Jestr

Journal of Engineering Science and Technology Review 15 (2) (2022) 69 - 78

JOURNAL OF
Engineering Science and Technology Review

Research Article

www.jestr.org

Prediction of GAS Concentration of BA-ELM based on Mapreduce

Yongliang Han*

Xi'an Research Institute Co. Ltd., China Coal Technology and Engineering Group Corp, Xi'an 710077, China

Received 2 March 2022; Accepted 30 May 2022

Abstract

Gas concentration monitoring is an important instrument for coal mine gas disasters, and gas concentration prediction is especially important for improving mine safety. This study proposed a gas concentration prediction method of bat algorithm (BA) optimized extreme learning machine (ELM) based on Mapreduce to accurately and quickly predict mine gas concentration. The parameter optimization algorithm of BA-ELM was proposed after analyzing the characteristics of the BA and the ELM algorithm, and the single-step prediction method of gas concentration was carried out utilizing Mapreduce. The accumulation error of gas concentration prediction in advance was analyzed on this basis. Then utilizing real-time error compensation a parallel prediction method for gas concentration was proposed, and the accuracy of this model was verified through simulation. Results show that the running time of the BA-ELM prediction method based on MapReduce is 21.36s, and the efficiency increases by 6.13 times more efficient than the BA-ELM method. When compared to the same period last year, the mean absolute error, mean relative error, and root mean square error decreases by 48%, 48%, and 49% respectively. The average absolute error of the parallel prediction model based on real-time error compensation is 0.030%, the average relative error is 8.050%, and the root mean square error is 0.034% at the 12th step. Meanwhile, the accuracy of a parallel prediction method of gas concentration based on real-time error compensation is higher than the model without real-time error compensation. The proposed method provides a good reference for gas concentration prediction and gas disaster warning.

Keywords: Mapreduce, Gas concentration, BA-ELM, Real-time error compensation, Intelligent prediction

1. Introduction

The concentration of mine gas is closely correlated with the safety production of coal mines, consequently, it is particularly crucial to accurately and effectively predict mine gas concentration and master its change rule [1]. The gas concentration prediction leverages the dynamically acquired monitoring data as the research object and excavates its internal regularity characteristics, so as to master its future change rules [2]. Using a variety of forecasting methodologies, a multitude of studies [3-5] have analyzed the dynamic change and development tendency of gas concentration and produced a wealth of findings.

With the growth of time and the increase of gas data samples, it is possible to predict gas concentration by utilizing a single model follows the variation characteristics of gas concentration to a certain extent. However, due to the theoretically inherent defects of the model, it is difficult to simultaneously meet the specifications for gas concentration prediction accuracy, operation efficiency, and prediction step length.

Therefore, scholars proposed to construct a combined prediction model of gas concentration by coupling various optimization methods [6-7]. However, there is still a significant discrepancy between the model's prediction accuracy, computational efficiency, leading steps, and the actual demand. Therefore, it is of great theoretical and practical value to probe how to construct an appropriate model for computing and improving the overall characteristics of the model.

Consequently, by combining the bat algorithm (BA) with the extreme learning machine (ELM) algorithm, the study proposes a multi-step gas concentration prediction method of BA-ELM based on Mapreduce, with the objective of predicting gas concentration more promptly and accurately and providing value for accident prevention references.

2. State of the art

At present, Scholars all over the world have conducted extensive research on gas concentration prediction and put forward a series of methods. Lv Pin et al. [9] constructed a dynamic gas concentration prediction model based on the gray theory and analyzed its application with examples. Liu et al. [10] constructed a model for predicting gas concentration based on the gray system and statistical theory. Due to the introduction of concepts such as gray derivative and background value, the whitening differential equation was transformed into the form of background variables, resulting in the low precision of the model. Liu et al. [11] proposed a multi-factor long short-term memory (LSTM) gas concentration prediction model, that integrated multi-source gas data, and analyzed the gas concentration change trend through tunnel wind speed, temperature, CO concentration, and historical gas concentration data. However, the model's generalization ability and prediction accuracy were of little value. Lai et al. [12] obtained the power exponential gray action that changed over time by optimizing the traditional gray model, and then proposed a power exponential gray gas concentration prediction model based on integrated learning. However, there was a major

*E-mail address: yaqingyxgs@163.com
ISSN: 1791-2377
doi:10.25103/jestr.152.10

error and making it difficult to be directly adopted in practical production. Shan et al. [13] combined the adaptive artificial immune system with particle swarm optimization to establish a multi-parameter parallel dual-adaptive artificial immune system (AIS)-particle swarm optimization (PSO) algorithm gas concentration prediction model, which improved the prediction accuracy, but the generalization ability was relatively limited. Li [14] proposed a multi-variable adaptive weighted least squares support vector machine gas prediction model based on an enhanced chaotic particle swarm optimization algorithm, which enabled the multi-step prediction of gas concentration. However, the maximum number of predicted steps was 5, rendering it more challenging to guarantee the prediction accuracy when the number of predicted steps increased. On the basis of the randomness and timing of gas concentration data, Zhao et al. [15] proposed a prediction method based on auto regressive and moving average model (RIMA)+generalized auto regressive conditional heteroskedasticity (GARCH), with high prediction accuracy but low calculation efficiency due to the lengthy calculation time. Slezak [16] established a prediction model for learning multi-sensor data sets based on feature extraction of sliding window and feature subset set of a rough set, with a lengthy operation time of 19 minutes per group of data. Dey [17] introduced a deep learning network to propose the t-distributed stochastic neighbor embedding (SNE)-variational auto encoder (VAE)-bidirectional (bi)-LSTM, with a small prediction error, but the impact of time series was not considered, so its generalization ability was insufficient. Grodzicka [18] proposed a linear equation to predict gas concentration one day in advance. The concentration of methane was predicted [19] at the sensor location up to 10 m in front of the longwall face and at the longwall outlet. BA is a swarm intelligence optimization algorithm proposed by Professor Yang from The University of Cambridge inspired by the bat echolocation behavior [20]. Compared to conventional optimization algorithms, BA is characterized by less parameter adjustment, strong search-ability, rapid convergence speed, and a simple structure; therefore it has been applied in numerous fields including engineering control, fuzzy intelligence, and fault identification. Mohammad [21] utilized the BA-artificial neural network (ANN) method to predict the shear strength of fiber reinforced polymer (FBR)-reinforced concrete beams, and in comparison, to the particle swarm optimization algorithm, the prediction accuracy was significantly enhanced. However, the mean absolute error is 13.34% and the root mean square error is 26.34%, indicating that the prediction results continue to deviate significantly. Bui [22] proposed a machine learning method combining least square support vector classification and bat algorithm to predict landslides, and the prediction accuracy was higher than that of support vector machine (SVM) and ANN methods. Based on adaptive neural fuzzy inference system (ANFIS) , harris hawks optimization (HHO), support vector regression (SVR) , and BA, Paryani [23] used ANFIS-HHO, ANFIS-BA, SVR-HHO, and SVR-BA hybrid models, respectively, to generate landslide risk maps, respectively, and the results indicated that SVR-HHO featured the highest accuracy.

