Multi-User Digital Platform for Data Mining, Decisions' Roots Design and Decisions' Root-Based Neural Networks Training

Aleksandr Alekseev[1, 2, 3][0000-0001-5033-6694]

[1] Perm Decision Making Support Center, Russia
[2] Perm National Research Polytechnic University, Russia
[3] Perm branch of the National Research University Higher School of Economics, Russia

Abstract. The multi-user digital platform is described, primarily performed for data mining and decision-making support. The main purpose of the suggested digital data mining platform is design of decisions' roots – graphical tool for decision making support, and training of decisions' root-based neural networks – new type of incompletely connected neural networks. Besides data mining function of the discussed platform, the last one is also a no-code integrated development environment and a database management system.

Keywords: data mining; multi-criteria decision making; decision making support; integrated rating mechanisms; decisions' roots; neural networks; digital platform.

1 Introduction

This article discusses the multi-user digital data mining platform developed based on the SaaS concept for following data mining services: intelligent analysis of users' data, searching for patterns and quantitative relationships in data, supporting decision-making of the user or other persons in whose interests the data analysis was carried out, and solving other tasks on the basis of data mining.

The main purpose of the digital data mining platform is design of decisions' roots [1] – graphical tool for decision making support based on data analysis and training of decisions' root-based neural networks (hereafter, DRB neural networks or DRB NN) – new type of incompletely connected neural networks [2]. The methodology of DRB neural networks is suggested in the study [3].

The key advantage of decisions' roots and DRB neural networks is their clarity and the ability to interpret and explain the results obtained, which can be used to decision-making support. Explain the results is ensured by the fact that the structure of the DRB NN is determined strictly by the values of the convolution matrices, which are part of the decisions' root. The distinctive feature of DRB NN is their efficient training even on small samples, which is especially relevant in industrial sector and medicine.

The digital data mining platform can be considered not only as a SaaS, but also as a PaaS, because digital data mining platform was developed as a no-code integrated development environment and a database management system.

2 Literature Review

In 2019, Cherepanov F.M. proposed a few methods and algorithms for preprocessing and identifying the most significant indicators in his PhD thesis [4].

In 2020, on the 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency, idea of interval data coding and then discrete data analysis for constructing neural networks was proposed [5].

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

The idea of discretization of the observation domain is not new (see, for example, [6–11]).

The reason for revisiting this idea are the achievements of the last few years. Thus, in 2020, two research teams proposed approaches to identifying integrated rating mechanisms (IRM) on complete and incomplete training sets [8, 12]. The first approach [12] is based on a one-hot encoding representation of the IRM. The second approach [5] represented a transformation of the truth table in matrix form seems like Karnaugh map. In 2021, the second approach was generalized [13] to the case of IRM with any structure of a binary criteria tree reflecting a sequence of functions in superposition. In the same 2021, in the studies [1, 14] already noted the possibility of using IRM as a data analysis tool. The study [15] demonstrated examples of IRM synthesis based on the analysis of socio-economic systems. In 2022, the study [16] demonstrated the use of IRM as a data analysis tool in medical and biological systems.

Back in 2020, the study [5] showed that any IRM can be represented in the form of an incompletely connected artificial neural network (ANN). Later in 2023, such neural networks type was named as decisions' root-based neural networks (DRB NN) [3]. The identified DRB NN completely reproduce training sets with discrete values. Accordingly, the research question arose – Will the founded DRB NNs have an accuracy acceptable for applied tasks if DRB NNs are trained on the initial data in a continuous form? In 2022-2023, software module [4] was developed to answer on above question.

In 2024, Kozhemyakin L.V. in his PhD thesis [16] showed that the search for a DRB neural network of acceptable quality require almost half the time compared to the iterative search for a non-fully connected neural network.

3 Methodology and concept of the digital data mining platform

Users are understood as individuals or representatives of legal entities who, under a license agreement, have the right to use the digital data mining platform software [2] as a service: intelligent analysis of user's data, searching for patterns and quantitative relationships in data, decision-making support, and solving other tasks on the basis of data mining.

According to the initial concept (Figure 1), an authorized user uploads his own data, in which, in his opinion, there is some pattern that should be quantitatively described, and eventually receives a ready-made neural network that quantifies causal and systemic relationships, as well as functional dependencies in the analyzed data.

From an algorithmic point of view, user data at the first stage is processed in a software module (hereafter, SW) for processing heterogeneous datasets in the form of discrete (categorical) data – SW Module 1 (see Figure 1), where the transition from continuous variables to discrete and from nominal to categorical is carried out. So, for quantifiable variables, interval coding and the construction of discontinuous reduction functions are used, establishing a connection between a set of continuous values of the indicator and discrete values. In fact, this is the division of the observation area of a variable into intervals and the representation of these intervals as an encoded variable. It is important that the user can independently configure the parameters of the conversion function for each quantitative variable. Unlike many programs used for data analysis, in the software module being developed, the number of intervals for different variables may differ, which is more in line with the needs of users to adequately describe the subject area they are studying. For nominal variables, a user-defined directory formation mode is implemented, where some nominal variables are combined by the user into a certain class characteristic of a certain category. Any of these approaches can be applied to ordinal variables i.e., both reduction functions and reference books can be constructed.

