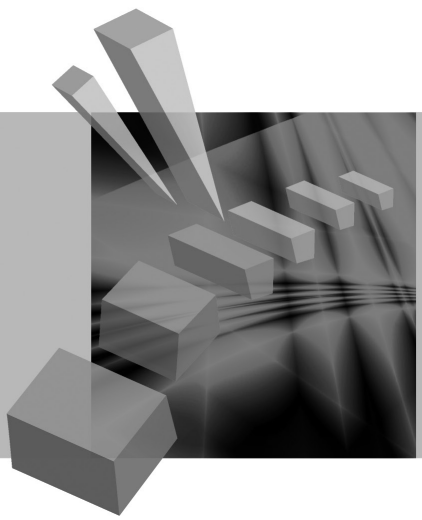
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
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 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Kamila Migdał-Najman
Uniwersytet Gdański

ZASTOSOWANIE JEDNOWYMIAROWEJ SIECI SOM DO WYBORU CECH ZMIENNYCH W GRUPOWANIU DYNAMICZNYM

Streszczenie: W artykule zaproponowano oryginalną procedurę wyboru cech w grupowaniu dynamicznym. Jako algorytm grupowania dynamicznego wykorzystano samouczącą się sieć neuronową typu GNG. Aby przyspieszyć i polepszyć wyniki grupowania, redukowano liczbę zmiennych, korzystając z sieci SOM. W eksperymencie symulacyjnym wykazano skuteczność takiego rozwiązania. W efekcie uzyskano znaczące przyspieszenie procesu grupowania dynamicznego bez utraty jakości grupowania.

Słowa kluczowe: sieć samoorganizująca się Kohonena (SOM), sieć typu gaz neuronowy (GNG), grupowanie dynamiczne.

1. Wstęp

Gwałtowny rozwój technik komputerowych powoduje między innymi lawinowe powiększanie się rozmiarów zbiorów danych, które wymagają analizy. Jednym z problemów, które zyskują na znaczeniu, jest rosnąca liczba cech zmiennych opisujących pojedynczą jednostkę. Tak duża liczba cech zmiennych w istotny sposób zmienia skalę problemów stojących przed analizą takich zbiorów danych. Malejące koszty zbierania i przechowywania danych, a także strategia: „Nie wiem, co jest ważne, zbieram wszystkie dane, jakie się tylko da", spowodowały łatwość w podejmowaniu decyzji o dodawaniu kolejnych cech do zbiorów danych. Wiele z analizowanych cech zmiennych może być nieistotnych dla struktury grupowej, inne mogą powielać te same informacje. Współliniowość i wysoka korelacja cech zmiennych jest niepożądaną cechą zbioru danych, utrudniając identyfikację struktury grupowej i znacząco zwiększając koszt samej analizy. Jak wskazuje między innymi G.W. Milligan [1994; 1996], podejście stosowane często przez twórców zbiorów danych, aby jednostki opisywać możliwie dużą liczbą cech zmiennych, jest nie tyko niepotrzebne, ale wręcz błędne.

Kolejną ważną cechą współczesnych zbiorów danych jest ich dynamiczny charakter, który wynika z bardzo dużej częstotliwości ich aktualizacji. W sieciach te-

lekomunikacyjnych czy systemach rejestrujących transakcje bankowe zbiór danych może być aktualizowany kilkaset razy na sekundę. Dynamiczny wzrost zawartości zbioru danych może powodować także dynamiczne zmiany jego struktury grupowej. Krytyczną cechą metod grupowania dynamicznego jest ich szybkość. Gdy liczba jednostek w zbiorze wzrasta o kilkaset, a nawet wiele tysięcy na sekundę, a grupowanie ma być aktualne w dowolnym momencie – szybkość grupowania staje się decydująca. Jedną z możliwych strategii prowadzących do zwiększenia szybkości grupowania jest redukcja zbędnych cech. Redukcja taka przynosi zwykle poprawienie jakości grupowania: zwiększenie homogeniczności skupień, zwiększenie heterogeniczności między skupieniami, łatwiejszą interpretację wyróżnionych skupień i znaczne skrócenie czasu analizy. Aby wybrać cechy grupujące, można dokonać ich grupowania, a następnie z każdej z wyróżnionych grup wybrać reprezentanta charakteryzującego się największą zdolnością do wyróżniania skupień. Jedną z możliwych do zastosowania metod grupowania cech zmiennych jest samoucząca się sieć neuronowa typu SOM (*Self-Organizing Map*). Celem prezentowanych badań jest weryfikacja własności jednowymiarowej sieci SOM w grupowaniu cech zmiennych. W szczególności weryfikowana będzie możliwość wykorzystania sieci tego typu w procesie grupowania dynamicznego.

