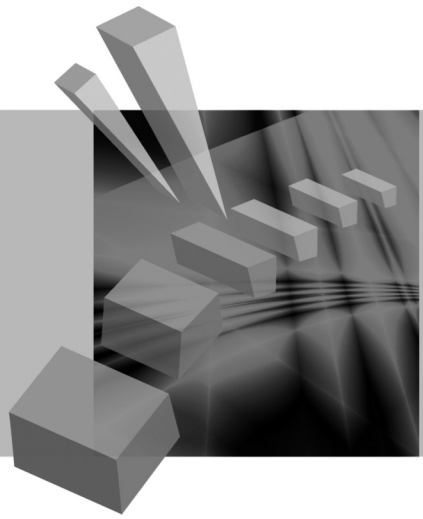
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Krzysztof Najman
Uniwersytet Gdański

GRUPOWANIE DYNAMICZNE Z WYKORZYSTANIEM SAMOUCZĄCYCH SIĘ SIECI GNG

Streszczenie: Wraz ze stale rozwijającą się techniką informatyczną dynamicznie zwiększa się także ilość danych zbieranych w różnych systemach komputerowych. Jedną z cech baz danych tworzonych dynamicznie w systemie *on-line* jest m.in. dynamicznie zmieniająca się struktura grupowa jednostek. Grupowanie jednostek w takiej bazie danych wymaga zastosowania specjalnych metod. W artykule sformułowano wymagania stawiane takiej metodzie, a także zaproponowano użycie w opisywanym celu samouczącej się sieci neuronowej typu GNG. W artykule na podstawie badań teoretycznych i eksperymentu badawczego weryfikowano hipotezę o przydatności sieci GNG w dynamicznym grupowaniu danych.

Słowa kluczowe: analiza skupień, grupowanie dynamiczne, samouczące się sieci neuronowe.

1. Wstęp

Wraz ze stale rozwijającą się techniką informatyczną dynamicznie zwiększa się ilość danych zbieranych w bazach danych. Jedną z cech współczesnych baz danych jest duża zmienność rejestrowanych jednostek. Obraz populacji zmienia się dynamicznie wraz z napływającym strumieniem danych. Struktura populacji badana w jednym momencie może się znacznie zmienić w kolejnym. Fakt ten powoduje określone komplikacje w bieżącej ocenie obserwowanej populacji. Jednym z problemów jest niestabilna struktura grupowa jednostek. Wraz z napływem nowych informacji znane i dobrze określone skupienia tracą na znaczeniu, rozmywają się w innych lub zanikają. Skupienia nieliczne lub słabo określone z czasem stają się bardziej liczne, nabierają znaczenia, stają się lepiej określone. Pojawiają się także całkowicie nowe skupienia. Wiele danych ma także swój czas życia, po przekroczeniu którego nie mają już znaczenia praktycznego i są z bazy danych usuwane.

Aby proces zmian struktury grupowej poprawnie obserwować, należy dokonywać grupowania i korekt w opisie skupień w sposób ciągły. Dla dużych baz danych jest to jednak kłopotliwe zarówno z technicznego, jak i z organizacyjnego punktu widzenia. Jednym z możliwych rozwiązań powyższego problemu jest zastosowanie metody grupowania pracującej bez przerwy. Metoda taka reagowałaby na każdą nową informację i automatycznie dokonywała niezbędnych korekt, minimalizując

koszty finansowe, czasowe i organizacyjne. Wydaje się, że samoucząca się sieć neuronowa typu *Growing Neural Gas* (GNG) może sprostać takim wymaganiom. Celem prezentowanych badań jest weryfikacja powyższej hipotezy.

2. Grupowanie statyczne a dynamiczne

Klasycznie w analizie skupień stosuje się podejście, które można nazwać statycznym, ponieważ w czasie procesu grupowania liczba i struktura grupowanych jednostek nie zmienia się. Sformułowane wnioski z takiego badania są aktualne na moment zakończenia zbierania danych. Dobrze zaplanowane i przeprowadzone badanie pozwoli uzyskać wiedzę o badanej populacji, która powinna być uniwersalna, a więc także trwała.

