Assessment of Investment Activity in the Regions

Mikhail Leizerovich Krichevsky, Julia Anatolevna Martynova

Saint-Petersburg State University of Aerospace Instrumentation (SUAI), Bolshaya Morskaya str., 67, Saint-Petersburg, 190000, Russia

Abstract **– This article presents the results of using machine learning methods to evaluate the investment activity of various Russian regions. The task was considered from two points of view: obtaining information about the class to which a particular region belongs, and forming a quantitative estimate of the investment activity of the regions. In the first case, the solution was obtained with the help of a neural network system implemented in the *MatLab* 2018*b*. In the second case, a hybrid system ANFIS was used, making it possible to generate a quantitative estimate of investment activity.**

Keywords **– investment activity, machine learning, cluster analysis, classification methods, evaluation of investment activity**

1. Introduction

In today's world, the problem of the innovative development of regions is especially important. Regional economic entities need to spend more and more efforts on carrying out scientific activities and conducting research to maintain their competitiveness and confidently cope with economic instability in the country [1].

DOI: 10.18421/TEM93-02
https://doi.org/10.18421/TEM93-02

Corresponding author: Mikhail Leizerovich Krichevsky, *Saint-Petersburg State University of Aerospace Instrumentation (SUAI), Bolshaya Morskaya str., 67, Saint-Petersburg, 190000, Russia.*
Email: mkrichevsky@mail.ru

Received: 29 January 2020.
Revised: 02 June 2020.
Accepted: 10 June 2020.
Published: 28 August 2020.

There is a need to develop innovative and investment activities in all business entities of the Russian Federation and to create a favorable climate for innovation and investment in order to increase the attractiveness of the regions for Russian and foreign investors.

The study [2] presents data on investment activity in Russia, which are calculated in accordance with the principles of modern statistical methodology. Each Russian region is described by a set of eight parameters. Considering this set of heterogeneous characteristics, it is difficult to draw an objective conclusion in terms of its investment activities. There are two possible directions to evaluate the investment activity of the region:

- formation of a class (rank) to which a specific region can belong;
- construction of a generalized evaluation of investment activity in the region, expressed as a unit value on a pre-selected scale.

The analysis of the first direction involved the use of a neural network approach. Thus, the formed neural network is first trained on the basis of examples, which is the initial data on the investment activity of particular regions. Next, an input feature vector is provided to the trained network, and the network decides whether the region belongs to one of the previously allocated classes.

In the second case, the *ANFIS* type neuro-fuzzy system is used, which allows for the formation of a quantitative estimate of the investment activity of the regions.

Machine learning methods are used to solve these tasks [3], [4], [5].

Next, the work is structured as follows: first, the principles of machine learning and tools that are suitable for solving the task at hand are described; then the results of using the selected methods to evaluate the investment activity of the regions are presented; in the final part, the obtained results are analyzed and the ways of further work are specified.

2. Methodology

Machine Learning

By "machine learning" we shall mean a technology that is defined in [3] as follows: "Optimization of the performance criterion of a model using data and past experience". In machine learning (ML), data plays an essential role, and the learning algorithm is used to discover and study knowledge or properties from data. The quality or quantity of the data set will affect the effectiveness of training and forecasting.

An important issue to be raised in the study of ML is the definition of the "training" term. To develop learning machines, you need to know what "learning" really means and how to determine success or failure. As one of the definitions of this direction, we can specify that ML is a modeling method that includes data [4], [5], [6]. This definition may seem too short, but essentially ML is a method that derives a model from data. Here, data means such information as documents, audio, images, etc. The model is the *final product* of ML.

The very name of ML reflects the fact that the described method analyzes the data and finds the model without human interference. This process is called "learning," because it resembles training with data to search for a model. Therefore, the data that is used in ML are training data.

ML methods are now used in various fields, for example, in risk assessment management, in production, in the development of recommendation systems [7], [8], [9].

Cluster Analysis

Cluster analysis belongs to uncontrolled (non-supervisory) ML methods and is a convenient way to identify homogeneous groups of objects called clusters. Objects (or observations) within a particular cluster have similar characteristics, but differ significantly from objects that do not belong to this group.

The most popular cluster analysis methods include [10], [11]:

- hierarchical methods;
- method k-means.

Each of these procedures differs in the approach to grouping the most similar objects into clusters.

