IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Piecewise-Linear Approach to Classification Based on Geometrical Transformation Model for Imbalanced Dataset

Anastasiya Doroshenko
ACS Department
Lviv Polytechnic National University,
Lviv, Ukraine
anastasia.doroshenko@gmail.com

***Abstract*— The article describes the method of cost-sensitive classification for imbalanced dataset based on neural-like structure of successive geometric transformations model using piecewise-linear approach to classification. The proposed method characterized by high learning speed and accuracy of classification.**

Keywords—data mining, classification, imbalanced data, neural-like structure of successive geometric transformations model, NLS SGTM, cost sensitive classification.

I. Introduction

Most of the working companies today have a large amount of accumulated data with information about customers, sales, orders, and more. Such data is a source of hidden knowledge, the ownership of which can provide the company with further growth, profits and customers. These tasks are mainly formulated as Data Mining tasks and one of the most popular among them is the task of classification. These tasks are formulated daily, in such spheres of life as how target marketing, medicine, telecommunication, insurance, chemical industry, bioinformatics and others. Researchers use different methods to solve classification problems. The most effective today are classification by decision tree induction, Bayesian classification, neural networks, support vector machines (SVM), deep learning methods [1-5].

The main requirements for the data mining methods, besides the high accuracy of the classification, are their speed and the ability to process huge amounts of data. One of the methods that well matched to these requirements is neural-like structure of successive geometric transformations model (NLS SGTM) [9,13,14].

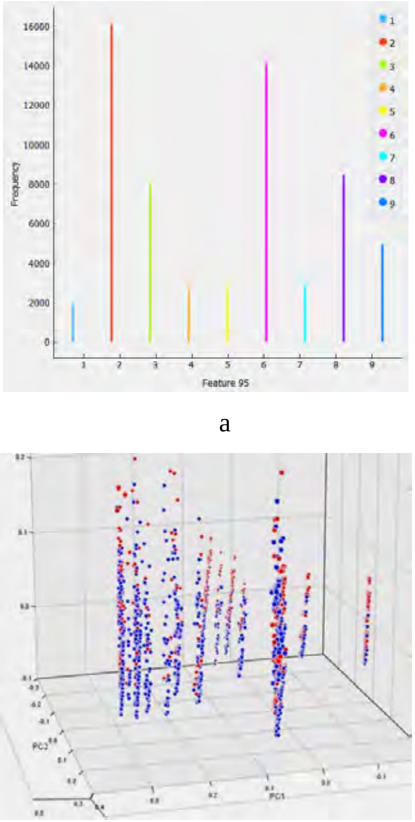
II. Statement of the Probleme

When solving the classification problem, there are often additional restrictions caused by the peculiarities of the subject area or specific task.

In addition, precisely when solving the classification problem for tasks that describe some kind of business or social processes, there are problems such as the imbalanced presentation of data and the different weight of classification errors.

A. Imbalanced representation of data

Imbalanced representation of data is a feature characteristic of such data mining task as classification. The number of instances of one class may differ by an order of magnitude from the number of instances of other classes, which greatly complicates the process of classification. In addition, the density of the distribution of instances of different classes in the space of signs may also be different (Fig.1). According to the hypothesis of compactness, it is assumed that objects belonging to the same class form certain clusters in the space of signs and can easily be separated by hyperplanes of a simple form. However, in many cases, in the systems under conditions of uncertainty, there is a mutual overlap of classes [1, 11-13].



a

b

Fig. 1. Data representation of the data mining task a) the frequency diagram that displays the number of objects in each class (for 9 classes) b) in the coordinates of three main components (for two classes)

978-1-5386-2874-4/18/$31.00 ©2018 IEEE

B. A different weight of errors

Depending on the condition of the task to be solved, each type of error can have its own weight. Often such conditions exist for classification tasks. Accordingly, when solving the classification problem, it is necessary to consider not only the accuracy of the recognition of each of the classes, but also their interdependence - in order to ensure that the weight of the total number of errors was minimal.