Z jakościowo inną sytuacją mamy do czynienia w grupowaniu dynamicznym. Grupowanie, w trakcie którego mogą się zmieniać grupowane jednostki i ich struktura grupowa, zostanie nazwane dynamicznym. Wnioski płynące z grupowania w danym momencie są aktualne na ten właśnie moment. Nie jest to sprzeczne z ideą pozyskiwania uniwersalnej wiedzy o badanej populacji. Istnieją bowiem populacje na tyle szybko się zmieniające, że kluczem do ich poznania i zrozumienia jest znajomość dynamiki i kierunków ich zmian.

Przykładem takiej populacji jest rejestr transakcji z udziałem kart kredytowych. System nadzorujący dokonywanie takich transakcji rejestruje setki tysięcy operacji dziennie. Poza innymi czynnościami system ten rozpoznaje transakcje nietypowe – być może związane z próbą oszustwa. Znane metody dokonywania oszustw mają zwykle znany profil[1], odróżniający je od transakcji prawidłowych. Gdy jednak pojawia się jakaś nowa metoda, w bazie danych transakcje wykonane z jej użyciem stają się nowym skupieniem o nieznanym profilu. Nowe skupienie należy rozpoznać natychmiast i stworzyć odpowiadający mu nowy profil. Gdy możliwość dokonywania oszustw nową metodą zostanie przez operatora rozpoznana, zostanie natychmiast zablokowana. Nowe przypadki się już nie pojawią, a skupienie takie wraz ze swoim profilem staje się obiektem o znaczeniu historycznym[2]. Wiedza o badanej populacji ma swój aspekt uniwersalny i trwały, którym są profile poprawnych transakcji. Ma także aspekt dynamiczny związany z napływem nowych informacji, usuwaniem informacji najstarszych i nowymi „pomysłami" oszustów.

Podobnie funkcjonują komputerowe systemy ochrony antywirusowej. Gdy aplikacja konsumencka nie rozpozna wirusa, ale podejrzewa jego istnienie, przesyła kod takiego programu do centrali firmy w celu jego weryfikacji. Tu następuje proces klasyfikacji kodu na podstawie znanych profili tysięcy wirusów. Jeżeli jest to kod nowego, nieznanego wcześniej wirusa, jego profil nie pasuje do profili znanych

[1] Profil ten może być ustalony we wstępnym grupowaniu statycznym lub na poprzednich etapach grupowania dynamicznego.

[2] Stanowi zwykle wzorzec do porównań nowych nietypowych transakcji.

wirusów. Zgłoszenia takie pojawiają się zwykle niemal jednocześnie z wielu stron świata. Jest to proces gwałtowny, gdyż ich liczba w czasie kilku godzin może wzrosnąć od zera do wielu dziesiątek tysięcy. Gdy zostanie przygotowana szczepionka, zostaje ona rozesłana do wszystkich użytkowników i problem automatycznie zanika. W bazie danych mamy nowe skupienie reprezentujące działanie nowego wirusa, jednak nowe obiekty przestają się pojawiać. Programy antywirusowe na komputerach użytkowników radzą sobie samodzielnie z problemem. Wiedza nabyta w drodze grupowania służy do rozwiązania chwilowego problemu. Gdy problem zostaje rozpoznany, znalezione zostaje jego rozwiązanie, a narzędzia informatyczno-techniczne uniemożliwiają jego ponowne pojawienie się.

Rozwiązywanie problemów powyższego typu wymaga umiejętności grupowania i klasyfikacji dużych i dynamicznie zmieniających się baz danych. Zastosowana metoda grupowania powinna charakteryzować się przynajmniej kilkoma cechami: 1) powinna być bardzo szybka. Jeżeli w bazie danych następuje wiele zmian w ciągu sekundy, w tym samym czasie musi być wykonane grupowanie; 2) powinna być oszczędna. Klasyczne metody grupowania wymagają np. wyznaczenia macierzy odległości między wszystkimi obiektami. Jeżeli są ich setki tysięcy, a czasem miliony, może to być niewykonalne w praktyce lub sprzeczne z warunkiem pierwszym; 3) powinna być wysoce autonomiczna. Sama szybkość zmian powoduje, że ewentualna ingerencja w algorytm lub jego parametry powinna być ograniczona do minimum. Przede wszystkim algorytm taki powinien autonomicznie ustalać liczbę skupień, powinien być niewrażliwy na pojedyncze jednostki nietypowe, a jednocześnie szybko tworzyć skupienie, gdy liczba obiektów nietypowych rośnie. Może to bowiem oznaczać pojawienie się nowej prawidłowości; 4) metoda musi się charakteryzować bardzo dobrymi własnościami uzyskanej struktury grupowej.

