Improved Extremely Randomized Trees Model for Fault Diagnosis of Wind Turbine

Bo Zhang

Department of Computer, North China Electric Power University, Baoding 071003, China;

Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding 071003, China.

Abstract

The operational data collected by wind farm have the characteristics of less fault data and high sample dimension. However, it is difficult for diagnostic model to accurately diagnose the fault type, and the high sample dimension will increase the cost of model training. In order to solve these two problems, this paper introduces Generative Adversarial Networks and Stack Sparse Autoencoder. Firstly, a small number of fault data are input into GAN neural network, and the new data approximate to the fault data are generated by learning the feature distribution of fault data, so as to improve the sample class imbalance in fault diagnosis of wind turbine. Secondly, the feature distribution of wind power data are extracted layer by layer by Stack Sparse Autoencoder, and the fault type related features in high-dimensional data is determined by combining the reconstruction error of each layer. Finally, the extreme random tree is constructed based on the reduced dimension data to realize the fault diagnosis of wind turbine. The results on a wind farm data set show that, compared with the traditional over sampling methods, the proposed method can effectively balance the fault sample set and improve the diagnostic accuracy of the diagnosis model in the case of insufficient fault samples.

Keywords

Fault Diagnosis; Generative Adversarial Networks; Imbalanced Data; Feature Dimensionality Reduction; Extremely Randomized Trees.

1. Introduction

Wind turbine has been operating in harsh natural environment such as thunderstorm, sun exposure, snow and so on for a long time. In addition, the internal structure of the wind turbine is complex, and each component is easy to be affected by different alternating loads, which is easy to cause various electrical and mechanical failures of wind turbine [1]. Once the wind turbine fails, it will lead to long-term shutdown, which not only affects the generating efficiency of the generator set, but also causes catastrophic accidents, resulting in huge economic losses. Therefore, real-time monitoring and diagnosis of wind turbine operating conditions is of great significance.

Currently, the mainstream wind turbine fault diagnosis methods are model-based method[2] and data-driven method[3]. The former establishes accurate mathematical or physical models[4] based on the physical characteristics of wind turbines and their subsystems, which can overcome the difficulty in obtaining fault data. However, due to the randomness of wind energy and the complexity of fault mechanism of wind turbines, accurate modeling is more difficult.The latter mostly uses supervisory control and data acquisition (SCADA) system to mine potential fault information from real-time collected operation data to realize the monitoring and diagnosis of unit status without establishing accurate model and system prior knowledge, so it is widely used[5,6]. At present, there are few researches on fault diagnosis of wind turbine. The main methods include Support Vector Machine (SVM)[7], Neural Network[8] and Swarm Intelligence algorithm[9]. However, these methods are based on sufficient and balanced distribution of sample data as the premise. However, in the actual wind farm, it is often expensive to obtain sufficient fault data. Therefore, most of the obtained data are normal data, and only a small part is fault data. Moreover, the wind power data obtained based on

SCADA often has a high dimension, and the class imbalance of data will affect it.The diagnostic performance of traditional machine learning methods[10], while high-dimensional data increases the cost of model training.

In accordance with the problem of data class imbalance, under sampling[11] or over sampling method[12] are mainly used to solve the problem.However, on the one hand, the under sampling method discards most class samples, which easily leads to the loss of important information. On the other hand, the traditional oversampling method is a simple copy of most class samples, without adding effective classification information[13]. Based on the synthetic minority over sampling technique (smote) and its improved method[14], samples are synthesized by using the local prior distribution information of samples, and the samples have no diversity. Therefore, this paper proposes to use generative adversarial networks (GANs)[15] to solve the sample class imbalance problem. GANs It is a new generation model. It can learn the probability distribution of the target data samples to generate forgery samples which are very similar to the target data samples. It is a new generation model that directly compares the distribution of the forged samples and the target samples to train and generate the generated samples that are most likely to approximate the real samples by means of confrontation, which effectively solves the traditional problem The over fitting problem caused by insufficient training samples in the generation process of generation model is rarely used to solve the imbalance problem of wind turbine fault diagnosis data.

Aiming at the problem of high data dimension and high model training cost, stack sparse autoencoder (SSAE) has the advantages of fast learning speed, good generalization performance and strong noise resistance compared with traditional machine learning methods[16]. Xue et al. [17] proposed a rolling bearing fault diagnosis method based on SSAE. Experiments show that the accuracy of the fault data processed by SSAE has been greatly improved, and it is more suitable for the feature extraction task in high-dimensional space. Therefore, in this paper, stacked sparse self encoder SSAE is used to reconstruct high-dimensional data. SSAE (Stack Sparse Autoencoder) adds sparsity limitation to hidden layer neurons on the basis of AE (Autoencoder), enhances the ability of data dimension reduction, and adds the lost packet technology on the basis of SAE (Sparse Autoencoder) to enhance the robustness of cascading between self coding networks.