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

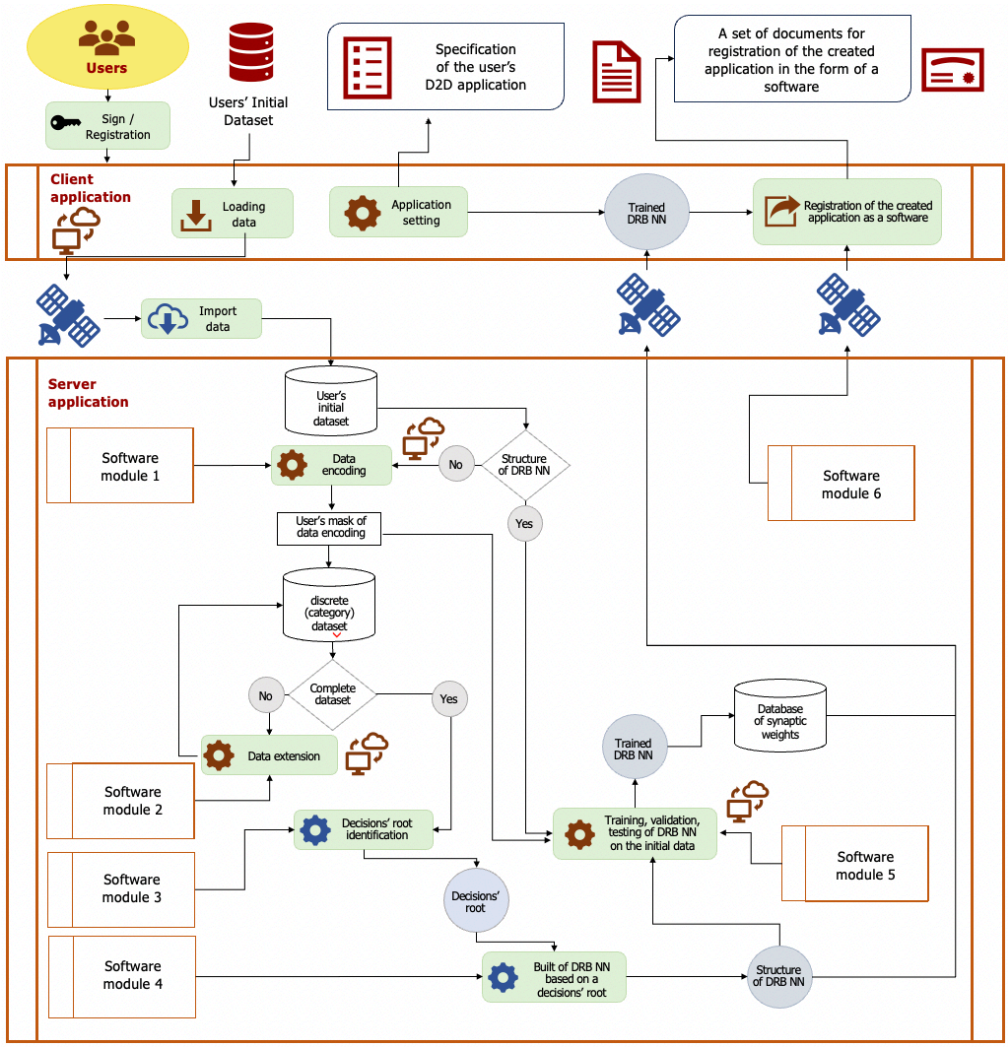


Fig. 1. The initial concept of the multi-user digital platform for data mining, decisions' roots design and decisions' root-based neural networks training

Data encoded with discrete (categorical) values is checked for completeness, where completeness of data means the presence of all possible combinations of arguments. If the data is incomplete, then it was supposed to launch a separate software module for determining unknown parameters in training sets to identify the decisions' root – SW Module 2 (see Fig. 1), as a result of which unknown data is determined by the user, based on the possibility of binding rows and columns of matrices obtained by constructing Karnaugh maps. If the data is complete, then the identification of the decisions' root is performed using the algorithms described in [8; 16] and the software module for identifying the integrated assessment mechanism based on a complete set of training examples – SW Module 3 (see Fig. 1).

The next stage of information processing is to represent the decisions' root in the form of an artificial neural network and determine its structure – this information processing is performed in the software module for building the structure of a neural network according to the given decisions' root – SW Module 4 (see Fig. 1). The final step of information processing is the training of the decisions' root–based neural networks in the software module – SW Module 5 (see Fig. 1).

As a result of development of functional software modules that are part of the software of the multi-user digital data mining platform, the author see a corrected

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

concept of its operation (Fig. 2).