3. Samouczące się sieci neuronowe typu GNG

Na podstawie wyników badań teoretycznych i empirycznych można sądzić, że sieci neuronowe typu GNG (*Growing Neural Gas*) mogą sprostać stawianym wymaganiom [Fritzke 1994; Prudent, Ennaji 2005; Jirayusakul, Auwatanamongkol 2007]. Wykazano, że sieć ta spełnia warunek oszczędności i jakości grupowania [Qin, Suganthan 2004; Najman 2009; 2010]. Sieć GNG jest oszczędna, ponieważ jej struktura dynamicznie dopasowuje się do skali złożoności badanego problemu. Liczba neuronów wzrasta od dwóch i jest powiększana, a czasem pomniejszana, do momentu ustabilizowania się własności sieci. Nowe neurony pojawiają się na sieci jedynie w tych miejscach przestrzeni, w których sieć popełnia największe błędy.

Jakość uzyskanej struktury grupowej przy zastosowaniu sieci GNG jest bardzo wysoka [Najman 2009; 2010; Migdał-Najman 2009]. Sieć potrafi rozpoznać skupienie o dowolnej konfiguracji w przestrzeni niezależnie od liczby wymiarów przestrzeni cech. Wyjątkiem jest sytuacja, gdy skupienia są nieseparowalne. Sieć GNG ma bowiem skłonność do łączenia takich skupień. Metody przycinania sieci GNG poprawiają, jednak nie rozwiązują całkowicie tego problemu.

Budowa sieci GNG jest także w wysokim stopniu autonomiczna, choć wymaga *a priori* ustalenia parametrów kontrolnych. Są to: maksymalna liczba neuronów, maksymalny czas życia neuronu, poziom błędu kwantyzacji przerywający proces samouczenia się sieci, liczba iteracji uczących, krok uczenia neuronu wygrywającego i drugiego najbliższego. W grupowaniu dynamicznym ustalenie większości z tych parametrów jest dość proste. Maksymalną liczbę neuronów można ustalić na wysokim poziomie. Ze względu na własność oszczędności sieci i tak nie zostanie on osiągnięty. Minimalny poziom błędu kwantyzacji można ustalić na poziomie 0. Liczbę iteracji uczących można ustalić na poziomie $+\infty$. Krok uczenia z kolei powinien być dostosowany do wymagań co do precyzji grupowania. Mały krok oznacza potencjalnie lepsze grupowanie kosztem czasu potrzebnego na jego uzyskanie. Z powyższych rozważań wynika, że jedynie dwa parametry mają realny wpływ na działanie sieci GNG przy grupowaniu dynamicznym. Są to maksymalny wiek neuronu odpowiadający za szybkość wzrostu i zanikania neuronów i krok uczenia.

Do zweryfikowania pozostaje kluczowy element dobrego algorytmu grupowania dynamicznego – tj. szybkość jego działania.

