Control of the Reliability of Textual Information in Documents Based on Neuro-Fuzzy Identification

Jumanov Isroil Ibragimovich
Doctor of Technical Sciences, Professor, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan

Tolipov Asliddin Erkinovich
Graduate student, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan

ABSTRACT

Constructive approaches, methods, algorithms and a software package for improving the reliability of transmission and processing of electronic documents in electronic document management systems based on soft computing have been developed. The methods of parametric identification, linear, non-linear functional dependencies, identifiers and approximators of fuzzy models, as well as neural networks have been studied. Knowledge bases of fuzzy rules and databases are implemented. Modified operators of the genetic algorithm are proposed, which are tied into the structure of mechanisms for fuzzy control of the reliability of texts of electronic documents.

KEYWORDS: *reliability of information, fuzzy model, neural network, parametric, structural identification, optimization.*

Relevance of the topic. Currently, there are software packages built on the basis of modern Data Mining approaches, fuzzy computing, neural networks (NN) and neuro-fuzzy networks (NFN) [1,2]. Packages that are positioned on the market as universal, such as Intsightful Miner, MATLAB, Megaputer, Microsoft SQL Server, Oracle Data Miner, SPSS Clementine, Statistica, are focused on solving a particular problem. Approaches aimed at improving the reliability of textual information in electronic document management systems based on the noted areas remain poorly understood. Moreover, of particular scientific interest is the study of methods and algorithms for parametric and structural identification based on fuzzy inference models and a neural network (NN) for detecting and correcting errors in the texts of electronic documents (ED) [3,4].

This work is devoted to the development of methods and systems for monitoring the reliability of the transmission and processing of texts in electronic document management systems based on soft computing technologies. Methods for parametric identification via fuzzy models, NN, implementation of databases (DB) and knowledge bases (KB) are proposed. Structural identification methods based on NFN with genetic tuning have been developed [6,7].

Improving the reliability of the transmission and processing of ED texts based on fuzzy identification. The solution of identification problems begins from the assignment the following restrictions: the number of output variables is 1; in the initial identification, mathematical functions are presented explicitly by $y = F(\vec{a}, \vec{x})$, where $\vec{a}$ is the vector of function parameters; $\vec{x}$ - is the vector of input variables; y - output variable whose domain of definition is a subset of the cartesian product R_m. Moreover, R is the set of real numbers, m is the number of input variables. For data

processing, measurements are presented in the form of a matrix of size $N \times (m+1)$, where N is the number of measurements, which are given in the range from 0 to 1.

In [8], it was proved that the methods of parametric and structural identification based on a neuro-fuzzy network (NFN) make it possible to improve existing approaches. At the same time, the identification algorithms and applied in it the heuristic search mechanisms used are aimed to improve convergence of both global and local searches. In the study, the results were obtained on the basis of modeling the dependences of the input data vectors - x and the output parameter - y in the identification model with a pseudo-linear approximation.

To optimize the U functional, the number of $K = 20$ points with the maximum estimation interval $U\max$ and the initial radius of the vicinity of $R_0 = 10^3 \div 10^4$, the reduction factor for the radius of the vicinity of $r = 0{,}999$, the maximum number of iterations of $IterMax = 2^{31}$.

Structural fuzzy identification. The identification model is represented as a tree, which peculiar its own rules. The identification tree has non-terminal nodes representing elementary functions and arithmetic operations, and terminal nodes of the tree are given by variables and constants incoming in the objective function record. The developed algorithm for structural identification with genetic tuning includes the following steps [9].

Step 1. Is generated K of random initial solutions in the form of functional trees. The K values and the depth of the initial decision trees are chosen. It is recommended to form initial trees with a depth equal to 3.

Step 2. For all K solutions in the population, the value of the correspondence function FC is calculated. After that, a roulette of probabilities is formed.

Step 3. The evolutionary process begins. It continues until the generation counter reaches a critical value. Values around 100 are recommended.

