Application of global optimization methods to increase the accuracy of classification in the data mining tasks

Doroshenko A. V.[0000-0002-7214-5108]

Lviv Polytechnic National University, Lviv, Ukraine
anastasia.doroshenko@gmail.com

Abstract. The article describes the solving of data mining task using neural-like structures of Successive Geometric Transformations Model (NLS SGTM). The main problems of this task are imbalanced dataset and different weigh of errors. Therefore, to take into account these features, the method of penalties and rewards was used, as well as a piecewise linear approach to classification. The supplement of the methods used by the final optimization procedure is proposed. The procedure of final optimization using simulated annealing.

Keywords: data mining, classification, imbalance problem, cost-sensitive learning, imbalanced data, principal components, neural-like structure of successive geometric transformations model, NLS SGTM, simulated annealing, analysis of the principal components, optimization methods.

1. Introduction

In the previous articles [1-2] the methods based on the combination of a Successive Geometric Transformations Model with the method of penalties and rewards was described. In addition, was developed the piecewise-linear approach to constructing separating surfaces in classification tasks.

The purpose of these methods is to solve the data mining tasks of the classification. The main features of these tasks are large-size datasets, imbalanced dataset and different weight of errors. The main goal of this research is increasing the accuracy of classification and minimize the number of penalty points.

In order to increase the accuracy of classification, we propose to supplement the developed methods with final optimization procedures, in particular by the method of random correction of decomposition elements and the method of simulated annealing.

2. Problem statement

Today the data mining tasks are widespread because most companies today have a huge amount of accumulated data with information about sales, customers, orders, and more. This information is a source of hidden knowledge. In turn, the possession of this knowledge allows this company to take a leading position in the market, to win

the competitive struggle. Among such tasks, one of the most popular is the task of classification. These tasks are formulated daily; in such spheres of life how as commerce, telecommunication, and chemical industry, target marketing, insurance, medicine, bioinformatics, and others. Researchers use different methods to solve classification problems [1, 2, 3, 4].

The main features of classification tasks in data mining are imbalanced data, different weight of error, huge amounts of data. These features require using some special additional methods to well-known methods of classification to provide high accuracy of classification.

Hereby as basic methods of classification, we used the neural-like structure of successive geometric transformations model (NLS SGTM) [5, 6]. As an additional method was used piecewise-linear approach [1] and cost-sensitive learning method [7, 8]. This allowed us to improve the classification accuracy and take into account the specifics of a specific task. This article proposes to apply global optimization methods to the neuro-like structure already trained as a result of previous experiments. This will allow us to find such parameters of the neural-like structure, in which the sum of points reaches the global max.

3. Increasing the accuracy of classification using random correction of decomposition elements.

3.1. Analysis of the principal components

The analysis of the principal components is the standard method used to reduce the dimensionality of data in statistical pattern recognition system and signal processing systems. However, it is also advisable to use the analysis of the principal components to solve the problems of data mining because of their high dimensionality [9, 10].

The main task of statistical recognition is the allocation of attributes - the process in which the data space is transformed into space of attributes, which theoretically has the same dimension as the input space. Conversions, however, are usually performed so that a reduced number of the most effective features can represent the data space. Consequently, only a substantial part of the information contained in the data remains, the dimension of the data is reduced. If this approach is applied to data mining task, we will reduce the size of the input data by extracting non-informative features without losing significant data. Consider a more detailed analysis of the principal components (in the theory of information is known as Karhunen-Loeve Transform) [11, 12].

Assume that there exists a vector X of dimension m, which we want to convey with the help of l numbers, where l <m. If we simply cut the vector x, this will cause the mean square error to be equal to the sum of the dispersions of the elements carved out of the vector x. It is necessary to find some linear transformation T, for which the value of the mean square error for the reduction of the vector X will be optimal. In this case, the transformation T must have the property of a small dispersion for its individual components. The analysis of the principal components maximizes the rate

of dispersion reduction and, accordingly, the probability of the correct choice [13, 14].

