International Journal of Innovative
Computing, Information and Control
Volume **20**, Number **2**, April **2024**

ICIC International ©2024 ISSN 1349-4198
pp. 575–590

UNBALANCED BIG DATA CLASSIFICATION BASED ON IMPROVED RANDOM FOREST ALGORITHM

Xin Zheng[1,*] and Li Huang[2]

[1]Artificial Intelligence Department
[2]Information Engineering College
Jiangxi University of Technology
No. 115, Ziyang Avenue, High-Tech Zone, Nanchang 330098, P. R. China
huangli@jxut.edu.cn
*Corresponding author: zhengxin@jxut.edu.cn

Received May 2023; revised September 2023

Abstract. *Big data analytics has developed rapidly in recent years and data mining has been a positive driver for development in all areas, but data in many areas is grossly unbalanced, and there are still many limitations to current research on classifying big data. To solve this problem, the study uses the K-means algorithm based on class distinction to approximately reduce the dimensionality of the data, and the untracked Kalman filter (UKF) algorithm with an adaptive traceless Kalman filter (Sage-Husa) to reduce the noise of the data. The noise-reduced and dimension-reduced data were obtained to improve the random forest algorithm (K-U-S-H-RF). However, during the study of classifying low-dimensional unbalanced data using K-S-H-RF, it was found that the random forest algorithm did not take account of the actual step-by-step of the data set and was not effective in classifying the data. For this reason, the study introduced cost sensitivity, cost error calculation for decision trees as well as voting. Random forest is parallelized with MapReduce idea to achieve optimum of K-S-H-RF. Then the study constructs an imbalanced big data classification model based on improved random forests. The model can effectively classify unbalanced big data and provide a new path for big data application in more fields, which has a positive effect on the development of the big data era.*
Keywords: Improved RF algorithm, Unbalanced data, Classification recognition

1. **Introduction.** The exponential rise of data in the network as a result of the quick advancement of information technology has steadily drawn attention to big data applications. Big data is a body of information that has grown beyond what can be acquired, stored, managed, and analyzed by conventional technology. Data removal for traditional relationships also includes unprocessed data such as videos, web pages, documents, audio, and emails that do not have a structural or semi-structural schema [1]. In the age of big data, data mining, which is the process of extracting meaningful information from incomplete, enormous, noisy, ambiguous, and random data, is a significant technology [2]. Many of these fields have seriously unbalanced data. The number of samples belonging to different categories is very different. Traditional machine learning methods usually use the global classification accuracy as the training target, and the performance in unbalanced data mining is not ideal [3]. The classification problem is an important problem in data mining. Imbalanced data classification can be widely used in credit card fraud detection, medical diagnosis, spam classification, information retrieval, text classification, smoke image detection, mechanical equipment failure prediction and other application

DOI: 10.24507/ijicic.20.02.575

fields [4]. There are many algorithms for classifying unbalanced data. At present, algorithms such as logistic regression, K-nearest neighbor classifier, random forest (RF), and Adaboost are relatively mature. One of the RF algorithms uses the concept of ensemble learning to integrate many trees, and the basic unit is the decision number [5]. In order to improve the RF algorithm and create a new algorithm, the study employs the K-means algorithm based on class distinction to roughly reduce the dimensionality of the data. It then combines the UKF algorithm with an adaptive traceless Kalman filter (Sage-Husa) to reduce data noise (K-U-S-H-RF). The study introduces cost-sensitive and MapReduce ideas to parallelize the RF and achieves optimization of K-U-S-H-RF. Synthesizing this, the study constructs an imbalanced big data classification model based on the improved RF. It is able to classify imbalanced big data effectively, provide technical support for data mining in various fields, and has a positive effect on the development of the big data era.

The main contributions in this study are as follows: 1) A dimensionality reduction method for calculating class differentiation is proposed; 2) An improved K-U-S-H-RF method is proposed to estimate costs more accurately; 3) K-U-S-H-RF algorithm was designed in parallel by using MapReduce programming idea. The main research contents in this study are the following. Firstly, aiming at the redundancy among features in high-dimensional unbalanced data set and the few ignored strong correlation features, K-means is used for feature clustering. The classification and differentiation of various characteristics of cocoon were calculated. Secondly, a cost-sensitive RF classification algorithm is proposed. The cost function is constructed according to the actual distribution of unbalanced data set, and the weight is introduced into the cost function. Thirdly, using the idea of MapReduce, parallel design of cost-sensitive RF algorithm is carried out. Triple parallel design is carried out in the modeling process of base classifier.

2. **Related Works.** In the era of big data, enterprises in various fields need to mine and analyze data to formulate development strategies. However, these massive data often have characteristics such as imbalance and high dimensionality. Scholars are increasingly interested in learning how to store, extract significant information from data, and classify data. To address the issue of an inaccurate classification learning algorithm caused by an imbalanced sample set of data used in medical diagnosis applications, Han et al. [6] accurately divided the minority samples according to the position of the minority samples, and synthesized the minority class (Mi C) samples by using the distribution-sensitive sample synthesis method, and proposed a method: distribution sensitive imbalanced data oversampling algorithm. To process the fault data of motor rolling bearing diagnosis, Hang et al. [7] used principal component analysis (PCA) to reduce the data dimension, and then used the oversampling technique (SMOTE) algorithm to classify and synthesize the unbalanced data. Based on the financial data of A-share listed companies in Shanghai Stock Exchange, Cong [8] used the Hellinger distance-based random forest algorithm (HDRF) and HDDT-based classifier to form an integrated method to mine and analyze a large amount of financial data. Pant et al. [9] found that dual support vector machines (TWSN) are often used for learning in imbalanced datasets, but are not suitable for large datasets. In search of a more suitable method for big data processing, a twin neural network (Twin NN) architecture for learning from large unbalanced datasets was proposed. Guha and Veeranjaneyulu [10] used the fuzzy c-means algorithm to sample the classifier, and used the clustering method and GA-based artificial neural network to predict the probability of company bankruptcy by considering different factors. In order to address the problem that traditional data analysis algorithms found difficult to effectively mine features and automatically produce accurate data when evaluating enormous multi-source

heterogeneous data at the limit of rotation, Yan et al. [11] proposed an automatic learning model using deep belief network (DBN), and accurately identify the rotor imbalance fault in the fault state.

