Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure

Roman Tkachenko [1] [0000-0002-9802-6799], Ivan Izonin [1] [0000-0002-9761-0096],
Natalia Kryvinska [2] [0000-0003-3678-9229], Valentyna Chopyak [3] [0000-0003-3678-9229],
Nataliia Lotoshynska [1] [0000-0002-6618-0070], Dmytro Danylyuk [3] [0000-0002-7661-6341]

[1] Lviv Polytechnic National University, Lviv, S. Bandera, 12, 79013, Ukraine
{roman.tkachenko,ivanizonin, natlotsv}@gmail.com
[2] University of Vienna, Universitätsring 1, 1010, Vienna, Austria,
natalia.kryvinska@univie.ac.at
[3] Danylo Halytsky Lviv State Medical University, 69 Pekarska str., Lviv, 79010, Ukraine
chopyakv@gmail.com, dimadanylyuk90@gmail.com

Abstract. The article proposes a new insurance medical cost prediction method. It is based on the piecewise-linear approach using the SGTM neural-like structure. Piecewise-linear approach provides high processing efficiency for large amounts of data, and the SGTM neural-like structure provides high accuracy and high-speed training procedure. The simulation of the proposed method using real data on health insurance costs and two SGTM neural-like structure cascades was performed. The high speed and accuracy of the proposed method were experimentally determined. The comparison of the proposed method was carried out with the existing methods, in particular, multilayer perceptron and the Common SGTM neural-like structure, which solved the task using all dataset. It was found that the worst results show a multilayer perceptron: the accuracy of its operation according to MAPE is more than 23% less than the accuracy of the proposed method, and the time of the training procedure lasts 51 times longer. The baseline method, Common SGTM neural-like structure, shows a higher learning speed but less precision - an 11% greater error than the developed method. The obtained results showed the possibility of using the proposed approach for the processing of a large amount of data, in particular in the fields of medicine, economics, materials science, service sciences.

Keywords: Approximation, Piecewise-linear Approach, Prediction Task, Neural-Like Structure, Successive Geometric Transformations Model, Insurance Costs.

1 Introduction

Insurance medicine is one of the pillars for the development of the health care system as in the world as in Ukraine [1, 2]. Mandatory and voluntary health insurance, which are provided by the law of Ukraine, are not sufficiently developed. A voluntary form

as the insurance companies' service takes a small share in the market of these services. The main reason for this is the distrust of the potential customers to this service.

The insurance market in Ukraine should be developed taking into account the best world trends. The personalized approach [2], in particular, potential risks identification, to identify individuals for whom it is necessary to intensify health management at the right time is a rather important problem. Its effective solution depends on many separate tasks, where one of the important is the individual health insurance costs prediction. This task is complicated by the individual data characteristics for each particular case [3, 4].

That is why the prediction task, in this case, should be based on a personalized Data-Driven approach, which will take into account many factors [4]. It is the large amount of data [5, 6], the influence of each individual factor [7] and multiparametric dependencies between variables [8] that are not fully studied. These factors necessitating the use of artificial intelligence tools to solve this task [9, 10].

In this paper, we propose the piecewise-linear approach to solving the individual health insurance cost prediction task. It will provide a number of advantages for the processing of large volumes of data, which is typical for this area [5]. Among the most important ones it should be noted the possibility of increasing the prediction accuracy by dividing the total dataset into sub-samples (clusters) of the same data type, and their effective processing.

2 Literature review and problem statement

There are many approaches to solving the health insurance costs predicting task. Regression methods, which are often used for these purposes, do not always provide sufficient precision. This is due to the actual characteristics of the data and the fact that in most cases they are not normally distributed and do not satisfy the assumption of homoscedasticity [5, 11]. In addition, some optimization methods in this class are quite time-consuming. The precision of the data-mining methods based on the decision trees and clustering depends on the accuracy of the chosen clustering method [3]. The neural network and neuro-fuzzy models do not always provide sufficient precision [12]. In addition, the iterative training algorithms, which are the basis of their work, require a long time to work [13]. This is a significant drawback for a large amount of data processing [14].

In [15], the piecewise-linear approach was proposed to solve a regression problem. Its benefits are evident in a large amount of data processing, where the accuracy of the method's work is critical. In [16, 17], for solving the high-speed classification task based on the piecewise stepping approximation, it is proposed to use the SGTM neural-like structure. It provides satisfactory results of the classification, and moreover, the non-iterative SGTM neural-like structure's training algorithm provides the high speed of the proposed method.

