PRZEGLĄD STATYSTYCZNY
R. LVIII – ZESZYT 3-4 – 2011

KRZYSZTOF NAJMAN

GRUPOWANIE DYNAMICZNE Z WYKORZYSTANIEM SIECI GNG

1. WSTĘP

Od początku lat 90-tych XX wieku obserwuje się stały, dynamiczny wzrost liczby baz danych i zbieranych w nich informacji. Rozwój gospodarek wielu krajów, spadek cen komputerów, większa liczba sprawnych aplikacji, rozwój różnych typów sieci komputerowych Intranet, a także rozwój globalnej sieci Internet spowodowały z jednej strony wzrost zapotrzebowania na informacje, a z drugiej stworzyły możliwości ich zbierania i przechowywania. Rozwój ten przyczynia się do poszerzania wiedzy o otaczającym nas świecie. Umożliwia między innymi lepsze zrozumienie zachodzących procesów społecznych, ekonomicznych, sprawniejsze zarządzanie państwem czy przedsiębiorstwem. Sama jednak rosnąca liczba zbieranych danych nie wpływa bezpośrednio na zwiększanie się zasobów wiedzy. Dane muszą być przetworzone w informacje, a dopiero te w wiedzę. Paradoksalnie, proces pozyskiwania wiedzy nie jest obecnie łatwiejszy, mimo ogromnego wzrostu liczby dostępnych danych. Sama objętość i własności istniejących baz stają się poważnym problemem w analizie zgromadzonych danych.

Jednym z problemów związanych z analizą danych zawartych we współczesnych bazach jest duża zmienność ich zawartości. Rejestracja nowych jednostek może następować w sposób ciągły, dynamicznie zmieniając obraz obserwowanej populacji. Dynamicznie może się także zmieniać struktura grupowa rejestrowanych jednostek. Wraz z upływem czasu i napływem nowych danych znane i dobrze określone skupienia mogą tracić na znaczeniu lub rozmyć się w innych. Skupienia zawierające nieliczne jednostki, słabo określone (rozmyte) mogą z czasem stać się bardziej liczne, dobrze określone a nawet dominujące. Mogą pojawić się również całkowicie nowe skupienia. Możliwe jest także, aby dane posiadały przypisany im czas ważności, po którym tracą swoje znaczenie i wpływ na bieżącą strukturę populacji.

Powyższe własności bazy danych powodują, że sama struktura grupowa populacji może być dynamiczna i zmieniać się w sposób ciągły. Aby poprawnie obserwować proces zmian struktury grupowej badanej populacji należy dokonywać grupowania obiektów i korekt w opisie skupień także w sposób ciągły – wraz z napływaniem nowych danych. Konieczne jest zastosowanie metody grupowania zdolnej do reagowania na każdą nową informację i automatycznie dokonującej niezbędnych korekt w opisie istniejącej struktury grupowej. Opis ten musi być dostępny w dowolnym momencie istnienia bazy danych.

Celem prezentowanych badań jest weryfikacja możliwości zastosowania samouczącej się sieci neuronowej o zmiennej strukturze typu *Growing Neural Gas* w dynamicznym grupowaniu jednostek.

2. GRUPOWANIE STATYCZNE A GRUPOWANIE DYNAMICZNE

W literaturze dotyczącej metod grupowania danych pojęcie grupowania dynamicznego pojawia się głównie w kontekście analizy szeregów czasowych (Wang X., Smith K., Hyndman R. [1997], Zamir O., Grouper O.E. [1999], Guha S., Mishra N., Motwani R., O'Callaghan L. [2000], Babcock B., Babu S., Datar M., Motwani R., Widom J. [2002]). Staje się ono elementem szczególnej analizy korelacji (Fenn D.J., Porter M.A., Mucha P.J., McDonald M., Williams S., Johnson N.F., Jones N.S. [2010]). Obserwuje się jednocześnie znaczną liczbę szeregów czasowych i analizuje ich podobieństwo, klasyfikując poszczególne szeregi do niewielkiej liczby klas. Samo pojęcie grupowania dynamicznego jest tu jednak rozumiane bardzo wąsko. Dynamiczne są, bowiem jedynie badane szeregi czasowe. Sam proces grupowania jest statyczny. Stosuje się tu klasyczne metody grupowania hierarchicznego lub metodę k-średnich. Rozszerzenie powyższego podejścia do grupowania dynamicznego można znaleźć w pracy: Guedalia I.D., London M., Werman M. [1999] gdzie autorzy dzielą metody grupowania na metody wsadowe (batch) i ciągłe (online). O metodach wsadowych mówi się wtedy, gdy dokonuje się grupowania po wprowadzeniu do bazy danych większego zbioru jednostek[1]. W takim przypadku ignoruje się wszystkie pośrednie struktury grupowe, które mogły się pojawić między momentami dokonywania analizy. O grupowaniu ciągłym mówi się, gdy grupowanie jest dokonywane każdorazowo po wprowadzeniu do bazy nowej jednostki. Autorzy ograniczają jednak swoją definicję zakładając, że nowe dane pojawiają się w czasie pojedynczo – w jednej chwili jedna jednostka. Zakładają także, że nowe jednostki mogą się pojawić jedynie w stałych odstępach czasu (Guedalia I.D., London M., Werman M. [1999]). Łatwo zauważyć zbieżność podejść z analizą szeregów czasowych, w których w danej chwili może się pojawić jedynie jedna nowa obserwacja. Grupowanie online jest więc także wąskim rozumieniem grupowania dynamicznego.

