PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Kamila Migdał-Najman
Uniwersytet Gdański

ANALIZA PORÓWNAWCZA STRUKTUR HIERARCHICZNYCH SKUPIEŃ UZYSKANYCH Z WYKORZYSTANIEM HYBRYDOWYCH METOD GRUPOWANIA

1. Wstęp

W badaniach empirycznych często występuje potrzeba wyróżnienia homogenicznych grup obiektów, które są bardziej podobne do obiektów tworzących daną grupę niż obiekty spoza tej grupy. W wielu badaniach występuje także potrzeba ustalenia hierarchii tych obiektów. Wśród metod grupowania wielowymiarowych obiektów do najbardziej znanych i stosowanych należą hierarchiczne metody aglomeracyjne. Metody te są stosunkowo najlepiej opracowane pod względem metodologicznym, pozwalają też na graficzną prezentację uzyskanej klasyfikacji w formie dendrogramu. Coraz częściej w badaniach empirycznych wyodrębnia się jednorodne grupy obiektów wielowymiarowych i ustala się ich hierarchię. Rozwój baz danych, zarówno pod względem ich złożoności, jak i objętości, powoduje, że wyzwaniem staje się nie tylko efektywne przechowywanie takich danych, lecz także ich analiza, zdolność interpretacji i wyciągania użytecznych wniosków. W efekcie, analizując duże zbiory danych, trudno wyodrębnić grupy obiektów na podstawie dendrogramów, a uzyskana struktura staje się często nieczytelna i niemożliwa do praktycznego zastosowania. Poważnym problemem jest także rozmiar samej macierzy odległości, którą poddaje się analizie. Jeszcze innym problemem uniemożliwiającym przeprowadzenie klasycznego grupowania hierarchicznego jest występowanie w macierzy danych licznych braków.

2. Proponowane rozwiązanie problemu

Jednym z możliwych rozwiązań pojawiających się problemów jest połączenie ze sobą zalet metod grupowania hierarchicznego z zaletami innych metod grupo-

wania. Propozycją taką są hybrydowe metody grupowania dużych zbiorów danych, łączące sztuczną sieć neuronową typu SOM (*self organizing map*) z klasycznymi metodami grupowania oraz metody optymalizacyjne z klasycznymi metodami grupowania (por. [Migdał-Najman 2007b]). Celem proponowanych badań jest prezentacja i opis dwustopniowych hybrydowych metod grupowania obiektów w rozpoznawaniu hierarchicznej struktury skupień. Badanie zostanie oparte na danych symulacyjnych. Do wyodrębniania skupień zostanie zastosowana sztuczna sieć neuronowa typu SOM oraz metody optymalizacyjne: *k*-średnich i *k*-medoids. Hierarchiczna struktura skupień uzyskanych z wykorzystaniem powyższych metod będzie następnie analizowana klasycznymi metodami aglomeracyjnymi. Sztuczna sieć neuronowa, metoda *k*-średnich oraz *k*-medoids nie dają w wyniku grupowania struktury hierarchicznej skupień. Posłużą jednak jako preprocesor, przygotowując dane dla metod aglomeracyjnych. Grupowanie hierarchiczne odbywać się będzie nie na danych pierwotnych, lecz na neuronach sieci, środkach ciężkości czy medoidach. Proponowane podejścia pozwolą na redukcję liczby obiektów do znacznie mniejszej liczby prototypowych skupień. Analizowane metody hybrydowe powinny zachować podstawowe zalety metod hierarchicznych (uporządkowaną strukturę skupień), redukując podstawowe wady związane z rozmiarem macierzy danych i nieczytelnością dendrogramu. Na podstawie mierników oceny podobieństwa wyników dwóch klasyfikacji dokonane zostanie porównanie zgodności klasyfikacji obiektów metodą hierarchiczną i hybrydową.