Hierarchical clustering classifies data on different scales by creating a cluster tree or dendrogram.

An important problem when applying cluster analysis is the decision regarding how many clusters should be derived from the data. Sometimes this number is known, but generally the number of clusters is unknown, and therefore you need to seek a compromise solution. The method *k*-means constructs exactly *K* different clusters located at as large distances from each other as possible. Their essence consists in that the classification process begins with the specification of some initial conditions (the number of clusters formed, the threshold for completing the classification process, etc.). Just as hierarchical cluster analysis, iterative methods have the problem of determining the number of clusters. In general, their number may not be known. Not all iterative methods require an initial specification of the number of clusters. But for the final solution of the question related to the structure of the target population, you can try several algorithms, changing either the number of clusters formed or the set proximity threshold for combining objects into clusters. Then it becomes possible to choose the best structuring.

Feature Selection

One of the common tasks of ML is the selection of predictors from a large list of candidates. With a large number of input parameters, learning of neural networks (NN) becomes difficult, and time costs increase.

The term "*curse of dimensionality*" usually refers to difficulties associated with fitting models, evaluating their parameters, or optimizing a multidimensional function with a large sample.

According to the authors, the feature selection methods are better implemented in the *Feature Selection and Variable Screening* module of *Statistica* 13 and are designed to process large sets of continuous and/or categorical predictors in tasks related to regression or classification. Here you can select a subset of predictors from a large list of candidates without assuming a degree of relations between predictors and dependent variables. This module can serve as an ideal preprocessor in ML, allowing you to select shortened sets of predictors for further analysis.

By default, this module for continuous dependent variables calculates the ratio of variances between categories to variance within a category (dependent variable) for intervals of predictor variables. Continuous dependent variables can also be "combined" and "transformed" into categorical variables to analyze the selection of characteristics. This option is especially useful when working with highly distorted continuous dependent variables, or when this variable contains extremely unusual values.

For continuous predictors, the program will divide the range of values in each predictor into *p* intervals (10 default intervals). Continuous variables are transcoded using a special algorithm. Categorical predictors are not transformed in any way.

Neural Networks

A neural network is a "black box" that reflects a situation with a completely unknown process, but includes observations (examples). Here we know inputs and output, but we need a base of examples to be used by the network for learning. Generally, NN is a machine simulating the way the brain processes a specific task [12]. This network is implemented using electronic components or modeled by a program run on a computer. Due to its learning and generalization capabilities, neural networks can be expressed as a mathematical representation of the architecture of the human brain.

Let us explain the principle of neural network technology on a one-layer network (Figure 1.).

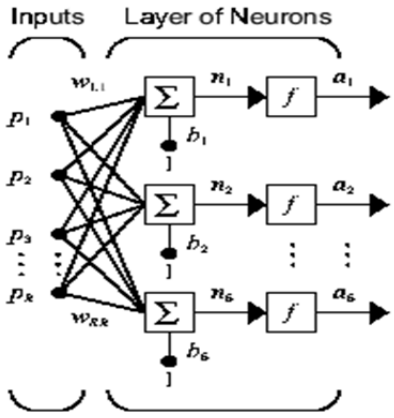

Figure 1. One-layer network [13]

An element with index *i* has an adder that receives weighted inputs and offsets to form the *ith* scalar output *n(i)*. Different outputs are combined to form the *Sth* element of the input vector **n**, which are the arguments of the activation function *f*. As a result, an output column vector **a** is formed at the network output [13].

The principle of operation of the network remains the same except that the outputs of the hidden layer are the inputs of the output one. After collecting data and forming a base of examples, we can easily solve the question of choosing the type of network, since a multilayer feedforward NN is a "*workhorse*" in neural network methods.

Neuro-Fuzzy system

A neuro-fuzzy system (NFS) integrates the principles of neural networks and fuzzy logic. The essence of the NFS is to determine the parameters of fuzzy systems using learning methods adopted in neural networks [14], [15]. The fuzzy logic mechanism can be implemented through such an algorithm as *Mamdani* or *Sugeno*. In this paper, the Sugeno algorithm is used to solve the problem.

3. Results

Table 1. shows a fragment of Rosstat data [2] on the investment activity of Russian regions.