In [6] described a comparative analysis of the solution of this problem using well-known methods such as Random Forest Leaner, Logistic Regression, SVM and NLS SGTM. It has been demonstrated that NLS SGTM gives a highest accuracy of classification. This article suggests methods for further improving the accuracy of classification for this problem.

However, article [6] did not take into account the constraints caused by the subject area, namely the fact that various mistakes (TF or FT) are of different weight during the classification. These restrictions are formulated by the customer when setting the task depending on the priorities of their business (for example, depending on what is more important - to identify a potential fraudster or not to lose a potential client). In [6] this cost-sensitive problem was solved by using method of penalties and rewards. This article describes the combination of two methods: method of penalties and rewards and piecewise-linear approach to classification using NLS SGTM.

III. Architecture and Learning Algorithm of Neural-Like Structure of Successive Geometric Transformations Model

A. Training NLS SGTM in supervisor mode

The supervisor mode is special because the components of the data vectors are divided into input and output, where the latter are known only for the elements of the training sample [12-14].

Accordingly to [9], the value of the coefficient $K_i^{(S)}$ required to execute the sequence of geometric transformations can be calculated only for the vectors of the training sample. In this case, it is sufficient to calculate the value of the coefficient $K_i^{(S)}$ based on

$$K_i^{(S)} = \frac{\sum_{j=1}^{n}(X_{i,j}^{(S)}.X_{b,j}^{(S)})}{\sum_{j=1}^{n}(X_{b,j}^{(S)})^2} \tag{1}$$

where $X_{i,j}^{(S)}$ - elements of input vector of features, $X_{b,j}^{(S)}$ - elements of basic vector of features, n - the number of input components of the vector and represent the pseudo coefficient as the approximation of the dependence (2), which is given tabular

$$K_i^{(S)} = f(\overline{K_i^{(S)}}) \tag{2}$$

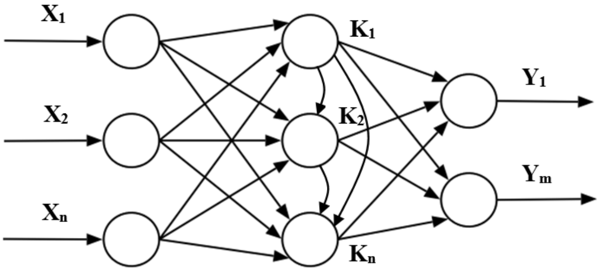


Fig. 2. The topology of NLS SGTM in a supervisor mode.

The topology of a NLS SGTM of this type is presented in Fig. 2, where linear or nonlinear neural elements can be used in the hidden layer, the function of which is the approximation of dependence (2).

For the case of a linear variant of the SNM, the function of activation of the neural elements has the form

$$K_i^{(S)} = \alpha \times \overline{K_i^{(S)}} \tag{3}$$

where α – the coefficient is calculated on the basis of the least squares criterion.

The use of neural-like structure of successive geometric transformations model in the classification mode involves the construction of a separating surface that provides the separation of given classes of objects. The construction of the hyperplane surface in the entire domain of the definition of the input variables has the following drawbacks and limitations:

- the number of vectors-implementations is too large for the simultaneous use of the entire training sample;
- the density of representation of the points of different classes in the realization space is approximately heterogeneous;
- separation of classes by a hyperplane does not take into account the essential nonlinearity of the tasks to be solved.

The use of linear classification methods for data mining is optimal in terms of the speed of solving the tasks, but they do not provide sufficient accuracy of the classification.

To solve this problem, a piecewise-linear approach for constructing separating surfaces based on a model of geometric transformations was developed. It allows for the nonlinearity of data mining tasks to be taken into account, but does not require a large amount of time to execute.

Another advantage of using the piecewise-linear approach of constructing separating surfaces is that by using it, by dividing the total sample into clusters, it is possible to process the entire sample for a reasonable time.

IV. Development of the Piecewise-linear Approach for the Construction of Separating Surfaces Based on Neural-Like Structure of Successive Geometric Transformations Model

Let us consider the use of the piecewise-linear approach for constructing separating surfaces in combination with the method of rewards and penalties for a problem with two classes.

