ISSN 1644-0757
www.acta.media.pl

Acta Sci. Pol. Oeconomia 14 (3) 2015, 95–104

DYNAMICAL CLUSTERING OF STREAMING DATA WITH A GROWING NEURAL GAS NETWORK

Kamila Migdał-Najman, Krzysztof Najman
University of Gdansk

Abstract. One of characteristic feature of contemporary data bases is their growing dynamics. The number of registered entities as well as their group structure tends to dynamically grow. In order to effectively determine the rapidly changing number and structure of clusters, appropriate methods of cluster analysis have to be applied. The paper presents the results of simulation research concerning the possibility of applying self-learning GNG neural networks in clustering data from data streams.

Key words: cluster analysis, data streams clustering, Growing Neural Gas network

INTRODUCTION

A characteristic feature of the developing economies of many countries is a constant increase in demand for information. The globalization of world markets and increasing competition forces companies to use the information accurate, precise, comparable, reliable and current. In addition to other features of the information, the increasing role of its topicality. Increasingly, it is no longer about information simply to date, but on information from the "last minute". Must contain the description of the phenomenon studied, the real "now".

Examples of such information may be recording data on financial transactions on stock exchanges worldwide. Active investor makes decisions based on what is happening in the market at the moment. A similar situation exists in the company responsible for the security of transactions on credit cards. Every day in the world there is a lot of millions of transactions using credit cards. However, some may be ineligible. When the owner of the card for many years. He made her small payments, and suddenly the system records a sample of transactions on a significant amount, perhaps a card has been stolen. Reaction institution supervising transactions must be taken immediately. Approve the transaction

Corresponding author: Kamila Migdał-Najman, University of Gdansk, Faculty of Management, Department of Statistics, Armii Krajowej 101, 81-824 Sopot, Poland, e-mail: kamila.migdał-najman@ug.edu.pl

or block? Every day, retailers are visited by hundreds of thousands of customers with different shopping preferences. The development of technology has made these institutions have extensive data sets on the individual customer purchase transactions. Such registration is carried out by electronic readers that automatically reads the barcode information on the purchased products. It is possible to automate this process searchable collections to find information about your preferences and shopping habits of customers, identification of combinations of products purchased together by customers. With even greater intensity stream of data we are dealing in computer systems that support the Internet. Every minute of users post on YouTube 48-hours movies. On Facebook at this time appears 684,478 new materials. Google search logs 2,000,000 questions from Internet users per second. Which Web pages open Internet users, how often, and in what order pass between successive links, which the advertisement was displayed at this point how much time the user spent on the site? The answer to these questions is essential in developing an optimal strategy for the promotion of the product, in the assessment of its costs and the expected results. This assessment is not easy. The number of existing web pages is huge and still growing dynamically.

One of the problems connected with the analysis of data contained in the contemporary data bases is high volatility of their content. New observations may be registered many times within a second, dynamically changing the image of the observed population. Not only the number of units may dynamically change, but also their group structure. In line with the flow of time and inflow of new data, the known and well defined clusters may lose their importance or be dissolved in other clusters. Clusters, which contain rare species, the clusters which are poorly defined (blurry) may become more numerous, better defined or even dominant. Completely new clusters may also emerge. It is also possible, that data have limited validity periods assigned to them, after which they lose their significance and their influence on the current structure of the population. In consequence it may lead to the disappearance of the existing clusters. To properly observe the studied population cluster structure change process it is necessary to continuously group particular objects and make corrections in the description of clusters – in line with the inflow of new data. It is indispensable to apply the clustering method which would be capable to react to each new information and to automatically make indispensable corrections in the description of the existing group structure. This description should be available at any time.

RELATED WORK

Since early 1990s a rapid growth in the number of data bases and their content could be observed. In many cases the inflow of new data is very fast and practically unlimited. A good example may be the registration of data about financial transactions at the world stock exchanges, recording of: financial transactions made with credit cards, transactions in Internet shops, credit applications in the headquarters of a bank, phone calls by a telecommunication company, log-ins in Internet services etc. Data of this kind are called data streams. They differ from typical, static data with several features. First of all the content of the data base changes in dynamic way, sometimes several hundred times per second and its size is unlimited. Secondly, the cluster structure of recorded objects may rapidly change. The number of clusters can also change. In the analysis of data streams the crucial role is played by the limitations of time and memory resources, leading to the equally unfavourable phenomenon of underfitting of classification models [Domingos and Hulten 2000]. The data stream may be so rapid and made of such a great number of objects, that there will be no possibility for it current analysis.