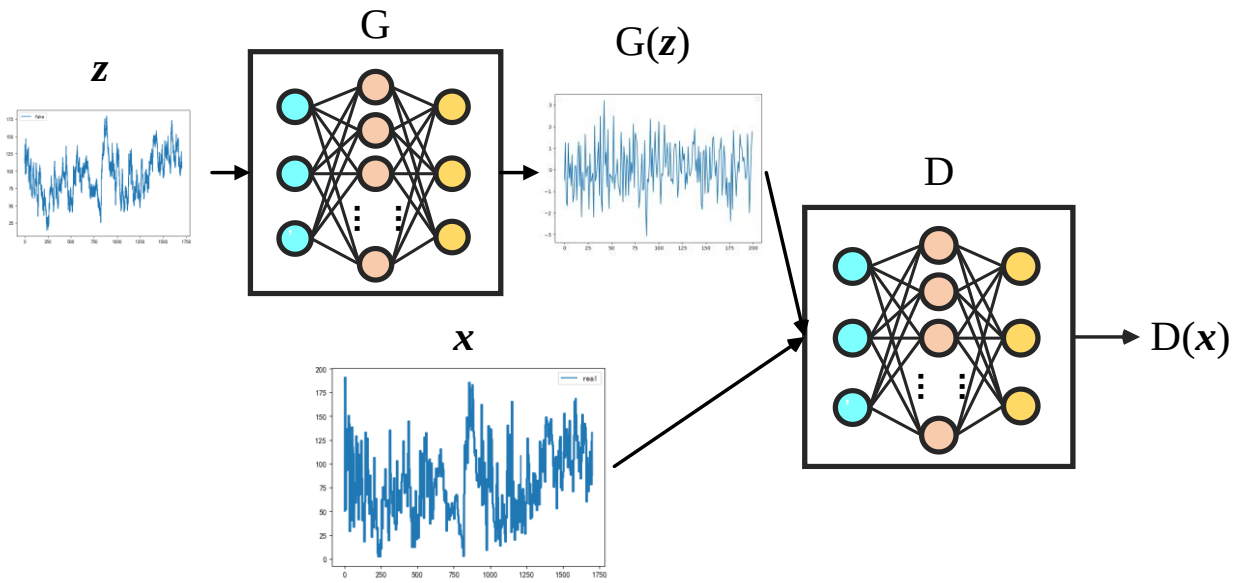

Fig.1 Basic structure of GAN

Extremely Randomized Trees (Extra-Trees, ET) are widely used in the field, Extra trees is an integrated machine learning algorithm[18] proposed by Pierre geurts and others in 2006. As a better classification algorithm applied in the field of fault diagnosis, it integrates multiple decision tree

models to form a classifier. Compared with neural network and support vector machine, it has the advantages of high detection accuracy, less adjustment parameters and easy parallel processing.

To sum up, in order to solve the problem of low fault diagnosis rate caused by unbalanced data and high dimension in wind turbine fault diagnosis, In this paper, a fault diagnosis model based on WGAN-SSAE-ET is proposed. Firstly, the non-equilibrium data is class balanced by using GAN method. Then, SSAE is used to reduce the dimension of high-dimensional data and construct an Extremely Randomized Trees. Finally, a fault diagnosis model which can effectively identify and process high-dimensional and unbalanced data is established. Compared with SVM, KNN and CNN, the results show that the performance of the proposed detection model is better, and the correctness and innovation are verified.

2. Related Work

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) was proposed by Ian goodfill in 2014. It is a generative model, which is divided into two parts: generator g and discriminator D.Among them, G learns the probability distribution of real samples and generates similar false samples; D judges the authenticity of samples by measuring the similarity between real samples and false samples. The network structure is shown in Figure 1. The training process of GANs can be understood as the process of game. G and D are the two sides of the game. G generates more real samples as much as possible to cheat D, and D tries to distinguish the samples generated by G as false samples. Through continuous optimization and iteration of G and D, the final model is in a dynamic balance: the samples generated by G are basically the same as the real samples, and the discrimination probability of D for samples is close to 0.5. The objective formula is defined as follows:

$$\min_{G}\max_{D}V(G,D)=E_{\boldsymbol{x}\square P_d}[\log D(\boldsymbol{x})] +E_{\boldsymbol{z}\square P_z}[\log(1-D(G(\boldsymbol{z})))] \tag{1}$$

Among them, V (G, D) is the loss function, P_d is the real sample distribution, P_z is a group of random noise obeying Gaussian distribution, $D(\cdot)$ is the probability that the sample is true, $G(\boldsymbol{z})$ is the false sample generated by G by inputting random noise variable z.

In order to solve the problems of gradient vanishing and mode collapse in traditional GAN algorithm [19], Wasserstein distance [20] is used to measure the difference between real fault sample distribution P_d and composite sample distribution P_g. It is defined as follows:

$$W(P_d,P_g)=\inf_{\gamma\sim\Pi(P_d,P_g)}E_{(x,y)\sim\gamma}\left[\|x-y\|\right] \tag{2}$$

Among them, $\Pi(P_d,P_g)$ is the set of joint distribution γ with P_d and P_g as edge distribution, $W(P_d,P_g)$ is the infimum of sample expectation under joint distribution γ, x and y are sample points in joint distribution γ.

Compared with the traditional JS distance, Wasserstein distance has superior smoothing characteristics, which can reduce the problem of gradient disappearing in the process of GAN training and improve the stability of training. In addition, there is no need to balance the training level of G and D in the process of GAN training. As long as the training of D is better, the samples generated by G will be more real and diverse.[21]

Since formula (4) cannot be solved directly, Kantorovich Rubinstein dual transformation is performed [20]:

$$W(P_d,P_g)=\sup_{\|D\|_L\le 1}E_{x\sim P_d}[D(\boldsymbol{x})]-E_{x\sim P_g}[D(\boldsymbol{x})] \tag{3}$$

Where, $\|D\|_L\le 1$ means that the Discriminator must satisfy the 1-Lipschitz condition.

