Journal on Artificial Intelligence
DOI:10.32604/jai.2021.016706
Article

Tech Science Press

An Anomaly Detection Method of Industrial Data Based on Stacking Integration

Kunkun Wang[1,2] and Xianda Liu[2,3,4,*]

[1]College of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang, 110159, China
[2]Key Laboratory of Networked Control Systems, Chinese Academy of Sciences, Shenyang, 110016, China
[3]Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China
[4]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China
[*]Corresponding Author: Xianda Liu. Email: liuxianda@sia.cn
Received: 09 January 2021; Accepted: 15 March 2021

Abstract: With the development of Internet technology, the computing power of data has increased, and the development of machine learning has become faster and faster. In the industrial production of industrial control systems, quality inspection and safety production of process products have always been our concern. Aiming at the low accuracy of anomaly detection in process data in industrial control system, this paper proposes an anomaly detection method based on stacking integration using the machine learning algorithm. Data are collected from the industrial site and processed by feature engineering. Principal component analysis (PCA) and integrated rule tree method are adopted to reduce the dimension of the process data, which can restore the original feature information of the data to the maximum extent. Random forest (RF), Adaboost, XGboost, SVM were selected as the first layer of basic learners. Logistic regression (LR) was used as the secondary learner to build the exception detection model based on stacking integrated method. TE data was used to train the base learner model and the integrated model. By comparing and analyzing the experimental results of between integrated model and each basic learning model. By comparing and analyzing the experimental results of the constructed anomaly detection model and the basic learning model, the accuracy of process data anomaly detection is effectively improved, and the false alarm rate of process data anomaly detection is effectively reduced.

Keywords: Industrial control system; anomaly detection; random forest; SVM; stacking

1 Introduction

As we all know, the industrial control system is the core of the entire industrial production process. It used to improve the automation level of industrial production, improved production efficiency and stability. At present, industrial control systems are widely used in important fields such as petrochemicals, smart grids, transportation, and manufacturing [1].

With the rapid development of the Industrial Internet in recent decades, industrial control systems and IT networks have been deeply integrated. The traditional control system has become open and fragile. Hackers attacked the industrial control system by attacking the network, attacking the industrial control system by assaulting the network. The traditional industrial control system is relatively closed and relatively independent, and it is not easy to be attacked [2]. The security risks in cyberspace are increasing. There is a large difference between industrial control system and IT network. There are a large number of private protocols in the industrial control system, and the network space has a unified protocol at each level. Due

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

to production requirements, industrial control systems pursue real-time and accuracy, and data delays and false alarms can cause serious attack consequences. The focus of IT network security is the confidentiality of information, followed by availability and integrity. The upgrade and replacement of industrial control systems requires a strong monetary price. It is not like network security that can be updated by downloading patches. Attacks on industrial control systems will directly affect the industrial production environment, causing a devastating blow to industrial production.

Since the 1990s, the Internet has grown exponentially, expanding and penetrating into various fields of economy and society explosively. It is currently widely used in new application fields such as industrial control systems, cloud computing, cloud services, and mobile payments. The risks of network security are also increasing [3–4]. Anomaly detection is an important tool for intrusion detection. By comparing and matching the behavior of the process data with the normal behavior library, it can be found whether the data is abnormal, which can effectively detect unknown network attacks. The methods used for anomaly detection mainly include statistical learning, machine learning, and deep learning. Among them, the application of machine learning is more extensive, through the continuous accumulation of data from industrial control systems, and then independent learning, which can effectively reduce the false alarm rate of anomaly detection. The idea of integrated was first proposed by Dasarathy et al. [5]. Boosting integration method was proposed by Schapire et al. [6], which promoted the weak learner to a strong one and trained the next base learner by constantly changing the weights of the misclassified samples, so as to achieve the optimal results eventually. In 1992, Wolpert proposed the Stacking Generalization Model (Stacking) [7].

