Kamila Migdał-Najman[*]
Krzysztof Najman[**]

Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej

Wstęp

Analiza skupień jest jednym z ważnych elementów statystycznej analizy danych wielowymiarowych. Wraz ze wzrostem liczby zakładanych baz danych i wzrostem ich objętości znaczenie analizy skupień systematycznie wzrasta. Tworzone zbiory zawierają czasem miliony jednostek opisanych setkami cech. Przykładem takich zbiorów mogą być rejestry operacji na kartach kredytowych, rejestry pytań w wyszukiwarkach internetowych, rejestr tras pokonywanych przez klientów w sklepach wielkopowierzchniowych czy opinie wyrażane na portalach społecznościowych. Obserwuje się coraz więcej zjawisk społeczno-ekonomicznych o coraz bardziej złożonej strukturze. Z tego powodu rosną także wymagania co do metod analizy skupień. Muszą sobie one radzić z dużą liczbą jednostek i cech, a także ich bardzo złożoną strukturą grupową. Za tą potrzebą podążają badania naukowe, poszukując coraz dokładniejszych, szybszych i bardziej oszczędnych metod grupowania danych.

Celem prezentowanych badań jest analiza porównawcza wybranych, klasycznych metod grupowania danych z autorskimi, hybrydowymi metodami opartymi na samouczących się sieciach neuronowych, takich jak: mapa samoorganizująca się (*Self Organizing Map*, SOM) i gaz neuronowy o zmiennej strukturze (*Growing Neural Gas*, GNG).

1. Klasyczne i hybrydowe metody grupowania

R.C. Tryon w 1939 roku jako pierwszy zaproponował pojęcie „analiza skupień" (*cluster analysis*), wydając książkę pod tym samym tytułem [Tryon, 1939]. Pojęcie to zostało wówczas przyjęte i zaakceptowane przez naukowców amerykańskich w naukach niebiologicznych. Analiza skupień oznacza współcześnie różnego rodzaju techniki i procedury numeryczne (rodzinę statystycznych metod klasyfikacji), które pozwalają na poszukiwanie i ocenianie wyodrębnionych skupień, tworzenie klasyfikacji i eksplorację danych.

Caralus Linnaeus w 1737 roku w opublikowanej pracy pt. *Genera Plantarum* napisał, że cała wiedza, jaką posiadamy o interesujących nas jednostkach, zależy od zastosowanych do ich obserwacji metod. Stosując metody analizy skupień, jesteśmy w stanie wyróżnić grupy podobnych do siebie jednostek. Im badane zbiory jednostek są liczniejsze i bardziej złożone, tym istotniejsze staje się tworzenie takich metod [Everitt i inni, 2011]. Według A. Hardy'ego to właśnie wybór właściwej metody, procedury grupowania ma istotny wpływ na odkrycie właściwej liczby skupień w zbiorze danych [Hardy, 1996].

W literaturze przedmiotu proponuje się różne systematyki metod grupowania: Jardine, Sibson [1968], Bergan [1971], Anderberg [1973], Sneath, Sokal [1973], Milligan [1980],

[*] Dr, Katedra Statystyki, Wydział Zarządzania, Uniwersytet Gdański, kmn@wzr.ug.edu.pl
[**] Dr, Katedra Statystyki, Wydział Zarządzania, Uniwersytet Gdański, knajman@wzr.ug.edu.pl

Milligan, Cooper [1987], Marek, Noworol [1987], Pociecha i inni [1988], Jajuga [1990; 1993], Mirkin [1996], Walesiak [1996], Gatnar, Walesiak [2004], Stąpor [2005] i inni. Jedną z syntetycznych systematyk jest podział metod grupowania na cztery grupy: 1) metody hierarchiczne [Ward, 1963], [Lance, Williams, 1966 a, b; 1967 a, b], [Johnson, 1967], [Gordon, 1987], 2) metody podziałowe [Ball, Hall, 1965], [Forgy, 1965], [MacQueen, 1967], [Hartigan, 1975], [Hartigan, Wong, 1979], 3) metody prezentacji graficznej [Greenacre, 1984], [Benzécri, 1992] i 4) metody kombinowane (hybrydowe) [Migdał Najman, 2007; 2008; 2012].

2. Klasyczne i hybrydowe metody grupowania

W prezentowanym badaniu grupę metod hierarchicznych reprezentuje aglomeracyjna metoda Warda (zastosowana z kwadratem odległości euklidesowej), a grupę metod podziałowych metoda k-średnich. Metoda k-średnich i metoda Warda są klasycznymi i najczęściej stosowanymi metodami analizy skupień. Są skuteczne, o ile liczba jednostek nie jest bardzo duża (dla metod hierarchicznych rzędu dziesiątek a podziałowych tysięcy jednostek), skupienia są sferyczne, separowalne, w danych nie występują liczne wartości nietypowe i znana jest istniejąca liczba skupień [Najman, 2008]. W eksperymencie założono, że liczba skupień nie jest *a priori* znana, co jest typowym zjawiskiem w badaniach empirycznych. W celu ustalenia liczby skupień [Milligan, Cooper, 1985; Migdał Najman, Najman, 2005; 2006 a, b; 2008] wykorzystano dwa współczynniki: sylwetkowy (*silhouette coefficient*, *silhouette index*, *SI*) [Rousseeuw, 1987] i Daviesa-Bouldina (*Davies-Bouldin index*, *DB*) [Davies, Bouldin, 1979]. Dla obu powyższych metod dokonywano grupowania na 2 do 12 skupień, każdorazowo wyznaczając wartość obu współczynników. Maksimum wartości współczynnika SI i minimum DB wskazuje optymalną ze względu na te wskaźniki liczbę skupień. Ponieważ uzyskane wskazania mogą różnić się między sobą, rejestrowano jakość grupowania uzyskaną przez obie metody grupowania z liczbą skupień wskazaną przez każdy ze wskaźników.