In order to ensure the continuity of 1-lipschitz in GAN, this paper improves the objective function of GAN by adding a regular term to the discriminator loss function to carry out gradient penalty [19]. The formula is as follows:

$$R = \lambda E_{\hat{x}\sim P_{\hat{x}}}[(\| \nabla_{\hat{z}} D(\hat{\boldsymbol{x}}) \|_2 - 1)^2] \tag{4}$$

Where, λ is the coefficient of regular term, $\hat{\boldsymbol{x}}$ is obtained by random sampling on the line between real sample $\boldsymbol{x}$ and composite sample $G(\boldsymbol{z})$, and $\|\cdot\|_2$ is 2-Norm.

According to the above contents, in the improved GAN(WGAN-GP) based on Wasserstein distance and gradient constraint, the loss functions of generator and discriminator are as follows:

$$L_G = -E_{z\sim P_z}(D(G(\boldsymbol{z}))) \tag{5}$$

$$\begin{aligned} L_D = {} & E_{\boldsymbol{z}\sim P_z}[\mathrm{D}(\mathrm{G}(\boldsymbol{z}))] - \mathrm{E}_{\boldsymbol{x}\sim P_d}[\mathrm{D}(\mathrm{G}(\boldsymbol{x}))] \\ & + \lambda E_{\hat{\boldsymbol{x}}\sim P_x}[(\|\nabla_{\hat{x}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2] \end{aligned} \tag{6}$$

2.2 Stack Sparse Autoencoder

Based on the characteristics of high latitude and complex data collected by SCADA system, this paper uses AE to extract the features of data. AE is a kind of neural network which uses back propagation algorithm to make the output value equal to the input value. It first compresses the input into the representation of latent space, and then reconstructs the output through this representation. The network model consists of two parts: encoder and decoder, in which the encoder is responsible for compressing the input into a potential spatial representation, and the decoder is responsible for reconstructing the latent spatial representation. As shown in Figure 2, AE network structure is divided into three layers, namely input layer, hidden layer and output layer. The input layer and hidden layer constitute the encoder network, and the hidden layer and output layer constitute the decoder network.

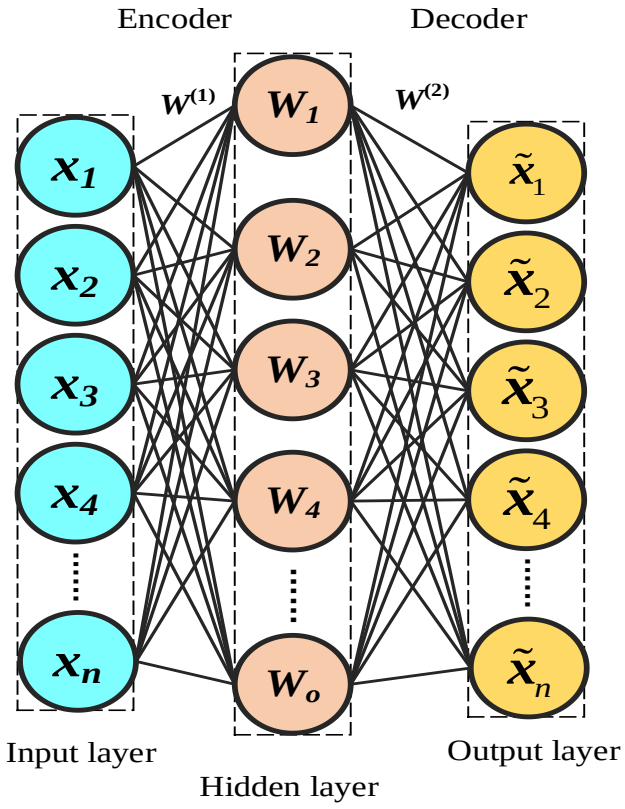

Fig.2 The structure of AE

Assuming that the input sample set is $X=\{x_1,x_2,x_3,\ldots,x_n\}$, the encoder operation formula is:

$$h_{Ei}=s(W_{jk}^{(1)}x_i+b_{jk}^{(1)}) \tag{7}$$

Among them, h_{Ei} is the potential spatial representation of the i-th sample after compression, $W_{jk}^{(1)}$ represents the weight between the j-th neuron in the input layer and the k-th neuron in the hidden layer, $b_{jk}^{(1)}$ represents the offset between the j-th neuron in the input layer and the k-th neuron in the hidden layer, and s represents the activation function.

The operation formula of decoder is as follows:

$$\tilde{x}_i=s(W_{jk}^{(2)}h_{Ei}+b_{jk}^{(2)}) \tag{8}$$

Among them, $\tilde{x}_i$ represents the i-th reconstruction sample, $W_{jk}^{(2)}$ Represents the weight between the j-th neuron in the hidden layer and the k-th neuron in the output layer, $b_{jk}^{(2)}$ Represents the offset between the j-th neuron in the hidden layer and the k-th neuron in the output layer.

