PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Krzysztof Najman
Uniwersytet Gdański

OCENA WPŁYWU PARAMETRÓW STERUJĄCYCH PROCESEM SAMOUCZENIA SIĘ SIECI GNG NA ICH ZDOLNOŚĆ DO SEPAROWANIA SKUPIEŃ

Streszczenie: Celem prezentowanych badań jest ocena wpływu wartości parametrów sterujących procesem samouczenia się sieci GNG na jakość uzyskanej klasyfikacji, szczególnie w sytuacji, gdy skupienia są słabo separowalne. Prezentowane badania opisują teoretyczny wpływ poszczególnych parametrów sterujących algorytmem GNG na strukturę i zdolność rozwiązywania problemów przez sieć. Zaprezentowano także wyniki badań symulacyjnych potwierdzających tezę, że optymalny wybór parametrów sterujących pozwala znacznie zwiększyć zdolność sieci GNG do poprawnej identyfikacji skupień.

1. Wstęp

Analiza skupień to jedna z podstawowych metod analizy danych wielowymiarowych. Wśród licznych metod stosowanych w analizie skupień swoje miejsce mają także metody wykorzystujące sztuczne sieci neuronowe. Szczególną ich klasą są samoorganizujące się sztuczne sieci neuronowe. Do sieci tego typu zalicza się sieć SOM (*Self Organizing Map*) zaproponowaną przez T. Kohonena [1997] i sieć GNG (*Growing Neural Gas*) zaproponowaną przez B. Fritzkego [1994]. Z rozważań teoretycznych, a także badań symulacyjnych [Najman 2009] wynika, że większy potencjał w grupowaniu danych wielowymiarowych prawdopodobnie ma sieć GNG. Ma wiele rzadko spotykanych zalet i kilka wad, które wymagają szczególnej uwagi badacza. Do głównych zalet sieci GNG należą:

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.

W konsekwencji stosowanie sieci GNG jest oszczędne, ponieważ brak w jej strukturze niepotrzebnych (martwych) neuronów, kluczowe własności skupień są

ustalane autonomicznie, a jej praca jest relatywnie szybka. Głównymi wadami sieci GNG są:

1) znaczna liczba parametrów sterujących pracą algorytmu;

2) brak formalnych kryteriów ustalania wartości tych parametrów, przy jednoczesnej dużej wrażliwości własności sieci na zmianę ich wartości;

3) słaba jakość grupowania dla słabo separowalnych skupień – szczególnie przy nieoptymalnie ustalonych parametrach uczenia sieci;

4) brak standardowego oprogramowania utrudniający powszechniejsze stosowanie tej metody analizy skupień.

Celem prezentowanych badań jest ocena wpływu wartości parametrów sterujących procesem samouczenia się sieci GNG na jakość uzyskanej klasyfikacji, szczególnie w sytuacji, gdy skupienia są słabo separowalne.

2. Parametry sterujące samouczeniem się sieci GNG

Istotą budowy sieci GNG jest konstrukcja sieci maksymalnie oszczędnej, bez zbędnych neuronów. Potrzeba ta wynika z doświadczeń z budową sieci o stałej strukturze, takich jak klasyczne sieci SOM i NG (*Neural Gass*). W sieciach tego typu ich struktura jest zakładana *a priori* i nie podlega modyfikacji w procesie samouczenia się. W konsekwencji, rozwiązując bardziej skomplikowane problemy (duża liczba obiektów i duża liczba cech), sieć musiała mieć duży rozmiar (wiele neuronów). Uczenie takiej sieci od pierwszej iteracji jest czasochłonne, a może się okazać, że jej rozmiar jest znacznie nadmiarowy. Pewna liczba neuronów może nie odpowiadać za klasyfikację jakichkolwiek obiektów. Biorą one udział w uczeniu się sieci jedynie jako odlegli sąsiedzi neuronów wygrywających (*best matching units*), spowalniając pracę całego algorytmu, utrudniając analizę wyników i nie wpływając na poprawę jakości grupowania. Neurony takie nazywamy martwymi. Sieci o zmiennej strukturze pozbawione są tej wady[1], ponieważ proces samouczenia się sieci GNG rozpoczyna się zawsze od 2 neuronów. Liczba ta jest zwiększana jedynie w tej części przestrzeni obiektów, w której sieć popełnia maksymalny błąd kwantyzacji. W algorytmie budowy sieci GNG należy więc zdefiniować pierwsze dwa parametry sterujące: 1. maksymalną zakładaną liczbę neuronów, po której osiągnięciu do struktury sieci nie są już dodawane kolejne neurony[2]; 2. poziom błędu kwantyzacji [Najman 2009, wzór 4, s. 198] przerywający proces samouczenia się sieci. Jego osiągnięcie będzie oznaczało „nauczenie się" sieci zadanego problemu z zakładaną dokładnością. Jest to zwykle bliska zeru liczba rzeczywista.

