UDC 519.87

Application of Information Technology for the Analysis of the Rating of University

Oksana N. Romashkova[*], **Yulia V. Gaidamaka**[†‡], **Ludmila A. Ponomareva**[*], **Igor P. Vasilyuk**[*]

[*] *Department of Applied Informatics, State Unitary Enterprise Moscow City Pedagogical University*
29, Sheremetevskaya str. Moscow, 127521, Russian Federation
[†] *Department of Applied Probability and Informatics*
Peoples' Friendship University of Russia (RUDN University)
6 Miklukho-Maklaya st., Moscow, 117198, Russian Federation
[‡] *Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences*
44-2 Vavilova st., Moscow, 119333, Russian Federation

Email: ox-rom@yandex.ru, gaydamaka_yv@rudn.university, ponomarevala@bk.ru, ipvkod@mail.ru

This paper builds a model of predicting the rating of the University on the basis of a neural network in IBM SPSS Statistics. The choice is due to the fact that the program contains gradient descent error function, which is able to automatically configure the network for data classification. The authors describe the modeling technique, a step-by-step algorithm for selecting the architecture of the network, setting its parameters, training and testing.

Experiment data of 1102 Russian universities and 123 indicators of their activity was used for this experiment.

A vector was supplied as an input for the network, the coordinates of which were the average total score of each University. Indicators were considered independent variables. 30 out of 123 indicators were left for the study by the method of correlation analysis. The number of input neurons was equal to the number of independent variables. The output layer contained the amount of neurons equal to the number of dependent variables. The activation function of neurons in the hidden and output layer is sigmoid.

The authors present the results of modeling. Using the constructed model, the input data was divided into clusters: "efficient", "inefficient". Centers of clusters were determined. The sample was split for two network architectures with different number of layers and neurons. The percentage of error on the control and training samples was calculated. Quality of the proposed model was evaluated using ROC (Receiver Operating Characteristic) curve.

Key words and phrases: neural networks, SPSS, multilayer perceptron, modeling, rating of universities.

In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the VIII Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Moscow, Russia, 20-Apr-2018, published at http://ceur-ws.org

1. Introduction

To build the prediction model, a large dataset with various dimensions was used (Table 1). In statistical methods of data processing it does not matter how the objective function's residual is minimized [1, 2], the model will remain unchanged. The question arises of choosing the optimal mathematical-statistical model for estimating the objective function. The authors decided to analyze the indicators of universities using a neural network [3–5].

Table 1

Fragment of experimental data

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19
U1	54,1	54,1	47,4	38,48	131	19,32	1,63	4,6	0	51,93	4,09	99,2	1,24	6 251	2 784	3 467	50,46	44,54	58,33
U2	61,1	61,1	54,97	42,95	21	10,14	4,28	7,43	37,5	53,07	7,43	88,81	0	1 918	879	1 039	58,92	45,83	69,5
U3	65,76	65,53	60,27	48,95	64	3,85	4,84	8,16	13,28	28,09	5,38	94,12	2,99	19 105	11 859	6 932	62,43	62,07	55,19
U4	62,54	61,34	59,47	49,06	68	4,49	4,28	7,02	100	24,64	5,5	99,68	1,24	15 255	7 542	7 713	57,35	49,44	66,14
U5	0	0	48,46	39,3	0	0	0,78	2,4	63,64	60,33	3,92	57,42	0	6 263	2 368	3 893	48,78	37,81	69,68
U6	61,69	61,69	56,02	52,68	0	0	0	0	0	0	0	90,48	0	1 129	407	722	58,5	36,05	76,79
U7	65,82	65,82	57,93	43,3	0	0	0	2,2	0	30	2,48	86,14	0	1 861	1 139	722	59,63	61,2	80,71
U8	64,72	64,72	61,55	63,12	0	0	0	0	0	0	0	100	0	3 045	999	2 046	61,21	32,81	78,19
U9	0	0	44,45	38,3	0	0	0	0	0	0	0	100	0	97	12	85	47,26	12,37	35,05
U10	0	0	47,95	38	0	0	0	0	0	0	0	0	0	801	221	580	48,12	27,59	86,89
U11	54,38	54,38	52,59	39,46	21	3,93	0,17	1,24	0	17,14	1,35	99,62	2,68	3 461	2 161	1 300	53,54	62,44	79,75
U12	59,34	59,34	54,01	46,15	39	4,24	2,22	3,86	100	43,15	3,59	90,46	0	6 991	3 826	3 123	57,3	54,73	67,29
U13	64,42	64,42	60,75	46,81	25	3,39	2,76	3,71	6,82	77,73	2,81	72,6	3,12	7 180	4 330	2 796	62,84	60,31	67,95
U14	0	0	45,8	33,1	0	0	0,23	0,23	0	100	0	100	0	2 201	140	2 047	48,12	6,36	79,78
U15	51,73	51,73	46,63	40,28	37	6,46	1,4	5,29	23,08	56	4,42	93,78	1,49	6 844	2 105	4 739	50,58	30,76	67,26
U16	0	0	0	0	0	0	0	0	0	0	0	100	5,93	429	0	429	48	0	30,54
U17	0	0	48,72	35,97	0	0	0	0	0	0	0	100	0	1 459	617	842	48,72	42,29	63,06