In order to ensure the consistency between the reconstructed data and the original data, the reconstruction error should be minimized. The reconstruction error calculation formula is as follows:

$$\min L(x_i,\tilde{x}_i)=\frac{1}{2}\left\|x_i-\tilde{x}_i\right\|_2 \tag{9}$$

For the deep self coding network with n samples as input and M network layers, the total loss function is as follows:

$$J(W,b)=\frac{1}{n}\sum_{i}^{n}L(x_i,\tilde{x}_i)+\frac{\lambda}{2}\sum_{l=1}^{m-1}\sum_{i=1}^{S_i}\sum_{j=1}^{S_j}(W_{ij}^{l})^2 \tag{10}$$

Among them, the first term of equation (5) represents the mean value of reconstruction error of the whole data set, and the second term is the weight attenuation term, in order to suppress the weight update speed and prevent over fitting. λ represents the weight attenuation parameter, S_i represents the number of neurons in layer i-th, W_{ij}^{l} represents the connection weight between layer l-th neuron i and layer (l+1)-th neuron j.

In the training process, a sparsity constraint is added to AE, that is, only part of the hidden layer neurons are activated at the same time, which constitutes SAE. This sparsity expression can extract the correlation features within the data more effectively. The average activation value of neurons in the hidden layer was calculated as follows:

$$\hat{\rho}_j=\frac{1}{n}\sum_{i=1}^{n}h(j,x_i) \tag{11}$$

Where n is the number of training samples, $h(j,x_i)$ represents the activation value of the hidden layer neuron j at a given input x_i.

When the activation value of neurons in the hidden layer is close to zero, it constitutes a sparsity limitation, and $\hat{\rho}_j=\rho,\rho$ is a sparse parameter. In order to make $\hat{\rho}_j$ as close as possible to ρ, a sparse penalty term is added to the loss function of AE:

$$J_p=\sum_{j=1}^{S_i}\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \tag{12}$$

Therefore, the loss function of SAE is as follows:

$$J_{SAE}(W,b)=J(W,b)+\beta J_p \tag{13}$$

Where, β is a super parameter, which is used to control the sparsity penalty factor

Multiple SAE are stacked to form a stacked sparse self encoder (SSAE). The input samples are extracted layer by layer to further reduce the data dimension. The loss function is minimized by the back-propagation algorithm to learn more representative and sparse features. The interference part of the original sample is removed and the sample information is retained as much as possible The low-dimensional features are used for classifier recognition, which can improve the training speed and classification performance of the classifier. The SSAE network structure is shown in Figure 3.

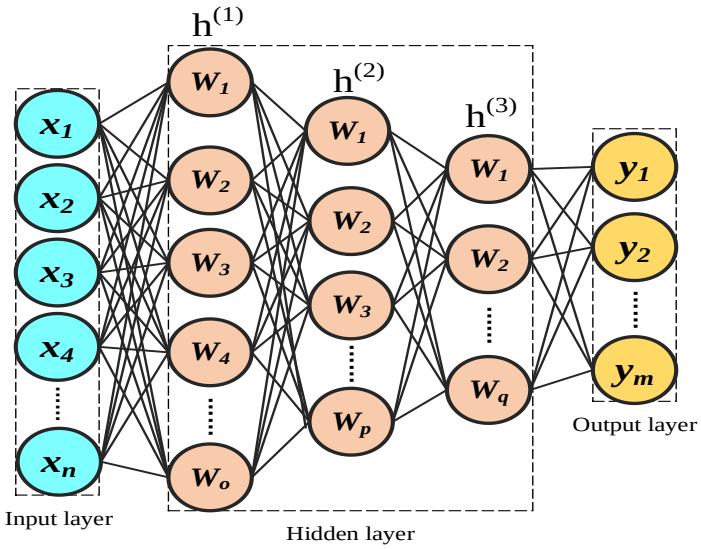

Fig.3 The structure of SSAE

2.3 Extremely Randomized Trees

Extremely Randomized Trees is an integrated machine learning algorithm proposed by Pierre geurts and others in 2006. It integrates multiple base classifiers and votes according to the prediction results of each base classifier, which is usually expressed as t (V, x, x),D) Where t is the classifier model, V is the number of base classifiers, X is the input sample, and D is the sample set

Step 1: given the original data sample set D, sample number n, feature number F. In the Extremely Randomized Trees classification model, each base classifier uses all the samples for training.

Step 2: generate base classifier according to cart decision tree algorithm. In order to enhance the randomness, f features are randomly selected from F features during each node splitting, and the optimal attribute is selected for each node for node splitting, and no pruning is performed in the splitting process. Step 2 is iterated over the split data subset until a decision tree is generated.

Step 3: repeat steps 1 and 2 for Z times to generate Z decision trees to form Extremely Randomized Trees.

Step 4: use test samples to generate prediction results for the generated Extremely Randomized Trees.

Extremely Randomized Trees algorithm is an improved algorithm of random forest algorithm. It improves the generalization ability of the model through the random selection of features and the randomness of node splitting.

3. Methods

GAN is a kind of generation model of learning real sample distribution by confrontation. This model can generate new samples with high quality without pre modeling. In the process of fault diagnosis, due to a few types of abnormal data, the distribution of data set used in intrusion detection is unbalanced. Therefore, this paper uses GAN to generate a small number of training samples to reduce the impact of unbalanced training samples on the accuracy of diagnosis.

SSAE is a deep learning method, which includes input layer, n hidden layer and output layer. It takes AE as basic unit and stacks layer by layer in order to form a deep network structure. It has the ability

of deep feature extraction. It can reduce the dimension of high-dimensional data as much as possible, get the most characteristic data, and get the reconstructed original data, which is easier to be learned by extra trees.