Let X be an m-dimensional random vector from the initial set of data. The mean value of this vector equal zero.

$$E(X)=0, \tag{1}$$

where E is the operator of statistical expectation. If X have a nonzero average, then you can calculate this value before the analysis begins. Let q be a unit vector with dimension m, the vector X is projected on vector q. This projection is defined as the product of the vectors X and q:

$$A=X^Tq=q^TX \tag{2}$$

with restriction

$$\|q\|=(q^Tq)^{0.5} = 1. \tag{3}$$

Projection A is a random variable that has an average value and variance that is related to random vector statistics X. Dispersion A equals

$$\sigma^2=E[A^2]=E[(q^TX)\,(X^Tq)]= q^TE[(XX^T]q = q^TRq \tag{4}$$

The matrix R of dimension m×m is the matrix of the correlation of the random vector X defined as the expectation of the product of the random vector X on itself:

$$R=E[XX^T] \tag{5}$$

The matrix R is symmetric, then:

$$R^T=R. \tag{6}$$

It follows from (6) that if a and b are arbitrary vectors with dimension $m\times 1$ then:

$$a^TRb=b^TRa \tag{7}$$

From equation (4) it follows that the projection A is a function of an odd vector q. So, we have:

$$\psi(q) =\sigma^2= q^TRq, \tag{8}$$

$\psi(q)$ – dispersion probe.

Directly the principal components are defined as follows. Let the data vector x be a realization of the random vector X. Since there are m possible values of the single vector q, then we must consider m possible projections of the data vector x. By formula (2)

$$a_j = q_j^Tx=x^Tq_j,\ j = 1,2,\ldots,m \tag{9}$$

where a_j - the projections of the vector x on the main guides, represented by single vectors q_j. These projections are called principal components, their number corresponds to the dimension of the data vector x. In this case formula (9) can be considered as the analysis procedure.

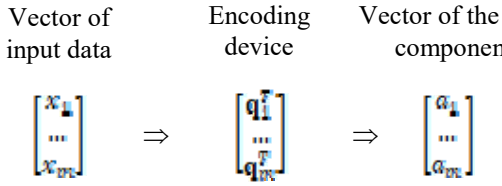

Fig. 1. Scheme of the procedure for analysis of the principal components. The encoding step

Allocation of the principal components (PC) is also proceeding on the outputs of the hidden layer of the neural-like structure of GTM. After selecting the PC, we create and teach an additional neural-like structure (Fig. 2), where the inputs are the vectors of the PC, and the outputs are the significance of the corresponding outputs of the initial training data [15, 16, 17].

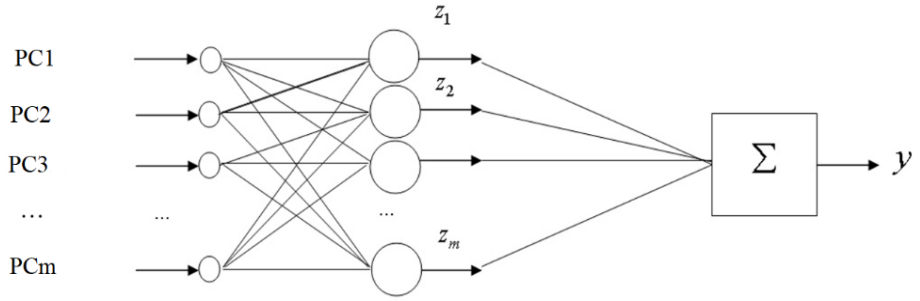

Fig. 2. The structure of neural network for obtaining decomposition elements

The feature of this neural network is the representation of the output of the synapse as $y = \sum_{i=1}^{m} z_i$ well as weak correlation z_i. In the mode of application on the inputs of the trained neural network, input vectors are:

$$0\,0\ldots0 \to y_0$$
$$1\,0\ldots0 \to y_1$$
$$0\,1\ldots0 \to y_2$$
$$\ldots\ldots\ldots\ldots$$
$$0\,0\ldots1 \to y_m$$

Define $\alpha_0 = y_0, \alpha_i = y_i - y_0, i = 1,2,...,m.$

Get the elements of decomposition $z_0 = y_0, z_i = PC_i\alpha_i, i = 1,2,...,m.$

Initially we have

$$y = \sum_{i=0}^{m} z_i \text{ or } y = \sum_{i=0}^{m} z_i k_i , \quad (10)$$

where z_i – the elements of decomposition, k_i =1, i = 1,2,…, m .