Table 1. Investment performance of the constituent territory of the Russian Federation in 2017 (taken from [2])

	Var1*10^3	Var2	Var3	Var4*10^3	Var5*10^3	Var6	Var7*10^3	Var8
Russian Federation	**472.2**	**56.0**	**34.2**	**140.5**	**37.1**	**68.1**	**108.8**	**104.4**
Central Federal District	**616.4**	**56.2**	**39.5**	**233.9**	**94.1**	**67.0**	**106.3**	**106.6**
Belgorod Oblast	470.8	55.4	32.2	109.6	26.4	72.2	89.7	91.6
Bryansk Oblast	233.7	55.1	28.0	79.8	6.8	63.6	45.0	78.1
Vladimir Oblast	281.4	54.0	28.6	117.3	15.6	68.6	57.5	110.8
Voronezh Oblast	360.4	55.5	33.3	121.5	15.0	73.6	126.0	100.1

Table 1. introduces the following definitions: Var1 - gross regional product per capita, RUB; Var2 - percentage of the working-age population in the total population,%; Var3 - percentage of people with higher education in the number of the employed,%; Var4, Var5 - size of the deposit of individuals in credit institutions of Russia per capita at the beginning of 2017 in ruble and foreign currency accounts, RUB; Var6 - percentage of profitable organizations in the total number of organizations, %; Var7 - capital investment per capita, RUB; Var8 - index of physical volume of capital investments in fixed assets, in % to the previous year. When compiling Table 1., we excluded the rows corresponding to three entities: the Nenets, Khanto-Mansiysk and Yamalo-Nenets autonomous districts, since we used data on the regions which included these districts. Thus, the original data matrix consists of 81 rows and 8 columns. Subsequently, the regions will be called objects, and the indicators will be called predictors.

The results of cluster analysis in the form of a dendrogram for 81 objects, each of which is characterized by 8 predictors, are shown in Figure 2. The horizontal scale of Figure 2. contains the numbers of objects, and the vertical scale specifies the fusion distance of objects.

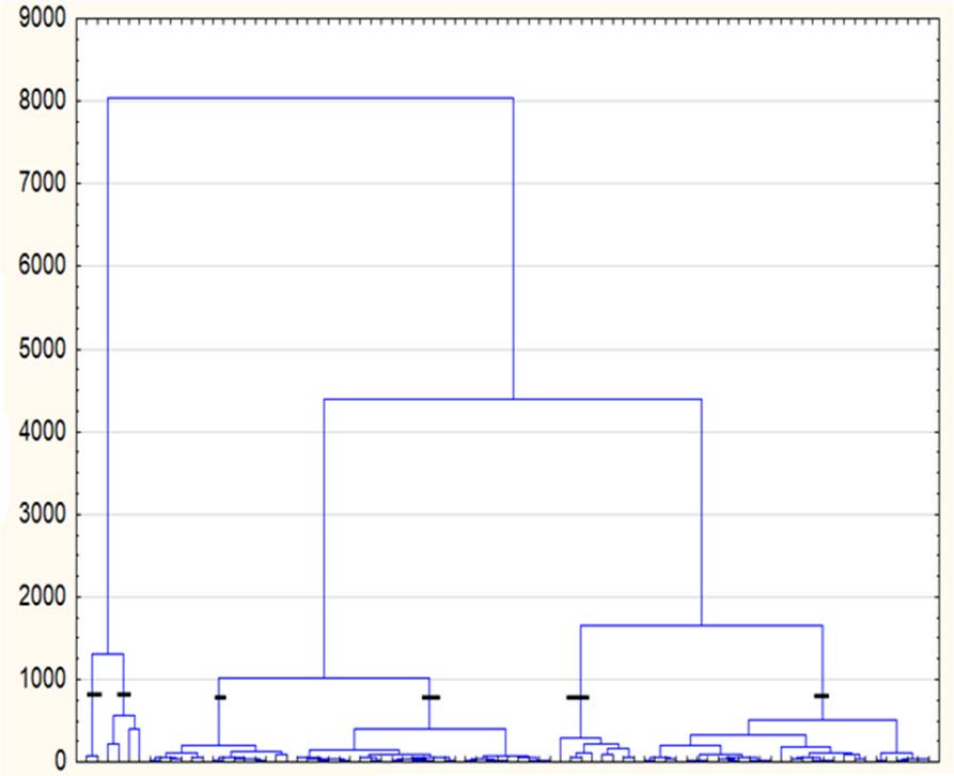

Figure 2. Dendrogram of objects (obtained by the authors)

An analysis of Fig. 2. allows us to identify 6 clusters obtained by cutting the dendrogram at a level slightly below 1,000 units (indicated by semi-bold segments). Let us use the method *k*-means to determine the composition of each of the clusters. As an example, tables 2-4 show the objects that fall into clusters ## 1, 2, and 3.