To group objects coming out of data streams it is necessary to apply special algorithms. They may be divided into four basic groups. The first group is made of the incremental or online classifiers. These are such algorithms as the Very Fast Decision Tree algorithm (VFDT) [Domingos and Hulten 2000] and its extension: Concept Drift Very Fast Decision Tree algorithm (CVFDT) [Hulten et al. 2001]. The second group is made of the multimodel algorithms such as Ensemble Classifiers (EC) [Kolter and Maloof 2003, Wang et al. 2003]. The third group of algorithms contains the low granularity Rule Based Classifier proposed by Wang et al. [2007]. This group contains as well the methods based on genetic algorithms – GA [Vivekanandan and Nedunchezhian 2011]. The fourth group of stream data analysis methods are Anytime algorithms. They were discussed for the first time by Dean and Boddy in 1988 [Dean and Boddy 1988]. In cluster analysis their applications were studied among others by Vlachos et al. [2003], as well as by Kranen et al. [2011, 2012].

In the study presented below, the data grouping algorithm based on the self-learning of neural network of the Growing Neural Gas (GNG) type will be presented. Compared to the classical Fritzke algorithm [Fritzke 1994] it has been substantially modified. A variable step of winning neuron learning and neurons connected to it was introduced. Learning step size was also made dependent on the rate of inflow of new information, distinguishing static and dynamic phase of the self-learning process. These changes significantly increase the speed of learning and quality of clustering network for streaming data. In the version described below it may be assigned as unsupervised version of the third and fourth group of data stream clustering algorithms.

CONSTRUCTION AND SELF-LEARNING ALGORITHM OF GNG NETWORK

A classic algorithm of construction and self-learning of GNG networks was proposed by Fritzke in 1994 [Fritzke 1994]. His idea reduced the typical problems with self-learning networks of the Self Organizing Map (SOM) type [Kohonen 1995]. Such networks have an *a priori* assumed structure, which does not change in the self-learning process. An optimal structure is, however, not known in advance. Another practical problem with the SOM network is the objective determination of the number and borders between clusters [Migdał-Najman and Najman 2008]. The GNG network was assumed to dynamically change its structure, adopting it to the real needs. It should also divide the existing clusters by itself. The self-learning algorithm of the GNG network, with appropriate modifications could be used to search for dynamically changing group structure of the observed objects (cases of database). Its essence may be presented in the following way:

Let D be an M-element of the set of objects in an n-dimensional space:

$$D=\{\xi_1,...,\xi_M\},\xi_i \in \Re^n \tag{1}$$

Each object $\xi \in D$ is described by n-element of set of data vectors (data vectors).

Let A be a k-element, n-dimensional set of neurons:

$$A=\{c_1,...,c_k\},c_i \in \Re^n \tag{2}$$

To each neuron $c \in A$ is connected a reference vector (w_c), which can be considered as vector of neuron coordinates in input space (input space).

The initial set of neurons is composed of two elements, $k = 2$. The self-learning process starts with the initiation of c_1 and c_2 neurons with random weights (co-ordinates in the space of analysed objects):

$$A=\{c_1,c_2\} \tag{3}$$

The connection between them and the age of the connection is set to 0.

From the set D is randomly selected one object ξ (data vector). Among the existing neurons the following ones are looked for: the neuron closest to the selected object and the second closest one:

$$s_1=\arg\min_{c\in A}\left\|\xi-w_c\right\|,\ s_2=\arg\min_{c\in A\setminus\{s_1\}}\left\|\xi-w_c\right\| \tag{4}$$