The aim of this work is the solution of the regression task based on the piecewise-linear approach using SGTM neural-like structure. Another difference from the works

[16, 17] is that the dataset division into a sub-sample occurs in a new, developed way, without the clustering algorithms usage.

3 Dataset Description

To solve the insurance prediction task the dataset from [18] is used. It collected 1338 observations about insurance costs in four USA regions. Dataset detailed characteristics are given in Table 1.

Table 1. Dataset characteristics.

Variable type	Variable name	Data characteristics
input	Age	from 18 to 64 years old; 39.2 is mean value
input	Gender	662 female and 676 male
input	Body mass index, kg/m^2	min. value: 15.96; max. value: 53.13; mean value: 30.66
input	Children	from 0 to 5; 1.095 is mean value
input	Smoking	1064 smokers and 274 no-smokers
input	Beneficiary's residential area	observations in USA: 364 in southeast; 324 in northeast; 325 in southwest; 325 in northwest
output	Insurance charges	min. value: 1122; max. value: 63770; mean value: 13270

The dataset visualization that was used for modeling are shown on fig. 1. The Orange software was used for this aim [19].

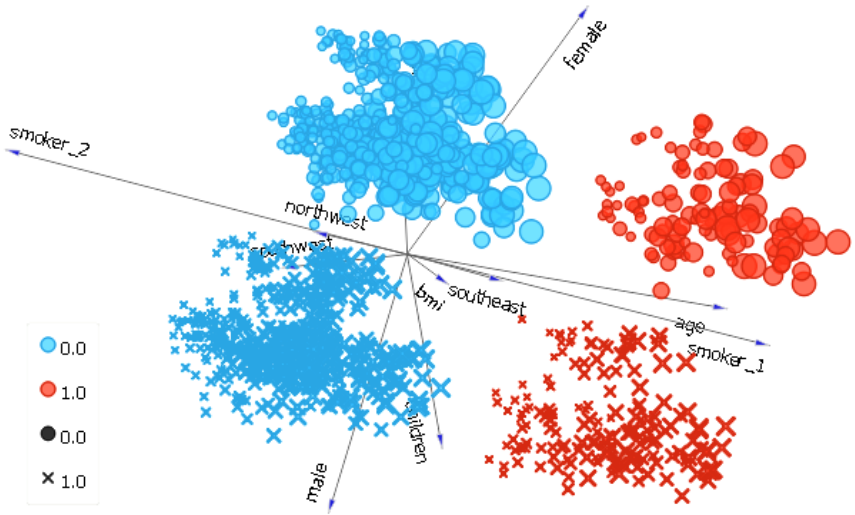


Fig. 1. Visualization of the data set using the FreeViz machine learning method (after optimization). The red colour denotes smokers, and the blue colour is not smokers, the cross represents men, the circle represents women. The size of the corresponding figure visualizes the age: the larger the figure's size, the greater the person's age. Colours and shapes are chosen randomly.

4 Piecewise-linear approach for prediction based on the SGTM neural-like structure

The features of the selected dataset are that the data is completely different. Insurance costs range from 1121.8739 to 63770.42801 monetary units. Its average value is 13270.42227 monetary units. This can be one of the reasons of why many regression techniques do not provide a sufficiently precise solution [20]. In addition, the impact of additional factors such as smoking, the number of children, body mass index, etc., on the value of individual health insurance costs is not fully understood. That is why the paper proposes a piecewise-linear approach to the solution of this task. It will increase the accuracy of prediction by processing each individual data cluster that contains homogeneous values.

The speed of such an approach depends on the chosen artificial intelligence instrument namely from the time of its training procedure. That is why the SGTM neural-like structure was chosen. It provides high-performance for large amounts of data processing while retaining high generalization properties [21]. This becomes possible due to the greedy non-iterative training algorithm [18, 21]. It provides the possibility of using the chosen tool to process large amounts of data, including for hardware implementation [22]. The piecewise-linear approach usage will increase the performance of each individual SGTM neural-like structure, which will ultimately provide an effective solution for the task. Fig. 2 shows the SGTM neural-like structure topology, which is used for modelling.