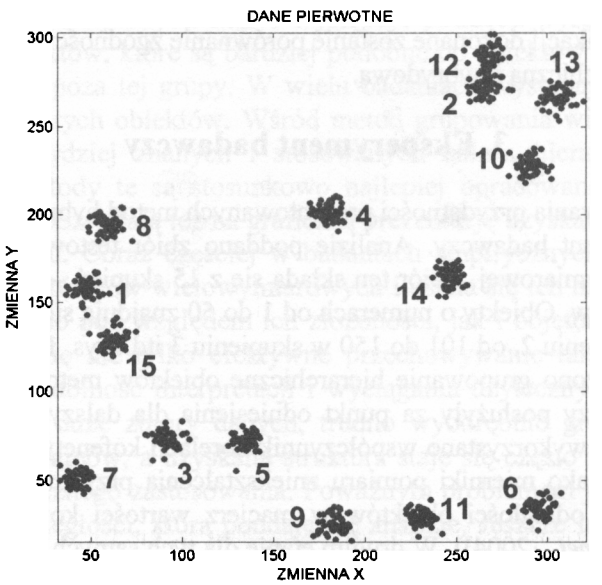
3. Eksperyment badawczy

Do zweryfikowania przydatności prezentowanych metod hybrydowych zaprojektowano eksperyment badawczy. Analizie poddano zbiór testowy 750 obiektów z przestrzeni dwuwymiarowej. Zbiór ten składa się z 15 skupień – w każdym skupieniu jest 50 obiektów. Obiekty o numerach od 1 do 50 znajdują się w skupieniu 1, od 51 do 100 w skupieniu 2, od 101 do 150 w skupieniu 3 itd. (rys. 1). Na etapie pierwszym przeprowadzono grupowanie hierarchiczne obiektów metodą średniej grupowej. Wyniki analizy posłużyły za punkt odniesienia dla dalszych etapów badawczych. W badaniu wykorzystano współczynnik korelacji kofenetycznej i sumę kwadratów odchyleń jako mierniki pomiaru zniekształcenia przy transformacji z wyjściowej macierzy odległości obiektów w macierz wartości kofenetycznych (por. [*Metody statystycznej*... 2004]). W drugim etapie dla tych samych 750 obiektów zbudowano sieć neuronową typu SOM o wymiarach 40×1 neuronów. W rezultacie uzyskano łańcuch neuronów i strukturę grupową obiektów wejściowych. Następnie w klasyczny sposób zbudowano hierarchię wyznaczonych skupień, opierając się na łańcuchu neuronów. Na etapie trzecim i czwartym zbiór testowy danych (750 obiektów) grupowano na 2 do 25 skupień metodą *k*-średnich i metodą *k*-medoids. Dla każdego zadanego podziału wyznaczono wartości wskaźnika jakości grupowania – tzw. wskaźnik sylwetkowy (*silhouette index*) – por. [Rousseeuw 1987; Migdał-Naj-

man, Najman 2006; Migdał-Najman 2006]). Liczbę skupień, która maksymalizuje wartość wskaźnika sylwetkowego uznano za optymalną liczbę skupień. Następnie w klasyczny sposób zbudowano hierarchię wyznaczonych skupień, opierając się na ich punktach ciężkości i medoidach. W etapie piątym porównano struktury hierarchiczne skupień i stopień podobieństwa uzyskanych klasyfikacji.

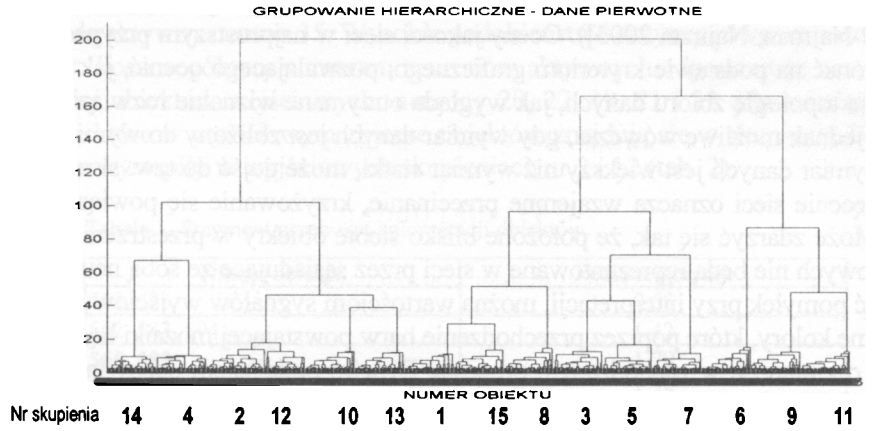
4. Wyniki badań

Kierując się kryterium maksymalnej wartości współczynnika korelacji kofenetycznej i minimalnej wartości sumy kwadratów odchyleń, dokonano wyboru metody klasyfikacji. Wybrano metodę średniej grupowej i odległość Czebyszewa. W wyniku grupowania hierarchicznego uzyskano dendrogram (rys. 2)[1]. Uzyskana struktura skupień jest zgodna z subiektywną oceną danych zbioru testowego. W zbiorze tym wyróżniono 15 skupień, które będą stanowić punkt odniesienia dla struktur uzyskanych w trzech proponowanych dwustopniowych hybrydowych metodach grupowania (sieć SOM + grupowanie hierarchiczne, metoda *k*-średnich + grupowanie hierarchiczne, metoda *k*-medoids + grupowanie hierarchiczne).