The above results focus on how to improve the prediction accuracy and operation efficiency. However, there are very few studies on the prediction accuracy, operation efficiency, and generalization ability based on the processing of massive gas monitoring data. Spatial-temporal correlation analysis of gas concentration is carried out firstly in this study. Based on the efficient global optimization ability of the BA algorithm, Mapreduce was employed to improve the parallelization processing ability of the model. Then the accumulation error of gas concentration prediction in advance was integrated to establish the BA-ELM gas concentration prediction method based on Mapreduce.

The remainder of this study is organized as follows. Section 3 describes the principle of combining BA with ELM and Mapreduce, analyzes the accumulation error of gas concentration advanced prediction, and proposes a gas concentration prediction method of BA-ELM based on Mapreduce. In section 4, a variety of methods for predicting and analyzing mine monitoring data are discussed. Section 5 provides a summary of this study and pertinent conclusions.

3. Methodology

3.1 BA optimization algorithm and ELM principle

(1) BA optimization algorithm

The basic principle of BA is as follows:

1) Bat individuals are able to sense and distinguish the differences between food, prey and background obstacles, etc., using ultrasound.

2) When a bat flies at a certain speed V_i at a certain position X_i, and searches for prey with a frequency of f_i, a variable wavelength of λ and a loudness of A_0, it can select the pulse wavelength and frequency emitted by itself according to the distance between the target and itself.

3) The variation of the sound emitted by Bat is irregular, and it is assumed that the variation is gradually reduced from the maximum (positive) A_0 to the minimum $A_{\min}$.

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta \tag{1}$$

$$V_i^t = V_i^{t-1} + (X_i^{t-1} - X^*) f_i \tag{2}$$

$$X_i^t = X_i^{t-1} + V_i^t \tag{3}$$

where t is the current iteration number, and β is the random variable evenly distributed between (0, 1).

$$X_{new} = X_{old} + \varepsilon A^t \tag{4}$$

where $\varepsilon \in [-1,1]$ is a random number and $A^t =< A^t >$ refers to the average loudness obtained by all Bats in a specified time period.

$$A_i^{t-1} = \alpha A^t \tag{5}$$

$$\gamma_i^t = \gamma_i^0 \left[1 - \exp(-\gamma t)\right] \tag{6}$$

where α and γ are an constant, and for any $0 < \alpha < 1$ and $\gamma > 0$, then $A_i^t \to 0, \gamma_i^t \to \gamma_i^0$. When $t \to +\infty$, $\alpha = \gamma$ =0.9. In the iteration process, the loudness and pulse rate of each Bat are set as the same value, and they are changed together by Equations (5) and (6). The termination condition of BA search is generally set according to either the number of iterations that reaches the maximum value or the precision of the search value that meets the requirements.

(2) Basic principle of ELM

ELM is a learning algorithm for single hidden layer feedforward neural network, which consists of three layers: input layer, hidden layer, and output layer.

Assuming there are Q learning samples $\{(x_l, y_l)\}_{l=1}^{Q}$, and $x_l \in R^r$, $y_l \in R^\psi$. If the number of hidden layer neurons is M, then the standard form of single-hidden layer feedforward neural network is as follows:

$$\sum_{j=1}^{M} \boldsymbol{\beta}_j \varphi(\boldsymbol{\omega}_j x_l + b_j) = f_M(x) \tag{7}$$

where $\boldsymbol{\omega}_j = [\boldsymbol{\omega}_{j1}, \boldsymbol{\omega}_{j2}, \cdots, \boldsymbol{\omega}_{jr}]^{\mathrm{T}}$ is the connection weight vector of the input layer and hidden layer; $\boldsymbol{\beta}_j = [\boldsymbol{\beta}_{j1}, \boldsymbol{\beta}_{j2}, \cdots, \boldsymbol{\beta}_{j\psi}]^{\mathrm{T}}$ is the connection weight vector of the hidden layer and output layer; $\varphi(\cdot)$ is the activation function, and the sigmoid function is usually selected as the model activation function. b_j is the bias of the jth node in the neuron node of the hidden layer, and $f_j = [f_{j1}, f_{j2}, \cdots, f_{j\psi}]^{\mathrm{T}}$ is the output vector.

According to the zero error approximation principle[25], there are b_j, $\boldsymbol{\omega}_j$, and $\boldsymbol{\beta}_j$ that convert equation (7) into the following one:

$$\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{Y} \tag{8}$$

$$\boldsymbol{H}(\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_M, \boldsymbol{b}_1, \cdots, \boldsymbol{b}_M, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_Q)$$

where $= \begin{bmatrix} \varphi(\boldsymbol{\omega}_1 \boldsymbol{x}_1 + \boldsymbol{b}_1) & \cdots & \varphi(\boldsymbol{\omega}_M \boldsymbol{x}_1 + \boldsymbol{b}_M) \\ \vdots & \vdots & \vdots \\ \varphi(\boldsymbol{\omega}_1 \boldsymbol{x}_Q + \boldsymbol{b}_1) & \cdots & \varphi(\boldsymbol{\omega}_M \boldsymbol{x}_Q + \boldsymbol{b}_M) \end{bmatrix}_{Q \times M}$,

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1^{\mathrm{T}} \\ \boldsymbol{\beta}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\beta}_M^{\mathrm{T}} \end{bmatrix}_{M \times \psi}, \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1^{\mathrm{T}} \\ \boldsymbol{y}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{y}_M^{\mathrm{T}} \end{bmatrix}_{M \times \psi}.$$

If the output threshold and weight are given randomly, $\boldsymbol{H}$ is the determined output matrix. At this point, the network training is finally transformed into the problem of solving the least square solution of $\boldsymbol{\beta}$, namely:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{H}^{+}\boldsymbol{Y} \tag{9}$$

where $\boldsymbol{H}^{+}$ is the Moore-Penrose generalized inverse of the output matrix of the hidden layer $\boldsymbol{H}$.