It should be noted, that in order to evaluate the effectiveness of the method two data samples are necessary. The first it is the training dataset, which is used to training of NLS SGTM and for calculating coefficients $K_i^{(S)}$. The second dataset it is test sample whose data was not used during the construction of the model of geometric transformations, but which are necessary to assess the accuracy of the classification.

In general, objective of standard classification is to obtain highest accuracy of classification. However, for the data mining tasks, we have to organize cost-sensitive learning with formulated cost matrix, class probability estimate, misclassification costs and other types of costs involved in the learning process [10].

In order to take into account such a feature of Data Mining as a different weight of errors, we use the method of rewards and penalties. The matrix of rewards and penalties in this case will have the following form (Table I). As a rule, for each specific subject area and task, customer or business analytic forms such a matrix separately.

TABLE I. MATRIX OF REWARDS AND PENALTIES

Matrix of Rewards and Penalties	**Values of Rewards and Penalties**	
	The vector is recognized as class 0	***The vector is recognized as class 1***
The vector belongs to class 0	r_{11}	p_{12}
The vector belongs to class 1	p_{21}	r_{12}

To increase the accuracy of the solution to the classification problem, it is proposed to combine the use of the method of rewards and penalties [5,8] and the tree of division into classes (clustering).

Using the division tree into classes, we can combine data vectors with similar inputs into separate clusters and analyze them independently of each other (Fig.3). After receiving the penalty points for each of the clusters, they are summed up. This approach makes it possible significantly improve the overall accuracy of the classification.

Consider the method of combining the method of rewards and penalties and dichotomy to solve the classification problems in more detail.

The piecewise-linear approach to constructing separating surfaces for solving the problems of classification on the basis of NLS SGTM using the method of rewards and penalties [5]:

1. Initially, for NLS SGTM as a training set we used all training data. The accuracy of classification we estimated as amount of penalty points calculated according to the method of rewards and penalties.

2. After that, both training and test sets, are divided into 2 clusters: vectors recognized by the NLS SGTM as "class 1" and vectors recognized as "class 2 ".

3. After clustering into the training samples obtained for each cluster, we substitute the real values of the outputs and repeat from step 1 all action for both clusters separately.

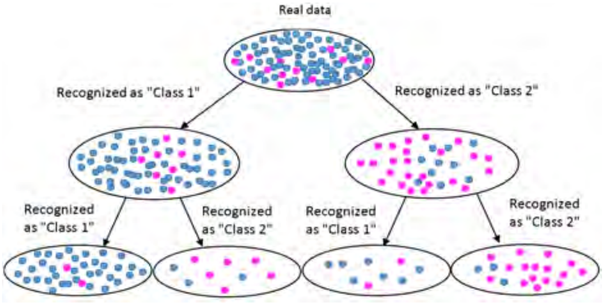


Fig. 3. The division tree into classes based on NLS SGTM

4. Repeat steps 1-3 until the amount of penalty points received for all lower-level clusters is not acceptable under the terms of the task. That is, the accuracy of the classification is sufficient.

5. When training of NLS SGTM for each cluster is finished and accuracy of the classification is sufficient we can use these networks for testing dataset.

So, due to the breakdown of data into clusters, we accelerate the implementation of the classification process, we can take into account a different weight of errors for a specific subject area and increase the accuracy of the classification.

V. EXPERIMENTAL RESULTS

This article describes the solving of classification task, which was formulated in [7]. The training sample describes the transactions carried out by credit card holders within two days and consists of 284,807 lines and 31 columns. For reasons of confidentiality, the dataset contains not original user data, but principal components obtained by the PCA method from the initial data (V1, ..., V28). Only two features: 'Time' (the number of seconds passing through each transaction) and 'Amount' have not been transformed.

Also, the dataset contains one target feature 'Class', which shows the client's affiliation to one of two classes - frauds or ordinary clients. The main feature of the dataset is that the data set is highly unbalanced - only 492 transactions out of 284807 (0.172% of all transactions) have the value of the target field 1, that is, customers are fraudulent.