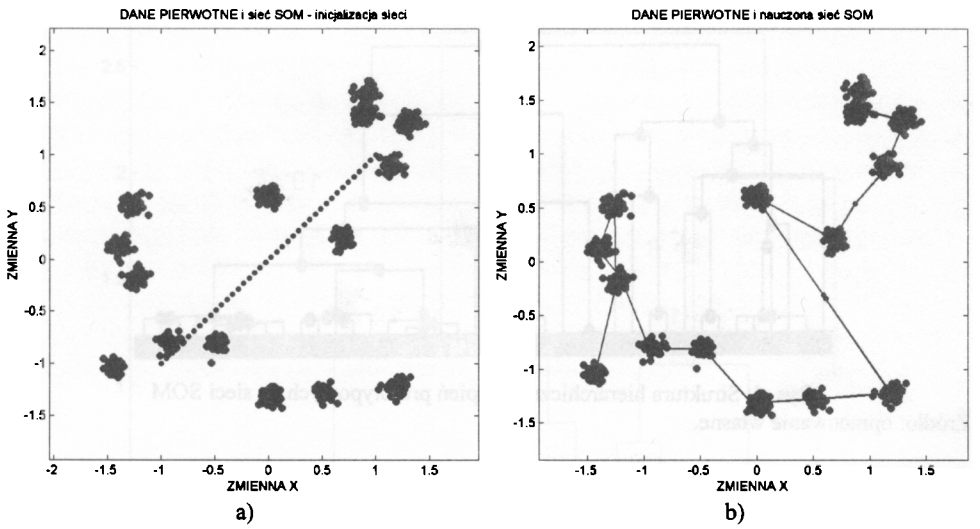


Rys. 1. Wykres rozrzutu zbioru testowego i numeracja skupień
Źródło: opracowanie własne.

[1] Wszystkie obliczenia wykonano w środowisku Matlab.



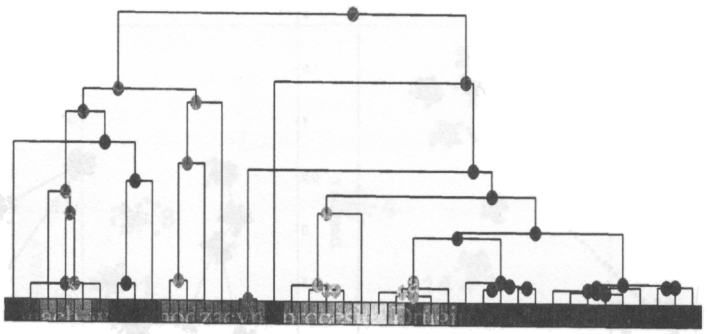
Rys. 2. Dendrogram hierarchii skupień zbioru testowego
Źródło: opracowanie własne.



Rys. 3. Obiekty wejściowe i sieć SOM a) przed samouczeniem się, b) po procesie samouczenia się sieci
Źródło: opracowanie własne.

Następnie zastosowano pierwszą dwustopniową metodę hybrydową (sieć SOM + grupowanie hierarchiczne). Na pierwszym stopniu zbudowano sieć neuronową typu SOM o wymiarach 40×1 neuronów, prostokątnej strukturze połączeń neuronów i funkcji sąsiedztwa typu *bubbel* (por. [*Sieci neuronowe*... 2000]). Sieć przed uczeniem metodą Kohonena typu *batch-train* prezentuje rys. 3a, a po uczeniu rys. 3b. Jakość