2. Grupowanie cech zmiennych

Problem wyboru zmiennych jest kluczowym zagadnieniem w klasyfikacji jednostek wielowymiarowych. Z tego powodu poświęcono mu w literaturze odpowiednio wiele miejsca [Gnanadesikan, Kettenring, Tsao 1995]. Zasadniczo wyróżnić można trzy podejścia do ustalania optymalnego zbioru cech zmiennych: 1) ważenie zmiennych – gdzie każdej zmiennej nadaje się wagę mówiącą o jej relatywnej ważności w opisie badanego problemu, 2) selekcję zmiennych – polegającą na tym, że ze zbioru zmiennych eliminuje się te, których potencjał dyskryminacyjny wydaje się najmniejszy; podejście to może być uznane za szczególny przypadek podejścia pierwszego, gdzie wagi zmiennych przyjmują jedynie wartości 0 – dla zmiennych odrzuconych i 1 – dla wybranych oraz 3) zastąpienie zmiennych oryginalnych przez zmienne sztuczne – jest to klasyczne statystyczne podejście bazujące na analizie głównych składowych [Walesiak 2005].

Optymalny zbiór cech zmiennych powinien zawierać w sobie jedynie te cechy, które istotnie różnicują badane jednostki. W badaniach statystycznych można wyróżnić wiele metod wyodrębniania jednorodnych grup zmiennych. W zagadnieniach związanych z grupowaniem cech na ogół przyjmowane są te same algorytmy taksonomiczne, jakie stosowane są w grupowaniu jednostek. W większości przypadków istniejące procedury postępowania polegają na pośrednim lub bezpośrednim wykorzystaniu miar podobieństwa między porównywanymi elementami. Do proponowanych w literaturze procedur grupowania cech zmiennych można zaliczyć: metodę Czekanowskiego, taksonomię wrocławską, metodę Prima, analizę wiązek

Gowera-Rossa, procedury aglomeracyjne z grupy Lance'a-Williamsa-Warda, metody obszarowe i inne. W taksonomii cech zmiennych wykorzystywane są również procedury oparte na macierzy związku, najczęściej macierzy korelacji. Do proponowanej grupy zaliczyć można: parametryczną metodę klasyfikacji cech Hellwiga, metodę Bekkera, metodę Łukackiej, grafowe procedury taksonomii cech zaproponowane przez Plutę, Bartosiewicz wraz z ich modyfikacjami, metodę Kinga, metodę Holzingera i Hermana i inne [Pociecha i in. 1998]. Możliwa jest również transformacja danych wielowymiarowych z wielowymiarowej przestrzeni na płaszczyznę i dokonywanie wyboru cech w zredukowanej przestrzeni. Do metod takich możemy zaliczyć: metodę głównych składowych, metodę głównych współrzędnych czy metodę współrzędnych dyskryminacyjnych. Możliwa jest również wizualizacja obiektów wielowymiarowych w formie rysunków symbolicznych i analizowanie grup zmiennych o podobnych własnościach. Do metod tego typu zaliczyć można: metodę rytów Andersona, krzywe Andrewsa, twarze Chernoffa – rozwijaną przez B. Flury i Riedwyla, lub metodę równoległych współrzędnych (*parallel coordinates*).