Wskazane wyżej próby definicji grupowania dynamicznego wydają się bardzo ograniczone. Nie biorą pod uwagę najważniejszego elementu odpowiadającego za dynamizm procesu grupowania – zmianę jednostek i ich struktury grupowej w czasie grupowania. W powyższych definicjach dynamiczne są jedynie grupowane obiekty, jak w grupowaniu szeregów czasowych, lub uważa się za dynamiczne sekwencyjne powtarzanie procesu grupowania, jak w grupowaniu ciągłym.

Wydaje się, że aby pełniej opisać rozpatrywany problem należałoby wziąć pod uwagę kryterium stałości jednostek i ich struktury grupowej w czasie procesu grupowania. Korzystając z tego kryterium można wyróżnić dwa podejścia do grupowania

[1] Grupowanie powtarza się, gdy do bazy zostanie wprowadzona z góry ustalona liczba jednostek, nie wiadomo jednak, kiedy się to stanie, lub grupowanie powtarza się, co pewien ustalony czas o ile zostanie przekroczona zakładana minimalna liczba nowych jednostek.

i to niezależnie od typu grupowanych obiektów: statyczne i dynamiczne. O grupowaniu statycznym można mówić, gdy podczas procesu grupowania ani zbiór jednostek ani ich struktura grupowa nie zmieniają się. W tej definicji mieszczą się wszystkie powyższe podejścia do grupowania (także te nazywane dynamicznymi). O grupowaniu dynamicznym można mówić wtedy, gdy w czasie procesu grupowania zmieniają się zarówno grupowane jednostki jak i ich struktura grupowa.

Aby grupowanie dynamiczne możliwe było do praktycznego zrealizowania, stosowana metoda grupowania powinna spełniać przynajmniej cztery warunki. Metoda taka musi być:

1. szybka,
2. oszczędna,
3. autonomiczna,
4. zapewniająca wysoką jakość grupowania dla dowolnej konfiguracji przestrzennej jednostek.

Warunek szybkości jest kluczowy. Jeżeli w ciągu sekundy w bazie danych może zmienić się kilkadziesiąt lub kilkaset jednostek, także zmiany struktury grupowej muszą być uwzględnione w tym samym czasie. Metoda musi być także oszczędna, ponieważ grupowaniu podlegają duże bazy danych, coraz częściej liczące miliony przypadków. Metoda grupowania musi oszczędzać zasoby pamięci komputera. Macierz odległości dla 15000 jednostek zajmuje 1717 Mb pamięci RAM. Musi być także oszczędna obliczeniowo. Jeżeli algorytm grupowania wymaga dużej liczby obliczeń, to może wymagać użycia superkomputera lub będzie zbyt wolny, aby nadążyć za zmianami w bazie danych. Metoda powinna być wysoce autonomiczna gdyż sama szybkość zmian w bazie danych powoduje, że ewentualna ingerencja w algorytm lub jego parametry powinna być ograniczona do minimum. W szczególności 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. Metoda musi się także charakteryzować bardzo dobrymi własnościami uzyskanej struktury grupowej niezależnie od konfiguracji przestrzennej jednostek

3. SAMOUCZĄCA SIĘ SIEĆ NEURONOWA GNG

Konstrukcję sieci samouczącej się typu Growing Neural Gas zaproponował B. Fritzke w 1994. (Fritzke B. [1994]). Jest to rozwinięcie znanych wcześniej sieci samouczących się typu SOM (*Self Organizing Map*) (Kohonen T. [1997]). Zasadniczą wadą sieci SOM jest stała struktura założona a priori przez badacza. Jest to szczególnie uciążliwe w dużych badaniach empirycznych gdzie może istnieć wiele skupień o złożonej strukturze przestrzennej. Aby w takim przypadku poprawnie wyróżnić skupienia sieć SOM może wymagać dużej liczby neuronów – trudno jednak przewidzieć ile. Konsekwencją dużych rozmiarów sieci jest długi czas uczenia się sieci i spadająca

wraz ze wzrostem skali badanego problemu efektywność procesu grupowania (Migdał Najman K., Najman K. [2008], Najman K. [2009]).