3.2 BA-ELM gas concentration prediction method

(1) BA optimizes ELM model parameters

BA possesses effective global optimization capability. By sending ultrasonic waves to simulate the dynamic searching and preying behavior of bats, the optimization and preying processes are coupled and connected, and the problem of solving the optimization target is transformed into the optimization problem of Bat preying position. In order to obtain the optimal w and b and improve the superiority of the ELM network, Random w and b are generated in the BA-ELM network.

1) Initialization of Bat parameters. The number of Bats is set as N, the initial position and initial velocity of the qth Bat is S_q and V_q, respectively, the pulse transmitting frequency range is $[\lambda_{\min}, \lambda_{\max}]$, the initial pulse rate is r_0, the pulse rate enhancement coefficient is γ, the attenuation coefficient for the loudness is α, the loudness scope is $[A_0, A_{\min}]$, the number of iterations is L, the training sample size is U, the sample size of the prediction set is V, the number of ELM hidden layer nodes is P, then each Bat contains optimization parameters w and b, and the mathematical expression of the qth Bat is:

$$N_q = [w_{11}, \cdots, w_{1U}, \cdots, w_{P1}, \cdots, w_{PU}, b_1, b_2 \cdots, b_P] \tag{10}$$

2) Suppose that the optimal position of the Bat population is S_N^*, and the fitness function is expressed by the mean square error of the prediction set:

$$\mathrm{Fit} = \sqrt{(\sum_{j=1}^{V} \left\| \sum_{q=1}^{P} \beta_q * (w_q * S_j + b_q) - t_q \right\|_2^2) / V} \tag{11}$$

where $j = 1, 2, \ldots, V$, V is a temporary variable.

3) According to the distance of food detected by echolocation, then the flight speed, pulse emission frequency, and position of Bat q in the υth iteration are updated as follows:

$$\lambda_q = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})\sigma \tag{12}$$

$$V_q^{\upsilon} = V_q^{\upsilon-1} + (S_q^{\upsilon} - S^*)\lambda_q \tag{13}$$

$$S_q^{\upsilon} = S_q^{\upsilon-1} + V_q^{\upsilon} \tag{14}$$

where υ is the current iteration number, and the random variable is $\sigma \in [0\ \ 1]$.

4) The Bat randomly generates a new position $S_{qM}^{'}$ around its selected solution. If the fitness $F(S_{qM}^{'})$ of Bat $S_2 = \{x_i(t) \mid i = 1, 2, \ldots, N+1\}$ is better than its own extreme fitness $F(S_{qM})$, then the position S_{qM} is updated.

$$S_{qM}^{'} = S_{qM} + \mu A^{\upsilon} \tag{15}$$

where $\mu \in [-1, 1]$ is a random number, A^{υ} is the average pulse loudness of all Bats in the current iteration number, M is the dimension of the search space.

$$\begin{cases} S^* = S_{\mathrm{qM}} \\ A_q^{\upsilon+1} = \lambda A_q^{\upsilon} \\ r_q^{\upsilon+1} = r_q^0[1 - \exp(\gamma\upsilon)] \end{cases} \tag{16}$$

5) If the fitness $F(S_{qM})$ of Bat q in the iterative process is better than the fitness $F(S^*)$ of global optimal $F(S^*)$, then the new position, pulse loudness, and rate of Bat q are updated as Equation (16).

6) If the iteration reaches the preset maximum value or meets the accuracy requirements, then the iteration will terminate. At this point, the optimal solution S^* is obtained, and the fitness function is $F(S^*)$, then the corresponding parameters w and b are the optimal parameter values.

Otherwise, the iterative search continues according to equations (12) to (16) until the termination conditions are satisfied.

(2) BA-ELM gas concentration prediction method

1) Spatio-temporal correlation analysis of gas concentration [25]. If the target monitoring site of gas concentration prediction is X_T, then the data sample set S_1 is jointly constructed for N monitoring sites with strong correlation obtained by improved dynamic clustering method. The data sample set can be expressed as follows:

$$S_1 = \{x_i(t) \mid i = 1,2,\ldots,N+1; t = 1,2,\ldots,n\} \tag{17}$$

where $N+1$ is the number of related sites and n is the length of gas time series of each site.

2) Assuming that the embedded dimension of phase space reconstruction is m, and the time delay is τ, the multi-variable phase space reconstruction method [25] is used to reconstruct the data of the sample set S_1, and reconstructed sample set S_2 can be obtained after the reconstruction of the sample set S_1. The reconstructed sample set can be defined in the following way:

$$S_2 = \{x_i(t) \mid i = 1,2,\ldots,N+1\} \tag{18}$$

$$x_i(t) = \{x_i(t), x_i(t-\tau), \ldots, x_i(t-(m-1)\tau)\} \tag{19}$$

$$t = \{(m-1)\tau + 1, (m-1)\tau + 2, \ldots, n\} \tag{20}$$

Thus, a phase point in a $M = (N+1) \times m$ dimensional phase space can be formed.

3) The reconstructed data samples are taken as training samples. BA-ELM parameter optimization method is used to learn the training samples and obtain the optimal weight matrix w and hidden layer threshold b of ELM.

4) The completed training model is applied used to calculate the prediction samples and the prediction results are output.

3.3 Research on BA-ELM gas concentration single-step prediction method based on Mapreduce

(1) Parallelization analysis of BA intelligent optimization algorithm

Individual bats are autonomous and cooperate with each other. For an individual Bat, the foraging process is essentially parallel, and BA itself possesses the attribute of parallelization.

(2) Parallel BA optimization algorithm based on Mapreduce

Fig. 1 shows the Parallel BA optimization process based on Mapreduce. This algorithm divides the initialized Bat population into R subpopulations, with each subpopulation's iterative process handled by a Mapreduce process, and each process independently completing the entire optimization process of serial BA. The specific process for evaluating fitness is as follows: the Map function is invoked to evaluate the fitness of each Bat individual in each subpopulation, and the fitness value of each Bat individual in each subpopulation is obtained and combined. Then whether the iteration termination conditions are met is determined. If the termination conditions are met, the optimal position will be output; otherwise, the intermediate results generated by the Map function will be transmitted to the Reduce function for reduction processing to complete the update of Bat parameters and generate new subpopulations. Under the new subpopulation condition, the Mapreduce process is further implemented to optimize parameters until the termination condition is met.

