PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Krzysztof Najman
Uniwersytet Gdański

ZASTOSOWANIE NIENADZOROWANYCH SIECI NEURONOWYCH TYPU *GROWING NEURAL GAS* W ANALIZIE SKUPIEŃ

1. Wstęp

Jedną z efektywnych metod analizy skupień są nienadzorowane sieci neuronowe samoorganizujące się (*Self Organizing Map*, SOM). Do najważniejszych zalet sieci SOM należą: ich nieparametryczność, niewrażliwość na występowanie wartości skrajnych, odporność na braki danych, a także brak apriorycznej konieczności ustalenia dokładnej struktury sieci (por. [Kohonen 1997; Migdał-Najman, Najman 2008]). Wadą szczególnie uciążliwą w dużych badaniach empirycznych jest duży rozmiar sieci, którego konsekwencją jest długi czas uczenia się sieci i spadająca efektywność grupowania. Jednym z powodów spadku efektywności grupowania sieci jest fakt, że w dużej sieci SOM wiele neuronów nie bierze udziału w rozpoznawaniu obiektów (tzw. efekt martwych neuronów), co wprowadza do analizy jedynie zbędny szum. W konsekwencji struktura sieci jest zwykle nadmiarowa, a proces samouczenia się sieci − niepotrzebnie długi. W analizie dużych problemów badacz napotka barierę informatyczną, gdy wielkość sieci będzie zbyt duża, aby móc uruchomić proces samouczenia, nawet na dobrze wyposażonym komputerze. Jednym z możliwych rozwiązań tego problemu jest konstrukcja sieci samouczącej się, która w procesie nauki sama, zgodnie z potrzebami, będzie korygować swoją strukturę. Proces samouczenia mógłby się rozpoczynać od sieci o minimalnych rozmiarach, a następnie sieć powiększałaby swoją strukturę zgodnie z wybranym kryterium lokalnej lub globalnej optymalności. Konstrukcję sieci tego typu zaproponował B. Fritzke w 1994 r., prezentując nienadzorowaną, samouczącą się sieć neuronową o zmiennej strukturze typu gaz neuronowy (*Growing Neural Gas*, GNG); por. [Fritzke 1994]. Wydaje się, że może być ona pozbawiona wymienionych wad sieci SOM. Jest to rozwinięcie idei klasycznej sieci SOM. Struktura takiej sieci zmienia się dynamicznie w procesie samouczenia się w taki sposób, że

nowe neurony są wstawiane do sieci jedynie w tym miejscu sieci, w którym występuje największy błąd rozpoznawania wzorców (błąd kwantyzacji).

Celem prezentowanych badań jest weryfikacja hipotezy o wysokim potencjale sieci typu GNG w analizie skupień. Przedstawione zostaną podstawy teoretyczne tej metody, jej własności, które na podstawie badan symulacyjnych będą poddane weryfikacji i ocenie.

2. Budowa sieci GNG

Istotą budowy sieci GNG jest konstrukcja sieci maksymalnie oszczędnej, bez zbędnych neuronów, rozłożonych jedynie w tej części przestrzeni, w której znajdują się obiekty. Sieć powinna mieć umiejętność odwzorowania skupień w przestrzeni o dowolnym wymiarze i o dowolnej konfiguracji w przestrzeni. Proces samouczenia się sieci GNG rozpoczyna się od zainicjowania dwóch neuronów $c1$ i $c2$ losowymi wagami (współrzędnymi z przestrzeni analizowanych obiektów):

$$A=\{c1,c2\}, \tag{1}$$

gdzie A to zbiór neuronów. Aby przyspieszyć proces uczenia, zwykle wagi dobiera się tak, aby początkowe dwa neurony były od siebie odległe − obejmowały znaczną część zmienności obiektów w przestrzeni (por. rys. 1). Następnie ustala się połączenie między tymi neuronami i wiek tego połączenia ustala się na 0. W kolejnym etapie testuje się kryteria stopu algorytmu. Podstawowymi kryteriami są: 1) osiągnięcie maksymalnej, założonej liczby iteracji, 2) osiągnięcie maksymalnego zakładanego błędu uczenia sieci (możliwe są różne kryteria formalne oceny błędu uczenia sieci (patrz poniżej), 3) osiągnięcie maksymalnej założonej liczby neuronów (maksymalnego rozmiaru sieci). Spełnienie któregokolwiek warunku kończy pracę algorytmu. Oczywiście w pierwszej iteracji żadne z kryteriów nie jest spełnione i proces przebiega dalej. Spośród wszystkich analizowanych obiektów losowo wybierany jest jeden:

$$D=\{\xi_1,...,\xi_M\}, \xi_i \in \Re^n, \tag{2}$$

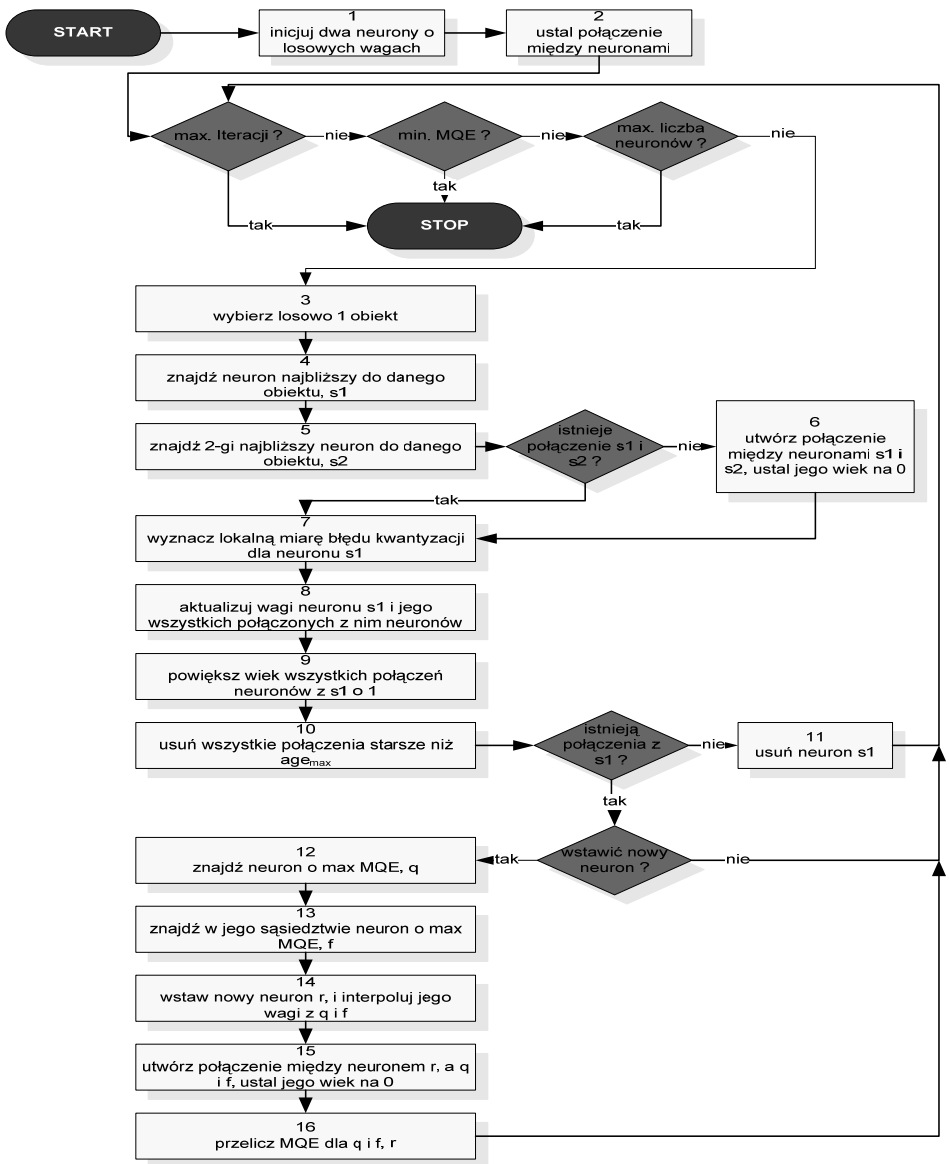
gdzie: D to zbiór M, n wymiarowych obiektów ξ. Wśród istniejących neuronów poszukiwane są: neuron najbliższy do wybranego obiektu i drugi najbliższy:

$$s1=\arg\min_{c\in A}\|\xi-w_c\| \quad s2=\arg\min_{c\in A\setminus\{s1\}}\|\xi-w_c\|, \tag{3}$$