The random forest (RF) algorithm is an integrated learning algorithm based on a decision tree-based learning model. It has high classification performance, which is widely used in research in various fields. Balli et al. [12] employed smart watch sensor data and RF algorithms to gather and classify human mobility data, and timely avert hazardous situations in tracking users. Their research target was the detection of human motion data in the sectors of medical care, fitness, and geriatric care. The security risk index and variable weights were calculated using the RF algorithm by Chen et al. [13]. They then used the method's model parameters to evaluate and forewarn against the security risks associated with large-scale group activities. Finally, the evaluation effect was verified through experiments. Georganos et al. [14] proposed a new RF geographic implementation, geographic random forest (GRF), for RF algorithms that cannot solve the spatially heterogeneous processes. Ao et al. [15] proposed a linear random forest algorithm to solve the problems of high cost and time-consuming direct measurement of formation properties. The benefits of the linear random forest technique in the well logging regression modeling were validated through investigation of its application in the model, and provided a more reasonable way for the further practice of the well logging regression model. Liang et al. [16] took African swine fever (ASF) outbreak data and WorldClim database meteorological data as the research object, combined the best feature selection method with RF, studied the relationship between ASF outbreak and weather, and predicted the outbreak. Based on this, a new ASF outbreak prediction model was constructed. Fitri et al. [17] used the RF algorithm to analyze the sentiment of the dynamic data published in social media, and compared the analysis effects of the Naïve Bayesian algorithm and the decision tree.

It can be learned from the above that there are many studies on classification processing and RF of unbalanced big data with wide applications, but the improvement of traditional RF algorithms applied in big data applications is not perfect. The study uses K-means and UKF algorithms with Sage-Husa to optimize the RF algorithm to obtain the improved K-U-S-H-RF and introduces cost-sensitive and MapReduce ideas to parallel the random forest. The improved K-U-S-H-RF algorithm is applied to imbalanced big data processing to classify high-dimensional imbalanced data and improve the recognition rate of few classes, which is of great significance to data mining and analysis in various industrial fields.

3. Construction of Big Data Classification Model Based on Improved K-U-S-H-RF.

3.1. Data processing and classification model construction.

There is redundancy between features in high-dimensional unbalanced datasets. It is very easy to ignore the problem of strong correlation features of Mi C [18]. The research uses the K-means algorithm and class discrimination to reduce the dimensionality of high-dimensional features and screen out effective and low-dimensional subset data. The K-means clustering algorithm first needs to select the initial cluster centres, then classify all data points, and finally keep adjusting the cluster centres in an iterative loop [19]. The K-means algorithm performs approximate dimensionality reduction on high-dimensional features as shown in Figure 1.

First, mark most categories and few categories separately, and characteristic attribute K is selected as the cluster center, and all attributes are divided into the cluster center

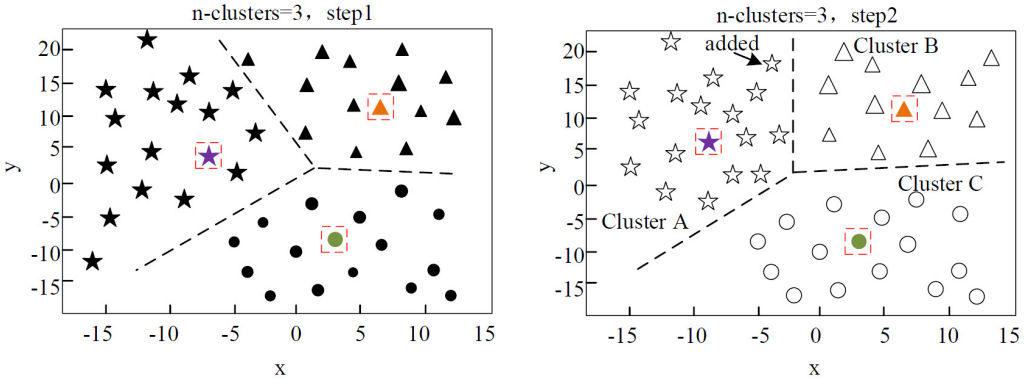


FIGURE 1. K-means clustering algorithm reduces the dimension of high-dimensional features.

according to the principle of the shortest distance, as shown in Formula (1).

$$\min\left\{\left\|X - Z_i(k)\right\|, i = 1, 2, \ldots, K\right\} = \left\|X - Z_j(k)\right\| = D_j(k) \tag{1}$$

In Formula (1), $Z_j(k)$ is the cluster center. K is the number of cluster centers. k indicates the number of clusters already clustered. After iteration, continue to calculate the distance from all points to the cluster center, and recalculate the cluster center according to the feature points in each cluster, as shown in Formula (2).

$$Z_j(k+1) = \frac{1}{N_j} \sum_{X \in S_j(k)} X, \quad j = 1, 2, \ldots, K \tag{2}$$

In Formula (2), if the new cluster center is equal to the previous cluster center, the algorithm converges and the calculation ends. To reduce duplication, improve similarity within the cluster, and decrease similarity between clusters, all features are classified into k clusters using Formulas (1) and (2). The mutual information representation regarding the category's features is shown in the following Formula (3).

$$MI\left(t_k, c_i\right) = \log \frac{P\left(t_k, c_i\right)}{P\left(t_k\right) P\left(c_i\right)} \tag{3}$$

In Formula (3), t_k is the attribute feature and c_i is the feature category. After the features are clustered according to the similarity, each characteristic in the clustering cluster has its class discrimination degree calculated, and Formula (4) can be obtained by combining Formula (3).

$$\alpha = \left|MI\left(t_k, c_i\right) - MI\left(t_k, c_j\right)\right| = \left|\log \frac{P\left(t_k, c_i\right)}{P\left(t_k\right) P\left(c_i\right)} - \log \frac{P\left(t_k, c_j\right)}{P\left(t_k\right) P\left(c_j\right)}\right|, \quad i \neq j \tag{4}$$

The degree of discrimination between the two categories is obtained by Formula (4). The degree to which a characteristic is classified into a category differs from its mutual information in the Mi C. The stronger the discrimination ability increases with the increase of differences. The importance of each feature in the cluster is sorted according to the degree of class discrimination, and the feature with the highest degree of class discrimination in the cluster is screened out, such as Formula (5).

$$CDHI_n = \max\left|MI\left(t_k, c_i\right) - MI\left(t_k, c_j\right)\right| = \max\left|\alpha\right| \tag{5}$$

In Formula (5), n is the number of clusters. After obtaining the class discrimination of each clustering data set, the features are deleted with small discrimination, and the features that are beneficial to the Ma C are selected. After filtering, the dimensionality-reduced