Step 4. For each generation, a cycle of K iterations is organized, during which descendants of solutions in the population are generated. Two random numbers from 0 to 1 are taken. In accordance with them, using a probability roulette, two parental individuals are determined. They are crossed, then with a certain probability, occur with a new solution going on random mutation.

Experiments have shown that a rational value for the probability of mutation is an 0.3. At high values of the mutation probability, the behavior of the GA becomes unstable and the distinguished branches of evolution do not receive further development; at lower probabilities, the effect of the mutation becomes insignificant.

Step 5. After the child is defined, FC is calculated for it.

If all K descendants have FC worse than the worst solution in the population, then evolution ceases.

If not, then, firstly, K descendants are added to the population, secondly, those members of the population who have not produced a single descendant for 4 generations are removed from it, and finally, thirdly, the probability roulette is recalculated.

Step 6. The algorithm proceeds to the next iteration of evolution. At the end of the evolutionary process, the solution in the population and the current group of descendants is chosen, which has the best value of FC.

The presented algorithm with genetic principles of modeling are synthesized with computational circuits of neuro-fuzzy modeling.

An analysis of the possibilities of synthesis of NFN and GA shows that the algorithms for checking the reliability of information are effective even for models with a relatively small number of parameters.

The possibilities of methods for controlling the reliability of data of non-stationary objects are expanded due to algorithms for parametric and structural identification based on NFN with genetic tuning.

Approaches and methods for their synthesis have been developed in the structure of an intelligent information reliability control system.

In fig. 1. shows a block diagram of a software package for controlling the reliability of transmission and processing of texts of documents, an electronic document management system (EDMS) in the Windows environment - an application.

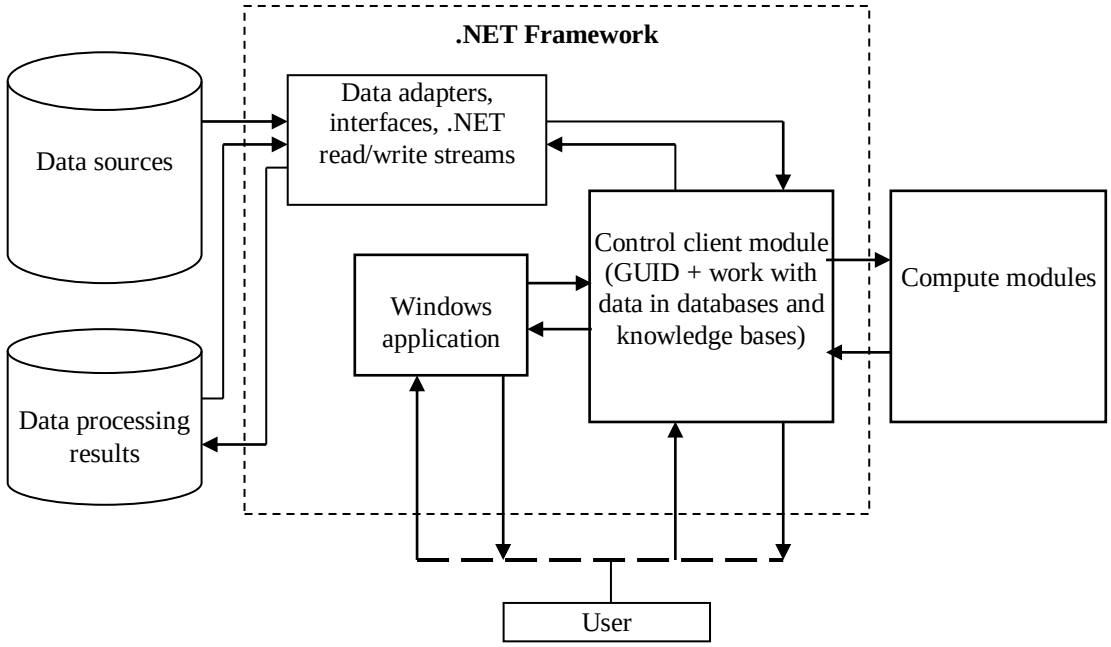

Fig. 1. Software package for improving the reliability of the transmission and processing of texts in the EDMS.

