Algorithms for big data mining of hub patent transactions based on decision trees

Aleksandr Zhukov[1], *Sergey* Pronichkin[1,2,3], *Yuri* Mihaylov[4], and *Igor* Kartsan[5*]

[1]Expert and Analytical Center, 33, Talalikhina str., Moscow, 109316, Russia
[2]Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 40, Vavilov Street, Moscow, 119333, Russia
[3]Central Economics and Mathematics Institute of Russian Academy of Sciences, 47, Nakhimovsky Prospekt, Moscow, 117418, Russia
[4]National University of Science & Technology MISiS, 4, Leninsky Prospect, Moscow, 119049, Russia
[5]Marine Hydrophysical Institute, Russian Academy of Sciences, 2, Kapitanskaya str., Sevastopol, 299011, Russia

Abstract. Dysfunctions of the patent supply and demand market have a negative impact on the sustainability of the national innovation system, which stimulates the growth of prices for knowledge-intensive products. It is necessary to establish a relationship between fiscal decisions regarding patent transactions and the prospects for the development of commercialization of the results of intellectual activity. One of the most promising methods for improving the accuracy of system analysis of big and semi-structured patent transaction data is the use of decision trees. Existing methods based on the error backpropagation method are quite slow, and their accelerated versions lose in training accuracy. To effectively solve the problem of forecasting the cost of hub patent transactions, parametric algorithms have been developed based on response bias and with the additional use of predicative structures of the model of successive geometric transformations. The optimal number of decision tree predicates has been established taking into account computational efforts and the accuracy of forecasting the cost of hub patent transactions. Based on evolutionary computing, the optimal values of the parameters of algorithms for big data mining of hub patent transactions have been established.

1 Introduction

The market of supply and demand for scientific knowledge is striking in its volume and growth rate. Unlike fundamental research, the applied research market is characterized by a high level of innovation and technology. Machine learning methods are increasingly used to solve decision-making problems [1, 2]. Such tasks include optimization and forecasting of patent costs [3]. The use of intelligent patent pricing strategies can significantly increase the profitability of a knowledge-intensive enterprise. In the context of digitalization with an

* Corresponding author: kartsan2003@mail.ru

increase in demand for complex IT solutions, the problem of forecasting the cost of patents has become relevant, especially when it comes to hub patents [4, 5]. Most often, the need to forecast the cost of a patent arises among large innovative companies specializing in the sale of licenses for hub patents.

Based on data on sales offers, it is possible to build various models for predicting the cost of hub patents, where the cost is affected by many independent factors [6, 7]. Often, such big data are characterized by many input features, depending on the type of patent: utility model, invention or industrial design. To effectively solve the forecasting problem, it is necessary to develop, apply and evaluate the results of the functioning of decision trees based on the response bias and with the additional use of predictive structures of the model of successive geometric transformations. To achieve this goal, it is necessary to develop a topology of decision trees, apply the developed complex for solving the regression problem, and compare the results of experimental studies with theoretical estimates.

2 Materials and methods

Royalty from the use of intellectual property is one of the main sources of income in the knowledge economy, primarily due to the significant business activity of innovators. Theoretical analysis of big data and determination of practical economic consequences for innovative enterprises from royalty payments for the non-exclusive use of intellectual property is important for the sustainable development of the digital economy. Otherwise, the establishment of royalties for innovative enterprises that do not correlate with their financial results can lead to crisis phenomena:

- a decrease in demand for high-tech products on the domestic market;
- a decrease in the financial stability of innovative enterprises;
- an increase in social tensions at innovative enterprises due to the forced reduction in the number of highly qualified employees;
- a decrease in the volume of patent transactions.

Dysfunctions of the market for supply and demand of scientific knowledge negatively affect the financial stability and ability to perform the main economic functions of innovators, which stimulates the growth of prices for science-intensive products and, accordingly, reduces the profitability of consumers of the results of intellectual activity.

One of the effective instruments for reducing the dysfunctions of the market for supply and demand of scientific knowledge is tax regulation of exclusive rights to the results of intellectual activity. Tax policy is an effective instrument for ensuring sustainable economic growth and a factor in balancing the interests of the state and innovative enterprises. It is an important component of innovation policy and is implemented through the established system of relations between the state and taxpayers in order to ensure that the state fulfills the functions assigned to it and allows it to use its resources and instruments of influence to achieve the goals of sustainable development of the knowledge economy. The implementation of the fundamental goal of tax policy is ensured through the tax mechanism - a set of techniques, methods, organization of tax relations covering all stages of the tax process and influencing the innovative development of enterprises.