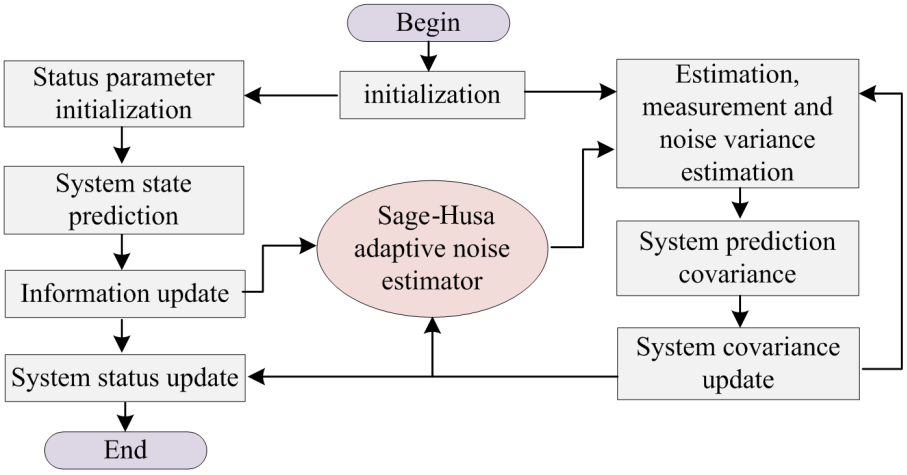


FIGURE 2. Data noise reduction process of UKF algorithm and Sage-Husa

data is obtained. After reducing the dimensionality of the data, the study combined the UKF algorithm with the adaptive unscented Kalman filter (Sage-Husa) to denoise the data. The noise reduction process is shown in Figure 2.

The UKF algorithm performs UT transformation on the nonlinear filter near the estimated point, and first generates the Sigma point. The generation process is as Formula (6).

$$\begin{cases} X_0 = \bar{X}, & i = 0 \\ X_i = \bar{X} + \left(\sqrt{(n+\lambda)P}\right)_i, & i = 1, 2, \ldots, n \\ X_i = \bar{X} - \left(\sqrt{(n+\lambda)P}\right)_i, & i = n+1, \ldots, 2n \end{cases} \tag{6}$$

In Formula (6), X is the Sigma point, n is the vector dimension, P is the variance, and λ is the scaling parameter. The weight after getting the point is shown in Formula (7).

$$\begin{cases} \omega_m^0 = \dfrac{\lambda}{n+\lambda} \\ \omega_c^0 = \dfrac{\lambda}{n+\lambda} + (1 - \alpha^2 + \beta), & i = 1, 2, \ldots, 2n \\ \omega_m^i = \omega_c^i = \dfrac{\lambda}{2(n+\lambda)} \end{cases} \tag{7}$$

In Formula (7), ω_m represents the average weight, ω_c represents the covariance, β represents the non-negative weight coefficient to be selected, and α represents a fixed parameter, generally set as $[0, 1]$. The state probability density function can be obtained by performing nonlinear mapping on the mean value and variance of the Sigma point set prediction. The following Formula illustrates the UKF nonlinear system (8).

$$\begin{cases} X(k+1) = f(x(k), W(k)) \\ Z(k) = h(x(k), V(k)) \end{cases} \tag{8}$$

In Formula (8), f is the nonlinear filling equation function, h is the nonlinear observation equation function, $W(k)$ and $V(k)$ are Gaussian white noises of two covariance matrices. The UKF algorithm is combined with the Sage-Husa filter to form the AUKF algorithm. The adaptive filter corrects the noise in real time while using the system for filtering and noise reduction. The data is input into the Sage-Husa adaptive noise estimator, and the recursive process is shown in the following Formula (9).

$$\begin{cases} \hat{r}_{k+1} = (1 - d_{k+1})\,\hat{Q}_k + d_{k+1}\left[\bar{Z}_{k+1} - \sum_{i=0}^{2n} \omega^i X^i\,(k+1|k)\right] \\ d_{k+1} = \dfrac{1-b}{1-b^{k+1}} \\ \hat{Q}_{k+1} = (1 - d_{k+1})\,\hat{Q}_k + d_{k+1}\left(K_{k+1}\varepsilon_{k+1}\varepsilon_{k+1}^T K_{k+1}^T + P_{k+1}\right) \\ \varepsilon_{k+1} = Z_{k+1} - \hat{Z}_{k+1} \end{cases} \tag{9}$$

In Formula (9), ε is the prediction error of the measurement result. b is the forgetting factor. After preprocessing the data, the imbalanced data is classified. The purpose of classification is to formulate a series of rules to accurately predict the category of new data, and there are a lot of unbalanced data in real life [20,21]. After reducing the dimensionality and noise of the data, the RF algorithm is used to classify the data. The RF algorithm trains, classifies, and predicts data using multiple decision trees. In the process of data classification, the weight of each variable is evaluated. The research introduces cost-sensitive learning to reduce the overall misclassification cost and address the imbalance issue in addition to the RF algorithm's improved ability to avoid over-fitting. Figure 3 depicts the RF algorithm's basic idea.

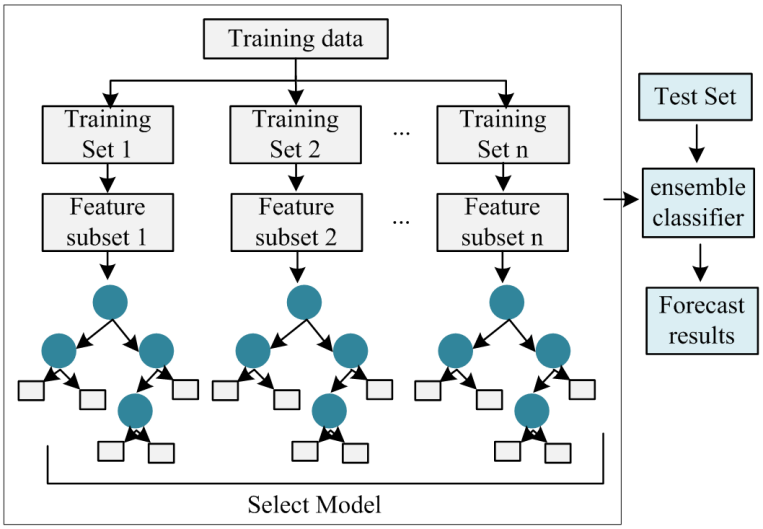


FIGURE 3. Schematic diagram of random forest

Before modeling, the Bagging method is used to extract sample sets with replacement and generate non-training sample sets that are different from each other. Then all feature spaces are screened, and the selected sample subsets and feature subsets are used to construct decision trees. First, the samples are trained, and the information gain rate is studied for attribute splitting. The information entropy of the sample subset is shown in the following Formula (10).

