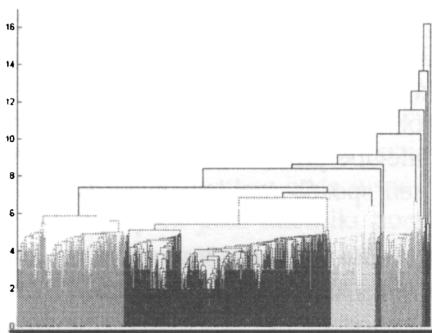


Kamila Migdał-Najman
Uniwersytet Gdański

PROPOZYCJA HYBRYDOWEJ METODY GRUPOWANIA DUŻYCH ZBIORÓW DANYCH WYKORZYSTUJĄCEJ SIĘĆ KOHONENA I TAKSONOMICZNE METODY GRUPOWANIA

1. Wstęp

Jednym z celów analizy wielowymiarowych zbiorów danych jest wyróżnienie homogenicznych grup obiektów. Grupy te skupiają obiekty, które ze względu na badane cechy są do siebie podobne. Wyróżnienie takich grup stanowi istotę analizy skupień i jest zagadnieniem dobrze znanym. Kolejnym celem analizy danych wielowymiarowych jest ustalenie hierarchii obiektów. Wśród metod klasyfikacji szczególne znaczenie zajmują hierarchiczne metody aglomeracyjne, które są stosunkowo najlepiej opracowane pod względem metodologicznym. Metody te odznaczają się cennymi walorami praktycznymi. Do zalet hierarchicznych metod aglomeracyjnych zaliczymy możliwość kontrolowania procesu klasyfikacji, a uzyskana hierarchia umożliwia dokładne określenie, jak są wzajemnie usytuowane poszczególne klasy oraz obiekty w nich zawarte. Dendrogram (rys. 1), będący graficzną formą prezentacji wyników klasyfikacji, pozwala na obserwowanie kolejności połączeń między klasami, co ma duże znaczenie praktyczne (por.: [Gatnar, Walesiak 2004]).



Rys. 1. Dendrogram opisujący hierarchię klientów pewnego przedsiębiorstwa
Źródło: opracowanie własne.

W badaniach empirycznych pojawia się coraz większa potrzeba połączenia obu tych aspektów w analizie zbiorów wielowymiarowych. Gwałtownie rozwijająca się informatyzacja życia ekonomicznego powoduje m.in. dynamiczny wzrost zawartości baz danych. Nierzadko bazy te zawierają dziesiątki tysięcy obiektów charakteryzowanych kilkudziesięcioma atrybutami. Jedną z konsekwencji analizy takich zbiorów danych jest to, że ustalenie hierarchii obiektów staje się technicznie trudne do przeprowadzenia, a uzyskana struktura jest często tak nieczytelna, że niemożliwa do interpretacji i praktycznego zastosowania. Jednocześnie badane zbiory danych charakteryzują się dużą liczbą skupień. Ze zbiorami danych tego typu mamy do czynienia zwłaszcza w badaniach marketingowych i rynkowych, a szczególnie w segmentacji klientów dużych instytucji takich jak banki. Odczytywanie dendrogramów uzyskanych z metod aglomeracyjnych staje się w tych warunkach bardzo utrudnione. Tak zwane odcięcie dendrogramu w celu ustalenia skupień może być niemożliwe do wykonania ze względu na wiele bardzo podobnych do siebie obiektów. Poważnym problemem staje się także rozmiar macierzy odległości, którą poddaje się analizie. Już 10 000 obiektów opisanych 100 cechami wymaga 800 Mb pamięci komputera w celu przechowania w tej pamięci macierzy odległości euklidesowych.

2. Proponowane rozwiązanie problemu

Jednym z możliwych rozwiązań obu tych problemów jest połączenie ze sobą zalet metod grupowania hierarchicznego z zaletami sieci neuronowej typu SOM. Sieć tego typu jest bardzo efektywna w analizie dużych zbiorów danych, jednocześnie ma dobre własności grupujące obiekty nawet przy występowaniu dużej liczby braków danych. Nie daje jednak w wyniku grupowania struktury hierarchicznej obiektów (por.: [Najman, Najman 2002; 2003]). Sieć taka mogłaby posłużyć jako preprocesor, wstępnie przygotowując „prototypy” skupień dla metod aglomeracyjnych.