uzyskanej sieci mierzono błędem topologicznym i dystorsji (por. [Kohonen 1997; Migdał-Najman, Najman 2003]). Oceny jakości sieci w najprostszym przypadku można dokonać na podstawie kryterium graficznego, pozwalającego ocenić, jak mapa zachowuje topologię zbioru danych, jak wygląda otrzymane wizualne rozwiązanie sieci. Jest to jednak możliwe wówczas, gdy wymiar danych jest zbliżony do wymiaru siatki. Gdy wymiar danych jest większy niż wymiar siatki, może dojść do tzw. skręcania sieci. Skręcenie sieci oznacza wzajemne przecinanie, krzyżowanie się powiązań neuronów. Może zdarzyć się tak, że położone blisko siebie obiekty w przestrzeni sygnałów wejściowych nie będą reprezentowane w sieci przez sąsiadujące ze sobą neurony. Aby uniknąć pomyłek przy interpretacji, można wartościom sygnałów wyjściowych przypisać różne kolory, które poprzez przechodzenie barw powstającej mozaiki będą sygnalizować o zmianach w zachodzącym procesie. Drugim podejściem oceny jakości sieci jest kryterium formalne, oparte na współczynnikach (miarach) pozwalających ocenić stopień odwzorowania obiektów wejściowych przez sieć. Na rys. 4 zaprezentowano dendrogram zbudowany na neuronach sieci SOM. Na rys. 3b neurony 2 i 6 (miejsce skręcenia sieci) odwzorowują obiekty skupienia 15. Na rys. 4 neuron 2 i 6 (drugi i szósty kwadrat od lewej) łączą się ze sobą, tworząc skupienie (neurony te mają podobne sygnały wyjściowe wyrażone zbliżoną kolorystyką).



Rys. 4. Struktura hierarchiczna skupień prototypowych na sieci SOM
Źródło: opracowanie własne.

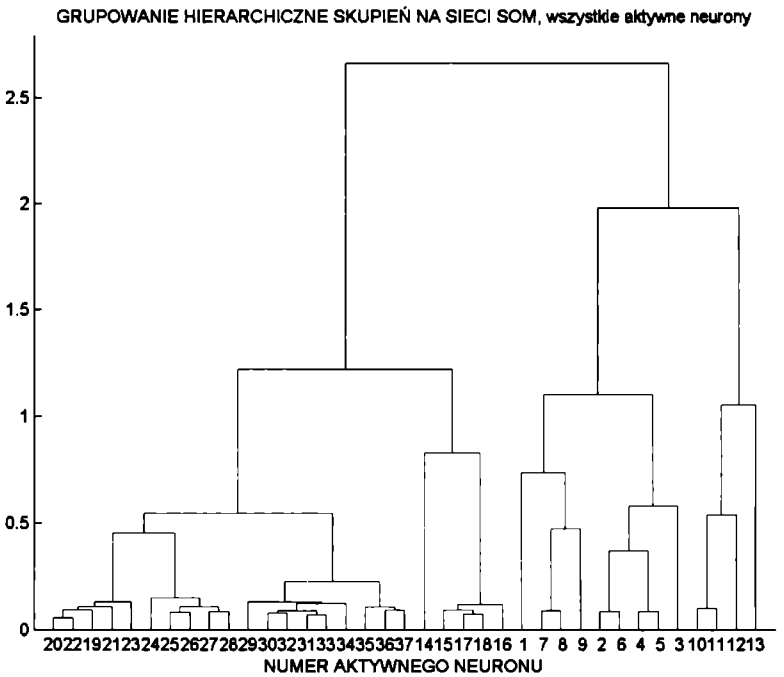
Jedną z technik wizualizacji sieci SOM jest histogram pobudzeń, który pozwala zaobserwować, ile obiektów wejściowych jest rozpoznawanych (por. [Deboeck, Kohonen 1998]) i reprezentowanych przez poszczególne neurony sieci SOM. Na przykład neuron 40 na histogramie pobudzeń, tj. 37 neuron aktywny (3 neurony nie odwzorowują obiektów zbioru testowego), odpowiada za identyfikację następujących obiektów: 562, 565, 568, 576, 577, 583, 584, 589 i 594 zbioru testowego. Obiekty wejściowe opisane przez neuron 40 (jeden neuron) stanowią skupienie prototypowe. Jednocześnie obiekty te są do siebie najbardziej podobne. Sposób łączenia tych obiektów przedstawia fragment przebiegu aglomeracji (tab. 1). Jest tu wysoka zgodność między wynikami grupowania hierarchicznego a grupowaniem na sieci

SOM. Obiekty 565, 589 w przebiegu aglomeracji połączyły się z obiektami, za które odpowiada aktywny neuron 35. Wysokie podobieństwo dotyczy również pozostałych obiektów. Poszczególnym numerom neuronów aktywnych odpowiadają następujące liczebności obiektów wejściowych: neurony 20, 22, 19, 21 i 23 odpowiadają za 50 obiektów, które w wyjściowym zbiorze obiektów znajdują się w skupieniu 10. Pewne różnice wystąpiły w wyjściowych skupieniach nr 2 i 12 (tab. 2).