[1] Szczegóły algorytmu budowy sieci GNG wraz ze schematem blokowym znajdują się w artykule [Najman 2009].

[2] W rzeczywistości osiągnięcie przez sieć tej liczby neuronów nie przerywa procesu samouczenia się. Wagi neuronów są nadal poprawiane, a neurony najdłużej niebiorące udziału w uczeniu się sieci, są usuwane. Po ich usunięciu efektywna liczba neuronów zmniejsza się i znowu mogą być dodane nowe neurony.

Oba te parametry są w wysokim stopniu arbitralne. Trudno bowiem ocenić, jak wiele neuronów będzie niezbędnych do rozwiązania problemu i na jakim poziomie błędu kwantyzacji można uznać problem za rozwiązany. Gdyby założyć, że dopuszczalny błąd kwantyzacji jest równy 0, a liczba neuronów jest dowolnie duża, proces samouczenia się sieci zakończyłby pracę przy liczbie neuronów równej dwukrotności liczby obiektów, w nieokreślonej liczbie iteracji i nieokreślonym czasie. Wynika to z tego, że błąd kwantyzacji sieci może być równy 0 jedynie wtedy, gdy współrzędne neuronów są identyczne jak współrzędne obiektów. Współrzędne te będą identyczne jedynie wtedy, gdy każdy obiekt będzie rozpoznawany przez dwa neurony. Będzie tak, ponieważ w sieci nie może pozostać pojedynczy, niepołączony z innymi neuron[3]. Pojawia się w tym momencie potrzeba wprowadzenia kolejnego, trzeciego parametru sterującego, którym jest liczba iteracji uczących[4]. Parametr ten ma zapobiegać takiej sytuacji, w której sieć uczy się bez końca, nie mogąc osiągnąć dostatecznie małego błędu kwantyzacji. Niestety ponownie trudno jednoznacznie zdefiniować, co to znaczy „zbyt długo". Może być to rozumiane jako orientacyjny czas pracy algorytmu, który użytkownik zgadza się poświęcić na samouczenie się sieci. Jeżeli w zadanej liczbie iteracji sieć nie osiągnie zakładanego poziomu błędu kwantyzacji, proces samouczenia się zostaje przerwany. Wszystkie te parametry, a więc liczba neuronów, dopuszczalny błąd kwantyzacji i maksymalna liczba iteracji, współdziałają ze sobą. Jeżeli bowiem chcemy osiągnąć bardzo mały błąd kwantyzacji, liczba neuronów i iteracji musi być znaczna. Wynika to z tego, że sieć musi zdążyć osiągnąć rozmiar umożliwiający wypełnienie przestrzeni obiektów, a także ich poprawne odwzorowanie. Szybkość osiągania odpowiedniego rozmiaru sieci jest indywidualną cechą każdego problemu. Zależy ona od liczby obiektów i ich konfiguracji w przestrzeni, a to są indywidualne cechy każdego zbioru danych. Szybkość tę można obserwować jedynie na drodze kolejnych eksperymentów. Zależy ona także od kolejnego parametru sterującego algorytmem GNG. Jest nim maksymalny wiek połączeń neuronów. W każdej iteracji wiek połączeń neuronów biorących udział w procesie samouczenia się jest zwiększany o 1. Wysokie wartości tego parametru oznaczają, że co prawda neurony modyfikują swoje wagi, jednak ich lokalny błąd kwantyzacji jest na tyle mały, że nie są w tym miejscu wstawiane nowe neurony. Wstawienie nowego neuronu wyzerowuje wiek wszystkich neuronów z nim połączonych. Neurony najstarsze są usuwane, ponieważ ich brak w minimalnym stopniu wpływa na łączny błąd kwantyzacji sieci, pozwala sąsiadom uczyć się szybciej i minimalizuje rozmiar sieci. Jeżeli maksymalny wiek neuronu zostanie ustalony na niskim poziomie, sieć nie będzie mogła się rozrastać, ponieważ wstępnie nauczone neurony będą usuwane,