Table 2. Composition of the first cluster (obtained by the authors)

Case No.	Object numbers and distance to the center
	Distance
C_18	110.05
C_73	102.11
C_78	38.62
C_81	86.53

Table 3. Composition of the second cluster (obtained by the authors)

Case No.	Object numbers and distance to the center
	Distance
C_20	24.23
C_21	26.14
C_24	42.12
C_25	20.28
C_28	65.35
C_46	36.20
C_67	23.27
C_74	28.31

Table 4. Composition of the third cluster (obtained by the authors)

Case No.	Object numbers and distance to the center
	Distance
C_59	13.71
C_79	13.71

Let us add another column to the observation matrix, which will indicate the cluster number. Note that Russia has eight federal districts, while the task at hand includes six clusters.

Next, we will use the *Feature Selection and Variable Screening* module to evaluate the significance of the predictors. The results of this ranking are shown in Figure 3. Let us select 4 predictors for further analysis: *Var1* (gross regional product per capita), *Var7* (fixed capital investment per capita), *Var5* and *Var4* (size of the deposit of individuals in credit institutions of Russia per capita at the beginning of 2017 in ruble and foreign currency accounts).

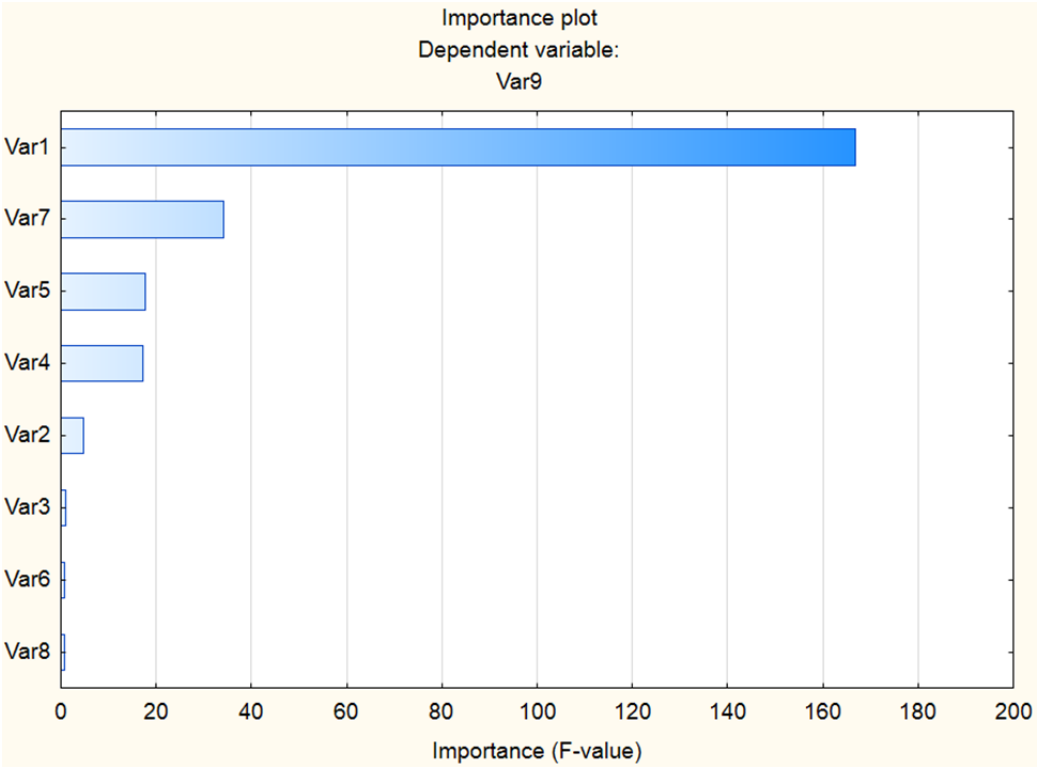

Figure 3. Histogram of the most important predictors (obtained by the authors)

The two-layer neural network created in the *MatLab* program is shown in Figure 4.