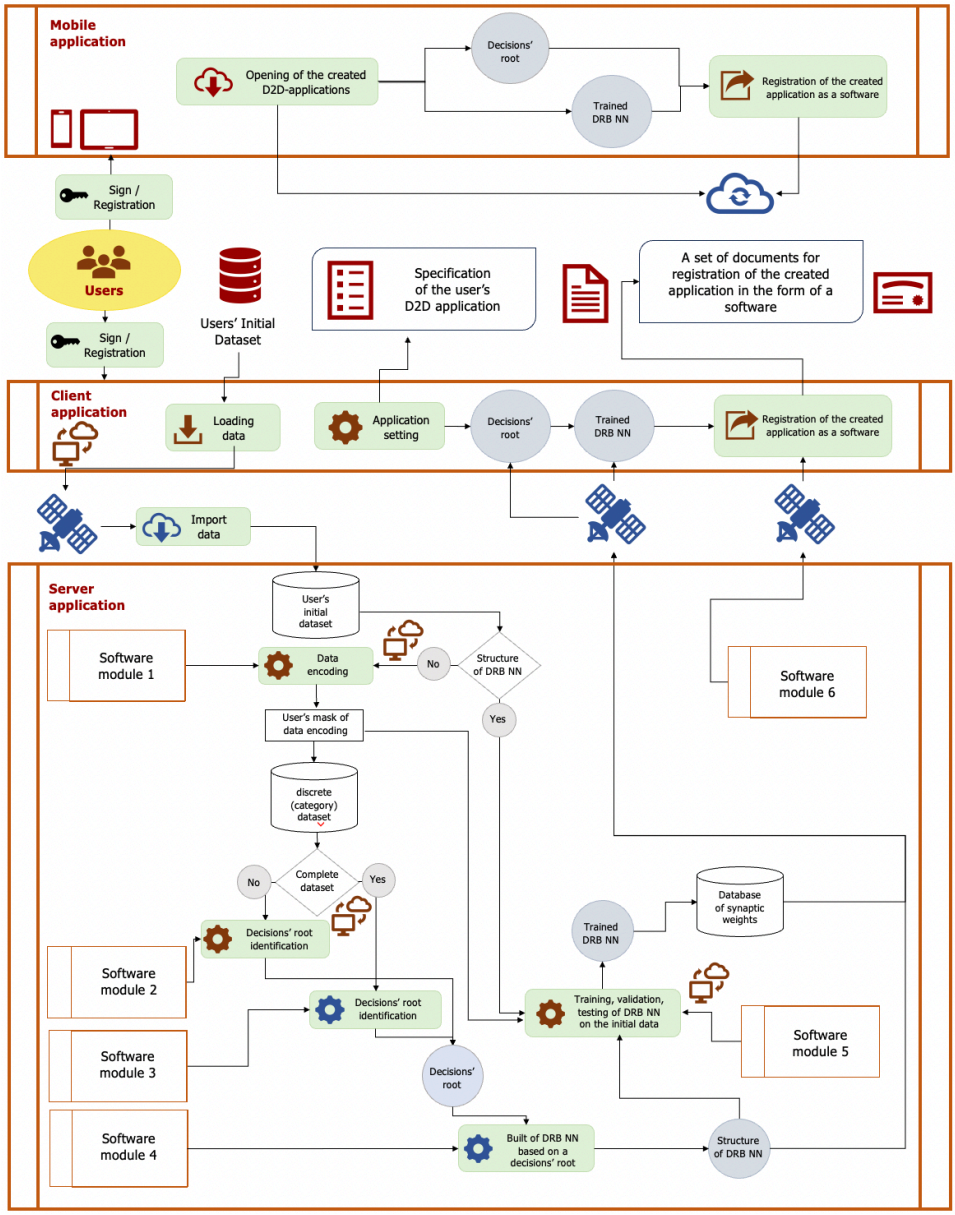


Fig. 2. Corrected concept of the multi-user digital platform for data mining, decisions' roots design and decisions' root-based neural networks training

The main change is the change in information flows, in particular, the organization of the process of determining unknown parameters in an incomplete training set depends on the structure of the criteria tree, therefore, software module 2 can be implemented according to the principle of operation of software module 3 to identify complex assessment mechanisms based on complete training sets, but differs in defining rows with missing data, which can be bound together, as well as by launching dialog boxes, in which the user is asked to choose which lines should be bound together with the variability of binding (see Figure 2). In the case of a complete data set, the software module for identifying the mechanisms of complex assessment based on complete training sets works without user participation i.e., in automatic

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

mode. Thus, the result of the work of both software modules is the desired mechanism of integrated assessment (the decisions' root).

In addition, not only the DRB neural network may be of value to the user in the end, but also the identified decisions' root as an independent solution, since it allows one to visualize a function approximating discrete (categorical) data in a compact form (see Figure 2 – client application). The found the decisions' root can already be used to support decision-making. However, it is necessary to consider the limited accuracy of the found function, since it works at user-defined intervals. It is the user who selects the encoding settings for quantitative variables in the dialog boxes and generates directories of nominal variables.

According to the author, another promising direction for the future development of a multi-user digital data mining platform is the creation of a mobile application (see Figure 2) in which the user can use ready-made applications in which cause-effect and system relationships, or functional dependencies have already been found in the analyzed data (if any), i.e. The user will already be able to use these dependencies to justify management decisions, demonstrate the discovered patterns in graphical form, but will not be able to perform data mining to search for such dependencies, since it is obvious that it will not be convenient to perform from a mobile device.

4 Results

In study [18] was considered data mining example that includes gross regional product (GRP) of Russian regions as explicable variable (GRP is measured in billion rubles) and following statistical indicators as explanatory variables: INTERNET – the share of households with Internet access, %; PC – the share of organizations that used personal computers, %; MONEY – the cost of implementing and using information and communication technologies, million rubles per 1000 people; WEB – the share of organizations that had a website on the Internet, %; ECONOM – the level of socio-economic development of the regions, measured in points in the rating of the news agency "RIA Novosti".

Let's consider the main steps of data mining by using the digital platform on the same example. Screen form at the import data step presents on the following figure (Fig. 3).