Istotą budowy sieci GNG jest konstrukcja sieci maksymalnie oszczędnej, bez zbędnych neuronów, rozłożonych jedynie w tej części przestrzeni, w której znajdują się obserwowane jednostki. Sieć w procesie samouczenia się powinna sama korygować swoją strukturę zgodnie z potrzebami. Dla prostej struktury grupowej (np. kilka sferycznych, separowalnych skupień) sieć powinna mieć niewielkie rozmiary niezależnie od liczby obserwowanych obiektów. Rozmiar sieci powinien wzrastać jedynie wraz ze wzrostem liczby skupień i stopniem komplikacji ich struktury przestrzennej (np. skupienia niesferyczne, słabo separowalne).

Wydaje się, że cele powyższe udało się osiągnąć w konstrukcji sieci GNG i algorytmie samouczenia się sieci[2]. Na podstawie wyników badań teoretycznych i empirycznych można sądzić, że sieci GNG spełniają warunek oszczędności, autonomiczności i jakości grupowania (Qin A.K., Suganthan P.N. [2004], Najman K. [2009, 2010]). Do zweryfikowania pozostaje kluczowy element dobrego algorytmu grupowania dynamicznego – to jest szybkość jego działania przy założeniu zachowania przez sieć warunków oszczędności, autonomiczności i wysokiej jakości grupowania.

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 10 do 15000 różnych jednostek, z których w jednym momencie istnieje maksymalnie 7500 jednostek. Baza ma zmienną strukturę grupową od 1 do 10 skupień (maksymalnie 9 skupień istniejących jednocześnie). Jednostki w skupieniach mają ograniczony czas istnienia. Gdy w bazie jest już 7500 jednostek, wraz z pojawianiem się nowych jednostek te istniejące najdłużej zostają systematycznie usuwane. Wszystkie jednostki opisane są dwoma cechami[5] mierzonymi na skali ilorazowej. Struktura grupowa wszystkich jednostek została zaprezentowana na rysunku 1.

Do grupowania dynamicznego zastosowano sieć GNG o nieograniczonej maksymalnej liczbie neuronów, maksymalnym błędzie kwantyzacji równym zero i nieograniczonej liczbie iteracji uczących. Parametry te gwarantują, że proces samouczenia się sieci będzie ciągły. Zostanie on zatrzymany sztucznie 100 iteracji uczących po wprowadzeniu do bazy ostatniej jednostki. Nowy neuron wstawiany jest co 70 itera-

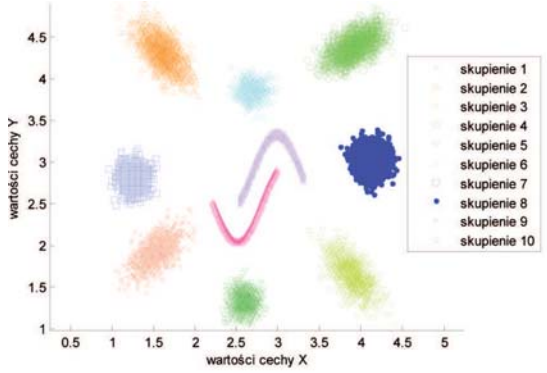
[2] Szczegóły algorytmu budowy sieci GNG, wraz ze schematem blokowym znajdują się w artykule Najman K. [2009].

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

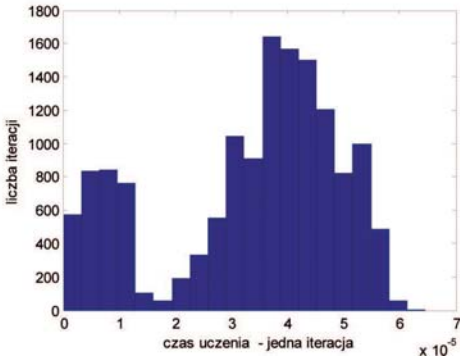
[4] Baza danych jest tu dynamiczna w takim sensie, w jakim zdefiniowano proces grupowania dynamicznego. Będzie się ona zmieniała w czasie procesu grupowania.

[5] Problem jest jedynie dwuwymiarowy z powodu niemożliwości wizualizacji sieci GNG dla wyższej liczby wymiarów. Sieć zachowuje ten sam wymiar neuronów, co wymiar przestrzeni cech badanych jednostek.