Obok znanych klasycznych procedur grupowania cech zmiennych można również zastosować sztuczne sieci neuronowe. Jedną z takich sieci neuronowych posiadających wysoki potencjał w grupowania jednostek i cech zmiennych jest samoorganizująca się sieć Kohonena, nazywana również mapą samoorganizującą się (*Self-Organizing Map* – SOM) [Kohonen 1995; 1997; 2001; Deboeck, Kohonen 1998; Kaski, Kangas, Kohonen 1998; Berthold, Hand 1999; Migdał-Najman, Najman 2008]). Sieć Kohonena należy do bardziej znanych nienadzorowanych modeli sztucznych sieci neuronowych. Sieć SOM tworzy nieliniową projekcję zbioru danych na siatkę, mapę Kohonena i zachowuje topologię zbioru wejściowego, tj. jednostki, które w przestrzeni wejściowej są do siebie podobne, na mapie SOM reprezentowane będą przez ten sam neuron lub neurony, które znajdują się blisko siebie. Jedną z istotnych własności sieci SOM jest możliwość wizualizacji wyników grupowania na macierzy ujednoliconych odległości, tzw. macierzy U. Posługiwanie się macierzą U do oceny zdolności dyskryminacyjnych analizowanego zbioru danych jest wysoce skuteczne. Jeżeli w zbiorze danych, w którym występuje wiele cech zmiennych, cechy te są w różny sposób i w różnym stopniu skorelowane ze sobą, istotne staje się wstępne pogrupowanie cech zmiennych. Jeżeli wyróżnimy wstępnie skupienia zmiennych o podobnych własnościach, z każdego skupienia zmiennych można wyeliminować zmienne o najmniejszym potencjale dyskryminacyjnym. Każde ze skupień zmiennych analizuje się niezależnie, co ułatwia analizę. Zastosowana procedura nie pozwoli na usunięcie wszystkich cech zmiennych, które mają podobne własności i znajdują się w jednym skupieniu. Może również wystąpić taka sytuacja, że z przyczyn merytorycznych niektóre zmienne będziemy chcieli zachować w badaniu lub przynajmniej jedną z każdego wyróżnionego skupienia. Do wyróżnienia skupień cech zmiennych można wykorzystać sieć SOM o topologii łańcucha.

3. Sieć GNG w grupowaniu dynamicznym

Grupowaniem dynamicznym można nazwać taki proces grupowania, w trakcie którego do zbioru danych non stop napływają, odpływają z niego lub jednocześnie napływają i odpływają jednostki, a ich struktura grupowa może się zmieniać. Taki zbiór danych podlega ciągłej aktualizacji, a proces grupowania nie zostaje przerwany.

Jedną z metod grupowania możliwą do wykorzystania w grupowaniu dynamicznym jest samoucząca się sieć neuronowa typu GNG (*Growing Neural Gas*) [Fritzke 1994] o zmiennej strukturze. W procesie samouczenia się sieci neurony wstawiane są w te obszary sieci, w którym występuje największy błąd rozpoznawania wzorców. Sieć tego typu bardzo szybko uczy się i sama poszukuje i rozpoznaje optymalną strukturę grupową (o ile skupienia są separowalne). Również posiada zdolność rozpoznawania skupień o dowolnej konfiguracji w przestrzeni cech, ale jednocześnie popełniać będzie niewielkie błędy na krawędziach skupień, w których trudno jest o jednoznaczne zaliczenie jednostki do skupienia. Sieć typu GNG nie pozwala na wizualizację danych i samej sieci, ale należy do grupy wyspecjalizowanych narzędzi analizy skupień i w tym zakresie w większości przypadków jest skuteczniejsza niż inne sieci samoorganizujące, jak np. sieć SOM [Najman 2011; 2012].

4. Eksperyment badawczy

Potencjał obu sieci może być wykorzystany łącznie. Sieć SOM może posłużyć do wyboru cech, które staną się podstawą grupowania dynamicznego z wykorzystaniem sieci GNG. Do weryfikacji postawionej hipotezy przygotowano eksperyment. Wygenerowano umowny zbiór danych, w którym cechy zmienne przygotowane zostały w taki sposób, aby tworzyć skupienia o gęstości rosnącej w kierunku centrum skupienia. Skupienia są sferyczne, separowalne, a ich centra znajdują się przeciętnie w odległości 2,5-krotności ich średnic. 131 razy (nazwijmy to krokiem) następowała aktualizacja zbioru danych: napływały, odpływały lub jednocześnie napływały i odpływały jednostki (dane). Każda jednostka opisana była 20 cechami zmiennymi. W pierwszym kroku w bazie było jedynie 20 jednostek, które należały do jednego skupienia. Natomiast w kroku ostatnim w zbiorze jednocześnie było 1588 jednostek, które znajdowały się w 4 skupieniach. Liczba skupień w zbiorze danych zmieniała się w 131 krokach od 1 do 5.