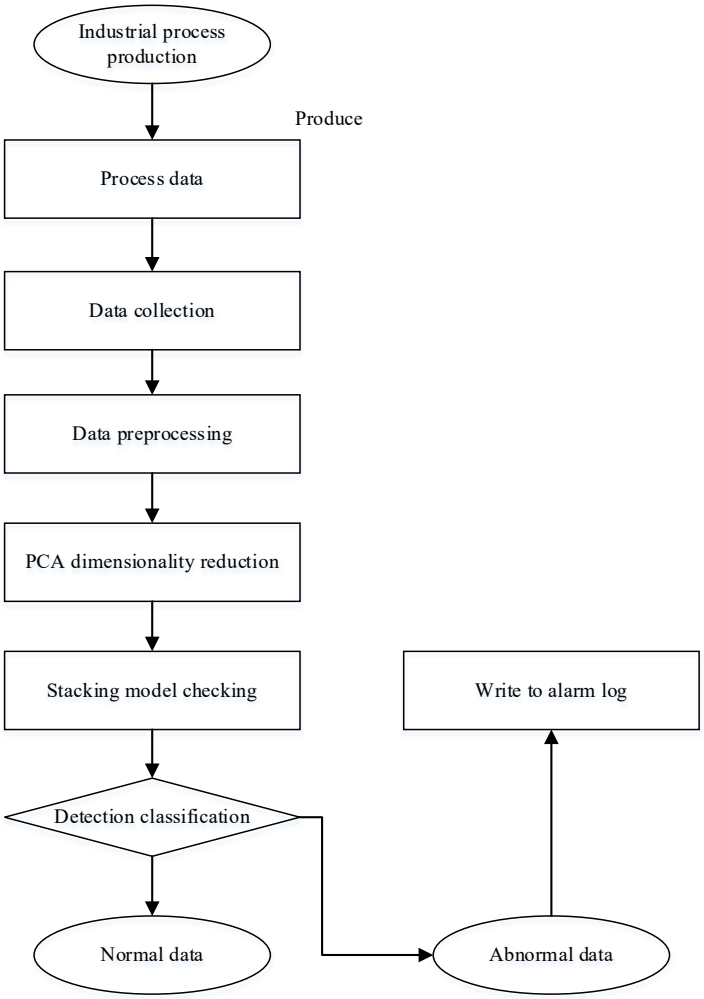

Figure 1: Flow chart of anomaly detection method

Random forest algorithm is an extended branch of Bagging algorithm [8]. The decision tree is selected as a simple learner and random attributes are added to the training and learning of the decision tree. Dietterich listed ensemble learning as the top four research directions in the field of machine learning in Aimagazine magazine [9]. In 2001, Zhou et al. put forward the idea of "selective integration" [10–11], pointing out that the number of basic learning will affect the prediction speed and increase the running time and storage space. Better results can be obtained by selecting some of the basic learners to build integrated learners [12]. The industrial production environment is characterized by noise and high data dimensions, and the constructed anomaly detection model should have high real-time performance. In order to reduce noise interference and prevent over-fitting, the constructed anomaly detection model should have good robustness [13]. This paper proposes an anomaly detection model based on stacking integrated supervised learning algorithm. First, the principal component analysis method is used to reduce the dimensionality of the data. By constructing a two-layer learner to train the data samples, the accuracy of anomaly detection can be improved.

The flowchart of process data exception detection method based on Stacking integration proposed in this paper is shown in Fig. 1. Collecting sequence data that produced in industrial production process. These data are matched and processed, engineering features are processed, dimensionality is reduced by principal component analysis method, and an anomaly detection model is constructed to detect the data, and the detection results are obtained. Through the test results, the industrial control system equipment is judged to be abnormal to prevent the occurrence of dangerous conditions.

This paper proposes an anomaly detection model based on Stacking integration for industrial data. Through PCA dimensionality reduction, high-dimensional and complex process data can be reduced to the dimensions required by the algorithm, based on Random Forest, SVM, XGboost, Adaboost. The Stacking integrated model composed of the base learner improves the accuracy rate and reduces the false alarm rate.

2 Build an Anomaly Detection Method Based on Stacking Integration

The theoretical basis of ensemble learning is PAC theory [13], strong learning and weak learning theory [14]. We choose a simple learner as the basic learner, and through some methods and strategies, the basic learner is integrated and converted into a strong learner [15]. There are three integrated frameworks for ensemble learning. The distribution is boosting, bagging and stacking. The experiment used in this article is the third Stacking integrated learning method.