Celem prezentowanych badań jest prezentacja i opis własności dwustopniowej metody grupowania obiektów opartej na powyższej idei. Do wyodrębnienia „prototypów” skupień zostanie użyta sieć neuronowa typu SOM (*self organizing map*), która wykazuje wysoką zdolność uogólniania podobieństwa obiektów wielowymiarowych, a także redukowania wymiarowości badanego problemu (por.: [Vesanto i in. 2000; Najman, Najman 2002; 2003]). Dzięki takiej wstępnej analizie zredukowana zostanie liczba obiektów do znacznie mniejszej liczby „prototypowych” skupień. Redukcji także ulegnie przestrzeń cech, ponieważ sieć SOM będzie jedynie dwuwymiarowa. W ten sposób problem wielkości macierzy danych zostanie znacznie zredukowany.

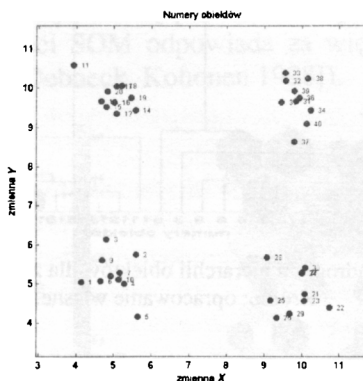
Struktura skupień – prototypów, wyodrębniona za pomocą sieci neuronowej, będzie następnie analizowana klasycznymi metodami hierarchicznymi. Grupowanie hierarchiczne będzie dotyczyło więc nie danych pierwotnych, a sieci neuronów.

W wyniku zastosowania tej procedury uzyska się hierarchiczną strukturę skupień – prototypów.

Z teoretycznego punktu widzenia proponowana metoda powinna zachować podstawowe zalety metod hierarchicznych (uporządkowana struktura podobieństwa obiektów), jednocześnie redukując ich podstawowe wady (nieczytelność dendrogramu, duże rozmiary macierzy odległości) ważne w analizie dużych zbiorów danych. Wadą metody będzie utrata pewnej części szczegółowej informacji o hierarchii wszystkich obiektów. Wynika to z tego, że wszystkie obiekty, które znajdują się po pierwszym etapie w jednym skupieniu – prototypie, choć nie są identyczne, na drugim stopniu analizy będą traktowane jak jeden obiekt.

3. Eksperyment badawczy

Aby zweryfikować hipotezę o przydatności prezentowanej metody, zaprojektowano eksperyment składający się z 3 etapów, oparty na zbiorze 40 obiektów w przestrzeni dwuwymiarowej. Zbiór ten ma czytelną strukturę grupową i hierarchiczną (rys. 2).



Rys. 2. Wykres rozrzutu danych analizowanych w eksperymencie
Źródło: opracowanie własne.

Etap 1. Przeprowadzone będzie klasyczne, hierarchiczne grupowanie obiektów. Za właściwy dla badanego typu danych uznany zostanie taki sposób klasyfikacji, który da minimalne zniekształcenie przy transformacji wyjściowej macierzy obiektów w macierz wartości kofenetycznych (wartości poziomego połączenia obiektów na dendrogramie). W badaniu wykorzystane będą współczynnik korelacji kofenetycznej i suma kwadratów odchyłeń jako mierniki pomiaru zniekształcenia przy transformacji z jednej macierzy na drugą. Uzyskana struktura hierarchiczna i skupień posłuży za punkt odniesienia dla drugiego etapu eksperymentu.