Na 131 aktualizacji zbioru danych zaobserwowano 110 faz statycznych i 21 faz dynamicznych. W fazie statycznej sieć ma tak ustalone parametry, aby uczyła się powoli, z maksymalną dokładnością. Jeżeli po aktualizacji zbioru danych wykryta struktura grupowa nie ulega pogorszeniu ze względu na przyjętą miarę jakości grupowania, sieć pracuje w fazie statycznej. W badaniu przyjęto, że jeżeli poziom wskaźnika sylwetkowego dla rozpoznanej struktury grupowej przez sieć GNG był powyżej poziomu 0,7, to sieć uczyła się w fazie statycznej. Natomiast jeżeli po zmianie jednostek wartość wskaźnika sylwetkowego spadała poniżej przyjęte-

Tabela 1. Grupowanie dynamiczne z wykorzystaniem sieci GNG i SOM od 19 do 33 kroku

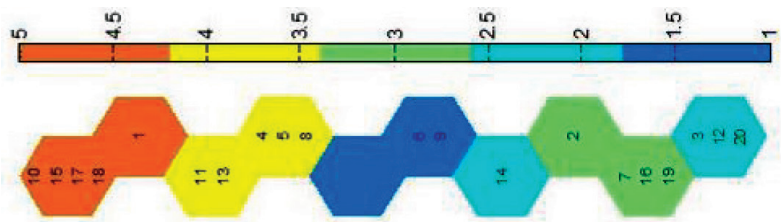
KROK	LICZBA PRZYPADKÓW JEDNOCZESNIE W BAZIE	LICZBA NEURONÓW SIECI GNG	PRAWDZIWA LICZBA SKUPIEN	LICZBA SKUPIEŃ GNG	RAND	CZAS FAZY STATYCZNEJ	CZAS FAZY DYNAMICZNEJ	ŁĄCZNA LICZBA ITERACJI	SILHOUETTE	LICZBA ODRZUCONYCH CECH	LICZBA CECH PO ODRZUCENIU
19	252	27	2	2	1	0,000496309	0	31420	1	0	20
20	271	25	2	2	1	0,000474908	0	33021	1	0	20
21	288	32	2	2	1	0,000485841	0	34622	1	0	20
22	292	32	2	2	1	0,000543037	0	36223	1	0	20
23	285	33	2	2	1	0,000544525	0	37824	1	0	20
24	289	32	2	2	1	0,00050936	0	39425	1	0	20
25	298	32	2	2	1	0,000517992	0	41026	1	0	20
26	304	33	2	2	1	0,000519728	0	42627	1	0	20
27	300	35	2	2	1	0,000529072	0	44228	1	0	20
28	315	31	2	2	1	0,000570513	0	45829	1	0	20
29	326	33	2	2	1	0,000567912	0	47430	1	0	20
30	353	35	2	2	1	0,000706461	0	49031	1	0	20
31	**361**	**22**	**2**	**2**	**1**	**0**	**0,00019835**	**51032**	**1**	**3**	**17**
32	371	22	2	2	1	0	3,2384E-05	2001	1	14	6
33	364	24	2	2	1	0,000394356	0	3602	1	0	6

Źródło: opracowanie własne.

go poziomu 0,7, sieć GNG przechodziła w **fazę uczenia dynamicznego**. W fazie tej parametry sieci pozwalają jej na szybkie uczenie się istotnie innych jednostek. Neurony są szybciej wstawiane i mają większy krok uczenia. Dla fazy statycznej przyjęto następujące parametry uczenia się sieci: nowy neuron wstawiany co $wiek_{max}$=150 iteracji, krok uczenia neuronu wygrywającego ε_b = 0,01 i jego najbliższego sąsiada ε_n = 0,005. Dla fazy dynamicznej odpowiednio: $wiek_{max}$= 50, ε_b = 0,1 i ε_n = 0,01. Każdorazowo po przejściu z fazy statycznej do dynamicznej dokonywany był wybór cech na podstawie sieci SOM o topologii łańcucha o 10 neuronach. Taka struktura sieci wynika bezpośrednio z konieczności bardzo szybkiego wyboru cech w grupowaniu statycznym. Proces samouczenia się sieci SOM jest wielokrotnie wolniejszy niż sieci GNG, jej struktura musi być więc maksymalnie prosta. Wybór dokonywany był zawsze spośród wszystkich 20 cech. Jeżeli po procesie wyboru liczba cech ulegała zmianie, dokonywano korekty we współrzędnych neuronów sieci GNG w ten sposób, że usuwano współrzędne odpowiedzialne za usuwaną cechę i dodawano wektor losowych współrzędnych w miejsce cechy dodawanej.