2.1 Stacking Algorithm

The thought of Stacking algorithm is to first train the base learner with the data set, and generate a new sample set through the base learner training. Use the new sample set to train the secondary learner. Stacking integrated base learner chooses heterogeneous algorithm as base learner. For training samples, cross-validation and leave-one-out methods are generally used for training. The data set is first divided into training set and test set. A cross-validation method is adopted to train and predict the base model. The prediction result of each base model is used as a feature of the training data of the secondary learner. The average of the prediction results is taken as a feature of the new sample, and the secondary The learner makes predictions. Use logistic regression algorithm as secondary learner [16–18]. Seewald proposed that using different attributes in MLR can make the algorithm more time-efficient and effective [19–20]. Fig. 2 is a schematic diagram of the constructed Stacking algorithm.

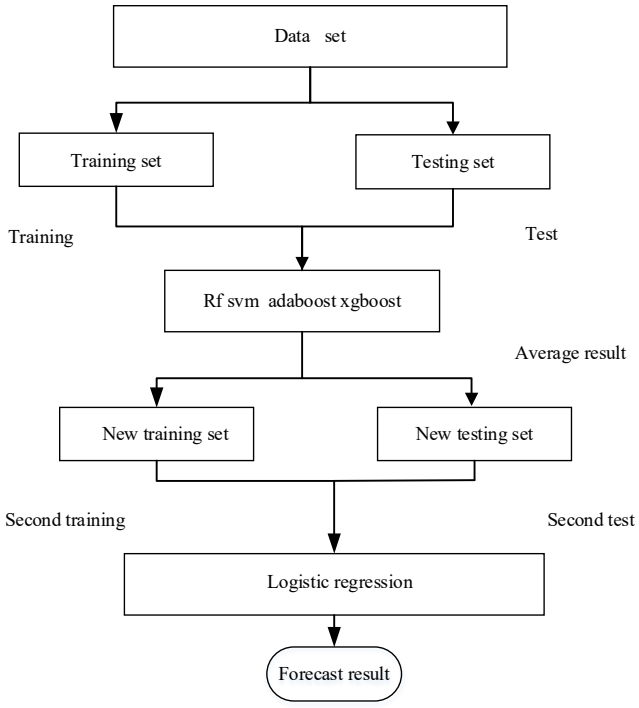

Figure 2: Principle diagram of stacking algorithm

2.2 Adaboost of the Base Learner Model

The idea of Adaboost algorithm is realized by changing the distribution of sample data. The weight of each sample next time is determined according to whether the classification of each sample in the previous training set is correct and the accuracy of the previous overall classification. The modified weight data set is trained by the next base learner, and finally the base learner obtained from each training is combined as the final learner.

2.3 Random Forest of Base Learner Model

Random forest algorithm is an extended algorithm of bagging algorithm, which uses decision tree as a base learner to construct bagging integration algorithm. The bootstrap method is used to generate m training sets, and then a decision tree is constructed for each training set. Part of the features are randomly selected, and the optimal solution is found for feature splitting among the extracted features. The random forest algorithm is relatively simple to implement, and the computational time is relatively small, but the learning ability is very powerful. The main parameters of the random forest algorithm are the number of decision trees n, the depth of the tree d, the lowest node gain m and so on. By adjusting these parameters, the random forest algorithm learns better.

2.4 SVM of Base Learner Model

The idea of the support vector machine model is to find a plane in the sample space of the data set to classify the data samples. Find the most suitable hyper-plane is a research crucial question. The linear equation dividing the hyper-plane is shown below:

$$W^T x + b = 0 \tag{1}$$

$$\mathrm{W} = (\mathrm{W1}; \mathrm{W2}; \ldots \mathrm{Wd}) \tag{2}$$

W is the normal vector, through which the direction of the hyper-plane is determined, and the distance between the hyper-plane and the origin is determined by the b displacement term. These two parameters can determine the hyper-plane required for classification. The hyper-plane can be expressed as (W, b), and

the distance from the data sample point to the hyper-plane is:

$$r = \frac{|W^T x+b|}{||W||} \tag{3}$$

If the hyper-plane (W, b) can correctly classify the samples in the data space, then:

$$W^T x + b \geq 1, \quad y_i = 1 \tag{4}$$

$$W^T x + b \leq -1, \quad y_i = -1 \tag{5}$$

At this time, the optimal hyper-plane is found.