$$E(S) = \sum_{i=0}^{n} p_i \log(p_i) \tag{10}$$

In Formula (10), p_i is the probability of the ith subset in the sample. The expected entropy after splitting is shown in Formula (11).

$$E(S, A) = \sum_{v \in X_A} \frac{|S_v|}{|S|} E\,(S_v) \tag{11}$$

In Formula (11), A is the selected feature, and S_v is the data sub A-block of size in the feature in the data set v. The information gain of attribute to data set is as Formula (12).

$$Gain(S, A) = E(S) - E(S, A) \tag{12}$$

Combined with Formula (12), the information gain rate can be obtained as Formula (13).

$$Gain\_ratio(S, A) = \frac{Gain(S, A)}{split\_\inf o_A(D)} \tag{13}$$

In Formula (13), $split\_\inf o_A(D)$ is the split information, representing the entropy of the training D set A. The split information formula is as Formula (14).

$$split\_\inf o_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{D} \tag{14}$$

Combining the above operations, after selecting attributes according to the selection principle for splitting, a decision tree model is constructed, such as Formula (15).

$$H(x) = \arg\max_y \sum_k I\left[h_k(x) = y\right] \tag{15}$$

In Formula (15), y is the classification result of the decision tree and $I(\,)$ is the indicator function. The test samples are voted on by all of the decision trees. The test sample's category label is determined by the category with the greatest aggregate score among all of the base classifiers that participated in the voting.

3.2. **Classification model construction based on improved K-S-H-RF.** The traditional K-S-H-RF algorithm needs to vote for the test samples as in all decision trees, and the base classifier with poor classification performance has a greater impact on the final result [22]. In order to address the imbalance issue for this study, cost sensitivity is introduced. However, because the traditional cost function construction uses Euclidean distance to calculate the sample distance rather than taking account of the actual distribution of the data set, the classifier's performance is less than ideal. Therefore, the misclassification cost is introduced in the attribute splitting of the decision tree, and the cost factor is constructed on how the samples were really distributed, and the average value of each feature column is taken, and the data set is expressed as a matrix, such as Formula (16).

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} & c \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nm} & c \end{bmatrix} \tag{16}$$

In Formula (16), c represents the category, the columns in the matrix are the characteristics of each data, and the rows represent the data samples. The calculation formula of the Ma C center is displayed in the formula below.

$$A_m = \frac{1}{n} \sum_{i=1}^{n} x_{im} \tag{17}$$

The center of the Ma C is obtained by Formula (17). Similarly, the center of the Mi C and the entire data set are calculated using the same method. After obtaining the category center, calculate the weight distance from the center to the center of the entire dataset. The information gain is used to measure the importance of each attribute in the Mi C and the Ma C. The formula is shown in the following Formula (18).

$$IG\left(x_k, c_i\right) = \sum_{c\in\{c_i,c_i\}} \sum_{x\in\{x_k,x_k\}} P(x,c) \log \frac{P(x,c)}{p(x), p(c)} \tag{18}$$

In Formula (18), $P(x, c)$ represents the probability of c including a feature in a category x, the probability $p(c)$ of a category c in the total data, and the probability of $p(x)$ is a feature in a data set x. All the features are got by using Formula (18) to generate weight vectors in the Ma C and the Mi C, and add the respective weight vectors when calculating the weight distance between the category and the training set. The weight distance is shown in the following Formula (19).

$$d_i = \sqrt{\sum_{j=1}^{m} w_i \left(A_{ij} - \bar{A}^2\right)} \tag{19}$$

In Formula (19), w_i is the weight value of the feature in the Ma C, A_{ij} is the category center of the Ma C, and $\bar{A}$ is the center of all sample sets. Similarly, Formula (19) is used to calculate the weight distance between the minority samples and the center of the data set. Coefficients for each category are defined. The formula is displayed in the formula below.

$$\gamma_i = \frac{\sum_{j=0}^{1} N_j}{N_i} \tag{20}$$

In Formula (20), the unbalanced data set is N. N_i is the number of data points in each category. Formulas (19) and (20) are combined to construct a cost function. The cost function is shown in Formula (21).

$$F(c_i, c_j) = \begin{cases} \gamma_i * \dfrac{d_i'}{d_j''}, & d_i' < d_j'' \\ \gamma_i * \dfrac{d_j'}{d_i''}, & d_i' < d_j'' \\ 0, & i = j \\ 1, & d_i' = d_j'' \end{cases} \tag{21}$$

In Formula (21), d_i' is the weight distance d_j'' between the category center and the center of the entire data set. c_i and c_j are the weight distance from the center of the entire data set. In the original data set, the Bagging method is used to extract sample sets to obtain different sample spaces.

The feature m is from the feature space of each original dataset. Calculate the cost reduction value, as shown in Formula (22).

$$\mathrm{Re}c = Mc - \sum_{i=0}^{n} Mc(A_i) \tag{22}$$

In Formula (22), Rec is the cost drop value generated after splitting, Mc is the cost without splitting, and A_i represents the feature. The final misclassification cost of the split point in Formula (22) is shown in Formula (23).

$$\sum_{i=0}^{n} Mc(A_i) = n * FP - \left(FP * \sum_{i=0}^{r} n_i + FN * \sum_{i=r+1}^{n} p_i \right) \tag{23}$$

In Formula (23), n is the number of the Ma C, and P is the number of the Mi C. After the descending value is obtained, the largest descending value is selected each time to split the node. Finally, a decision tree classifier is generated. In the base classifier composition stage, for unbalanced data, each decision tree uses the AUC value for performance

evaluation, uses the AUC value for weighted voting, and assigns the obtained weights to the base classifier. Finally, the output formula of the RF classifier is obtained, such as Formula (24).

$$H(x) = \arg \max_{y \in \{-1,1\}} \sum_{k} a_k I\left[h_k(x) = y\right] \tag{24}$$

In Formula (24), h_k represents the decision tree model, $I(\,)$ represents the indicator function, and a_k is the voting weight of the decision tree. The traditional RF algorithm does not adopt a parallel method in building the base classifier. The slow efficiency of RF building results from the requirement to wait for the training of the preceding classifier to be finished before training the next model. When solving big data problems, it seriously affects the performance of the algorithm. Therefore, the parallel design of the algorithm is studied. The idea of MapReduce is introduced into the study. In the RF, each submodel is regarded as a "sub-problem", that is, the Map process. The establishment of all base classifiers is completed in the Map process, and there are two parallel processes for attribute splitting. Reduce summarizes the base classifiers obtained in the Map process, writes each base classifier to the distributed file system, and completes the modeling of the final integrated classifier. After all decision trees are constructed, the Out-of-Bag samples of each decision tree are tested to obtain the AUC values. Then the voting process is weighted for a number of Map functions here, and the weight values are counted using the Reduce process to obtain the categories predicted by the samples. The flow of the study's classification algorithm for imbalanced big data is shown in Figure 4.