The ET algorithm integrates the classifiers formed by multiple decision tree models, which has the advantages of high detection accuracy, less parameters and easy parallelization. However, when the dimension of the data set is too high, the training time of the algorithm is long and the detection accuracy is low; when the distribution of the data set is unbalanced, the detection results of the algorithm tend to favor the majority of samples. Considering the characteristics of high dimension and class imbalance of wind power data, GAN and SSAE are combined to process wind power data according to the shortcomings of traditional ET algorithm, so as to improve the classification accuracy of ET algorithm. After using GAN to generate a small number of samples, the generated samples are combined with the original data set to form a new and balanced data set. The bagging algorithm is used to sample the new data set to generate multiple balanced data subsets, and then uses SSAE to reduce the feature dimension of each data subset, and each reduced data sample corresponds to each decision tree for training. In the stage of fault diagnosis, voting is carried out by combining the classification results of each decision tree. Finally, all decision trees are collected to form a forest and the classification results are obtained. The fault diagnosis model based on WGAN-SSAE-ET is constructed. The overall framework is shown in Figure 4.

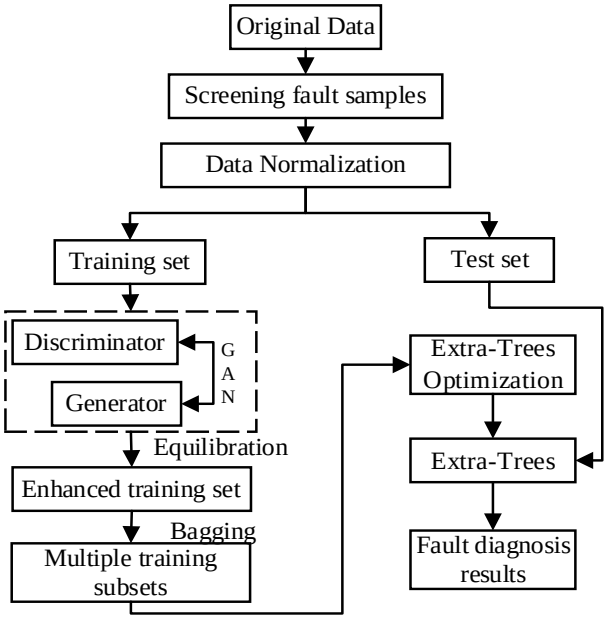

Fig.4 The framework of WGAN-SSAE-ET

3.1 Expansion of Minority Training Data

This paper uses GANto generate countermeasure network to generate the data of health state which is less in training data

Step 1: firstly, five real data sets with health status of f1, f2, f3, f4 and f5 are separated respectively.

Step 2: according to the input format of GANmodel, the 86 dimension data is converted into 10 × 10 matrix vector, and the remaining 14 dimensions are supplemented with 0.

Step 3: given a 100 dimensional Gaussian noise s with a value range of [-1,1], the noise is sent to the generator to generate false samples, and the generated data and real data are sent to the discriminator for training.

Step 4: according to the set number of iterations, the discriminator is trained iteratively until the discrimination result is optimal. At this time, the parameters of the discriminator are fixed and the discrimination results are fed back to the generator.

Step 5: according to the set number of iterations, the training iteration of the generator is carried out until the worst discrimination result is obtained. At this time, the parameters of the generator are fixed.

Step 6: repeat step 4 and step 5 until GANmodel is balanced.

Step 7: combine the generated minority data as expanded samples with the original data, reorganize the expanded samples into 100 dimensional features, and take the first 86 dimensions of data as the expanded samples to get the balanced training data set.

3.2 SSAE Training Process

SSAE is used to extract features from the expanded data, as shown in Figure 6

Step 1: build the first AE, set each rule $\theta_j : x_1, \ldots, x_n, \theta_j$ is the hidden layer neuron of the object network; $x_1, \ldots, x_n$ is the input layer neuron set.

Step 2: determine the connection weight $W\theta_j$ between θ_j and $x_1, \ldots, x_n$. When the input neuron corresponds to the activation element in the rule, then W = 1; otherwise, W = 0, and the weight of residual which has little relationship with θ_j is set to 0. The neuron deviation was set as a random value.

Step 3: use back propagation algorithm to train network and update connection weights.

Step 4: repeat step 1-3 for each AE until all AE have completed training.

3.3 Model Training

The GAN model is used to generate minority class samples, SSAE feature extraction and extra trees algorithm are used to construct parallel design.As shown in Figure 7, the whole parallelization design idea is as follows:

Step 1: firstly, the data sets captured on the network are digitized and normalized, and then the GAN model is expanded with a few samples.

Step 2: integrate the minority class samples generated by GAN model with the original data samples to obtain a new and balanced data set. Through bagging algorithm, the new data set is randomly sampled to generate multiple data subsets with equal number and balanced distribution.

Step 3: each data subset is extracted by SSAE to get the new data subset after reconstruction.

Step 4: each data subset trains the corresponding decision tree model according to the generation mode of the decision tree.

Step 5: gather all decision trees to form Extra-Trees.

4. Experiments

4.1 Data Acquisition and Processing

The experimental hardware environment is Intel (R) Xeon (R) silver 4214CPU@2.2GHzThe software environment is windows 10 operating system, the experimental tools are Python 3.7, and the deep learning framework is tensorflow and keras.