For classification tasks, appropriate indicators are optimized (percentage of properly classified specimens, number of penalty points, average arithmetic or mean square error) by random correction of elements of decomposition of coefficients k_i. Random correction of coefficients k_i can be performed simultaneously for all components or for each component independently. It should be noted that component-based random correction is appropriate for prediction since it has been experimentally confirmed that the components are practically independent [18, 19].

Consider a more detailed algorithm for random correction.

3.2. Increasing the accuracy of the classification tasks based on correction of decomposition elements by random correction

Let us consider the case of optimizing the accuracy of the classification problem solving for a case where the recognized data belong to one of the two classes.

Re-coding the input file: outputs of the elements belonging to class 1, assign 1, the outputs of the elements belonging to class 2, assign -1.

The problem of recognition is solved as a prediction problem by the formula (3), but after obtaining a value, we analyze it by the formula (11):

$$y \to \begin{cases} class1, y \ge 0, \\ class2, y < 0. \end{cases} \tag{11}$$

For the task of recognizing the criteria by which optimization can be carried out, there may be the number of penalty points, the percentage of incorrectly classified representatives of class 1, the percentage of incorrectly classified representatives of class 2.

Then the percentage of incorrectly classified representatives of classes is calculated as follows:

$$ErrC1 = 100\%\ NWC1/NC1,\ ErrC2 = 100\%\ NWC2/NC2 \tag{12}$$

where $NC1$ is the number of representatives in class 1 in the training sample, $NC2$ – number of representatives of class 2 in the training sample, $NWC1$ – number of incorrectly classified representatives of class 1, $NWC2$ – number of incorrectly classified representatives of class 2.

In order to calculate the number of penalty points, it is necessary to determine the fines that count for erroneous recognition. So if $P1$ is a fine charged if the element of class 1 is recognized as an element of class 2 and $P2$ is a fine charged if a class 2 element is recognized as an element of class 1.

Then the value of the penalty function - the total number of received penalty points (PP) is:

$$PP = ErrC1P1 + ErrC2P2. \tag{13}$$

Optimization method for classification tasks based on correction of decomposition elements by random correction:

1. Initial values $k_i = 1$.

2. We calculate the value of the error, which is optimization (PP, $ErrC1$ or $ErrC2$).

3. Using a random number generator with a uniform distribution, we choose the value ΔD from the range $(-D, D)$.

4. Calculate new values $k_i = k_i + \Delta D$.

5. Calculate the value of outputs with new k_i .

6. Convert the resulting value to the designation of the class to which this element belongs.

7. Calculate the new value of the error, which is optimization (PP, $ErrC1$ or $ErrC2$).

8. Compare the value of the calculated error with the pre-calculated value. If the new value is less than the previous one, we will remember it as well as the current coefficients k_i. Go to step 3. Otherwise, go to step 3 without remembering.

9. Continue optimization until the predefined desired optimization value is reached, or until the time t has expired.

After performing the optimization method by randomly correcting the decomposition elements, the resulting coefficients are used for further classification.

4. Application of global optimization methods to increase the accuracy of classification in the tasks of the data mining

4.1. Method of simulated annealing

Annealing method is an algorithmic analogue of the controlled cooling process. It was proposed in 1953 by N. Metropolis and refined by numerous followers. Today it is considered one of the few methods by which one can practically find the global minimum of functions of several variables. Consider a more detailed method of simulated annealing [20, 21].