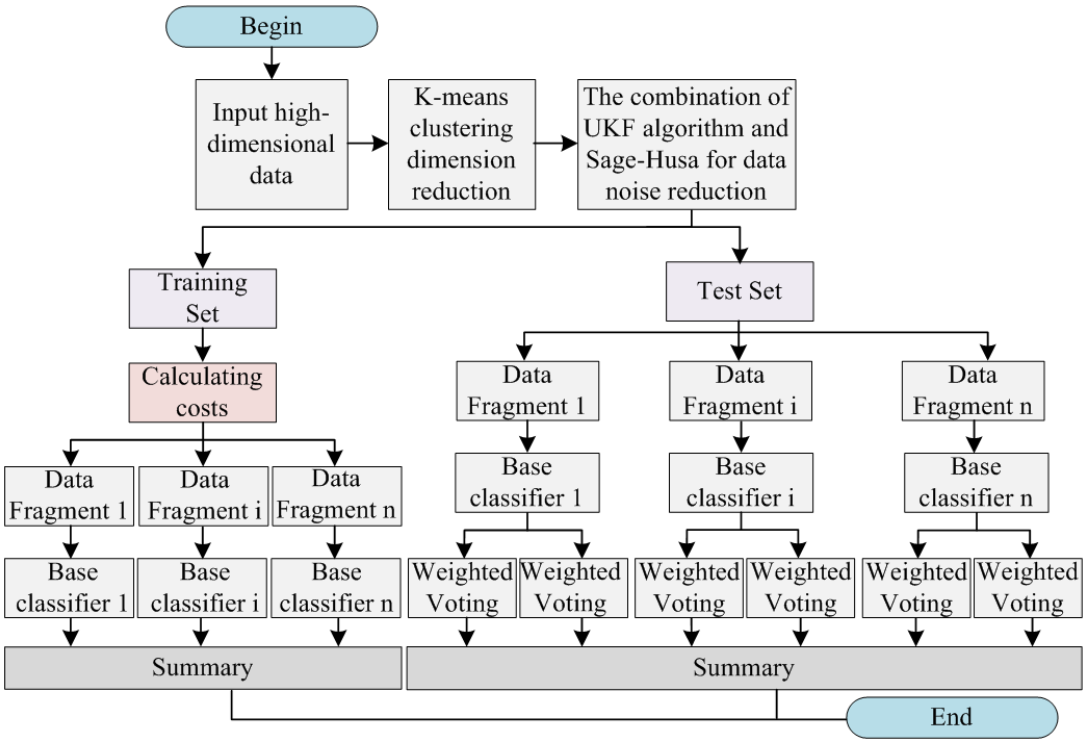


FIGURE 4. Unbalanced big data classification algorithm flow

In Figure 4, MapReduce encapsulates the function of realizing distributed computing. Users only need to process input data. The MapReduce distributes tasks to different work nodes for execution. The algorithm is defined by two functions, namely, Map function and Reduce function. In this paper, in the MapReduce framework, the establishment of decision tree is parallelized twice. The construction process and the split feature selection

process are completed in the Map task, including the establishment of base classifier and the two parallel processes of attribute splitting. There is also a parallelization in the final voting process, because the voting of each decision tree is also independent of each other and does not interfere with each other. Finally, it is summarized in the Reduce process. Each base classifier is written to the distributed file system to complete the modeling of the final integrated classifier. Integrating the above, the study uses K-means and class distinction to achieve approximate dimensionality reduction for high-dimensional data. The cost-sensitive and MapReduce ideas are introduced to improve the random forest algorithm to identify imbalanced data. An imbalanced big data classification model based on the improved random forest algorithm is constructed to achieve the mining and analysis of high-dimensional imbalanced data.

4. **Performance Analysis of Classification Model Based on Improved K-S-H-RF Algorithm.** With the advancement of science and technology, people's ability to obtain stable data has been greatly improved. Classification is one of the tasks in the field of data mining, which is the process of extracting hidden information from a huge amount of data. Data used in the experiment are obtained from UCI database with four data sets: Amazon_initial, 20-Newgroups, waveform and Covtype. Amazon_initial contains 10,000 features and 50 categories, which are selected to accumulate as a minority class and the rest as a majority class. 20-Newgroups is a set of high-dimensional samples frequently used by text, with a total of 19,997 sample points and 100,000 features. 20-Newgroups contains 20 category tags. Waveform contains 15,000 data sets and Covtype data set contains 28,000 sample numbers. In the data set, there are relatively few important features. It is unfair to calculate the construction cost of Euclidean distance from the class center to the whole data set. This paper introduces weight distance and uses information gain to measure the importance of each attribute in majority class and minority class. The calculation method is shown in Formula (25).

$$IG\left(x_k, c_i\right) = \sum_{c\in\{c_i,c_i\}} \sum_{x\in\{x_k,x_k\}} P(x,c) \log \frac{P(x,c)}{p(x), p(c)} \tag{25}$$

In Formula (25), c_i represents the feature, and the information gain of the feature is IG. The size of the IG value is proportional to the amount of contribution that the feature provides to the classification. In this paper, the information gain value is used to rank the features. Then the weights are assigned according to the ranking results.

To realize the efficient classification of HdUD, the study uses K-means combined with a class discrimination algorithm (CDHI) to make HdUD less dimensional and uses the improved RF algorithm to classify the data, thus constructing an imbalanced big data classification model based on improved RF algorithm. To verify the dimensionality reduction effect, its performance was measured by the AUC value. The CDHI algorithm was compared with traditional feature selection algorithms such as (chi-square detection) CHI algorithm, information gain (IG) algorithm, and mutual information (MI). Select two classifiers, NB and RF, and use the CDHI algorithm to select AUC values with different numbers of features in each feature cluster for comparison. The details are shown in Figure 5.