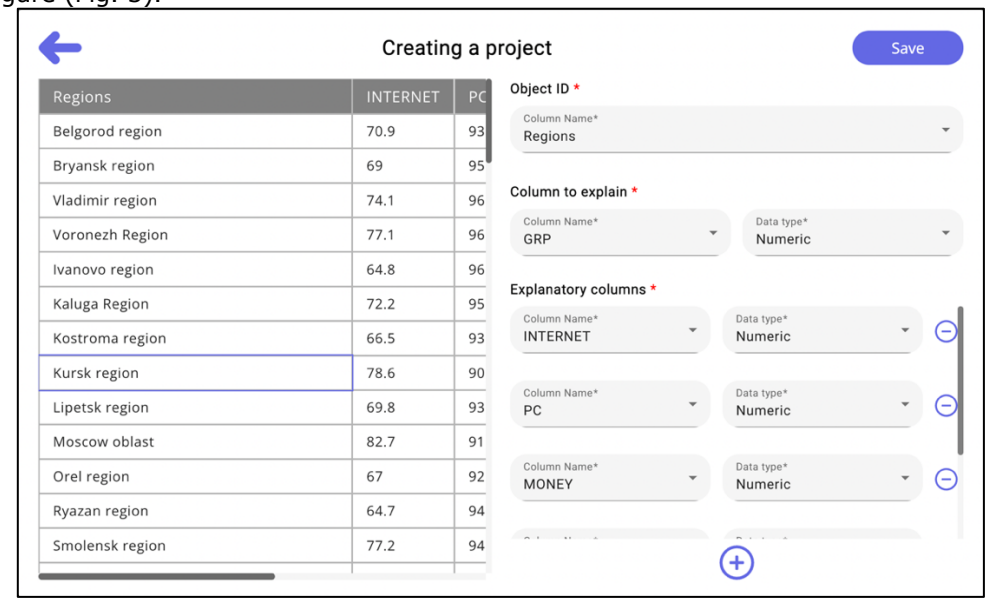


Fig. 3. Screen of the digital data mining platform at the step of creating a project – import dataset and setting

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

On this step it is necessary to choose the column that will be use as object identificatory, the column including values of explicable variable, and columns corresponded to explanatory variables.

Next step includes interval encoding for both type of variables: explicable and explanatory. User can to derive observation domain to any numbers of intervals with accordance to his expert view on the research space, in following figure (figure 4) showed two intervals with equal distances: the first interval is mean GRP is less than boundary point 2928.095, and the second – more than 2928.095.

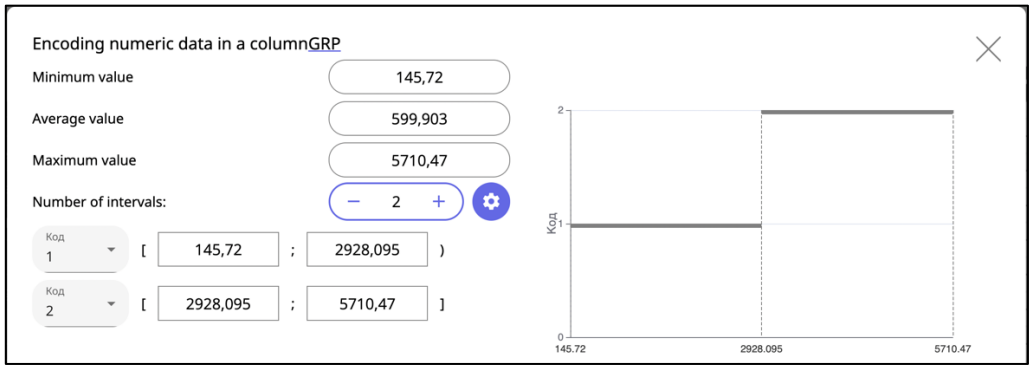
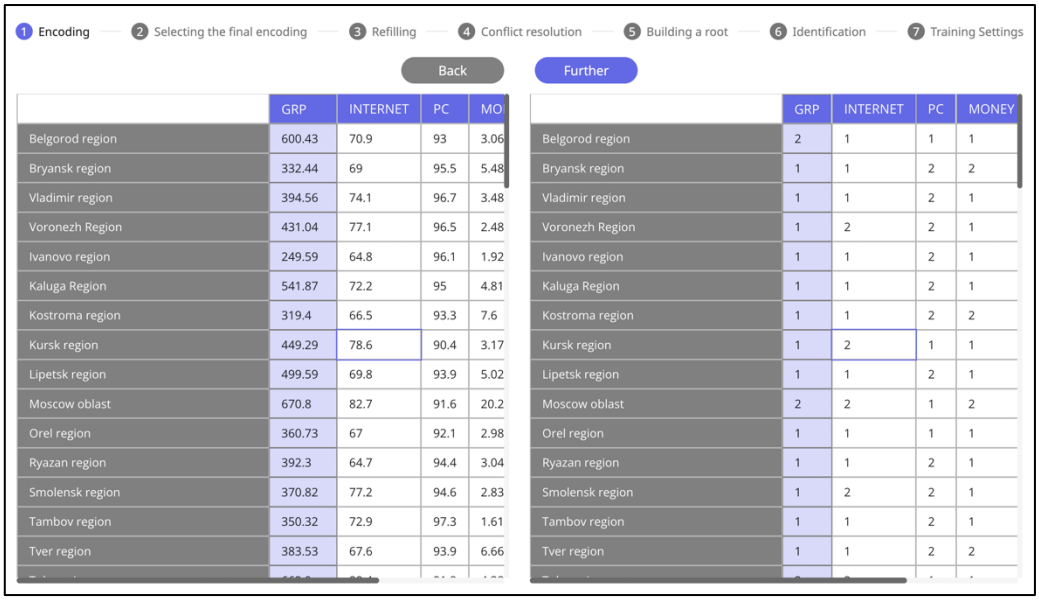


Fig. 4. Screen of the digital data mining platform at the step of interval encoding

As result of this step, user obtained a dataset with discrete (categorical) values (Fig. 5). Encoded dataset on Fig. 5, right part) makes it possible to use methods of the discrete mathematics to searching for patterns and quantitative relationships in data by approximation of discrete function. Note that any multi-valued discreate function of many variables can be presented as a superposition of set discreate functions of two variables.



