# Initial database creation for scientific and technical solutions efficiency assessment based on artificial intelligence approach

*Alexandra* Khalyasmaa[1,*], *Elena* Zinovieva[1], *Stanislav* Eroshenko[1] and *Daria* Shatunova[1]

[1]Ural Federal University, 620002 Mira Street 19, Ekaterinburg, Russian Federation

**Abstract.** This paper is devoted to the problems and features of database creation in intelligent systems for assessing the efficiency of scientific and technical solutions. The system data model developed by the authors and the principles of its operation are described. This paper also considers the process of training sampling and the analysis of various teaching methods for solving the presented problem. The implementation of the developed model is made using mathematical modeling. The initial data was the data of applications for grants in the field of technical sciences related to the fuel and energy complex.

## 1 INTRODUCTION

In the modern conditions, when new scientific approaches and subjects based on a general scientific and interdisciplinary approach appear, there is an urgent need for a fundamentally new system for obtaining reliable estimates to solve the multicriterial, multifactor and difficult for formalization task of assessing the efficiency and feasibility of scientific and technical projects.

The authors of this paper are implementing a project to develop a model and software for a learning and adaptive analytical system that allows operating large amounts of data and is able to perform a differentiated evaluation of the efficiency of scientific and technical solutions and technologies, depending on the likelihood of their feasibility under initial data incompleteness. This paper is devoted to the main problems of the database creation in systems for assessing the efficiency of scientific and technical solutions implemented on artificial intelligence methods.

Currently, modern automated systems allow not only processing and analyzing large data arrays but also their accumulating, and modern mathematical methods, such as methods of artificial intelligence, allow finding new implicit laws and correlations in them. It seems that a large amount of data should help improve the accuracy of calculations but practically it is not always true, since in this case the problem becomes multicriterial where the optimizing is a multistep (not always sequential or hierarchical) optimization task under uncertainty. The data accumulation leads to the need for training, continuous refinement and changes in the structure of data presentation: new objects are revealed, new properties appear, new relationships arise.

To solve the above problems and to create an automated system for assessing the efficiency of scientific and technical solutions, in this research it was decided to form it as a decision support system (DSS). Any DSS is based on a database designed for solving data entry, storage and analysis problems [1].

When using artificial intelligence methods in DSS, the task of the database creation becomes more complicated in view of the complexity of implementing artificial intelligence methods themselves and integrating them into the general structure of the system. It is the quality, structure, and algorithms of the database that essentially determine the quality of the entire DSS system to a greater extent. That is why this paper is devoted to the problems and features of the database creation in intelligent systems for assessing the efficiency of scientific and technical solutions.

## 2 Database creation features and its structure

Modern intelligent systems consist of two types of heterogeneous objects: databases and knowledge bases. The presentation and processing of data (facts, tables, graphs, networks, texts, etc.) is carried out by algebraic methods and knowledge base models (predicates, frames, semantic networks, rules) are built based on a declarative approach which greatly complicates the relationship between the database and knowledge base in one system [2].

### 2.1 Database

First of all, databases are used for the data storage and affording the access and manipulation of stored data [3].

[*] Corresponding author: lkhalyasmaa@mail.ru

The basis of any database is data suitable for use and analysis at various stages of decision making [4,5].

The database creation is always preceded by collecting the up-to-date information on the necessary topics, which should be analyzed and structured in accordance with the selected criteria [6]. As practice shows, the initial selection of information is the most difficult and it is on it that the efficiency of future work depends. This problem is characterized by the following main factors [7]:

– compatibility of objective and subjective characteristics of data;

– multicriteriality: the need to simultaneously consider both quantitative and qualitative criteria for assessing the data;

– multiplicity of selection process.

The main problems in the process of database design are ensuring integrity, consistency, restorability, security and efficiency. A database has the integrity property if it satisfies certain constraints on the values and structure of the data. Data security refers to the protection against unintentional access to data. Restorability is the ability to restore the data integrity after any system failure. Consistency is the property of the database in relation to a certain set of users which means the database ability to responds to their requests in the same way at any time. Efficiency is the ability to use computing resources in a user-friendly time during the execution of database applications using the minimum amount of external memory for data storage.

The system for assessing scientific and technical solutions based on artificial intelligence methods developed by the authors of the paper is described in detail in [8-11]. At the stage of data collection, the input data volume for the database creation is determined and estimated. Data generation is a sequential process that includes pre-processing of data and the creation of the database itself.

Preliminary stage of data processing is due to the need for data scaling, normalization, as well as automated identification and verification of some data such as papers from the list of specialized scientific citation databases Scopus, Elibrary, etc.

In this case, to solve the problem of assessing scientific and technical solutions using artificial intelligence, the database of the developed system essentially serves as a training sample.