[3] Jeżeli w procesie samouczenia się sieci GNG zostanie usunięty jeden neuron, a był on połączony jedynie z jednym innym neuronem, ten, który pozostał, jest automatycznie usuwany z sieci. Z tego powodu nigdy w sieci nie pozostanie neuron niepołączony z żadnym innym neuronem.

[4] Parametrem tym może być alternatywnie czas pracy algorytmu wyrażony np. w sekundach.

znacznie zwiększając błąd kwantyzacji. W konsekwencji w pobliżu usuwanego neuronu wstawiony będzie nowy neuron i proces będzie się powtarzał aż do osiągnięcia maksymalnej liczby iteracji. Ustalenie zbyt wysokiego maksymalnego wieku neuronów skutkuje tym, że liczba neuronów rośnie bardzo szybko. Szybki wzrost liczby neuronów powoduje szybki spadek błędu kwantyzacji, jednak nie przez samouczenie się sieci i optymalizację struktury.

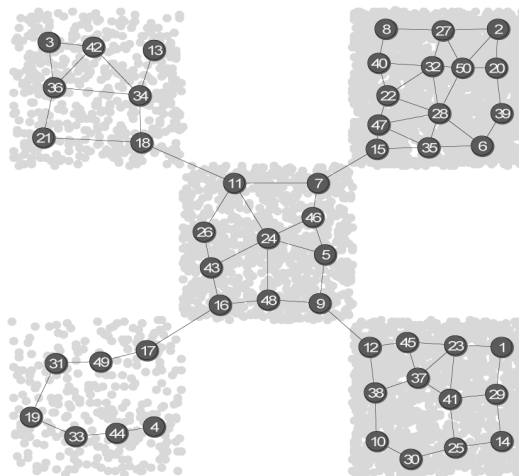
Szybkość wypełniania przestrzeni obiektów zależy od kolejnych parametrów uczenia się sieci. Są to: krok uczenia się neuronu wygrywającego (uczącego się) i krok uczenia się wszystkich neuronów połączonych z neuronem wygrywającym. Ustalenie wartości tych parametrów także nie jest łatwe. Przyjmuje się, że krok uczenia się neuronu zwycięskiego jest większy niż pozostałych neuronów. Wynika to z tego, że neuron wygrywający powinien znacznie zbliżyć się do odwzorowywanego obiektu, zmniejszając błąd kwantyzacji, a jednocześnie pozostałe uczące się neurony mogły dość dobrze odwzorowywać inne obiekty, więc nie powinno się ich zanadto „psuć". Jeżeli wartości te są bardzo małe[5], to jest szansa na osiągnięcie bardzo małych błędów kwantyzacji sieci, jednak tylko kosztem bardzo dużej liczby iteracji. Jeżeli wartości te będą nadto duże, uczenie się będzie co prawda bardzo szybkie (niewielka liczba iteracji znacznie zmniejsza błąd kwantyzacji), jednak błąd kwantyzacji będzie malał jedynie do momentu, w którym wymagane poprawki współrzędnych neuronów będą większe niż krok uczenia. W dalszych iteracjach sieć utraci zdolność uczenia się[6]. Wartości kroków uczenia można wstępnie oszacować na podstawie macierzy odległości między obiektami. Powinny one być ułamkiem przeciętnej odległości między obiektami. Optymalne ustalenie wszystkich parametrów uczenia sieci nie jest zadaniem łatwym i mimo pewnych wskazówek teoretycznych należy je ustalać indywidualnie dla każdego badanego zagadnienia.