	GRP	INTERNET	PC	MO
Belgorod region	600.43	70.9	93	3.06
Bryansk region	332.44	69	95.5	5.48
Vladimir region	394.56	74.1	96.7	3.48
Voronezh Region	431.04	77.1	96.5	2.48
Ivanovo region	249.59	64.8	96.1	1.92
Kaluga Region	541.87	72.2	95	4.81
Kostroma region	319.4	66.5	93.3	7.6
Kursk region	449.29	78.6	90.4	3.17
Lipetsk region	499.59	69.8	93.9	5.02
Moscow oblast	670.8	82.7	91.6	20.2
Orel region	360.73	67	92.1	2.98
Ryazan region	392.3	64.7	94.4	3.04
Smolensk region	370.82	77.2	94.6	2.83
Tambov region	350.32	72.9	97.3	1.61
Tver region	383.53	67.6	93.9	6.66

	GRP	INTERNET	PC	MONEY
Belgorod region	2	1	1	1
Bryansk region	1	1	2	2
Vladimir region	1	1	2	1
Voronezh Region	1	2	2	1
Ivanovo region	1	1	2	1
Kaluga Region	1	1	2	1
Kostroma region	1	1	2	2
Kursk region	1	2	1	1
Lipetsk region	1	1	2	1
Moscow oblast	2	2	1	2
Orel region	1	1	1	1
Ryazan region	1	1	2	1
Smolensk region	1	2	2	1
Tambov region	1	1	2	1
Tver region	1	1	2	2

Fig. 5. Screen of the digital data mining platform at the finished step of interval encoding

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

Superposition of a set functions can be presented as an oriented graph (function tree) corresponded to sequence of functions, leaves of such graph are variables. That, next step in the digital data mining platform includes building of binary tree with named leaves (function tree) that are part of the decisions' root. User can to convolute any pair of initial variables and new nodes created by convolution. In following figure showed graph has four convolution nodes: the first one corresponded to convolution of the variables "PC" and "WEB"; at the same level of tree, convolution of the variables "INTERNET" and "MONEY" create second convolution node "INTERNET + MONEY"; on the second level of tree node "PC + WEB" convoluting with variable "ECONOM" create third node "PC + WEB+ ECONOM"; the last node corresponding to convolution of the node "PK + WEB+ ECONOM" and node "INTERNET + MONEY" (Fig. 6).

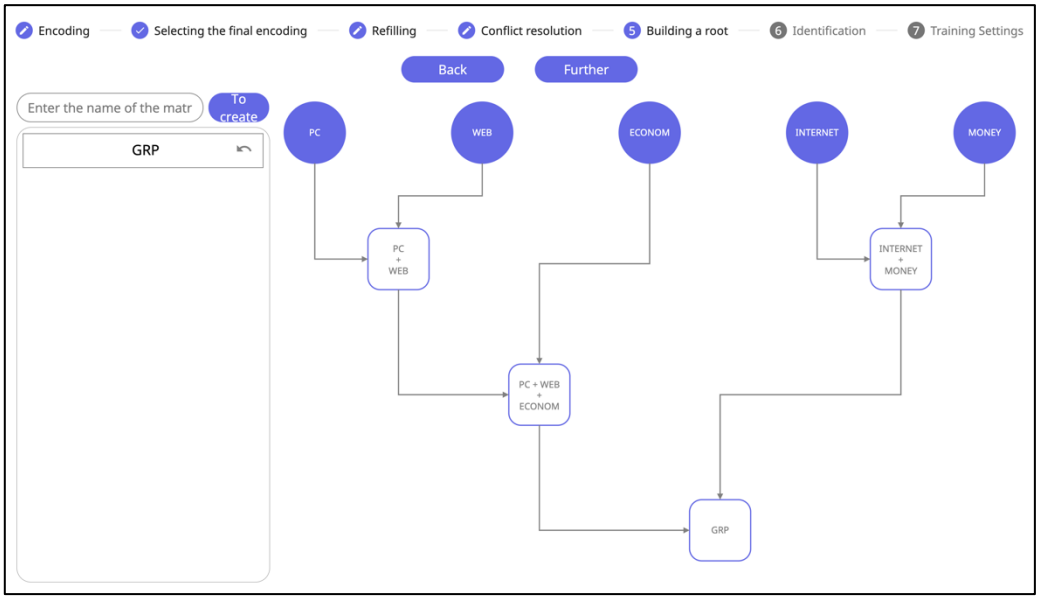


Fig. 6. Screen of the digital data mining platform at the step of decisions' root structure building

For built decisions' root structure of criteria tree (function tree), table with all variables sets and corresponded them function value is sorted in order corresponded to the list of leaves names. Next step includes creating of the matrix from table of discrete function. Obtained matrix seems like a Karnaugh map, but with deleted duplicates.

On the following step, user should to determinate unknown elements – the user should strive to glue the rows of the matrix in such a way as to obtain the smallest number of rows. It should be borne in mind that duplicates in columns will also be deleted.

Below (figure 7) showed that in first row forth unknown element (marked by symbol "~" named "Tilda") make it determinates like in second completely known row; third row has three unknown elements, but third element has value 1 like in second completely known row, in fourth row third element has value 2, that so third row couldn't be glue with fourth row.