From Figure 5(a), with the increase of the number of feature samples, the AUC values of each algorithm have increased. When the number of features is 100, the CDHI algorithm's AUC value is 0.800, the IG algorithm's AUC value is 0.780, which is 0.020 less than that of the CDHI algorithm, the MI algorithm's AUC value is 0.776, which is 0.024 lower than that of the CDHI algorithm, and the CHI algorithm's AUC value is 0.657, 0.143 less than that of the CDHI algorithm. When the number of feature values is 1500, the

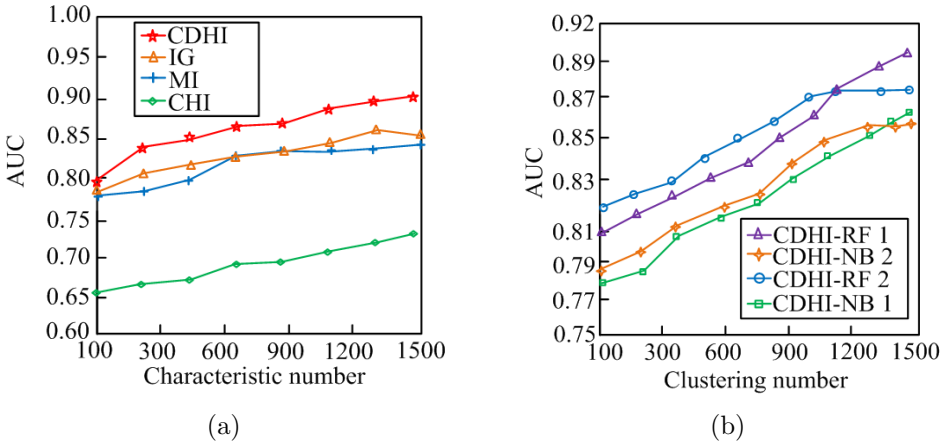


FIGURE 5. Comparison of AUC values

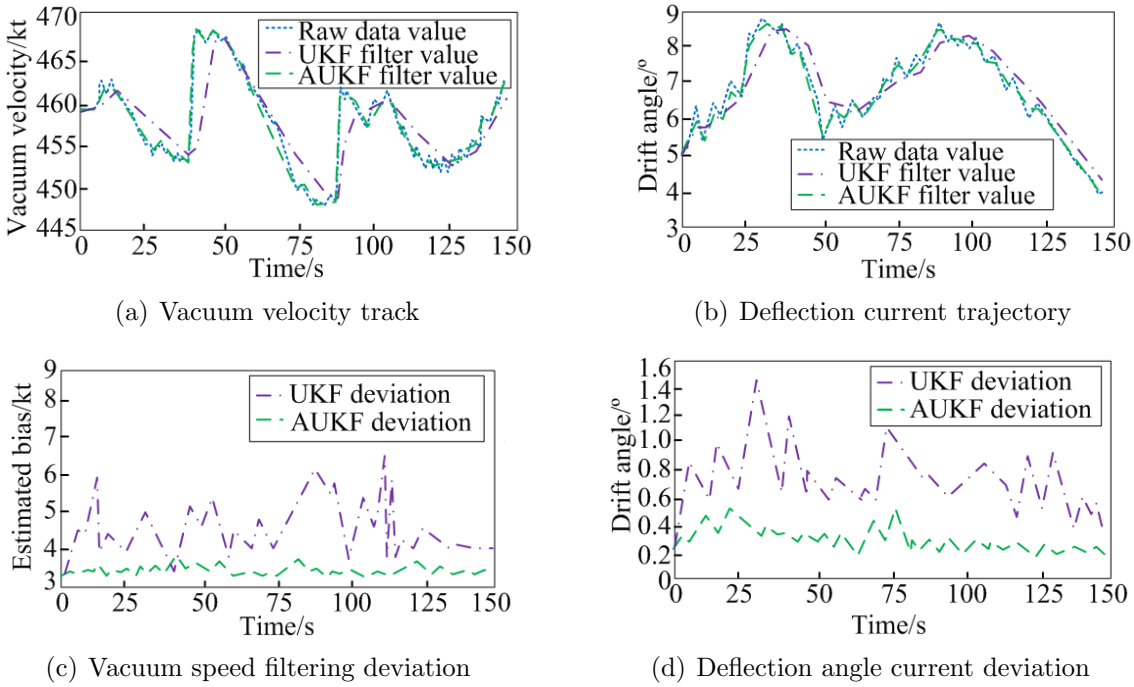
CDHI algorithm's AUC value is 0.902, the IG algorithm's AUC value is 0.839, 0.063 less than that of the CDHI algorithm, and the MI algorithm's AUC value is 0.825, which is 0.077 lower than that of the CDHI algorithm, and the CHI algorithm's AUC value is 0.728, which is 0.174 lower than that of the CDHI algorithm. The classification accuracy of the CDHI algorithm is higher, and the classification performance is better than other algorithms. From Figure 5(b) that CDHI-RF1 and CDHI-NB1, CDHI-RF2 and CDHI-NB2 are the features with the largest class discrimination in each cluster, as the number of clusters increases, the features themselves will have interference from non-strongly correlated features. Therefore, CDHI-RF2 and CDHI-NB2 with a large number of features will change from a higher AUC value to a slightly lower value at the beginning in CDHI-RF1 and CDHI-NB1. Based on the above content, it can be seen that the CDHI algorithm has a good classification performance in feature selection, and the number of features also affects its classification effect.

In order to compare the noise reduction effects of the AUKF and UKF algorithms and compare the filter trajectory and deviation under the conditions of true air speed and drift angle, the research compares the noise reduction effects of the data using the AUKF and UKF algorithms, as shown in Figure 6.

From Figure 6, the filtered curve is smoother. The two algorithms are effective for data noise reduction processing, among which the UKF filter curve deviates more from the original data track than the AUKF filter curve. It can be seen from the two figures of Figures 6(c) and 6(d) that the deviation of AUKF is smaller, and the change range is not large. The filter deviation is maintained between 3.5 and 3.7 at true air speed, and the deviation value is 0.2 in the case of bias angle. It fluctuates between 0.5 and UKF deviation value which is constantly changing with the increase of the number of features, and the range of change is very large. The filter deviation remains floating between 3.6 and 6.7 at true air speed, the deviation value variables between 0.4 and 1.5 under the condition of bias angle. The results showed that the data processed by AUKF is more accurate and trustworthy because it is closer to the genuine value.

Next, the study introduces root mean square error (RMSE), signal-to-noise ratio (SNR) and smoothness (R) to evaluate the noise reduction effect, as shown in Table 1.