gdzie: w_c to współrzędne neuronu c. Neuron $s1$ jest nazywany neuronem wygrywającym. Za miarę odległości obiektów od neuronów należy przyjąć wybraną metrykę odpowiednią dla skal pomiarowych wykorzystanych w badaniu. Jeżeli pomiar był dokonany na skali ilorazowej lub przedziałowej, to zwykle stosuje się odległość euklidesową. Po znalezieniu tych dwóch neuronów sprawdza, się czy były one połączone. Jeżeli nie, to tworzy się takie połączenie i ustala się jego wiek na 0 ite-

racji. Następuje teraz etap uczenia się neuronów $s1$ i $s2$. W pierwszym kroku wyznaczana jest lokalna miara błędu sieci dla neuronu $s1$:

$$\Delta E_{s1} = \left\| \xi - w_{s1} \right\|^2 . \tag{4}$$



Rys. 1. Algorytm uczenia się sieci GNG

Źródło: opracowanie własne.

Jest to klasyczny błąd kwantyzacji. Poszukuje się następnie wszystkich neuronów połączonych z neuronem s1 i aktualizuje się ich wagi:

$$\Delta w_{s1} = \varepsilon_b \left(\xi - w_{s1} \right) \qquad \Delta w_i = \varepsilon_n \left(\xi - w_{s1} \right) \quad \left(\forall i \in N_{s1} \right), \tag{5}$$

gdzie: *i* oznacza *i*-ty połączony z wygrywającym neuron (por. [Jirayusakul, Auwatanamongkol 2007]). Wiek połączeń między wszystkimi neuronami, których wagi zostały zaktualizowane, zwiększa się o 1. Następnie usuwa się wszystkie połączenia między neuronami starsze niż założony maksymalny wiek połączenia. Sprawdza się następnie, czy neuron *s*1 pozostał połączony z jakimkolwiek innym neuronem. Jeżeli utracił wszystkie połączenia, to wraca się do testowania warunków stopu. Jeżeli połączenia istniały, to rozpoczyna się procedurę wstawiania nowego neuronu. Szuka się neuronu o maksymalnym błędzie kwantyzacji q i jego neuronu najbliższego f. Nowy neuron r wstawia się między neuronami q i f, tworząc jego wagi przez interpolację wag neuronów q i f:

$$A = A \cup \{r\}, \quad w_r = \left(w_q + w_f \right) / 2 . \tag{6}$$

Jednocześnie modyfikuje się połączenia między neuronami, usuwając połączenie między q i f, następnie łącząc neurony q z r i f z r. Wiek tych połączeń ustala się na 0. Wyznacza się także błąd kwantyzacji dla nowego neuronu:

$$E_r = \left(E_q + E_f \right) / 2 . \tag{7}$$

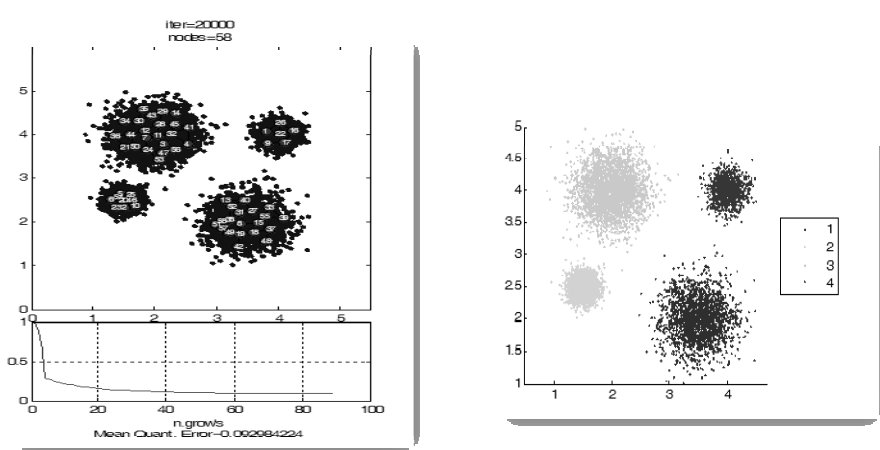
Jest to ostatni etap algorytmu, po którym wraca się do testowania warunków stopu.