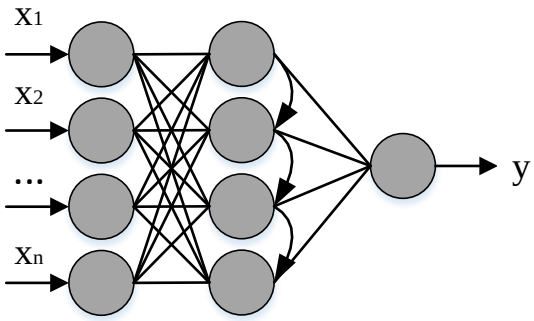


Fig. 2. SGTM neural-like structure's topology.

The description of the training and test modes is described in [20].

Figure 3 shows the structural scheme of the proposed method of piecewise-linear regression. The main idea of this approach is to divide the training and test samples into clusters (partial sub-samples of data that do not overlap) by comparing the predicted values for each i-1 (i = 2, 3, … , N) step of the method with the mean value of the training sample of the i-1 data cluster. Each cluster is fed to a separate SGTM neural-like structure where the calculation takes place. The number of clusters is doubled with each step of the method. The division takes place before obtaining the high

accuracy (weighted value for all clusters) or to the set stopping criteria. It should be noted that data vectors in clusters are not repeated.

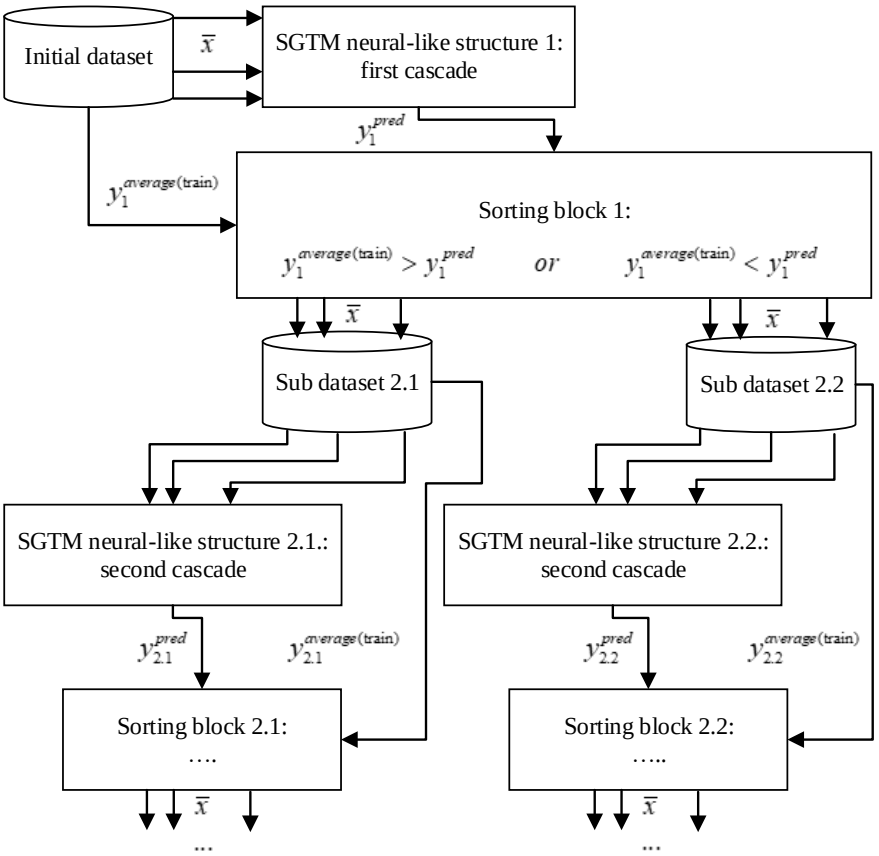


Fig. 3. Flowchart of the proposed method.

The flowchart in Figure 3 was built for two division steps (initial status and one division), but for a large amount of data processing, it is necessary to use more division steps.

The advantages of the proposed method are:

- the high-speed of the training procedure;
- the improved prediction accuracy;
- the ability to the efficiently process large amounts of data;
- the possibility of applying the proposed approach both for regression and for classification tasks.

5 Modelling and results

For the implementation of experimental studies, a dataset was adapted to the kind that enables the use of machine learning methods by separating incoming independent variables into several additional (binary system). In particular, the Smoker column was divided into two - female-smoker and male-smokers. Similar procedures were carried out with other independent variables. As a result, we received the final data set that was used for modeling [23].

The MAPE and MAE [23] metrics were used to assess the quality of the result. Modelling was carried out on the console software that was developed by the authors on a laptop with the following characteristics: the Core i5-6200U CPU, 2.40 GHz.