It is obvious that the same problem can be almost always solved by various methods. The choice of the mathematical approach used (method, algorithm, etc.) influences the accuracy, speed of operation and structure of the developed model. But no matter how effective the used algorithm is, the result will not give the desired result if the initial data presented by a training sample is poor-quality, unprocessed and unprepared.

That is why the issue of database creation is so important in any tasks using artificial intelligence. The sequence of the training sampling is described in detail by the authors of this paper in [8]. In this paper, more attention is paid to the functioning of the developed system from the standpoint of its trainability.

## 3 Learning techniques

In order to solve the problem of scientific and technical solutions effectiveness assessment within the framework of the presented study the authors of the paper analyzed three of the most popular and relevant types of machine learning approaches:

- unsupervised learning;
- supervised learning;
- semi-supervised learning.

Semi-supervised learning - implies both labeled and unlabeled data. Most frequently a small amount of labeled and a significant amount of unlabeled data are used. Semi-supervised learning is a compromise between unsupervised learning (without any labeled learning data) and supervised learning (with a fully labeled learning sample).

Despite the fact that these methods are fundamentally different, the presented task within the framework of the study is an integral one and involves solving two fundamentally different aspects:

- analysis of technologies and new scientific and technical solutions that have analogues worldwide;

- analysis of new technologies and new scientific and technical solutions (newly introduced, not having analogues worldwide); in other words, the analysis is carried out in conditions of uncertainty;

If in the first case it is obvious that this is a task implying supervised learning, in the second case it is a task with unlabeled data, where it is impossible to assign the value “output” in the “input-output” bundle, since the technology is completely new and there is no such sample. But for any expert, it is obvious that even if there is no absolutely similar technology in some sense there may be related technologies and approaches. And in this case, neither supervised learning, nor unsupervised learning in an explicit form is an effective tool to apply.

The above hypothesis was confirmed by a series of computational experiments where two different methods were analyzed: neuro-fuzzy inference and gradient boosting decision trees.

### *2.3.1 Neuro-fuzzy inference*

The rules in Takagi-Sugeno’s neuro-fuzzy inference approach are not presented in a form of belonging of the output variable to the given fuzzy sets, but in the form of functional dependencies

$$R^{(k)}: \text{if } (x_1 \textit{ is } D_1^k \textit{ and} \ldots \textit{and } x_N \textit{ is } D_N^k), \textit{ then}$$
$$y = f^{(k)}(x_1, ..., x_N) \qquad (1)$$

where $R^{(k)}$ is the fuzzy rule set; $k = 1,...,K$, where $K$ is the number of fuzzy rules; $D_i^k$ - fuzzy sets, where $D_i^k \in X_i \subset R,\ i = 1,...,N$; $x_1,...,x_N$ - input variables; $y$ - output variable, presented as a function of input variables.

As a result of the application of this model, a fuzzy neural network of the required structure was created. In Takagi-Sugeno neuro-fuzzy inference, the linear function of the input variables is used as the final rule.

$$y_i(x) = p_{i0} + \sum_{j=1}^{N} p_{ij} x_j, \qquad (2)$$

where $N$ is the number of parameters of the set $X$; $p_{i0}, p_{i1}, ..., p_{iN}$ - parameters of the Takagi-Sugeno polynomial [9].

Neuro-fuzzy inference, as a method based on neural networks, involves the development of not only a training sample, but also the architecture of the neural network itself along with the definition of optimal membership functions.

### *2.3.2 Gradient boosting decision trees*

The gradient booster trees algorithm is implemented in such a way that, firstly, the basic algorithm $b_0$ is initialized. For $n = 1, \ldots, N$ the following steps are taken:

- the shifting vector $S$ is calculated, which shows how to correct the predictions of the already constructed composition of trees in order to reduce the error on the training set:

$$s_n = \left(-2(a_{n-1}(x_1) - y_1), \ldots, -2(a_{n-1}(x_l) - y_l)\right); \qquad (3)$$

- the basic algorithm $b_n$ is built by approximating its output on the training sample to the shifting vector $S_i$:

$$b_N(x) = \arg\min_b \frac{1}{l}\sum_{i=1}^{l}\left(b(x_i) - s_i\right)^2 = \sum_{j=1}^{J}\left[x \in R_{Nj}\right] b_{Nj}; \quad (4)$$

- after the algorithm is found, it is added to the composition:

$$a_n(x) = a_{n-1}(x) + \eta \sum_{j=1}^{J}\left[x \in R_{Nj}\right] b_{Nj}. \qquad (5)$$

Then the previous three steps are performed until the stopping criterion is fulfilled. As a result, the output of the algorithm is a piecewise constant function that describes the function of technical and scientific solutions efficiency assessment.