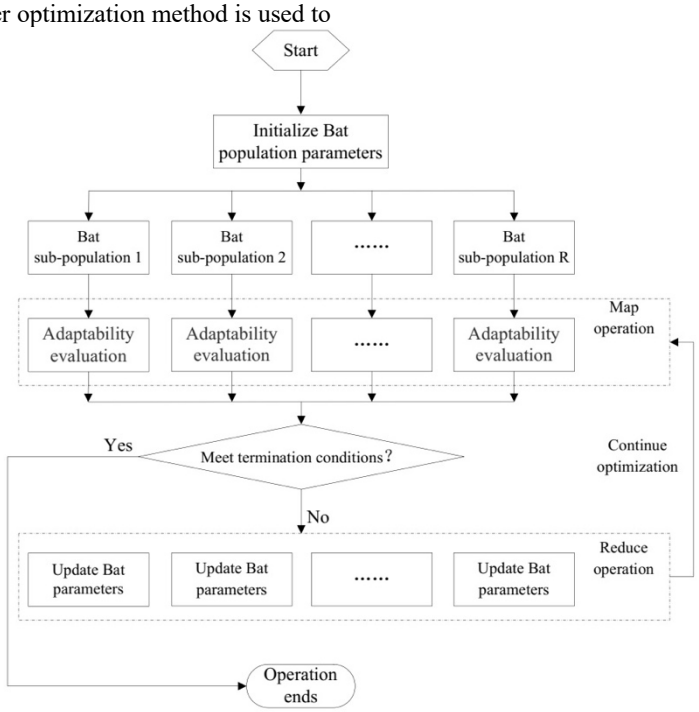

Fig. 1. Parallel BA optimization process based on Mapreduce

(3) BA-ELM gas concentration prediction method based on Mapreduce

1) Design idea of BA-ELM algorithm based on Mapreduce

The Map function is utilized to complete the ELM network parameter training by relying on the node where it is located. The ELM input layer and hidden layer connection weights and hidden layer thresholds are output locally if the training accuracy requirements are met or the maximum number of iterations has been reached. The Reduce function is employed to reduce the intermediate output results of each Map function, thereby the optimal ELM connection weight

and threshold are obtained. The design idea of the algorithm is as follows.

a. Map function

After setting the BA population parameters and ELM basic parameters, the basic parameters are passed to each slave node, and the fitness of each node is then calculated according to the network structure of ELM. If the fitness does not meet the accuracy requirement, the BA algorithm is adopted to optimize the input weights and threshold of the ELM, and the input weights and threshold value are adjusted through repeated iterations until the required termination conditions are satisfied. Among them, the local data on the node exists as key-value pairs of <key, value>. The Map function repeatedly and iteratively adjusts the input weights and thresholds of ELM through fitness evaluation, and the output result is an intermediate key-value pair consisting of input weights and thresholds meeting <key1, value1>.

b. Reduce function

The < key1, value1> intermediate result key values generated by the Map function are forwarded to the Reduce function for reduction processing. After reduction operation, input weights and thresholds of ELM network whose output result is <key2, value2> are generated. In addition, the updated global network input weights and thresholds are saved in the distributed system file HDFS of the Hadoop platform and used as the initial values of the Map function in the subsequent Mapreduce process for iterative optimization.

After repeated iterations, the Mapreduce process is carried out. If the change range of input weights and thresholds of the BA-ELM algorithm is small or meets the maximum iterative value, and the fitness evaluation has met or is very close to reaching accuracy requirements, then the training is complete and the final network input weights and thresholds can be used as input parameters of the ELM network for the following prediction.

2) Realization of BA-ELM gas concentration prediction algorithm based on Mapreduce

First of all, the data set of gas concentration is prepared to obtain the basic data of gas concentration prediction; Secondly, the BA-ELM network structure based on Mapreduce is trained, Mapreduce is applied to program the model framework; the parameters of ELM are optimized by BA, and the optimal input weights and thresholds of ELM are obtained. Thirdly, the trained BA-ELM model is adopted to predict gas concentration in advance. The gas concentration prediction process of BA-ELM based on Mapreduce is shown in Figure 2.

a. Preparation of gas concentration data sample set

After completing the data preprocessing of the original gas concentration data set, assuming that the target monitoring site of gas concentration prediction is X_T, and parallel dynamic clustering method is applied to get N monitoring sites near the target site with a strong correlation. Then all site monitoring data are extracted to form the data sample set $S_1 = \{x_i(t) \mid i = 1,2,\ldots,N+1; t = 1,2,\ldots,n\}$, where $N+1$ are several related sites, and n is the gas time series length for each site. Assuming that the embedded dimension of phase space reconstruction is m, and the time delay is τ, the multi-variable phase space reconstruction method is used to reconstruct the data of the sample set S_1, and S_2 can be obtained after the reconstruction of the sample set S_1, thus forming a $M = (N+1) \times m$ multi-dimensional sample space. The reconstructed multidimensional data samples are stored in HDFS, a distributed file system of the Hadoop platform, so as to provide basic data for the gas concentration prediction model.

b. Training of BA-ELM network structure based on MAPReduce

The specific training steps are as follows:

The parameters of the limit vector machine are set. The master node (Job Tracker) randomly generates the initial limit vector machine parameters, determines the input weights and thresholds of the limit vector machine randomly within the interval [-1,1], uploads them to the Hadoop platform, and saves them in the distributed file system HDFS.

Initialize the Bat population, including setting the number of Bats as N, the pulse transmitting frequency range as $[\lambda_{\min}, \lambda_{\max}]$, the initial pulse rate as r_0, the pulse rate enhancement coefficient as γ, the loudness attenuation coefficient as α, the loudness range as $[A_0, A_{\min}]$, the iteration times as L, and the training sample size as U.

Population segmentation: N individuals of the Bat population are divided into R sub-populations and assigned to each sub-node (Task Tracker), and the basic parameters of initializing the Bat population are obtained for each sub-node.

Map operation: the Map function of each node is invoked for each node, the fitness of each Bat population is evaluated to determine the fitness value of each individual Bat in a population, and the master node incorporates various fitness values. At this point, it judges whether the termination conditions are satisfied: when satisfied the ELM optimal input weights and thresholds will be output, otherwise, the next operation will be carried out.

Reduce operation: the Reduce function receives the resultant location information generated by the Map function transmitted from the primary node, reads the intermediate result file, updates the parameters of Bat individuals, and iterates to execute a new round of the Mapreduce process.

