UDC 004.9; 004.8
DOI: 10.32626/2308-5916.2023-24.35-44

Ivan Izonin, Ph.D. in Engineering

Lviv Polytechnic National University, Lviv, Ukraine,
University of Birmingham, Birmingham, United Kingdom

AN ENSEMBLE METHOD FOR THE REGRESSION MODEL PARAMETER ADJUSTMENTS: DIRECT APPROACH

Intelligence analysis of tabular datasets in the field of biomedical engineering is a complex task. This is explained both by the multidimensional datasets and the complex relationships between the components of the set, and by the high price of the error in the prediction. The task becomes more difficult in the case of limited data for training, which often occurs in this field. This is due to the enormous time, material, or human resources required to collect enough data to implement training procedures with classical machine learning tools. This paper presents a new approach to solving this task. The author has developed a new ensemble method for the regression model parameters adjustments (direct approach) with the possibility of cyclically increasing the accuracy of intellectual analysis of short datasets. The basis of the method is the use of the rational fraction and two machine learning algorithms for its parametric identification. Modeling of the method's efficiency on a real-world short set of data from the field of biomedical engineering demonstrated the high accuracy of the developed method's operation. In particular, the prediction accuracy of the General Regression Neural Network was increased by more than 14% (based on the coefficient of determination. That is why the developed method can be used to solve various applied biomedical engineering tasks in the case of the need to analyze small amounts of data.

Keywords: *prediction, small data, accuracy improvement, cascade ensemble, parameters adjustments, direct approach, biomedical engineering, surrogate model.*

Introduction. Intelligent analysis of biomedical datasets by machine learning tools is a difficult task due to many features of such data, in particular [1, 2]:

- the multiparametric nature of such datasets;
- the need to take into account medical, biological, engineering, and technical features of biomedical datasets;
- complex non-linear interconnections inside of the tabular dataset;
- the presence of both numerical and categorical features;
- the presence of a large number of omissions, anomalies and outliers that occur during data collection;
- etc.

All this significantly affects the accuracy and generalization properties of machine learning (ML) tools. The task becomes more complicated when it is necessary to analyze short datasets with all the features described above [3]. Similar tasks, with a limited amount of data for the implementation of the training procedure, are increasingly occurring in various directions of scientific research in the field of biomedical engineering [4]. However, the existing ML-based tools do not provide sufficient forecasting accuracy.

This paper aims to develop a new method for the regression model parameters adjustments (direct approach) with the possibility of cyclically increasing the accuracy of intellectual analysis of short datasets. The basis of the method is the use of the rational fraction and two machine learning algorithms for its parametric identification.

Approximation by rational fractions has a number of advantages over other types of approximation [5, 6]. In particular, rational fractions can provide:

- the possibility of approximation of complex functions with high accuracy over a wide range of parameter values;
- faster convergence because they converge not only in points but also in the intervals between points, where the function values can be large;
- the possibility of effective approximation of functions that change rapidly at some points, since rational fractions can take into account the peculiarities of the behavior of the function at these points;
- a smoother interconnection between the values of the function at individual points (reduction of the Runge effect);
- avoiding emissions at the edges of the range.

Formulation of the problem. Let the table of a short biomedical dataset consist of vectors of observations $x_{i,1},\ldots,x_{i,m} \to y_i$. Let's introduce a generalized notation for the existing tabular dependency:

$$F(x_{i,1},\ldots,x_{i,m}) \to y_i . \quad (1)$$

For the case of real-world data, the multiparameter dependence (1) can be approximated with a certain accuracy by a function $f(x_{i,1},\ldots,x_{i,n})$ using the selected machine learning method. As a result, we will get the predicted value for each *i*-th vector y_i^{pred}. However, in many cases, the prediction results by the chosen machine learning method, in particular by an Artificial Neural Network (ANN), turn out to be unsatisfactory.

Therefore machine learning task is to apply a step-by-step adjustments of the response signal y_i^{pred} using known y_i as input attributes

only in the training mode and using the received parameters of the rational fraction formula to implement the prediction.

Methods. Lets us consider in detail the main steps of the developed ensemble approach for the implementation of the method when modeling short datasets in the field of biomedical engineering.