Podejście hybrydowe reprezentują trzy metody, które powstały z połączenia samoorganizującej się sieci neuronowej typu SOM z metodą hierarchiczną, metodą podziałową i samoorganizującą się siecią nauronową typu GNG. Dla porównania zaprezentowano także wyniki grupowania uzyskane dla samodzielnie zastosowanych sieci samoorganizujących się typu SOM i GNG.

Sieci neuronowe typu SOM i GNG należą do względnie nowych metod analizy skupień, które są dość rzadko stosowane w praktyce[1]. Sieć SOM może być skuteczną metodą analizy skupień, o ile uda się ustalić strukturę sieci konieczną dla rozwiązania postawionego przed nią problemu [Kohonen, 1995; 1997; 2001; Deboeck, 1998; Migdał Najman, 2009]. Kluczowymi parametrami sieci SOM jest liczba neuronów, rozmiar i rodzaj siatki (kwadratowa czy heksagonalna) oraz rodzaj i zasięg sąsiedztwa. Ze względu na brak wiedzy *a priori* o optymalnej strukturze sieci testuje się różne jej struktury i ocenia za pomocą odpowiednich miar. W badaniu wyznaczono wartości: błędu kwantyzacji – *Q*, błędu topograficznego – *T*, błędu dystorsji – *D* i udział martwych neuronów w ogólnej liczbie neuronów w sieci – *VM* [Vesanto i inni, 2002]. Na ich podstawie wyznaczono całkowity błąd sieci jako ich funkcję o postaci:

[1] Wydaje się, że wynika to głównie ze złożoności obliczeniowej algorytmów ich tworzenia i samouczenia się. Z tego powodu brakuje profesjonalnego oprogramowania komputerowego, które pozwoliłoby szerszej grupie badaczy na ich praktyczne stosowanie. Autorzy niniejszego opracowania w swoich publikacjach dotyczących sieci SOM i GNG korzystają z oryginalnego oprogramowania napisanego w środowisku Matlab i biblioteki procedur numerycznych SOM Toolbox.

$$OE = 0,2Q + 0,6T + 0,15D + 0,05VM \quad (1)$$

Wagi zastosowane przy wartościach błędów składowych wynikają z ich ważności dla jakości grupowania. Każdorazowo budowano sieć SOM, stosując cztery funkcje sąsiedztwa: gaussowską, uciętą gaussowską, wykładniczą i prostokątną.

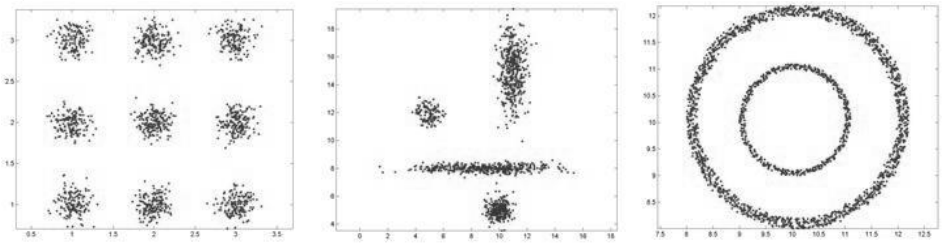
W wyniku procesu samouczenia się sieci uzyskuje się macierz ujednoliconych odległości **U**, która jest podstawą do grupowania jednostek [Migdał Najman, Najman, 2005; 2006 a]. Zwykle grupowania tego dokonuje się w oparciu o graficzną prezentację macierzy **U** na rysunku nazywanym mapą SOM (zobacz rysunek 3a). Ten sposób grupowania posiada tę wadę, że oparty jest na subiektywnym osądzie badacza, który kreśli linie separujące skupienia widoczne na mapie SOM. Aby uniknąć subiektywnego wyznaczania linii podziału mapy SOM na skupienia, proponuje się zastosowanie podejścia hybrydowego. Sieć SOM może być potraktowana jako preprocesor, upraszczający badany problem, redukujący wymiar przestrzeni do dwóch i liczbę jednostek do liczby neuronów, która jest małym ułamkiem liczby jednostek. Na neuronach, które są traktowane jako nowe abstrakcyjne jednostki, można przeprowadzić ponowne grupowanie metodą, która pozwala na obiektywne wskazanie liczby skupień i wyznaczenie granic między nimi. W prezentowanym badaniu zastosowano trzy podejścia hybrydowe: sieć SOM + metoda Warda, sieć SOM + metoda k-średnich i sieć SOM + samoorganizującą się sieć GNG. Dla pierwszych dwóch metod liczbę skupień ustalano ponownie wskaźnikiem SI i DB.