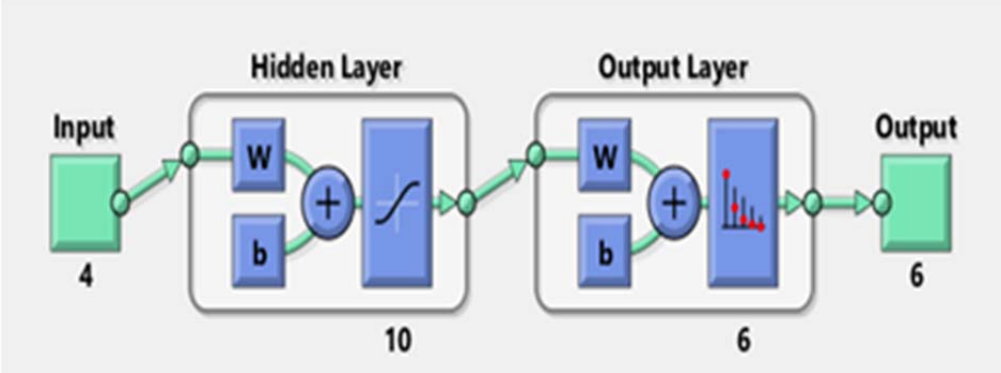

Figure 4. Neural network (obtained by the authors)

The initial data set of 81 objects is divided into 3 parts in the proportion of 70%, 15% and 15%:

- *Training* sample, (57 objects), which is used to train the network;
- *Validation* sample (12 objects), used to evaluate the generalization ability of the network and stop the learning process in case the improvement of generalization ability is stopped;
- *Testing* sample, (12 objects), designed for an independent measure of network characteristics during and after training.

Some of the learning results are presented below. Figure 5. shows three learning curves corresponding to training, validation, and test samples.

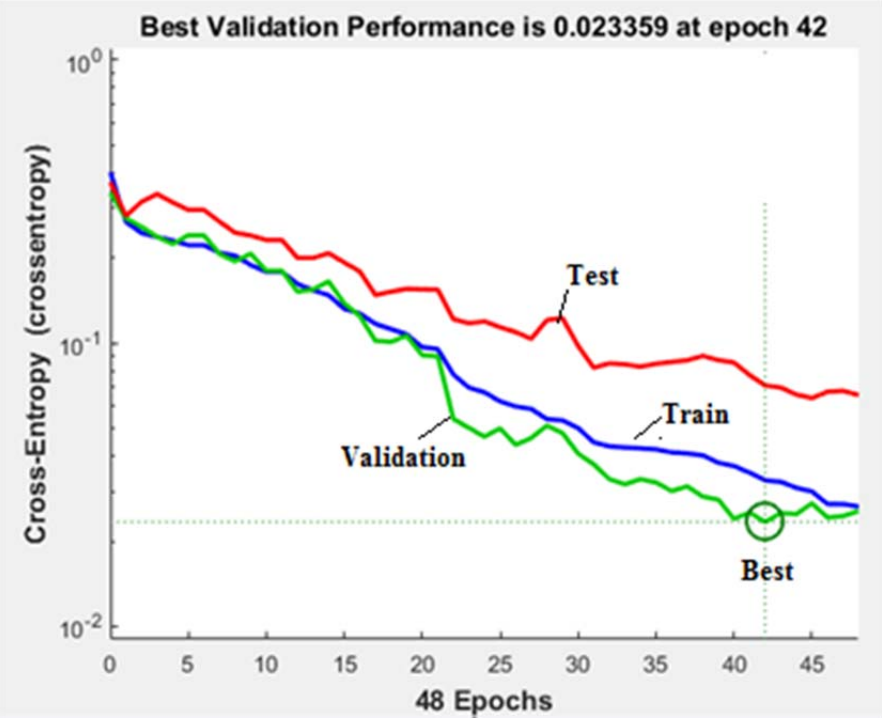

Figure 5. Network learning curves (obtained by the authors)

The vertical axis of the graph in Fig. 5 contains the *cross-entropy* value, which estimates the behavior of the network taking into account the target and output values. Minimizing cross-entropy leads to classifiers of good quality. The horizontal axis indicates the number of training epochs, showing that at the 42th epoch the generalization of the network declined, so the network learned in this period.