The s_1 neuron is called the winning neuron. As a measure of the distance of objects from neurons (data vectors and reference vectors) it is necessary to adopt the appropriate metrics, corresponding to the measurement scales used in the study. If these neurons are not connected, such connection should be created and its age set to 0. The age of connection is the number of consecutive iterations in which the neuron is not the winning neuron. Then a learning stage of the s_1 and s_2 neurons is initiated. In the first step a local measure of the network error for the s_1 neuron is determined:

$$\Delta E_{s_1}=\left\|\xi-w_{s_1}\right\|^2 \tag{5}$$

It is a classical quantisation error. Then all neurons connected with the s_1 neuron are looked for and their coordinates are updated:

$$\Delta w_{s_1}=\varepsilon_b\left(\xi-w_{s_1}\right),\ \Delta w_i=\varepsilon_n\left(\xi-w_i\right),\ \left(\forall i\in N_{s_1}\right) \tag{6}$$

where: i means the i-th neuron connected with the winning one [Jirayusakul and Auwatanamongkol 2007]. The speed of learning of the winning neuron (moving of neuron s_1 and s_2 toward object ξ) is determined by ε_b and other connected neurons by ε_n. The age of connections between all neurons, whose coordinates have been updated is increased by 1. Then all connections between neurons older than the maximum seaget (age$_{\max}$) are removed. It is then checked, whether the s_1 neuron remained to be connected with any other neuron. If it lost all connections, it is removed.

The above procedure is repeated for successive drawn objects ξ. If the number of random objects so far is equal to a multiple parameter λ, the procedure for inserting a new neuron begins. If the connections existed, the procedure for inserting a new neuron is initiated. The neuron with the maximum quantisation error q is looked for and the neuron f closest to it. The new neuron r is placed between the q and f neurons, creating its coordinates by interpolation of coordinates of the q and f neurons:

$$A = A \cup \{r\}, \quad w_r = \left(w_q + w_f\right)/2 \tag{7}$$

At the same time the connections between neurons are modified by removal of the connection between q and f, and then linking the neuron q with r and the neuron f with r. The age of those connections is set at 0. The quantisation error for the new neuron is also determined:

$$E_r = \left(E_q + E_f\right)/2 \tag{8}$$

where:

$$E_q = \left\|\xi - w_q\right\|^2, \quad E_f = \left\|\xi - w_f\right\|^2 \tag{9}$$

This is the last stage of the algorithm, when the stop conditions are tested, as follows: achieving the assumed maximum number of iterations – $it_{\max}$, achieving a minimum assumed learning error of networks – $MQE_{\min}$ (Mean Quantisation Error) and reaching maximum assumed number of neurons – $k_{\max}$. The fulfillment of any condition ends the algorithm.

Among the drawbacks of the algorithm it is possible to identify a number of parameters, which have to be determined *a priori*.These are: maximum number of iterations, maximum number of neurons and the maximum age of connections, minimum network error and the frequency of new neuron λ iteration. These parameters are difficult to determine *a priori* because of the lack of simple formal dependencies between them and the quality of cluster reconstruction. In the process of dynamic clustering the problem becomes somewhat simpler. Since the self-learning process must have a continuous character, it is easier to determine the key values of parameters. If $k_{\max} > 2M$, $it_{\max} = Inf$, $MQE_{\min} < 0$, the self-learning process of the GNG network will not be automatically interrupted.

Further modifications result from the possibility of the change in the existing group structure in time. The algorithm should behave differently, when the structure is not subject to change, and when it changes. To attain the adaptability of the self-learning process of the GNG networks it is necessary to correct λ the size of learning step ε_b and ε_n [Najman 2011a, 2011b]. The change in the learning step can be accomplished either discretionally (one value of the learning step for the static phase and one for the dynamic phase) or functionally (one value of the learning step for the static phase and e.g. a linear function of change of the learning step for the dynamic phase). To additionally accelerate the process of recognition of new clusters it is possible to increase the frequency of insertion of new neurons. The distinction of phases may proceed on the basis of the measurement of the quality of the achieved group structure. One of the many known indices may be used to that end, such as e.g. Silhouette Coefficient [Kaufman and Rousseeuw 1990] or others. If the value of the index is subject to a discretional change, it is usually connected with the appearance of a new cluster or division of an existing cluster into parts. In both cases changes of group structure may be significant and the network should switch to the dynamic phase.