The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection [7].

According to the subject area a matrix was formed. By analyzing this matrix, it can be seen that a properly classified vector that belongs to the "fraud" class has a much greater weight than a properly classified "ordinary client" vector. At the same time, the case where an ordinary customer is classified as a fraud has the highest number of penalty points (Table II) [8].

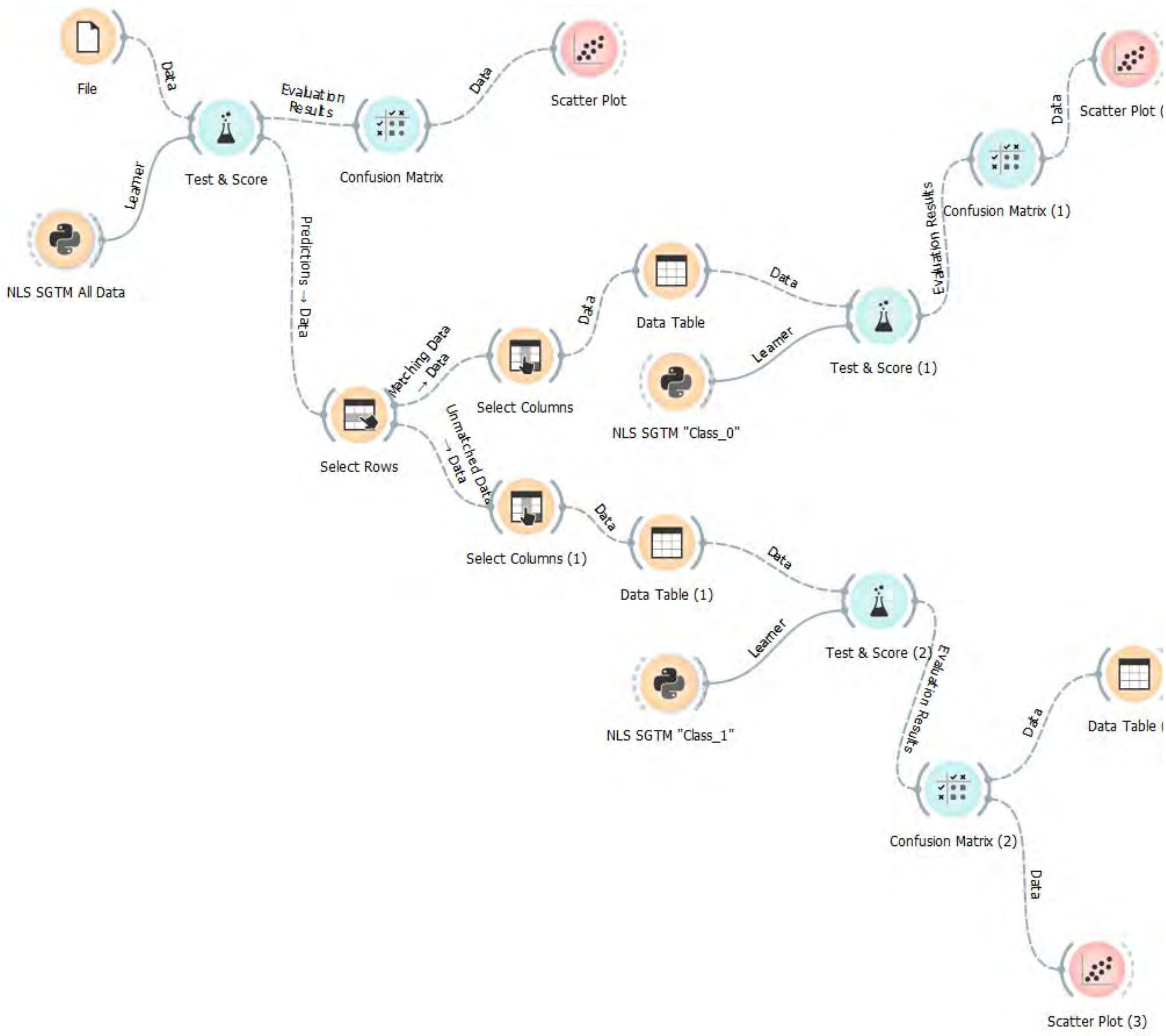


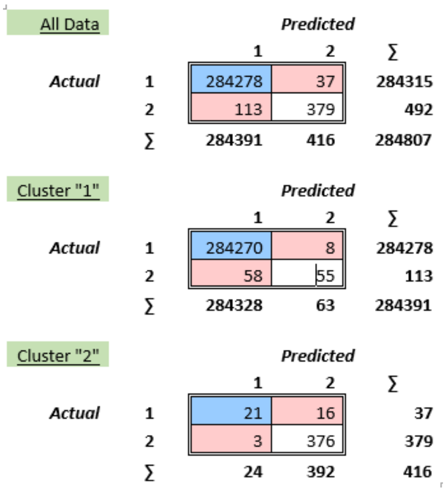
Fig. 4. The structure of the workflow in Orange

TABLE II. MATRIX OF REWARDS AND PENALTIES FOR SOLVING TASK

Matrix of Rewards and Penalties	**Values of Rewards and Penalties**	
	The vector is recognized as class 0	***The vector is recognized as class 1***
The vector belongs to class 0	1	-3
The vector belongs to class 1	-2	5

The structure of the workflow in Orange is shown in Fig.4.

Analyzing the results of the classification (Fig.5), we can see that despite the fact that for the sample as a whole the accuracy of the classification was very high in consequence of the application of the piecewise-linear approach accuracy further increased.



All Data

		Predicted 1	Predicted 2	∑
Actual	1	284278	37	284315
	2	113	379	492
	∑	284391	416	284807

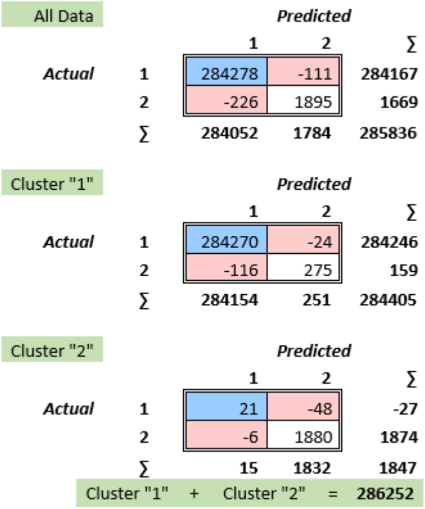
Cluster "1"

		Predicted 1	Predicted 2	∑
Actual	1	284270	8	284278