The data in Table 1 shows that in both cases, the SNR obtained by UKF is smaller than that of AUKF, the root mean square error is larger than that of AUKF, and after filtering by UKF, the smoothness of the data is larger than that of AUKF. The comprehensive comparative analysis of the data shows that the AUKF selected in the study can retain



(a) Vacuum velocity track

(b) Deflection current trajectory

(c) Vacuum speed filtering deviation

(d) Deflection angle current deviation

FIGURE 6. Comparison of noise reduction effects

TABLE 1. Evaluation of noise reduction effect

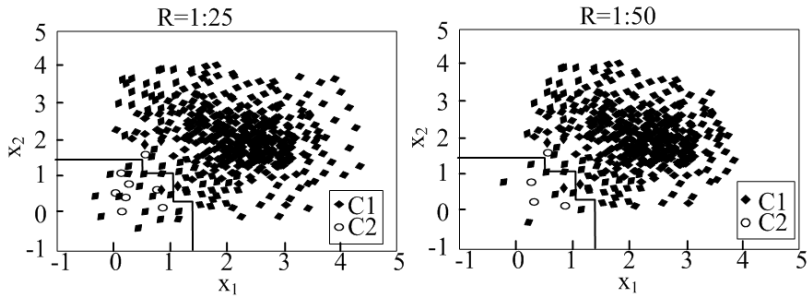
Evaluation criteria	Vacuum velocity filtering		Bias angle filtering	
	AUKF	UKF	AUKF	UKF
RMSE	1.4965	10.5432	0.5689	1.6895
SNR	61.5874	44.6895	32.4698	23.5692
R	1.4689	8.3245	1.6948	3.4597

more effective data information after processing the original data, and can reduce the noise of the data.

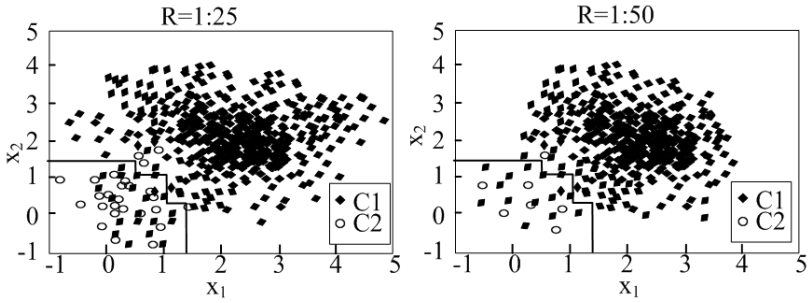
The classification model before and after introducing cost sensitivity is explored when the degree of data imbalance is 1 : 25 and 1 : 50 to confirm the validity of introducing cost sensitivity in the process of utilizing the RF algorithm to classify imbalanced data after dimensionality reduction. Figure 7 illustrates a comparison of the recognition effect.

From Figure 7, before and after the introduction of cost-sensitivity, as the degree of imbalance increases, the small class recognition rate solution gradually decreases. When the imbalance ratio reaches 1 : 25, the model before the introduction of cost-sensitivity is 64.8%. The recognition rate of the small class is 63.2%, and the recognition rate of the model after the introduction of cost sensitivity is 78.5%. The comprehensive comparison shows that after the introduction of cost sensitivity, the RF algorithm is less affected by the degree of data imbalance. The recognition rate of small classes is also increased. In terms of recognition rate and generalization performance, it performs better than the conventional RF algorithm.

For further verification research, the accuracy rate, recall rate (Re), execution time, and AUC value are introduced to evaluate the model performance. Four imbalanced data sample sets with unbalanced class distribution were selected and classified using the improved K-U-S-H-RF imbalanced big data classification model (Model 1) and the more



(a) Before introducing cost sensitivity



(b) After introducing cost sensitivity

FIGURE 7. Scatter chart of recognition before and after introducing cost sensitivity under different balances

TABLE 2. Comparison of classification of four models

Project	Dataset SN	Accuracy (%)	Re (%)	AUC	Execution time (ms)	F-measure
Model 1	1	94.586	91.132	0.802	4085	0.825
	2	93.565	92.468	0.854	3729	0.846
	3	92.973	91.985	0.824	4156	0.878
	4	94.051	92.069	0.896	3943	0.849
Model 2	1	90.589	88.469	0.785	19454	0.796
	2	90.238	87.285	0.759	18796	0.785
	3	89.568	88.023	0.792	19632	0.768
	4	89.569	89.464	0.774	20110	0.694
Model 3	1	89.379	88.233	0.798	19584	0.699
	2	88.958	87.692	0.658	20115	0.692
	3	90.023	86.985	0.691	21054	0.701
	4	88.491	86.999	0.725	20036	0.658
Model 4	1	88.269	87.456	0.721	21058	0.669
	2	89.465	86.987	0.695	22345	0.667
	3	90.369	87.263	0.693	21937	0.682
	4	91.466	88.021	0.701	22018	0.701

commonly used classification models, including convolutional neural network-based classification model (Model 2), integrated neural network-based classification model (Model 3), and linear analysis-based classification model (Model 4). All four models were computed in parallel using MapReduce. The results are shown in Table 2.

From Table 2, the Model 1's average accuracy is 93.80%, and the Model 2's average accuracy is 89.99%, which is 3.81% lower than Model 1. The Model 3's average accuracy is 89.21%, which is 4.59% lower than Model 1. The Model 4's average accuracy rate is 89.89%, which is 3.91% lower than the recognition and classification accuracy rate of Model 1. For other evaluation values, the value of Model 1 is greater than that of the other three models. Therefore, Model 1 has better performance and can better classify unbalanced data.

5. **Conclusion.** With the rise of big data, all kinds of tools in daily lives are getting more intelligent. Economic development is also accelerated through data, and data mining is an essential research area. However, there is usually a serious imbalance in the data in various fields. To achieve the dimensionality reduction of high-dimensional unbalanced data, the study uses the K-means clustering algorithm based on class distinction to reduce the dimensionality of the data, and combines the UKF algorithm and Sage-Husa to improve the RF to obtain the improved K-U-S-H-RF algorithm. During the study, it was found that the K-U-S-H-RF algorithm did not take account of the actual step-by-step of the dataset and did not work well in classifying the data. For this reason, the study introduced cost sensitivity, cost error calculation for decision trees as well as voting, and parallelized random forests by introducing MapReduce ideas to achieve optimization of the algorithm. Overall, the study constructs an imbalanced big data classification model based on improved random forests. The experimental analysis shows that the average recognition classification accuracy of Model 1 is 93.80%, the AUC value is 0.844, the Recall value is 91.91%, the average running time is 3978 ms, and the F-measure value is 0.850. The model constructed in the study can accurately and effectively classify high-dimensional unbalanced data. The research model performed well while categorizing vast volumes of data; however, there is opportunity for speed enhancement, which can be investigated in future studies. In unbalanced big data classification, the selection of appropriate features is crucial for classification performance. The research may not fully consider the correlation between sample imbalance and features. In the future, new feature selection and feature combination methods can be explored to improve the performance of random forest algorithms in unbalanced big data classification.