Trzydzieści pierwszych kroków, w których aktualizowany był zbiór danych, to faza statyczna. W kroku trzydziestym w zbiorze znajdowały się 353 jednostki opisane 20 cechami zmiennymi, które sieć GNG pogrupowała na dwa skupienia (por. tab. 1). Sieć bezbłędnie rozpoznaje strukturę grupową. Wartość współczynnika Randa wyniosła jeden, oznaczając idealną zgodność i przynależność każdej jednostki do właściwego skupienia. Dla wszystkich kroków (1-31) wskaźnik sylwetkowy był powyżej przyjętego progu 0,7.

W 31 kroku wskaźnik sylwetkowy w momencie dołączania nowych danych do zbioru uzyskał wartość poniżej ustalonego progu 0,7. Zgodnie z przyjętym założeniem sieć GNG przechodzi do fazy dynamicznej i rozpoczyna się grupowanie cech. Pierwsze grupowanie 20 cech zmiennych przeprowadzone zostało na początku fazy 31. Dla 361 jednostek i 20 cech zmiennych przeprowadzono grupowanie cech na bazie sieci SOM. Zbudowano sieć SOM o rozmiarze 10 × 1, z gaussowską funkcją sąsiedztwa o zasięgu 2, która uczona była w 300 iteracjach. Na niej wyróżniono 5 skupień. Liczebności poszczególnych skupień wyniosły odpowiednio: 5, 5, 2, 4, 4. Na rysunku 1 zaprezentowano sieć SOM uzyskaną w 31 kroku.



Rys. 1. Sieć SOM o topologii łańcucha z 31 kroku

Źródło: opracowanie własne.

Do oceny zdolności dyskryminacyjnej cech zmiennych zastosowano współczynnik koncentracji bazujący na entropii [Migdał-Najman, Najman 2008]. Poziom współczynnika koncentracji dla poszczególnych cech zmiennych w wyróżnionych 5 skupieniach przedstawiono w tab. 2. W badaniu przyjęto, że eliminowane z dalszego badania będą te cechy ze skupień, których poziom współczynnika koncentracji będzie mniejszy niż 0,047. Założono również, że każde skupienie ma reprezentować co najmniej jedna cecha, niezależnie od uzyskanego poziomu współczynnika koncentracji. W wyniku zastosowania powyższej procedury do dalszego etapu fazy dynamicznej wyróżniono 17 cech zmiennych.

Tabela 2. Wynik grupowania cech na podstawie sieci SOM w 31 kroku

Skupienie	Cechy zmienne w wyróżnionych skupieniach	Współczynnik koncentracji cech zmiennych w wyróżnionych skupieniach
1	**10**, **15**, 17, 18, 1	**0,044236**, **0,04679**, 0,049802, 0,05404, 0,049311
2	11, 13, 4, 5, 8	0,053148, 0,047658, 0,052681, 0,047348, 0,054756
3	6, 9	0,051526, 0,048408
4	14, 3, 12, 20	0,048196, 0,047453, 0,051896, 0,049524
5	2, 7, **16**, 19	0,048458, 0,048195, **0,046713**, 0,049546

Źródło: opracowanie własne.

361 jednostek i 17 cech zmiennych (bez cechy 10, 15, 16) pogrupowano za pomocą sieci GNG (22 neurony) na 2 skupienia. Uzyskano idealne grupowanie ze współczynnikiem Randa równym jeden i wskaźnikiem sylwetkowym równym jeden.