AN EXPERIMENTAL STUDY: DATA SETS

In the empirical research artificial datasets were used. Altogether 3,235 dynamic data bases were generated. Each base consisted of 20,000 objects with random group structure (spherical and separable focus) consisting of 2 to 20 clusters, 2- to 40-dimensional. For each database was separately build a new GNG network. Clustering started with 100 randomly selected objects from the database. After each iteration of network learning, new objects were added to the database at the same time removing those that have been longest inside. Number of added and deleted objects in each case was determined randomly (add max 40 objects, remove max 15 objects) in such a way that the current size of the database was gradually increased. The initial database is treated as a population, and its passage, clustered at the moment, as a test. In this way, the appropriate changes of the number of objects and the structure of the cluster in time were provided. During all experiments, the maximum number of clusters existing at the same time was 15, and the maximum number of objects grouped at the same time was 1,345.

In the process of clustering was observed, in each iteration, compatibility of the achieved group structure with the known, established *a priori* model, measured with corrected ratio Rand (Rand Adjusted Statistic) [Rand 1971] the achieved group structure measured with the Silhouette Coefficient, the time of single learning iteration measured in seconds (values given in seconds refer to a typical personal (PC) computer equipped with an Intel(R) Core(TM) i5-2434M CPU 2.4 GHz processor), the current number of neurons and objects in the database. Because it is believed that when the SC value is less than 0.7, a set of objects has a difficulties to identify the structure of the cluster [Kaufman and Rousseeuw 1990], the switching from the static to the dynamic phase was effected when the value of the Silhouette Coefficient fell below 0.7. In the static phase the learning step was adopted at the level $\varepsilon_b = 0.02$ and $\varepsilon_n = 0.004$, while in the dynamic phase $\varepsilon_b = 0.06$ and $\varepsilon_n = 0.001$. New neuron was inserted into the network every 200 iterations ($\lambda = 200$). These parameters were determined experimentally, taking into account the extreme values of object features.

EXPERIMENTAL RESULTS

Summing up the results of experiments, it can be said, that the compatibility of grouping with the known model was very high (Table 1). The average value of the Rand Adjusted Statistic was higher than 0.95. Lower values were received only for clusters 2, 3 and 14, 15. However, these were only exceptional and temporary situations, at low number

Table 1. Quality and speed of learning of the GNG network in relation to the number of existing clusters

Clusters	2	3	4	5	6	7	8
Mean Adjusted Rand Index	0.751	0.918	0.969	0.983	0.988	0.988	0.990
Mean Silhouette Coefficient	0.502	0.736	0.860	0.911	0.933	0.935	0.944
Mean Time 1 Iteration	4.1 E-04	4.2 E-04	4.3 E-04	4.3 E-04	4.4 E-04	4.3 E-04	4.3 E-04
Clusters	9	10	11	12	13	14	15
Mean Adjusted Rand Index	0.987	0.989	0.985	0.976	0.959	0.928	0.831
Mean Silhouette Coefficient	0.945	0.939	0.930	0.923	0.885	0.816	0.758
Mean Time 1 Iteration	4.3 E-04	4.3 E-04	4.3 E-04	4.4 E-04	4.4 E-04	4.5 E-04	4.5 E-04

Source: Own research.

of clusters they were connected with their excessive division while at a high number of clusters they were connected with their excessive connection.

The achieved group structure may be considered as a good one. The Silhouette Coefficient has attained an average value of more than 0.85. The values lower than 0.8 were attained rarely and temporarily in identical situations as was the case with lower values of the Rand Adjusted Statistic. The switch of the network to the dynamic phase resulted in a radical improvement of the group structure after several learning iterations.

The average time of learning in a single iteration amounted to about 0.00042 seconds. It has slightly increased in line with the increase in the number of clusters. This growth was, however, not directly connected with the existing number of clusters. At a high number of clusters the group structure was subject to more frequent change. It entailed a more frequent switching of the self-learning process from the static phase into a dynamic phase. In the dynamic phase the neurons are inserted and removed more often and this process is responsible for the increase in the time of learning. In the static phase the learning time in a single iteration does not depend on the number of clusters.

