ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2012
Seria: ORGANIZACJA I ZARZĄDZANIE z. 63a Nr kol. 1891

Katarzyna GRUBER-SZYDŁO
Akademia Medyczna we Wrocławiu
Klinika Chorób Zawodowych i Nadciśnienia
Jacek GRUBER, Michał SZYNKARUK
Politechnika Wrocławska
Instytut Informatyki

# PRZETWARZANIE WIELOWYMIAROWYCH DANYCH MEDYCZNYCH Z ZASTOSOWANIEM ROSNĄCEGO GAZU NEURONOWEGO

**Streszczenie.** W artykule omówiono możliwości usprawnienia przetwarzania danych medycznych i zwiększenia wydajności tego procesu przez wprowadzenie początkowego etapu eksploracji danych z wykorzystaniem rosnącego gazu neuronowego GNG (ang. *Growing Neural Gas*). Przetwarzanie danych medycznych charakteryzuje się dużą złożonością i występującymi w nim trudnościami, ponieważ analizowane dane są wielowymiarowe i dotyczą dopiero poznawanych zależności i zjawisk. W pracy wskazano, dlaczego rosnący gaz neuronowy pozwala osiągać lepsze rezultaty niż inne popularne sztuczne sieci neuronowe uczone metodą nienadzorowaną.

# THE PROCESSING OF MULTIVARIATE ASSOCIATIONS OF MEDICAL DATA USING THE GROWING NEURAL GAS

**Summary.** Increasing economies of medical data analysis process by entering the initial phase of exploration data using the growing neural gas (GNG), has been discussed in this article. The processing of medical data is characterized by high complexity and difficulties occurring in the analyzed data as multidimensional and concern only cognized relationships and phenomena. It was pointed out why the growing neural gas achieves better results than other popular artificial neural network learned by unattended.

## 1. Wprowadzenie

Proces opracowywania danych medycznych metodami statystycznymi powinien być poprzedzony etapem eksploracji danych. Zastosowanie eksploracji służy wydobyciu podstawowej wiedzy o zależnościach i związkach w danych. Wprowadzenie etapu

eksploracji, wykonywanego metodą nienadzorowaną, pozwala na zwiększenie niepełnej wiedzy eksperta na temat zjawisk niedostatecznie zbadanych lub nieodkrytych, ułatwia, znacznie przyspiesza i zmniejsza koszty późniejszych etapów procesu przetwarzania przez badacza danych metodami statystycznymi. Pozyskiwanie wiedzy z danych realizowane jest często metodami grupowania, umożliwiającymi odkrywanie związków w danych oraz filtrację danych niepowiązanych. Filtracja pozwala wyeliminować z dalszych etapów procesu badawczego zmienne niepowiązane, jako niemające znaczenia, w zasadzie zaburzające poznawanie zależności w danych.

Do odkrywania struktury w danych medycznych dobrze nadają się sztuczne sieci neuronowe (SSN). Dzięki SSN możliwe jest wskazanie potencjalnych związków między danymi, zarówno za pomocą analizy wyników ilościowych klasyfikacji, jak i ich reprezentacji graficznej. Zobrazowanie danych na wykresach ułatwia zaplanowanie dalszych badań, np. metodami statystyki matematycznej. Ponieważ celem poznawania struktury danych medycznych jest zwykle odkrycie zjawisk słabo znanych lub w ogóle nieznanych, konieczne jest zastosowanie nienadzorowanej SSN. Jako że liczba neuronów w sieci jest mniejsza od liczby wszystkich przypadków podawanych na wejście sieci, możliwe jest grupowanie danych. Grupowanie umożliwia badaczowi poznawanie struktury w wielowymiarowych danych medycznych, a ekspertowi pozyskiwanie wiedzy o aktualnie odkrywanych zależnościach.

Dobrym wyborem przy odkrywaniu struktury w danych jest zastosowanie rosnącego gazu neuronowego (ang. *Growing Neural Gas – GNG*), ze względu na jego inkrementacyjny charakter, eliminujący potrzebę definiowania początkowego rozmiaru sieci, oraz satysfakcjonujące dostosowywanie się do danych podawanych na wejściu sieci neuronowej [2].

Oprócz odkrywania struktury zależności w danych, rosnący gaz neuronowy GNG pozwala również eliminować z dalszych badań dane pozostające bez wpływu na pozostałe parametry, dzięki rywalizacji neuronów w trakcie szybkiego procesu adaptacji do danych podawanych na wejściu sieci neuronowej. Dane nieistotne można odrzucić na podstawie analizy wag poszczególnych neuronów. Dla danej cechy, jeżeli w wielu neuronach odpowiadająca jej waga nie mieści się w zakresie od wartości minimalnej do maksymalnej parametru, można powiedzieć, że ta cecha ma znikomy wpływ na to, jak dane zostaną pogrupowane.

