Research on anomaly data mining method of new energy field stations based on improved Adaboost algorithm

Wang Nan1*, Wang Yanzhuo1, Cheng Yan1, Guan Ti2, Ma Qiang1, Sun shumin1, Guan Yifei1, Wang Yuejiao1, Wang Shibo1

1State Grid Shandong Electric Power Company Electric Power Research Institute, Jinan, Shandong, 250002, China
2State Grid Shandong Electric Power Company, Jinan, Shandong, 250000, China
*Corresponding author’s e-mail: wangnantju@163.com

Abstract: Traditional anomalous data mining methods require a lot of prior knowledge, which leads to low data mining integrity and efficiency. For this reason, a new energy field abnormal data mining method based on improved Adaboost algorithm is proposed. After pre-processing the new energy field data, the algorithm is improved by introducing dynamic weight parameters to address the shortcomings of the Adaboost algorithm. After calculating the degree of data anomaly using the direct push belief machine, the neural network is used to reduce the error value of the Adaboost algorithm, and finally the output of the Adaboost algorithm is used to realize abnormal data mining. The simulation experiment proves that the researched abnormal data mining method has high data integrity and high efficiency.

1. Introduction

Currently, new energy sources are widely used throughout my country to undertake energy supply tasks, and many new energy stations have been established. New energy stations are an important part of the new energy supply system, and a large amount of control command data and monitoring data are summarized in the new energy stations[1]. In order to ensure the normal operation of new energy plants, it is necessary to dig out abnormal data from massive data in depth, high speed and accuracy.

The distance-based anomaly data mining method mentioned in the literature[2] takes different distances between data as the measure and achieves anomaly data mining by setting different distance thresholds. However, this method is only applicable to data in general massive database for mining, and there are limitations in its use. The density-based anomalous data mining method mentioned in the literature[3] determines whether the data is anomalous by calculating the number of data in a given range according to a certain threshold value. This method requires several experiments to determine the best range threshold, is costly in data processing, and is only applicable to anomaly mining cases with less data. The statistical-based abnormal data mining method mentioned in the literature[4] needs to clarify that the normal data in the object to be processed as a whole obeys some known or approximately known probability distribution model before processing, and then realize abnormal data mining according to the statistical principles. However, this data mining method can only deal with single-dimensional data objects in actual processing, and requires more a priori knowledge. The daily operation data of new energy stations are highly heterogeneous and the data volume is extremely large, which puts high demands on the reliability and data processing efficiency of the anomaly data mining
method. Therefore, it is crucial to use other advanced algorithms or technical processing tools to improve the performance of the anomaly data mining method.

To this end, this paper proposes a new energy field abnormal data mining method based on the improved Adaboost algorithm. After collecting and preprocessing the new energy field data, the neural network improved Adaboost algorithm is used to detect and mine abnormal data. The experimental results show this method can effectively improve the degree and efficiency of excavation.

2. Research on anomaly data mining method of new energy field stations based on improved Adaboost algorithm

2.1. New energy station data pre-processing

Due to the different sources of data collected in the new energy stations, the large amount of heterogeneous data affects the efficiency improvement of subsequent abnormal data mining. Therefore, it is necessary to preprocess the mined data before the new energy station abnormal data mining. In the process of transmitting and receiving energy supply instructions, monitoring the energy supply of new energy and maintaining the normal operation of the station, a large amount of data is transmitted through different data channels, and the transmission process, data collection and monitoring equipment will produce unavoidable noise data. Firstly, the wavelet threshold method is used to de-noise the data of new energy stations.

If the new energy field station data sequence can be expressed in the form of formula (1), noise can be removed by performing discrete wavelet transform on the original data sequence containing noise data\(^5\).

\[
s(k) = f(k) + n(k), k = 0, 1, 2, \ldots, N - 1
\]

In formula (1), \(s(k)\) is the data sequence of new energy stations that has not been processed; \(f(k)\) is the noise-free data series of new energy stations; \(n(k)\) is the noise data sequence contained in the data sequence of new energy station. The discrete wavelet transform\(^6\) of the original data sequence of the new energy station can be obtained in the Formula (2):

\[
w_s(j, k) = w_f(j, k) + w_n(j, k), j = 0, 1, \ldots, J; k = 0, 1, \ldots, N - 1
\]

In formula (2), \(w_s(j, k)\) is the wavelet coefficient on the corresponding scale of the unprocessed data series of new energy stations; \(w_f(j, k)\) is the wavelet coefficient on the corresponding scale of the data series of new energy stations without noise. \(w_n(j,k)\) is the wavelet coefficient on the corresponding scale of the noise data sequence; \(J\) is the maximum decomposition scale of discrete wavelet transform; \(N\) is the length of the data sequence. According to the specific data scale of the new energy station, the appropriate wavelet basis function and the number of decomposition layers are selected to complete the denoising process. After denoising, the data need to be normalized after filling the gaps.