The compatibility of classification with the model depends to a certain degree on the number of dimensions of the space (Table 2). The value of the Rand Adjusted Statistic remains at a level higher than 0.9, apart from two-dimensional sets. In two-dimensional sets, the network identified local changes in the object density and has excessively divided the existing clusters independently from the number of clusters. Similarly, the cluster structure, independently from the number of dimensions, has been relatively well recognised. The average value of the Silhouette Coefficient amounted to over 0.75.

The time of one learning iteration of the network depends barely on the dimension of the data. It is related to the time of computing the distance between the object and the neuron. The learning time-related to the dimension of data attained approximately the following values:

$$time = -1.3E - 007 \times dimension^2 + 5.2E - 006 \times dimension + 0.00039 \tag{10}$$

Table 2. Quality and speed of learning of the GNG network depending on the number of dimensions

Dimensions	2	4	6	8	10	12	16
Mean Adjusted Rand Index	0.83	0.90	0.91	0.91	0.93	0.91	0.91
Mean Silhouette Coefficient	0.57	0.72	0.75	0.77	0.78	0.78	0.79
Mean Time 1 Iteration	3.9 E-04	4.0 E-04	4.1 E-04	4.1 E-04	4.2 E-04	4.1 E-04	4.2 E-04
Dimensions	20	24	28	32	36	40	
Mean Adjusted Rand Index	0.91	0.92	0.91	0.90	0.92	0.93	
Mean Silhouette Coefficient	0.80	0.81	0.81	0.80	0.82	0.83	
Mean Time 1 Iteration	4.3 E-04	4.3 E-04	4.4 E-04	4.4 E-04	4.4 E-04	4.4 E-04	

Source: Own research.

The time of a single iteration of learning of the network depends on the number of objects in the data base. This relationship has a linear character:

$$time = 5E - 008 \times objects + 0.00043 \qquad (11)$$

For 10,000 cases it would amount to some 9.3E-4 seconds, for 100,000 cases some 0.0054 seconds and for 1,000,000 cases about 0.504 seconds. Then, may by assumed, that the network is learning relatively quickly.

The time of single learning iteration of the network depends linearly on the number of neurons in that network (Table 3). This relationship is clearly discernible, since the Pearson's Correlation Coefficient amounts to 0.8. This is not a problem for the GNG network since the number of neurons does not depend on the number of objects but only on the number of clusters only and the level of complication of the cluster structure. For this reason, in the experiment, the average number of neurons depends in a small extent on the number of grouped objects. Averaging the number of neurons on the number of

Table 3. Learning speed of the GNG networks and the number of neurons related to the number of cases

Number of cases	300	350	400	450	500	550	1,000
Mean Time 1 Iteration	4.7 E-04	5.0 E-04	4.8 E-04	5.2 E-04	4.9 E-04	4.8 E-04	4.3 E-04
Mean number of Neurons	30.1	33	27	35.7	33.5	32.5	22
Number of cases	1050	1100	1150	1200	1250	1300	1350
Mean Time 1 Iteration	4.4 E-04	4.9 E-04	4.6 E-04	4.7 E-04	4.2 E-04	5.4 E-04	5.2 E-04
Mean number of Neurons	28.5	33	32.5	31.5	28	25	34

Source: Own research.

dimensions and the number of existing clusters from 300 to 1,350 the number of objects very similar values were obtained. In all simulations, the minimum number of neurons was 5 and the maximum – 47.

CONCLUSIONS

The experiment allows us to think, that the self-learning networks of the GNG type may be an effective instrument for objects dynamic clustering. In the overwhelming majority of cases, a high compatibility of clustering with the model and the achieved group structure was being achieved throughout the time of operation of the network. The network has confirmed its economical character. It is difficult to assess the frequency of insertion of neurons (λ).

The main problem of GNG network construction remains to determine the parameters of its construction. Some parameters controlling the operation of algorithm in dynamic clustering loses its relevance ($it_{\max}$, $MQE_{\min}$, $k_{\max}$). Other parameters may be subject to self-regulation (ε_b and ε_n). It is also difficult to determine the threshold value at which the network should switch from the dynamic into the static phase. All the received results concern as well the data which do not contain any meaningful noise and concern only separable clusters. The above mentioned problems will be subject of further research.