3. Wpływ parametrów sterujących samouczeniem się sieci GNG na jakość grupowania

Grupowanie obiektów wielowymiarowych metodą GNG jest bardzo skuteczne, o ile tylko skupienia obiektów są dobrze separowalne. Jak pokazano w artykule [Najman 2009], gdy skupienia obiektów są rozmyte, sieć GNG może nie rozdzielić ich prawidłowo. Na rysunku 1 zaprezentowano typowy przykład takiej sytuacji. Badane obiekty wyraźnie grupują się w 5 skupień, jednak skupienia graniczą ze sobą w przestrzeni.

[5] Małe w stosunku do przeciętnych odległości między obiektami w przestrzeni. Odległości te są mierzone odpowiednimi do skal pomiarowych cech metrykami.

[6] Jest to znany z metod optymalizacyjnych problem stałego kroku uczenia. W algorytmie GNG, jak dotąd, nie testowano zmiennego kroku uczenia się neuronów. Poprawkę taką można łatwo zaimplementować, rozpoczynając proces samo uczenia się od względnie dużej wartości kroku uczenia, systematycznie zmniejszając go wraz ze spadkiem lokalnych błędów kwantyzacji.



Rys. 1. Problem z rozdzieleniem skupień graniczących ze sobą – eksperyment 1
Źródło: opracowanie własne.

Sieć GNG zastosowana do wyróżnienia skupień ma następujące parametry[7]: MN = 50, LI = 5000, WP = 100, EW = 0.09, E2W = 0.06. Jak można zaobserwować, neurony są rozproszone w przestrzeni obiektów zgodnie z ich rozkładem w przestrzeni. Wynik samouczenia się tej sieci jest jednak niezadowalający, ponieważ skupienia nie zostały rozdzielone. Powtarzając wielokrotnie budowę sieci o takich parametrach, poza niewielkim przesunięciem pozycji neuronów, nie można poprawić wyniku klasyfikacji. W praktyce stosowania sztucznych sieci neuronowych stosunkowo często można spotkać się z opinią, że jeżeli nie można rozwiązać jakiegoś problemu, to znaczy, że jest zbyt mało neuronów lub (i) zbyt mało iteracji uczących[8]. Aby zweryfikować to twierdzenie dla sieci GNG, przeprowadzono eksperyment. W kolejnych próbach zwiększano liczbę neuronów i liczbę iteracji uczących. Parametry algorytmu zaprezentowano w tab. 1, a wyniki kolejnych prób na rys. 2, 3 i 4. Wnioski płynące z eksperymentów są jednoznaczne. Samo zwiększanie liczby neuronów i liczby iteracji uczących nie prowadzi do poprawy własności sieci. Sieć jest co prawda coraz bardziej wrażliwa i potrafi wyróżniać coraz więcej szczegółów, jednak są to szczegóły o znaczeniu lokalnym. Szczególnie kolejne sieci nie potrafią rozróżnić istniejących skupień. Potrafią jednak rozpoznać lokalne zagęszczenia obiektów, traktując je jako miniskupienia.

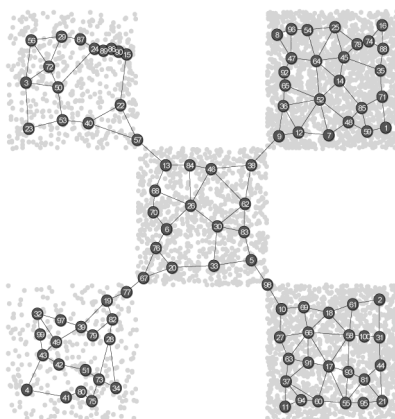
[7] Dla skrócenia zapisu w artykule przyjęto następujące oznaczenia: MN – maksymalna liczba neuronów, LI – liczba iteracji uczenia, EW – szybkość uczenia się neuronu zwycięskiego, E2W – szybkość uczenia się drugiego najlepszego neuronu, WP – maksymalny wiek połączenia.