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

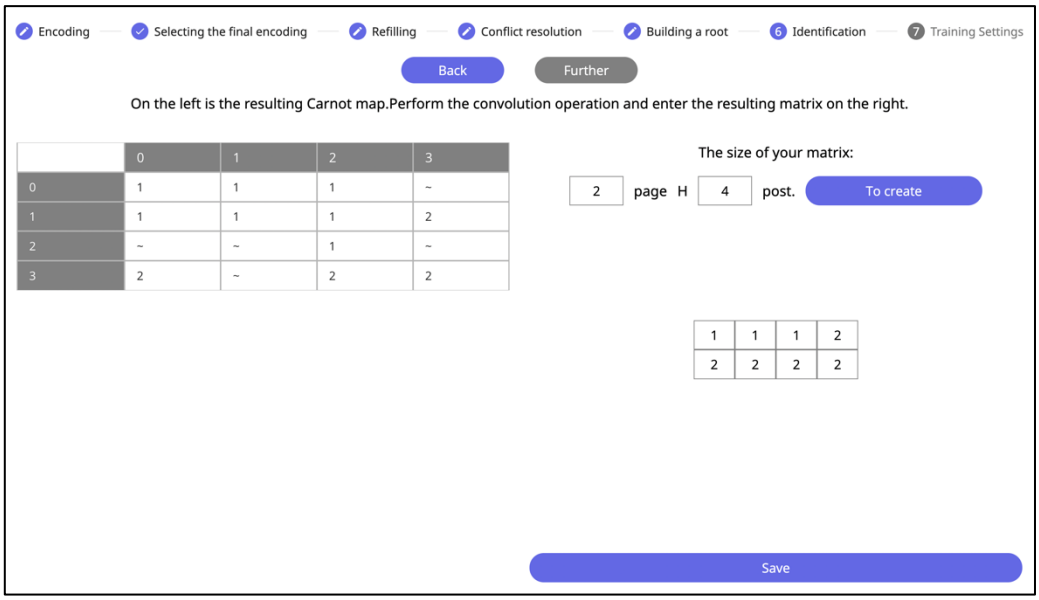


Fig. 7. Screen of the digital data mining platform at the step of the decisions' root identification

In fourth row second unknown element could be determine 1 or 2. We could see that after glue of first three rows the first and the third columns will be have a set {1, 2}. So, if user determine unknown element by 1, then we obtain a new column with set {1, 1}, but if user determine unknown element by 2, then we obtain existed column with set {1, 2}. The user should strive to glue the columns of the matrix in such a way as to obtain the smallest number of columns too. That is why in this case, the final matrix after removing of duplicates in columns obtain the matrix $\{m_{11}=1, m_{12}=2, m_{21}=2, m_{22}=2\}$. This matrix corresponds to the root node named as "GRP". The remaining matrices could be obtained automatically by performing algorithm suggested in the study [14] (Fig. 8).

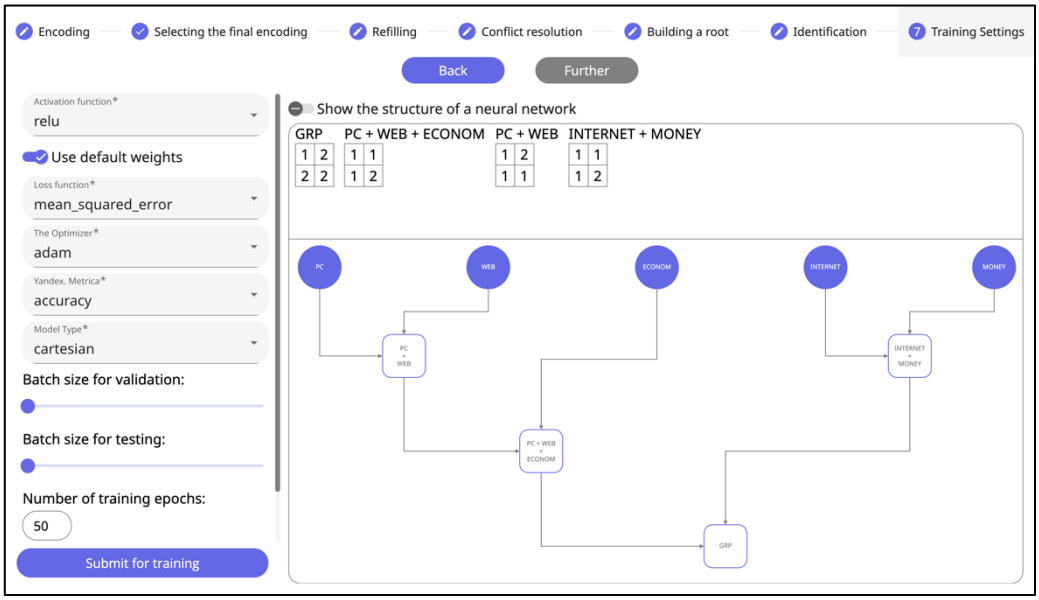


Fig. 8. Screen of the digital data mining platform at the step of the identified decisions' root and set of the convolution matrices

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

Above was defined that any multi-valued discreate function of many variables can be presented as a superposition of set discreate functions of two variables. In the same time, any discrete function of two variables can be presented in form of a binary matrix. Functions of two variables correspond to the nodes of binary graph (function tree). Decisions' root – is discrete function graphically represented in form of oriented tree and set of the convolution matrices. That, decisions' root is useful graphical tool for presentation complex discrete functions. Moreover, above was noted that any decision root make be represented in form of incompletely connected neural network – decisions' root-based neural network (Fig. 9).

