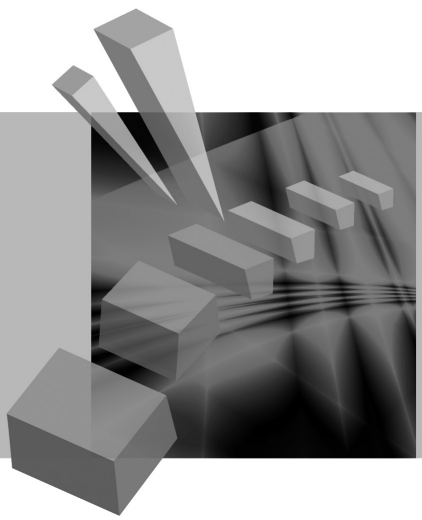
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
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

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

Kopiowanie i powielanie w jakiejkolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2013

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

Wersja pierwotna: publikacja drukowana
Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Krzysztof Najman
Uniwersytet Gdański

SAMOUCZĄCE SIĘ SIECI GNG W GRUPOWANIU DYNAMICZNYM ZBIORÓW O WYSOKIM WYMIARZE

Streszczenie: W artykule przedstawiono wyniki badań symulacyjnych dotyczących zastosowania samoorganizujących się sieci neuronowych typu GNG w grupowaniu dynamicznym danych o wysokim wymiarze. Przeprowadzono niezależnie dwa eksperymenty na wygenerowanych danych, dokonując pomiaru szybkości uczenia się sieci w zależności od wymiaru danych. Rezultaty badań wydają się potwierdzać przydatność sieci tego typu w grupowaniu dynamicznym.

Słowa kluczowe: sieć typu gaz neuronowy (GNG), grupowanie dynamiczne, analiza skupień.

1. Wstęp

Jedną z cech współczesnych zbiorów danych jest ich ogromna dynamika. Wyraża się ona szybkim przyrostem liczby zarejestrowanych jednostek, a także coraz większą liczbą rejestrowanych cech. W każdej minucie użytkownicy portalu YouTube zamieszczają 48 godzin filmów. Na Facebooku w tym czasie pojawia się 684 478 nowych materiałów. Wyszukiwarka Google rejestruje 2 000 000 pytań od internautów na sekundę [Khanna, Smith 2012]. Każdy z powyższych serwisów rejestruje dziesiątki, a czasami setki parametrów dotyczących użytkowników i samej transmisji danych.

Po roku 2000 szybki spadek cen nośników danych i wzrost technicznych możliwości ich przechowywania spowodował przekonanie, że nie ma już praktycznych ograniczeń w ilości zbieranych danych. Co więcej, w wielu zbiorach danych rejestruje się przypadki na podstawie pomiarów dokonanych z wykorzystaniem bardzo precyzyjnych urządzeń. Mogło się wydawać, że spełniło się marzenie statystyków o praktycznie nieograniczonej liczbie dokładnych i aktualnych obserwacji. Szybko się jednak okazało, że wykreowana przez istniejące możliwości presja spowodowała tak szybki rozrost liczby zbiorów danych, liczby obserwowanych cech, jak również liczby rejestrowanych przypadków, że pojawiły się nowe, nieznane wcześniej problemy. Wzrost ilości rejestrowanych przypadków okazuje się szybszy niż spa-

dek cen ich przechowywania. Przechowywanie danych staje się więc coraz bardziej kosztowne. Ogromna liczba przypadków, a także bardzo szybki napływ nowych, powodują problemy z ich analizą. Dotyczy to zarówno problemów skali (np. jak wyznaczyć macierz odległości między miliardami jednostek?), jak i czasu przetwarzania danych (ile można przeznaczyć czasu na klasyfikację danej jednostki, gdy pojawia się ich 50 000 na sekundę?).