In the process of energy supply of new energy stations, some data received and stored by the stations will be missing due to the action of many factors. According to different data types, the corresponding methods are used to fill the missing data\(^7\). For the periodic missing data, the data law of adjacent periods is used to fill; For data that can be ignored and will not affect data mining results, missing data filling processing can be avoided. For data with association rules, missing data can be filled according to established association rules. After the data filling process, the data will be normalized according to the formula (3).
In formula (3), $x'$ is the standardized data; $\bar{x}$ is the average value of the data to be processed; $x$ is the data after filling processing; $\sigma_s$ is the standard deviation of the data to be processed. After data preprocessing, the Adaboost algorithm was improved and optimized.

2.2. Adaboost algorithm improved processing

Adaboost is an iterative algorithm that improves on the Boosting polynomial augmentation learning algorithm, which can further improve the prediction accuracy of all current machine learning algorithms by addressing the difficulty of constructing strong learners directly\(^8\). By adjusting the weights of the sample set and the weights of the weak classifier, several classification sub-models with lower accuracy or resolution are constructed, trained, and integrated to form a classification model with higher accuracy or resolution, thus shortening the steps of generating classification models with higher accuracy or resolution and effectively reducing the difficulty, and also making the classification model less demanding on the sample set data of the object processed by the algorithm. By regulating the weights of the classification sub-models, the processing accuracy of the target classification model is adjusted in the process of multiple iterations, and the classification processing capability of the final generated model for the data is improved. In the iterative process of Adaboost algorithm, firstly, the sample data weights are initialized according to the size of the training sample set and a classification sub-model with lower accuracy or resolution is obtained. Through multiple iterations, several classification sub-models with poor classification ability are obtained. According to and according to the classification accuracy of each classification sub-model, different weights are assigned to the classification sub-models. After the iteration of the Adaboost algorithm is completed, all the classification sub-models with known weights are combined to generate a classification model with strong accuracy or resolution by some linear weighting\(^9\). The Adaboost algorithm is mainly applied to solve classification problems, but the algorithm is also applied to solve two-class problems, multi-class single-label problems, multi-class multi-label problems, large-class single-label problems, linear regression problems, and other regression problems. The Improved Adaboost algorithm, on the other hand, is based on the traditional Adaboost algorithm and reduces the error rate of the traditional Adaboost algorithm in sub-giving processing by combining other algorithms, thus improving the performance of the Adaboost algorithm as a whole.

The Adaboost algorithm forms a classification model with high classification accuracy by adjusting the weights of sub-classification models generated by different iteration stages during the iteration process. Traditional Adaboost algorithm, however, in this process, can't accurate control algorithm is the number of iterations, and to the son of different iterative link to generate the value assignment of classification model does not have fixed rules, lead to the resulting higher classification accuracy of classification model, although has the higher classification accuracy, but the processing efficiency is very low. Therefore, aiming at the problems existing in the traditional Adaboost algorithm studied above, this paper improved and optimized the traditional Adaboost algorithm by increasing the threshold value and distributing the momentum parameter\(^10\) by weight.

According to the data processing requirements of the traditional Adaboost algorithm, a known training data set $\{X, Y\}$ is given, where $Y$ is the attribute category label corresponding to each data
in the training data set. Then, by assigning the same weight value to each data in the training data set, the weight distribution of the data in the training data set is initialized to obtain the initial weight value of the data in the training data set. According to the iterative process of traditional Adaboost algorithm, sub-classifiers with low classification accuracy are generated iteratively. From all the currently generated sub-classifiers with low classification accuracy, the sub-classifier with the highest accuracy is selected as the base sub-classifier of the \( t \) iteration, and the error rate of the initial weight distribution of the training data set is calculated based on the weight value of the base sub-classifier\[11-14\]. The specific calculation formula is as formula (4):

\[
D_t = \sum w_i \cdot H_t(e_i)
\]

In formula (4), \( e_i \) is the error rate of initial weight distribution of training data set; \( w_i \) is the weight of the data in the training data set during the round algorithm iteration; \( H_t \) is the selected base sub-classifier. Then, the first-order weight of the sub-classifier can be calculated according to the formula (5):

\[
a_t = \frac{1}{2} \ln \frac{1 - e_i}{e_i}
\]