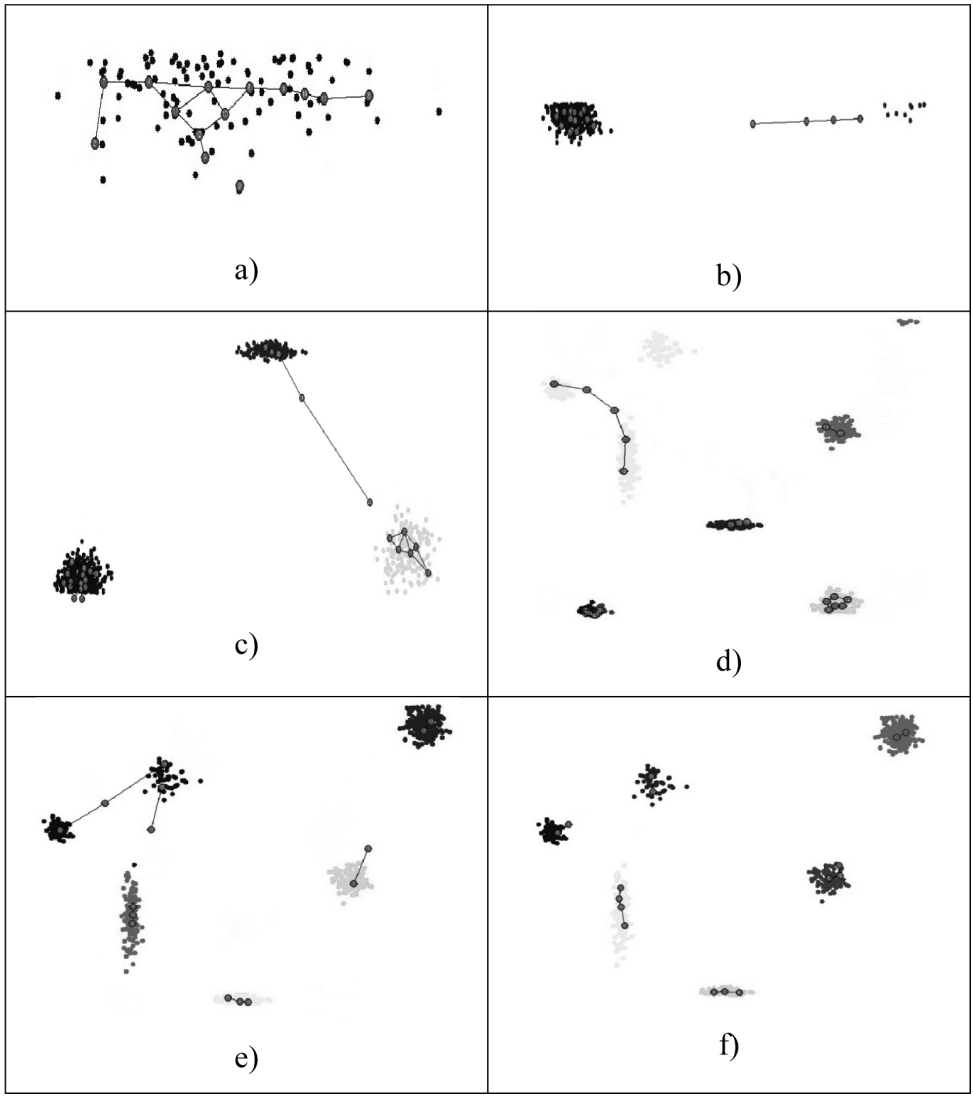
4. Eksperyment badawczy

W celu weryfikacji możliwości zastosowania sieci neuronowej typu GNG w grupowaniu dynamicznym przygotowano eksperyment[3]. Wygenerowano dynamiczną bazę danych[4], złożoną z od 10 do 90 000 jednostek, o zmiennej strukturze grupowej od 1 do 8 jednocześnie istniejących skupień. Jednostki w skupieniach mają ograniczony czas ważności. Gdy w bazie jest już 90 000 jednostek, istniejące najdłużej zostają usunięte. Dla wstępnie istniejących 10 jednostek zbudowano sieć GNG o następującej strukturze: maksymalna liczba neuronów wynosi 1000, maksymalna liczba iteracji jest nieskończona, czas życia neuronu wynosi 51 iteracji, nowy neuron wstawiany jest co 40 iteracji, krok uczenia neuronu wygrywającego wynosi 0,05, krok uczenia drugiego najlepszego neuronu wynosi 0,01, minimalny błąd kwantyzacji jest równy zeru. Po 1000 iteracji wstępnych, dla których sieć uczyła się pierwszych 10 jednostek, co sekundę dołączano do bazy danych losową liczbę (od 1 do 100) nowych jednostek, nie przerywając pracy sieci GNG. Na rysunku 1a pokazano jednostki bazy danych po pierwszych kilku sekundach jej istnienia wraz z siecią GNG. Na tym etapie istnieje tylko jedno skupienie. Po kolejnych kilkunastu sekundach pojawiają się nowe jednostki leżące daleko od pierwszego skupienia.