Samoorganizująca się sieć neuronowa typu GNG jest skutecznym narzędziem grupowania danych o dowolnym rozkładzie przestrzennym, o ile skupienia są separowalne [Fritzke, 1994; Najman, 2009; 2010; 2011]. Sieć ta nie wymaga wielu założeń dotyczących jej struktury, ponieważ sama modyfikuje ją w procesie samouczenia się. Założono jedynie parametry kończące proces samouczenia się sieci: maksymalną liczbę neuronów N=100, błąd kwantyzacji Q=0,0001 i maksymalną liczbę iteracji uczących K=1000. Aby zredukować problem z rozdzieleniem skupień nieseparowalnych w eksperymencie, zastosowano zaproponowany przez Najmana [Najman, 2011] wariant algorytmu samouczenia się sieci GNG ze zmiennym krokiem uczenia.

3. Zbiory testowe

Aby zrealizować postawiony w badaniu cel, przygotowano odpowiedni eksperyment badawczy. Wygenerowano trzy zbiory danych o zróżnicowanej liczbie jednostek, skupień i strukturze grupowej. Zbiór pierwszy charakteryzuje się bardzo prostą strukturą grupową o sferycznych i separowalnych skupieniach. Gęstość jednostek w skupieniu rośnie w kierunku centrum skupienia. W danych nie ma szumu ani wartości nietypowych. Zbiór ten jest punktem odniesienia dla dalszych analiz, ponieważ jest wzorcowo łatwym do analizy przypadkiem (zobacz rysunek 1). Zbiór drugi charakteryzuje się bardziej złożoną strukturą grupową. Składa się z czterech nieseparowalnych liniowo skupień. W danych nie ma szumu, ale na krańcach dwóch skupień istnieje pewna liczba wartości nietypowych. Wyróżnienie skupień jest tu znacznie trudniejsze, ponieważ dwa z nich tworzą bardzo rozciągnięte elipsy. Jednostki należące do elipsoidalnych skupień, znajdujące się na ich odległych krańcach, mogą znajdować się w kilkukrotnie większej odległości od centrum skupienia niż jednostki należące do innych skupień (zobacz rysunek 1).

Rysunek 1. Analizowane zbiory testowe



zbiór 1 zbiór 2 zbiór 3

Źródło: Opracowanie własne.

Zbiór trzeci jest złożony z dwóch skupień, które mają kształt koncentrycznych okręgów. Skupienia znacząco różnią się liczbą jednostek i wielkością zajmowanej przestrzeni. Rozdzielenie takich skupień jest bardzo trudne, ponieważ nie posiadają sferycznej struktury i związanej z nią gęstości jednostek. W każdym wymiarze ich rozkład jest względnie równomierny. Jedno skupienie zawiera się w drugim i nie istnieje niedomknięta krzywa pozwalająca odseparować skupienia.

4. Analiza wyników badania

Pierwszy zbiór danych został poprawnie pogrupowany przez wszystkie metody. Skorygowany współczynnik Randa [Rand, 1971] jest w każdym przypadku równy 1. W każdym przypadku udało się poprawnie ustalić liczbę skupień (zobacz tablica 1). Wynik ten nie jest zaskoczeniem. Potwierdza on jedynie, że w przypadku prostych struktur przestrzennych skupień skuteczne są wszystkie metody analizy skupień.

W przypadku drugiego zbioru danych wyniki są bardziej zróżnicowane. Metoda k-średnich wyróżniła skupienia sferyczne, mimo że większość jednostek znajduje się w skupieniach elipsoidalnych. Nie udało się także poprawnie wskazać liczby istniejących skupień. Wskaźnik SI wskazał 3 skupienia a DB aż 9. Najwięcej problemów sprawiały skupienia elipsoidalne, które były błędnie dzielone na kilka mniejszych. Wartość skorygowanego współczynnika Randa wyniosła odpowiednio 0,62 dla podziałów współczynnikiem SI i 0,53 dla podziałów współczynnikiem DB. Uzyskane grupowania przedstawiono na rysunku 2a, 2b. Wyniki uzyskane metodą Warda są podobnej jakości jak uzyskane metodą k-średnich. Oba indeksy liczby skupień błędnie wskazały liczbę skupień, odpowiednio: współczynnik sylwetkowy 2, a DB 8 (zobacz rysunek 2c).

Tablica 1. Analiza porównawcza wyników grupowania

Zbiór	A	B	C	A	B	C
Metoda	Skorygowany współczynnik Randa			Rzeczywista liczba skupień		
				9	4	2
				Wykryta liczba skupień		
k-średnich + SI	1,00	0,62	0,09	9	3	12
k-średnich + DB	1,00	0,53	0,09	9	9	12
Ward + SI	1,00	0,52	0,07	9	2	12
Ward + DB	1,00	0,61	0,07	9	8	12
SOM + k-średnich + SI	1,00	0,62	0,00	9	3	3
SOM + k-średnich + DB	1,00	0,62	0,08	9	3	12
SOM + Ward + SI	1,00	0,65	0,11	9	3	12
SOM + Ward + DB	1,00	0,63	0,11	9	8	12
SOM + GNG	1,00	1,00	1,00	9	4	2

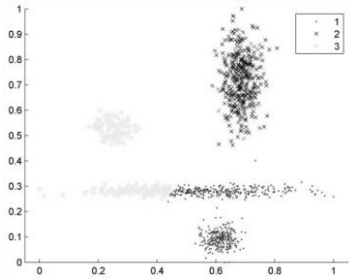
GNG	1,00	0,95	1,00	9	6	2

Źródło: Opracowanie własne.