Postępowanie to gwarantuje, że nowe neurony powstają w tej części przestrzeni obiektów, w której sieć jest najmniej dopasowana do obiektów. Jednocześnie w tej części sieci, która jest najlepiej dopasowana, neurony się nie uczą (nie modyfikują swoich wag) i tracą połączenia z innymi neuronami, a w konsekwencji są usuwane. Nie nastąpi taka sytuacja, gdy część obiektów będzie bardzo dobrze odwzorowana przez sieć, a część − bardzo źle. Sieć ma oszczędną strukturę. W początkowej fazie uczy się bardzo szybko (mało neuronów i połączeń). Do oczywistych wad algorytmu zaliczyć należy szereg parametrów, które należy ustalić *a priori*. Są to: maksymalna liczba iteracji, maksymalna liczba neuronów i maksymalny wiek połączeń. Jeżeli liczba iteracji będzie zbyt mała, to nie nastąpi prezentacja sieci wszystkich obiektów, a więc uczenie będzie niekompletne. Zbyt duża spowolni zaś proces uczenia się sieci – sieć dobrze nauczona będzie po prostu usuwała najlepszy neuron – jego brak spowoduje jednocześnie, że ta część przestrzeni będzie najgorzej odwzorowywała obiekty, a w konsekwencji niemal w to samo miejsce zostanie wstawiony nowy neuron. Jeżeli ustalimy maksymalną liczbę neuronów na zbyt niskim poziomie, to nie uzyskamy optymalnego odwzorowania obiektów na sieci. Zbyt duża liczba nie jest groźna dla algorytmu, ponieważ sieć GNG osiąga pewną równowagę, gdy sieć poprawnie odwzorowuje obiekty. Jeden

neuron jest usuwany i jeden dodawany, nie powodując nadmiernego wzrostu sieci. Warto jednak pamiętać, że średni błąd kwantyzacji dla wszystkich neuronów może być osiągnięty jedynie wówczas, gdy liczba neuronów różna jest dwukrotności liczby obiektów. Aby neuron pozostał na sieci, musi być połączony co najmniej z jednym sąsiadem. Jeżeli maksymalny wiek połączeń będzie zbyt krótki, to neurony dobrze nauczone będą natychmiast usuwane i algorytm utraci zbieżność. Zbyt wysoka wartość tego parametru spowoduje, że niepotrzebne neurony nie będą usuwane. W konsekwencji sieć będzie się bardzo szybko rozrastać aż do maksymalnej liczby neuronów i zakończy proces uczenia się. Parametry te są trudne do optymalnego ustalenia *a priori*, ponieważ brak prostych formalnych zależności między nimi a jakością uzyskanego grupowania.

3. Własności sieci GNG, badania symulacyjne

W trywialnych przykładach nie ujawniają się zalety sieci GNG. Sieć uczy się bardzo szybko, na każdy obiekt przypada po kilka neuronów i po niewielkiej liczbie iteracji proces uczenia zatrzymuje się. Zwykle każdy obiekt jest traktowany jako osobne skupienie, a liczba neuronów jest większa niż obiektów. To oczywiście zaprzecza postulatom analizy skupień, a także oszczędności algorytmu. W badaniach większej skali tych problemów już nie ma. Liczba neuronów jest w naturalny sposób mniejsza od liczby obiektów. Jeżeli skupienia są separowalne i zajmują w przestrzeni niewielki obszar, to liczba neuronów jest mała i tworzą one dobrze separowalne skupienia (por. rys. 2). Jeżeli skupienia mają porównywalną liczbę obiektów, a różnią się zajmowaną przestrzenią, to skupienia duże będą odwzorowywane przez większą liczbę neuronów niż skupienia małe (por. rys. 3).

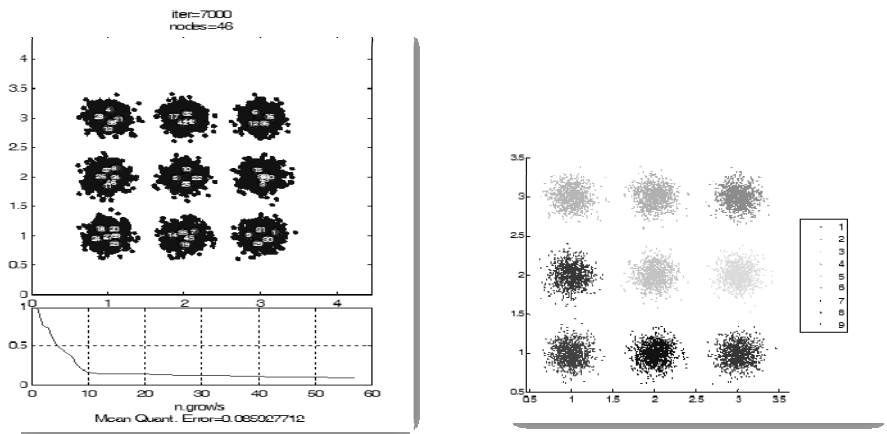


Rys. 2. Pierwszy zbiór testowy

Źródło: opracowanie własne.

