

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

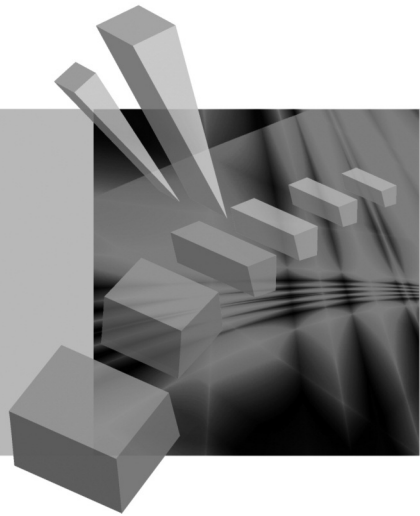
RESEARCH PAPERS

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 Pocięcha,
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](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

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

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

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka, Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska, Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński, Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz, Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel, Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień, Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębkowska, Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan, Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska, Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka, Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański, Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk, Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk, Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura, Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk, Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski, Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka, Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk, Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarc , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

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 profil¹, 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 historycznym². 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.

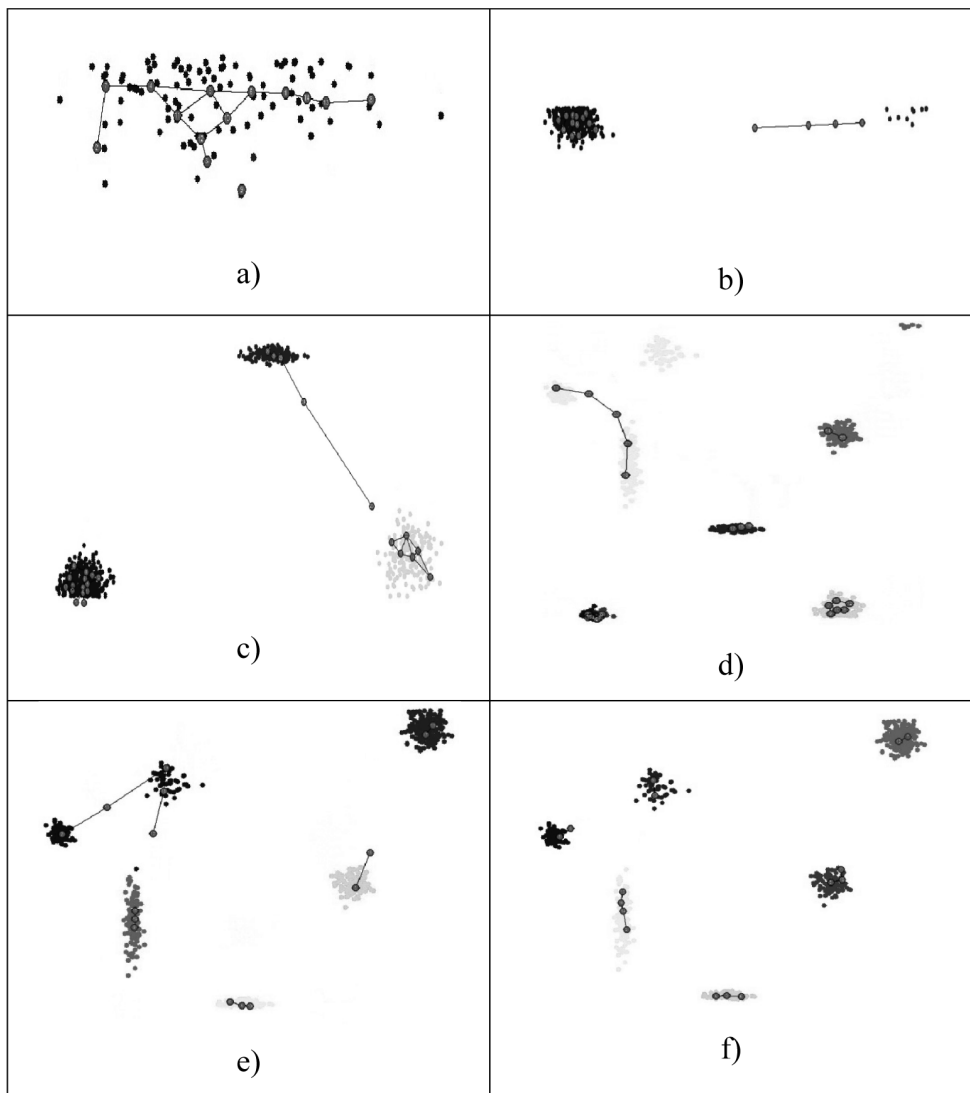
Do zweryfikowania pozostaje kluczowy element dobrego algorytmu grupowania dynamicznego – tj. szybkość jego działania.

4. Eksperyment badawczy

W celu weryfikacji możliwości zastosowania sieci neuronowej typu GNG w grupowaniu dynamicznym przygotowano eksperyment³. Wygenerowano dynamiczną bazę danych⁴, 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 zero. 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.