In order to ensure the processing efficiency of Adaboost algorithm, weight distribution momentum parameter is introduced. The momentum parameters are distributed by weights, and the direction is generated according to a certain direction to speed up the learning of Adaboost classifier, so as to reduce the iteration of the algorithm with incorrect direction and improve the efficiency of the algorithm. The initial weight distribution of the converted training data set is optimized by the momentum method\[15\] to satisfy the relationship of formula (6):

\[
W_{r+1} = W_r + V_r
\]

In formula (6), \( V_{r+1} \) is the momentum parameter allocated by the iteration weight added by the momentum method. The calculation formula of this parameter is formula (7):

\[
V_{r+1} = -\gamma L(W_t) + \mu
\]

In formula (7), \( \gamma \) is the learning rate of Adaboost algorithm; \( \mu \) is the coefficient of momentum parameter term of weight distribution; \( L(W_t) \) is the distribution gradient value of the centralized weight of training data. For data samples with unclear category identification to be classified and recognized, in order to avoid excessive iteration times, resulting in the increase of weight of corresponding category items and affecting the processing efficiency of Adaboost algorithm, classification iteration threshold is introduced to control the increase of sample weight of category identification for unknown classification recognition\[16\]. When the sample weight of the unidentified classification identification category is greater than the threshold value, the corresponding sample weight is cleared to reduce the interference of such samples to the classifier. In order to avoid the redundancy of sub-classifier in the process of multiple iterations, another threshold is introduced as the adjustment threshold of algorithm iteration. When the number of iterations of the Adaboost algorithm reaches the iteration threshold, the iteration of the algorithm is stopped. By introducing the iteration threshold of the algorithm, the classification weight of the sub-classifier is obtained when the algorithm correctly classifies the data, so as to obtain a better sub-classifier. After the improvement and optimization of Adaboost algorithm, the data anomaly degree is calculated by direct push reliability machine, so as to detect the abnormal data in the new energy station data.
2.3. Data anomaly degree calculation

The direct push reliability machine is mainly used to determine the degree of anomaly in the data\cite{17}, so as to divide the data to be processed. According to the formula (8), the anomaly degree of new energy station data is depicted:

$$\alpha_i = \sum_{j=1}^{k} D_j$$

(8)

In formula (8), $k$ is the number of data in the nearest neighbor selected in the data set of new energy stations; $D$ is the nearest neighbor distance of the sample data. According to the formula (8), the anomaly degree of the data is determined, and under the independent co-distribution condition \cite{18}, the P value of the sample data to be processed is obtained. The data can be divided according to the process in Figure 1.

![Figure 1 schematic diagram of direct push reliability machine division data](image)

After dividing the abnormal degree of new energy station data according to the direct push reliability machine, the neural network combined with the improved Adaboost algorithm was used to mine the abnormal data of new energy station.

2.4. New energy station abnormal data mining is realized

The neural network selected in this paper adopts single hidden layer\cite{19}, and the number of hidden layer nodes is selected according to the empirical formula. The empirical formula for calculating the number of hidden layer nodes is as formula (9):

$$K < \sum_{i=1}^{n} C_i^{n_i}$$

(9)

In formula (9), $K$ is the number of samples, $n_i$ is the number of hidden layer node. When $i > n_i$, $C_i^{n_i} = 0$. Where, the relationship between the number of hidden layer nodes $n_i$ and the number of input layer nodes and output layer nodes\cite{20} is shown in the formula (10).

$$n_i = \sqrt{k + h} + d$$

(10)

In formula (10), $h$ is the number of nodes at the output layer of the neural network, and $d$ is the constant in the interval\cite{1,10}. In the hidden layer, the improved Adaboost algorithm is used to classify and process the new energy station data that determines the degree of abnormality\cite{21}. The sub-classifier generated by the improved Adaboost algorithm generates the final classifier according to the weighted linear combination formula of formula (11):

$$y(x) = \text{sign} \left( \sum_{i=1}^{M} a_i(x) \right)$$

(11)

In the formula (11), $M$ is the total number of iterations of the Adaboost algorithm. In the neural network, each layer is connected by connection weight\cite{22}, and the output of each layer serves as the input of the next layer. The calculation formula of connection weight is as formula (12).
After the basic structure of the neural network is determined, the training samples are used to train the neural network to minimize the mean square error between the actual output value and the expected output value. The neural network was initialized, and the gradient search technique was used to input training samples to the neural network. The input signals were processed from the input layer through the hidden layer and then transmitted to the output layer. Each layer of neurons only affected the state of the next layer of neurons. If the desired output cannot be obtained in the output layer, it will be transferred to back propagation, and the error of the output signal will be returned along the original connection path. By modifying the connection weight between neurons in each layer, the error will be minimized, and the neural network training will be completed. The neural network parameters determined after the training are used to mine the data of the new energy stations. Then the output result of the neural network output layer combined with the improved Adaboost algorithm is the abnormal data of the mined new energy station data. So far, the new energy station anomaly data mining method based on the improved Adaboost algorithm is completed.