The advantages of neural network modeling include the ability to work with data with different measurement scales and the possibility of approximating any continuous function [6].

The implementation of the model through a neural network can be performed using various programs. The authors selected IBM SPSS Statistics 25 because of their commercial availability.

The object of research is the performance indicator of Russian universities.

The subject of the study is the process of predicting the rating of the university.

The aim of the research is the methodological aspects of constructing a neural network model for predicting the rating of the university using the tools – the IBM SPSS package.

The scientific novelty of the research consists in the development of methods and algorithms for analyzing and predicting the evaluation of the activity of the university with the use of neural networks [7–9].

The work is of practical importance, since it contains a methodology for constructing a model and setting up a multilayer perceptron in the IBM SPSS Statistics [10, 11].

2. Experimental data

The initial data for modeling is presented in Table 1. Objects of the study are 1102 Russian universities. This sample includes all state universities and private higher education institutions head units of the Moscow region. Properties of objects – 123 indicators of the work of universities.

For example:

I.1.1 (Average score of the Unified State Examination of students, accepted according to the results of the Unified State Examination for full-time education according to the bachelor's and specialist programs at the expense of the corresponding budgets of the budget system of the Russian Federation, point);

I.2 (The average score of USE students of the University, taken according to the results of the USE for full-time education under the Bachelor's and Specialist programs at the expense of the corresponding budgets of the budget system of the Russian Federation, with the exception of people who have entered special rights and within the quota of the target admission, score);

I.2.16 (Number of grants received for the reporting year per 100 NDP, units);

10 (Total amount of R & D performed by own forces, thousand rubles);

11 (The total amount of work, services related to scientific, scientific and technical, creative services and development, made by own forces, thousand rubles);

12 (Total number of publications of the organization per 100 NDP, units);

13 (Number of business incubators, units);

14 (Number of technoparks, units);

15 (Number of centers for collective use of scientific equipment, units);

16 (Number of small enterprises, units);

17 (Total number of post-graduate students, people);

18 (The proportion of post-graduate students studying in full-time,%).

Table 1 has the headings: “Name”, “Results of performance evaluation”, “Scorecard”:

- References;
- Name of the educational organization;
- Region;
- Departmental affiliation;
- Website;
- Organization profile;
- Information about the parent educational organization;
- Name of the educational organization;
- Region.

3. Problem statement

Based on these indicators, to predict the value of the target binary variable — whether the work of the university will be effective. Using the IBM SPSS Statistics, build a neural network that divides the input data into clusters and identifies their centers. According to the trained network, determine to which cluster the new input vector will belong.

The input vector (dependent variables) is the average total score collected by the institution. Independent variables (factors) are indicators (“Results of performance evaluation”) that have been coded for ease of presentation in the table in accordance with program requirements, for example:

P.1. – Educational activity;
P.2. – Scientific-research activity;
P.3. – International activity;
P.4. – Financial and economic activity;
P.5. – Salary of the teaching staff;
P.6. – Employment.