Rysunek 1 Struktura grupowa



Rysunek 2 Rozkład czasu uczenia jednej iteracji



Źródło: opracowanie własne

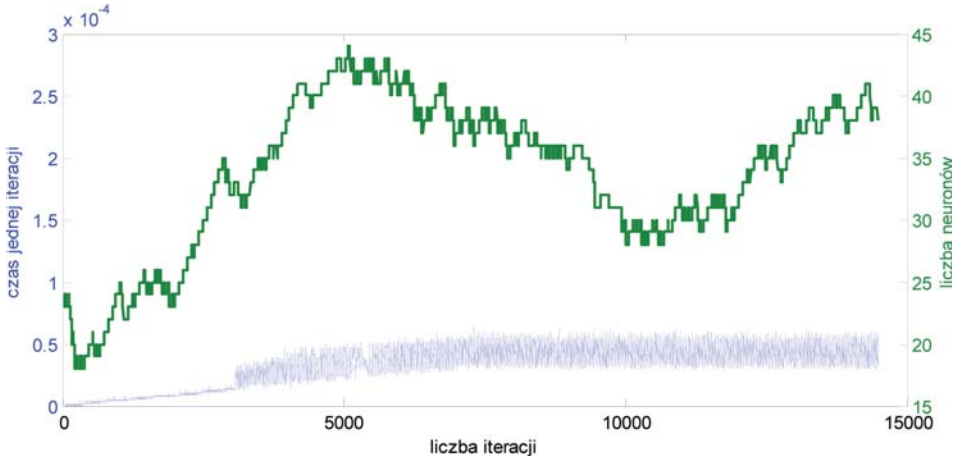
cji przy maksymalnym czasie życia neuronu równym 71 iteracji[6]. Zastosowano stały krok uczenia neuronu wygrywającego równy 0,01 i drugiego najlepszego równy 0,009. Pierwsza z wartości jest nieznacznie większa niż przeciętna odległość euklidesowa między obserwowanymi jednostkami, co zapewnia dużą szybkość przesuwania się neuronu zwycięskiego w przestrzeni. Druga wartość jest rzędu przeciętnej odległości między jednostkami, co pozwala na dokładne dopasowanie neuronu do obserwowanej jednostki. Po wprowadzeniu do bazy nowej jednostki po 142 iteracjach uczących (dwukrotnie wstawiony nowy neuron i usunięty najstarszy) dokonywany jest pomiar zgodności uzyskanej struktury grupowej ze wzorcem w oparciu o skorygowany współczynnik Rand'a.[7] Mierzony jest czas wykonania każdej iteracji uczącej.

Ponieważ jednostek w bazie jest 15000, a po wstawieniu każdej wykonywane są 142 iteracje uczące, to w eksperymencie sieć wykonała łącznie 142*15000=2130000 iteracji uczących. Łączny czas wszystkich iteracji uczących wyniósł niemal dokładnie 70 minut, co oznacza, że przeciętny czas jednej iteracji uczącej wyniósł 0,000033 minuty. Sieć jest w stanie wykonać ponad 30328 iteracji uczących na minutę i ponad 507 na sekundę. Oznacza to w praktyce ponad 3 procesy dodawania i usuwania neuronu na sekundę. Ponieważ jeden neuron może odpowiadać za poprawne grupowanie wielu jednostek, oznacza to możliwość uwzględnienia wielu nowych jednostek co sekundę. Sieć powinna zauważyć w tym czasie zmianę w strukturze grupowej wywołaną przez nowe jednostki. Rozkład czasów wykonania pojedynczej iteracji uczącej został zaprezentowany na rysunku 2. Jest on wyraźnie bimodalny. Pierwsze maksimum dotyczy jedynie początkowego okresu pracy sieci. Gdy jednostek jest bardzo niewiele

[6] Dłuższy czas życia neuronu niż częstotliwość wstawiania nowych neuronów powoduje zwykle, że sieć szybciej reaguje na pojawienie się nowych jednostek a jednocześnie mniej chętnie pozbywa się zbędnych neuronów.

[7] Każda sekwencja składająca się ze wstawienia nowej jednostki do bazy, 142 iteracji uczących sieci i prezentacji wyników jest w badaniu nazwana iteracją roboczą. Na wszystkich wykresach prezentowane są liczby iteracji roboczych.