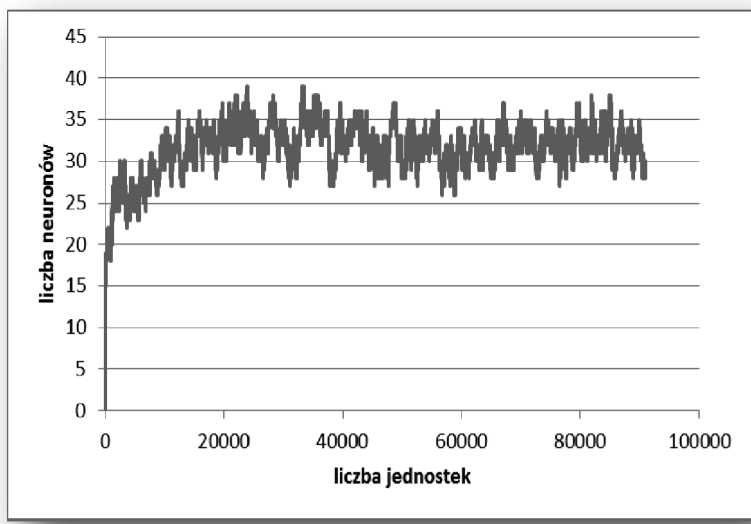
Rys. 1. Przebieg grupowania dynamicznego

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.

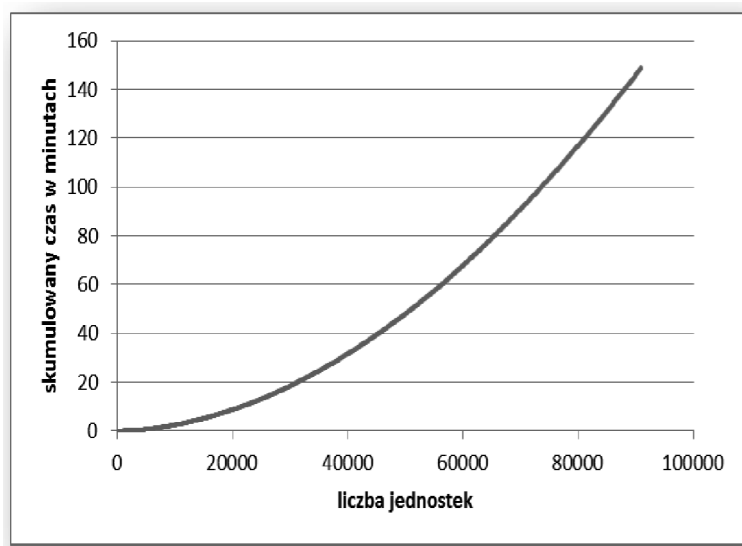


Rys. 2. Liczba neuronów na sieci GNG a liczba jednostek

Źródło: opracowanie własne.

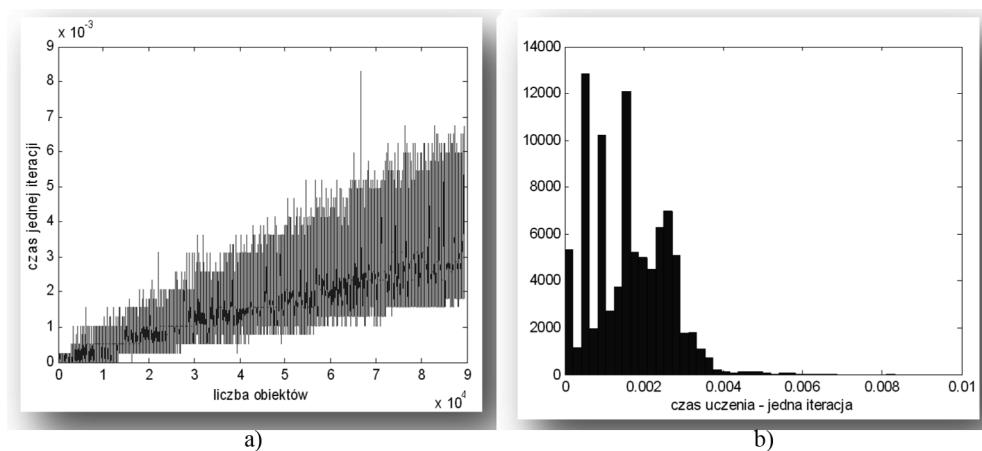
Czas uczenia się sieci jest wysoce niezależny od liczby jednostek w bazie danych. W jednej iteracji uczą się jedynie dwa neurony: neuron wygrywający (znajdujący się najbliższej obserwowanej jednostki) i drugi najbliższy. Liczba modyfikowanych neuronów nie zależy więc od liczby jednostek ani od rozmiaru 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.



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 rozdzielania skupień słabo seprowalnych. Warto także zauważyć, że nie istnieje w tej chwili oprogramowanie komputerowe umożliwiająca szerszej grupie badaczy analizę własności tej sieci, jak również stosowanie jej w praktyce⁵. 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.