Parameters of SGTM neural-like structure are as follows: 11 input variables, 11 neurons in the hidden layer, 1 output. Two steps of the algorithm were used, which resulted in the receipt of two data clusters. The results of the training and test procedures of the SGTM neural-like structure for both data clusters are given in Table 1. Weighted results are obtained by the ratio of the sum of the errors products on the respective test subsample dimensionality for both clusters to the overall dimension of the respective sample. These results are also presented in Table 1.

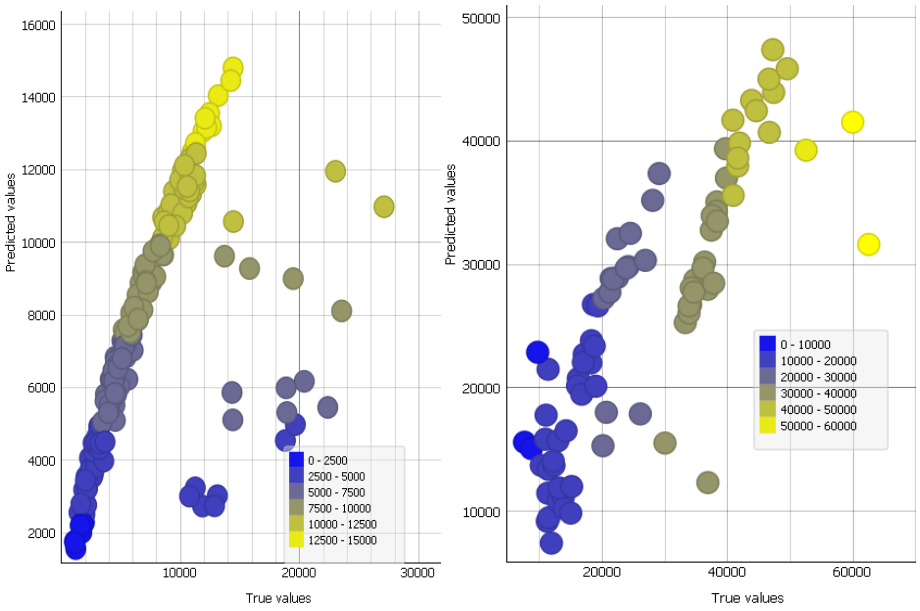
Table 2. Results for two clusters.

Cluster	**Training sample's dimensionality**	**Test sample's dimensionality**	**MAPE, %**	**MAE**
		Training mode		
First cluster	714	185	32.6357786	2462.195359
Second cluster	356	83	26.4181306	5150.517939
Weighted value	***1070***	***268***	***30.5671032***	***3356.627919***
		Test mode		
First cluster	714	185	32.8850502	2415.148908
Second cluster	356	83	25.15974353	5767.230675
Weighted value	***1070***	***268***	***30.60400373***	***3453.293634***

Fig. 4 shows the scatter plot of predicted results for both data clusters for a visual assessment of the proposed approach.

As can be seen from figure 4, the first cluster show worse prediction results due to nearly double the greater number of observations and large deviations, nearly 20 of them. Nevertheless, the weighted value for both clusters for test and training modes shows good and very close results.

This indicates the adequacy of the proposed model and the possibility of its practical application for solving the task.



a) b)

Fig. 4. Predicted results obtained for: a) first cluster, b) second cluster.

6 Comparison and discussion

Comparison of the work of the developed method occurred with known methods. The multilayer perceptron is chosen as the most similar topology to the SGTM neural-like structure, but iterative type. The SGTM neural-like structure is chosen as the basic comparison tool because it is used to build a committee. In addition, this allows comparing the predictive efficiency of a whole sample and its individual clusters (proposed approach). The simulation results for all methods are presented in Table 2.

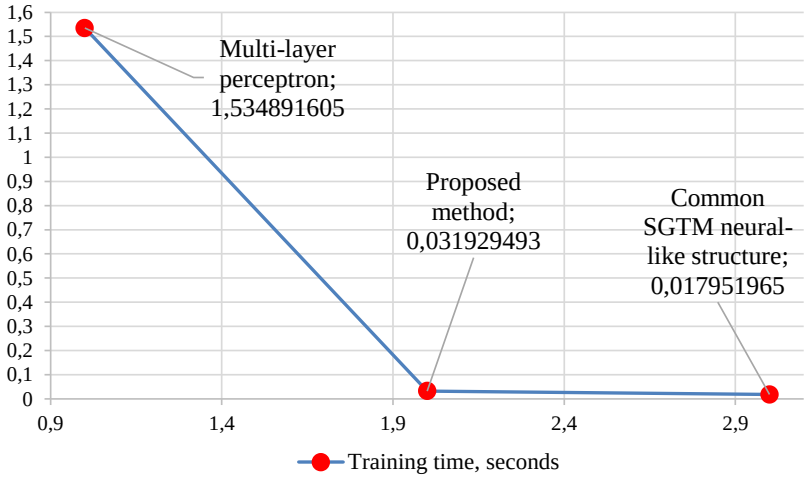
Table 2. Comparison of the predicted results.