4. Theoretical part

For modeling, the multilayer perceptron network architecture was used [12–14]. The choice is due to the presence of the learning algorithm-the occurrence of a local minimum (gradient descent) of the error function. This algorithm allows automatic configuration of the network for data classification [15–17].

Stages of building a network:

- assess the significance of the indicators and determine the range of change in their values;
- prepare data for modeling;
- design the network architecture – determine the number of layers and the number of neurons in each layer;
- training;
- testing.

5. Experimental research

Before the simulation, the data was checked for abnormal emissions in values, duplicates were deleted, etc. [18]. This data went beyond the reasonable bounds of value of the indicators and tested the distribution for the whole sample. Excel was used for finding out whether the outliers are or errors. The frequency of occurrence of each individual experimental value was calculated. Thus, typos, missing and unexpected values were detected.

Since the experimental sample is large, it was difficult to construct a histogram taking any form. Therefore, the nature of data distribution was determined by a graphical method: construction of quantile graphs (Fig. 1).

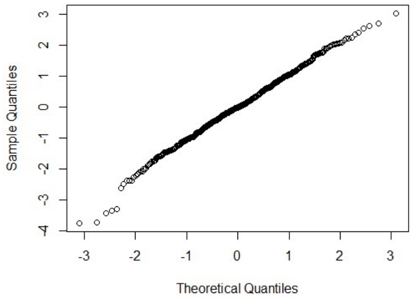

Figure 1. A graph of quantiles for a set consisting of 1102 observations

The graph shows the quantiles of two distributions – empirical (i.e. based on the analyzed data) and theoretically expected standard normal distribution. The quantiles are lined up at an angle of 450. Based on this, the authors concluded that the distribution of the studied data is normal.

More details of this important phase of the analysis are not described in the article.

At the stage of preliminary data preparation, 30 were left for the study in order to reduce the sample size by correlation analysis from 123 indicators.

A hyperbolic tangent or sigmoid function is usually used as an activation function. Activation function is a function that calculates the output signal of an artificial neuron. Sigmoid – is an increasing everywhere differentiable s-shaped nonlinear function with saturation, which allows you to amplify weak signals without saturating with strong signals. The activation function decides on the activation of the neuron and makes it easier to train the network with the method of reverse propagation of the error.

In the preparation of quantitative variables, the domain of definition and the value of the activation function were taken into account. The activation function – the sigmoid has the range of values $(0, 1)$ [19]. In SPSS, normalization was used to bring the data to the interval $(0, 1)$. The value of factors (x) is recalculated in accordance with the formula $[x-(\min-\varepsilon)]/[(\max+\varepsilon)-(\min-\varepsilon)]$, where "min" is the minimum value of the variable for all observations, "max" is the maximum value, ε — correction to reduce the range of values of variables. The domain of the function is the whole numerical axis [20].

The number of neurons of the input layer of the network is equal to the number of independent variables — 30. Each dependent variable is assigned to one output neuron. The number of hidden layers is determined automatically by SPSS. The activation function of the neurons of the hidden and output layer is the sigmoid.

In order to assess the accuracy of the constructed model, part of the sample from training was deleted. Thus, the data was divided into three parts in proportion: 60% —

training, 20% — control and 20% — test. The control sample served to estimate the accuracy, and the test sample demonstrated the operation of the neural network for clustering data. The separation was done randomly by the program. The learning control took place in a mini-packet mode, in which the algorithm for back propagation of the error is a stochastic gradient descent. Rule for stopping network learning: the maximum number of steps without changing the error. The parameters “interval center” and “interval offset”, which set the range of initial values of the weights of the neural network, were taken equal to 0 — the center of the interval, and the offset from 0.5 to 1.5.

6. Results achieved

The number of hidden layers and the number of neurons in these layers was selected automatically by the program, two models with different network architectures were built (Table 2).