Iteration of the Mapreduce process: repeatedly perform Map and Reduce operations. If the change range of input weights and thresholds of the BA-ELM algorithm is small or reaches the iterative maximum, the optimal input weights and thresholds of ELM can be obtained.

c. Build an advanced single-step prediction model of gas concentration based on BA-ELM

The optimal input weights and thresholds obtained through model training are then incorporated into the BA-ELM model, and the BA-ELM prediction model is used to make an advance prediction of gas concentration. The structure diagram of the prediction model is shown in Fig. 2.

Assuming $f(x)$ is the gas concentration prediction model based on BA-ELM, the single-step prediction modeling method is adopted, then the advanced prediction of the one-step gas concentration is:

$$x_{n+1} = f(x_n, \cdots, x_{n-M+1}) \tag{21}$$

$$f(x) \rightarrow \left[\begin{array}{cccc|c} x_1 & x_2 & \Lambda & x_M & x_{M+1} \\ x_2 & x_3 & \Lambda & x_{M+1} & x_{M+2} \\ O & O & O & O & O \\ x_{n-M} & x_{n-M+1} & \Lambda & x_{n-1} & x_n \\ \hline x_{n-M+1} & x_{n-M+2} & \Lambda & x_n & \end{array}\right] \longrightarrow x_{n+1}$$

Fig. 2. Structure diagram of gas concentration advance single-step prediction

3.4 Real-time error compensation method for gas concentration multi-step prediction

(1) Cumulative error analysis of gas concentration ahead prediction

The prediction error is the difference between the actual value and the predicted value, which contains the characteristics of the internal dynamic change of the gas concentration time series. Therefore, effectively analyzing the accumulation error of gas concentration prediction and constructing a reasonable real-time error compensation strategy for gas concentration are essential for improving the prediction accuracy and reliability.

When performing the multi-step advance prediction of gas concentration time series, the prediction value of the early stage is adopted as the input sample of the later stage to realize the gradual prediction, and the prediction error is bound to accumulate and expand in the transmission process before and after. To further analyze the impact of cumulative error, the variance decomposition concept is introduced into the square loss function, and the time series model of gas concentration monitoring value is assumed as follows:

$$x_{k+1} = g(X_{k-l+1}^{k}) + \xi(0,\sigma^2) \tag{22}$$

where ξ is the random error and σ^2 is the variance. Suppose that the observed value of rolling prediction window length h is $(y_1, y_2, \cdots, y_h)$, then:

$$y_1 = x_t\,,\ \ y_2 = x_{t+1}\,, \cdots,\ \ y_h = x_{t+h-1} \tag{23}$$

Assuming that the corresponding value generated by the time series model $g(X)$ is $(y_1^*, y_2^*, \cdots, y_h^*)$, then:

$$y_i = y_i^* + e(0,\sigma^2) \tag{24}$$

Further assuming that the h step-ahead prediction value of the prediction model $f(x)$ is $(f_1, f_2, \cdots, f_h)$, then the mean square error (MSE) at the kth step of prediction can be expressed as:

$$MSE(k) = E\left[(y_k - f_k)^2\right] \quad k \in [1,h] \tag{25}$$

Then the MSE of the prediction model $f(x)$ at each step of prediction can be expressed as:

$$MSE(k) = (E(f_k) - y_k^*)^2 + E\left[(f_k - E(y_k^*))^2\right] + E\left[(y_k - y_k^*)^2\right] \tag{26}$$

where each error to the right of the equal sign is successively represented as the square deviation between the actual model and the prediction model, the variance of the prediction model, and the random error arising from time series noise.

As using BA-ELM to carry out rolling prediction error analysis is complicated, the prediction model is assumed as follows:

$$x_t = \alpha x_{t-1} + \beta x_{t-2} + e \tag{27}$$

where the mean of the random error e is 0 and the variance is σ^2.

Assuming that the coefficients a_1 and a_2 of the model can be accurately estimated, the first-step prediction output of the progressive prediction can be obtained, as follows:

$$y_1 = \alpha x_{t-1} + \beta x_{t-2} + e_1 = f_1 + e_1 \tag{28}$$

The mean square error corresponding to the first-step-prediction is:

$$MSE(1) = E\left[(y_1 - f_1)^1\right] = E(e_1^2) \infty \sigma^2 \tag{29}$$

In this way, the mean square errors of the prediction of the second and third steps are:

$$MSE(2) \infty (a_1^2 + 1)\sigma^2 \tag{30}$$

$$MSE(3) \infty (a_1^4 + a_1^2 + a_2^2 + 1)\sigma^2 \tag{31}$$

According to the preceding equation, the error gradually accumulates and expands as the number of predicted steps increases. Therefore, constructing a real-time dynamic error compensation strategy based on the predicted value is an essential measure in enhancing the prediction accuracy of the model.

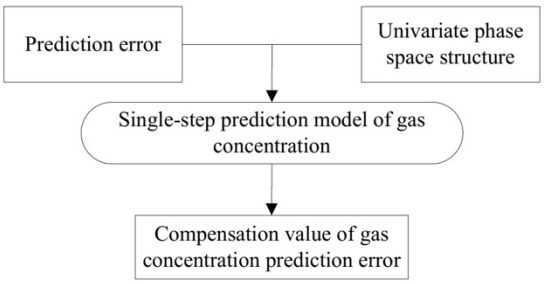

Fig. 3. Flow chart of Real-time error compensation model of gas concentration

(2) Real-time error compensation model for gas concentration prediction

The cumulative error analysis reveals that the prediction error of the model is composed of self error and random error, both of which are predictable. Therefore, there is room for improvement in the prediction accuracy of the model. With the gradual advance of later prediction, the predicted value of gas concentration will lag behind the actual value due to the cumulative influence of prediction errors. The objective of a real-time error compensation strategy for gas concentration is to use the prediction model to correct and compensate the prediction results of gas concentration in real time. Based on this, the MapReduce-based BA-ELM gas concentration single-step prediction method proposed in Section 3.3 is adopted as the real-time compensation model for prediction error, the prediction error of gas concentration is taken as the fundamental data source of the model, and the single-variable phase space reconstruction theory is adopted to reconstruct the prediction error data. The real-time error compensation model is applied to learn the reconstructed data, so as to correct and compensate each prediction result in real time. The flow chart of the real-time error compensation model of gas concentration is shown in Fig. 3.