Ilustracją powyższych problemów może być rejestracja zapytań kierowanych do wyszukiwarki internetowej. Powiedzmy, że użytkownik chce znaleźć w Internecie informacje o konferencji SKAD w 2012 r. i zada w wyszukiwarce pytanie: „konferencja SKAD2012". Sam tekst składa się z 40 bajtów. Jeżeli do tego wyszukiwarka zarejestruje adres IP komputera, z którego zadano pytanie, dokładną datę i czas, dodatkowe informacje dotyczące oprogramowania użytkownika, to okaże się, że to proste pytanie będzie zapisane w ponad 100 bajtach. Liczba ta wydaje się bardzo mała w porównaniu z powszechnie dostępnymi terabajtowymi pamięciami masowymi. Jest to jednak pozór. W tabeli 1 zaprezentowano liczbę zapytań do wyszukiwarki i wielkość uzyskanego zbioru danych w wybranych momentach.

Tabela 1. Liczba zapytań do wyszukiwarki i wielkość uzyskanego zbioru danych

Czas	Liczba zapytań	Objętość archiwizowanych danych w TB	Koszt nośnika danych
1 sekunda	2 000 000	0,0002	0,07 zł
1 minuta	120 000 000	0,0109	4,37 zł
1 godzina	7 200 000 000	0,6548	261,93 zł
1 doba	172 800 000 000	15,7161	6 286,43 zł
1 rok	63 072 000 000 000	5736,3650	2 294 545,99 zł

Źródło: opracowanie własne.

Jak można zauważyć, zarówno wielkość zbioru danych, jak i jego rozmiar już pierwszego dnia rejestracji danych są bardzo duże. Przyjmując przeciętną cenę nośnika danych o pojemności 1 TB na poziomie 400 zł, należałoby kupować ich ponad 15 dziennie, ponosząc z tego tytułu ponad 6286 zł kosztów. Po roku jest to już ponad 2 miliony zł. Koszt ten uwzględnia jedynie ceny nośników. W rzeczywistości byłby on dużo większy, gdyż rosłyby także lawinowo koszty infrastruktury informatycznej. Potrzebne jest odpowiednio duże pomieszczenie, kilometry kabli, setki komputerów, a także praca wielu informatyków i techników.

Drugim wyzwaniem w analizie skupień opisanych powyżej baz danych jest ogromna szybkość napływu nowych danych. Jeżeli rejestruje się tysiące jednostek na sekundę, to po kilku czy kilkunastu sekundach mogą się pojawić całkowicie nowe struktury. W tym samym czasie struktury istniejące mogą zaniknąć.

Aby sprostać powyższym wymaganiom, należy zastosować specjalny algorytm grupowania danych. Powinien on charakteryzować się przynajmniej czterema cechami. Musi być bardzo szybki. Jeżeli w bazie danych następuje wiele zmian w ciągu sekundy, w tym samym czasie musi być wykonane grupowanie. Powinien być oszczędny. 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. Musi być wysoce autonomiczny. Sama szybkość zmian 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. Po czwarte musi się charakteryzować dobrymi własnościami uzyskanej struktury grupowej. Warunek ten jest jednak na drugim planie. Ważniejsze jest, aby nawet popełniając błędy, nadążyć za napływem danych, niż żeby idealnie grupować, ale dane już historyczne. Jeżeli w zbiorze rejestruje się 2 mln nowych jednostek na sekundę, a na przeciętnym komputerze grupowanie metodą *k*-średnich takiej liczby przypadków zajmuje około 2,3 sekundy, to algorytm taki staje się nieskuteczny niezależnie od jakości uzyskanego grupowania. W tym czasie bowiem w bazie zarejestrowanych zostanie 4,6 mln nowych jednostek.