Algorithm of simulated annealing method:

1. Start the process from the starting point at a given initial temperature $T=T_{max}$.
2. As long as $T>0$, repeat L times the following actions:
   - choose a new solution w' from the vicinity w;
   - calculate the change of target function; $\Delta = E(w')- E(w)$;
   - if $\Delta \leq 0$, take $w = w'$; otherwise, if $\Delta>0$, take $w = w'$ with probability $\exp(-\Delta/T)$ by generating a random number R from the interval (0,1), then comparing it with the value $\exp(-\Delta/T)$; if $\exp(-\Delta/T) > R$, take a new solution $w = w'$; in other case - ignore it.
3. Reduce the temperature ($T \leftarrow rT$) using the reduction coefficient r, selected from the interval (0,1) and return to step 2.
4. After lowering the temperature to zero, apply one of the deterministic methods (Levenberg-Marquardt algorithm, error-return algorithm, fastest-speed algorithm, etc.) to achieve the minimum of the target function.

The concept of "temperature" in this algorithm is quite formal, since the presented optimization model is only a mathematical analogy of the annealing process.

The efficiency of the annealing algorithm has an extremely high impact with the choice of parameters such as the initial temperature T_{max}, the coefficient of reduction of temperature r and the number of cycles L, performed at each temperature level.

The main problem is to determine the threshold level optimal for each annealing simulation process. For some practical tasks, this level may have different meanings, but the overall range remains unchanged. As a rule, the initial temperature is selected so as to ensure the implementation of about 50% of the subsequent random changes in

the solution. Therefore, knowledge of the pre-distribution of such changes makes it possible to estimate the initial temperature approximately.

Numerous computer experiments [22] prove that in the case where the time limit is small, the best results give a single implementation. If simulation can be long lasting, then statistically better results can be achieved thanks to the multiple implementation of the annealing simulation if the value of the coefficient *r* is close to 1.

If we compare genetic algorithms with an annealing algorithm, then, in spite of the significant external difference between the algorithms, they are essentially similar in nature. An annealing algorithm according to [23] can be considered a genetic algorithm with a population consisting of one instance. Consequently, an algorithm for simulating annealing of a metal can be regarded as an algorithm that has only a mutation operation, but not cross-linking.

In addition, if we compare these two algorithms from the applied point of view, then it should be noted that, according to Kohonen's study [7], in the case when the initial solution is sufficiently close to optimal, the annealing algorithm of the metal has significant advantages over the genetic algorithms from a computational point of view.

Since in our study the initial data are pre-processed by methods of fines and incentives with sampling alignment and piecewise linear classification on the basis of the model of geometric transformations, then the initial solution of the problem is sufficiently close to the optimal. Accordingly, in this case, it is more appropriate to choose for optimization of the solution.

Improvement of accuracy of the decision of tasks of intellectual data analysis on the basis of correction of elements of decomposition by algorithm of simulated annealing

Let's consider a more detailed algorithm of simulated annealing of metal in combination with methods of fines and incentives and piecewise linear classification on the basis of a model of geometric transformations.

Fig. 3 depicts the structural scheme of the developed neural network based on the model of geometric transformations, where $x_1, x_2, \ldots, x_n$ primary features of classification objects – input data, PC_1, PC_2, ..., PC_n – the principal components derived from input data, $w_1, w_2, \ldots, w_n$ – weight coefficients, y – an output that indicate on belonging to certain classes.

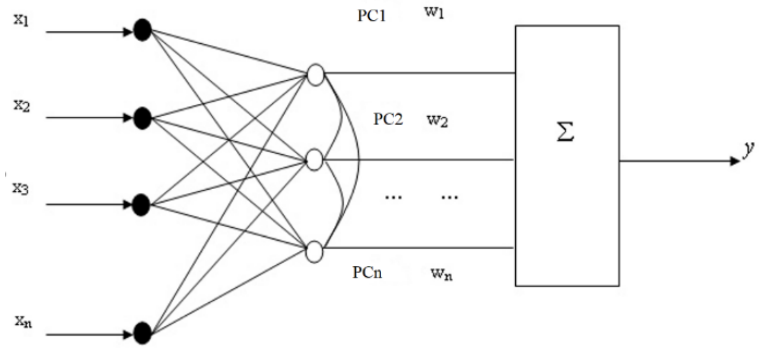

Fig. 3. The scheme of neural-like structure GTM

Functioning of such a neural network can be described by the formula 14.