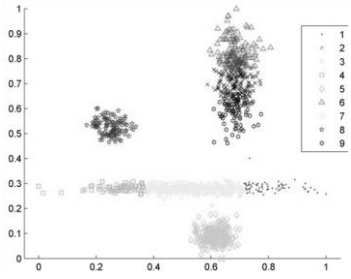
Należy zauważyć, że nawet uważna obserwacja dendrogramu (zobacz rysunek 2d) nie pozwala ustalić poprawnej liczby skupień równej 4. Uzyskane wartości skorygowanego współczynnika Randa są równe 0,52 dla liczby skupień ustalanej współczynnikiem SI i 0,61 dla współczynnikiem DB. Wyniki grupowania metodą Warda przedstawiono na rysunku 2e, 2f.

W kolejnym etapie badania budowano sieci SOM. Testowano sieci o rozmiarze od 4x4 do 20x20 neuronów o połączeniach kwadratowych i heksagonalnych. W każdej próbie testowano cztery funkcje sąsiedztwa: gaussowską, uciętą gaussowską, wykładniczą i prostokątną.

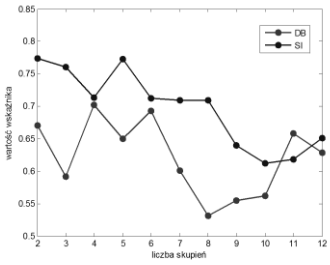
Rysunek 2. Zbiór drugi – wyniki grupowania uzyskane metodą k-średnich i Warda



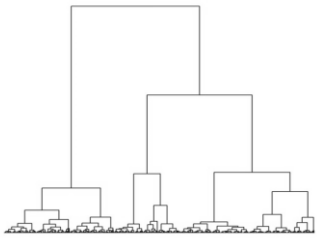
a) wynik grupowania dla metody k-średnich + SI



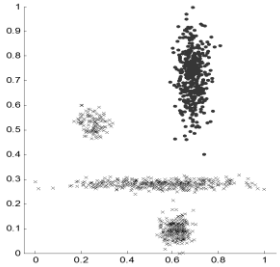
b) wynik grupowania dla metody k-średnich + DB



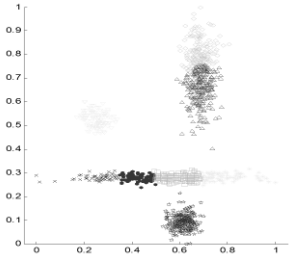
c) wartości współczynników SI i DB w metodzie Warda w podziale od 2 do 12



d) wynik grupowanie metodą Warda



e) wynik grupowania dla metody Warda + SI



f) wynik grupowania dla metody Warda + DB

Źródło: Opracowanie własne.

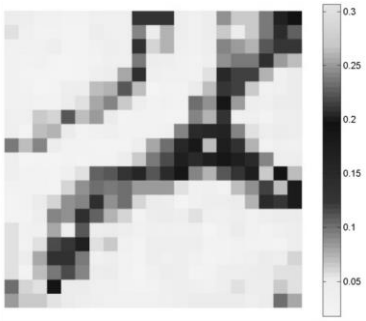
Minimalny błąd sieci *OE*=0,2 uzyskano dla sieci o kwadratowej strukturze połączeń, o wymiarze 11x11 neuronów, z sąsiedztwem gaussowskim. Sieć ta pozwoliła na bezbłędne rozpoznanie[2] istniejących 4 skupień, co jest wyraźnie widoczne na rysunku 3a. Jasne

[2] Jest to oczywiście ocena subiektywna. Na mapie SOM należałoby narysować granice skupień tak jak je badacz widzi, następnie ustalić położenie każdej jednostki na mapie SOM, uzyskując ostateczny wynik grupowania.

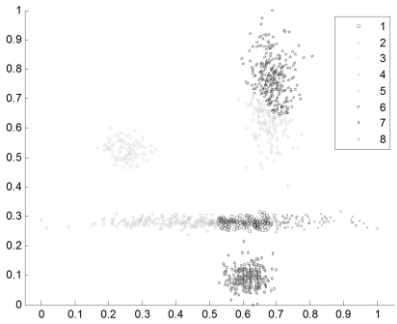
płaszczyzny mapy odpowiadają małym odległościom między jednostkami, a więc skupieniom, oddzielonym ciemnymi pasami większych odległości.