Currently, the measures taken by regulatory authorities to ensure the collection of taxes on income in the form of an exclusive right are ineffective. The search for opportunities to develop a knowledge economy should be carried out at the macro level with a distinction between the macrofinancial and macroeconomic spheres. In countries with a developed knowledge-based economy, a distinction is made between the paid and unpaid tax burden, that is, tax debts and debts to target funds. In this case, the ratio of tax revenues to the gross national product is used, including the cost of goods and services produced by residents of the country outside its borders.

For countries with a developed economy, a redistribution of the tax burden towards indirect taxes is typical. For developing countries, the main share of the tax debt is formed by payments to the state budget, but in recent years there has been a clear tendency towards an increase in the share of tax debt on payments to local budgets. The greatest problematic situations arise at the stage of interaction of public administration in the field of taxation with the management systems of economic entities. This is especially evident when implementing procedures for monitoring the payment of tax debt and its collection within the framework of patent transactions (inconsistency of state administration, conflict of interests, contradictions in the coordination of amounts of tax liabilities, the emergence of various types of risks, etc.).

Within the framework of patent transactions, each classification of a hub patent is equal to the distribution of financial indicators by fiscal policy areas and can be supplemented by additional analyses of performance indicators. It should be noted that at the macro level, assigning individual indicators to a particular classification group is a labor-intensive task. For an innovative enterprise to stably fulfill its own obligations at the macro level, the necessary conditions are maintaining the basis for the formation of public finance funds and the relative constancy of expenditure on research and development.

Reducing tax rates or even establishing a zero tax rate on patent transactions does not solve the general problem of their double taxation. Double taxation involves taxing one tax object of an individual taxpayer with the same (similar) tax for the same period of time, or taxing the same tax base by different actors with comparable taxes. It is necessary to establish a relationship between the management (fiscal) decisions of the state regarding patent transactions and the prospects for the development of commercialization of the results of intellectual activity within related sectors of the knowledge-based economy. For effective fiscal regulation of patent transactions, the following conditions must be met:

- consider the theoretical foundations and economic consequences for economic entities in the case of double taxation of patent transactions, as well as measures to eliminate them;
- identify significant differences in the formation of income from hub patent transactions, which should affect the mechanism of their taxation;
- analyze the dynamics of the volume of rent payments for the non-exclusive use of the results of intellectual activity paid by innovative enterprises, predict the cost of patents, economic losses for knowledge-intensive industries from increasing taxes;
- propose possible ways to solve the problem of double taxation of patent transactions while ensuring a balance of interests of knowledge-intensive industries of the digital economy and state interests.

When simultaneously levying a tax on patent transactions of transnational companies for the non-exclusive use of the results of intellectual activity, double legal taxation results, when the same taxpayer is taxed with comparable (similar) taxes in relation to the same taxable object two or more times in one period. Moreover, double taxation is not a mathematical doubling of the tax amount, but an excessive increase in the tax burden on taxpayers (transnational innovative enterprises).

Double taxation can only be eliminated by legislative regulation. The grounds for avoiding double taxation must be laid down through the system of taxation principles in the current legislation. Taxation of patent transactions violates the principles of: one-time payment; equality (introduces tax discrimination against transnational innovative enterprises) and economic justification (establishes the tax in a way that reduces the competitiveness of the taxpayer).

In addition to the theoretical groundlessness of introducing a tax on patent transactions of transnational corporations simultaneously with rent payments for the use of the results of intellectual activity, there are practical negative economic consequences for science-

intensive sectors of the knowledge-based economy. If executive authorities, when implementing measures to decentralize power, seek to increase financial support for the budget, they must act in a different way than introducing double taxation of hub patents through the introduction of an additional tax for innovative sectors.

The volumes of tax payments that are currently collected from patent users are understated and the lion's share of the rent is appropriated by transnational corporations and is not redistributed, as in advanced countries, in the interests of society. A more significant increase in the rates of rent payments for the non-exclusive use of the results of intellectual activity as a component of taxes, which already occurs annually, automatically leads to an increase in prices for science-intensive products of innovative enterprises, and further along the chain, without actually affecting the cost structure of transnational corporations.

The following may be ways to solve the problem of increasing the receipt of rent payments to the budget and maintaining the financial stability of knowledge-intensive sectors of the knowledge economy:

- exemption from taxation of patent transactions with a simultaneous increase in the share of deductions from rent payments for the non-exclusive use of the results of intellectual activity obtained within the framework of the implementation of state or municipal contracts;
- establish by law categories of hub patents that should not be taxed because they cannot, due to regulatory restrictions, generate the expected income from their use.