	2	58	55	113
	∑	284328	63	284391

Cluster "2"

		Predicted 1	Predicted 2	∑
Actual	1	21	16	37
	2	3	376	379
	∑	24	392	416

Fig. 5. Results of classification by clusters (in vectors)



All Data		Predicted		
		1	2	∑
Actual	1	284278	-111	284167
	2	-226	1895	1669
	∑	284052	1784	285836

Cluster "1"		Predicted		
		1	2	∑
Actual	1	284270	-24	284246
	2	-116	275	159
	∑	284154	251	284405

Cluster "2"		Predicted		
		1	2	∑
Actual	1	21	-48	-27
	2	-6	1880	1874
	∑	15	1832	1847

Cluster "1" + Cluster "2" = 286252

Fig. 6. Results of classification by clusters according to the matrix of rewards and penalties (in points)

Fig. 6 shows that the sum of the points obtained for the correct classification by clustering is greater than for the initial sample. It is also worth noting that the initial data was very unbalanced and the number of instances of the lower class was only 0.17% of the total number of training vectors.

VI. Conclutions

The piecewise-linear approach to classification based on geometrical transformation model for imbalanced dataset is developed. This method realizes cost-sensitive classification for such tasks of data mining as fraud detection, home loan or credit applications, medical diagnostic, marketing research where some classes are rare but with big impact. The proposed method characterized by high learning speed and accuracy of classification.