Программный комплекс повышения достоверности текстов в ЭД. Программный комплекс, функционирует как Windows-приложение **Constellation**, которое непосредственно запускает пользователь и выполняется в среде.NET Framework. После инициализации всех необходимых параметров, приложений для загрузки комплекса и проведения локальных пользовательских настроек обращается к базовому объекту Keel, объявленному в сборке **Constellation.Core**.

Software complex for increasing the reliability of texts in ED. The software package functions as a **Constellation** Windows application, which is directly launched by the user and performed in the.NET Framework environment. After initializing all the necessary parameters, the applications for loading the complex and making local user settings access the Keel base object declared in the **Constellation.Core** assembly [10,11].

Constellation.Core is the main control module that implements the user interface, where the initial data is read from the database and the results of data processing are ensured.

The main module (.NET module) is divided into two assemblies - a Windows application and a DLL library called by it, which makes it possible to increase the modularity and scalability of programs.

The modeling process in the identification complex is presented as the application one or the other operations on admissible documents.

A set of documents derived from the same experiment is built into a tree, at the root of which is the initial observation matrix of the experiment, which is considered the initial phase of the GA functioning. The different branches of this tree correspond to different branches of the evolutionary modeling process. All documents generated by the system are saved and at any time, at the request of the user, he can return to old documents and start a new branch of model building [12].

Without violating the requirements for the format and grammar of an XML-based language, we have proposed the following format extensions for representing calculation results:

1) the element Node corresponds to the node of the tree representation of the function, type is the type of the node: function is the main elementary function (the value value is one of {pow, exp, ln, sin, tg, arcsin, arctg}), operator is the operator (value $\in$ { +, –, ×, composition of functions}), variable is a variable (the value here is the index of the variable), parameters is the parameter to be determined (the value here is the index of the parameter), constant is the already known function constant (the value is the value of the constant).

If a Node element describes a node rather than a leaf, it may contain other Node elements within it.

2) support for logging preliminary data processing in the form of a procedure - parameters:

<DataPreprocessingProcedure>

<Name>< **Procedure name** ></Name>

... parameter list...

<Parameter name = "<**name** >" value = "<**meaning** >">

</ DataPreprocessingProcedure >

The results of evaluation of already built models are recorded in a similar way.

When testing the effectiveness of an intelligent information reliability control system, two groups of experiments were carried out. Experimental studies are aimed at checking the effectiveness of control over the reliability of data transmission and processing based on the use of the developed algorithms for parametric identification based on NFN, structural identification based on the synthesis of NFN with GA, as well as checking the effectiveness of the synthesized algorithms for parametric and structural identification [13].

The first group is devoted to testing the method of nonlinear parametric identification. At this stage, linear, intralinear (including polynomial) and a number of purely nonlinear models were built, which were then compared with the reference ones.

Identification was carried out by traditional methods (for linear and intralinear dependences - regression analysis, for purely nonlinear ones - numerical solution of nonlinear systems), using only local optimization methods.

It has been determined that identification algorithms built on the basis of regression models are inferior in terms of accuracy and speed of data processing to algorithms for linear and reducible models [14]. And the proposed neuro-fuzzy identification allows you to significantly expand the spheres of modeling for nonlinear dependencies and provides the required accuracy and authenticity of data processing.

The second group of experiments is aimed at obtaining the results of structural identification based on the synthesis of NFN with GA. The implementation results were compared with the approach of structural identification according to given formulas.

The efficiency of using fuzzy models, neural networks by combining with GA in identification is analyzed in tables and illustrated by graphs of dependency curves of information reliability on the following parameters: the probability of information transmission errors, the standard deviation of identification errors for various models and adaptable parameters of models of a non-stationary object.