Maintaining the financial stability of innovative enterprises should be based on scientifically sound tools for predicting the cost of patent transactions. To predict the cost of patents, many studies use machine learning tools, in particular artificial neural networks [8]. Artificial neural networks have proven themselves to be successful in decision support tasks [9]. They automatically generate highly accurate models that can analyze big and weakly structured data, but despite a large number of successful applications, these methods have a number of shortcomings [10, 11].

Training of neural networks is reduced to solving a multi-criterial and multi-parameter problem of minimizing operational errors. The developed training algorithms based on the backpropagation method are quite slow, and their accelerated versions lose in training accuracy. In addition to traditional algorithms, optimization methods based on genetic algorithms or simulated annealing algorithms are used [12, 13]. Given the specifics of the task of analyzing big patent transaction data, the use of artificial neural networks for them is ineffective in the sense of too great a complexity of training and debugging.

Linear neural networks of the single-layer perceptron type and neural-like structures based on the model of successive geometric transformations [14, 15] are quite attractive for use in the problems of analyzing big patent transaction data, since they do not require a complex and lengthy procedure for setting parameters, and training is fast and effective. Consequently, for the class of linear problems, these types of neural networks can be recommended for use in the conditions of using modern computer tools. With respect to problems that are represented by nonlinear response surfaces, the accuracy of their solution by such tools in most cases turns out to be unsatisfactory.

Radial basis function neural networks perform approximation by a combination of hyperspheres, unlike neural networks that approximate the response surface by a combination of sigmoid surfaces [16, 17]. Radial basis function neural networks have obvious advantages over sigmoid neural networks, in particular, a significantly simpler structure setup, faster and more efficient training. The disadvantages of radial neural networks include the inability to perform extrapolation due to the large volumes of networks caused by the need to use a significant number of radial function centers in the tasks of analyzing big data from hub patent transactions.

A significant advantage of probabilistic neural networks is the absence of debugging and training procedures [18, 19]. However, the accuracy of such neural networks is quite low, so they are recommended to be used only for preliminary data analysis. Probabilistic neural networks do not have good extrapolation capabilities [20-22].

3 Results

One of the most promising and modern methods for improving forecasting accuracy is the use of decision trees. The topic of decision trees has been well studied in machine learning. The mathematical justification for the idea of decision trees is the probabilistic assessment of the correct classification in accordance with the probabilities of correct decisions of individual tree nodes. One of the key conditions for obtaining good results in the use of decision trees is the use of methods as nodes whose error probabilities do not exceed half and which are uncorrelated with each other. If each of the tree nodes is able to predict the result with an error probability of less than half and there is no correlation between the results of applying the methods, then the final result will be better than using each of the methods separately.

The main idea of the developed algorithms for big data mining of hub patent transactions based on decision trees is to linearize the response surface defined by the data of the available training sample. For this purpose, the surface obtained using the approximation method (with quite noticeable errors) is fed to the input of the decision tree. An even higher level of linearization, and therefore greater accuracy of modeling, is achieved by feeding the outputs of several predicates to the inputs of the decision tree, for which the vectors of the input signals are randomly shifted relative to the moving point.

Let a sample for preparing a decision tree (the first sample) be given from L_1 vectors and a sample for applying the decision tree (the second sample) from L_2 vectors. It is important to note that for the second sample the value of the original component is unknown.

The algorithmic implementation of the proposed scientific and methodological approach involves the following steps:

1. The components of the hub patent transaction vectors are normalized, resulting in vectors for the first sample of the form z_{lk}, where $l = 1, \dots, L_1$; $k = 1, \dots, K$.

2. Based on the results of calculating responses for all transactions of the first sample, the optimal values of the Gaussian function range coefficient are selected using the selected deviation criterion. Unlike existing approaches, in this case a training stage is introduced.

3. A set of deviations is randomly generated, which will be of the same type for each l-th vector of the first sample and the g-th vector of the second sample δ_{pk}, where $p = 1, \dots, P$; $k = 1, \dots, K$. For each pair of p and k, the value δ_{pk} is a random number in a given range. The range is the same for the first and second samples of hub patent transactions. It is necessary that the deviation range be less than one. The value P is selected and specifies the number of inputs of the decision tree expansion.

4. Based on each l-th vector of the first sample of hub patent transactions, $P + 1$ additional input vectors z_{lk}^{p} are formed, where $z_{lk}^{p} = z_{lk} + \delta_{pk}$, $l = 1, \dots, L_1$; $k = 1, \dots, K$; $p = 1, \dots, P$.