---

[3] Niestety nie jest dostępna publicznie empiryczna, dynamiczna baza danych.

[4] Baza danych jest tu dynamiczna w takim sensie, w jakim zdefiniowano proces grupowania dynamicznego. Będzie się ona zmieniała w czasie procesu grupowania. Niestety w dokumencie papierowym nie jest możliwe graficzne przedstawienie procesu działania bazy danych. Kilka ważnych etapów jej działania pokazano na rys. 1.



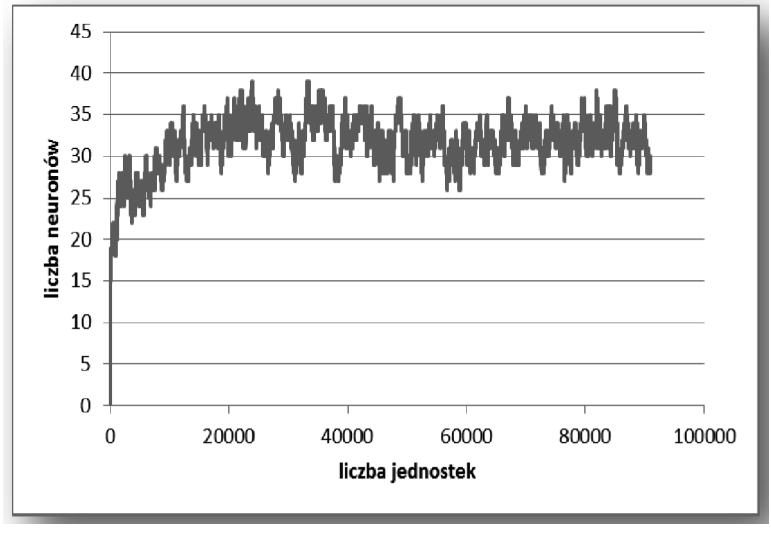
Rys. 1. Przebieg grupowania dynamicznego

Źródło: opracowanie własne.

Na rysunku 1b przedstawiono moment, w którym sieć przestaje ignorować pojedyncze nowe obiekty i zaczyna się uczyć nowego skupienia. Po pewnym czasie istnieją już trzy dobrze wykształcone skupienia, których łączna liczebność przekracza 90 000. Jednostki z pierwszego skupienia zaczynają być usuwane. Na rysunku 1c można zaobserwować moment zanikania skupienia pierwszego. Po kolejnych

sekundach zanikać zaczynają także jednostki z drugiego skupienia, a jednocześnie kształtują się kolejne 4 skupienia. Sieć GNG zaczyna redukować liczbę neuronów odwzorowujących jednostki w pierwszych dwóch skupieniach, przenosząc je do nowo powstałych skupień. Na rysunku 1f przedstawiono ostateczny wynik grupowania uzyskanego po 1000 iteracji od wstawienia ostatniej nowej jednostki do bazy danych.

Obserwując liczbę neuronów w stosunku do liczby grupowanych jednostek, należy stwierdzić, że sieć jest bardzo oszczędna i stabilizuje się na poziomie 32-35 neuronów. Zależność tę prezentuje rys. 2.