Zaimplementowana na bazie gazu neuronowego metoda GNG posłużyła autorom do oceny ex post sprawności oraz możliwości obniżenia poniesionych nakładów pracy i czasu na wykonanie opracowania danych medycznych metodami statystycznymi. Dane dotyczą poszukiwania zależności wielowymiarowych między: stężeniami białek szoku termicznego

w komórkach krwi, stężeniami metali ciężkich i metaloidów w płynach ustrojowych, stężeniem wolnych protoporfiryn erytrocytarnych oraz wskaźnikiem masy ciała. Dane zebrano dzięki specjalistycznym badaniom oraz odpowiednio skonstruowanej ankiecie oraz zapisano je w standardowej postaci cyfrowej. Dane zebrano od ponad stu mężczyzn z populacji o długotrwałej ekspozycji zawodowej na metale ciężkie oraz z podobnie licznej grupy mężczyzn z populacji bez narażenia zawodowego.

Zbudowana aplikacja sieci GNG, wyposażona w interfejs użytkownika z unikalną funkcjonalnością prezentacji dwóch kolejnych stanów końcowych eksploracyjnego procesu przetwarzania danych (stanu bieżącego i stanu poprzedniego), implementująca GNG, może być wykorzystywana przez specjalistów (ekspertów i badaczy z różnych dziedzin) do eksploracji i analizy wielowymiarowych, niezależnych prób danych, nie tylko medycznych.

## 2. Problemy statystycznej obróbki danych medycznych

Biorąc pod uwagę ontologię medycyny, semantykę poszukiwanych zależności w danych medycznych oraz często niepełną wiedzę eksperta w poznawanych dopiero obszarach, poszukiwania związków oraz innych metod eksploracji danych i data mining zwykle ogranicza się do przestrzeni jedynie kliku lub kilkunastu zmiennych lub atrybutów wybieranych arbitralnie przez badacza. W tak wybranych przestrzeniach próbuje się następnie klasyfikować wektory danych i w ten sposób odkrywać w tych danych związki zmiennych oraz filtrować zmienne i odpowiadające im dane, eliminując niektóre z dalszych badań. W takim postępowaniu przy wyborze przestrzeni, istnieje ryzyko przeoczenia istotnych zmiennych i danych branych do badań.

Pojawia się potrzeba wykonywania badań wielu wariantów przestrzeni z różnymi zmiennymi i o zróżnicowanej liczbie wymiarów. Badacze wykonują je zwykle od razu mocnym aparatem statystycznym i nierzadko równolegle kilkoma statystycznymi metodami. Taka metodyka „eksploracji statystycznej” jest nie tylko niezwykle kosztowna, ale wprowadza do badań chaos. Przede wszystkim jednak ma cztery inne znaczące wady.

Pierwsza z tych wad, zasadnicza, to konieczność prowadzenia badań na grupach atrybutów lub zmiennych i sprawdzania, czy statystyki w grupach różnią się od siebie i czy istnieją między statystykami i rozkładami prób zależności, np. korelacyjne lub regresyjne. Grupy te muszą być wcześniej wykreowane. Istnieje tu ryzyko zagubienia ważnych, a często w zasadzie poszukiwanych klas. Takie klasy można by wykryć lub spostrzec jakąś dobrej

jakości, w sensie pewnej globalnej miary, dedykowaną, nienadzorowaną metodą uczenia rozpoznawania związków w danych, najlepiej z możliwością douczania na nowych danych.

Druga wada – „eksploracji statystycznej" – wiąże się z istniejącym w metodyce badań statystycznych paradygmatem podziału analizowanych danych na badaną grupę podstawową i grupę odniesienia. Taka narzucona klasyfikacja sprawia, że nie zdołamy odrzucić takich atrybutów i zmiennych oraz odfiltrować danych z nimi powiązanych, pomiędzy którymi związki i zależności nie zachodzą i z tego powodu są bez znaczenia lub zaciemniają obraz badań i wyniki. Ważna jest też możliwość wykrycia związków oczywistych, dobrze znanych ekspertowi, ponieważ pozwala to na sprawdzenie, czy zastosowana metoda lub model statystyczny powoduje ich odkrycie. Taka weryfikacja pozwala stwierdzić, która metoda lub model statystyczny, z możliwych do zastosowania w późniejszych etapach przetwarzania danych jest semantycznie właściwy.

Trzecią niedogodnością w procesie badawczym analizy danych bez wykonania etapu eksploracji jest konieczność wskazania możliwości grupowania i kreowania klasyfikacji szczegółowych wewnątrz grupy podstawowej i wewnątrz grupy odniesienia. Tak kreowane klasyfikacje umożliwiają coraz bardziej szczegółowe badania statystyczne związków pomiędzy zmiennymi. Ponieważ wskazywanie grup odbywa się w warunkach niepełnej wiedzy eksperta i znacznej niewiedzy badacza, więc badanie związków statystycznych między zmiennymi trzeba wykonać dla bardzo wielu możliwych klasyfikacji i grup. Taki proces jest długotrwały, kosztowny, trudny do automatyzacji i kontroli.