Table 1 Distribution of the samples

Health Status	Sample Size	Training Set	Test set	Unbalanced rate
f0	2515	1761	754	—
f1	183	128	55	13.74
f2	230	161	69	10.93
f3	217	152	65	11.59
f4	160	112	48	15.72
f5	129	90	39	19.50
Total	3434	2404	1030	—

The experimental data were provided by a wind farm in Hebei Province. Six kinds of health states were selected, including f0 as normal state and f1~f5 as abnormal state, which respectively indicated that the position encoder of yaw system was abnormal, the yaw slip rate was too fast, the yaw sensor was faulty, the over temperature protection sensor of yaw system was invalid, and the temperature protection of motor of yaw system triggered abnormal. There are 3404 samples in the data set. The training set and the test set are divided by a ratio of 7:3. The training set contains 2404 samples and the test set contains 1030 samples. Known data imbalance rate is defined as the ratio of majority class to minority class. According to table 1, compared with normal samples, the maximum unbalance rate of various fault samples is 19.50, and the minimum is 10.93. In this paper, the average value of 14.30 is taken as the imbalance rate of the overall class distribution of fault data set to enhance the samples. Finally, the sample set is normalized to ensure the stability of GAN training:

$$\tilde{x}_{ik} = \frac{x_{ik} - x_{min}}{x_{max} - x_{min}} \tag{14}$$

Among them, x_{ik} is the k-th eigenvalue of the sample, $\tilde{x}_{ik}$ is the normalized eigenvalue, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature respectively.

4.2 Evaluating Indicator

For the multi-classification problem, we prefer the model to have a more accurate judgment for the minority class. However, because the sample size of the minority class is relatively small, its misjudgment and missed judgment have less impact on the overall classification accuracy, so it is unreasonable to only use the classification accuracy rate to measure the classification performance of the model. In order to ensure the accurate and comprehensive evaluation of the model performance, the evaluation system based on confusion matrix is adopted in this paper. The definition of confusion matrix is shown in Table 3.

Table 2 Confusion matrix in fault diagnosis of yaw system

Diagnosis / Real	f0	f1	f2	f3	f4	f5
f0	d00	d01	d02	d03	d04	d05
f1	d10	d11	d12	d13	d14	d15
f2	d20	d21	d22	d23	d24	d25
f3	d30	d31	d32	d33	d34	d35
f4	d40	d41	d42	d43	d44	d45
f5	d50	d51	d52	d53	d54	d55

As shown in Table 2, when fault diagnosis is conducted, d_{ii} represents the sample size consistent with the actual state and the diagnosis state, and d_{ij} represents the sample size in which the actual state i is wrongly diagnosed as state j. Therefore, the Precision(P) and the Recall(R) are introduced. Among them, the Precision represents the proportion of correctly diagnosed samples in a certain class of samples diagnosed by the algorithm, which reflects the accuracy of the algorithm; the Recall represents the proportion of a certain type of samples diagnosed, reflecting the comprehensiveness of the algorithm. When evaluating the classification performance of the model for unbalanced samples, F1-score and G-mean are usually used as evaluation indexes. F1-score is the harmonic average of precision and recall. The higher the F1-score is, the better the classification effect is. P, R, F1 score and g-mean are as follows.

$$P_i = \frac{d_{ii}}{\sum_{j=1}^{m} d_{ji}} \tag{15}$$

$$R_i = \frac{d_{ii}}{\sum_{j=1}^{m} d_{ij}} \tag{16}$$

$$F1 = \frac{2\sum_{i=1}^{m} P_i * R_i}{m(\sum_{i=1}^{m} P_i + \sum_{i=1}^{m} R_i)} \tag{17}$$

$$G_{G-\text{mean}} = \frac{\sum_{i=1}^{m} \sqrt{P_i * R_i}}{m} \tag{18}$$

Where, m is the number of sample categories.

4.3 Experimental Results and Analysis

The training set is used as the input of GAN to train it so that it can learn the original data distribution P_d. In the initial stage of training, the distribution of the samples generated by the generator is not similar to the actual sample distribution. With the increase of training rounds, the generating ability of the generator increases gradually, and the distribution of composite samples and the actual sample distribution gradually tend to be consistent. Fig. 5 shows the change of the loss of the generator and discriminator of GAN with the increase of training rounds, in which the abscissa is the training round and the ordinate is the loss value. It can be seen that with the increase of training times, both of them gradually change to the ideal state and finally in the dynamic equilibrium state, which shows that the improvement of GAN by using Wasserstein distance and gradient constraint has good effect, the model is easier to converge and the stability is increased.