(3) Gas concentration multi-step prediction method based on real-time error compensation

Based on the research on gas concentration single-step prediction and real-time error compensation, a multi-step gas concentration prediction method based on real-time error compensation was proposed. The specific steps are as follows:

1) The historical monitoring data of gas concentration are preprocessed according to the spatio-temporal correlation analysis method [25] to complete the abnormal processing of historical monitoring data, including data abnormality, data missing, as well as data noise processing; The dynamic parallel clustering method is then employed to acquire the data sets of each monitoring station with a strong correlation with the target monitoring station. The final step is to reconstruct the data group composed of four monitoring stations via the multivariable phase space reconstruction method.

2) The reconstructed data serve as the training sample for the BA-ELM gas concentration single-step prediction based on Mapreduce. The optimal parameters of the model are obtained through training, while the prediction error and advance prediction value of the training sample are obtained with the trained model.

3) The same prediction method in step 2) is adopted to construct the real-time error compensation model of the prediction results. The single-variable phase space reconstruction is undertaken to the prediction error of the training sample, and concurrently the reconstructed data are utilized as the real-time error compensation model of the training sample to obtain the network structure of the optimal model and acquire the target gas concentration monitoring compensation error predicted in advance. At this juncture, the advanced one-step predicted value of the target monitoring site plus the compensation error to produce the advanced one-step predicted value.

4) Univariate phase space reconstruction is carried out for each gas monitoring site, and the final single-step-ahead prediction value of each monitoring site is obtained according to the same method in 2) and 3).

5) According to the rolling multi-step prediction strategy, the one-step-ahead prediction values of $N+1$ monitoring sites are taken as the input samples of the BA-ELM gas concentration single-step prediction model based on MapReduce, and the two-step prediction values of the target monitoring station are obtained by applying the BA-ELM gas concentration single-step prediction model based on MapReduce. Meanwhile, the real-time error compensation model is utilized to calculate the advanced two-step prediction value, and the sum of the two values is the final prediction value.

6) The advanced H-step prediction value of the target monitoring station can be obtained directly by iteration cycle.

4 Result Analysis and Discussion

The gas concentration monitoring data samples of a mine for 31 consecutive days were collected (the sampling interval was 1min), and the data of each monitoring station were preprocessed. For instance, No.1735 gas monitoring site was selected as the target site (prediction site), and four other roadway monitoring sites with a strong spatio-temporal correlation with No.1735 gas monitoring site were obtained by the dynamic parallel clustering analysis method, namely No.1734, 1732, 1713 and 1724 monitoring sites sequentially.

First and foremost, the original monitoring data from the gas sensor were preprocessed to minimize the impact of data loss and data noise. Secondly, a dynamic parallel clustering analysis method was used to obtain a roadway gas concentration monitoring site with a close correlation to the target monitoring site to jointly form a multivariate data set, and the data of the multivariate gas concentration was reconstructed with the phase space reconstruction theory. Finally, the reconstructed data were imported into the BA-ELM prediction model based on Mapreduce to obtain the advanced gas concentration prediction. Applying the Java assembly language, BA and ELM intelligent algorithms, and the Mapreduce programming framework, the gas concentration prediction model based on real-time error compensation had been developed. Then seven computers were designated for the Hadoop data analysis platform, and a distributed cluster was built in the LAN. The parameters of the example were set as follows: the number of parallel computing nodes is 6, the number of Bat populations is $N=200$, the pulse emission frequency is $\lambda \in [0,2]$, the initial pulse rate is $r_0=0.1$, the pulse rate enhancement coefficient is $\gamma=0.9$, the loudness attenuation coefficient is $\alpha=0.9$, the loudness range is $\alpha=0.9$, and the iteration times are $L=400$.

Firstly, the time delay of No. 1735 target site was obtained by multi-variable phase space reconstruction, and the embedded dimension was $m=9$. The reconstruction parameters of the target monitoring site were taken as the standard to reconstruct the data of other monitoring sites, and a total of 44631×36 dimension data sample space was generated. The first 44571 groups were selected as the training sample set and the rest as the test sample set. When using the trained model for prediction, the current input prediction sample was added to the training sample set for each advanced prediction, and the previous sample in the sample set was eliminated to keep the length of the training sample set unchanged before and after prediction, so as to achieve the advance prediction of No.1735 gas station successively. According to the multi-step advance prediction process of gas concentration based on real-time error compensation, the prediction results of this model were compared with those of the BA-ELM prediction model and Mapreduce BA-ELM gas concentration prediction model (real-time compensation without error), then the mean absolute error (MAE), mean relative error (MRE), root mean square error (RMSE), and running time (T) were used to evaluate the model performance. The prediction results are shown in Fig. 4.

As can be seen from Fig. 4 and Table 1, the two models showed different performance characteristics due to different prediction processes and mechanisms: as for the prediction error, both the two prediction models have high prediction accuracy on the whole, which reflected that after ELM parameters are optimized by BA, the impact of ELM model parameters on prediction accuracy is effectively avoided. Compared with the BA-ELM serial prediction method, the mean absolute error (MAE), mean relative error (MRE), and root mean square error (RMSE) of the BA-ELM parallel prediction method based on Mapreduce are reduced by 48%, 48%, and 49%, respectively, indicating that after the parallel BA based on Mapreduce is adopted to optimize ELM parameters, the obtained model has high prediction accuracy and model generalization ability. In terms of operation efficiency, when the number of operation nodes is 6, the running time of the parallel prediction method based on Mapreduce BA-ELM is 21.36s, while the single-point serial prediction time is 152.31s, the acceleration ratio is

T_s =152.31/21.36≈7.13, and the operation efficiency is improved by 6.13 times, which reflect that the training efficiency of the parallel prediction method based on Mapreduce is substantially improved, thus effectively enhancing the overall computing efficiency of the model.