To analyze the testing error of the algorithm $a$ on the sample $X^l$, we used the average error $Q$ in percent [%], which is calculated by the formula:

$$Q(\mathrm{a}, \mathrm{X}^l) = \frac{1}{l}\sum_{i=1}^{l}\left|a(\mathrm{x}) \neq \mathrm{y}^*(\mathrm{x})\right| \cdot 100\%, \qquad (6)$$

where $\left|a(\mathrm{x}) \neq \mathrm{y}^*(\mathrm{x})\right|$ is the error indicator; $\mathrm{y}^*(\mathrm{x})$ - target dependence; $l$ - the number of observations.

### *2.3.Semi-supervised learning*

The task of labeling the data for system learning often requires a qualified person, but as mentioned above, some data may not be labeled due to the absence of an expert. One of the key tasks of the project is to obtain reliable estimates for “new” technologies, scientific and technical solutions (firstly introduced and having no analogues worldwide) under the conditions of incomplete initial dataset. In this case, the use of semi-supervised learning can be a good alternative in order not to divide the analyzed task into two subtasks: supervised learning and unsupervised learning.

In this case, from the group of semi-supervised learning algorithms, the most appropriate for the presented task is the self-training algorithm.

Step by step procedure is given as follows:

– there is a data set $T = \{x_1..., \mathrm{x}_i\}$ of independent equally distributed examples with labeled data and unlabeled data $U = \{u_1..., u_j\}$;

– the function *f* is trained using the set of labeled data *T*;

– a forecast function *P = f(U)* is created;

– the condition is checked: $\textit{if } P_i > \alpha \textit{ then add } (x,\ f(\mathrm{x}))\ \text{to}\ T$

– retrain $f$ on $T$.

In essence, this algorithm is a wrapper over an arbitrary learning method. The goal of semi-supervised learning is to use this combined information to achieve the best grading performance results.

## 3 Case study

A computational example is implemented using three different artificial intelligence approaches to carry out analysis of the effectiveness of scientific and technical solutions, submitted to the scientific foundation since 2014 as grant applications in the sphere of fuel and energy.

The accuracy of the settings of the developed system was estimated on the basis of the training and testing samples. The training sample consists of data on 134 applications according to the criteria of systematicity and efficiency of the project, described in Table 1. The testing sample includes data on 72 scientific and technical solutions. In case of semi-supervised learning, in order to confirm the above stated hypothesis, 34 projects (out of 134 in the training set) were unlabeled.

The simulation was carried out in MatLab software package (for neuro-fuzzy inference) and in Jupyter Notebook software package (for gradient boosting and the semi-supervised learning approaches).

To evaluate the effectiveness of scientific and technical solutions four possible assessments of each of the sub-criterions are provided:

- effective planning for this sub-criterion (criterion), when the sub-criterion (criterion) fully meets all the stated requirements (S1);
- insufficiently effective planning for this sub-criterion (criterion), when the sub-criterion (criterion) meets the claimed requirements to a greater extent (S2);
- ineffective planning for this sub-criterion (criterion), when the sub-criterion (criterion) meets the stated requirements to a lesser extent (S3);
- absolutely ineffective planning for this sub-criterion (criterion), when the sub-criterion (criterion) absolutely does not meet the stated requirements (S4).

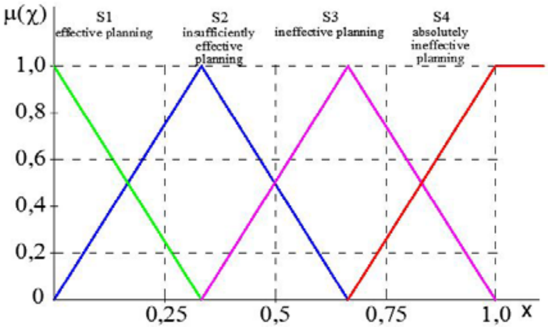

Fig. 1. The membership functions for evaluation by subcriteria

Table 1 describes the criteria and sub-criteria, on the basis of which a comparative analysis of the models developed by the authors for evaluating the effectiveness of scientific and technical solutions based on artificial intelligence approaches was performed.

Table 2 presents the results of the comparative analysis of the models developed by the authors for evaluating the effectiveness of scientific and technical solutions based on artificial intelligence methods.

From Table 2 it can be seen that with the same sizes of the training and testing samples, the average error of training and testing differs greatly for all of the considered methods. Obviously, the training error is lower than the testing error for all three methods, since, the identification accuracy is higher is carried out for the labeled data.

The testing error of the semi-supervised learning algorithm is lower, which confirms the hypothesis described at the beginning of the study.

