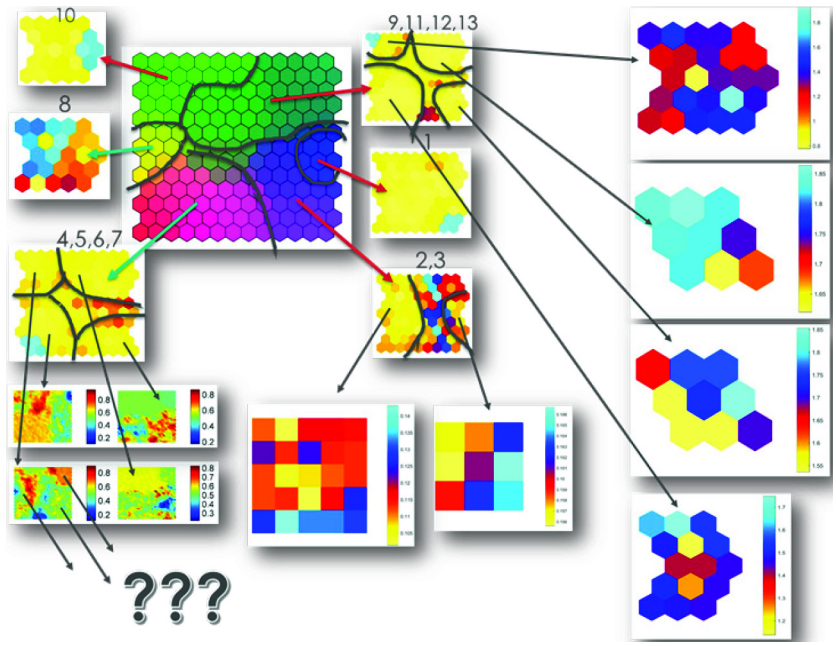
Sieci drugiego poziomu pozwoliły zaobserwować mikrostrukturę skupień istniejącą wewnątrz makroskupień wyróżnionych na pierwszym poziomie hierarchii. Tam, gdzie istnieją jeszcze głębsze struktury, można procedurę kontynuować (rys. 4c). Jednocześnie niektóre neurony nie pozwalają na wykrycie i interpretację skupień, ponieważ faktycznie ich nie ma (rys. 4b).

4. Konstrukcja dynamicznych sieci H_dSOM

W podejściu dynamicznym w budowie sieci H_dSOM badacz ma największą swobodę. Nie ustala liczby sieci, ich rozmiarów ani nawet liczby stopni w hierarchii sieci. Każda z budowanych sieci dopasowuje się dynamicznie do istniejących potrzeb, a więc do liczby i struktury przestrzennej jednostek, które mają być odwzorowane. Wykorzystuje się tu zwykle wariant sieci SOM o dynamicznie zmieniającej się strukturze, nazywany Growing SOM (GSOM). W sieci GSOM proces uczenia się sieci rozpoczyna się od minimalnej struktury, a więc sieci o rozmiarze 2×2 neurony. W kolejnych iteracjach procesu samouczenia się dodawane są wiersze lub kolumny neuronów, odpowiednio w tych częściach sieci, które mają największy udział w ogólnym błędzie kwantyzacji sieci [Kohonen 1995; Migdał-Najman, Najman 2014]. Proces iterowania kończy się, gdy osiągnięta zostanie maksymalna, założona struktura sieci lub w kolejnych iteracjach nie zmienia się błąd kwantyzacji sieci (sieć osiągnęła stabilne rozwiązanie). W ten sposób ogranicza się wpływ subiektywnego ustalania rozmiaru sieci przez badacza.

Na pierwszym poziomie budowy sieci H_dSOM, gdy analizie podlegają wszystkie badane jednostki, sieć SOM przybiera zazwyczaj znaczne, liczące setki neuronów rozmiary. Rozmiar ten wynika jednocześnie z liczby samych jednostek i z ich struktury przestrzennej. Na kolejnych poziomach deglomeracji dla każdej wyróżnionej struktury przestrzennej tworzone są nowe, zwykle znacznie mniejsze sieci i poszukiwane są nowe, bardziej lokalne struktury. Proces zagłębiania się jest kontynuowany tak długo, jak długo możliwe jest wyróżnianie możliwych do identyfikacji struktur przestrzennych jednostek. Ich istnienie można wykryć, stosując jeden z wielu wskaźników homogeniczności jednostek lub heterogeniczności skupień. Należą do nich wskaźniki entropii, współczynniki koncentracji przestrzennej, wskaźnik sylwetkowy i inne [Migdał-Najman, Najman 2013].

Proces powstawania sieci H_dSOM, w wariancie dynamicznym, dla danych analogicznych jak w pierwszym przykładzie pokazano na rys. 5. W centrum rysunku znajduje się sieć SOM, która powstała dla wszystkich badanych jednostek. Ma ona rozmiar 13×13 neuronów i heksagonalną strukturę połączeń neuronów. Na wizualizacji pokazano na skali barwnej podobieństwo jednostek odwzorowywanych przez neurony (*similarity coloring*) [Migdał-Najman, Najman 2013, 2014]. Pozwalają one wyróżnić przynajmniej sześć makrostruktur przestrzennych (oznaczonych na sieci). Dla każdej z nich tworzone są nowe sieci, których mapy ujednoliconych



Rys. 5. Hierarchiczna deglomeracyjna dynamiczna sieć SOM

Źródło: opracowanie własne.