Taking into account the fact of the availability of short data samples, the approximation of the multiparameter dependence (1) is performed using a General Regression Neural Network (GRNN), for which there is essentially no training procedure. The basic steps of GRNN implementation are the following [7, 8]:

1. Calculation of the selected similarity metric between the current vector $x_{i,j}$, $i=\overline{1,n}$, $j=\overline{1,m}$ and each vector of the support (training) data sample $x_{l,j}$, $l=\overline{1,N-1}$. In the most common variant of GRNN implementation, Euclidean distance is used:

$$E_{i,l}=\sqrt{\sum_{j=1}^{m}(x_{i,j}-x_{l,j})^2}. \quad (2)$$

2. Calculation of Gaussian functions from (2):

$$G_{i,l}=\exp(-\frac{(E_{i,l})^2}{\sigma^2}). \quad (3)$$

3. Obtaining the desired response based on the following expression:

$$y_i^{pred}=\frac{\sum_{l=1}^{n-1}y_l G_{i,l}}{\sum_{l=1}^{n-1}G_{i,l}}. \quad (4)$$

Therefore, the use of (4) to approximate (1) is due to a number of advantages of this ANN for the analysis of short datasets. In particular, this type of ANN:

- does not provide for a training procedure in the classical sense of the word;
- ensures high speed of work during the analysis of short datasets;
- has the highest generalization properties among all existing ANN's topologies;
- requires searching of only one parameter of its efficient operation.

The disadvantage of the GRNN, even when analyzing short datasets (for which it is essentially intended), is low prediction accuracy. This is precisely where the problem of increasing the accuracy of its use for solving the stated task arises.

Having exact values of the output parameters y_i, predicted values y_i^{pred} and assuming that we have the necessary corresponding values $F\left(x_{i,1},\ldots,Fx_{i,n}\right)$, we apply the rational fraction formula:

$$y_i = \frac{y_i^{pred}}{1+F\left(x_{i,1},\ldots,x_{i,n}\right)}. \quad (5)$$

Performing simple transformations on (5) we can obtain:

$$F\left(x_{i,1},\ldots,x_{i,n}\right) = \frac{y_i^{pred}}{y_i} - 1. \quad (6)$$

Let's introduce a new notation:

$$z_i = \frac{y_i^{pred}}{y_i} - 1. \quad (7)$$

We will use the second machine learning algorithm to approximate dependence (6) using (7):

$$g(x_{i1},\ldots,x_{in}) \to z_i, \quad (8)$$

The main purpose of this step is to get z_i^{pred}. As a second machine learning algorithm to obtain z_i^{pred} from (8) used SVR with RBF-kernel. Such a choice is conditioned [9]:

- high speed of operation of the training algorithm;
- high prediction accuracy due to the use of RBF-kernel;
- high efficiency of data analysis of both small and large volumes;
- the possibility of working in automatic mode.

Having meaning y_i^{pred} from (4) and z_i^{pred} from (8) we can perform the first adjustment of the sought initial value $y_i^{(1)}$ (5) using the following expression [10]:

$$y_i^{(1)} = y_i^{pred} / (z_i^{pred} + 1). \quad (9)$$

Let's denote it as basic method.

After analyzing the algorithm described above, it can be seen that (9) can be further adjusted cyclically through the use of (7) and (8). That is, it is possible to obtain an additional increase in the prediction accuracy by performing a cyclic substitution of values (9) $y_i^{(t)}, t = 1,...,T$ (T are the number of iterations) in (7) instead y_i^{pred} and that calculation (9). Let us refer to this as the improved method.

In [11], it can be seen that with each level of the cascade, both training and application errors will decrease. However, the training error will

decrease until a certain point, and then it will start to increase. This will correspond to the model of optimal complexity according to the research of Prof. Ivakhnenko [12]. That is why the stopping criterion for the iterative procedure (7)-(9) will be the iteration when the user-selected error grows in the method training mode.

The structural diagram of the implementation of the developed method for the regression model parameters adjustments (direct approach) with the possibility of cyclically increasing the accuracy of the intellectual analysis of short datasets can be displayed as a cascade ensemble of two machine learning algorithms. It is shown in Fig. 1.

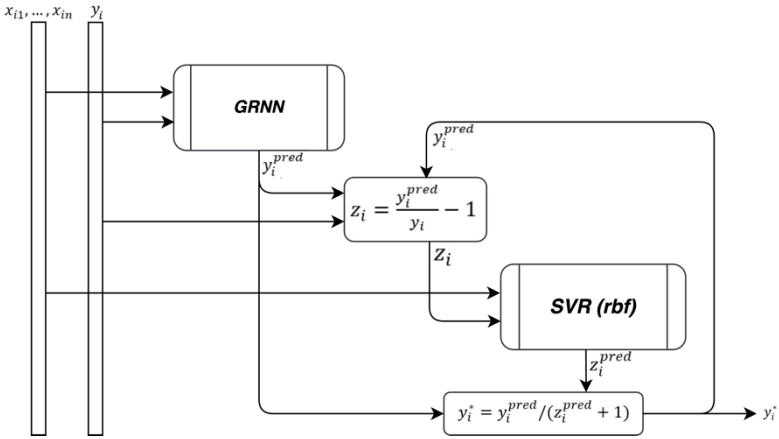


Fig. 1. *Structural and functional scheme of the developed method implementation for the regression model parameters adjustments (direct approach) with the possibility of cyclical improvement of the accuracy of intellectual analysis of short datasets*