Rysunek 3 Liczba neuronów i czas uczenia jednej iteracji



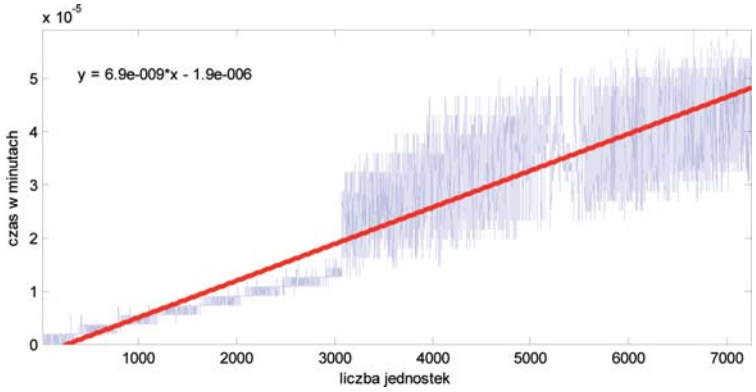
Źródło: opracowanie własne

i neuronów także, czasy uczenia są bardzo krótkie i wynoszą poniżej 0,00001 minuty. Przeciętne czasy wykonania pojedynczej iteracji uczącej dla wszystkich 15000 iteracji roboczych pokazano na rysunku 3. Ich zróżnicowanie rośnie w początkowej fazie budowy sieci. Wzrost ten nie jest jednak bezpośrednio związany z liczbą neuronów a raczej ze strukturą grupową jednostek. W początkowej fazie jest tylko jedno a później dwa skupienia, które są sferyczne i gęste. Neuronów jest niewiele stąd duża szybkość uczenia się i małe wahania czasu pojedynczej iteracji uczącej. Wyraźny skok następuje w momencie, gdy rozpoczynają się pojawiać jednostki należące do skupień o złożonej strukturze przestrzennej (dwa u-kształtne skupienia w centrum rysunku 1). Gdy sieć jest większa a struktura grupowa bardziej złożona sieć uczy się szybko w iteracjach, w których nie są wymieniane neurony a wolniej, gdy wstawiany lub usuwany jest neuron. Czas potrzebny na usunięcie neuronu jest krótszy niż czas potrzebny na jego wstawienie, co jest związane z liczbą koniecznych do wykonania operacji[8]. W ten sposób powstają 3 różne typy iteracji, różniące się także czasem wykonania, co jest widoczne na rysunku. Istotną obserwacją jest także niewielka zależność czasu iteracji od liczby neuronów i liczby jednostek. Wzrost czasu uczenia jednej iteracji jest w przybliżeniu liniową funkcją liczby grupowanych jednostek. Zależność ta zaprezentowana jest na rysunku 4.

W całym czasie pracy sieci maksymalna liczba neuronów wyniosła 44. Jest to bardzo niewielka liczba wskazująca, że jeden neuron odwzoruje nawet 170 jednostek w bazie danych. Przez większość czasu pracy sieci liczba neuronów była jeszcze

[8] Aby usunąć neuron należy znaleźć ten, który przekracza maksymalny czas życia, usunąć go i połączyć ze sobą neurony, z którymi połączony był ten usuwany. Aby wstawić nowy neuron należy znaleźć neuron o największym błędzie kwantyzacji i drugi najgorzej dopasowany, wstawić między nie nowy neuron, wyznaczyć dla niego przeciętny błąd kwantyzacji na podstawie błędów neuronów, między które jest wstawiany nowy, zaktualizować macierz czasu życia neuronów (powiększyć jej rozmiar).

Rysunek 4 Zależność czasu uczenia jednej iteracji od liczby grupowanych jednostek