$$y = \sum_{i=1}^{n} PCw_i \tag{14}$$

The method of simulating annealing of a metal is proposed to be used to optimize weight coefficients so that the resulting amount of penalty points is minimal, that is, the optimization parameter is the amount of penalty points [15, 17].

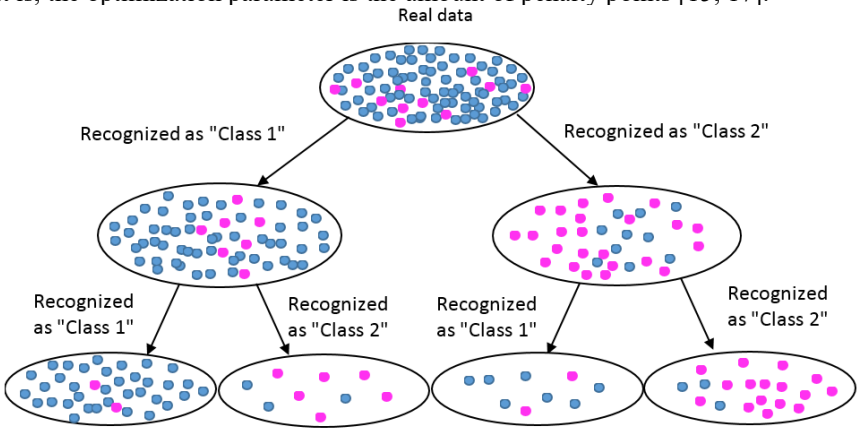

Fig. 4. The division tree into classes based on NLS SGTM

As can be seen from the flowchart describing the solution of the data mining problem by combining the method of fines and incentives, simulated annealing and piecewise linear approach, a modified annealing method for which the target function is the number of penalty points, will be applied separately for each cluster [22, 24].

Accordingly, if we have a two-step division into clusters for a sample of n classes, then we will have division into clusters, which is depicted in Fig. 4, and for each of the clusters a modified annealing algorithm will be implemented.

5. Experimental results

This article describes the solving of classification task, which was formulated in [1]. The training sample describes the transactions carried out by credit card holders within two days and consists of 284,807 lines and 31 columns 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 fraud has the highest number of penalty points (Table 1).

Then we used the modified method of imitation of annealing of the metal with parameters: initial temperature T=Tmax=20895, L=100, R is a random number from the interval (0,1), r=0,9.

Table 1. Matrix of penalties and rewards solving task

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

After lowering the temperature to zero, apply one of the deterministic methods (Levenberg-Marquardt algorithm, error-return algorithm, fastest-speed algorithm, etc.) to achieve the minimum of the target function. The results of classification are described (Fig 5, Fig.6).

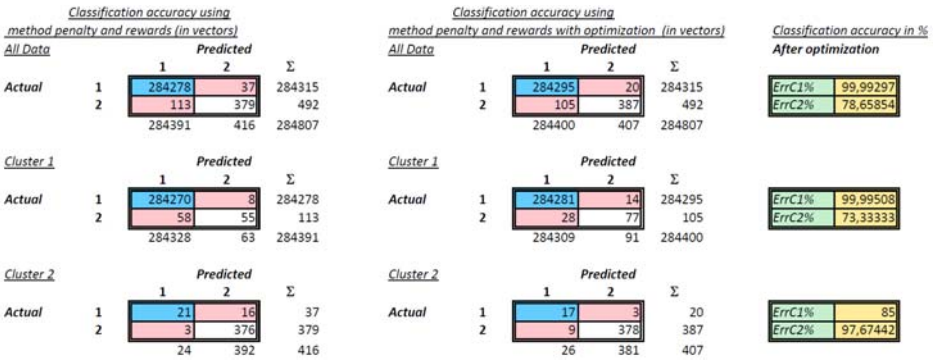

Fig. 5. Results of classification using NLS SGTM with the method of penalties and rewards and the method of simulated annealing (in vectors)