[8] To twierdzenie opiera się na założeniu, że biologiczne mózgi mają miliony neuronów i dlatego potrafią rozwiązywać złożone problemy. Algorytmy uczenia sztucznych sieci neuronowych są przy tym wolno zbieżne, więc trzeba sieci uczyć długo. W ogólnym przypadku oba te twierdzenia są błędne, ponieważ znane są z badań empirycznych bardzo proste sieci zdolne do rozwiązywania złożonych problemów, a algorytmy uczenia sieci są ciągle udoskonalane i ich szybkość ciągle rośnie.

Tabela 1. Parametry sieci dla eksperymentów 2, 3 i 4

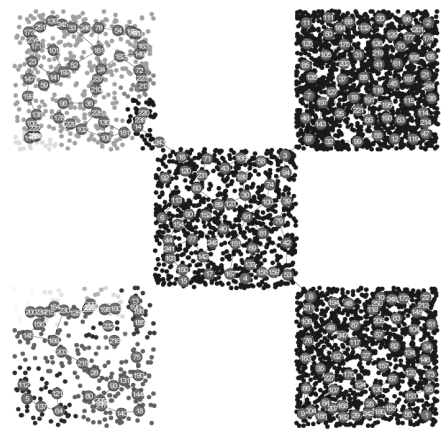
Parametr	Eksp. 2	Eksp. 3	Eksp. 4
MN	100	250	1500
LI	15 000	50 000	5 000 000
EW	0,09	0,09	0,09
E2W	0,06	0,06	0,06
WP	100	100	100
Liczba skupień	1	10	781

Źródło: opracowanie własne.

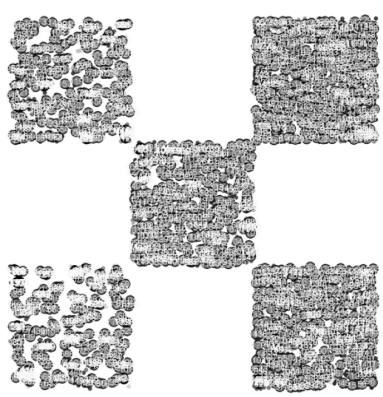


Rys. 2. Wyniki eksperymentu 2

Źródło: opracowanie własne.



Rys. 3. Wyniki eksperymentu 3

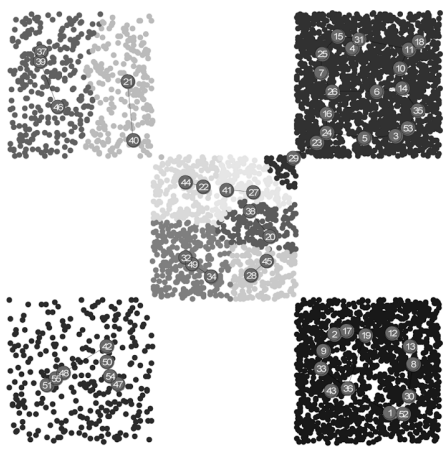


Rys. 4. Wyniki eksperymentu 4

Źródło: opracowanie własne.

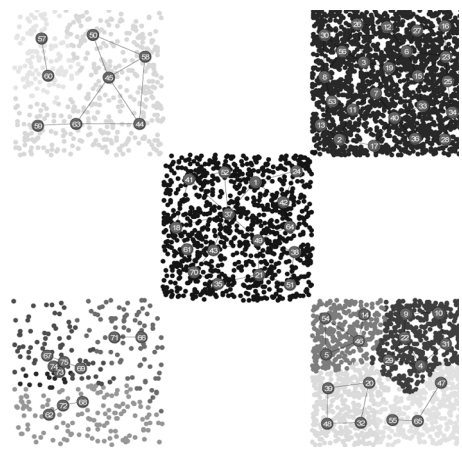
W skrajnym przypadku, gdy liczba neuronów wynosi 1500, sieć rozróżniła 781 skupień. Jedynym rzeczywistym efektem takiego postępowania jest znaczne wydłużenie czasu samouczenia się sieci.