Acknowledgement. The research is supported by two Science and Technology Project of Jiangxi Provincial Department of Education: Research on unbalanced classification algorithm and application of automotive big data based on machine learning (No. GJJ212012), and Research on Crowdsourcing Test Report Mining Technology Based on Image Recognition (No. GJJ202011).

REFERENCES

[1] K. Sridharan, G. Komarasamy and S. D. M. Raja, Hadoop framework for efficient sentiment classification using trees, *IET Networks*, vol.9, no.5, pp.223-228, 2020.

[2] A. Sungheetha and R. S. Rajendran, Big data analysis and perturbation using data mining algorithm, *Journal of Soft Computing Paradigm*, vol.3, no.1, pp.19-28, 2021.

[3] Y. X. Yu, X. D. Yu, Q. Z. Cheng, L. Tang and M. Q. Shen, The association of serum vitamin K2 levels with Parkinson's disease: From basic case-control study to big data mining analysis, *Aging (Albany NY)*, vol.12, no.16, pp.16410-16419, 2020.

[4] S. Lee, Y. Hyun and M. J. Lee, Groundwater potential mapping using data mining models of big data analysis in Goyang-si, South Korea, *Sustainability*, vol.11, no.6, 1678, 2019.

[5] D. Tao, P. Yang and H. Feng, Utilization of text mining as a big data analysis tool for food science and nutrition, *Comprehensive Reviews in Food Science and Food Safety*, vol.19, no.2, pp.875-894, 2020.

[6] W. Han, Z. Huang, S. Li and Y. Jia, Distribution-sensitive unbalanced data oversampling method for medical diagnosis, *Journal of Medical Systems*, vol.43, no.2, pp.1-10, 2019.
[7] Q. Hang, J. Yang and L. Xing, Diagnosis of rolling bearing based on classification for high dimensional unbalanced data, *IEEE Access*, vol.7, no.21, pp.79159-79172, 2019.
[8] W. Cong, Study of financial warning ensemble model for listed companies based on unbalanced classification perspective, *International Journal of Intelligent Information Technologies*, vol.16, no.1, pp.32-48, 2020.
[9] H. Pant, M. Sharma and S. Soman, Twin neural networks for the classification of large unbalanced datasets, *Neurocomputing*, vol.343, no.28, pp.34-49, 2019.
[10] A. Guha and N. Veeranjaneyulu, Prediction of bankruptcy using big data analytic based on fuzzy c-means algorithm, *IAES International Journal of Artificial Intelligence*, vol.8, no.2, pp.168-174, 2019.
[11] J. Yan, Y. Hu and C. Guo, Rotor unbalance fault diagnosis using DBN based on multi-source heterogeneous information fusion, *Procedia Manufacturing*, vol.35, no.11, pp.1184-1189, 2019.
[12] S. Balli, E. A. Sağbaş and M. Peker, Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm, *Measurement and Control*, vol.52, nos.1-2, pp.37-45, 2019.
[13] Y. Chen, W. Zheng, W. Li and Y. Huang, Large group activity security risk assessment and risk early warning based on random forest algorithm, *Pattern Recognition Letters*, vol.144, no.4, pp.1-5, 2021.
[14] S. Georganos, T. Grippa, A. N. Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff and S. Kalogirou, Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modeling, *Geocarto International*, vol.36, no.2, pp.121-136, 2021.
[15] Y. Ao, H. Li, L. Zhu, S. Ali and Z. Yang, The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling, *Journal of Petroleum Science and Engineering*, vol.174, no.3, pp.776-789, 2019.
[16] R. Liang, Y. Lu, X. Qu, Q. Su, C. Li, S. Xia, Y. Liu, Q. Zhang, X. Cao, Q. Chen and B. Niu, Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data, *Transboundary and Emerging Diseases*, vol.67, no.2, pp.935-946, 2020.
[17] V. A. Fitri, R. Andreswari and M. A. Hasibuan, Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm, *Procedia Computer Science*, vol.161, pp.765-772, 2019.
[18] M. Jain, G. Kaur and V. Saxena, A K-Means clustering and SVM based hybrid concept drift detection technique for network anomaly detection, *Expert Systems with Applications*, vol.193, no.1, 116510, 2022.
[19] F. H. Awad and M. M. Hamad, Improved k-means clustering algorithm for big data based on distributed smartphoneneural engine processor, *Electronics*, vol.11, no.6, 883, 2022.
[20] Y. Guo, D. Yang, Y. Zhang, L. Wang and K. Wang, Online estimation of SOH for lithium-ion battery based on SSA-Elman neural network, *Protection and Control of Modern Power Systems*, vol.7, no.1, pp.40-41, 2022.
[21] S. Wang, P. Ren, P. Takyi-Aninakwa, S. Jin and C. Fernandez, A critical review of improved deep convolutional neural network for multi-timescale state prediction of lithium-ion batteries, *Energies*, vol.15, no.14, 5053, 2022.
[22] K. R. Ummah, T. Karlita, R. Sigit, E. M. Yuniarno, I K. E. Purnama and M. H. Purnomo, Effect of image pre-processing method on convolutional neural network classification of COVID-19 CT scan images, *International Journal of Innovative Computing, Information and Control*, vol.18, no.6, pp.1895-1912, 2022.

Author Biography

Xin Zheng received her Bachelor of Engineering degree in Computer Science and Technology from Nanchang University, China in 2003, and her Master's degree in Computer Technology from Nanchang University, China in 2010.

From 2003 to now, she has been working in Jiangxi University of Technology, engaged in the teaching and scientific research of computer majors and big data majors.

She has published 9 academic papers, participated in the compilation and publication of 2 academic works and textbooks, led and participated in 10 research projects, authorized 1 invention patent, and 2 utility model patents.

Li Huang obtained a Bachelor's degree in Computer Science and Technology from Jiangxi Normal University in 2003 and a Master's degree in Software Engineering from Nanchang University in 2010.

From 2003 to present, she worked at Jiangxi University of Technology, engaged in teaching and research work in the field of software engineering.

She has published 8 academic papers, participated in the compilation and publication of 1 academic textbook, led and participated in 10 research projects, authorized 1 invention patent, and 4 utility model patents.