Czwarta wada bezpośredniego, „statystycznego procesu eksploracji danych", polega na tym, że w celu wykonania obliczeń statystycznych badacz musi podejmować dość niejasne semantycznie decyzje co do tego, które zmienne są niezależne, a które są zależne w badaniach istotności różnic statystyk zmiennych w poszczególnych grupach. Podobnie zawsze pozostaje niepewność odnośnie trafności grupowania i klasyfikacji w statystycznych obliczeniach istotności korelacji zmiennych i atrybutów w grupach oraz przy budowaniu wielo-wymiarowych modeli regresji dla grup.

Podsumowując cztery powyższe wady, należy stwierdzić, że utrudnienia w doborze grup oraz w przebiegu i trafności procesu klasyfikacji wywierają negatywny wpływ na poprawność semantyczną oraz zapewnienie pełnej kontroli i klarownej metodyki procesu analizy danych oraz uniemożliwiają automatyzację tego procesu. Dotyczy to zarówno wyszukiwania związków między zmiennymi i w danych poprzez analizę istotności różnic statystyk danych, jak też istnienia korelacji, kowariancji, tworzonych wielowymiarowych modeli regresji zmiennych oraz zagadnień identyfikacji zmiennych niepowiązanych i filtracji danych.

Skalę trudności w systematycznym podejściu do przetwarzania danych medycznych zwiększa fakt, że badane próby atrybutów i zmiennych nie mają prawie nigdy rozkładów normalnych ani nie pochodzą z populacji o takich rozkładach, co dodatkowo przyczynia się do braku ich czytelności. Eksploracja takich danych i wykonywanie zadań data mining w przestrzeniach takich danych, bez odpowiedniego automatycznego aparatu uczenia nienadzorowanego, jest praktycznie niemożliwe, w większości przypadków również dla eksperta. Nawet w podejściu przetwarzania danych medycznych dwuwymiarowych metodami statystycznymi, w zasadzie zawsze wymagane są systematyczne statystyczne badania nieparametryczne wariancji, korelacji i kowariancji prób. Metoda odkrycia związków za pomocą przeglądania danych i w ich reprezentacjach na wykresach jest całkowicie nieskuteczna. Duży wpływ na wybór odpowiedniej metody i modelu eksploracji danych ma liczność przetwarzanych danych. Dane pozyskiwane w większości ontologii medycznych zwykle nie są zbyt liczne i zawierają od kilkudziesięciu danych w próbie, wymaganych jako minimum w standardowych procedurach statystycznych, do stu lub rzadziej kilkuset elementów w próbie.

Reasumując, potrzebny jest aparat do pozyskania wiedzy z danych medycznych jeszcze przed zapoczątkowaniem szczegółowych badań danych metodami statystycznymi.

Sztuczna sieć neuronowa, o modelu rosnącego gazu neuronowego, wydaje się być dobrą metodą do eksploracji danych medycznych. Udało się potwierdzić znaczące klasyfikacje oraz odkryć ex post tendencje zależności stężenia obronnych białek szoku termicznego od czynników i warunków narażenia organizmu na negatywne działanie metali i metaloidów, zależności będących przedmiotem szczegółowych badań w ramach rozprawy doktorskiej [3].

## 3. Sztuczne sieci neuronowe

Sztuczne sieci neuronowe realizowane są jako modele matematyczne bądź symulacyjne. Niekiedy spotyka się ich realizacje sprzętowe [9]. Sieci te składają się z neuronów, które powstały na wzór ich rzeczywistych odpowiedników. Obecny stan wiedzy nie pozwala na pełne odzwierciedlenie pracy biologicznych neuronów. Niemniej, według Ryszarda Tadeusiewicza, dysponując jedynie uproszczonym modelem, można dokonać próby znalezienia interpretacji, która przybliżyłaby nas do istoty rzeczy [8]. O podstawach biologicznych traktuje wiele publikacji, omawiających tematykę SSN [1, 6, 8, 9]. Liczne zastosowania i duża liczba publikacji na temat sieci neuronowych, że są cennym narzędziem,

pomagającym rozwiązywać problemy z różnych dziedzin. Stanisław Osowski tłumaczy, z czego wynika ich szerokie uznanie [6]. Według tego autora, SSN „stanowią uniwersalny układ aproksymacyjny odwzorowujący wielowymiarowe zbiory danych". W trakcie procesu uczenia sieć wytwarza strukturę, dzięki której możliwe jest rozwiązywanie różnego typu zadań [9]: predykcji, klasyfikacji, kojarzenia danych, analizy danych, filtracji sygnałów i optymalizacji. Spotyka się też inne klasyfikacje wykonywanych zadań, np.: poszukiwania związku, odkrywania pojęć, kompresji, sterowania.

W niniejszej pracy wykorzystano rosnący gaz neuronowy do grupowania i filtracji, na przykładzie danych medycznych. Grupowanie danych metodą nienadzorowanej sztucznej sieci neuronowej nabiera szczególnego znaczenia w sytuacji, gdy wiedza eksperta na temat zjawisk niedostatecznie zbadanych nie pozwala na samodzielne grupowanie badanych przypadków. Dzięki algorytmowi grupowania możliwe jest wskazanie potencjalnych związków między danymi. Zobrazowanie danych na wykresach ułatwia zaplanowanie dalszych badań, np. metodami statystyki matematycznej. Rosnący gaz neuronowy GNG pozwala także na wyeliminowanie z dalszych badań danych pozostających bez wpływu na pozostałe parametry. Nieistotne dane można odrzucić na podstawie analizy wag poszczególnych cech danych przypadków. Dane odrzucamy, jeśli wagi danego parametru nie mieszczą się w zakresie od wartości minimalnej do maksymalnej dla tego parametru.