Method	**MAPE, %**	**MAE**
Multi-layer perceptron	53.34111109	4751.013931
Common SGTM neural-like structure	41.91968982	4232.394402
Proposed method	**30.60400373**	**3453.293634**

As can be seen from Table 2, the best results were obtained using the proposed piecewise-linear approach. The application of several neural-like structures for pro-

cessing each data cluster individually is increased prediction accuracy by more than 11 % compared to the common SGTM neural-like structure.

In addition, the investigation of the training procedure duration for all investigated methods has been carried out. Figure 5 shows the duration of training procedures for three methods. Since data vectors in clusters do not repeat, it is possible to implement distributed parallel computing, which significantly reduces the duration of the training procedure of the developed method. However, the stepwise division into clusters require to add SGTM neural-like structure works in each step of the procedure. As noted above, two steps were taken to simulate the proposed method. That is why the time for the training procedure of the developed method was calculated as the sum of the training procedure duration in the first step with the greatest training time of one of the two neural-like structures of the second step.



.

Fig. 5. Comparison of the training time.

As can be seen from the figure, the multi-layered perceptron showed the greatest work time. Despite the fact that the developed method for the two steps of the algorithm requires the implementation of three training procedures, the multi-layer perceptron works 51 times slower than developed. This is due to the non-iteration training algorithm of the developed method. In addition, as shown in the figure, the training time of the proposed method does not significantly differ with the time of Common SGTM neural-like structure. An additional explanation is that the Common SGTM neural-like structure processed all dataset, while the second step of the developed method involved the processing of only its parts. Therefore, the total time indicator was not very large. However, while processing considerably larger amounts of data, where the number of steps for sample divide to clusters obviously needs to be larger, the time of the training procedure will also increase.

Further research will be conducted in the direction of applying the hybrid variants of the SGTM neural-like structure as well as finding more efficient cluster division procedures to improve the accuracy of the method.

Conclusion

In the article, the new method of insurance medical costs prediction is proposed. The task is solved using the piecewise-linear approach because of the peculiarities of the data sample. In order to effectively implement this approach in terms of improving the accuracy and speed of its operation, it is suggested to use a non-linear SGTM neural-like structure to process each individual data cluster. The high speed of its work that is based on the non-iterative training algorithm, as well as high generalization properties, the developed approach resulted in an improvement of the accuracy by 11% in comparison with the basic method. Moreover, it showed satisfactory time characteristics of the work. This is confirmed by an experimental comparison of the developed method with analogues (multilayer perceptron and common SGTM neural-like structure) according to MAPE and MAE.

Based on the foregoing, it can be argued that the developed approach provides an opportunity to efficiently use it for the large amounts of data processing, in particular for regression and classification tasks in various fields.

References

1. Guo, X., Gandy, W., Coberley, C., Pope, J., Rula, E., Wells, A.: Predicting Health Care Cost Transitions Using a Multidimensional Adaptive Prediction Process. Population Health Management, Vol. 18, No. 4, 9-18, (2015) DOI: 10.1089/pop.2014.0087
2. Melnykova, N., Markiv, O.: Semantic approach to personalization of medical data. In: 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), Lviv, pp. 59-61 (2016)
3. Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., Wang, G.: Algorithmic Prediction of Health-Care Costs, Operayions research, vol. 56, no. 6, 6-18, (2008) DOI: 10.1287/opre.1080.0619
4. Perova, I., Pliss, I.: Deep Hybrid System of Computational Intelligence with Architecture Adaptation for Medical Fuzzy Diagnostics. International Journal of Intelligent Systems and Applications (IJISA), 9(7), 12-21, (2017).
5. Cucciare, M. A., O'Donohue, W.: Predicting future healthcare costs: how well does risk-adjustment work? J Health Organ Manag. 2006;20(2-3):150-62 (2006)
6. Shakhovska, N., Veres, O., Bolubash, Y., Bychkovska-Lipinska, L.: Data space architecture for Big Data managering. In: 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies, Lviv, pp. 184-187 (2015)
7. Chyrun, L., Vysotska, V., Kis, I., Chyrun, L.: Content Analysis Method for Cut Formation of Human Psychological State. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2018, pp. 139-144.
8. Babichev, S., et al.: Development of a technique for the reconstruction and validation of gene network models based on gene expression profiles. EasternEuropean Journal of Enterprise Technologies, Vol 1, No 4 (91), 19-32 (2018) DOI: 10.15587/1729-4061.2018.123634
9. Hu, Zh., Bodyanskiy, Ye., Tyshchenko, O.: A Cascade Deep Neuro-Fuzzy System for High-Dimensional Online Possibilistic Fuzzy Clustering. In: XI-th International Scientific