2.5 Xgboost of Base Learner Model

Xgboost algorithm is a kind of Boosting algorithm family. It is improved on the basis of GBDT. It uses regression tree as the basic learner, divides the regression tree into structure part and leaf weight part, and chooses the negative gradient direction to adjust the performance of the regression tree. Regularizing the training samples can reduce the risk of over-fitting. Data can be processed in parallel during operation, which can shorten the algorithm running time Good results for classification and regression problems.

2.6 Logistic Regression of Secondary Learners

Logistic regression algorithms are mainly used in classification problems to find a suitable classification function. The classification function is used to predict the discriminant result of the input data. Construct a loss function. The loss function is used to describe the deviation between the predicted result of the data and the actual situation, and all deviations are processed to evaluate the performance of the algorithm.

The logistic regression formula is expressed as:

$$h(x) = \frac{1}{1+e^{-(W^T x+b)}} \tag{6}$$

The logistic regression algorithm has fast training speed and simple form, which is convenient for the adjustment of output results.

3 Experimental Data Feature Processing

3.1 Experimental Data Description

The data used in this article is the TE Tennessee-Eastman Process Data Set. The entire data set consists of training set and test set.

The TE data set consists of 22 different simulation data. Each data sample contains 52 observation variables, that is, the data dimension is 52 dimensions, and the data is divided into normal data and abnormal process data. In this experiment, normal operating condition data and 4 abnormal operating condition data are selected as data samples. The data is shown in Fig. 3 below.

V_304	LV_305	LV_308	LV_310	LV_312	LV_401	LV_402	TV_1321	TV_1326	TV_302	COUNTER13	FUHEFCS13PER	NetLoad13_A	NetLoad13_B	NetLoad13_C
600.0	600.0	600.0	600.0	600.0	600.0	600.0	600.0	600.0	600.0	600.000000	600.000000	600.000000	600.000000	600.00000
0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.0	1.5	0.0	54.062339	6.923333	1.041067	0.320133	0.32320
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.078039	0.859469	0.009183	0.003266	0.01569
0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.0	1.5	0.0	0.000000	5.000000	1.040000	0.320000	0.32000
0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.0	1.5	0.0	29.950100	6.000000	1.040000	0.320000	0.32000
0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.0	1.5	0.0	59.899700	7.000000	1.040000	0.320000	0.32000
0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.0	1.5	0.0	77.324400	8.000000	1.040000	0.320000	0.32000
0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.0	1.5	0.0	99.899100	9.000000	1.120000	0.400000	0.40000

Figure 3: TE data set

There are a total of 52 nodes in the setting of various equipment parameters in the TE production process, that is, the initial dimension of TE process data is 52 dimensions, which respectively include the factors of feed volume, control valve flow, separator flow, reactor temperature and pressure. Real-time simulation of data in the production process was obtained by using Matlab tool simulation. The data

description is shown in Tab. 1 below.

Table 1: Process data set description

	Role	Level	Count
1	Input	float64	43
2	Input	Int	9
3	Input	Object	1

The constructed data set is 52 dimensions and has no labels. The data needs to be labeled according to the working conditions. The experimental data selected 1 normal working condition and 4 abnormal working conditions. Add the normal working condition data set to label column 10, and label the 2 abnormal working conditions in label columns 1 and 2, respectively. The data of various working conditions are uniformized. Fig. 4 shows the visualization and uniform distribution of processed data labels.