W literaturze światowej autorzy nie znaleźli prac na temat zastosowania rosnącego gazu neuronowego przy badaniach nad zależnościami między stężeniem białek szoku termicznego (ang. *Heat shock protein – Hsp*) a stężeniem metali ciężkich (ołowiu – Pb, kadmu – Cd) i metaloidów (arsenu – As) w płynach ustrojowych u osób zawodowo na nie narażonych. Do wykonania analizy wymienionych zależności zaimplementowano algorytm GNG w języku programowania Java. Implementację wykonano na platformie Standard Edition 6, na podstawie pracy Bernda Fritzke [2]. Wykorzystano dane medyczne dołączone do pracy doktorskiej poświęconej wymienionemu zagadnieniu [3]. Badane dane podlegają konwersji do standardowego formatu CSV, stosowanego w najważniejszych pakietach statystycznych oraz narzędziach do analizy data mining.

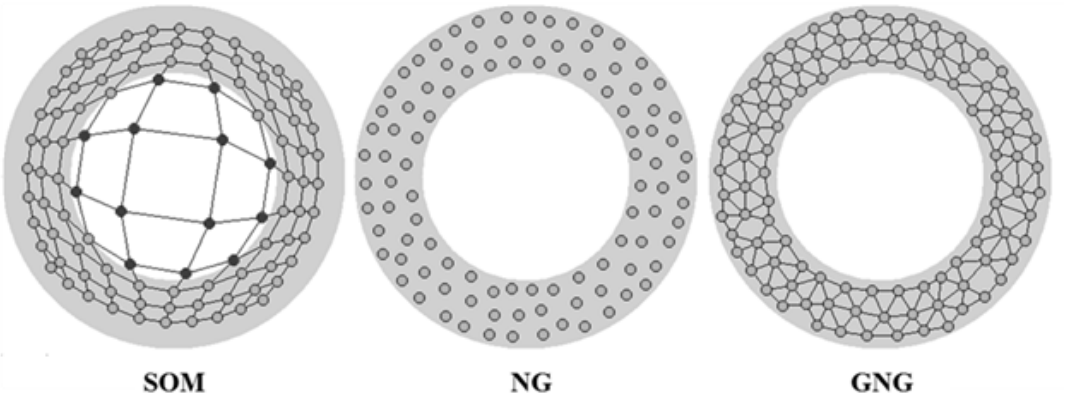
## 4. Wybór sztucznej sieci neuronowej dla zadanego problemu

Gaz neuronowy (ang. *Neural Gas – NG*) jest sztuczną siecią neuronową, zainspirowaną samoorganizującą się mapą wprowadzoną w 1991 roku przez Thomasa Martinetza i Klausa Schultena [5]. NG jest prostym algorytmem znajdowania optymalnych reprezentacji danych

na podstawie wektorów cech. Algorytm został nazwany gazem neuronowym ze względu na dynamikę tych wektorów cech w procesie adaptacji, które rozprowadzają się jak gaz w przestrzeni danych. Gaz neuronowy można stosować do problemów kompresji lub do problemów kwantyzacji wektorowej, np. do rozpoznawania mowy, przetwarzania obrazów lub rozpoznawania wzorców. Ponieważ jest to algorytm silnie zbieżny, więc jest alternatywą stosowania algorytmu k-średnich do analizy skupień. NG jest kolekcją swobodnych neuronów. Poruszają się one w przestrzeni, w której zostały zdefiniowane, ze swobodą ograniczoną współzawodnictwem. Każdy z neuronów ma przypisane pewne wektory wag $w_i$. Składowe wektora $w_i$ są równocześnie interpretowane jako współrzędne punktu $w_i$ w przestrzeni wielowymiarowej liczb rzeczywistych $R_N$. Metoda NG polega na budowaniu sieci ze swobodnych neuronów z kolekcji. Gaz neuronowy jest najskuteczniejszą metodą organizacji neuronów w sieć neuronową.

Podstawą rosnącego gazu neuronowego GNG jest opracowanie przez Thomasa Martinetza i Klausa Schultena idei gazu neuronowego NG [4-5]. Zaletą GNG w stosunku do gazu neuronowego NG jest inkrementacyjny charakter modelu, który eliminuje potrzebę definiowania początkowego rozmiaru budowanej sieci.