W kolejnej serii eksperymentów modyfikowano nie tylko liczbę neuronów i liczbę iteracji, ale także krok uczenia się neuronu wygrywającego i drugiego najlepszego. Jeżeli krok uczenia jest mały, jest szansa na bardzo dobre odwzorowanie przez neurony struktury danych. Jednocześnie uczenie się sieci jest wolniejsze, a więc wymaga większej liczby iteracji. Parametry te należy modyfikować w zsynchronizowany sposób. Na rysunkach 5, 6 i 7 zaprezentowano wyniki eksperymentów.

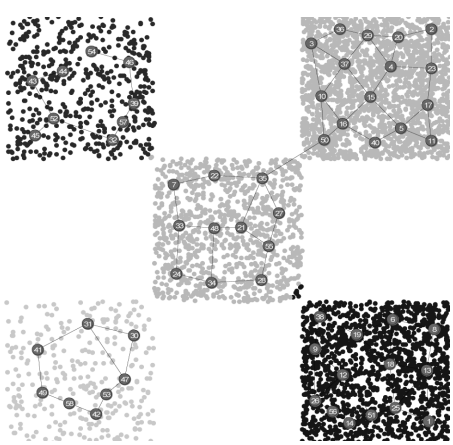


Rys. 5. Wyniki eksperymentu 5

Źródło: opracowanie własne.



Rys. 6. Wyniki eksperymentu 6



Rys. 7. Wyniki eksperymentu 7

Źródło: opracowanie własne.

Tabela 2. Parametry sieci dla eksperymentów 5, 6 i 7

Parametr	Eksp. 5	Eksp. 6	Eksp. 7
MN	55	75	60
LI	140 000	100 000	1 000 000
EW	0,09	0,07	0,05
E2W	0,06	0,006	0,0006
WP	100	100	100
Liczba skupień	10	12	4

Źródło: opracowanie własne.

Krok uczenia w kolejnych eksperymentach został zmniejszony do poziomu minimalnych odległości euklidesowych między obiektami. W badanym zbiorze danych minimalna odległość między obiektami była równa 0.00067, największa = 4.15, średnia = 1.5271.

Wnioski z powyższych eksperymentów wskazują, że ważniejsze od liczby neuronów jest właściwe ustalenie kroku uczenia. Wyniki eksperymentu 7 są znacznie lepsze niż wszystkich pozostałych. Liczba neuronów jest niewielka i wynosi zaledwie 60. Ponieważ krok uczenia jest tu bardzo mały, liczba iteracji uczących musi być duża.

4. Wnioski

Wydaje się, że przeprowadzone eksperymenty wyraźnie wskazują na wagę, jaką należy przykładać do poprawnego ustalenia parametrów samouczenia się sieci GNG. Samo zwiększanie liczby neuronów i liczby iteracji nie daje żadnych pozytywnych efektów. Parametry te są wtórne w stosunku do kroku uczenia neuronu wygrywającego i kolejnego. Optymalizacja kroków uczenia w znaczny sposób wpływa na poprawę jakości grupowania, mimo że w prezentowanym przykładzie wszystkich skupień nie udało się poprawnie rozdzielić. Warto także zauważyć, że mimo znacznej liczby iteracji wymaganej przy małym kroku uczenia, ograniczenie liczby neuronów powoduje, że czas samouczenia się sieci jest krótki.

Literatura

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

Kohonen T., *Self-Organizing Maps*, Springer Series in Information Sciences, Springer-Verlag, Berlin Heidelberg 1997.

Najman K., *Zastosowanie nienadzorowanych sieci neuronowych typu Growing Neural Gas w analizie skupień*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, UE, Wrocław 2009.

THE ASSESSMENT OF INFLUENCE OF PARAMETERS CONTROLLING THE SELF-LEARNING PROCESS OF GNG NETWORK ON ITS ABILITY TO SEPARATE CLUSTERS

Summary: The aim of research is the assessment of influence of parameters controlling the self-learning process of GNG network on the quality of received classification, especially in a situation when clusters are weakly separated. Research describes the theoretical influence of individual parameters controlling the GNG algorithm on the structure and the ability to solve problems through GNG network. The paper also presents the results of simulation research confirming the thesis that the optimal control parameters choice enables increasing the GNG network's ability to identify clusters.