Table 2

Neural network models with different architectures

Size of training sample	Network Architecture		Percent of erroneous forecasts	
	Hidden layers	Number of neurons	Teaching	Verification
441 (60%)	1	10	18.5%	17.9%
441 (60%)	2	200	18%	18%

Calculations showed that the number of layers and neurons do not greatly affect the quality of the model. As a result of the study, the sample was divided into two clusters (Table 3). The percentage of errors on the training and control samples is almost the same, which indicates a well-trained network.

Table 3

Results of the classification

	Predicted		Percentage of correct
Sample	1 cluster	2 cluster	
Teaching	56.9%	43.1%	81.7%
Control	56.8%	43.2%	82.1%
Verification	56.2%	43.8%	82.2%

Using the ROC (Receiver Operating Characteristic) curve, you can estimate the quality of the constructed model. The diagonal line in the graph (Fig. 2) is the indicators of the lack of informative model. The more the curve is bent the better the network is trained. It is considered that the coefficient of the area of the curve in the range 0.9–1.0 indicates a very good quality of the model. As a result of constructing the neural network the indicator reached 0.97.

As for the interpretation of the model for the experimental data, the results of partitioning into clusters using a neural network matched with experimental observations.

The first cluster of “effective university” included all public and private institutions of higher education that carried out 4 or more monitoring indicators.

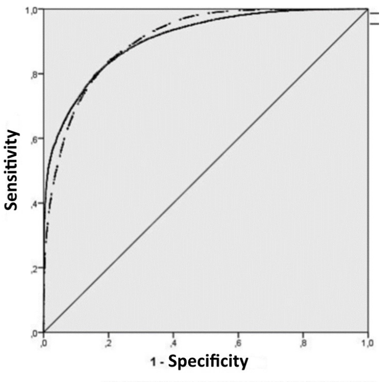

Figure 2. ROC – curve for the constructed model

7. Conclusion

The authors considered the methodology for modeling the rating of universities by the example of building a neural network in the IBM SPSS Statistics. This technique can be an alternative to statistical methods for studying similar experimental data [21–23].

References

1. L. A. Ponomareva, V. L. Kodanev, Development of the module of the corporate information system “Educational environment of the university” on the basis of cloud technologies, In the collection: Informatics: problems, methodology, technologies, the collection of materials of the XVII international scientific and methodical conference, **5** (2017) 393–398, in Russian.
2. O. N. Romashkova, A. I. Morgunov, Information System for the Assessment of the Activity Results of Moscow Secondary Educational Institutions, Bulletin of Peoples Friendship University of Russia, series Informatization of Education, no. 3 (2015) 88–95, in Russian.
3. L. A. Ponomareva, P. E. Golosov, Development of a mathematical model of the educational process in the university for improving the quality of education, Fundamental Research, no. 2 (2017) 77–81, in Russian.
4. O. N. Romashkova, T. N. Ermakova, Education Quality Monitoring in a Comprehensive Secondary Insitution with the Use of Modern IT-based Means Bulletin of Peoples Friendship University of Russia, series Informatization of Education, no. 4 (2014) 10–17, in Russian.
5. Y. Orlov, D. Zenyuk, A. Samuylov, D. Moltchanov, Y. Gaidamaka, K. Samouylov, S. Andreev, O. Romashkova, Time-dependent sir modeling for d2d communications in indoor deployments, Proceedings – 31st European Conference on Modelling and Simulation, ECMS. (2017) 726–731.
6. A. A. Drozdova, A. I. Guseva, Modern Technologies of E-learning and its Evaluation of Efficiency, Procedia – Social and Behavioral Sciences, **237** (2017) 1032–1038.