Stosując hybrydową sieć neuronową typu SOM, uzyskano zróżnicowane wyniki. Stosując na drugim stopniu metodę k-średnich i Warda, jakość grupowania jest niska, choć lepsza niż dla tych samych metod stosowanych samodzielnie. Skorygowane współczynniki Randa kształtują się na poziomie 0,62 do 0,65. Żadna z dwóch zastosowanych metod nie pozwoliła na poprawne wskazanie liczby skupień niezależnie od zastosowanego współczynnika liczby skupień. W metodzie SOM + k-średnich wyróżniono 3 skupienia dla obu wskaźników liczby skupień. W metodzie SOM + metoda Warda wyróżniono 3 skupienia dla współczynnika SI a 8 dla współczynnika DB.

Rysunek 3. Zbiór drugi – mapa SOM i wynik grupowania dla hybrydy sieć SOM + metoda Warda + DB



a) mapa SOM

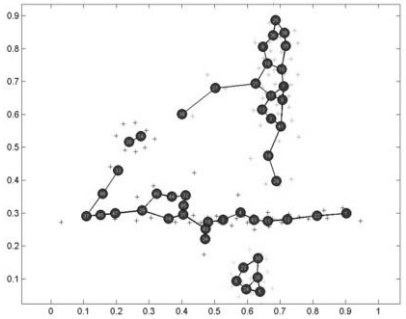


b) wynik grupowania hybrydy: sieć SOM + metoda Warda + DB

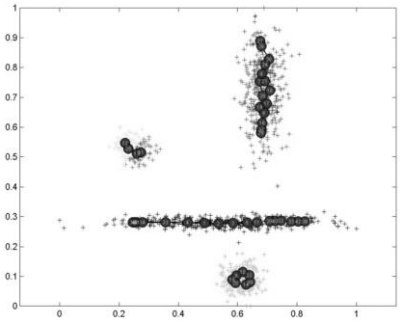
Źródło: Opracowanie własne.

Znacząco lepsze wyniki uzyskano dla hybrydy: sieć SOM + sieć GNG. Metoda ta pozwoliła na bezbłędne wyznaczenie liczby skupień i bezbłędne wyróżnienie skupień. Skorygowany współczynnik Randa osiągnął wartość 1. Sieć GNG zbudowaną na neuronach sieci SOM przedstawiono na rysunku 4a. Kształt wyróżnionych skupień odbiega nieco od ich kształtu rzeczywistego, ponieważ w sieci SOM znajduje się kilka martwych neuronów. Nie odpowiadają one za żadną jednostkę, jednak sam fakt ich istnienia powoduje ich odwzorowanie przez sieć GNG. Sieć GNG zastosowana samodzielnie także uzyskała dobry wynik. Współczynnik Randa jest równy 0,95. Sieć okazała się jednak bardzo wrażliwa na lokalne zmiany w gęstości jednostek i oddzieliła niewielką ich liczbę, tworząc dodatkowe dwa skupienia (zobacz rysunek 4b). Warto zauważyć, że sieć GNG idealnie odwzorowała strukturę przestrzenną badanych jednostek, wyróżniając dwa skupienia rozciągnięte w przestrzeni i dwa sferyczne.

Rysunek 4. Zbiór drugi – hybryda: sieć SOM + sieć GNG i sieć GNG



a) wynik grupowania hybrydy: sieć SOM + sieć GNG



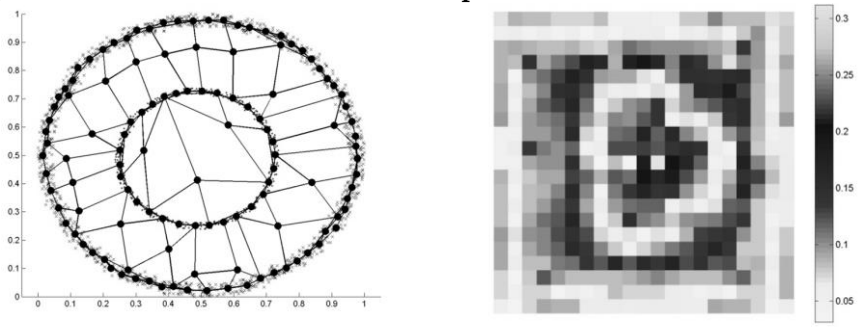
b) grupowanie w oparciu o sieć GNG

Źródło: Opracowanie własne.

Przypadek zaprezentowany w trzecim zbiorze danych jest szczególnie trudny. Skupienia mają bardzo nietypową konfigurację. Nie istnieje żadna niedomknięta krzywa pozwalająca je rozdzielić. Żadna z analizowanych klasycznych metod nie pozwoliła, nawet w przybliżeniu, rozdzielić istniejących skupień. Skorygowane współczynniki Randa kształtują się na poziomie 0,07 dla metody k-średnich i 0,09 dla metody Warda. Nie można także poprawnie wyznaczyć istniejącej liczby skupień. Zarówno metoda k-średnich, jak i metoda Warda wskazały niepoprawnie liczbę skupień równą 12, a więc maksymalną liczbę skupień, na którą próbowano rozdzielić jednostki.