Jedną z metod możliwych do zastosowania w grupowaniu jednostek rejestrowanych w dynamicznie zmieniających się bazach danych jest sieć neuronowa typu gazu neuronowego o zmiennej strukturze (*Growing Neural Gas*, GNG) [Fritzke 1994; 1995; Migdał-Najman 2009; Najman 2009]. W dotychczasowych badaniach wykazano, że zapewnia ona grupowanie wysokiej jakości [Netto i in. 2012]. Jest także wysoce autonomiczna, gdyż nie wymaga apriorycznego ustalenia jej struktury czy liczby istniejących skupień [García-Rodriguez i in. 2012]. Jest ponadto oszczędna, ponieważ w procesie samouczenia się osiąga jedynie taką wielkość struktury, która jest niezbędna do odwzorowania badanego zbioru danych. Nie wymaga ani dużej pojemności pamięci komputera, ani znacznej mocy obliczeniowej do grupowania nawet kilku milionów jednostek [Najman 2009; 2010; 2011a; 2011b; 2012]. Wykazano także, że szybkość sieci jest bardzo wysoka. W niewielkim stopniu zależy ona od liczby istniejących skupień i liczby jednostek w bazie danych [Najman 2012]. Celem bieżących badań jest ocena wpływu liczby rejestrowanych cech jednostek na szybkość procesu samouczenia się sieci GNG.

2. Eksperyment badawczy

Aby zrealizować cel badania, przygotowano dwie niezależne symulacje. W pierwszej wygenerowano 3800 zbiorów danych, złożonych z od 2 do 20 skupień (2, 3, 4, …, 20, łącznie 19 wariantów), od 2 do 40 cech (2, 4, 6, 8, …, 20, łącznie 20 wariantów) i od 2000 do 20 000 jednostek (2000, 4000, 6000, …, 20 000, łącznie 10 wariantów). W drugim eksperymencie wygenerowano jeden zbiór danych, złożo-

ny z 50 skupień, 2 mln jednostek, każda opisana przez 202 cechy. Wszystkie zbiory miały charakter dynamiczny. Dane napływały w losowych interwałach czasowych od 10 do 2000 jednostek na sekundę. Każdy przypadek posiadał swój własny czas ważności (generowany losowo od 0,1 do 5 sekund), co powodowało usuwanie ze zbioru danych przypadków oznaczonych jako nieaktualne. Dane usuwane były nieco rzadziej niż rejestrowane nowe przypadki, co gwarantowało wzrost liczby przypadków podlegających grupowaniu. Każdorazowo po dołączeniu do zbioru nowych danych rejestrowano wartość skorygowanego współczynnika Randa [Rand 1971], wartość wskaźnika sylwetkowego [Kaufman, Rousseeuw 1990], a także liczbę iteracji uczących wykonanych od poprzedniej rejestracji i dokładny czas poprzedniej aktualizacji. Pozwoliło to na bieżące kontrolowanie jakości grupowania przez porównanie jego stanu ze znanym wzorcem, ocenę jakości uzyskanej struktury grupowej, a także czas pojedynczej iteracji uczącej sieci GNG.

W eksperymencie pierwszym dla wszystkich zbiorów o danej liczbie cech, niezależnie od liczby jednostek i skupień, wyznaczono średni czas pojedynczej iteracji uczącej, przeciętną wartość skorygowanego współczynnika Randa (RAC) i średnią wartość wskaźnika sylwetkowego (SC). Wyniki prezentuje tab. 2. Zauważyć należy, że wartość skorygowanego współczynnika Randa utrzymywała się na poziomie 0,9 i wyższym. Wartość wskaźnika sylwetkowego utrzymywała się powyżej poziomu 0,8. Oba te wskaźniki pozwalają sądzić, że jakość grupowania jest wysoka. Można także zaobserwować, że czas iteracji wzrasta wraz ze wzrostem liczby wymiarów. Jest to jednak wzrost bardzo powolny. Największe przyrosty można zaobserwować przy małej liczbie wymiarów. Gdy ich liczba jest względnie duża, dodanie kolejnego wymiaru w coraz mniejszym stopniu wpływa na czas pojedynczej iteracji uczącej.