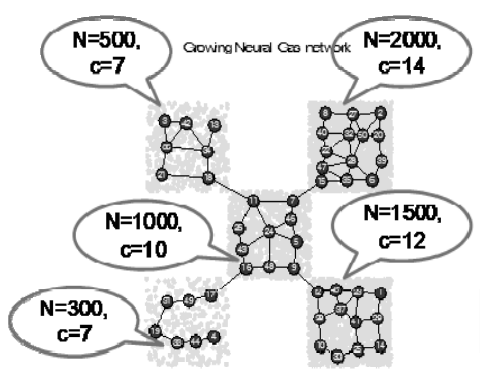
Jednocześnie liczba neuronów odwzorowujących jedno skupienie rośnie wraz ze wzrostem liczby obiektów w skupieniu, nawet jeżeli skupienia zajmują identyczną przestrzeń.

Na rysunku 4 znajduje się 5 skupień różniących się liczbą obiektów odp. N = 300, 500, 1000, 1500, 2000. Liczba neuronów odwzorowujących obiekty w tych skupieniach jest równa odp. c = 7,7,10,12,14. Zależność ta nie jest liniowa. Liczba neuronów przyrasta bardzo wolno wraz ze wzrostem liczby obiektów w skupieniu.



Rys. 3. Drugi zbiór testowy

Źródło: opracowanie własne.



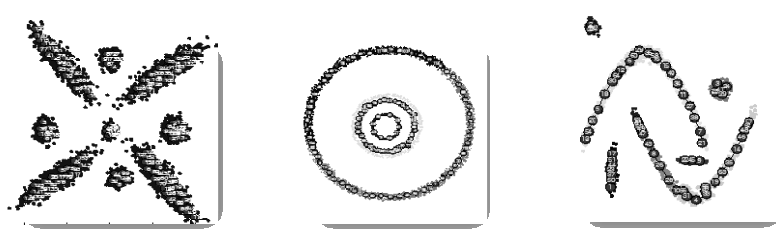
Rys. 4. Liczba obiektów w skupieniach a liczba neuronów odwzorowująca obiekty

Źródło: opracowanie własne.

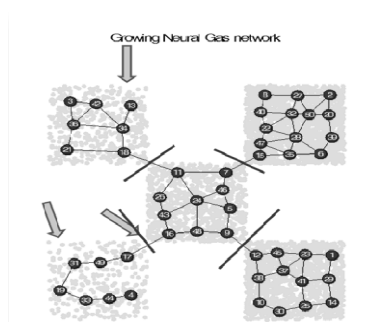
Wspomniane dwie cechy są bardzo mocną stroną sieci GNG. Wskazują na jej oszczędność obliczeniową. Nowe neurony są wstawiane jedynie wtedy, gdy są naprawdę potrzebne, i tam, gdzie są naprawdę potrzebne. Sieć nie ma tendencji do nadmiernego rozrostu. Struktura sieci tworząca się w trakcie procesu samouczenia się sieci ma także tę zaletę, że pozwala sieci dopasować się do dowolnego kształtu skupień w przestrzeni.

Na rysunku 5 przedstawiono 3 zbiory przykładowych obiektów i nauczone sieci GNG. Jak można zauważyć, nie ma żadnych zbędnych neuronów w tej części przestrzeni, w której nie ma żadnego obiektu. Sieć dopasowała się do bardzo skomplikowanych kształtów skupień (por. rys. 5).

Poza wymienionymi zaletami sieć GNG ma także wiele trudnych do wyeliminowania wad. Badania symulacyjne wskazują, że sieć GNG nie jest w stanie rozróżnić skupień o strukturze rozmytej. Jeżeli odległości między obiektami z różnych skupień są porównywalne do typowych odległości między obiektami w skupieniach, to sieć GNG nie rozróżni tych skupień (por. rys. 6). Na rysunku 6 strzałkami zaznaczono przykładowe odległości między obiektami w skupieniach, które są znacząco większe niż odległości między obiektami sąsiednich skupień. W konsekwencji sieć wszystkie obiekty traktuje jak jedno skupienie. Liniami na tym wykresie zaznaczono połączenia neuronów, które powinny zaniknąć, aby skupienia zostały rozdzielone.



Rys. 5. Dopasowanie sieci GNG do dowolnego kształtu skupień w przestrzeni
Źródło: opracowanie własne.