Sieć SOM zbudowana dla trzeciego zbioru ma rozmiar 11x11 neuronów połączonych w sieć kwadratową z gaussowską funkcją sąsiedztwa. Całkowity błąd sieci *OE* wyniósł 0,23. Na mapie SOM podobnie jak w poprzednim zbiorze można bez trudu zauważyć istnienie dwóch skupień i ich konfigurację w przestrzeni. Wizualna obserwacja sieci i mapy SOM pozwala dokonać poprawnego wnioskowania (zobacz rysunek 5a, 5b). Zastosowanie podejścia hybrydowego sieci SOM z obu klasycznymi metodami grupowania ponownie nie powala na poprawne wyróżnienie skupień. Skorygowane współczynniki Randa w tym przypadku kształtują się na poziomie 0,001 do 0,11. Nie można także ustalić poprawnej liczby skupień. Hybrydy te wskazują błędny podział, wyróżniając 3 lub 12 klas (zobacz tablica 1).

Rysunek 5. Zbiór trzeci – sieć SOM i mapa SOM

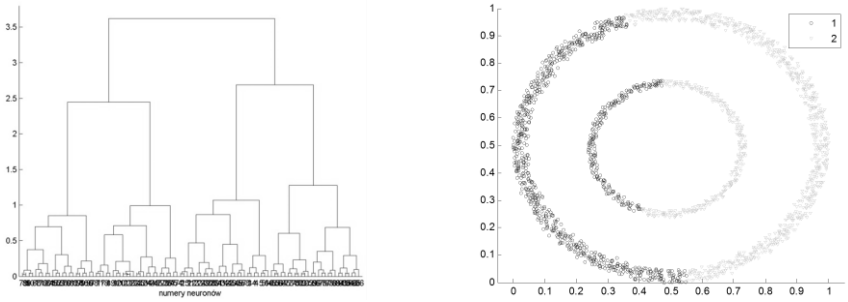


a) sieć SOM b) mapa SOM

Źródło: Opracowanie własne.

Na rysunku 6a pokazano dendrogram uzyskany metodą Warda zastosowaną na neuronach sieci SOM. Gdyby nawet subiektywnie oceniając dendrogram dokonać odcięcia na poziomie wiązania równym 3, wyróżniając dwa skupienia, nie można uzyskać poprawnej struktury grupowej. Na rysunku 6b pokazano wynik takiego podziału.

Rysunek 6. Zbiór trzeci – metoda Warda na neuronach sieci SOM i jej wynik grupowania



a) dendrogram dla hybrydy: sieć SOM + metoda Warda + SI

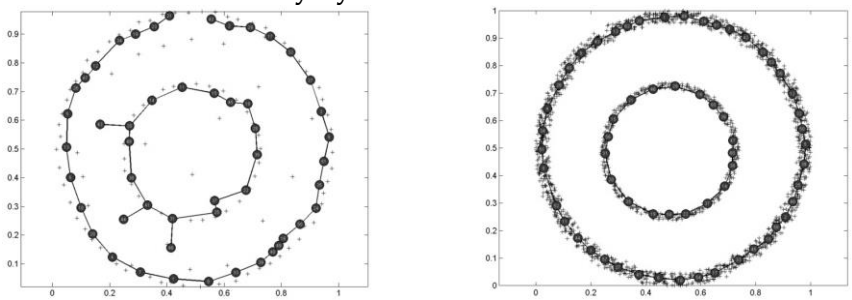
b) wynik grupowania hybrydy: sieć SOM + metoda Warda + SI

Źródło: Opracowanie własne.

Zupełnie inne efekty osiągnięto dla hybrydy: sieć SOM + sieć GNG. Metoda ta pozwoliła na bezbłędne rozpoznanie zarówno liczby skupień, jak i ich struktury przestrzennej. Uzyskano wartość skorygowanego współczynnika Randa na poziomie 1. Na rysunku 7a przedstawiono sieć GNG zbudowaną na neuronach sieci SOM.

Zastosowanie samodzielnej sieci GNG ponownie pozwoliło na poprawne ustalenie liczby istniejących skupień i ich rozdzielenie. Współczynnik Randa wyniósł 1. Sieć GNG na tle badanych jednostek prezentuje rysunek 7b.

Rysunek 7. Zbiór trzeci – hybryda: sieć SOM + sieć GNG i sieć GNG



a) wynik grupowania hybrydy: sieć SOM + sieć GNG b) grupowania w oparciu o sieć GNG

Źródło: Opracowanie własne.

Zakończenie

Wyniki badań symulacyjnych zwykle nie posiadają waloru dowodu takiego jak rozważania czysto formalne. Dostarczają jednak cennych informacji dotyczących badanych zagadnień przynajmniej w zakresie uwzględnionym w założeniach symulacji. W prezentowanym badaniu testowano trzy zbiory danych o charakterystycznych własnościach i zbadano zdolność wybranych metod analizy skupień do poprawnej ich oceny. Wykazano, że w przypadku prostych struktur przestrzennych skupień zastosowanie złożonych algorytmów grupowania nie daje żadnego zysku. Struktury te są łatwo rozpoznawalne także przez klasyczne metody. Dodatkowo są one łatwiejsze do zastosowania i dostępne w typowych pakietach oprogramowania statystycznego. Jednak w przypadku bardziej złożonych struktur przestrzennych jednostek metody klasyczne mogą zawodzić. Wykazano, że w takich przypadkach sieci hybrydowe SOM + GNG posiadają znacząco większy potencjał do wyróżniania skupień od metod klasycznych i hybrydowych SOM + metoda klasyczna. Wykazano także, że w testowanych przypadkach sieć GNG charakteryzuje się najlepszymi własnościami wyróżniania skupień spośród badanych metod.