To determine whether a region object belongs to a particular class of 6 selected clusters, let us first save the neural network in the *MatLab* workspace. Next, let us apply the parameters of the region object to its input through the command window, for example, VN2=[175;17;102;27]. The network gives a response in the form of a column vector, the largest value of which determines the class:

```
>> fit = net(VN2)
fit =
 0.0000
 0.0000
 0.0000
 0.0001
 0.2930
 0.7069
```

In this case, the network assumes that the object belongs to class 6.

Let us use the *ANFIS* system to solve the second task, which lies in finding a quantitative estimate of investment activity. Let us divide the sample into two parts: training (60 lines) and testing (21 lines) and load them into the system.

The resulting neuro-fuzzy system is shown in Figure 6.

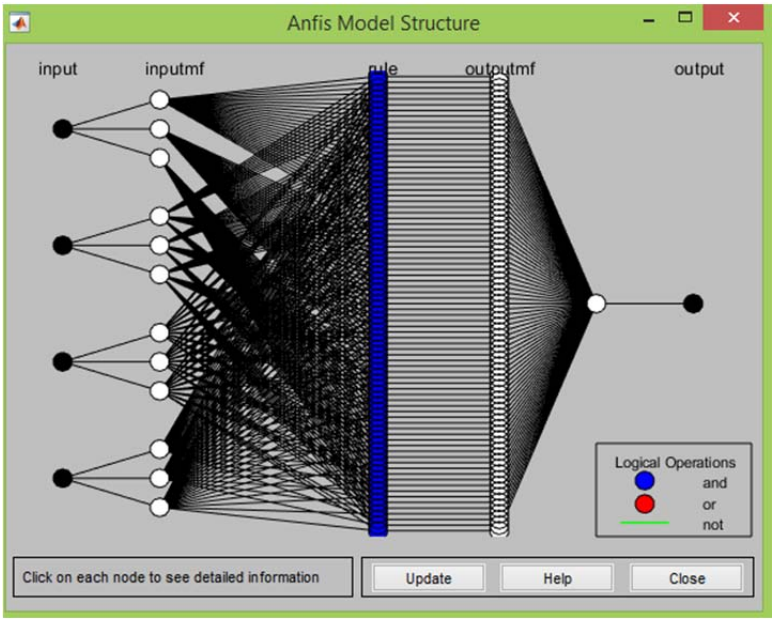

Figure 6. Neuro-fuzzy system (obtained by the authors)

The *ANFIS* system automatically creates a base of rules, which in this case consists of 81 rules. The results of testing this system are shown in Figure 7., which shows that the testing data is close to the results created by the *ANFIS* system.

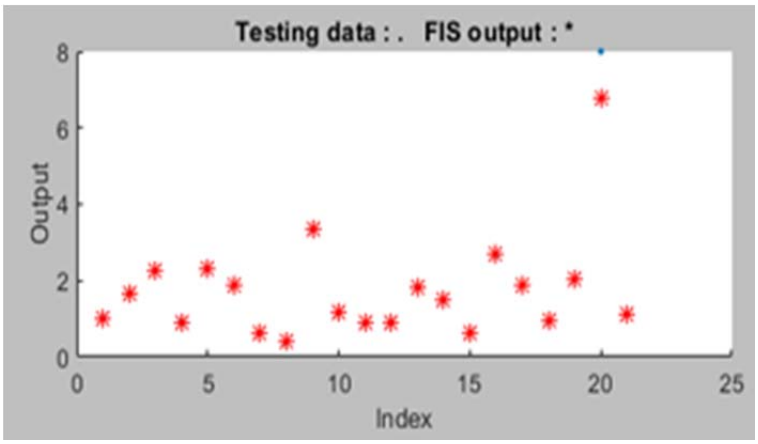

Figure 7. Results of testing (obtained by the authors)

Now, let us consider the fuzzy part of the *ANFIS* system to find a quantitative estimate of investment activity through the option of viewing the rules (here a 10-point scale is selected). Figure 8. shows a fragment of a figure showing such a result: when applying the values of the four predictors indicated above the input columns to the system input, the system gives an estimate of 1.66 points.