and Technical Conference "Computer Science and Information Technologies" (CSIT2016), September 6-10, Lviv, Ukraine, pp.119-122 (2016).
10. Bodyanskiy Y., Vynokurova O., Pliss I., Peleshko D.: Hybrid Adaptive Systems of Computational Intelligence and Their On-line Learning for Green IT in Energy Management Tasks. In: Kharchenko V., Kondratenko Y., Kacprzyk J. (eds) Green IT Engineering: Concepts, Models, Complex Systems Architectures. Studies in Systems, Decision and Control, vol 74. Springer, Cham (2017) doi.org/10.1007/978-3-319-44162-7_12
11. Dronyuk, I., Fedevych, O., Poplavska, Z.: The generalized shift operator and non-harmonic signal analysis. In: 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), pp. 89-91. Lviv (2017)
12. Lytvyn, V., Vysotska, V., Peleshchak, I., Rishnyak, I., Peleshchak, R.: Time dependence of the output signal morphology for nonlinear oscillator neuron based on Van der Pol model. International Journal of Intelligent Systems and Applications, 4, 8-17 (2018)
13. Hu, Zh., Bodyanskiy, Ye., Tyshchenko, O.: A Cascade Deep Neuro-Fuzzy System for High-Dimensional Online Possibilistic Fuzzy Clustering. In: XI-th International Scientific and Technical Conference "Computer Science and Information Technologies" (CSIT2016), September 6-10, Lviv, Ukraine, pp.119-122 (2016).
14. Shakhovska, N. B., Bolubash, Y. J., Veres, O. M.: Big data federated repository model," The Experience of Designing and Application of CAD Systems in Microelectronics, Lviv, pp. 382-384 (2015) doi: 10.1109/CADSM.2015.7230882
15. Eriksson, K., Estep, D., Johnson, C.: Piecewise-linear Approximation. In: Applied Mathematics: Body and Soul. Springer, Berlin, Heidelberg, 741-753 (2004)
16. Doroshenko, A.: Piecewise-Linear Approach to Classification Based on Geometrical Transformation Model for Imbalanced Dataset. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2018, pp. 231-235.
17. Tkachenko, R., Doroshenko, A.: Classification of Imbalanced Classes using the Committee of Neural Networks. In: 2018 XIIIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), Lviv, pp. 400-403 (2016)
18. Medical Cost Personal Datasets.: https://www.kaggle.com/mirichoi0218/insurance. last accessed 10/20/2018.
19. Demsar, J., et al.: Orange: Data Mining Toolbox in Python. J. Mach. Learn. Res. 14, 2349–2353 (2013)
20. Medical Cost Personal Datasets. Kernels: https: //www.kaggle.com/mirichoi0218/insurance/kernels last accessed 10/20/2018.
21. Tkachenko, R., Izonin, I.: Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations. In: Hu, Z.B., Petoukhov, S., (eds) Advances in Computer Science for Engineering and Education. ICCSEEA2018. Advances in Intelligent Systems and Computing, vol. 754, Springer, Cham, 578-587 (2018)
22. Tsmots, I., Teslyuk, V., Teslyuk, T., Ihnatyev, I.: Basic Components of Neuronetworks with Parallel Vertical Group Data Real-Time Processing. In: Shakhovska N., Stepashko V. (eds) Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, vol 689. Springer, Cham (2018)
23. Tkachenko, R., Izonin, I., Vitynskyi, P., Lotoshynska, N., Pavlyuk, O.: Development of the Non-Iterative Supervised Learning Predictor Based on the Ito Decomposition and SGTM Neural-Like Structure for Managing Medical Insurance Costs. *Data* 3, 46 (2018).