Źródło: opracowanie własne

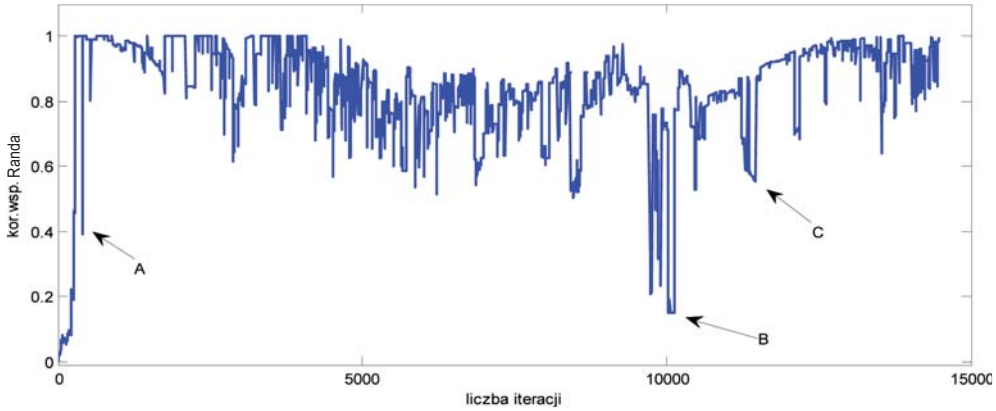
mniejsza i liczyła przeciętnie 33 neurony. Liczba ta w niewielkim tylko stopniu zależy od liczby jednostek w bazie. Obserwując liczbę neuronów w kolejnych iteracjach roboczych (por. rysunek 3) łatwo zaobserwować, że wzrost liczby neuronów zależy w największym stopniu od złożoności samej struktury grupowej obserwowanych jednostek. W początkowej fazie pracy sieci, gdy liczba jednostek rośnie i kształtuje się 5 pierwszych skupień liczba neuronów systematycznie rośnie. Powyżej pięciotysięcznej iteracji liczba jednostek już nie wzrasta i przez pewien czas nie komplikuje się także struktura grupowa. Można wtedy zaobserwować spadek liczby neuronów. Ponowny wzrost liczby neuronów obserwowany w końcowej fazie pracy sieci jest ponownie związany ze zwiększeniem się komplikacji w strukturze grupowej jednostek. Jest to spowodowane zanikaniem początkowych skupień. Sieć przeznacza część neuronów na odwzorowywanie jednostek w starych skupieniach, nawet gdy istnieje tam tylko kilka jednostek, a jednocześnie coraz lepiej odwzorowuje jednostki z nowych dużych skupień.

Interesujące jest kształtowanie się wartości skorygowanego współczynnika Randa w kolejnych iteracjach roboczych[9]. Wartość współczynnika Randa w początkowych iteracjach jest niska, co związanej jest z losowymi początkowymi współrzędnymi neuronów. Z tego powodu sieć wymaga pewnej liczby iteracji, aby zacząć rozpoznawać jednostki. Po około 220 iteracjach neurony przesunęły się w kierunku obserwowanych jednostek i następuje skokowy wzrost wartości współczynnika Randa do 1. Po niewielkiej liczbie iteracji następuje gwałtowny spadek do poziomu 0,4 (oznaczony na rysunku 5 symbolem A) co jest związane z pojawieniem się nowych jednostek w nowym skupieniu. Już po jednej iteracji roboczej sieć nauczyła się rozpoznawać nowe skupienie. Sytuacja taka powtarza się wielokrotnie w całym czasie pracy sieci. Skokowe zmiany wartości współczynnika Randa są zawsze wywołane dużą zmianą w strukturze grupowej jednostek. Szczególnie trudne momenty zostały zaznaczone na

[9] Wartości skorygowanego współczynnika Randa zawierają się w przedziale <0,1>, gdzie 0 oznacza całkowitą niezgodność badanego podziału ze wzorcem a 1 całkowitą zgodność.

rysunku 5 symbolami B i C. We wszystkich tych momentach następuje znaczna zmiana struktury grupowej jednostek. Struktura grupowa w iteracji roboczej oznaczonej literą B jest przedstawiona na rysunku 6.

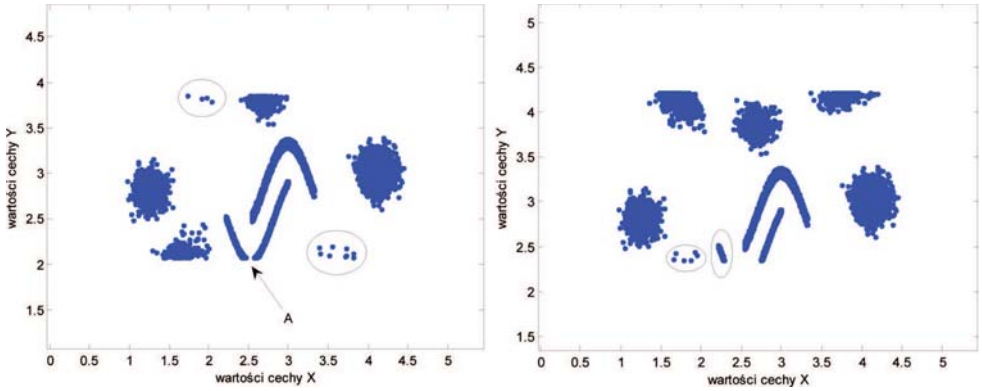
Rysunek 5 Wartości skorygowanego współczynnika Randa