REFERENCES

Dean, T., Boddy, M.S. (1988). An analysis of time-dependent planning. [In:] Proceedings of the seventh National Conference on Artificial Intelligence, AAAI, St. Paul, 49–54.

Domingos, P., Hulten, G. (2000). Mining high-speed data streams. [In:] Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, 71–80.

Fritzke, B. (1994). Growing cell structures – a self-organizing network for unsupervised and supervised learning. Neural Networks, 7, 1441–1460.

Hulten, G., Spencer, L., Domingos, P. (2001). Mining time-changing data streams. [In:] KDD'01 Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, NY, 97–106.

Jirayusakul, A., Auwatanamongkol, S. (2007). A supervised growing neural gas algorithm for cluster analysis. International Journal of Hybrid Intelligent Systems, 4, 129–141.

Kaufman, L., Rousseeuw, P.J. (1990). Finding groups in data: a introduction to cluster analysis. Wiley, New York.

Kohonen, T. (1995). Self-organizing maps. Springer-Verlag, Berlin-Heidelberg.

Kolter, J.Z., Maloof, M.A. (2003). Dynamic weighted majority: a new ensemble method for trakking concept drift. [In:] Proceedings of the third IEEE International Conference on Data Mining, Los Alamitos, 123–130.

Kranen, P., Assent, I., Baldauf, C., Seidl, T. (2011). The ClusTree: Indexing micro-clusters for anytime stream mining. Knowledge and Information Systems, 29, 249–272.

Kranen, P., Assent, I., Seidl, T. (2012). An index-inspired algorithm for anytime classification on evolving data streams. Datenbank-Spektrum, Springer DASP, 12, 43–50.

Migdał-Najman, K., Najman, K. (2008). Data analysis, machine learning and applications, applying the Kohonen self-organizing map networks to selecting variables. [In:] C. Preisach,

H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds), Studies in Classification. Data Analysis and Knowledge Organization. Springer Verlag, Berlin-Heidelberg, 45–54.

Najman, K. (2011a). Propozycja algorytmu samouczenia się sieci neuronowych typu GNG ze zmiennym krokiem uczenia. Taksonomia 18, Klasyfikacja i analiza danych – teoria i zastosowania. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, 176, 282–289.

Najman, K. (2011b). Dynamical clustering with Growing Neural Gas networks. Statistical Review, 3–4, 231–242.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846–850.

Vivekanandan, P., Nedunchezhian, R. (2011). Mining data streams with concept drifts using genetic algorithm. Artificial Intelligence Review, 36, 163–178.

Vlachos, M., Lin, J., Keogh, E.J., Gunopulo, D. (2003). A wavelet-based anytime algorithm for k-means clustering of time series. [In:] ICDM Workshop on Clustering High Dimensionality Data and its Applications. SIAM Data Mining, San Francisco.

Wang, H., Fan, W., Yu, P.S., Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. [In:] Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, NY, 226–235.

Wang, P., Wang, H., Wu, X., Wang, W., Shi, B. (2007). A low-granularity classifier for data streams with concept drifts and biased class distribution. IEEE, Transactions on Knowledge and Data Engineering, 19, 1202–1213.

GRUPOWANIE DYNAMICZNE STRUMIENI DANYCH Z ZASTOSOWANIEM SIECI TYPU GROWING NEURAL GAS

Streszczenie. Jedną z charakterystycznych cech współczesnych zbiorów danych jest ich dynamika. Liczba zarejestrowanych obiektów, jak również ich struktura grupowa potrafi zmienić się wielokrotnie w ciągu sekund. W celu skutecznego wykrycia liczby skupień i struktury grupowej rejestrowanych obiektów konieczne staje się zastosowanie specjalnych metod analitycznych. W artykule przedstawiono wyniki badań symulacyjnych w zakresie możliwości zastosowania samouczących się sztucznych sieci neuronowych typu GNG w grupowaniu strumieni danych.

Słowa kluczowe: analiza skupień, grupowanie strumieni danych, sieci typu Growing Neural Gas

Accepted for print: 10.08.2015

For citation: Migdał-Najman K., Najman K. (2015). Dynamical clustering of streaming data with a Growing Neural Gas network. Acta Sci. Pol., Oeconomia, 14 (3), 95–104.