Tabela 1. Fragment przebiegu aglomeracji obiektów

Numer obiektu	Poziom łączenia
562, 583	0,68
562, 583, 594	1,42
565, 589	1,79
576, 584	1,79
562, 583, 594, 568	2,37
562, 583, 594, 568, 577	3,29
562, 583, 594, 568, 577, 576, 584	5,66

Źródło: opracowanie własne.



Rys. 5. Dendrogram hierarchii neuronów aktywnych na sieci SOM
Źródło: opracowanie własne.

Na drugim stopniu (pierwszej dwustopniowej hybrydowej metody grupowania) obiekty przygotowane w powyższy sposób pogrupowano, korzystając z grupowania

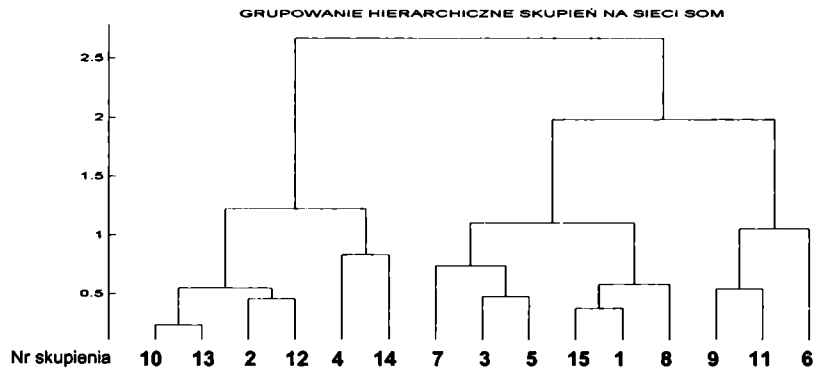
Tabela 2. Przyporządkowanie liczby mapowanych obiektów do skupień prototypowych

Numer neuronu aktywnego						20	22	19	21	23	24	25	26	27	28	29	30	32
Liczba mapowanych obiektów						10	14	8	15	3	4	17	11	6	12	7	10	11
Numer skupienia						SK10 $n_{10}=50$					SK13 $n_{13}=50$					SK2		
31	33	34	35	36	37	14	15	17	18	16	1	7	8	9	2	6		
8	8	15	19	13	9	50	6	19	10	15	50	24	26	50	27	23		
SK2 $n_2=59$			SK12 $n_{12}=41$			SK4	SK14 $n_{14}=50$				SK7	SK3		SK5	SK15			
4	5	3	10	11	12	13												
18	32	50	17	33	50	50	np. SK7 – skupienie numer 7											
SK1		SK8	SK9		SK11	SK6												

Źródło: opracowanie własne.

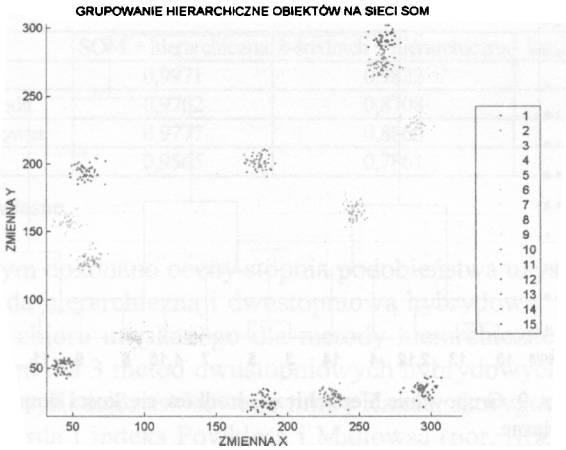
hierarchicznego. Grupowane były wszystkie wyróżnione skupienia prototypowe. Kierując się kryterium maksymalnej wartości współczynnika korelacji kofenetycznej i minimalnej wartości sumy kwadratów odchyleń, dokonano wyboru metody klasyfikacji. Wybrano metodę średniej grupowej i kwadrat odległości euklidesowej. Na rys. 5 zaprezentowano dendrogram pozwalający na wyróżnienie 15 skupień, a na rys. 6 i 7 uzyskaną hierarchię 15 skupień. Klasyfikację i liczbę odwzorowanych obiektów wejściowych przez neurony aktywne uzyskane metodą hybrydową zaprezentowano w tab. 2.