W trakcie procesu uczenia, w każdej chwili można dokonać zamrożenia wag, jeżeli uznamy, że sieć wystarczająco dobrze dostosowała się do przypadków podawanych na wejściu, a jej rozmiar, czyli liczba neuronów, jest wystarczający. Jeżeli po wykonaniu wstępnych badań, gdy rosnący gaz neuronowy przeszedł etap uczenia oraz mapowania danych, otrzymamy nowe zestawy danych, to sieć można douczyć, biorąc pod uwagę nowe przypadki. Sposób wstawiania nowych neuronów i modyfikacji istniejących powoduje, że proces ten będzie przebiegał lepiej i szybciej niż w przypadku rozpoczęcia procesu uczenia sieci od nowa. Ponieważ wykonanie badań medycznych omówionych w pracy [3] było kosztowne, nie było możliwości otrzymania nowego zestawu danych i zebrany zestaw danych był próbą niezależną. Biorąc też pod uwagę niezbyt dużą ilość posiadanych danych medycznych, nie podejmowano prób douczania rosnącego gazu neuronowego.




Rys. 1. Porównanie działania sieci SOM, NG i GNG
Fig. 1. Comparison of the effects of SOM, NG and GNG network
Źródło: [12]

Ponadto, jako zaletę rosnącego gazu neuronowego wymienia się fakt, że definiowane dla niego parametry są stałe, w przeciwieństwie do tych w gazie neuronowym czy samoorganizującej się mapie (ang. Self Organizing Map – SOM) [2], co znacząco ułatwia dostrajanie sieci. Warto też zauważyć, że neurony wchodzące w skład NG i GNG lepiej pokrywają przestrzeń danych niż neurony należące do SOM. Fakt ten można dostrzec m.in. uruchamiając aplikację [12-14], sporządzoną w Instytucie Neuroinformatik na Uniwersytecie Ruhr w Bochum. Na wejściu aplikacji podawano dwuwymiarowe dane będące współrzędnymi x oraz y. Wagi neuronów przyjmowały przed etapem uczenia wartości z przestrzeni zaznaczonej zarówno kolorem białym, jak i kolorem jasnoszarym. Rozmiar maksymalny każdej z tych sieci, licząc w neuronach, ograniczono do 100. Przyjęto wartości domyślne pozostałych parametrów w modelach badawczych. Algorytm zatrzymano po 100 tysiącach iteracji. Na rys. 1 widać dostosowanie się neuronów do danych wejściowych w przetestowanych trzech modelach sieci.

Podając na wejście sieci jedynie dwuwymiarowe dane i rzutując je na wykres, można w łatwy sposób dostrzec, czy sieć dobrze się do nich dostosowała. Każdy może być ekspertem i stwierdzić, czy neurony „ułożyły się" równomiernie w przestrzeni wejściowej. Na rys. 1 można zauważyć, że zarówno NG, jak i GNG dobrze poradziły sobie z tym zadaniem, w przeciwieństwie do SOM, w którym kilka neuronów zaznaczonych kolorem czarnym źle dostosowało się do danych podawanych na wejście. Świadczy o tym ich „położenie" poza przestrzenią zaznaczoną kolorem jasnoszarym. Ponadto, neurony sieci SOM, „znajdujące się" w szarym polu, nie są rozłożone równomiernie.

W przypadku wielowymiarowych danych nie wystarczy analiza jedynie wykresów, a czasami nawet analiza wag poszczególnych neuronów oraz przypadków przypisanych do

danego neuronu po etapie uczenia i mapowania danych. Na sformułowanie tego wniosku pozwala doświadczenie autorów w badaniach nad danymi medycznymi za pomocą własnej implantacji GNG.

## 5. Opis rosnącego gazu neuronowego

Rosnący gaz neuronowy GNG składa się z zestawu A węzłów. Każda jednostka c należąca do A, ma w sobie wektor wag w. Ponadto, sieć zawiera N połączeń pomiędzy jednostkami. Połączenia te nie mają ustalonej wagi, służą tylko temu, aby zachować topologiczne sąsiedztwo. Główną ideą opisywanej metody jest sukcesywne dodawanie nowych jednostek do wstępnej, małej sieci. Określenie miejsca wstawienia nowej jednostki bazuje na analizie pewnej lokalnej miary błędu. Każdy nowy neuron jest dodawany w pobliżu tych elementów, które charakteryzują się największym błędem lokalnym. Model rosnącego gazu neuronowego zaczyna działanie z dwoma neuronami i „wstawia" nowe.

Algorytm sztucznej sieci neuronowej GNG opisuje się następująco [4-5]:

1. Algorytm rozpoczyna działanie ze zbiorem elementów zawierającym dwa neurony, $i$=1,2 – liczba początkowa neuronów. Obie jednostki mają losowe wartości wag $w_i$, $i$=1,2. Dobór wag $w_i$, jest w zakresie zmienności analizowanego zbioru danych. Następuje inicjacja początkowego pustego zbioru połączeń między neuronami $C$ (brak jakichkolwiek połączeń między neuronami).
2. Podawany jest sygnał wejściowy $x$.
   Znajdowane są dwa neurony, których odległość od wektora $x$ jest najbliższa $s_1$ (neuron zwycięzca) oraz $s_2$ , przy czym $d_{s1} \leq d_{s2}$. Neuron $s_1$ nazywany jest zwycięzcą, gdy $\|x - s_1\| < \|x - s_2\|$. Ogólnie, wyłanianie neuronu zwycięzcy polega na ustawieniu go w relacji słabego porządku rosnącego (ranking), na pozycji $d_0$ , następująco:
   $d_0 < d_1 < ... < d_M$ , $i = 1,2,...,M$. Przy czym $d_m = \|x - w_{m(i)}\|$ oznacza odległość $i$-tego neuronu zajmującego w wyniku sortowania $m$-tą pozycję w szeregu za neuronem zwycięzcą, który został ustawiony na pozycji $d_0$.
3. Jeżeli połączenie pomiędzy neuronami $s_1$ oraz $s_2$ nie istnieje w zbiorze $C$, należy je dodać: $C = C \cup \{(s_1, s_2)\}$.