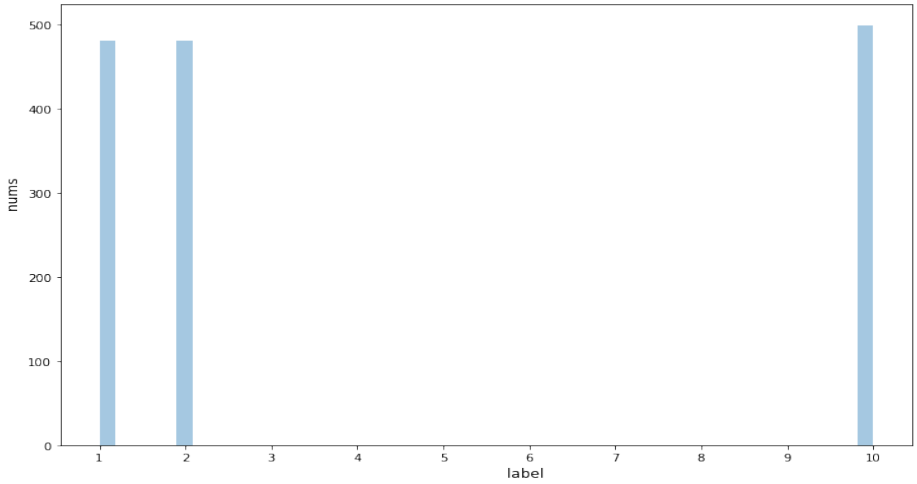

Figure 4: Data view after labeling and normalization

3.2 Data Missing Value Processing

Due to various disturbances in the data in the industrial production process, certain dimensions of the data will be lost. The data with missing dimensionality will produce distortion and produce bad results for the training of the model. Perform missing value query on the data, discard the data with too many missing values, and fill in the data with a small amount of missing values according to the average method, or select 0 to fill in.

3.3 Use Principal Component Analysis to Reduce Dimensionality of Data

Principal component analysis is a widely used dimensionality reduction algorithm. The idea of dimensionality reduction is to convert n-dimensional data into new k-dimensional data. The new k-dimensional features are called principal components. The principle of PCA is to re-establish coordinate axes on the basis of n-dimensional feature data, and construct new k-dimensional feature data. Find a group of orthogonal coordinate axes from the original data space, make the variance of each group of coordinate axes orthogonal, and similar features will be merged. Reduce the dimensionality of the data, and the data after the dimensionality reduction is completely new. PCA dimensionality reduction steps are as follows:

1a) Perform mean normalization on continuous raw data to ensure that the data magnitude of each dimension is the same.

1b) Find the covariance matrix of the feature:

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1} \quad (7)$$

1c) Obtain eigenvalues and eigenvectors according to SVD.

1d) Arrange from largest to smallest characteristic value.

1e) Select k high variance features.

Use the PCA method to reduce the dimensionality of the data, and use the loop iteration method to determine the value of K. If the value of K is too large, the dimensionality reduction effect is not obvious and the training time is too long. If K is too small, it will cause data distortion. The experimental results of different values of K are shown in Fig. 5.

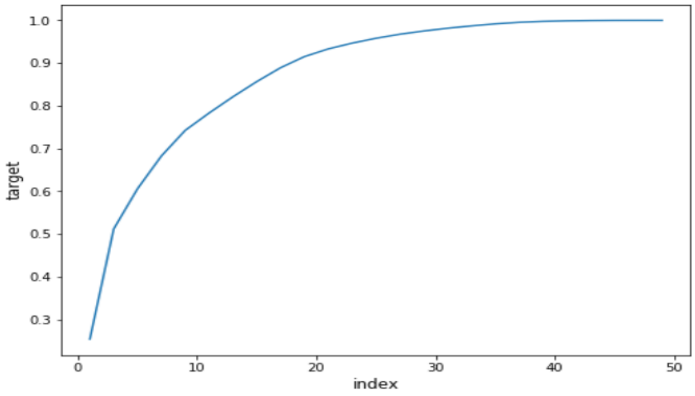

Figure 5: Graph of K value

It can be seen from the experimental curve result graph that when K is 35, the reduction rate can reach 0.95, and the experimental effect is better.