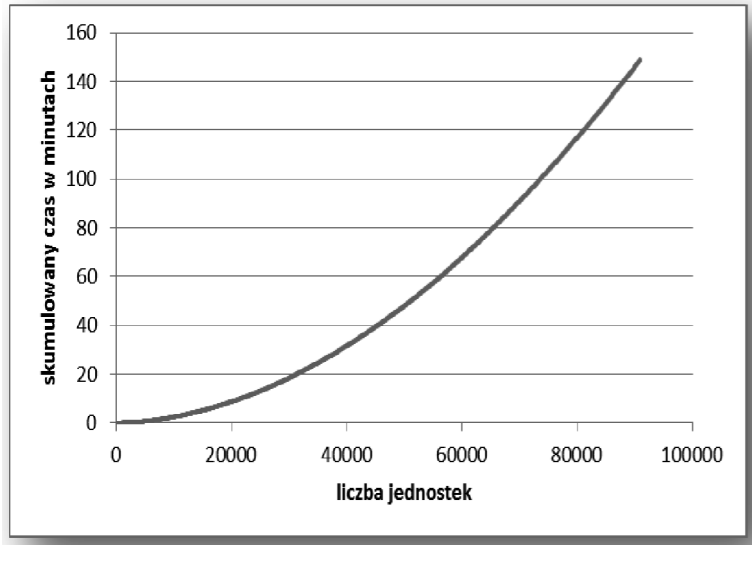


Rys. 2. Liczba neuronów na sieci GNG a liczba jednostek

Źródło: opracowanie własne.

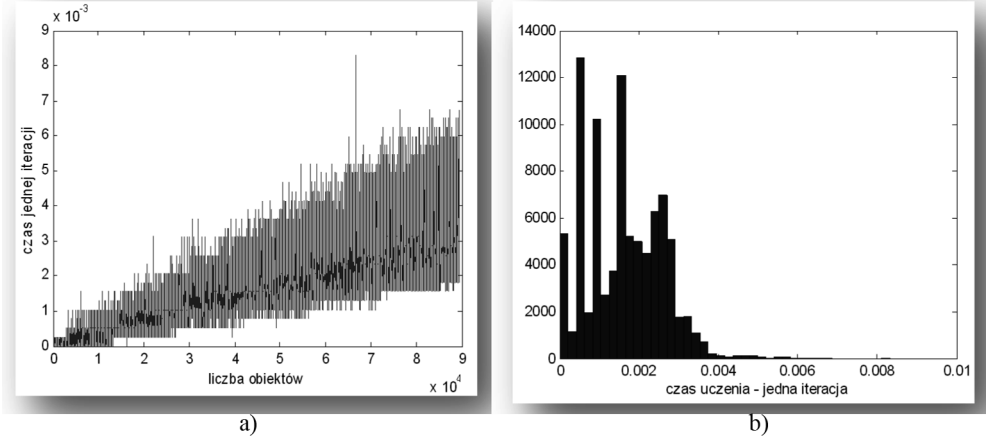
Czas uczenia się sieci jest wysoce niezależny od liczby jednostek w bazie danych. W jednej iteracji uczą się jedynie dwa neurony: neuron wygrywający (znajdujący się najbliżej obserwowanej jednostki) i drugi najbliższy. Liczba modyfikowanych neuronów nie zależy więc od liczby jednostek ani od rozmiaru samej sieci. Proces uczenia się sieci GNG nieco zwalnia wraz ze wzrostem rozmiaru sieci, ale jest to związane z koniecznością aktualizacji macierzy wieku neuronów i macierzy połączeń neuronów, których rozmiar zależy od liczby neuronów. Zależność czasu uczenia sieci od liczby jednostek w bazie danych przedstawia rys. 3. Wzrost rozmiarów sieci wpływa na rozproszenie czasu uczenia sieci. Na rysunku 4a pokazano czas jednej iteracji uczącej w zależności od liczby jednostek w bazie danych. Na rysunku 4b z kolei pokazano rozkład czasu pojedynczej iteracji uczącej. Wraz ze wzrostem liczby jednostek czas jednej iteracji uczącej wzrasta i coraz bardziej się różnicuje. Jednocześnie na histogramie pojawiają się wyraźne piki przy bardzo krótkich cza-

sach uczenia. Głębsza analiza pracy sieci wskazuje, że gdy nie są wstawiane ani usuwane neurony z sieci, czasy pojedynczych iteracji są bardzo krótkie i nie zależą od liczby jednostek. Jednocześnie w momencie wstawiania lub usuwania neuronu z sieci czas ten wydłuża się wraz ze wzrostem liczby jednostek i neuronów.