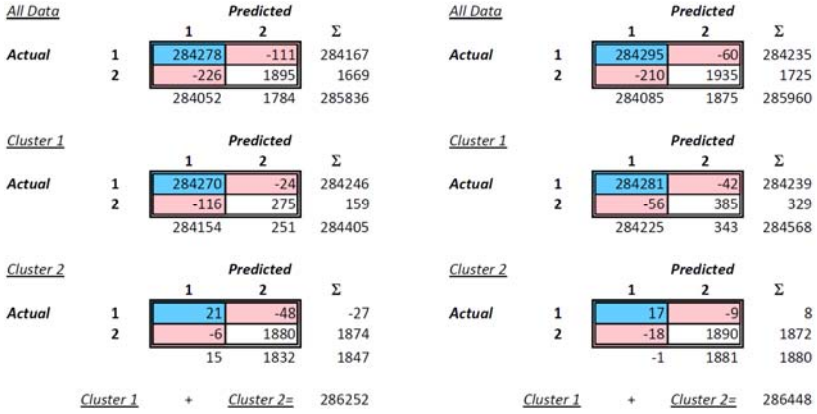

Fig. 6. Results of classification using NLS SGTM with the method of penalties and rewards and the method of simulated annealing (in points)

6. Conclusion

The application of global optimization methods to classification in the data mining tasks allowed to increase the accuracy of classification, especially in combination with other methods of classification, such as neural-like structure of successive geometric transformations model. Also, the method of simulated annealing successfully combined with such methods as piecewise-linear approach and cost-sensitive learning method. The application of the method of simulated annealing made it possible to reach the point of the global maximum and minimize the amount of penalty points for this task.

References

1. Doroshenko, A.: Piecewise-Linear Approach to Classification Based on Geometrical Transformation Model for Imbalanced Dataset. In: Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, pp. 231-235 (2018)
2. Tkachenko, R., Doroshenko, A.: Classification of Imbalanced Classes using the Committee of Neural Networks. In: Proceedings of the XIIIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), Lviv, 11-14 September 2018. (2018)
3. Tepla, T.L., Izonin, I.V., Duriagina, Z.A., Tkachenko, R.O., Trostianchyn, A.M., Lemishka, I.A., Kulyk, V.V., Kovbasyuk, T.M.: Alloys selection based on the supervised learning technique for design of biocompatible medical materials Archives of Materials Science and Engineering, 93/1, pp. 32-40 (2018)
4. Izonin, I., Trostianchyn, A., Duriagina, Z., Tkachenko, R., Tepla, T., Lotoshynska N.: The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production. International Journal of Intelligent Systems and Applications (IJISA), Vol.10, No.9, pp.40-47 (2018).
5. Tkachenko, R., Yurchak, I., Polishchuk, U.: Neurolike networks on the basis of Geometrical Transformation Machine. In: Proceedings of the 2008 International Conference on Perspective Technologies and Methods in MEMS Design, Polyana, pp. 77-80 (2008). doi: 10.1109/MEMSTECH.2008.4558743
6. Polishchuk, U., Tkachenko, P., Tkachenko, R., Yurchak, I.: Features of the autoassociative neuro like structures of the geometrical transformation machine (GTM). In: Proceedings of the 2009 5th International Conference on Perspective Technologies and Methods in MEMS Design, Zakarpattya, pp. 66–67 (2009)
7. Pozzolo, A.D., Caelen, O., Johnson, R.A., Bontempi, G.: Calibrating probability with underdamping for unbalanced classification. In: Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, pp. 159–166 (2015) doi: 10.1109/SSCI.2015.33
8. Liu, X.: A Benefit-Cost Based Method for Cost-Sensitive Decision Trees. In: Proceedings of the WRI Global Congress on Intelligent Systems, Xiamen, pp. 463-467 (2009).
9. Ebied, H. M.: Feature extraction using PCA and Kernel-PCA for face recognition. In: Proceedings of the 2012 8th International Conference on Informatics and Systems (INFOS), Cairo, pp. MM-72-MM-77 (2012)
10. Subbotin, S.: Quasi-Relief Method of Informative Features Selection for Classification. In: Proceedings of the 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, pp. 318-321 (2018).
11. Amar, A., Leshem, A., Gastpar, M.: A greedy approach to the distributed Karhunen-Loève transform. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, pp. 2970-2973 (2010)
12. Mattera, D., Palmieri, F., Haykin, S.: An explicit algorithm for training support vector machines. In: Proceedings of the IEEE Signal Processing Letters, vol. 6, no. 9, pp. 243-245 (1999).
13. Riznyk, O., Yurchak, I., Povshuk, O.: Synthesis of optimal recovery systems in distributed computing using ideal ring bundles. In: Proceedings of the 2016 XII Interna-