Rysunek 1. Źródło: opracowanie własne

Rysunek 6 Struktura grupowa jednostek po 10000 iteracji

Rysunek 7 Struktura grupowa jednostek po 11000 iteracji



Rysunek 2. Źródło: opracowanie własne

Jest to bardzo trudna struktura grupowa. Jedno ze starych skupień zanika i składa się z zaledwie 10 jednostek (oznaczone elipsą w prawym dolnym rogu na rys.6). Jednocześnie pojawia się nowe skupienie (oznaczone elipsą w górnym lewym rogu na rys.6). W tym samym czasie u-kształtne skupienie w centrum wykresu zaczyna zanikać i rozdziela się na dwa niezależne skupienia (oznaczone symbolem A). Podobnie w momencie oznaczonym na rysunku 5 literą C następuje znacząca zmiana struktury grupowej pokazana na rysunku 7, gdy zanikają jednocześnie dwa skupienia (oznaczone

elipsami). Warto zauważyć, że gdy przez pewną liczbę iteracji struktura grupowa jest względnie stabilna (zmieniają się jednostki, ale nie zmieniają skupienia) sieć szybko poprawia swoją jakość i uzyskuje bardzo dobre wyniki grupowania. Przeciętna wartość współczynnika Randa wynosi w całym eksperymencie 0,85. Przez ponad 10% iteracji roboczych jego wartość jest równa 1. Przez 42% iteracji roboczych wartość współczynnika przekracza 0,9. Można uznać, że są to dobre wyniki grupowania jak na tak złożoną sytuację.

5. PODSUMOWANIE

W badaniach dotyczących zastosowania samouczącej się sieci neuronowej o zmiennej strukturze typu GNG skupiano się dotychczas, na jakości uzyskanej struktury grupowej. Wykazano eksperymentalnie, że o ile skupienia są separowalne sieć tego typu może bezbłędnie określić strukturę grupową jednostek niezależnie od konfiguracji przestrzennej skupień. W badaniach tych podkreślano także wysoką autonomiczność sieci GNG. Nie wymaga ona ustalenia liczby skupień, może pracować w sposób ciągły, ponieważ nie wymaga ustalenia żadnych parametrów zatrzymujących pracę sieci. Wymaga jednak zdefiniowania a priori wartości kilku parametrów. Można to zrobić jedynie eksperymentalnie gdyż nie ma ogólnych, formalnych przesłanek ustalania ich wartości. Sieć ta z samego założenia jest oszczędna w stosunku do sieci o stałej liczbie neuronów, gdyż inteligentnie i dynamicznie dopasowuje swoją strukturę do rzeczywistych potrzeb. Nie prowadzono jednak badań nad szybkością uzyskiwania wyników przez sieć GNG oceniając jedynie, że szybkość ta jest wystarczająca do jej praktycznego stosowania. Ponieważ w grupowaniu dynamicznym szybkość działania jest kluczowym elementem wymaga on znacznie głębszego zbadania. Niestety ze względu na dużą komplikację samego algorytmu samouczenia się sieci GNG wyznaczenie jego złożoności obliczeniowej[10] jest niezwykle trudne. Aby ocenić jej szybkość działania w praktycznych zastosowaniach posłużono się eksperymentem. Nie ma on siły dowodu formalnego, niemniej wskazuje, że w badaniach empirycznych zbliżonych do warunków eksperymentu jest to metoda bardzo szybka. Oceniono wpływ liczby neuronów, liczby i struktury przestrzennej skupień oraz liczby grupowanych jednostek na szybkość uczenia się sieci GNG. Na przeciętnym komputerze[11] sieć GNG może wykonać tysiące iteracji uczących na sekundę nawet w najmniej korzystnych warunkach branych pod uwagę w eksperymencie. Mimo zróżnicowanej liczby skupień istniejących w danym momencie, złożonej struktury grupowej części skupień, wzrost czasu wykonania jednej iteracji, w zależności od liczby jednostek i stopnia komplikacji grupowania jest

[10] Dla niektórych algorytmów istnieje możliwość wyliczenia łącznej liczby elementarnych działań matematycznych (najczęściej dodawania) niezbędnych do jego realizacji. Liczba ta jest nazywana złożonością obliczeniową algorytmu.

[11] Wszystkie obliczenia zostały wykonane na komputerze wyposażonym w procesor: Intel(R) Core(TM) i5 CPU M 450 @ 2.40GHz

w eksperymencie niemal liniowy. Pozwala to ocenić szybkość pracy sieci w innych podobnych badaniach, lecz o innej liczbie grupowanych jednostek.