4. Ustaw wiek połączeń między $s_1$ i $s_2$ na wartość 0 („odświeżanie" połączenia): $wiek(s_1, s_2) = 0$.
5. Dodawany jest kwadrat odległości pomiędzy sygnałem wejściowym i zwycięzcą do lokalnego błędu: $\Delta E_{s1} = \|x - w_{s1}\|^2$.
6. Modyfikuj wektory wag zwycięzcy i jego bezpośrednich sąsiadów: $\Delta w_{s1} = e_b(x - w_{s1}), \Delta w_i = e_n(x - w_i)$, $i = 1,2,...$, oraz $e_b, e_n$ – odległości między wektorami wyliczone na podstawie miary euklidesowej.
7. Zwiększ o 1 wiek połączeń wychodzących z neuronu $s_1$: $wiek(s_1, i) = wiek(s_1, i) + 1$, dla wszystkich *i*, $i = 1,2,...$.
8. Usuń te połączenia, których wiek przekroczył zadaną wartość $wiek_{\max}$. Usuń neurony, które nie mają połączeń.
9. Jeżeli liczba wygenerowanych sygnałów wejściowych jest wielokrotnością parametru $\lambda$ ($\lambda$ >0, promień sąsiedztwa parametryzujący funkcję sąsiedztwa neuronów, modyfikującą wektory wag neuronów), dodaj nowy neuron w następujący sposób:
   - znajdź neuron *q* z maksymalnym błędem $E_q, q \in \{1,2,...\}$,
   - spośród sąsiadów neuronu *q* znajdź neuron *f* z maksymalnym błędem $E_f$, $f \in \{1,2,...\}$,
   - wstawienie nowego neuronu *r* w połowie drogi między neuronami i wyznacz wartości jego wag jako $w_r = (w_q + w_f)/2, r \in \{1,2,...\}$,
   - utwórz połączenia pomiędzy neuronem *r* i neuronami *q* oraz *f* i usuń bezpośrednie połączenie pomiędzy neuronem *q* i *f*.
   - zmniejsz wartość błędu neuronów *q* i *f*: $\Delta E_q = -\alpha E_q, \Delta E_f = -\alpha E_f$, $\alpha > 0$, współczynnik parametryzujący wartość błędu lokalnego (promień sąsiedztwa) neuronu),
   - wyznacz wartość błędu nowego neuronu *r*: $E_r = (E_q + E_f)/2$.
10. Zmodyfikuj wartość błędu wszystkich neuronów: $\Delta E_c = -\beta E_c, c = 1,2,...$ $\Delta E_q = -\alpha E_q, \Delta E_f = -\alpha E_f$, $\alpha > 0$ – współczynnik parametryzujący wartość błędu lokalnego (promień sąsiedztwa) wszystkich neuronów.
11. Jeżeli kryterium stopu (np. wielkość sieci lub inna miara jakości) nie jest spełnione – powrót do punktu 2.

Można zastosować jedno bądź kilka kryteriów warunku stopu algorytmu, m.in.: rozmiar sieci, globalny błąd sieci. W omawianej w niniejszym artykule implementacji GNG, jako

kryterium warunku stopu przyjęto rozmiar sieci. Jeżeli rozmiar rosnącego gazu neuronowego osiągnął wcześniej określoną wartość maksymalną, to następowało zakończenie procesu uczenia. Odległości między wektorami wyliczono na podstawie miary euklidesowej. Przy każdej iteracji wybierany jest losowo sygnał wejściowy *x*. Wykorzystano w tym celu generator liczb pseudolosowych, dostępny w środowisku programistycznym Java Development Kit. Po etapie uczenia wszystkie wagi neuronów są zamrażane i następuje etap mapowania danych, tzn. wybierany jest zbiór danych, z którego każdy przypadek jest przypisywany do danego neuronu. O tym, gdzie zostanie on przypisany, decyduje odległość między wagami. Na potrzeby wykonanych badań wykorzystano ten sam zbiór danych zarówno na etapie uczenia, jak i na etapie mapowania. W opisywanej metodzie wykorzystano podejście iteracyjne, nie dokonano natomiast podziału na epoki. Uznano, że losowanie wektora wejściowego w każdej iteracji z całego zbioru uczącego będzie bardziej naturalne. Sieci SOM, NG i GNG można porównać, biorąc pod uwagę m.in. czas uczenia lub błąd globalny. W omawianej implementacji GNG błąd jest wyznaczany jako suma najmniejszych odległości pomiędzy danym neuronem a najbliższym mu przypadkiem pochodzącym z całego zbioru uczącego. Dużym usprawnieniem GNG w stosunku do podstawowego algorytmu gazu neuronowego NG jest to, że gdy sieć jest już nauczona, nic nie stoi na przeszkodzie, by ją douczyć. Odbywa się to poprzez dodanie nowych neuronów i połączeń, zgodnie z wcześniej opisanym algorytmem. Stosując algorytm GNG, dokonujemy grupowania elementów, pozyskując w ten sposób odpowiednią wiedzę z posiadanych danych. Grupowanie zachodzi dzięki konkurencji między neuronami w trakcie procesu adaptacji do danych podawanych na wejście sieci. Bez wsparcia z użyciem jakiegoś mechanizmu klasyfikacji ciężko dostrzec jakiekolwiek zależności w danych wielowymiarowych.