Literatura

1. Anderberg M.R. (1973), *Cluster analysis for applications*, Academic Press, New York, San Francisco, London.
2. Ball G.H., Hall D.J. (1965), *ISODATA, a novel method of data analysis and pattern classification*, Technical report NTIS AD 699616, Stanford Research Institute, Stanford, Menlo Park CA.
3. Benzécri J.P. (1992), *Correspondence Analysis Handbook*, New York: Marcel Dekker.
4. Bergan T., (1971), *Survey of numerical techniques for grouping*, „Bacteriological Reviews", Vol. 35, No. 2

5. Davies D.L., Bouldin D.W. (1979), *A cluster separation measure*, „IEEE Transactions on Pattern Analysis and Machine Intelligence", PAMI-1, No. 2.
6. Deboeck G., Kohonen T. (1998), *Visual explorations in finance with Self-Organizing Maps*, Springer-Verlag, London.
7. Everitt B.S., Landau S., Leese M., Stahl D. (2011), *Cluster analysis*, 5th edition, John Wiley & Sons, Ltd., Chichester.
8. Forgy E.W. (1965), *Cluster analysis of multivariate data: efficiency vs. interpretability of classifications*, Streszczenia referatów wygłoszonych na Spring Meeting of ENAR, na Florida State University at Tallahassee, Floryda, 29.04-1.05, „Biometrics", Vol. 21, No. 3.
9. Fritzke B. (1994), *Growing cell structures – a self-organizing network for unsupervised and supervised learning*, „Neural Networks", Vol. 7, No. 9.
10. Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.
11. Gordon A.D. (1987), *A review of hierarchical classification*, „Journal of the Royal Statistical Society", Series A (General), Vol. 150, No. 2.
12. Greenacre M.J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.
13. Hardy A. (1996), *On the number of clusters*, „Computational Statistics & Data Analysis", Vol. 23, No. 1.
14. Hartigan J. (1975), *Clustering algorithms*, John Wiley & Sons, New York.
15. Hartigan J., Wong M. (1979), *Algorithm AS136: a k-means clustering algorithm*, „Applied Statistics", Vol. 28, No. 1.
16. Jajuga K. (1990), *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa.
17. Jajuga K. (1993), *Statystyczna analiza wielowymiarowa*, PWN, Warszawa.
18. Jardine N., Sibson R. (1968), *The construction of hierarchic and non-hierarchic classifications*, „The Computer Journal", Vol. 11, No. 2.
19. Johnson S.C. (1967), *Hierarchical clustering schemes*, „Psychometrika", Vol. 32, No. 3.
20. Kohonen T. (1995, 1997, 2001), *Self-Organizing Maps*, Springer-Verlag, Heidelberg, Berlin.
21. Lance G.N., Williams W.T. (1966 a), *Computer programs for hierarchical polythetic classification* („Similarity analysis"), „The Computer Journal", Vol. 9, No. 1.
22. Lance G.N., Williams W.T. (1966 b), *A generalized sorting strategy for computer classifications*, „Nature", Vol. 212.
23. Lance G.N., Williams W.T. (1967 a), *A general theory of classificatory sorting strategies, I. Hierarchical systems*, „The Computer Journal", Vol. 9, No. 4.
24. Lance G.N., Williams W.T. (1967 b), *A general theory of classificatory sorting strategies: II. Clustering systems*, „The Computer Journal", Vol. 10, No. 3.
25. Lamirel J., Mall R., Cuxac P., Safi G. (2011), *Variations to incremental growing neural gas algorithm based on label maximization*, „Neural Networks, The 2011 International Joint Conference on".
26. MacQueen J.B. (1967), *Some methods for classification and analysis of multivariate observations*, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1.
27. Marek T., Noworol C. (1987), w: Brzeziński J. (red.), *Wielozmiennowe modele statystyczne w badaniach psychologicznych*, PWN, Warszawa-Poznań.