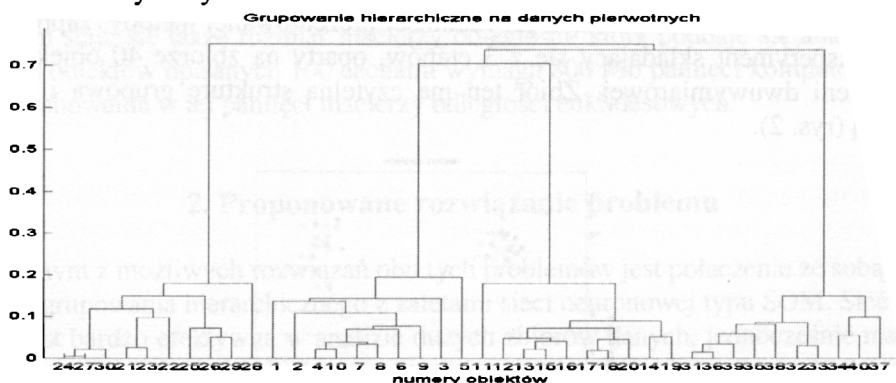
Etap 2. Dla tych samych obiektów zbudowana zostanie sieć SOM o małych wymiarach. Sieć optymalizowana będzie tak, aby błędy topograficzny i dystorsji

były jak najmniejsze (por.: [Najman, Najman 2003]). W efekcie powstanie pewna liczba skupień prototypowych, które powinny być złożone z obiektów najbardziej do siebie podobnych. Następnie uzyskane skupienia – prototypy będą grupowane klasyczną metodą grupowania hierarchicznego.

Etap 3. Porównane zostaną wyniki uzyskanych klasyfikacji.

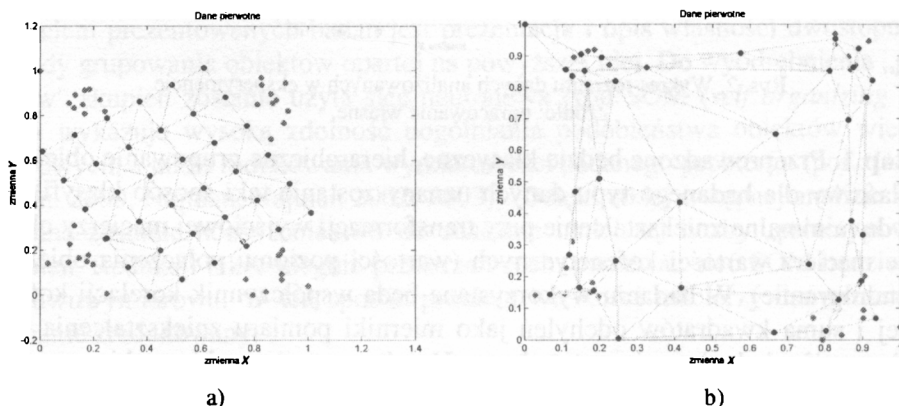
4. Wyniki badań

W etapie pierwszym, kierując się kryterium maksymalnej wartości współczynnika korelacji kofenetycznej i minimalnej wartości sumy kwadratów odchyłeń, dokonano wyboru metody klasyfikacji opartej na metodzie średniej klasowej i odległości Czebyszewa. W wyniku grupowania hierarchicznego uzyskano dendrogram zaprezentowany na rys. 3.



Rys. 3. Dendrogram hierarchii obiektów dla zbioru testowego

Źródło: opracowanie własne.



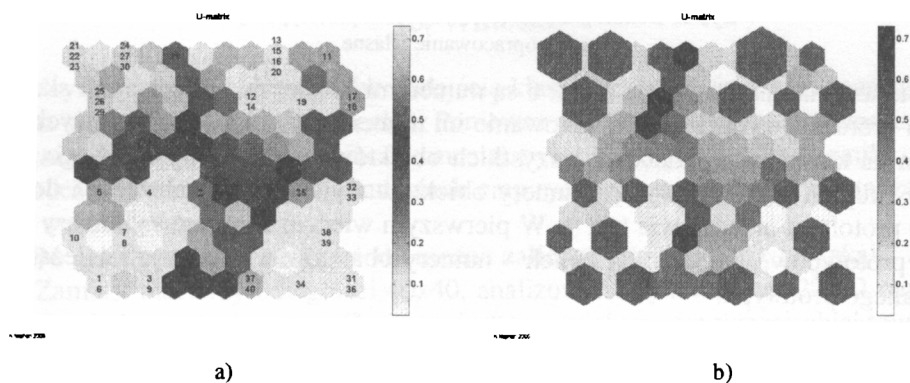
Rys. 4. Obiekty wejściowe i sieć SOM a) przed samouczeniem się, b) po procesie samouczenia się sieci

Źródło: opracowanie własne.