Rys. 6. Sieć GNG a skupienia o strukturze rozmytej
Źródło: opracowanie własne.

Innym ważnym problemem jest silna zależność zdolności do wyróżniania skupień sieci GNG od liczby neuronów. Zbyt mała liczba neuronów spowoduje niemożliwość rozdzielenia skupień, ponieważ sieć nie będzie zdolna do usuwania połączeń między neuronami. Jednocześnie zbyt duża liczba neuronów spowoduje, że sieć zacznie wyróżniać jako skupienia każde lokalne zagęszczenie obiektów w przestrzeni. Duża liczba neuronów powoduje, że sieć traci zdolność do uogólniania. Niestety sterowanie liczbą neuronów jest utrudnione w algorytmie uczenia się sieci GNG. Możliwe jest jedynie założenie maksymalnej liczby neuronów i maksymalnej liczby epok, w których neuron nieuczący się nie zostaje usunięty. O rzeczywistej liczbie neuronów decydują konfiguracja obiektów i stosunek szybkości dodawania nowych neuronów do szybkości ich usuwania w sieci.

4. Wnioski

Sieć samoucząca się jest ważnym narzędziem analizy skupień. Ma wiele rzadko spotykanych zalet i kilka wad, które wymagają szczególnej uwagi badacza. Główne zalety sieci GNG są następujące: 1) zdolność do samodzielnej modyfikacji struktury sieci zgodnej z przyjętym kryterium optymalizacji; 2) sieć w znacznym stopniu autonomicznie ustala optymalną liczbę neuronów; 3) sieć całkowicie samodzielnie ustala liczbę skupień; 4) potrafi dopasować się do każdego kształtu skupień; 5) uczenie się sieci jest relatywnie szybkie; 6) nie wymaga dużej mocy obliczeniowej. Najważniejszymi wadami sieci GNG są: 1) znaczna liczba parametrów sterujących pracą algorytmu; 2) brak formalnych kryteriów ustalania wartości tych parametrów przy jednoczesnej dużej wrażliwości sieci na zmianę ich wartości; 3) słabe własności sieci dla nieseparowalnych skupień; 4) brak standardowego oprogramowania utrudniający powszechniejsze stosowanie tej metody analizy skupień. Wady 1 i 2 do pewnego stopnia można zredukować, powtarzając analizę kilkukrotnie przy różnych zestawach parametrów, każdorazowo mierząc homogeniczność uzyskanych skupień jednym ze znanych wskaźników (por. [Migdał-Najman, Najman 2005]). Ponieważ sieć uczy się bardzo szybko, nie jest to bardzo uciążliwe postępowanie.

Niezależnie od wad w ocenie autora zalety sieci GNG zdecydowanie przeważają. Pełniejsza ocena tej sieci wymaga dalszych badań, jednak już teraz można stwierdzić, że jest to cenne narzędzie w analizie skupień.

Literatura

Fritzke B. (1994), *Growing cell structures − a self-organizing network for unsupervised and supervised learning*, „Neural Networks" , vol. 7, no 9, , s. 1441-1460.

Jirayusakul A., Auwatanamongkol S. (2007), *A supervised growing neural gas algorithm for cluster analysis*, „International Journal of Hybrid Intelligent Systems", 4, s. 129-141.

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

Migdał-Najman K., Najman K. (2005), *Analityczne metody ustalania liczby skupień*, Prace Naukowe AE we Wrocławiu nr 1076, AE, Wrocław, s. 256-273.

Migdał-Najman K., Najman K. (2008), *Data analysis, machine learning and applications, Applying the Kohonen Self-Organizing Map Networks to Selecting Variables*, *Studies in Classification, Data Analysis and Knowledge Organization*, C. Presisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker, Springer Verlag, Heidelberg, Berlin, s. 45-54.

APPLYING OF GROWING NEURAL GAS NEURAL NETWORKS IN CLUSTER ANALYSIS

Summary

One of the more effective methods in cluster analysis are unsupervised neural networks, for example Self Organizing Map, SOM. The problem which can appear in large data sets is a priori the network's structure. SOM could be time consuming and require powerful computers, it has tendency to twine and possess many neurons which do not take part in learning. It seems that unsupervised growing neural gas (GNG) with dynamic structure does not have these disadvantages.

The main goal of research presented in this paper is hypothesis verification that the GNG network has large potential in cluster analysis. Theoretical principles, properties of this method, simulation research and opinions are presented.