3.4 Feature Selection on Data

Use the integrated rule tree to perform feature selection on the reduced data. The idea of the rule tree model is to divide the data set features into multiple school-scale data groups by selecting the appropriate split point, and when selecting the split point, it contains all the features of the data and the division of a single feature split point. The Gini coefficient and cross entropy are generally selected as the method of measuring purity. The regular tree model is to generate different tree models by using different characteristics and the uncertainty of random sampling to ensure the generalization performance of the results to prevent the model from falling into over-fitting. The constructed tree model has good robustness to process data. According to the Gini coefficient change value of each feature in different rule tree models as the basis of feature importance, the feature importance index map is generated.

In this experiment, the principal component analysis method and the integrated rule tree feature selection method are selected to perform dimensionality reduction operations on the data, which can reduce the dimensionality of the data while retaining important features, which can better restore the original features of the data and make the model The training effect is better.

4 Analysis of Results

4.1 Model Training

Adjust the parameters of the basic model. This paper selects the method of combining loop iteration and cross-validation to adjust the parameters of the basic model. Select 10 times of cross-validation, you will get 10 different data sets, and output 10 results. The parameters of the optimal result are selected as the parameters of the basic learner model.

4.2 Comparative Analysis of Model Results

Precision and deviation are used to evaluate model performance. The precision rate is calculated as shown in Tab. 2.

Table 2: Forecast result matrix

The true situation	Forecast result	
	True example	False example
True example	TP	FN
False example	FP	TN

Precision rate:

$$\mathbf{P} = \frac{TP}{TP+FP} \tag{8}$$

P represents the precision, TP represents the correct positive example of the prediction, and FP represents the negative example of the correct prediction. The distribution uses TE data to train the base learner model and the integrated model, and compares and analyzes the experimental results of each model.

The random forest of the base learner is optimized by using the grid search method. The main parameters of the random forest are the number of decision trees n estimators, the maximum depth of the tree max depth, and the number of decision trees is the parameter of the grid search method to find the optimal parameters. The experimental results are shown in Fig. 6.

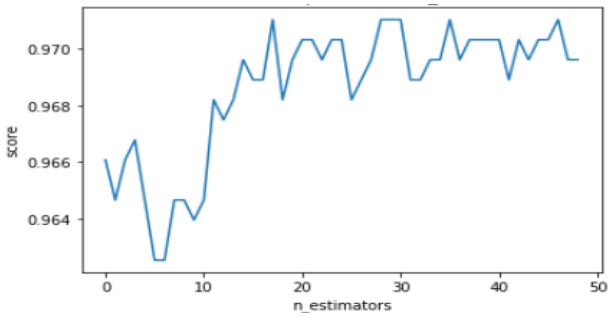

Figure 6: Grid search method to find optimal parameters

As can be seen from the above figure, the number of decision trees ranges from 0 to 50. When the number of decision trees is set to 30, the accuracy of the random forest algorithm is the highest, which is 0.971. At this time, the maximum depth of the tree is set to 2, and the number of jobs is –1. The learning curve query of random forest is shown in Fig. 7.

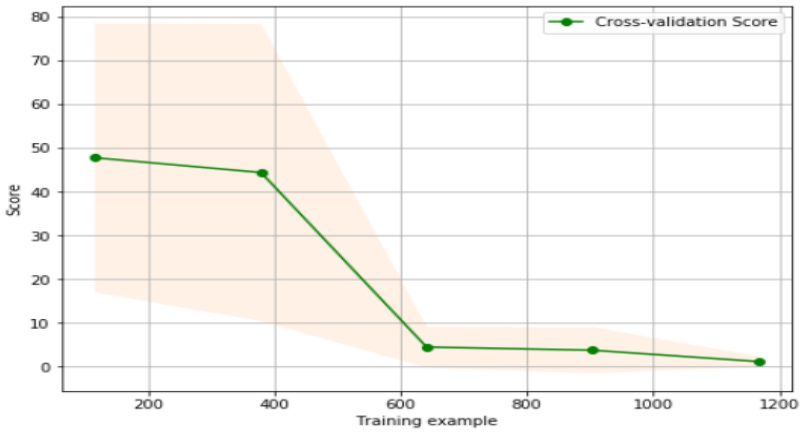

Figure 7: Random forest learning curve

It can be seen from the learning curve that as the training samples increase, the cross-validation bias gradually decreases, and tends to stabilize when the number of samples reaches 800.