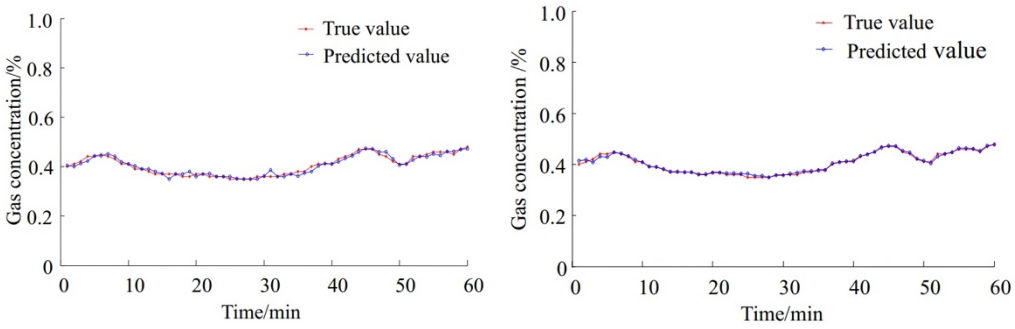

Fig. 4. Single-step prediction results of gas concentration. (a). Single step prediction based on BA-ELM. (b). BA-ELM single-step prediction based on Mapreduc

Table 1. Performance evaluation of model prediction

Model type	MAE/%	MRE/%	RMSE/%	T/s
BA-ELM single-step prediction	0.0079	0.0195	1.00	152.31
Mapreduce+BA-ELM single-step prediction	0.0041	0.0101	0.51	21.36

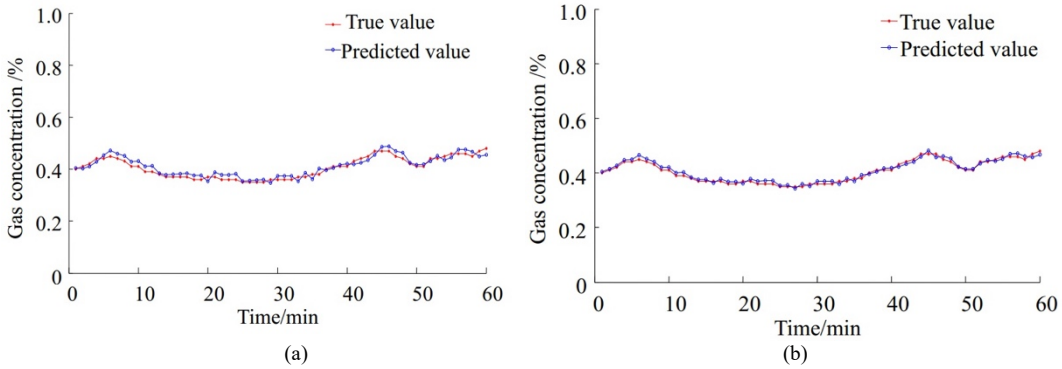

Fig. 5. Prediction results of gas concentration 7 steps ahead. (a). Prediction based on Mapreduce 7 steps ahead. (b). Prediction based on real-time error compensation 7 steps ahead

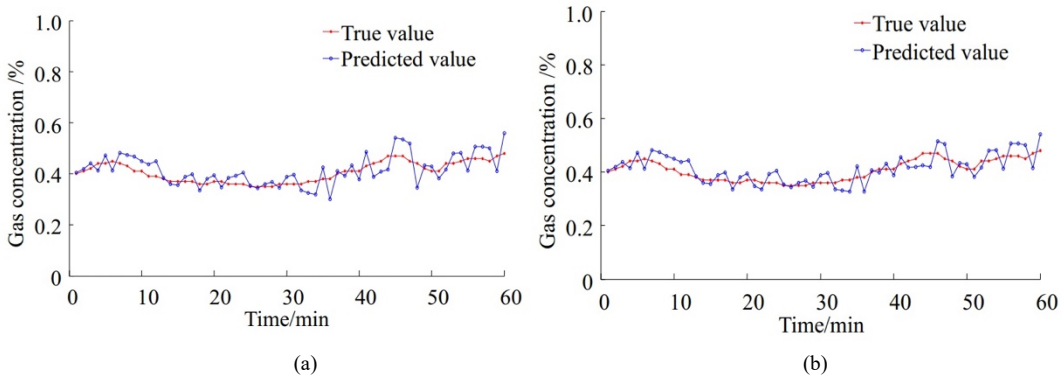

Fig. 6. Prediction results of gas concentration 12 steps ahead. (a). Prediction based on Mapreduce 12 steps ahead. (b). Prediction based on real-time error compensation 12 steps ahead

The high accuracy and efficiency of the single-step prediction model serve as the premise and guarantee for multi-step ahead prediction. However, when multi-step ahead prediction is utilized, two significant issues must be resolved: error accumulation and predictable step size For this reason, the gas concentration parallel prediction model based on real-time error compensation is compared with the parallel prediction model without real-time error compensation to verify the effectiveness of the real-time error compensation model and determine the number of ahead steps of the model. Tables 2 and 3 show the prediction performance evaluation of the BA-ELM model based on Mapreduce and Mapreduce+BA-ELM model based on error compensation. Fig. 5 and Fig. 6 show the prediction results of gas concentrations of 7 and 12 steps ahead.

Table 2. Prediction performance evaluation of BA-ELM model based on Mapreduce

Advanced step size	MAE/%	MRE/%	RMSE/%	T/s
7 steps	0.014	3.435	0.015	21.42
9 steps	0.018	4.518	0.020	21.47
10 steps	0.027	7.663	0.033	21.51
12 steps	0.035	9.537	0.041	21.58

Table 3. Prediction performance evaluation of Mapreduce+BA-ELM model based on error compensation

Advanced step size	MAE/%	MRE/%	RMSE/%	T/s
7 steps	0.008	2.136	0.009	39.46
9 steps	0.013	3.265	0.014	39.57
10 steps	0.022	5.405	0.025	40.12
12 steps	0.030	8.050	0.034	40.31

According to Tables 2-3 and Figures 5-6, the prediction accuracy of the two models decreases gradually as the number of leading steps, and the accuracy deteriorates as the number of the predicted steps increases; the worse the accuracy is, primarily because the prediction error accumulates and expands with the number of steps increases. By comparing the prediction results of the two models, it can be observed that the MAE, MRE and RMSE of the parallel prediction model based on real-time error compensation are lower than those of the parallel prediction model without real-time error compensation for any step length, reflecting that the real-time error compensation model is an effective measure to suppress the prediction error transmission. Regarding the selection of predictable step size, a relative error of less than 9% is adopted as the evaluation benchmark, and the MAE, MRE, and RMSE differences of the parallel prediction model based on real-time error compensation at the 12th step are 0.030%, 8.050%, and 0.034%, respectively, with high prediction accuracy. Therefore, after repeated trials, the predictable step size of the model can be considered as 12, namely, the prediction results can satisfy the actual requirements of the project within 12 steps; Under the condition of the same step size, the prediction results of the parallel prediction model with real-time error compensation are superior to than those of the parallel prediction model without real-time error compensation, confirming that adopting a real-time error compensation model is an effective measure to improve the prediction step size of the model in the process of leading multi-step prediction. Due to the increasing complexity algorithm during operation, the parallel prediction model based on real-time error compensation is inefficient in terms is insufficient in terms of operational efficiency.