odległości pokazano wokół sieci centralnej. Na trzech z nich, oznaczonych 1, 8, 10 (numery skupień jak na rys. 3a) nie można już wyróżnić żadnych struktur. Na sieci oznaczonej 9-11-12-13 wyraźnie można zauważyć istnienie przynajmniej czterech regionów podobieństwa jednostek, co świadczy o istnieniu skupień niższego rzędu. Podobnie na sieci oznaczonej 2-3 widoczne są dwa takie obszary, a na sieci oznaczonej 4-5-6-7 cztery. Dla tych sieci budowane są w kolejnym kroku deglomeracji kolejne sieci, w liczbie równej liczbie zaobserwowanych obszarów podobieństwa jednostek. Sieci te pokazane są na zewnętrznej części rysunku. Wszystkie budowane sieci mają różne struktury, które powstały w autonomicznym procesie samouczenia się sieci GSOM. Analiza przedstawionych wizualizacji pozwala zauważyć wszystkich 13 faktycznie istniejących skupień. Widoczne są także mikrostruktury dostrzegalne tylko w odpowiednio małej skali, co sugeruje, że proces eksploracji mógłby być kontynuowany (sieci oznaczone znakiem zapytania).

5. Zakończenie

Samouczące się deglomeracyjne sieci neuronowe typu H_dSOM mają wiele zalet w analizie skupień. Pozwalają obserwować struktury równie szerokie, a nawet ogólniejsze niż pojedyncza sieć SOM. Od sieci na pierwszym poziomie nie wyma-

ga się bowiem wykrycia wszystkich skupień, a jedynie podstawowej makrostruktury obiektów. Jednocześnie sieci H_dSOM pozwalają znajdować i eksplorować mikrostruktury trudne do uchwycenia w sieci SOM. Proces deglomeracji pozwala także obserwować strukturę hierarchiczną skupień i lokalnych mikrostruktur. Chociaż buduje się tu wiele sieci, proces ten jest szybki. Sieć pierwszego poziomu jest zwykle mniejsza niż w pojedynczej sieci SOM, a na dalszych poziomach struktura przestrzenna jednostek, a także ich liczba, nie wymagają zwykle dużych rozmiarów sieci. Sieci te mają niewielkie rozmiary, dlatego szybkość ich uczenia się jest duża. Łącznie szybkość całej analizy jest porównywalna z szybkością analizy pojedynczej sieci SOM lub od niej większa. Budowa każdej sieci jest osobnym procesem, co zmniejsza wymagania wobec zasobów komputera, ponieważ możliwa jest bieżąca archiwizacja sieci. Jest to element nie bez znaczenia w analizie bardzo dużych zbiorów danych. Jednocześnie, ze względu na swoją konstrukcję, sieć H_dSOM zachowuje wszystkie zalety klasycznej sieci SOM (tab. 1).

Tabela 1. Własności sieci H_dSOM

Własności	H_dSOM
Struktura sieci	zmienna/zależna od modelu
Liczba krytycznych parametrów sterujących	duża/zależna od modelu
Szybkość uczenia	porównywalna z SOM/zależna od modelu
Wymagania pojemności pamięci	przeciętne/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/bardzo głęboka

Źródło: opracowanie własne.

Zalety te okupione są jednak pewną liczbą wad. Największą jest prawdopodobnie bardzo duża liczba parametrów koniecznych do założenia *a priori*. Sieci buduje się wiele, a dla każdej z nich trzeba ustalić przynajmniej pięć parametrów: maksymalny rozmiar sieci, poziom błędu kwantyzacji dla dodawanej nowej warstwy neuronów, rodzaj i zasięg sąsiedztwa, strukturę połączeń neuronów. Oznacza to w praktyce konieczność wprowadzania uproszczeń, zakładając np. te same wartości parametrów dla wszystkich sieci, co nie jest strategią optymalną. Kolejnym problemem jest bieżące profilowanie uzyskiwanych grup, co jest konieczne dla podjęcia decyzji o dalszej deglomeracji. Gdy sieci jest wiele, konieczne staje się automatyzowanie tego procesu, co także nie jest strategią optymalną z punktu widzenia ogólnej zdolności sieci do analizy badanego problemu. Inną wadą, jednak

o dużym znaczeniu praktycznym, jest całkowity brak szerzej dostępnego oprogramowania, które pozwalałoby wykonywać tego typu analizy. Wszyscy autorzy badań przygotowują takie oprogramowanie we własnym zakresie. Wydaje się jednak, że zalety w ogólnym bilansie przeważają i warto włączyć sieci H_dSOM do zestawu technik analitycznych stosowanych w analizie skupień.

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.

Fritzke B., 1991a, *Let it grow – self organizing feature maps with problem dependent cell structure*, [w:] *Proceedings of the International on Artificial Neural Networks, ICANN'91*, Helsinki, s. 403–408.

Fritzke B., 1991b, *Unsupervised clustering with growing cell structures*, [w:] *Proceedings of the International Joint Conference on Neural Networks, IJCNN'91*, Seattle, WA, s. 531–536.

Fritzke B., 1996, *Growing self-organizing networks – why?*, [w:] M. Verleysen (red.), *Proceedings of European Symposium on Artificial Neural Networks, ESANN'96*, Bruges, s. 61–72.

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 Eko-

nomicznej 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.

Migdał-Najman, Najman K., 2014, *Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 327, Taksonomia 22: *Klasyfikacja i analiza danych – teoria i zastosowania*, s. 131–138.

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.