W kroku 32 po kolejnej aktualizacji danych nastąpiło kolejne pogorszenie jakości grupowania, które zasygnalizowane zostało przez wskaźnik sylwetkowy. Ponownie rozpoczęto grupowanie cech zmiennych, ale tym razem opisujących 371 jednostek. Do etapu grupowania dynamicznego na podstawie sieci GNG wytypowano jedynie 6 cech. Mimo odrzucenia w 32 kroku aż 14 cech zmiennych, jakość grupowania badanych jednostek nie pogorszyła się (wskaźnik sylwetkowy przyjął poziom równy 1). Wyniki grupowania dynamicznego na podstawie sieci GNG dla kolejnych kroków (do kroku 66) przedstawiono w tab. 3.

Tabela 3. Grupowanie dynamiczne z wykorzystaniem sieci GNG i SOM od 34 do 66 kroku

KROK	LICZBA PRZYPADKÓW JEDNOCZESNIE W BAZIE	LICZBA NEURONÓW SIECI GNG	PRAWDZIWA LICZBA SKUPIEN	LICZBA SKUPIEŃ GNG	RAND	CZAS FAZY STATYCZNEJ	CZAS FAZY DYNAMICZNEJ	ŁĄCZNA LICZBA ITERACJI	SILHOUETTE	LICZBA ODRZUCONYCH CECH	LICZBA CECH PO ODRZUCENIU
34	395	27	2	2	1	0,000401663	0	5203	1	0	6
35	421	23	2	2	1	0	0,00015291	2001	1	12	8
36	453	23	2	2	1	0	0,00028543	2001	1	14	6
37	454	23	2	2	1	0	0,000276709	2001	1	14	6
38	453	28	2	2	1	0,000400115	0	3602	1	0	6
39	454	32	2	2	1	0,00044333	0	5203	1	0	6
40	448	33	2	2	1	0,000460403	0	6804	1	0	6
41	469	35	2	2	1	0,000475227	0	8405	1	0	6
42	481	33	2	2	1	0,000456347	0	10006	1	0	6
43	500	32	2	2	1	0,000471292	0	11607	1	0	6
44	504	38	2	3	0,94	0,000465996	0	13208	0,33333333	0	6
45	507	36	2	2	1	0,000471993	0	14809	1	0	6
46	528	31	3	2	0,99	0,000463776	0	16410	0,92334266	0	6
47	533	31	3	5	0,98	0,000454442	0	18011	0,5	0	6
48	540	27	3	3	1	0,000435877	0	19612	1	0	6
49	540	27	3	3	1	0,000506541	0	21213	1	0	6
50	562	30	3	3	1	0,000641817	0	22814	1	0	6
51	575	31	3	3	1	0,000525943	0	24415	1	0	6
52	599	33	3	3	1	0,000478034	0	26016	1	0	6
53	609	33	3	4	0,96	0,000457441	0	27617	0,5	0	6
54	607	33	3	3	1	0,000456135	0	29218	1	0	6
55	612	28	3	3	1	0,000446837	0	30819	1	0	6
56	620	29	3	3	1	0,000446017	0	32420	1	0	6
57	635	30	3	2	0,98	0,0004378	0	34021	0,848	0	6
58	633	27	3	2	0,98	0,000431766	0	35622	0,943	0	6
59	664	27	3	3	1	0,000437603	0	37223	0,714	0	6
60	685	23	2	2	1	0	3,85663E-05	2001	1	10	10
61	687	30	2	2	1	0,000487983	0	3602	1	0	10
62	691	33	2	2	1	0,000479416	0	5203	1	0	10
63	696	31	2	2	1	0,000508788	0	6804	1	0	10
64	707	34	2	2	1	0,000519326	0	8405	1	0	10
65	713	30	2	2	1	0,000502707	0	10006	1	0	10
66	720	31	2	2	1	0,000493951	0	11607	1	0	10

Źródło: opracowanie własne.

5. Wnioski

Proponowana metoda selekcji cech zmiennych posiada wiele zalet. Należy do procedur całkowicie autonomicznych, jest zgodna z filozofią *data mining*. Jest tym efektywniejsza, im liczba cech zmiennych jest większa. Selekcja części cech wpływa na skrócenie czasu procesu grupowania opartego na sieci GNG. Zaoszczędzony czas może zostać wówczas spożytkowany na zwiększenie liczby neuronów sieci i dokładniejsze grupowanie. Uczenie się jednowymiarowej sieci SOM jest również bardzo szybkie i w niewielkim stopniu wpływa na łączny czas grupowania. Selekcja nawet znacznej liczby cech nie musi powodować zmniejszenia jakości grupowania.