Na etapie trzecim zastosowano drugą dwustopniową hybrydową metodę grupowania (metodę *k*-średnich + grupowanie hierarchiczne). Na pierwszym stopniu metody hybrydowej podzielno zbiór testowy danych na 2 do 25 skupień metodą *k*-średnich. Dla każdego zadanego podziału wyznaczono wartości wskaźnika jakości grupowania (wskaźnika sylwetkowego). Największą wartość w oparciu o wskaźnik sylwetkowy uzyskano dla podziału na 13 skupień (rys. 8). Na drugim stopniu zbudowano hierarchię wyznaczonych skupień, opierając się na ich punktach ciężkości (rys. 9). Kierując się kryterium maksymalnej wartości współczynnika korelacji kofenetycznej i minimalnej wartości sumy kwadratów odchyleń, dokonano wyboru metody klasyfikacji. Wybrano metodę średniej grupowej i kwadrat odległości euklidesowej.



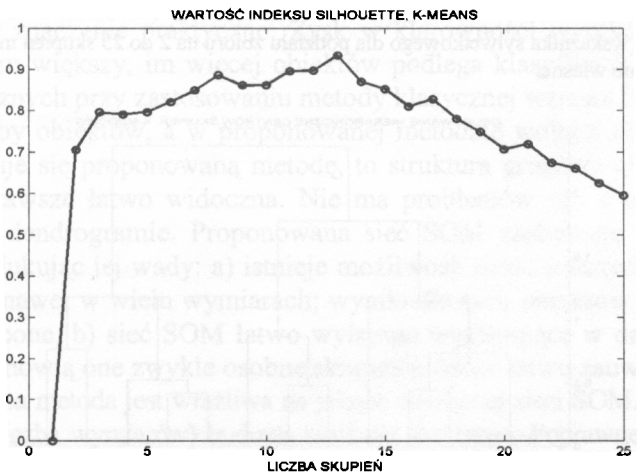
Rys. 6. Grupowanie hierarchiczne skupień prototypowych na sieci SOM

Źródło: opracowanie własne.

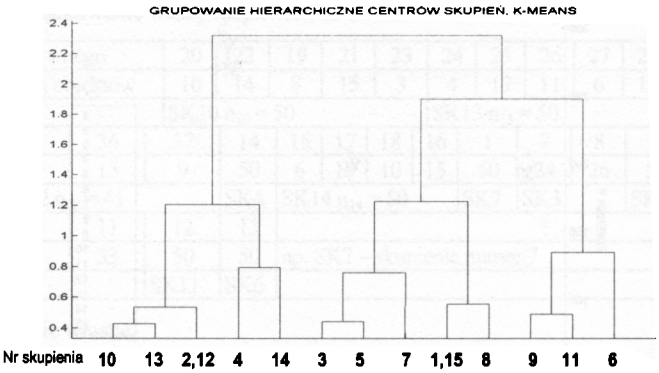


Rys. 7. Klasyfikacja obiektów do skupień uzyskanych dla sieci SOM i grupowania hierarchicznego
Źródło: opracowanie własne.

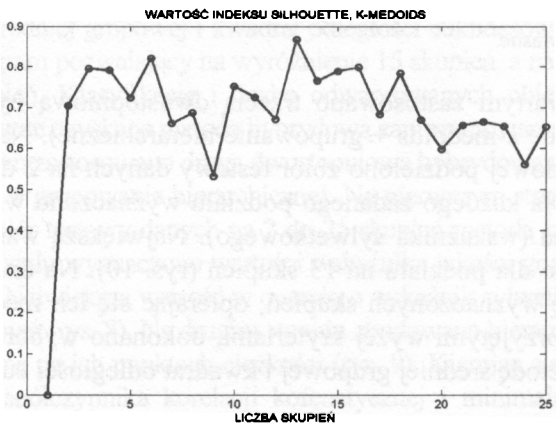
Na etapie czwartym zastosowano trzecią dwustopniową hybrydową metodę grupowania (metodę *k*-medoids + grupowanie hierarchiczne). Na pierwszym stopniu metody hybrydowej podzielono zbiór testowy danych na 2 do 25 skupień metodą *k*-medoids. Dla każdego zadanego podziału wyznaczono wartości wskaźnika jakości grupowania (wskaźnika sylwetkowego). Największą wartość w oparciu o wskaźnik uzyskano dla podziału na 13 skupień (rys. 10). Na drugim stopniu zbudowano hierarchię wyznaczonych skupień, opierając się ich medoidach (por. rys. 11). Kierując się przyjętymi wyżej kryteriami, dokonano wyboru metody klasyfikacji. Wybrano metodę średniej grupowej i kwadrat odległości euklidesowej.