Przeprowadzono badania 258 zdrowych mężczyzn, bez możliwości zidentyfikowania poszczególnych osób: grupa zasadnicza obejmuje 138 pracowników hut miedzi, zawodowo narażonych na działanie ołowiu i arsenu, z których 108 było dodatkowo środowiskowo narażonych na kadm. Złożone i czasochłonne badanie medyczne [3] pozwoliło ocenić stężenie wybranych białek szoku termicznego w komórkach ludzi poddanych przewlekłej, złożonej ekspozycji na metale ciężkie i metaloidy w małym stężeniu. Miało ono nowatorski charakter, gdyż dotychczas oceniano jedynie krótkotrwały wpływ pojedynczych metali i metaloidów w dużym stężeniu na stężenie białek szoku termicznego w komórkach zwierząt i w hodowlach komórkowych [7]. Te wartościowe wyniki zostały potwierdzone rezultatami w postaci obserwacji graficznych i wartości wyliczonych w analizie data mining wartości

wymienionych wskaźników za pomocą zaimplementowanej sieci GNG, przy czym wyniki otrzymano w stosunkowo krótkim czasie.

Po etapie uczenia zarówno wagi wszystkich neuronów, jak i wartości połączeń między nimi pozostają niezmienione. Umożliwia to przypisywanie danych przypadków medycznych do poszczególnych neuronów. Liczba neuronów jest mniejsza od liczby przypadków, więc dane są pogrupowane. Takie grupowanie może sugerować kierunek dalszych badań nad powiązaniami i zależnościami między danymi za pomocą silniejszych matematycznych metod poznawczych i analitycznych. Grupowanie przypadków ma na celu spostrzeżenie istnienia potencjalnych związków pomiędzy poszczególnymi parametrami.

Rosnący gaz neuronowy pozwala badaczowi odkryć, który z atrybutów prawdopodobnie nie ma wpływu na pozostałe. Autorom udało się również w relatywnie krótkim czasie dokonać odkrycia cennych i zaskakujących, zarówno dla ekspertów, jak i dla badaczy zależności, a także braku zależności w badanych danych medycznych. Przykładowo, podczas badania zależności pomiędzy współczynnikiem masy ciała BMI (ang. *Body Mass Index*) a stężeniem ołowiu, kadmu, arsenu, Hsp27, Hsp60, Hsp70 oraz stężeniem białych (WBC) i czerwonych krwinek (RBC) wykryto, że BMI nie ma wpływu na grupowanie przypadków. Odkryte spostrzeżenia uwiarygodnione są faktem, że w literaturze światowej autorzy nie znaleźli prac, które potwierdzałyby wpływ współczynnika BMI na którykolwiek z wziętych do tego badania parametrów. Okazało się, że w formułowaniu podobnych spostrzeżeń i odkryć zasadnicze znaczenie ma analiza wag poszczególnych neuronów w tworzonych modelach GNG, a nie tylko studiowanie wykresów i graficznych ilustracji zależności danych otrzymanych z działania gazu GNG.

## 6. Zakończenie

Aplikacja zawierająca implementację GNG oraz przejrzysty i wartościowy funkcjonalnie interfejs użytkownika może ułatwić lekarzom oraz innym specjalistom z dziedziny medycyny odkrywanie potencjalnych zależności w danych wielowymiarowych. Rosnący gaz neuronowy może być stosowany do analizy różnych danych, nie tylko medycznych.

W procesie badawczym za pomocą GNG dużą rolę odgrywa analiza wykresów. Mając pogrupowane wielowymiarowe dane i rzutując je na wykres, można łatwiej dostrzec istniejące zależności zmiennych. Dzięki działaniu GNG można także odnaleźć i wyeliminować z dalszych badań te atrybuty, które nie mają silnego wpływu na grupowanie

danych. Wyniki uzyskane za pomocą narzędzia z implementacją sieci GNG mogą stanowić podstawę do dalszych badań przy użyciu metod i technik z obszaru klasycznej statystyki matematycznej oraz mocniejszych metod i technik data mining.