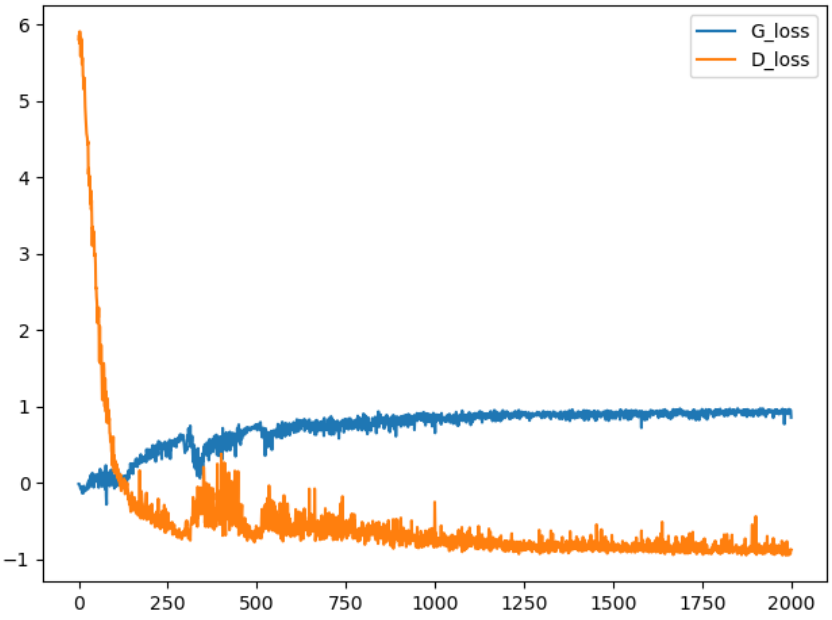

Fig.5 Generator's loss curve and Discriminator's loss curve

Taking the average unbalance rate of 14.30 as the standard, the samples of fault state f1, f2, f3, f4, f5 of yaw system are enhanced by trained GAN model. Taking F1 as an example, the active power characteristics of generator are selected for sample visualization and compared with actual samples, as shown in Figure 6.

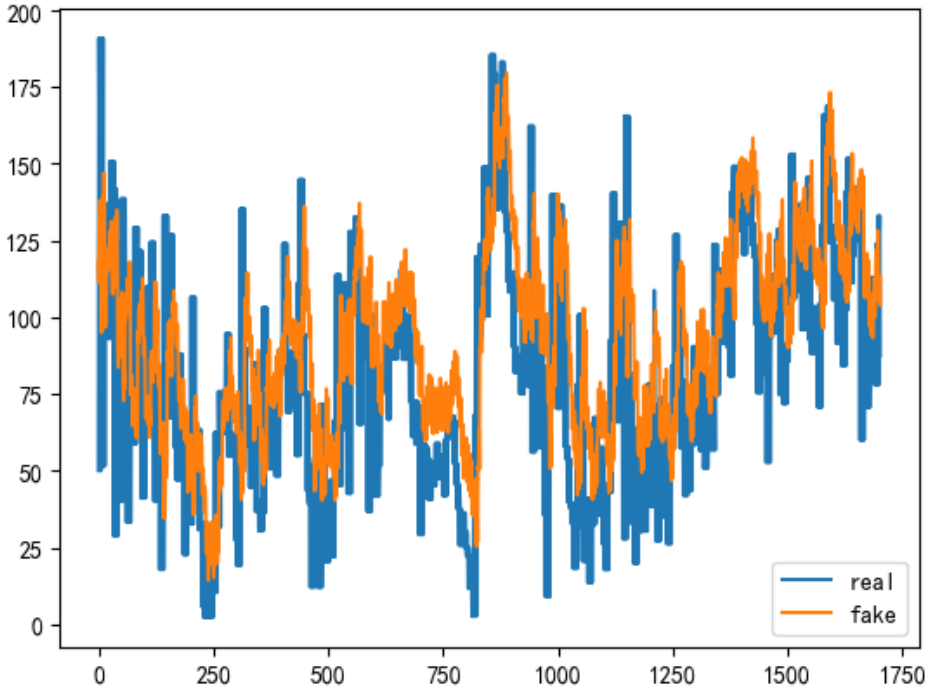

Fig.6 Comparison between real samples and synthetic samples

It can be seen from the figure that after 2000 rounds of training, the samples generated by the generator and the real samples have the same trend, but there are differences in amplitude, indicating that the samples generated by the improved GAN have diversity.

After the fault samples are generated by GAN, the synthesized samples are combined with the original training set as the enhancement training set, and the dimension reduction is carried out by SSAE. The dimension reduction visualization is shown in Fig. 7. The dimension reduced data is input into the Extremely Randomized Trees for fault diagnosis. The forest scale of the algorithm is 200, the depth of decision tree is 12, and the weight is 1, 3, 2, 5, 2.5 respectively. Table 3 and table 4 show the confusion matrix of test set classification before and after balance.

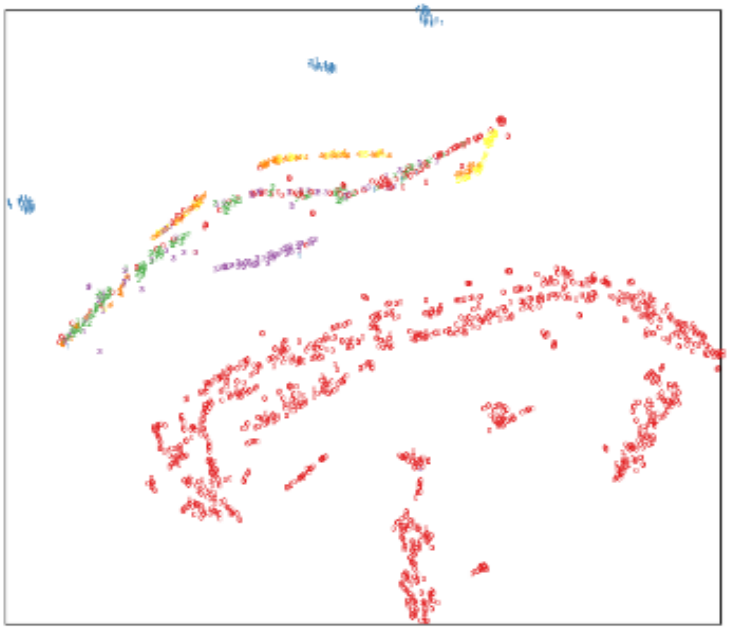

Fig.7 Data dimensionality reduction visualization

Table 3 Confusion matrix of experimental results(before)