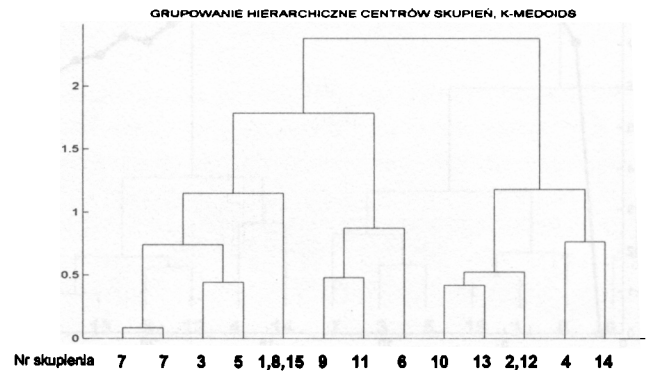
Rys. 8. Wartości wskaźnika sylwetkowego dla podziału zbioru na 2 do 25 skupień metodą *k*-średnich
Źródło: opracowanie własne.



Rys. 9. Grupowanie hierarchiczne środków ciężkości skupień
Źródło: opracowanie własne.



Rys. 10. Wartości wskaźnika sylwetkowego dla podziału zbioru na 2 do 25 skupień metodą *k*-medoids
Źródło: opracowanie własne.



Rys. 11. Grupowanie hierarchiczne medoid skupień
Źródło: opracowanie własne.

Tabela 3. Wskaźniki oceny podobieństwa wyników klasyfikacji

Wskaźnik	SOM + hierarchiczna	*k*-średnich + hierarchiczna	*k*-medoids + hierarchiczna
Miara Randa	0,9971	0,9822	0,9622
Skorygowana miara Randa	0,9762	0,8708	0,7502
Indeks Fowklesa i Mallowsa	0,9777	0,8866	0,7863
Współczynnik Jaccarda	0,9565	0,7861	0,6259

Źródło: opracowanie własne.

Na etapie piątym dokonano oceny stopnia podobieństwa uzyskanych klasyfikacji obiektów metodą hierarchiczną i dwustopniową hybrydową: przynależności do skupień badanego zbioru uzyskanego dla metody hierarchicznej i przynależności obiektów do skupień dla 3 metod dwustopniowych hybrydowych. Do oceny podobieństwa dwóch klasyfikacji zastosowano miarę Randa, skorygowaną miarę Randa, współczynnik Jaccarda i indeks Fowklesa i Mallowsa (por. [Rand 1971; Fowkles, Mallows 1983; Migdał-Najman 2007]). W tab. 3 przedstawiono udziały prawidłowo sklasyfikowanych obiektów do skupień dla trzech analizowanych dwustopniowych metod hybrydowych. Najwyższą zgodność porównywanych klasyfikacji uzyskano dla pierwszej dwustopniowej hybrydowej metody łączącej sztuczną sieć neuronową typu SOM z hierarchiczną metodą grupowania.

5. Wnioski

Uzyskana struktura hierarchiczna skupień podejścia klasycznego i opartego na sieci SOM jest identyczna. Uzyskana struktura hierarchiczna obiektów jest niemal identyczna ze strukturą uzyskaną z wykorzystaniem sieci SOM. Powstają małe różnice, ponieważ grupowanie na sieci SOM odbywa się nie na obiektach, a na „metaskupieniach", tworzących się przy poszczególnych neuronach. Skala różnic ma niewielkie znaczenie praktyczne. Zysk w klarowności wyników i szybkości analizy jest tym większy, im więcej obiektów podlega klasyfikacji. Skala problemów numerycznych przy zastosowaniu metody klasycznej wzrasta liniowo wraz ze wzrostem liczby obiektów, a w proponowanej metodzie wolniej niż logarytmicznie. Jeśli stosuje się proponowaną metodę, to struktura grupowa obiektów (jeżeli istnieje) jest zawsze łatwo widoczna. Nie ma problemów np. z poszukiwaniem „odcięcia" na dendrogramie. Proponowana sieć SOM zachowuje zalety metody klasycznej, redukując jej wady: a) istnieje możliwość analizy danych z występującymi brakami nawet w wielu wymiarach; wyniki dla tych obiektów mogą być jednak zniekształcone, b) sieć SOM łatwo wyłapuje występujące w danych wartości nietypowe. Stanowią one zwykle osobne skupienie, które łatwo zauważyć.