5. For each additional vector, a forecast of prices for hub patents is obtained:

$$f_{lp} = \frac{\sum_{l=1}^{L_1 - 1} f_l e^{-\frac{H_{pl}^2}{\sigma^2}}}{\sum_{l=1}^{L_1} e^{-\frac{H_{pl}^2}{\sigma^2}}} \quad (1)$$

where H_{pl} is the Hamming distance between vectors z_{lk} and z_{lk}^{p} for $p = 1, \dots, P$.

6. The extended l-th vector of the first sample z_{lk} is formed, $f_{lp} \to f_l$; the l-th vector is removed from the formulas for calculating f_{lp} from the first sample. For the $(P+1)$-th vector, the price forecast for hub patents is obtained:

$$f_{lP+1} = \frac{\sum_{l=1}^{L_1-1} f_l e^{-\frac{E_{P+1l}^2}{\sigma^2}}}{\sum_{l=1}^{L_1} e^{-\frac{E_{P+1l}^2}{\sigma^2}}} \quad (2)$$

where H_{pl} is the Euclidean distance between vectors z_{lk} and z_{lk}^{P+1}.

7. The expansion of the decision tree is performed after the formation of predicates on the vectors of the first sample L_1.

8. Based on the sliding vector from the second sample, P additional input vectors are formed using the sets of deviations introduced for the second sample z_{gk}^p, where $z_{gk}^p = z_{gk} + \delta_{pk}$, $g = 1, \dots, L_2$; $k = 1, \dots, K$; $p = 1, \dots, P$.

9. For each additional vector of the second sample, the predicted value of prices for hub patents is found:

$$f_{gp} = \frac{\sum_{g=1}^{L_2-1} f_g e^{-\frac{D_{pg}^2}{\sigma^2}}}{\sum_{g=1}^{L_2} e^{-\frac{D_{pg}^2}{\sigma^2}}} \quad (3)$$

where D_{pg} is the Hausdorff distance between the vectors z_{gk} and z_{gk}^p for $p = 1, \dots, P$.

The developed algorithms were used to solve the problem of forecasting prices for hub patents using a sample of transaction data. Price forecasting was based on three independent groups of features corresponding to the cost, comparative and income methods of patent evaluation.

The data sample contained 1500 vectors, of which 1000 vectors were used in the decision tree preparation mode, and 500 vectors, respectively, in the application mode. A number of experiments were conducted to select the optimal parameters for the algorithms. The first of them involves searching for such a number of decision tree predicates that would provide an optimal result. The experiment was conducted with a change in the number of decision tree predicates from 10 to 50. The assessment was carried out using the root-mean-square error. The smallest error in the application mode of the developed algorithms was obtained for the case when the number of decision tree predicates was 30. In addition, in this case, the minimum deviation was obtained between the error values in all versions of the hub patent transaction analysis algorithms. This indicates that the proposed model has achieved a certain optimal complexity. With a further increase in the number of decision tree predicates, a significant increase in the error of the forecasting mode is observed. Moreover, such an error in the forecasting mode is significantly greater than in the training mode. This may indicate overfitting of the proposed decision tree when selecting a large number of its predicates.

The following experiment involved finding the best parameters for the algorithms:

- the magnitude of the Gaussian function range;
- deviation values on the basis of which decision tree predicates are formed.

For this purpose, optimization based on genetic algorithms was applied. The minimum error value was obtained for the minimum value of deviations from which the decision tree predicates are formed.

As experiments have shown, the accuracy of the developed algorithms is significantly higher than for individual predicates. Certain differences between the experimental results and the theoretical ones are explained by the presence of an insignificant correlation between the parameters of the decision tree predicates.

4 Discussion

In order to maintain a balance between ensuring the economic efficiency of innovative enterprises and strengthening the economic and financial self-sufficiency of scientific organizations, it is necessary to apply effective rental instruments for involving the results of intellectual activity in economic circulation.

Setting inflated rent payments for the special use of hub patents is unjustified from both theoretical and practical points of view. In such a case, the basic principles of diffusion of open innovations (equality of actors and economic justification) are violated, the financial stability of the economic activity of innovative enterprises is undermined, and the process of bankruptcy of knowledge-intensive small business organizations spreads.

In the context of taxation of income in the form of exclusive rights to the results of intellectual activity, the amount of tax debt does not directly depend on the level of innovative development of the enterprise. A significant factor influencing the total amount of tax debt is the type of tax. Tax debt is formed mainly due to value added tax and income tax of innovative enterprises.