Uzyskana struktura jest całkowicie zgodna z subiektywną oceną danych. W zbiorze tym wyróżniono 4 wyraźne skupienia zawierające po 10 obiektów, które będą stanowić punkt odniesienia dla struktury uzyskanej z metody dwustopniowej.

W etapie drugim zastosowano metodę dwustopniową. Na pierwszym stopniu zbudowano sieć neuronową typu SOM o wymiarach 6x6 neuronów, o heksagonalnej strukturze połączeń neuronów i gaussowskiej funkcji sąsiedztwa (por.: [Duch 2000]). Sieć przed uczeniem metodą Kohonena typu *batch-train* prezentuje rys. 4a, a po uczeniu – rys. 4b. Strukturę sieci optymalizowano tak, aby zminimalizować błędy topograficzne i dystorsji (por.: [Kohonen 1997]).

Poza oceną parametryczną jakości uzyskanej sieci dokonuje się także jej wizualizacji. Najczęściej stosowaną techniką wizualizacji sieci SOM jest **macierz U** (*unified distance matrix, U-matrix*), nazywana macierzą ujednoliconych odległości (rys. 5a). Odległość między sąsiadującymi neuronami zaznaczono różnymi odcieniami. Można zauważyć, że odległości w macierzy U są tego samego rzędu jak na dendrogramie. **Histogram pobudzeń** (*hit histogram*), nazywany również diagramem pobudzeń (*diagram hit's*), pozwala zaobserwować, ile obiektów wejściowych rozpoznawanych i reprezentowanych jest przez poszczególne neurony na sieci SOM, tzn., które neurony są aktywniejsze, częściej pobudzane od pozostałych neuronów. Wielkość wyróżnionych sześciokątów wskazuje na siłę pobudzenia (rys. 5b). Im sześciokąt jest większy, tym dany neuron sieci SOM odpowiada za większą liczbę obiektów z przestrzeni wejściowej (por.: [Deboeck, Kohonen 1998]).



Rys. 5. Ocena sieci SOM a) macierz U, b) histogram pobudzeń

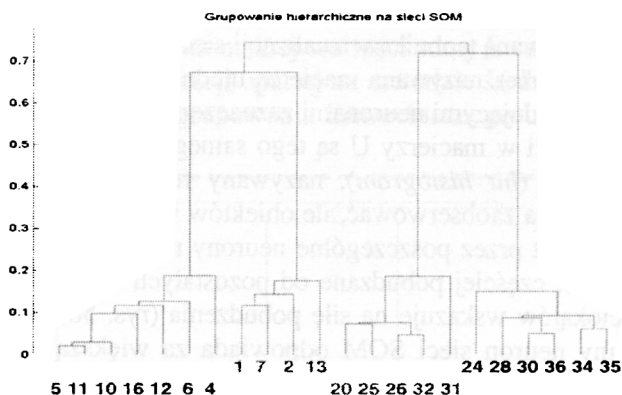
Źródło: opracowanie własne.

Obiekty wejściowe opisywane przez jeden neuron stanowią skupienie prototypowe. Takimi skupieniami prototypowymi są np. obiekty (21,22,23), (25,26,29), (24,27,30) i obiekt 28. Leżą one na macierzy U, odpowiednio w tych samych miejscach, co oznacza, że ten sam neuron odpowiada za identyfikację każdego obiektu w danym skupieniu prototypowym. Jednocześnie na rys. 3 widzimy, że obiekty te są do siebie najbardziej podobne i tak właśnie łączą się kolejno, tworząc w sumie

jedno skupienie. Jest tu pełna zgodność między wynikiem grupowania hierarchicznego a grupowaniem opartym na sieci SOM. Dotyczy to także wszystkich pozostałych skupień.