Proponowana metoda jest wrażliwa na jakość uzyskanej sieci SOM. W dużych badaniach (duża liczba wymiarów) budowa sieci nie jest łatwa. Poprawne działanie sieci zależy od wielu parametrów: zastosowanej metody uczenia, założonej topologii, funkcji sąsiedztwa, zasięgu sąsiedztwa. Uzyskana struktura hierarchiczna skupień podejścia

klasycznego i opartego na środkach ciężkości i medoidach jest wysoka, choć wyraźnie słabsza niż z sieci SOM. Podejście to zachowuje wszystkie wady metod podziałowych, jak wrażliwość na występowanie wartości nietypowych, wrażliwość na warunki startu algorytmu, konieczność eliminacji obiektów z brakami danych.

Literatura

Deboeck G., Kohonen T. (1998), *Visual Explorations in Finance with Self-organizing Maps*, Springer-Verlag, London.

Fowkles E.B., Mallows C.L. (1983), *A Method for Comparing Two Hierarchical Clusterings*, „Journal of the American Statistical Association", 78.

Kohonen T. (1997), *Self-organizing Maps*, Springer Series in Information Sciences, Springer-Verlag, Berlin, Heidelberg, 1997, s. 85.

Metody statystycznej analizy wielowymiarowej w badaniach marketingowych (2004), red. E. Gatnar, M. Walesiak, AE, Wrocław, s. 322-329.

Migdał-Najman K. (2006), *Ocena wyniku grupowania w oparciu o indeks silhouette. Konkurencyjność polskich przedsiębiorstw na rynku UE wybrane aspekty*, Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego nr 2, s. 111-120.

Migdał-Najman K. (2007a), *Propozycja hybrydowej metody grupowania dużych zbiorów danych wykorzystującej sieć Kohonena i taksonomiczne metody grupowania*, [w:] Taksonomia 14, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1169, AE, Wrocław, s. 305-313.

Migdał-Najman K. (2007b), *Charakterystyka mierników oceny podobieństwa wyników podziałów*, Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego nr 3, s. 191-201.

Migdał-Najman K., Najman K. (2003), *Próba zastosowania sieci neuronowej typu SOM w badaniu przestrzennego zróżnicowania powiatów w Polsce*, „Wiadomości Statystyczne" nr 4, s. 72-84.

Migdał-Najman K., Najman K. (2006), *Wykorzystanie indeksu silhouette do ustalania optymalnej liczby skupień*, „Wiadomości Statystyczne" nr 6, str. 1-10.

Rand W.M. (1971), *Objective Criteria for the Evaluation of Clustering Methods*, „Journal of the American Statistical Association", vol. 66, nr 336.

Rousseeuw P.J. (1987), *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis*, „J. Comput. Appl. Math." 20.

Sieci neuronowe (2000), red. W. Duch, J. Korbicz, L. Rutkowski, R. Tadeusiewicz, PAN, Akademicka Oficyna Wydawnicza EXIT, Warszawa, s. 183.

COMPARATIVE ANALYSIS OF HIERARCHICAL STRUCTURES OF CLUSTERS BASED ON HYBRID CLUSTERING METHODS

Summary

The article is mainly designed to study the effect of joining the hierarchical agglomerative clustering with the neural network type of SOM (Self Organizing Map), k-means algorithm and k-medoids algorithm. First, the original data set is represented using a smaller set of prototype clusters through neurons SOM, centroids and medoids which allow the efficient use of hierarchical agglom-

erative clustering to divide the prototypes into groups. The reduction of the computational cost is especially important for hierarchical algorithms allowing clusters of arbitrary size and shape. Second, the hybrid methods allow a rough visual presentation, classify original data to clusters and interpretation of the clusters. The clustering results using hybrid methods as an intermediate step were also comparable with the results obtained directly from the data.