Rys. 3. Czas uczenia sieci GNG a liczba jednostek

Źródło: opracowanie własne.



a) b)

Rys. 4. a) Czas jednej iteracji uczącej w zależności od liczby jednostek, **b)** Rozkład czasu pojedynczej iteracji uczącej

Źródło: opracowanie własne.

5. Wnioski

Prezentowane rozważania teoretyczne, a także przedstawiony eksperyment badawczy pozwalają mieć nadzieję, że samoucząca się sieć neuronowa typu GNG może być skutecznym narzędziem dynamicznego grupowania danych. Wykazano, że spełnia ona podstawowe wymagania stawiane metodzie grupowania dynamicznego. Jest bardzo szybka. Zmierzone czasy uczenia w eksperymencie na przeciętnym komputerze PC wahały się od 0,0001 do 0,006 sekundy na jedną iterację uczącą, co oznacza do 10 000 iteracji uczących na sekundę. Jest to bardzo duża liczba umożliwiająca grupowanie tysięcy jednostek w tym czasie. Sieć jest bardzo oszczędna i nie rozwija swojej struktury ponad rzeczywiste potrzeby. Jest wysoce autonomiczna. Większość parametrów sterujących jej pracą nie wymaga modyfikacji w trakcie procesu samouczenia się. Jest także bardzo skuteczna w rozpoznawaniu struktury grupowej, o ile tylko skupienia są separowalne. W prezentowanym eksperymencie, w którym skupienia były separowalne, sieć rozpoznała je bezbłędnie w ciągu najwyżej kilku sekund pracy.

Dalszych badań wymagają kwestie regulacji kroku uczenia neuronów, który może wymagać modyfikacji w procesie samouczenia się sieci, gdy pojawią się w bazie danych obiekty jakościowo inne od przewidywań badacza. Dalszych badań wymaga także problem rozdzielenia skupień słabo seprowalnych. Warto także zauważyć, że nie istnieje w tej chwili oprogramowanie komputerowe umożliwiające szerszej grupie badaczy analizę własności tej sieci, jak również stosowanie jej w praktyce[5]. Niezależnie od wskazanych dalszych kierunków badań wydaje się, że proponowana metoda może być z powodzeniem włączona do zestawu procedur grupowania danych stosowanych w badaniach empirycznych.

Literatura

Jirayusakul A., Auwatanamongkol S., *A supervised growing neural gas algorithm for cluster analysis*, „International Journal of Hybrid Intelligent Systems" 2007.

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

Migdał-Najman K., *Analiza porównawcza własności nienadzorowanych sieci neuronowych typu Self Organizing Map i Growing Neural Gas w analizie skupień*, [w:] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, Taksonomia 16, Wydawnictwo UE, Wrocław 2009.

Najman K., *Zastosowanie nienadzorowanych sieci neuronowych typu Growing Neural Gas w analizie skupień*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, Taksonomia 16, Wydawnictwo UE, Wrocław 2009.

Najman K., *Ocena wpływu parametrów sterujących procesem samouczenia się sieci GNG na ich zdolność do separowania skupień*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, Taksonomia 17, Wydawnictwo UE, Wrocław 2010.

[5] W pracy wykorzystano autorskie oprogramowanie przygotowane w systemie Matlab.

Qin A.K., Suganthan P.N., *Robust growing neural gas algorithm with application in cluster analysis*, „Neural Networks" 2004, nr 17 t. 8-9.

Prudent Y., Ennaji A., *An incremental growing neural gas learns topologies*, Proceedings of International Joint Conference on Neural Networks, 2005.

A DYNAMIC GROUPING BASED ON SELF-LEARNING GNG NETWORKS

Summary: Along with the constantly evolving IT technology, there is a rapid increase of the amount of data collected in different computer systems. One of the characteristics of the dynamically created online databases is, inter alia, the structure of a dynamically changing group of individuals. The grouping in such a database requires the use of special methods. This article formulates the requirements for such a method and proposes the use in the described purpose self-learning GNG type networks. The article, based on theoretical studies and research experiment, verifies the hypothesis about the suitability of GNG network in a dynamic clustering.

Keywords: cluster analysis, dynamic clustering, self-learning neural networks.