Na drugim stopniu obiekty przygotowane w ten sposób zostały pogrupowane z wykorzystaniem hierarchicznego grupowania obiektów. Grupowane były wszystkie wyróżnione skupienia prototypowe.

Podobnie jak na etapie pierwszym zastosowano metodę średniej klasowej z odległością Czebyszewa. W wyniku grupowania hierarchicznego uzyskano dendrogram zaprezentowany na rys. 6.



Rys. 6. Dendrogram hierarchii skupień prototypowych

Źródło: opracowanie własne.

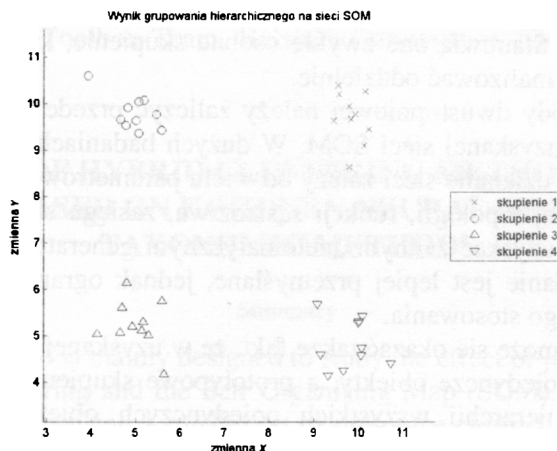
Numerzy na dendrogramie z rys. 6 są numerami kolejnymi nadawanymi skupieniom prototypowym. Przyporządkowanie im numerów obiektów wyjściowych pozwala na klasyfikację końcową wszystkich obiektów ze zbioru wyjściowego. Numery skupień prototypowych i numery obiektów wyjściowych należących do danego prototypu przedstawia tab. 1. W pierwszym wierszu zapisane są numery skupień prototypów, pod każdym z nich – numery obiektów wejściowych należących do danego prototypu.

Tabela 1. Przyporządkowanie numerów obiektów wejściowych do skupień prototypowych

1	2	4	5	6	7	10	11	12	13	16	20	24	25	26	28	30	31	32	34	35	36
21	25	5	10	1	24	4	7	3	28	2	12	37	13	19	35	34	11	17	32	38	31
22	26			6	27		8	9			14	40	15					18	33	39	36
23	29				30									16							
														20							

Źródło: opracowanie własne.

Prezentowany dendrogram (rys. 6) pozwala na wyróżnienie 4 skupień, a wyniki w tab. 1 umożliwiają dokonanie klasyfikacji wszystkich obiektów wejściowych do klas.



Rys. 7. Klasyfikacja obiektów wejściowych oparta na metodzie dwustopniowej
Źródło: opracowanie własne.

Klasyfikacja obiektów metodą dwustopniową jest identyczna z klasyfikacją uzyskaną klasyczną metodą grupowania hierarchicznego (rys. 7).

5. Wnioski

Uzyskana zgodność wyników podejścia klasycznego i dwustopniowego oparte-
go na sieci SOM jest niemal idealna. Ponieważ prezentowany przykład miał służyć
jedynie ilustracji działania metody, musiał być bardzo prosty. Jak wynika z do-
świadczeń autorki, w dużych badaniach empirycznych ta zgodność jest także bar-
dzo wysoka¹.

Do zalet prezentowanej metody można zaliczyć wyraźny zysk na „skali” anali-
zy. Zamiast macierzy odległości 40x40, analizowana była macierz 22x22 neurony
(pozostałe były nieaktywne). Zysk ten jest tym większy, im więcej obiektów pod-
lega klasyfikacji. Skala problemów numerycznych przy zastosowaniu metody kla-
sycznej wzrasta liniowo wraz ze wzrostem liczby obiektów, a w proponowanej me-
todzie wolniej niż logarytmicznie. Dzięki zastosowaniu proponowanej metody
zawsze łatwo widoczna będzie struktura grupowa obiektów (jeżeli istnieje). Nie ma
problemów z poszukiwaniem „odcięcia” na dendrogramie.