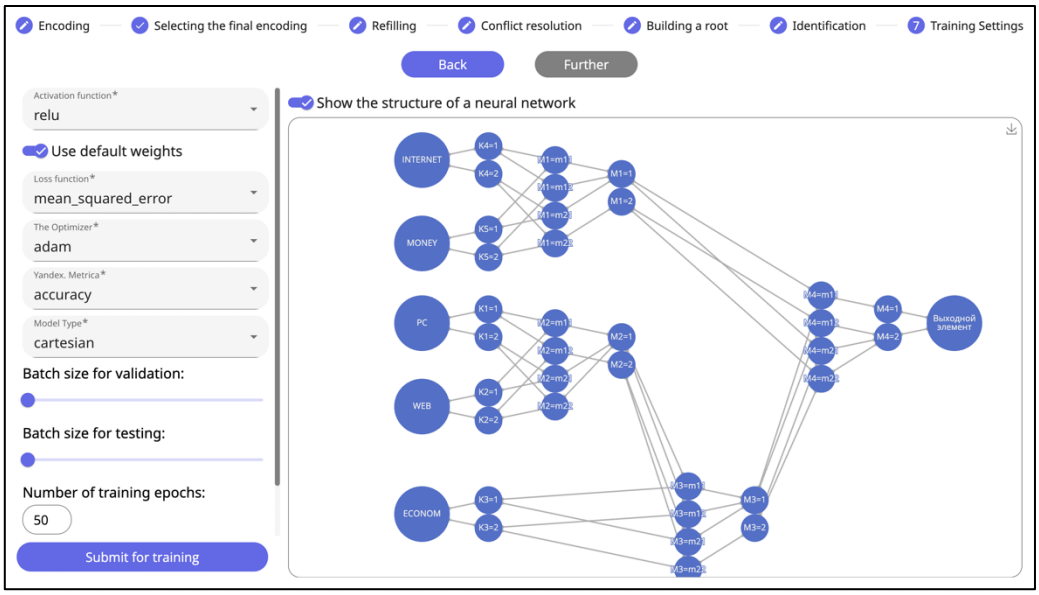


Fig. 9. Screen of the digital data mining platform at the step of setting of training of the decisions' root-based neural network

5 Conclusion

In conclusion, it is worth noting that discussed the digital data mining platform was developed as a no-code integrated development environment and a database management system. The digital data mining platform provides a software module for preparing a set of documents for registration of a computer program – SW Module 6 (see Figure 1, Figure 2), which records User actions.

The User of the digital data mining platform owns the copyright to the mathematical model of the decisions' root and the mathematical model of the decisions' root-based neural network founded in the digital data mining platform and exclusive rights to the User application created in the digital data mining platform, which implements the data processing sequence in the digital data mining platform in accordance with the models.

All User actions in the process of data mining (formation of reference books of nominal variables, determination of parameters of functions for reducing quantitative variables, selection of the structure of the function tree, determination of unknown values in the encoded data set when identifying the root matrix) in digital data mining platform are perceived as an author's contribution to the creation of Models and User applications. The User application is the result of the intellectual activity of the User and can be registered as a computer Program. A set of independent materials that determine the decisions' root found by the user (the structure of the function tree, convolution matrices) and the structure of the neural network (the number of hidden

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

layers, the number of neurons on each layer, connections between neurons), as well as a set of synaptic coefficients of the neural network, can be registered by the user as a database.

Users could obtain their own web applications in form of the digital decisions' root (Fig. 10) and DRB neural network (Fig. 11).

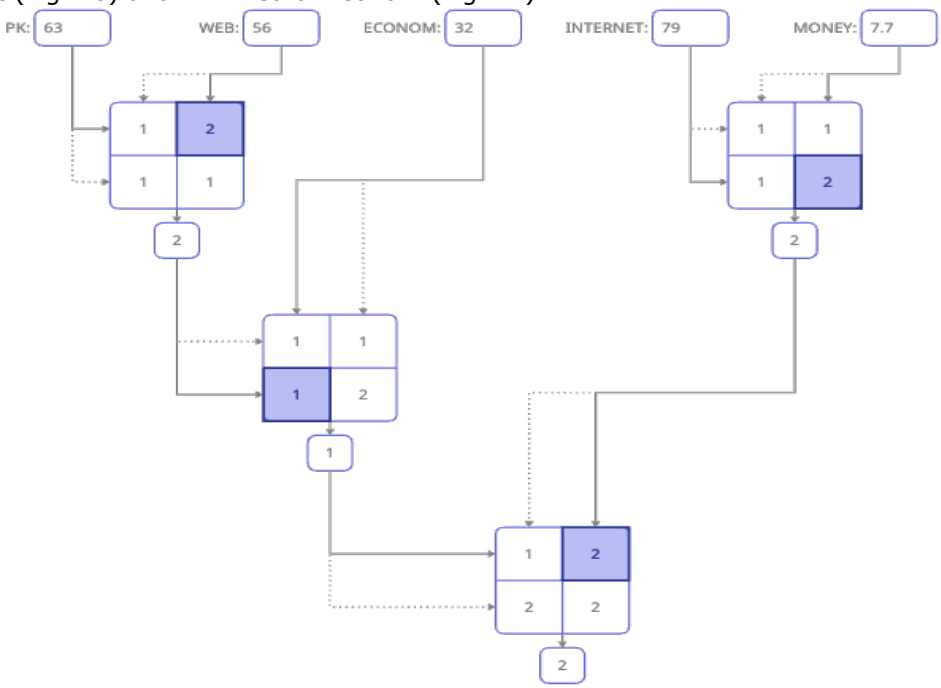


Fig. 10. Example of the decisions' roots identified in the digital data mining platform

Explain the results is ensured by the fact that the structure of the DRB NN is determined strictly by the values of the convolution matrices, which are part of the decisions' root.

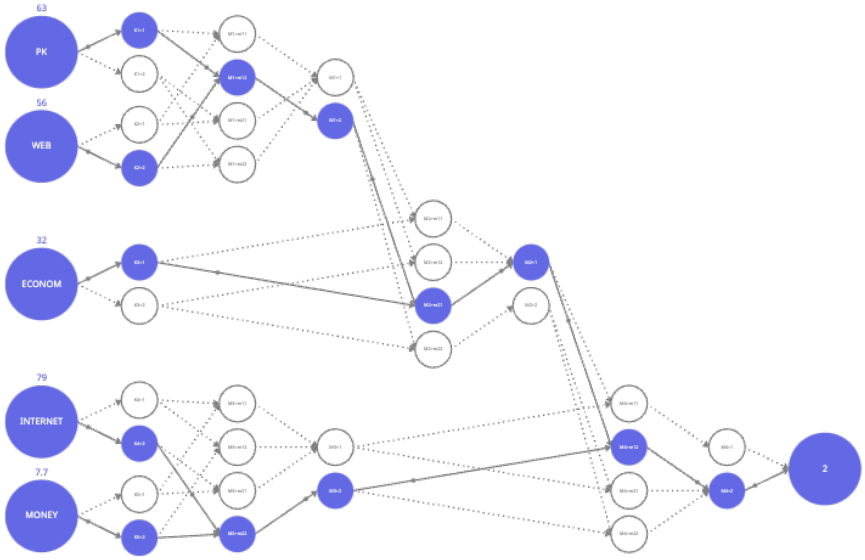


Fig. 11. Example of the decisions' root-based neural network identified in the digital data mining platform

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

The key advantage of decisions' roots and DRB neural networks is their clarity and the ability to interpret and explain the results obtained, which can be used to decision-making support.