Trzeba zaznaczyć, że podobne wyniki analizy data mining danych można otrzymać, stosując inne typy sieci neuronowych, uczone metodą nienadzorowaną, jak choćby samoorganizująca się mapa czy gaz neuronowy w oryginalnym wydaniu, bez wprowadzania (zawsze przecież możliwych do wykonania) modyfikacji. Jednakże różnica między takimi sieciami a siecią GNG polega na tym, że wspomniane sieci nie mogą być douczane, w podobny sposób, jaki jest realizowany w rosnącym gazie neuronowym. Ewentualna modyfikacja sieci SOM i NG wiązałaby się ze sporym nakładem pracy, a otrzymane wyniki mogłyby być niepewne bądź gorsze od tych otrzymanych za pomocą GNG. Douczanie o którym mowa w przypadku rosnącego gazu neuronowego, jest ważne szczególnie wtedy, gdy dysponujemy niewielką ilością danych i jedynie na ich podstawie chcemy dostrzec pewne zależności bądź odrzucając te atrybuty, które mają niski wpływ na pozostałe. Proces uczenia w każdej chwili można zatrzymać, jeżeli uznamy rozmiar sieci lub osiągnięty błąd globalny sieci za satysfakcjonujący. Może się zdarzyć, że po pewnym czasie otrzymamy nową porcję danych. Można wtedy wykorzystać możliwość kontynuacji procesu uczenia utworzonej wcześniej sieci na podstawie danych.

Powodem podjęcia implementacji rosnącego gazu neuronowego było to, że w najważniejszych stosowanych komercyjnie systemach do analizy statystycznej danych i analizy data mining, takich jak Statistica firmy Statsoft Inc. [11] oraz SAS Enterprise Miner firmy SAS Institute Inc. [10], brak jest implementacji GNG. Jednakże, jak to uzasadniono w niniejszym artykule, sieć neuronowa GNG jest bardzo użyteczna i pożądana na wstępnym etapie opracowywania danych medycznych, gdy w zasadzie brak jest możliwości nadzorowania opracowywania danych przez eksperta lub brak jest wiedzy eksperckiej. Ważnym atutem interfejsu wykonanej implementacji GNG jest zdolność pamiętania stanów końcowych poprzedniego i bieżącego wykonanych obliczeń, co umożliwia porównywanie i ocenę postępu w obliczeniowym procesie badawczym.

**Bibliografia**

1. Duch W., Korbicz J., Rutkowski L., Tadeusiewicz R.: Sieci neuronowe, tom 6. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2000.

2. Fritzke B.: A Growing Neural Gas Network Learns Topologies. MIT Press, Cambridge MA 1995.
3. Gruber-Szydło K.: Wybrane białka szoku termicznego i skierowane przeciwko nim przeciwciała u narażonych na metale ciężkie. Akademia Medyczna we Wrocławiu, Wydział Lekarski, Klinika Chorób Wewnętrznych, Zawodowych i Nadciśnienia Tętniczego, rozprawa doktorska, Wrocław 2009.
4. Krętowska M.: Sztuczne sieci neuronowe. Materiały do wykładu – wykład 11. Politechnika Białostocka, Wydział Informatyki, Instytut Informatyki, http://aragorn.pb.bialystok.pl/~gkret/SSN/SSN_w11.PDF, 17/04/2011.
5. Martinetz T., Schulten K.: A neural gas network learns topologies. Artifcial Neural Networks. Elsevier, 1991, p. 397-402.
6. Osowski S.: Sieci neuronowe do przetwarzania informacji. Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 2006.
7. Rossi M.R., Somji S., Garrett S.H., Sens M.A., Joginder N., Sens D.A.: Expression of hsp27, hsp60, hsc70, and hsp70 stress response genes in cultured human urothelial cells (UROtsa) exposed to lethal and sublethal concentrations of sodium arsenite. Environ Health Perspect 110, (2002), pp. 1225-1232.
8. Tadeusiewicz R.: Neurocybernetyka teoretyczna. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa, 2009.
9. Tadeusiewicz R.: Sieci neuronowe. Akademicka Oficyna Wydawnicza RM, Warszawa 1993.
10. SASEnterprise Miner. http://www.sas.com/technologies/analytics/datamining/miner/neuralnet/, 17/04/2011.
11. Statistica Automated Neural Networks. http://www.statsoft.com/products/statistica-Automated-Neural-Networks/, 17/04/2011.
12. Neural Competitive Models Demo – GNG. http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/DemoGNG/GNG_2.html, 17/04/2011.
13. Neural Competitive Models Demo – NG. http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/DemoGNG/NG_2.html, 17/04/2011.
14. Neural Competitive Models Demo – SOM. http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/DemoGNG/SOM_2.html, 17/04/2011.

**Abstract**

It was pointed out why the growing neural gas (GNG) achieves better results than other popular unsupervised neural networks. GNG neural algorithm was supplemented by a global measure assessing the quality of the learning process of the network. Built on the neural gas method and the implementation of GNG has been used to study the relationship of medical data concerning the effects of heat shock proteins. The few examples of the processing shown how to interpret GNG obtained by grouping and filtering the results and how to explore the structure of multidimensional data. The application can also be used by specialists in other areas of science and applications to explore and analyze multidimensional data independent trials.