W wyniku analizy uzyskujemy strukturę hierarchiczną obiektów, co w wielu
badaniach empirycznych odgrywa znaczną rolę.

Kolejną zaletą metody dwustopniowej jest możliwość analizy danych z wystę-
pującymi brakami danych, nawet w wielu wymiarach, co jest zupełnie niemożliwe

¹ Bardziej złożone przykłady można znaleźć na stronie WWW autorki: <http://panda.bg.univ.gda.pl/~Najman>.

w podejściu klasycznym. Sieć SOM łatwo identyfikuje występujące w danych wartości nietypowe. Stanowią one zwykle osobne skupienie, które łatwo wyłączyć z dalszej analizy i analizować oddzielnie.

Do wad metody dwustopniowej należy zaliczyć przede wszystkim jej wrażliwość na jakość uzyskanej sieci SOM. W dużych badaniach budowa sieci nie jest łatwa. Poprawne działanie sieci zależy od wielu parametrów: zastosowanej metody uczenia, założonej topologii, funkcji sąsiedztwa, zasięgu sąsiedztwa. Optymalnej sieci nie można uzyskać żadnym „automatycznym generatorem sieci”, co jest jej zaletą, gdyż badanie jest lepiej przemyślane, jednak ogranicza to możliwość jej powszechniejszego stosowania.

Pewną wadą może się okazać także fakt, że w uzyskanej strukturze hierarchicznej mamy nie pojedyncze obiekty, a prototypowe skupienia. Oznacza to, że nie można ustalić hierarchii wszystkich pojedynczych obiektów, gdyż wszystkie obiekty leżące w jednym skupieniu prototypowym będą traktowane jak obiekty identyczne. W niektórych zastosowaniach może to być pewnym ograniczeniem. Z tego powodu także uzyskana struktura hierarchiczna nie jest identyczna ze strukturą uzyskaną dzięki zastosowaniu metody klasycznej. Różnice jednak dotyczą zwykle pojedynczych obiektów bardzo podobnych do siebie.

Reasumując, wydaje się, że proponowana metoda dwustopniowa zachowuje zalety metody klasycznej i redukuje jej wady.

Literatura

- Deboeck G., Kohonen T. (1998), *Visual Explorations in Finance with Self-Organizing Maps*, Springer-Verlag, London.
- Duch W., Korbicz J., Rutkowski L., Tadeusiewicz R. (red.) (2000), *Sieci neuronowe*, PAN, Akademicka Oficyna Wydawnicza EXIT, Warszawa, s. 183.
- Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław, s. 322-329.
- Kohonen T. (1997), *Self-Organizing Maps*, Springer Series in Information Sciences, Springer-Verlag, Berlin Heidelberg, s. 85.
- Najman K., Najman K. (2002), *Zastosowanie sieci neuronowej typu SOM do wyboru najatrakcyjniejszych spółek na WGPW*, Prace Naukowe AE we Wrocławiu nr 952, AE, Wrocław, s. 493-499.
- Najman K., Najman K. (2002), *Zastosowanie sieci neuronowej typu SOM do wyboru najatrakcyjniejszych spółek na WGPW na bazie wskaźników analizy technicznej*, Rynek kapitałowy – skuteczne inwestowanie, t. II, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Międzyzdroje, s. 403-417.
- 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.

Vesanto J., Himberg J., Alhoniemi E., Parhankangas J. (2000), *SOM Toolbox for Matlab 5*, SOM Toolbox Team, Helsinki University of Technology, ESPOO, Finland.

PROPOSAL OF HYBRID CLUSTERING METHOD A LARGE DATASETS BASED ON KOHONEN NEURAL NETWORK AND TAXONOMIC METHODS

Summary

The present article is mainly designed to study the effect of join the hierarchical agglomerative clustering and the Self Organizing Map (SOM). First, the original data set is represented using a smaller set of prototype clusters, which allows efficient use of hierarchical agglomerative 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 2-D grid allows rough visual presentation, classify original data to clusters and interpretation of the clusters. The clustering results using SOM as an intermediate step was also comparable with the results obtained directly from the data.