Warto zauważyć, że w prezentowanych badaniach nie uwzględniono wpływu liczby cech opisujących jedną jednostkę (problem wymiarowości) na szybkość uczenia się sieci. Wydaje się także, że w grupowaniu dynamicznym istnieje możliwość wprowadzenia zmiennych wartości niektórych parametrów sieci GNG (np. czas życia neuronu, liczba iteracji między kolejnymi momentami wprowadzania nowego neuronu do sieci, krok uczenia neuronów), zależnych od bieżącej dynamiki zmian liczby jednostek w bazie i dynamiki zmian ich struktury grupowej. Problemy te będą elementem dalszych badań autora.

Przedstawiony powyżej eksperyment wskazuje, że samoucząca się sieć neuronowa typu GNG spełnia wszystkie wymagania stawiane metodzie grupowania dynamicznego i warto włączyć ją do zbioru procedur stosowanych przez badaczy w analizie skupień, szczególnie w grupowaniu dynamicznym.

LITERATURA

[1] Babcock B., Babu S., Datar M., Motwani R., Widom J. [2002], *Models and issues in data stream systems. In Proceedings of the Twentyfirst ACMSIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, June 3-5, Madison, Wisconsin, USA, 1-16, ACM.

[2] Fenn D.J., Porter M.A., Mucha P.J., McDonald M., Williams S., Johnson N.F., Jones N.S. [2010], *Dynamical Clustering of Exchange Rates*, arXiv:0905.4912v2.

[3] Fritzke B. [1994], *Growing cell structures - a self-organizing network for unsupervised and supervised learning*, Neural Networks, 7, 9, 1441-1460.

[4] Guedalia I.D., London M., Werman M. [1999], *An on-line agglomerative clustering method for non-stationary data*, Neural Computation, 11, 2, 521-540.

[5] Guha S., Mishra N., Motwani R., O'Callaghan L. [2000], *Clustering data streams*. In IEEE Symposium on Foundations of Computer Science (FOCS), 359-366.

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

[7] Migdał Najman K., Najman K., Data Analysis [2008], *Machine Learning and Applications*, Applying the Kohonen Self-organizing Map Networks to Selecting Variables, Studies in Classification, Data Analysis and Knowledge Organization, Presisach C., Burkhardt H., Schmidt-Thieme L., Decker R., Springer Verlag Berlin Heidelberg, 45-54.

[8] Najman K. [2009], *Zastosowanie nienadzorowanych sieci neuronowych typu Growing Neural Gas w analizie skupień*, Prace Naukowe UE we Wrocławiu, nr 47, 196-205.

[9] Najman K. [2010], *Ocena wpływu parametrów sterujących procesem samouczenia się sieci GNG na ich zdolność do separowania skupień*, Klasyfikacja i analiza danych – teoria i zastosowania Taksonomia 17, Prace Naukowe UE we Wrocławiu nr 17, 296-2010.

[10] Qin A. K. i Suganthan P. N. [2004], *Robust growing neural gas algorithm with application in cluster analysis*, Neural Networks, 17, 8-9, 1135-1148.

[11] Wang X., Smith K. Hyndman R. [1997], *Characteristic-based clustering for time series data*, Data mining and knowledge discovery, 13, 3, 335-364.

[12] Zamir O., Etzioni O. [1999], Grouper: A dynamic clustering interface to web search results, *Proceeding of WWW8*, Toronto, Canada.

GRUPOWANIE DYNAMICZNE Z WYKORZYSTANIEM SIECI GNG

Streszczenie

Od początku lat 90-tych XX wieku obserwuje się stały, dynamiczny wzrost liczby baz danych i zbieranych w nich informacji. Obserwuje się też stały wzrost zapotrzebowania na informacje, a z drugiej strony stały wzrost możliwości ich zbierania i przechowywania. Jedną z własności niektórych baz danych jest ich dynamiczna, zmieniająca się w czasie struktura grupowa. W artykule przedstawiono przegląd podstawowych koncepcji grupowania dynamicznego i zaproponowano jego nową definicję. Wskazano także praktyczną metodę realizacji grupowania dynamicznego opartą na samouczącej się sieci neuronowej typu GNG. Przedstawiono wyniki badań symulacyjnych nad własnościami takiej sieci w grupowaniu dynamicznym.

słowa kluczowe: analiza skupień, grupowanie dynamiczne, sieci GNG

THE DYNAMIC GROUPING USING GNG NEURAL NETWORK

SUMMARY

Since early 90s of 20th century has seen a steady and dynamic growth databases and collected information. There has also been a steady increase in demand for information, on the other hand growth in collection and storage information. One of the properties of some databases is their dynamic, changing during the group structure. The article presents an overview of the basic concepts of dynamic grouping and its proposed new definition. It was also a practical method to implement dynamic grouping based on self-learning neural network type of GNG. The results of simulation studies are presented in a dynamic grouping.

keywords: cluster analysis, dynamical clustering, GNG networks