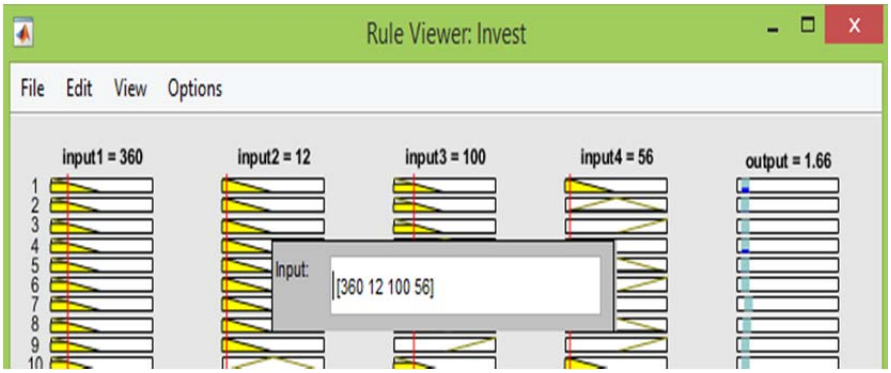

Figure 8. System work (obtained by the authors)

4. Discussion

Our findings show the application of machine learning methods to evaluate the investment activity of various Russian regions. We have considered the solution from two points of view: finding information about the class to which a particular region belongs, and forming a quantitative estimate of the investment activity of the regions. In the first case, we have obtained a solution using a neural network system implemented in the *MatLab 2018b* software product. In the second case, we have used the *ANFIS* neural-fuzzy system, also implemented in *MatLab 2018b*, allowing us to form a quantitative estimate of investment activity. Here, the solution of the problem using only the main components is of interest. Moreover, it is possible to conduct further analysis of the study of the fuzzy inference system using *Simulink* tools.

5. Conclusion

Thus, the article shows that the task of evaluating the investment activity of the regions can be solved using technologies implemented on neural networks and a neuro-fuzzy system. These techniques are part of machine learning, which, in turn, is part of artificial intelligence.

Acknowledgements

The study was carried out with the financial support of the Russian Foundation for Basic Research in the framework of the scientific project No. 18-010-00338A.

References

[1]. Mil'skaya, E.A., &Bychkova, A. V.(2017). Analysis and Evaluation of Innovation and Investment Activities Potential of Economic Entities (For Example, The Northwestern Federal District). *St. Petersburg State Polytechnical University Journal. Economics*, 10(2), 44-53.

[2]. *Investment activity in Russia: conditions, factors, trends.*(2018). Moscow: Federal State Statistics Service. Retrieved from: https://www.gks.ru/bgd/regl/b18_112/Main.htm. [accessed: 15 January 2020].

[3]. Alpaydn, E. (2010). Introduction to machine learning. The Massachusetts Institute of Technology Press.

[4]. Kim, P. (2017). Matlab deep learning. *With Machine Learning, Neural Networks and Artificial Intelligence*, *130*, 21.

[5]. Daumé III, H. (2012). A course in machine learning. *Publisher, ciml. info*, *5*, 69.

[6]. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[7]. Chandrinos, S. K., Sakkas, G., & Lagaros, N. D. (2018). AIRMS: A risk management tool using machine learning. *Expert Systems with Applications*, *105*, 34-48.

[8]. Stanula, P., Ziegenbein, A., & Metternich, J. (2018). Machine learning algorithms in production: A guideline for efficient data source selection. *Procedia CIRP*, *78*, 261-266.

[9]. Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, *97*, 205-227.

[10]. Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

[11]. Everitt, B.S., Landau, S., &Leese, M.(2011). *Cluster Analysis*. New York:Wiley&Sons.

[12]. Haykin, S. (2010). *Neural Networks and Learning Machines, 3/E*. Pearson Education India.

[13]. Hudson, M., Hagan, M. T., & Demuth, H. B. (2014). Neural Network Toolbox, User's Guide, MATLAB, Mathworks.

[14]. Mewada, K. M., Sinhal, A., & Verma, B. (2013). Adaptive neuro-fuzzy inference system (ANFIS) based software evaluation. *International Journal of Computer Science Issues (IJCSI)*, *10*(5), 244-250.

[15]. Dubey, S. K., & Jasra, B. (2017). Reliability assessment of component based software systems using fuzzy and ANFIS techniques. *International Journal of System Assurance Engineering and Management*, *8*(2), 1319-1326.