3. Simulation experiment research

This paper studies the new energy field abnormal data mining method based on the improved Adaboost algorithm, and conducts experiments to test the effectiveness of the model.

3.1. Experiment content

In this experiment, the anomaly data mining method mentioned in literature[2] and the conventional data mining method mentioned in literature[3] are selected as comparison items, and compared with the new energy station anomaly data mining method based on improved Adaboost algorithm studied in this paper. According to the mining completion degree of the three data mining methods, the performance of the three methods is compared.

3.2. Source of experimental data

This experiment test data provided by the new energy terminal data warehouse, first, the history of a certain period of time from the original database operations run the database data extraction to experiment, and the data transformation, data cleaning and data analysis, such as operation, combined with the history of the terminal operation monitoring report, to determine the proportion of abnormal data test data set.

3.3. Test platform construction

The test uses a simulation platform equipped with Hadoop distributed system architecture, which is characterized by high fault tolerance and provides high throughput to process the data required by the test. The test platform consists of an IBM server and 4 hosts, among which the IBM server is configured with 16G memory, dual gigabit network card, the host is configured with I7 processor and 8G memory. The computers are connected by gigabit switch, and all the network cables use gigabit network cables.

The operating system of all the physical machines used above is 64-bit ubuntu10.04 and Hadoop version 2.2.0. The operation parameters are selected as shown in the table 1.

| Parameter name | The parameter value | instructions |
|----------------|---------------------|--------------|
| block.size     | 11274618            | HDFS data block size, the smallest unit of data transfer |
| replication    | 2                   | Number of file backups |

Table 1 Selection of Hadoop operation parameters
### Table 1

| data.dir          | /Hadoops1 | Data files are stored locally in directories |
|-------------------|-----------|---------------------------------------------|
| mapred.reduce.parel|el.copy   | The parallel copier that can be started in the Reduce phase obtains the output of the data |
| Io.sortmb         | 200       | Increase the upper memory limit for sorting in Reduce |

### 3.4. Test results and analysis of data mining methods

Three data mining methods were used to mine abnormal data from the new energy station selected by the experiment, and different experimental data volumes and data mining processing time were set. Compare the experimental data and draw the corresponding conclusion. The experimental results are shown in Figure 2.

![Graph](image1.png)

(a) Test results of the proposed method

![Graph](image2.png)

(b) Literature [2] method test results
The three groups of test results a, b, and c in Figure 2 correspond to the proposed method, the abnormal data mining method mentioned in literature[2] and the experimental results of genetic algorithm for the regular data mining method mentioned in literature[4]. a shows that the anomalous data mining completion degree always remains between 85% and 99% for different sizes of data volumes within different data processing times, and the data mining completion fluctuates little and is relatively stable; under the same conditions, the abnormal data mining completion fluctuates between 70% and 82% in graph b. Graph c shows that the abnormal data mining completion fluctuates between 68% and 92%, but when the tested data volume is 10GB and 20GB, the fluctuation is larger and the mining completion shows a significant change. Moreover, the data mining completion degree of the methods in literature[2] and literature[4] reached its peak for the first time significantly later than the proposed method, indicating that the methods in literature[2] and literature[4] have low processing efficiency. Comparing the three sets of results shows that under the same conditions, when the proposed method is used for processing, the degree of completion and efficiency of abnormal data mining is higher.

4. Conclusion
Aiming at the problems of traditional data mining methods, this paper studies an abnormal data mining method for new energy stations based on the improved Adaboost algorithm, and verifies that this method has better performance than traditional methods through simulation experiments. In the future research, the performance of the data mining method should be further improved from the perspective of solving the defects of the neural network used in this method.

Acknowledgments
This project was supported partially by science and technology project of State Grid Shandong Electric Power Company, named Research on key technologies of data mining and security protection for supporting and regulating cloud platform - topic 4: state assessment and deepening application of new energy data (No. 520626190011).

Reference
[1] Fabrizio Marozzo, Domenico Talia, Paolo Trunfio. A Workflow Management System for Scalable Data Mining on Clouds.[J]. IEEE Trans. Services Computing,2018,11(3): 480-492.
[2] Jung-Hyun Lee, Maher Ibrahim Sameen, Biswajeet Pradhan, et al. Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods[J]. Geomorphology,2018,303: 284-298.
[3] Evan Y. W. Yu, Anke Wesselius, Christoph Sinhart, et al. A data mining approach to investigate food groups related to incidence of bladder cancer in the BLadder cancer Epidemiology and Nutritional Determinants International Study[J]. British Journal of Nutrition, 2020, 124(6): 611-619.