Results and discussion. Modeling of the method was performed using a short set of biomedical data taken from [4]. The dataset contains 35 observations and 5 attributes. It is designed to predict the compressive strength of trabecular bone for patients with osteoarthritis and hip replacements [4]. The dataset is randomly divided into training and test samples (28 and 7 observations, respectively). Data were normalized using the maximum element normalization scheme in each column. Experimental studies were performed using the author's software in the Python language. A GRNN was chosen as the first machine learning algorithm. It requires the setting of only one parameter, the smooth factor, which was selected in this paper using the differential evolution optimization method in the interval [0.001, 10]. The optimal value of the smooth factor is equal to 0.0848561538566178. SVR with RBF kernel was chosen as the second machine learning algorithm in the developed cascade scheme. The operat-

ing parameters of this method are as follows: gamma='scale', coef0=0.0, epsilon=0.001, max_iter=–1.

The composition of the developed method for the regression model parameters adjustments (direct approach), proposed above, was used to simulate the work of both the basic and the improved (with the additional use of the iterative procedure) algorithmic implementation of the method. To determine the highest accuracy of the improved method, the paper runs the method with different numbers of iterations (from 1 to 100). To visualize the results of this step, Fig. 2 shows the change in the value of the RMSE when the number of iterations of the method changes from 1 to 30 (the number of iterations greater than 30 significantly increases the value of the error and will reduce the informativeness of Fig. 2).

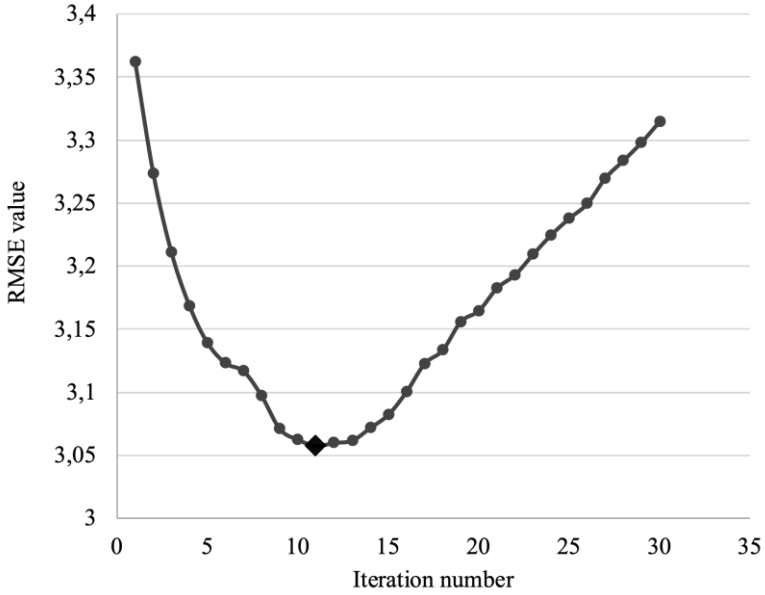


Fig. 2. *RMSE values for different numbers of iterations of the improved version of the developed method*

As can be seen from Fig. 2 (marked with a red color), the smallest error value of the method was obtained when using 11 iterations. With a further increase in the number of iterations, as can be seen from the graph, the accuracy of the method decreases significantly. The same results were obtained for all other efficiency indicators used in the paper. That is why these iterations are chosen as stopping criteria of the method.

Table 1 summarizes the numerical values of various performance indicators as for basic variant of the implementation of the method [10] as

well as the version, improved in this paper, due to the use of an iterative procedure for the regression model parameters adjustments (direct approach). A description of the performance indicators used in this paper is given in [11].