Optimize the parameters of the base learner model adaboost algorithm. The research direction of this paper is the problem of multi-class abnormality detection of process data, so SAMME.R is selected as the default parameter, a single-level decision tree is selected, finally the number of trees is determined by the grid search method. The experimental results are shown in Fig. 8.

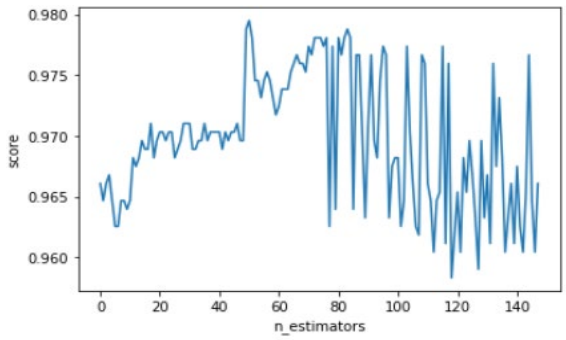

Figure 8: Adaboost parameter optimization accuracy chart

According to the graphical results, when the number of decision trees is 51, the highest accuracy rate is 0.978.

This experiment uses the accuracy rate and the false alarm rate as the experimental evaluation criteria. The experimental comparison results are shown in Fig. 9.

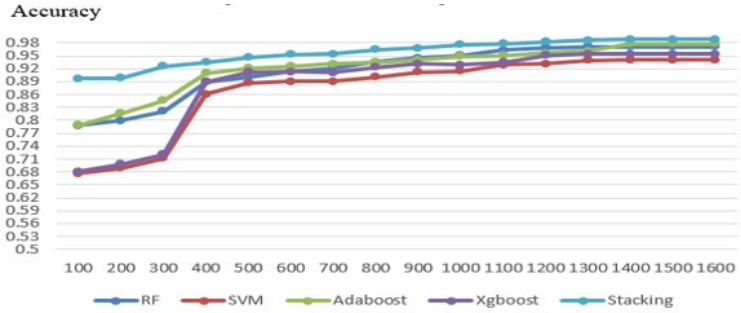

Figure 9: Comparison chart of experimental model accuracy

From the comparison chart of the experimental results of the base model and the integrated model, it can be seen that the curve starts to converge after training with 1000 sample sets. Compared with the base model, the ensemble test model has significantly improved accuracy and faster convergence speed, and good experimental results have been obtained. Tab. 3 below is a comparison table of the accuracy rate and false alarm rate of each model. The accuracy rate is the proportion of the number of correctly classified samples to the total number of samples.

Table 3: Accuracy comparison

Learner Model	Accuracy	Error
Adaboost	0.978	0.022
Svm	0.941	0.059
Xgboost	0.954	0.046
Random Forest	0.971	0.029
Stacking Ensemble	0.988	0.012

From the comparison of anomaly detection accuracy results, it is found that the accuracy of the built integrated anomaly detection model is effectively improved, and the false alarm rate is greatly reduced. This paper proposes to build a process data anomaly detection model based on a supervised learning algorithm. Through feature engineering processing of process data, principal component analysis is used to reduce the dimensionality of the data, and then the Stacking integrated anomaly detection model is built, which effectively improves the anomaly detection Accuracy, this model method can be applied to anomaly detection of industrial and manufacturing process data, and can improve the effect of anomaly detection.

5 Conclusion

This paper proposes an anomaly detection model based on Stacking integration for process data production conditions. Through PCA dimensionality reduction, high-dimensional and complex process data can be reduced to the dimensions required by the algorithm, based on random forest, svm, xgboost, adaboost The Stacking integrated model composed of the base learner improves the accuracy rate and reduces the false alarm rate. It can judge the operating status of the industrial equipment according to whether the process data is abnormal, and help the staff to take security measures more quickly according to the abnormal situation. The built ensemble model still has many directions worth studying. Improving the accuracy and reducing the running time of the algorithm are the next research interests.