Zaproponowana procedura selekcji cech zmiennych w grupowaniu dynamicznym może być również zastosowana z innymi metodami grupowania danych i posłużyć jako preprocesor. Proponowane podejście wymaga subiektywnego ustalania minimalnej liczby cech zmiennych w wyróżnionych skupieniach cech, uzyskanych na podstawie sieci SOM. Nie ma żadnych merytorycznych wskazówek, jaka ta liczba powinna być. Wymaga również ustalenia wartości progowej współczynnika koncentracji, poniżej którego cechy zmienne zostaną odrzucone. Należy również zwrócić uwagę, że uzyskane wyniki grupowania dynamicznego będą wrażliwe na jakość samej sieci SOM. W opinii autorki wydaje się, że sieć SOM może być wykorzystana do selekcji cech zmiennych w grupowaniu dynamicznym i z powodzeniem może być stosowana w praktyce.

Literatura

Berthold M., Hand D.J., *Intelligent Data Analysis*, Springer-Verlag, Berlin Heidelberg, 1999, s. 253.

Deboeck G., Kohonen T., *Visual Explorations in Finance with Self-Organizing Maps*, Springer-Verlag, London 1998, s. 159.

Fritzke B., *Growing cell structures – a self-organizing network for unsupervised and supervised learning*, „Neural Networks", 7, 9, 1994, s. 1441-1460.

Gnanadesikan R., Kettenring J.R., Tsao S.L., *Weighting and selection of variable for cluster analysis*, „Journal of classification", 12, 1995, s. 113-136.

Kohonen T., *Self-Organizing Maps*, Springer-Verlag, Berlin, Heidelberg 1995, 1997, 2001.

Kaski S., Kangas J., Kohonen T., *Bibliography of self-organizing map (SOM) papers: 1981-1997*, „Neural Computing Surveys", 1, 1998, s. 102-350.

Milligan G.W., *Issues in applied classification: selection of variables to cluster*, Classification Society of North America, News Letter, November Issue 37, 1994.

Milligan G.W., *Clustering Validation: Results and Implications for Applied Analyses*, [in:] P. Arabie, L. Hubert, G. DeSoete (eds.), *Clustering and Classification*, River Edge, NJ, World Scientific, 1996, s. 341-375.

Migdał-Najman K., Najman K., *Applying the Kohonen Self-Organizing Map Networks to Select Variables*, [in:] C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (eds.), *Data Analysis, Machine Learning and Applications*, Springer-Verlag, Berlin, Heidelberg 2008, s. 45-54.

Migdał-Najman K., *Analiza porównawcza własności nienadzorowanych sieci neuronowych typu Self Organizing Map i Growing Neural Gas w analizie skupień*, [w:] Taksonomia 16, *Klasyfikacja i analiza danych – teoria i zastosowania,* Prace Naukowe UE we Wrocławiu nr 47, 2009, s. 205-213.

Najman K., *Grupowanie dynamiczne i wykorzystaniem sieci GNG*, „Przegląd Statystyczny", nr 3-4, 2011, 231-241.

Najman K., *Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG*, [w:] Taksonomia 19, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe UE we Wrocławiu nr 242, 2012, s. 361-369.

Pociecha J., Podolec B., Sokołowski A., Zając K., *Metody taksonomiczne w badaniach społeczno-ekonomicznych*, PWN, Warszawa 1998, s. 102-110.

Walesiak M., *Problemy selekcji i ważenia zmiennych w zagadnieniach klasyfikacji*, [w:] Taksonomia 12, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe UE we Wrocławiu nr 1076, 2005, s. 106-118.

APPLYING THE ONE-DIMENSIONAL SOM NETWORK TO SELECT VARIABLES IN DYNAMIC CLUSTERING

Summary: In the article the author proposes an original procedure for selecting the features in dynamic clustering. The author verifies the potential of the dynamic clustering method, such as: self-learning neural network type of GNG. To speed up and improve the results of the clustering, the author reduces the number of variables using SOM network. The simulation experiment shows the effectiveness of this approach. This approach allows a considerable speed up of the process of dynamic clustering without losing the quality of clustering.

Keywords: Self Organizing Map (SOM), Growing Neural Gas (GNG), dynamic clustering.