Tabela 2. Średni czas jednej iteracji uczącej, wskaźników RAC i SC w eksperymencie pierwszym

Liczba cech	2	4	6	8	10	12	16	20	24	28	32	36	40
Średnia RAC	0,93	0,90	0,91	0,91	0,93	0,91	0,91	0,91	0,92	0,91	0,90	0,92	0,93
Średnia SC	0,87	0,82	0,85	0,87	0,88	0,88	0,89	0,80	0,81	0,81	0,83	0,82	0,83
Średni czas jednej iteracji (s^{-4})	3,93	4,01	4,08	4,08	4,17	4,14	4,24	4,31	4,34	4,36	4,38	4,42	4,41

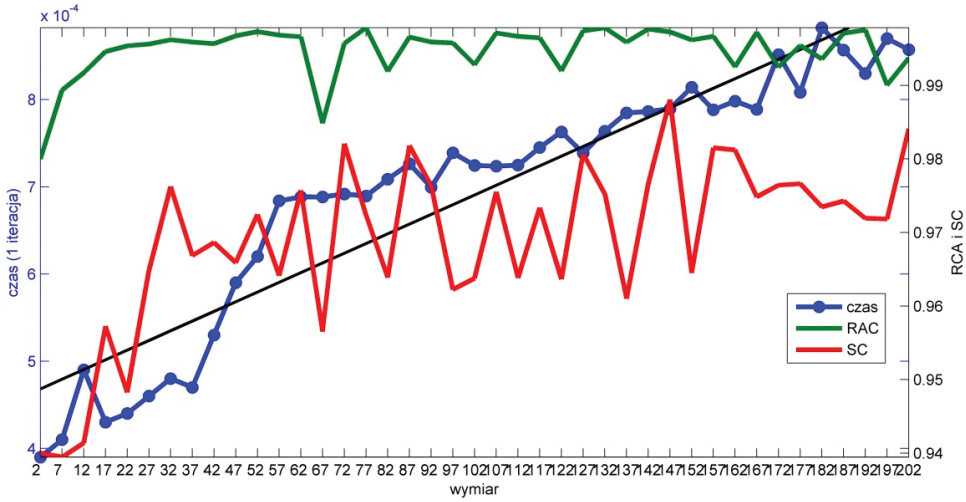
Źródło: opracowanie własne.

Obserwowany wzrost czasu obliczeń wiąże się bezpośrednio z czasem niezbędnym do wyznaczenia odległości między jednostką a neuronem. Im więcej wymiarów, tym mniejsze znaczenie ma dodanie kolejnego.

W pierwszym eksperymencie obserwowano wiele zbiorów danych o zróżnicowanej strukturze grupowej, lecz względnie niewielkiej liczbie wymiarów. Z tego powodu przeprowadzono także drugi eksperyment. Tym razem obserwacji podlegał jeden zbiór danych złożony z nieporównanie większej liczby jednostek, a także pię-

ciokrotnie większej liczby wymiarów. Ten sam zbiór podlegał grupowaniu, biorąc pod uwagę kolejno 2,7,12,17, …, 202 (41 wariantów) cech zmiennych.

Wyniki pomiarów czasu wykonania pojedynczej iteracji zaprezentowano na rys. 1.



Rys. 1. Czas jednej iteracji uczącej, wskaźniki RAC i SC w eksperymencie drugim

Źródło: opracowanie własne.

Podobnie jak w pierwszym eksperymencie wartości skorygowanego współczynnika Randa i wskaźnika sylwetkowego utrzymywały się na poziomie bliskim 1, co świadczy o wysokiej jakości grupowania. Można także zaobserwować wzrost czasu uczenia się sieci w pojedynczej iteracji wraz ze wzrostem liczby wymiarów. Jest on jednak stosunkowo wolny i najszybszy przy małej liczbie wymiarów. Różnica między czasem wykonania jednej iteracji dla 2 i 47 wymiarów wynosi 0,00022 sekundy. Dla 102 i 202 wymiarów jest to już tylko 0,00012 sekundy.