[4] Brindley, Goulding, Wilson. Generating vague neighbourhoods through data mining of passive web data[J]. International Journal of Geographical Information Science, 2018, 32(3): 498-523.

[5] Luo Changyong, Yu He, Yang Tao, et al. Data Mining and Systematic Pharmacology to Reveal the Mechanisms of Traditional Chinese Medicine in Recurrent Respiratory Tract Infections’ Treatment[J]. Evidence-Based Complementary and Alternative Medicine, 2020, (1): 1-13.

[6] Chen Yan, Zhengqun Wang, Chunlin Xu. Gentle Adaboost algorithm based on multi-feature fusion for face detection[J]. The Journal of Engineering, 2019, 2019(15): 609-612.

[7] Xiaofu Sun, Yue Li. Denoising of desert seismic signal based on synchrosqueezing transform and Adaboost algorithm[J]. Acta Geophysica, 2020, 68: 403-412.

[8] Seol Dahee, Jung Soonho, Kim Bora, et al. Data Mining of Heterogeneous Electrical Conduction in the Electrode Components of Fuel Cells.[J]. ACS applied materials & interfaces, 2020, 12(20): 23576-23583.

[9] Maize Mine: A Data Mining Warehouse for the Maize Genetics and Genomics Database.[J]. Frontiers in plant science, 2020, 11: 592730-592730.

[10] Chen Yan, Zhengqun Wang, Chunlin Xu. Gentle Adaboost algorithm based on multi-feature fusion for face detection[J]. The Journal of Engineering, 2019, 2019(15): 3313-3322.

[11] Aboozar Taherkhani, Georgina Cosma, T.M. McGinnity. AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning[J]. Neurocomputing, 2020, 404: 351-366.

[12] Xiaobin Xu, Haibin Duan, Yanjie Guo, et al. A cascade adaboost and CNN algorithm for drogue detection in UAV autonomous aerial refueling[J]. Neurocomputing, 2020, 408: 121-134.

[13] Tuan M. Pham, Danh C. Doan, Eckhard Hitzer. Feature Extraction Using Conformal Geometric Algebra for AdaBoost Algorithm Based In-plane Rotated Face Detection[J]. Advances in Applied Clifford Algebras, 2019, 29(4): 1-19.

[14] S. Saravanakumar, P. Thangaraj, A Computer Aided Diagnosis System for Identifying Alzheimer’s from MRI Scan using Improved Adaboost[J]. S. Saravanakumar; P. Thangaraj, 2019, 43(3): 76.

[15] Lamia Ben Amor, Imene Lahyani, Mohamed Jmaiel. AUDIT: AnomaloUs data Detection and Isolation approach for mobile healThcare systems[J]. Expert Systems, 2020, 37(1): e12390.

[16] Yi-Ting Chen, Edward W. Sun, Yi-Bing Lin. Merging anomalous data usage in wireless mobile telecommunications: Business analytics with a strategy-focused data-driven approach for sustainability[J]. European Journal of Operational Research, 2020, 281(3): 687-705.

[17] Minchul Kang, Charles A. Day, Anne K. Kenworthy. A novel computational framework for D ( t ) from Fluorescence Recovery after Photobleaching data reveals various anomalous diffusion types in live cell membranes[J]. Traffic, 2019, 20(11): 867-880.

[18] Matthew R. Pintar, William J. Resetarits. Aquatic Beetles (Coleoptera) of the University of Mississippi Field Station, Lafayette County, Mississippi, USA.[J]. The Coleopterists Bulletin, 2020, 74(2): 351-369.

[19] Ciarán D. Beggan, Laurence Billingham, Ellen Clarke. Estimating external magnetic field differences at high geomagnetic latitudes from a single station[J]. Geophysical Prospecting, 2018, 66(6): 1227-1240.

[20] Upala Mukhopadhyay, Debasis Majumdar, Debabrata Adak. Evolution of dark energy perturbations for Slotheon field and power spectrum[J]. The European Physical Journal C: Particles and Fields, 2020, 80(6): 593.

[21] Kamyr Mehranzamir, Mahdi Davarpanah, Zulkurnain Abdul-Malek, et al. Discriminating cloud to ground lightning Flashes based on wavelet analysis of Electric field signals[J]. Journal of Atmospheric and Solar-Terrestrial Physics, 2018, 181: 127-140.