tional Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH), Lviv, pp. 220-222 (2016)
14. Li, Xin, Qian, Xu, Wang, Ziqiang: Classification rule mining using feature selection and genetic algorithm. In: Proceedings of the 2009 Asia-Pacific Conference on Computational Intelligence and Industrial Applications, Wuhan, pp. 107-110 (2009)
15. Yue, S., Kaizhang, Liu, W., Wang, Y.: A Fuzzy Clustering Approach Using Reward and Penalty Functions. In: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, pp. 18-21 (2009)
16. Zhiwei, Ye, Juan, Yang, Xu, Zhang, Zhengbing, Hu: Remote Sensing Textual Image Classification based on Ensemble Learning. International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.8, No.12, pp.21-29 (2016).
17. Tkachenko, R., Cutucu, H., Izonin, I., Doroshenko, A., Tsymbal, Y.. (2019) Non-iterative Neural-like Predictor for Solar Energy in Libya. In: Ermolayev V., Suárez-Figueroa M., Yakovyna V., Mayr H., Nikitchenko M., Spivakovsky A. (eds) In: Proceedings of the 14-th International Conference ICTERI 2018. Volume I: Main Conference. Kyiv, Ukraine, May 14-17, pp.35-45 (2018).
18. Oleynik, A., Subbotin, S.: Parametrical synthesis of neural network models based on the evolutionary optimization. In: Proceedings of the 2009 10th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics, Lviv-Polyana, pp. 335-338 (2009).
19. Feiden, D., Tetzlaff, R.: Iterative annealing: a new efficient optimization method for cellular neural networks. In: Proceedings of the 2001 International Conference on Image Processing (Cat. No.01CH37205), Thessaloniki, Greece, 2001, pp. 549-552 vol.1.
20. Granville, V., Krivanek, M., Rasson, J.-P: Simulated annealing: A proof of convergence. IEEE Transactions on Pattern Analysis and Machine Intelligence. 16 (6): 652–656 (1994) doi:10.1109/34.295910.
21. Huang, K., Hsieh, Y.: Very fast simulated annealing for pattern detection and seismic applications. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Vancouver, BC, pp. 499-502 (2011)
22. Shafti, L. S., Pérez, E.: Data Reduction by Genetic Algorithms and Non-Algebraic Feature Construction: A Case Study. In: Proceedings of the 2008 Eighth International Conference on Hybrid Intelligent Systems, Barcelona, pp. 573-578 (2008)
23. Zhengbing, Hu., Bodyanskiy, Y.V., Tyshchenko, O. K., Boiko, O. O.: An Ensemble of Adaptive Neuro-Fuzzy Kohonen Networks for Online Data Stream Fuzzy Clustering. International Journal of Modern Education and Computer Science(IJMECS), Vol.8, No.5, pp.12-18 (2016). doi: 10.5815/ijmecs.2016.05.02