Table 1

Performance indicators of both researched algorithmic implementations of the developed cascade method for the regression model parameters adjustments (direct approach)

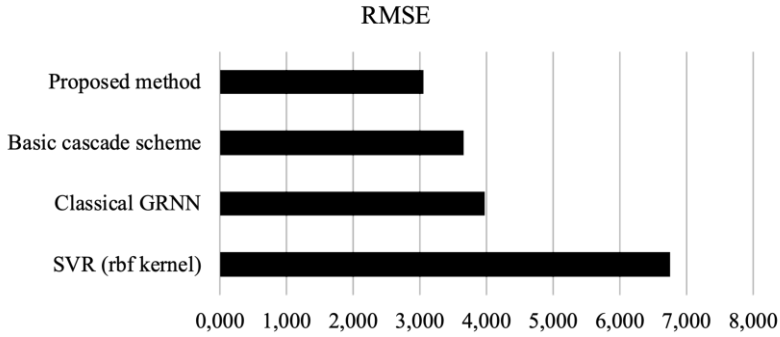
Performance indicators	**Basic method**	**Improved method**
MaxE	6,37	4,76
MAE	2,83	2,51
MSE	13,33	9,35
MedAE	2,65	3,05
RMSE	3,65	3,06
MAPE	0,22	0,2
R^2	0,69	0,79
Training time (seconds)	0,004	0,013

As can be seen from Table 1, the improved version of the cascade method provides significantly higher prediction accuracy. In particular, the use of the iterative procedure provided a reduction of the RMSE error by more than 16% and a reduction of the maximum residual error by more than 25% compared to the basic version of the developed method. Despite this, the improved version shows a significantly higher time required to implement the training procedure. However, since we are talking about the analysis of short datasets, this drawback can be eliminated.

To evaluate the effectiveness of the developed method (both of its algorithms), the paper compares its accuracy with both machine learning methods that form it: GRNN and SVR with RBF kernel.

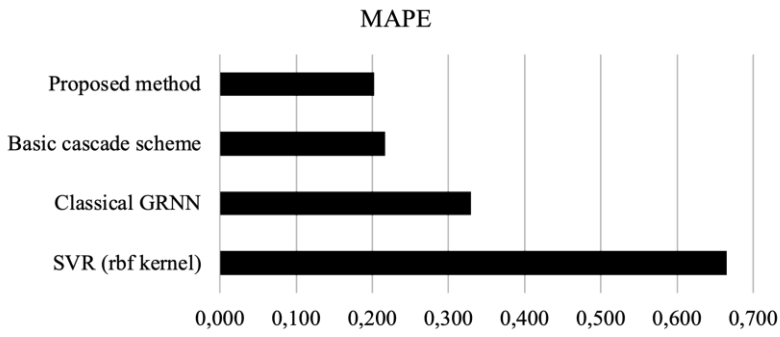
Fig. 3 shows the results of comparison of all studied methods based on RMSE and MAPE.

As can be seen from Fig. 3, SVR with RBF kernel demonstrates the lowest accuracy in terms of both efficiency indicators. Significantly better results were obtained due to the use of GRNN. However, the R2 = 64% value obtained for this method is not a satisfactory result. The accuracy of both algorithmic implementations of the developed method, which is designed to increase the accuracy of GRNN, is quite high. In particular, R2 = 69% for the basic version of the developed method implementation [10], and R^2 = 78%, for the version of the method implementation improved in this paper. Such a high increase in accuracy (more than 14%) allows the use of an improved method when solving applied tasks in biomedical engineering in the case of the need to analyze short datasets.



	SVR (rbf kernel)	Classical GRNN	Basic cascade scheme	Proposed method
■ RMSE	6,759	3,977	3,652	3,058

a)



	SVR (rbf kernel)	Classical GRNN	Basic cascade scheme	Proposed method
■ MAPE	0,665	0,330	0,216	0,203

b)

Fig. 3. *Comparison of the prediction accuracy of all studied methods based on: a) RMSE; b) MAPE*

Conclusions. This paper considers the currently relevant problem of intelligent analysis in the case of short datasets. The use of classical machine learning tools does not ensure the adequacy of predictions in the case of a limited training sample. To eliminate this shortcoming, this paper describes the developed cascade method for the regression model parameters adjustments (direct approach) based on the use of the rational fraction formula. Experimental modeling on a short set of biomedical data demonstrated a significant increase in the accuracy of GRNN when solving the

stated task due to the use of the cyclic version of the developed cascade method. This ensures the possibility of its use in practice.

Acknowledgment. The British Academy's Researchers at Risk Fellowships Programme supports this research.