Conclusion. Thus, the results of research and development of methodological foundations for improving the reliability of the transmission and processing of ED texts are as follows. An algorithm for the parametric identification of nonlinear models has been developed, which is less sensitive to the behavior of the function than algorithms based on methods for the numerical solution of systems of nonlinear equations. An algorithm for structural identification with genetic tuning has been developed, which allows optimizing the identification of time series and nonlinear dependencies of inputs and output when using fuzzy inference and NN models. A software package has been designed as part of modules for performing parametric identification of linear, non-linear models, regularization, as well as structural identification and non-linear dependencies of inputs and outputs of fuzzy models.

The software package has been tested on control test cases and tested in solving practical problems of checking the spelling of natural languages.

References

1. Espitia, H., Machón, I., & López, H. (2022). Design and Optimization of a Neuro-Fuzzy System for the Control of an Electromechanical Plant. Applied Sciences, 12(2), 541.
2. Borah, T. R., Sarma, K. K., & Talukdar, P. H. (2015, September). Retina recognition system using adaptive neuro fuzzy inference system. In 2015 International Conference on Computer, Communication and Control (IC4) (pp. 1-6). IEEE.
3. Rustamov, S., Mustafayev, E., & Clements, M. A. (2013, April). Sentiment analysis using Neuro-Fuzzy and Hidden Markov models of text. In 2013 Proceedings of IEEE Southeastcon (pp. 1-6). IEEE.
4. Isroil, J., & Khusan, K. (2020, November). Increasing the Reliability of Full Text Documents Based on the Use of Mechanisms for Extraction of Statistical and Semantic Links of Elements. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-5). IEEE.
5. Jumanov, I. I., & Karshiev, K. B. (2020, May). Mechanisms for optimization of detection and correction of text errors based on combining multilevel morphological analysis with n-gram models. In Journal of Physics: Conference Series (Vol. 1546, No. 1, p. 012082). IOP Publishing.
6. Jumanov, I. I., & Karshiev, K. B. (2019). Increasing information validity using natural redundancy and n-grams of natural language. International journal of advanced research in science, engineering and technology, 6(9), 10937-10945.
7. Shi, G., Wei, Q., & Liu, D. (2017). Optimization of electricity consumption in office buildings based on adaptive dynamic programming. Soft Computing, 21(21), 6369-6379.

8. Jumanov, I. I., & Xolmonov, S. M. (2021, February). Optimization of identification of non-stationary objects due to information properties and features of models. In IOP Conference Series: Materials Science and Engineering (Vol. 1047, No. 1, p. 012064). IOP Publishing.
9. Jumanov, I. I., Safarov, R. A., & Xurramov, L. Y. (2021, November). Optimization of micro-object identification based on detection and correction of distorted image points. In AIP Conference Proceedings (Vol. 2402, No. 1, p. 070041). AIP Publishing LLC.
10. Холмонов, С. М., & Абсаломова, Г. Б. (2020). Методы и алгоритмы повышения достоверности текстовой информации электронных документов. Science and world, 43.
11. Djumanov, O. I., Kholmonov, S. M., & Shukurov, L. E. (2021). Optimization of the credibility of information processing based on hyper semantic document search. Theoretical & Applied Science, (4), 161-164.
12. Akhatov, A. R. (2018). Implementation of the fuzzitic semantic hypernet based on the graphic models to provide the reliability of information in the systems of electronic document circulation. Chemical Technology, Control and Management, 2018(3), 108-113.
13. Djumanov O.I., Kholmonov S.M., Shukurov L.E. Increasing the credibility of information in the systems of electronic circulation of documents of enterprises // International Journal of Advanced Research in Science, Engineering and Technology, India, Vol. 8, Issue 2, February 2021, pp. 16940-16943.
14. Jumanov, I. I., & Kholmonov, S. M. (2020). Optimization of identification under the conditions of low reliability of information and parametric uncertainty of non-stationary objects. Chemical Technology, Control and Management, 2020(5), 104-112.