5. Conclusions

This study proposes a Mapreduce-based BA-ELM gas concentration prediction method to improve the accuracy, generalization ability, operation efficiency, and prediction step size of gas concentration prediction, as well as to accurately and rapidly grasp the variation tendency of mine gas concentration. The method is based on the basic principles of BA and ELM and combines Mapreduce and real-time error compensation. The concept was demonstrated using specific cases. The following conclusions could be drawn:

(1) The Mapreduce+BA-ELM single-step prediction method is significantly more efficient than the BA-ELM single-step prediction method.

(2) After the parallel BA optimization of ELM parameters based on Mapreduce, the model exhibits high prediction accuracy and model generalization ability.

(3) In the leading multi-step prediction process, the error length of the parallel prediction model based on real-time error compensation is smaller than that of the parallel prediction model without real-time error compensation. Adopting the real-time error compensation model is an effective measure for improving the prediction step size of the model.

The study presented a new understanding of gas concentration prediction by combining theoretical research with practical production, and the resulting model established guiding significance for the prevention of mine gas accidents. Due to the preliminary establishment of the real-time error compensation method, the prediction accuracy would decrease after 12 steps as the number of steps increases. The method of real-time error compensation will be further studied to increase the number of predicted steps based on improving the accuracy of the prediction thereby gaining a deeper understanding of the law of gas concentration prediction.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 51704148, 52174117 and 52004117).

References

1. Zhu. Z., Zhang H., Han. J., "Prediction of Coal and Gas Outburst Based on PCA-BP Neural Network". *China Safety Science Journal*, 23(4), 2013, pp.45-50.
2. Li. L., "Research of coal and gas outburst prediction based on data mining technology". *Master thesis of Liaoning Technical University*, China, 2011, pp.1-2.
3. Gao. W., Zhang. Y., Gao S., Prediction of gas content in unmined area based on BP neural network. *Coal of Shanxi*, 39(1), 2020, pp.77-80.
4. Zhang. Z., Zhu. Q., Li Q., "Prediction of Mine gas concentration in heading face based on Keras long short term memory network". *Safety and Environmental Engineering*, 28(1), 2021, pp.61-67,78.
5. Zhang. Y. Guo. H., Tu. H., "Gas concentration prediction based on neural network with random hidden weight". *Computer Engineering and Science*, 41(4), 2019, pp.699-701.
6. Zhang. X., Lai. W., Xue. S., Application of MI and SVM in coal and gas outburst prediction. *China Safety Science Journal*, 31(1), 2021, pp.5-80.
7. Xu. M., Gao. S., Cao. Y., "Forecast of gas content based on gray theory and multiple regression analysis". *Safety in Coal Mines*, 49(9), 2018, pp.211-214.
8. Li S., Li M., Pan. S., "Research on prediction model of gas concentration based on RNN in coal mining face". *Coal Science and Technology*, 48(1), 2020, pp.33-38.
9. Lyu. P., Ma. Y., Zhou. X., "Research and application on dynamic forecasting model of gas consistence in top corner". *Journal of China Coal Society*, 31(4), 2006, pp.461-465.
10. Liu. G., Duan. F., "A study on prediction of coalmine gas concentration based on comparisons between gray system and statistics". *Computer Development and Application*, 22(6), 2009, pp.1-3.
11. Liu. Y., Yang. C., "LSTM gas concentration prediction model based on multiple factors". *Journal of safety science and technology*, 18(1), 2022, pp.108-113.
12. Lai X., Xia. Y., Zheng. W., "Improved grey prediction of gas concentration sequence based integrated learning". *Journal of safety science and technology*, 17(7), 2021, pp.16-21.

13. Shan. Y., Gao. Z., "Study on double adaptive AIS-PS0 based on model for gas concentration soft-sensing". *Computer Simulation*, 37(1), 2020, pp. 338-393.
14. Li D., Sun. Z., Li M., "AWLSSVM gas prediction research based on chaotic particle swarm optimization". *Safety in Coal Mines,* 51(8), 2020, pp.193-205.
15. Zhao. M., He. A, Qu. S., "Research on time series prediction method of gas data on fully mechanized mining face". *Industry and Mining Automation*, 45(6), 2019, pp.80-85.
16. Slezak, D., Grzegorowski, M., "A framework for learning and embedding multi-sensor forecasting models into a decision support system". *Information Sciences*, 451-452, 2018, pp.112-133.
17. Dey, P., Saurabh, K., Kumar, C., "t-SNE and variational auto-encoder with a bi-LSTM neural network based model for prediction of gas concentration in a sealed-off area of underground coal mines". *Soft Computing*, 25, 2021, pp. 14183–14207.
18. Grodzicka, A., Badura, H., Shaidurova, N., Sentyakov, K., Sviatskii, V., One-Day "Prognoses of Methane Concentrations for the 102 Longwall in the 325/1 Seam in the "W" Coal Mine Operating in a Continuous System". *New Trends in Production Engineering*, 3, 2020, 169-185.
19. Badura, H., Zmarzły, M., Trzaskalik, P., Korshunov, A. I., "Study of the Accuracy of Methane Concentration Forecasts in the Area of Longwall Outlet B-3 in Seam 407/1 in KWK "Borynia-Zofiówka" Ruch Zofiówka". *New Trends in Production Engineering*, 3, 2020, pp. 120-130.
20. Yang X. *Nature-Inspired Metaheuristic Algorithms*, Luniver Press , 2010. pp. 1-10.
21. Mohammad, N., Babak, A., Alireza, L., "Predicting shear strength in FRP-reinforced concrete beams using bat algorithm-based artificial neural network". *Advances in Materials Science and Engineering*, 2021, 2021, pp.1-13.
22. Bui, D. T., Hoang, N. D., Nguyen, H., Tran, X. L., "Spatial prediction of shallow landslide using Bat algorithm optimized machine learning approach: A case study in Lang Son Province, Vietnam". *Advanced Engineering Informatics*, 42, 2019, pp.1-14
23. Paryani, S., Neshat, A., Pradhan, B., "Improvement of landslide spatial modeling using machine learning methods and two Harris hawks and bat algorithms". *The Egyptian Journal of Remote Sensing and Space Sciences*, 24, 2021, pp.845-855.
24. Huang, G.B., Siew C.K., "Extreme learning machine with randomly assigned RBF kernels". *International Journal of Information Technology*, 11(1), 2005, pp.16-24.
25. Han, Y. L., "*Research on intelligent prediction of mine gas based on big data*". Doctoral Dissertation of Liaoning Technical University, China, 2016, pp.34-46.