Funding Statement: This work is supported by projects: “Industrial Internet security standard system and test verification environment construction” of Industrial Internet Innovation and Development Project in 2018 and “Shenyang Science and Technology Development” [2019] No. 66 (Z191001).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. K. Bi and S. L. Xu, “Research status and development of intrusion detection technology,” *Software Guide*, vol. 9, no. 11, pp. 152–154, 2010.

[2] Y. X. Lai, Z. H. Liu, X. T. Cai and K. X. Yang, “Overview of intrusion detection research in industrial control system,” *Journal on Communications*, vol. 38, no. 2, pp. 143–156, 2017.

[3] K. Y. Zhang, T. M. Chen and C. Yan, “Research progress on industrial control system security and anomaly detection,” *Information Security Research*, vol. 3, no. 7, pp. 624–632, 2017.

[4] K. Yim, A. Castiglione, J. H. Yi, M. Migliardi and I. You, “Cyber threats to industrial control systems,” in *Proc. of the 7th ACM CCS Int. Workshop on Managing Insider Security Threats*, Colorado, USA, pp. 79–81, 2015.

[5] S. N. Atluri and S. Shen, “Global weak forms, weighted residuals, finite elements, boundary elements & local weak forms,” in *The Meshless Local Petrov-Galerkin (MLPG) Method*, 1st ed., vol. 1. Henderson, NV, USA: Tech Science Press, pp. 15–64, 2004.

[6] B. V. Dasarathy and B. V. Sheela, “A composite classifier system design: concepts and methodology,” in *Pro. of the IEEE*, vol. 67, no. 5, pp. 708–713, 1979.

[7] L. K. Hansen and P. Salamon, “Neural network ensembles,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[8] L. Breiman, “Stacked regressions,” *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.

[9] A. Alhussain, H. Kurdi and L. Altoaimy, “A neural network-based trust management system for edge devices in peer-to-peer networks,” *Computers, Materials & Continua*, vol. 59, no. 3, pp. 805–815, 2019.

[10] T. G. Dietterich, “Machine learning research: four current directions,” *AI Magazine*, vol. 18, no. 4, pp. 97–136, 1997.

[11] B. J. Gu, W. L. Xiong and Z. H. Bai, “Human action recognition based on supervised class-specific dictionary learning with deep convolutional neural network features,” *Computers, Materials & Continua,* vol. 63, no. 1, pp. 243–262, 2020.

[12] G. Y. Yang, J. Q. Zeng, M. K. Yang, Y. F. Wei and X. Q. Wang, "Ott messages modeling and classification based on recurrent neural networks," *Computers, Materials & Continua*, vol. 63, no. 2, pp. 769–785, 2020.

[13] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[14] X. M. Chang, Y. Qiu, S. T. Su and D. L. Yang, "Data cleaning based on stacked denoising autoencoders and multi-sensor collaborations," *Computers, Materials & Continua*, vol. 63, no. 2, pp. 691–703, 2020.

[15] Y. P. Zhang and L. Zhang, "Machine learning theory and algorithm," Beijing, China: Science Press, 2012.

[16] B. Hossain, T. Morooka, M. Okuno, M. Nii, S. Yoshiya *et al.,* "Surgical outcome prediction in total knee arthroplasty using machine learning," *Intelligent Automation & Soft Computing*, vol. 25, no. 1, pp. 105–115, 2019.

[17] A. K. Seewald, "How to make Stacking better and faster while also taking care of an unknown weakness," in *Proc. of the Nineteenth Int. Conf.*, Sydney, Australia, pp. 554–561, 2002.

[18] M. E. Mamoun, Z. Mahmoud and S. Kaddour, "SVM model selection using PSO for learning handwritten arabic characters," *Computers, Materials & Continua*, vol. 61, no. 3, pp. 995–1008, 2019.

[19] Y. Freund and R. E. Schapire, "Special invited paper, additive logistic regression: a statistical view of boosting: discussion," *The Annals of Statistics*, vol. 28, no. 2, pp. 391–393, 2020.

[20] Z. Q. Wang, R. Jiao and H. P. Jiang, "Emotion recognition using WT-SVM in human-computer interaction," *Journal of New Media*, vol. 2, no. 3, pp. 121–130, 2020.