From the point of view of macro-financial stability analysis, the ratio of tax revenues to revenues from the sale of hub patents provides sufficient grounds for drawing conclusions regarding long-term trends in budget replenishment and identifying potential risks. However, tax revenues are not an indicator associated with the level of innovative development of scientific organizations. The level of non-payment of tax liabilities and the level of shadowing of patent transactions also reflect the economic situation and the real macro-financial stability of innovative enterprises.

The negative impact of taxpayers' tax arrears on the stability of the tax system and the effectiveness of its development determines the objective need to develop effective fiscal instruments to minimize it. In terms of macrofinancial analysis, it is necessary to identify and implement the most adequate indicator for recording negative trends and potential threats of hub patent transactions.

5 Conclusion

Algorithms for big data mining of hub patent transactions based on decision trees have been developed. Decision trees are constructed based on response surface shifts. Additionally, predictive structures of the model of successive geometric transformations have been introduced into the algorithmic implementation of decision trees, ensuring high-speed training and increasing the accuracy of the algorithms. A structural diagram of the operation of the developed algorithms is provided. The modeling of the algorithms' operation is carried out by solving the problem of forecasting prices for hub patents.

A number of experiments were conducted to select the optimal parameters of the proposed algorithms. The optimal number of decision tree predicates was determined taking into account computational costs and forecast accuracy. The best values of two algorithm parameters were determined by optimization using the evolutionary computation method: the magnitude of the biases and the range of the Gaussian function. A comparison with existing regression methods was made. High accuracy of operation was determined based on the standard deviation of the proposed algorithms compared to the considered methods.

The developed algorithms can be used to solve various problems of analyzing big data of the results of intellectual activity with increased accuracy.

The work was carried out within the framework of the state assignment of the Ministry of Education and Science of Russia on the topic " Methods and algorithms for monitoring, forecasting and expertise of higher education institutions' activities using artificial intelligence " № 124013000662-7.

References

1. S. Huang, P. Wang, Z. Lai, Computer Methods in Applied Mechanics and Engineering **432**, 117445 (2024).
2. H. Liao, Y. He, X. Wu, Information Fusion **100**, 101970 (2023).
3. Y. Ye, S. Xu, M. Mariani, Chaos, Solitons & Fractals **160**, 112234 (2022).
4. C. Hsieh, C. Lin, L. Lu, World Patent Information **78**, 102297 (2024).
5. Z. Cai, D. Ma, R. Zhou, Technological Forecasting and Social Change **208**, 123666 (2024).
6. S. Mishra, World Patent Information **65**, 102024 (2021).
7. X. Xiang, Y. Geng, Journal of Environmental Management **368**, 122193 (2024).
8. Z. Chang, W. Guo, L. Wang, Expert Systems with Applications **256**, 124895 (2024).
9. F. Mumali, Computers & Industrial Engineering **165**, 107964 (2022).
10. G. Vidal, R. Caiado, L. Scavarda, Computers & Industrial Engineering **174**, 108777 (2022).
11. A. Cammarano, V. Varriale, F. Michelino, Technological Forecasting and Social Change **209**, 123811 (2024).
12. A. Garcia, S. Lorenzo, J. Ripoll, Physica A: Statistical Mechanics and its Applications **639**, 129637 (2024).
13. Y. Kaplan, Engineering Applications of Artificial Intelligence **136**, 109034 (2024).
14. E. Pitz, K. Pochiraju, Engineering Applications of Artificial Intelligence **134**, 108622 (2024).
15. M. Xia, X. Zhao, X. Hu, Advanced Engineering Informatics **62**, 102721 (2024).
16. Y. Wang, Y. Wang, M. Tie, Engineering Applications of Artificial Intelligence **104**, 104393 (2021).
17. S. Langer, Journal of Multivariate Analysis **182**, 104696 (2021).
18. A. Khater, E. Gaballah, M. Bardin, ISA Transactions **152**, 191-207 (2024).
19. A. Nappa, M. Quartulli, I. Azpiroz, Ecological Informatics **82**, 102723 (2024).
20. Y. Li, W. Wang, T. Okaze, Sustainable Cities and Society **115**, 105837 (2024).
21. M. Mohseni, S. Zargarzadeh, N. Arjmand, Journal of Biomechanics **162**, 111884 (2024).
22. V.S. Averyanov, I.N. Kartsan, Voprosy cybersecurity **2(54),** 65-72 (2023).