28. Migdał Najman K., (2007), *Propozycja hybrydowej metody grupowania dużych zbiorów danych wykorzystującej sieć Kohonena i taksonomiczne metody grupowania*, Taksonomia 14, Prace Naukowe AE we Wrocławiu, nr 1169.
29. Migdał Najman K. (2008), *Analiza porównawcza struktur hierarchicznych skupień uzyskanych z wykorzystaniem hybrydowych metod grupowania*, Taksonomia 15, Prace Naukowe Uniwersytety Ekonomicznego we Wrocławiu, nr 7.
30. Migdał Najman K. (2009), *Analiza porównawcza własności nienadzorowanych sieci neuronowych typu self organizing map i growing neural gas w analizie skupień*, Taksonomia 16, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 47.
31. Migdał Najman K. (2012), *Propozycja hybrydowej metody grupowania opartej na sieciach samouczących*, Taksonomia 19, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 242.
32. Migdał Najman K., Najman K. (2005), *Analityczne metody ustalania liczby skupień*, Taksonomia 12, Prace Naukowe AE we Wrocławiu, nr 1076.
33. Migdał Najman K., Najman K. (2006 a), *Analityczne metody ustalania liczby skupień w rozmytych zbiorach danych*, Taksonomia 13, Prace Naukowe AE we Wrocławiu, nr 1126.
34. Migdał Najman K., Najman K. (2006 b), *Wykorzystanie indeksu silhouette do ustalania optymalnej liczby skupień*, „Wiadomości Statystyczne", nr 6.
35. Migdał Najman K., Najman K. (2008), *Wykorzystanie wskaźnika Dunna do ustalania optymalnej liczby skupień*, „Wiadomości Statystyczne", nr 11.
36. Milligan G.W. (1980), *An examination of the effect of six types of error perturbation on fifteen clustering algorithms*, „Psychometrika", Vol. 45, No. 3.
37. Milligan G.W., Cooper M.C. (1985), *An examination of procedures for determining the number of clusters in a data set*, „Psychometrika", Vol. 50, No. 2.
38. Milligan G.W., Cooper M.C. (1987), *Methodology review: clustering methods*, „Applied psychological measurement", Vol. 11, No. 4.
39. Mirkin B.G. (1996), *Mathematical classification and clustering*, Kluwer Academic Publishers, Dordrecht, The Netherlands.
40. Najman K. (2008), *Symulacyjna analiza wpływu wyboru kryterium optymalności podziału obiektów na jakość uzyskanej klasyfikacji a algorytmach k-średnich*, Taksonomia 15, Prace Naukowe Uniwersytety Ekonomicznego we Wrocławiu, nr 7.
41. Najman K. (2009), *Zastosowanie nienadzorowanych sieci neuronowych typu Growing Neural Gas w analizie skupień*, Taksonomia 16, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 47.
42. Najman K. (2010), *Ocena wpływu parametrów sterujących procesem samouczenia się sieci GNG na ich zdolność do separowania skupień*, Taksonomia 17, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 17.
43. Najman K. (2011), *Propozycja algorytmu samouczenia się sieci neuronowych typu GNG ze zmiennym krokiem uczenia*, Taksonomia 18, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 176.
44. Pociecha J., Podolec B., Sokołowski A., Zając K. (1988), *Metody taksonomiczne w badaniach społeczno-ekonomicznych*, PWN, Warszawa.
45. Rand W.M. (1971), *Objective criteria for the evaluation of clustering methods*, „Journal of the American Statistical Association", Vol. 66, No. 336.
46. Rousseeuw P.J. (1987), *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, „Journal of Computational and Applied Mathematics", Vol. 20, No. 1.

47. Sneath P.H.A., Sokal R.R. (1973), *Numerical taxonomy, the principles and practice of numerical classification*, W.H. Freeman and Company, San Francisco.
48. Stąpor K. (2005), *Automatyczna klasyfikacja obiektów*, Akademicka Oficyna Wydawnicza EXIT, Warszawa.
49. Tryon R.C. (1939), *Cluster analysis*, New York: McGraw-Hill.
50. Vesanto J., Himberg J., Alhoniemi E., Parhankangas J. (2002), *SOM Toolbox for Matlab 5*, SOM Toolbox Team, Helsinki University of Technology, ESPOO, Finland.
51. Walesiak M. (1996), *Metody analizy danych marketingowych*, PWN, Warszawa.
52. Ward J.H. (1963), *Hierarchical grouping to optimize an objective function*, „Journal of the American Statistical Association", Vol. 58, No. 301.

Streszczenie

Celem prezentowanych badań jest analiza porównawcza wybranych klasycznych metod grupowania danych z autorskimi, hybrydowymi metodami opartymi na samouczących się sieciach neuronowych typu *Self Organizing Map* (SOM) i *Growing Neural Gas* (GNG). Wykazano eksperymentalnie, że w wyróżnianiu złożonych struktur przestrzennych badanych jednostek najczęściej stosowane metody klasyczne mogą być nieskuteczne. Znacząco lepsze wyniki spośród uwzględnionych metod uzyskano dla hybrydowej sieci neuronowej typu SOM+GNG. Wykazano także, że w testowanych przypadkach samodzielnie stosowana sieć GNG charakteryzuje się najlepszymi własnościami wyróżniania skupień spośród wszystkich badanych metod.

Słowa kluczowe

analiza skupień, sztuczne sieci neuronowe, samoorganizująca się mapa (SOM), gaz neuronowy o zmiennej strukturze (GNG), hybrydowa samoucząca się sieć neuronowa SOM+GNG

A comparative analysis of selected methods of cluster analysis in the grouping units with a complex group structure (Summary)

The aim of this article is a comparative analysis of selected classical clustering methods, with hybrid methods based on self-learning neural network type Self Organizing (SOM) and Growing Neural Gas (GNG). It has been shown experimentally that in distinguishing complex spatial structures of the units, most commonly used classical methods may be ineffective. The better results were obtained from the included methods for a hybrid neural network SOM+GNG. It was also shown that the test cases used alone GNG network is characterized highlighting the best of the properties of clusters of all of these methods.

Keywords

cluster analysis, artificial neural networks, self-organizing map (SOM), growing neural gas (GNG), hybrid self-learning neural network SOM+GNG