3. Wnioski

Wyniki prezentowanych symulacji nie mają charakteru dowodu formalnego. Uzyskane wartości zależą od konkretnego komputera, języka programowania i umiejętności programisty. Zależą także od samych danych, ich ilości i struktury grupowej. Każdy eksperyment opisuje rzeczywistość jedynie w stopniu założonym przez badacza, a więc ograniczonym. Wydaje się jednak, że uzyskane rezultaty potwierdzają wysoki potencjał sieci GNG w grupowaniu danych zmieniających się dynamicznie, niezależnie od ich wymiaru. Nawet 1000-wymiarowy zbiór danych złożony z 50 skupień, zmieniający ponad 1000 jednostek na sekundę, może być analizowany w tempie 0,00268 sekundy na iterację. W badaniach empirycznych oznacza to setki

iteracji uczących na sekundę, co wystarcza do grupowania nawet bardzo dużych zbiorów o złożonej strukturze.

Literatura

Fritzke B., *Growing cell structures – a self-organizing network for unsupervised and supervised learning*, „Neural Networks", 7, 9, 1994, s. 1441-1460.

Fritzke B., *A growing neural gas network learns topologies*, Advances in Neural Information Processing Systems, 7th edn., MIT Press, Redmond, Washington 1995.

García-Rodriguez J., Angelopoulou A., García-Chamizo J.M., Psarrou A., Escolano S.O., Giménez V.M., *Autonomous growing neural gas for applications with time constraint: optimal parameter estimation*, "Neural Networks", 32, s. 196-208, 2012.

Kaufman L., Rousseeuw P.J., *Finding Groups in Data: a Introduction to Cluster Analysis*, Wiley, New York 1990.

Khanna P., Smith A., *Jobs of the feature*, „Foreign Policy", 13 October, 2012.

Migdał-Najman K., *Analiza porównawcza własności nienadzorowanych sieci neuronowych typu Self Organizing Map i Growing Neural Gas w analizie skupień*, [w:] Taksonomia 16, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe UE we Wrocławiu nr 47, 2009, s. 205-213.

Najman K., *Zastosowanie nienadzorowanych sieci neuronowych typu Growing Neural Gas w analizie skupień*, [w:] Taksonomia 16, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe UE we Wrocławiu nr 47, 2009, s. 196-204.

Najman K., *Ocena wpływu parametrów sterujących procesem samouczenia się sieci GNG na ich zdolność do separowania skupień*, [w:] Taksonomia 17, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 17, 2010, s. 296-304.

Najman K., *Grupowanie dynamiczne i wykorzystaniem sieci GNG*, „Przegląd Statystyczny", nr 3-4, 2011a, s. 231-241.

Najman K., *Propozycja algorytmu samouczenia się sieci neuronowych typu GNG ze zmiennym krokiem uczenia*, [w:] Taksonomia 18, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, 2011b, s. 282-289.

Najman K., *Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG*, [w:] Taksonomia 19, *Klasyfikacja i analiza danych – teoria i zastosowania,* Prace Naukowe UE we Wrocławiu nr 242, 2012, s. 361-369.

Netto S.M.B., Silva A.C., Nunes R.A., Gattass M., *Automatic segmentation of lung nodules with growing neural gas and support vector machine*, "Computers in Biology and Medicine", 42, 11, 2012, s. 1110-1121.

Rand W.M., *Objective criteria for the evaluation of clustering methods*, „Journal of the American Statistical Association", 66, 336, 1971, s. 846-850.

SELF-LEARNING NEURAL NETWORK OF GNG TYPE IN THE DYNAMIC CLUSTERING OF HIGH-DIMENSIONAL DATA

Summary: In the article the author presents the results of simulation research that involves the use of self-organizing neural networks of GNG type in the dynamic clustering high-dimensional data. The author performed two independent experiments on the generated data. He measured the learning speed of neural networks depending on the size of the data. It seems that the results of research confirm the usefulness of GNG neural network in the dynamic clustering.

Keywords: Growing Neural Gas (GNG) network, dynamic clustering.