References

[1] C. C. Aggarwal and P. S. Yu, "A Survey of Uncertain Data Algorithms and Applications," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, pp. 609-623, May 2009.

[2] A. J. Chamatkar and P. K. Butey, "Implementation of Different Data Mining Algorithms with Neural Network," 2015 International Conference on Computing Communication Control and Automation, Pune, pp. 374-378, 2015. doi: 10.1109/ICCUBEA.2015.78

[3] D. Zhu, H. Jin, Y. Yang, D. Wu and W. Chen, "DeepFlow: Deep learning-based malware detection by mining Android application for abnormal usage of sensitive data," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, pp. 438-443, 2017. doi: 10.1109/ISCC.2017.8024568

[4] Ye. Bodyanskiy, I. Perova, O. Vynokurova, and I. Izonin “Adaptive Wavelet Diagnostic Neuro-Fuzzy System for Biomedical Tasks,” 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 299-303, February 20 – 24, 2018.

[5] O. Riznik, I. Yurchak, E. Vdovenko and A. Korchagina, "Model of stegosystem images on the basis of pseudonoise codes," *VIth International Conference on Perspective Technologies and Methods in MEMS Design*, Lviv, pp. 51-52, 2010.

[6] R. Tkachenko, . Doroshenko, I. Izonin, Y. Tsymbal, and B. Havrysh, “Imbalance Data Classification via Neural-like Structures of Geometric Transformations Model: Local and Global Approaches,” In: Hu, Z. B., Petoukhov, S., (eds) Advances in Computer Science for Engineering and Education. ICCSEEA2018. Advances in Intelligent Systems and Computing. Springer, Cham, vol.754, pp.112-122, 2018. https://doi.org/10.1007/978-3-319-91008-6_12

[7] A. D. Pozzolo, O. Caelen, R. A. Johnson and G. Bontempi, "Calibrating Probability with Undersampling for Unbalanced Classification," *IEEE Symposium Series on Computational Intelligence*, Cape Town, pp. 159-166, 2015. doi: 10.1109/SSCI.2015.33

[8] J. Wang, P. Zhao and S. C. H. Hoi, "Cost-Sensitive Online Classification," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 10, pp. 2425-2438, Oct. 2014.

[9] R. Tkachenko and I. Izonin “Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations”, In: Hu, Z.B., Petoukhov, S., Advances in Computer Science for Engineering and Education. ICCSEEA2018. Advances in Intelligent Systems and Computing. Springer, Cham (2018).

[10] S. Ghosh, A. Ray, D. Yadav and B. M. Karan, "A Genetic Algorithm Based Clustering Approach for Piecewise Linearization of Nonlinear Functions," 2011 International Conference on Devices and Communications, Mesra, pp. 1-4, 2011.

[11] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009. doi: 10.1109/TKDE.2008.239

[12] R. Tkachenko, H. Cutucu, I. Izonin, A. Doroshenko, and Yu. Tsymbal ‘Non-iterative Neural-like Predictor for Solar Energy in Libya,” In: Ermolayev, V., Suárez-Figueroa, M. C., Ławrynowicz, A., Palma, R., Yakovyna, V., Mayr, H. C., Nikitchenko, M., and Spivakovsky, A. (Eds.): ICT in Education, Research and Industrial Applications. Proc. 14-th Int. Conf. ICTERI 2018. Volume I: Main Conference. Kyiv, Ukraine, May 14-17, pp.35-45, 2018, CEUR-WS.org

[13] U. Polishchuk, P. Tkachenko, R. Tkachenko and I. Yurchak, "Features of the auto-associative neurolike structures of the geometrical transformation machine (GTM)," 2009 5th International Conference on Perspective Technologies and Methods in MEMS Design, Zakarpattya, Ukrane, pp. 66-67, 2009.

[14] R. Tkachenko, I. Yurchak and U. Polishchuk, "Neurolike networks on the basis of Geometrical Transformation Machine," 2008 International Conference on Perspective Technologies and Methods in MEMS Design, Polyana, Ukrane, pp. 77-80, 2008. doi: 10.1109/MEMSTECH.2008.4558743