7. V. S. Kireev, Development of fuzzy cognitive map for optimizing e-learning course, Communications in Computer and Information Science, **706** (2017) 47–56.
8. V. Kireev, A. Silenko, A. Guseva, Cognitive competence of graduates, oriented to work in the knowledge management system in the state corporation “rosatom”, Journal of Physics: Conference Series, **781** (1) (2017) 012060, doi:10.1088/1742-6596/781/1/012060.
9. Y. Attali, M. Arieli-Attali, Gamification in assessment: Do points affect test performance? Computers & Education, **83** (2015) 57–63, doi:10.1016/j.compedu.2014.12.012.
10. L. A. Barnett, Developmental benefits of play for children. Journal of Leisure Research, no. 22 (1990) 138–153, URL: https://www.researchgate.net/publication/232469836_Developmental_benefits_of_play_for_children.
11. M. Blasi, S. C. Hurwitz, S. C. Hurwitz, For Parents Particularly: To Be Successful–Let Them Play!, Childhood Education, **79** (2) (2002) 101–102, doi:10.1080/00094056.2003.10522779.
12. V. A. Potatorum, Informatization of education as a problem of culture, Man and culture, no. 3 (2015) 1–40, in Russian, doi:10.7256/2409-8744.2015.3.15247.
13. A. D. Ursul, T. A. Ursul Education for sustainable development: first results, problems and prospects, Sociodynamics, no. 1 (2015) 11–74, in Russian doi:10.7256/2409-7144.2015.1.14001.
14. D. B. Elkonin, Game and mental development, Almanac of the Institute of correctional pedagogics of RAO, no. 28 (2017) 32–66, in Russian.
15. T. E. Gololobova, S. V. Cheskidov, E. N. Pavlicheva, Topical issues of automation of activity of educational Department of the University on the example of IMIAN GAOU IN Moscow state pedagogical University, Information resources of Russia, no. 2 (2017) 24–28, in Russian, URL: https://elibrary.ru/item.asp?id=21970410.
16. E. I. Prokhorov, L. A. Ponomareva, E. A. Permyakov, M. I. Kumskov, Fuzzy classification and fast rejection rules in the structure-property problem, Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications) **23** (1) (2013) 130–138, URL: https://elibrary.ru/item.asp?id=20517066.
17. O. N. Romashkova, L. A. Ponomareva, Model of educational process in high school using Petri nets, Modern information technologies and it education **13** (2) (2017) 131–139, in Russian, doi:10.25559/SITITO.2017.2.244.
18. L. A. Ponomareva, K. R. Litvinova, V. I. Gorelov, Comparative analysis of the Russian rating systems of the University assessment, in the collection: Methods, mechanisms and factors of international competitiveness of national economic systems collection of articles of the International scientific and practical conference: in 2 parts (2017) 55–58, in Russian, URL: https://elibrary.ru/item.asp?id=30378977.
19. O. N. Romashkova, L. A. Ponomareva, Model of effective management of the United educational system (structure), New information technologies in scientific researches materials of the XXI all-Russian scientific and technical conference of students, young scientists and specialists. Ryazan state radio engineering University (2017) 16–18, in Russian, URL: https://elibrary.ru/item.asp?id=30521101.
20. L. A. Ponomareva, V. L. Kodanev, S. V. Cheskidov, Model of management of process of development of competences in educational organizations, New information technologies in scientific research materials of the XXII all-Russian scientific-technical conference of students, young scientists and specialists. Ryazan state radio engineering University (2017) 20–22, in Russian, URL: https://elibrary.ru/item.asp?id=30521104.
21. L. A. Ponomareva, O. N. Romashkova, I. Vasilyuk, Conceptual model of changing the rating assessment of the University, in the collection: Methods, mechanisms and factors of international competitiveness of national economic systems. Collection of articles of the international scientific-practical conference: in 2 parts (2017) 75–77, in Russian, URL: https://elibrary.ru/item.asp?id=30378981.
22. L. A. Ponomareva, P. E. Golosov, A. B. Mosyagin, V. I. Gorelov, Method of effective

management of competence development processes in educational environments, Modern science: actual problems of theory and practice. Series: Natural and technical Sciences, no. 9 (2017) 48–53, in Russian, URL: https://elibrary.ru/item.asp?id=30281545.

23. L. A. Ponomareva, G. M. Kochergina, E. N. Perelygina, The use of information and communication technologies in the study of banking in College, in the collection: Theoretical and applied issues of science and education. Collection of scientific works on the materials of the International scientific-practical conference: in 16 parts. (2015) 104–107, in Russian, doi:10.17117/na.2015.02.083.