**Table 1.** Criteria for assessing the efficiency of scientific and technical decisions

| Criteria | Subcriterion |
|---|---|
| Systematicity | Presence of the project manager, duration and functions of his participation in the project |
| | Presence of the team and all required competencies for the project implementation |
| | Duration of team work and the team experience of the project implementing |
| | Number and quality of events in which the team participated with this project |
| | Number and quality of publications on the project |
| | Quality and quantity of presentation of materials on the project |
| Effectiveness | Level of scientific significance of the expected result |
| | Evaluation of patent search and competitive technologies |
| | Level and quality of intellectual property protection |
| | Risk assessment of project implementation |

**Table 2.** Comparative analysis of the models of technical and scientific solutions efficiency assessment

| Indicators | Neuro-fuzzy inference | XGBoost | Self-training algorithm |
|---|---|---|---|
| The number of pairs in the training sample, pcs. | 134 | | |
| The number of pairs in the testing sample, pcs. | 72 | | |
| Average training accuracy, % | 0.78 | 0.88 | 0.91 |
| Average testing accuracy, % | 0.71 | 0.82 | 0.86 |

## 4 Conclusion

The task of evaluating the effectiveness of scientific and technical solutions is reduced to the task of identifying stable groups (according to a different set of variables), each of which combines objects with similar characteristics. In this case, the main problem lies not so much in the identification of belonging to certain a group, as in the definition of the objects with an atypical set of parameters for which it is not obvious either they belong to one or another group or to several groups simultaneously.

When scientific and technical solutions based on new technologies (newly introduced and/or having no analogues worldwide) are included into the training sample, the algorithms of neuro-fuzzy inference and gradient boosting decision trees significantly reduce its identification accuracy compared to the semi-supervised learning algorithm. Therefore, when analyzing solutions on “new” technologies, it is more efficient to use self-training algorithm. Obtained accuracy of identification on a testing sample (86%) is a very good result for such a small amount of training and testing samples, which also confirms the effectiveness of the developed approach.

The reported study was supported by RFBR, research project No. 17-06-00320_a.

## References

1. A.A. Zuenko, A.Ya. Fridman, B.A. Kulik. Intelligent databases: survey of results obtained within the project 4.3 of the programme № 15 of the chair of ras. URL: https://cyberleninka.ru/article/n/intellektualnye-bazy-dannyh-rezultaty-vypolneniya-proekta-4-3-programmy-15-pran (Date of circulation: 13.11.2018).
2. A.A. Barseghyan, M.S. Kupriyanov, I.I. Kholod, MD Tess, S.I. Elizarov. Analysis of data and processes: a tutorial. Publisher: SPb.: BHV-Petersburg, 2009. 512 p.
3. A.D. Gonchar Comparative analysis of databases and knowledge bases (ontologies) is applicable to the modeling of complex processes. Modern scientific research and innovation.2014. № 5. URL: http://web.snauka.ru/issues/2014/05/34325
4. Aksyonov K., Antonova A., Goncharova N. (2018) Choice of the Scheduling Technique Taking into Account the Subcontracting Optimization. Advances in Signal Processing and Intelligent Recognition Systems. SIRS 2017. Advances in Intelligent Systems and Computing, vol 678. pp. 297-304.
5. Aksyonov K., Bykov E., Aksyonova O., Goncharova N., Nevolina A. The architecture of the multi-agent resource conversion processes. UKSim-AMSS 11th European Modelling Symposium on Mathematical Modelling and Computer Simulation. Manchester, England, 20 - 22 November 2017. Pp. 61-64.
6. I. Nezhvinsky. Types of databases, their advantages and disadvantages. 2017. [Electronic resource]. URL: https://www.syl.ru/article/365055/tipyi-baz-dannyih-ih-preimuschestva-i-nedostatki (Date of circulation: 11/13/2018).
7. Databases: a tutorial. E.I. Chigarin. Samara: SSAU Publishing House, 2015. 208 p.
8. Khalyasmaa, A.I., Zinovieva, E.L., Eroshenko, S.A. Problems of Developing Decision Rules in Decision Support Systems for Assessing Innovative Solutions. Proceedings of the 3rd International Conference Ergo-2018: Human Factors in Complex Technical Systems and Environments, Ergo 2018.8443854, P. 21-24
9. Khalyasmaa, A.I., Zinovieva, E.L., Eroshenko, S.A. Formation Features of Criterias for Assessing the Feasibility of Innovative Technical Solutions. Proceedings of the 3rd International Conference Ergo-2018: Human Factors in Complex Technical Systems and Environments, Ergo 2018. 8443920, p. 16-20
10. Khalyasmaa, A.I., Zinovieva, E.L. Intelligent decision support system for technical solutions efficiency assessment. Proceedings of 2017 IEEE 2nd International Conference on Control in Technical Systems, CTS 2017. 8109537, p. 247-250
11. Khalyasmaa, A.I., Zinovieva, E.L., Eroshenko, S.A. A set of criteria for scientific and technical solutions assessment. Proceedings of 2017 IEEE 2nd International Conference on Control in Technical Systems, CTS 2017. 8109505, p. 122-125.