References:

1. Perova I., Bodyanskiy Y. Adaptive human machine interaction approach for feature selection-extraction task in medical data mining. *IJC*. 2018. P. 113-119. DOI: 10.47839/ijc.17.2.997.
2. Krak I., Kuznetsov V., Kondratiuk S. and other. Analysis of Deep Learning Methods in Adaptation to the Small Data Problem Solving. *Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making*. Cham: Springer International Publishing, 2023. Vol. 149. P. 333-352. DOI: 10.1007/978-3-031-16203-9_20.
3. Dolgikh S. Modeling of Small Data with Unsupervised Generative Ensemble Learning. *CEUR-WS.org*. 2022. Vol. 3302. P. 35-43.
4. Shaikhina T., Khovanova N. A. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine*. 2017. Vol. 75. P. 51-63. DOI: 10.1016/j.artmed.2016.12.003.
5. Holub A. P., Lysenko L. O. Апроксиманти типу Паде деяких класів функцій кількох змінних. 2017. Vol. 69. № 5.
6. Pelekh Y., Kunynets A., Beregova H., Magerovska T. Methods for solving the initial value problem with a two-sided estimate of the local error. *Fìz.-mat. model. ìnf. tehnol*. 2021. № 33. P. 88-92. DOI: 10.15407/fmmit2021.33.088.
7. Вітинський П. В., Ткаченко Р. О., Ізонін І. В. Ансамбль мереж GRNN для розв'язання задач регресії з підвищеною точністю. *Науковий вісник НЛТУ України*. 2019. Vol. 29. № 8. DOI: 10.36930/40290822.
8. Duda P., Jaworski M., Rutkowski L. Online GRNN-Based Ensembles for Regression on Evolving Data Streams. *Advances in Neural Networks – ISNN 2018*. Cham: Springer International Publishing, 2018. P. 221-228. DOI: 10.1007/978-3-319-92537-0_26.
9. Medykovskvi M., Pavliuk O., Sydorenko R. Use of Machine Learning Technologys for the Electric Consumption Forecast. *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*. 2018. P. 432-435. DOI: 10.1109/STC-CSIT.2018.8526617.
10. Izonin I., Tkachenko R., Shcherbii O. and other. An Approximation Cascade Scheme via Rational Fractions for Biomedical Data Analysis. *2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT)*. Lviv: IEEE, 2023. P. 1-4. DOI: 10.1109/CSIT61576.2023.10324122.
11. Izonin I., Tkachenko R., Holoven R. and other. SGD-Based Cascade Scheme for Higher Degrees Wiener Polynomial Approximation of Large Biomedical Datasets. *MAKE*. 2022. Vol. 4. № 4. P. 1088-1106. DOI: 10.3390/make4040055.
12. Ivakhnenko A. G. Development of models of optimal complexity using self-organization theory. *International Journal of Computer and Information Sciences*. 1979. Vol. 8. № 2. P. 111-127. DOI: 10.1007/BF00989666.

АНСАМБЛЕВИЙ МЕТОД УТОЧНЕННЯ ПАРАМЕТРІВ МОДЕЛІ РЕГРЕСІЇ: ПРЯМИЙ ПІДХІД

Інтелектуальний аналіз табличних наборів даних у галузі біомедичної інженерії являється складним завданням. Це пояснюється як багатовимірними наборами даних і складними взаємозв'язками між компонентами набору так і висока ціна помилки у прогнозуванні. Задача стає складнішою у випадку обмеженості даних для навчання, що часто виникає у цій галузі. Це пов'язано з величезними часовими, матеріальними чи людськими ресурсами, необхідними для збору достатньої кількості даних для реалізації процедур навчання класичним інструментарієм машинного навчання. У цій статті представлено новий підхід до розв'язання цієї задачі. Автором розроблено новий ансамблевий метод уточнення параметрів моделі регресії (прямий підхід) із можливістю циклічного підвищення точності інтелектуального аналізу коротких наборів даних. В основі методу покладено використання формули раціонального дробу та двох алгоритмів машинного навчання для її параметричної ідентифікації. Моделювання роботи методу на реальному короткому наборі даних з галузі біомедичної інженерії продемонстрував високу точність роботи розробленого методу. Зокрема, автору вдалося підвищити точність прогнозу нейронної мережі узагальненої регресії на більші ніж 14% (на основі коефіцієнту детермінації. Саме тому, розроблений метод можна використовувати для розв'язання різноманітних прикладних задач біомедичної інженерії у випадку необхідності аналізу даних малих обсягів.

Ключові слова: *прогнозування, нейронна мережа узагальненої регресії, малі дані, підвищення точності, каскадний ансамбль, прямий підхід, біомедична інженерії, сурогатна модель.*

Отримано: 19.11.2023