The distinctive feature of DRB NN is their efficient training even on small samples, which is especially relevant in industrial sector and medicine. In [16] showed that the search for a DRB neural network of acceptable quality require almost half the time compared to the iterative search for a non-fully connected neural network.

Acknowledgments

Development of the digital data mining platform software [2] was supported by the Foundation for Assistance to Small Innovative Enterprises (FASIE), Agreement No. 4801GS1/80326 dated 2022/11/24.

Development of digital education case and tutorial for students [18] was supported by the Vladimir Potanin Foundation, Agreement No. GSGK-094/24 dated 2024/03/05.

References

1. Korgin N.A., Sergeev V.A. Decisions' Root—Yet Another Tool for Ordinal Data Analysis. URL: https://youtu.be/b4dF7znmVyo, last accessed 2021/27/11.
2. Alekseev A.O., Kozhemyakin L.V., Makkaveev A.V. Software Module for Designing and Training Decisions Root-based Neural Networks. Certificate of the Computer Program RU 2023685360 from 2023/11/27
3. Alekseev A., Kozhemyakin L., Nikitin V., Bolshakova J. Data Preprocessing and Neural Network Architecture Selection Algorithms in Cases of Limited Training Sets – On an Example of Diagnosing Alzheimer's Disease. Algorithms. 2023. Vol. 16. Iss. 5. Art. 219. DOI 10.3390/a16050219.
4. Cherepanov F.M. Methods for improving the efficiency of neural network recommender systems in conditions of limited sample sizes with complex correlation links (using the example of diagnostics and prognosis of human cardiovascular diseases). Ph.D. Thesis. Perm, 2019, 136 p.
5. Alekseev A.O. Identification of Integrated Rating Mechanisms Based on Training Set. Proceedings of 2020 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency. SUMMA 2020. Virtual, Lipetsk. IEEE. 2020, pp. 398-403. DOI 10.1109/SUMMA50634.2020.9280751.
6. Vereskun V.D., Guda A.N., Butakova M.A. Data mining: Discretization of attribute values using the theory of rough sets and clustering. Rostov State Transport University Bulletin, 2018, 3, 76–84.
7. Ching J.Y., Wong A.K.C., Chang K.C.C. Class-dependent discretization for inductive learning from continuous and mixed mode data. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1995. Vol. 17, no. 7. pp. 641-651
8. Dougherty J., Kohavi R., Sahami M. Supervised and unsupervised discretization of continuous features. Proceedings 12th International conference on machine learning, 1995. pp. 194-202.
9. Catlett J. On changing continuous attributes into ordered discrete attributes. Proceedings of the European working session on learning. Berlin, Germany, 1991. pp. 164-178.
10. Fayyad U.M., Irani K.B. Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the 13th International joint conference artificial intelligence, 1993. pp. 1022-1027.
11. Liu H., Setiono R. Feature selection via discretization. IEEE Transactions of Knowledge and Data Engineering. 1997. Vol. 9. № 4. pp. 642-645
12. Burkov V.N., Korgin N.A., Sergeev V.A. Identification of Integrated Rating Mechanisms as Optimization Problem. Proceedings of 2020 13th International Conference Management of Large-Scale System Development. MLSD 2020. 2020. DOI 10.1109/MLSD49919.2020.9247638.
13. Alekseev A.O. Identification of Integrated Rating Mechanisms with Non-serial Structures of Criteria Tree. Proceedings of 2021 3nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency. SUMMA 2021. Virtual. Lipetsk. IEEE. 2021, pp. 599-603. DOI 10.1109/SUMMA53307.2021.9632221.
14. Sergeev V.A., Korgin N.A. Identification of integrated rating mechanisms as an approach to discrete data analysis. IFAC-PapersOnLine. 2021. Vol. 54, No. 13. pp. 134-139. DOI 10.1016/j.ifacol.2021.10.433

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.

15. Kozhemyakin L.V., Alekseev A.O., Nikitin V.N. Application of Decisions' Roots for Data Analysis on Example of Dataset with Magnetic Susceptibility Values of the Brain Veins and the Alzheimer's Disease. Proceedings of 2022 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency SUMMA 2022. Lipetsk, 09-11 November 2022. IEEE, 2022. pp. 425-427. DOI 10.1109/SUMMA57301.2022.9973977.
16. Kozhemyakin L.V. Methods and algorithms for preprocessing limited data and assessing the quality of incompletely connected neural networks when they are built based on decisions' roots. Ph.D. Thesis. Perm, 2024, 160 p.
17. Alekseev A.O. et al. Decisions' Roots and Trees Design for Graphical Representation of a Discrete Function Found Based on Interval Coding of Statistical Indicators. KubSAU Scientific Journal, 2024. no. 10 (204). – available at: http://ej.kubagro.ru/2024/10/pdf/61.pdf, DOI 10.21515/1990-4665-204-061
18. URL: https://do.pstu.ru/enrol/index.php?id=3261, last accessed 2024/12/30

© The Author(s). JDS 6(2), 2024. Published by ICS, licensed under CC BY 4.0.