Diagnosis / Real	f0	f1	f2	f3	f4	f5
f0	730	0	6	5	7	0
f1	1	54	0	0	0	0
f2	9	0	60	0	0	0
f3	5	0	3	56	1	0
f4	1	0	0	0	47	0
f5	9	0	0	0	30	0

Table 4 Confusion matrix of experimental results(after)

Diagnosis / Real	f0	f1	f2	f3	f4	f5
f0	751	0	0	2	2	0
f1	0	55	0	0	0	0
f2	0	0	69	0	0	0
f3	0	0	0	64	0	0
f4	0	0	0	0	48	0
f5	0	0	0	0	0	39

It can be seen from table 4 that the misdiagnosis rate before balance is relatively high, especially the state f5 is all misdiagnosed, and the classification effect has a large deviation; after the class balance by GAN, it can be found from table 5 that although there are still some misdiagnosis cases before the balance, the misdiagnosis situation has been greatly reduced. The specific performance is that the state f1 ~ f5 can be fully diagnosed, and the state f1 is also greatly improved. This shows that the proposed method has good performance in fault diagnosis of wind turbine yaw system.

In order to further illustrate the advantages of this method, three traditional oversampling methods, SVM, KNN and CNN, are selected for comparative experiments, and comprehensive evaluation is made by three evaluation indexes, namely Precision, F1-score and G-mean. The above methods are all carried out in the same experimental environment, and the average non-equilibrium rate of 14.30

is taken as the index of sample generation. Table 5 shows the comparison of fault diagnosis results under different methods.

Table 5 Comparison of fault diagnosis results under different over-sampling methods

Algorithm name	Precision	F1-score	G-mean
original	0.7179	0.7833	0.7430
GAN-SVM	0.7830	0.7722	0.7841
GAN-KNN	0.8857	0.8853	0.8866
GAN-CNN	0.9073	0.9288	0.9303
this paper	0.9884	0.9936	0.9937

As can be seen from table 5, compared with the original data set, the classifier performance after class sample balancing has been improved. Among them, the precision rate, F1 score and g-mean of CNN algorithm are improved by 26.39%, 18.58% and 25.22% respectively; KNN algorithm is improved by 23.38%, 13.02% and 19.34% respectively; SVM algorithm has the lowest improvement in this data set, and the three indicators are improved by 9.07%, - 1.42% and 5.54% respectively; and the Extremely Randomized Trees algorithm trained by SSAE algorithm has the most performance improvement, 67%, 26.85% and 33.75% respectively. Therefore, compared with the traditional machine learning method, the WGAN-SSAE-ET model proposed in this paper can better solve the problems of sample class imbalance and high latitude of SCADA data attributes in wind turbine fault diagnosis.

5. Conclusions

Aiming at the problem of unbalanced fault data and high data dimension of wind turbine, a new fault diagnosis model of wind turbine based on WGAN-SSAE-ET is proposed in this paper. The WGAN neural network is used to learn the real distribution of fault samples, generate composite samples that conform to the fault change rules, and enhance the original sample set. It not only retains the detection accuracy of most types of data samples, but also improves the diagnostic accuracy of a small number of samples In the Extremely Randomized Trees algorithm, each decision tree makes a vote to decide the abnormal category. The experimental results show that the Extremely Randomized Trees fault diagnosis model of WGAN-SSAE is better than SVM, KNN and CNN algorithm. The problem of class imbalance and high dimensional data in the process of wind turbine fault diagnosis is effectively solved.

References

[1] LI Hui, HU Yaogang, LI Yang, et al. Overview of condition monitoring and fault diagnosis for grid-connected high-power wind turbine unit[J]. Electric Power Automation Equipment, 2016, 36(1):6-16.

[2] ZHAO Hongshan, DONG Yeye, SONG Peng, DENG Chun. Fault detection method of wind turbine yaw system based on model[J]. Acta Energiae Solaris Sinica, 2020, 41(05): 142-149.

[3] LI Han, XIAO Deyun. Overview of data driven fault diagnosis methods [J]. Control and Decision, 2011, 26 (01): 1-9,16.

[4] ZHAO Hongshan, ZHANG Jianping, WANG Guilan, et al. State estimation based fault detection of hydraulic variablepitch system for wind turbines[J]. Automation of electric power systems, 2016, 40(22):101-104.

[5] HOU Zhongsheng, XU Jianxin. Review and Prospect of data driven control theory and method [J]. Acta Automatica Sinica, 2009,35 (06): 650-667.

[6] JIN Xiaohang, SUN Yi, SHAN Jihong, WU Genyong. Overview of wind turbine fault diagnosis and prediction technology [J]. Chinese Journal of Scientific Instrument, 2017,38 (05): 1041-1053.

[7] GAO Feng, DENG Xingxing, LIU Qiang, YANG Xiyun, WU Xiaojiang. Pitch angle fault diagnosis of large wind turbine electric pitch system [J]. Acta Energiae Solaris Sinica, 2020,41 (05): 98-106.

[8] BANGALORE P, TJERNBERG L B. An artificial neural network approach for early fault detection of gearbox bearings[J]. IEEE Transactions on Smart Grid, 2017, 6(2): 980-987.

[9] ZHANG C, HE Y, YUAN L, et al. Analog circuit incipient fault diagnosis method using DBN based features extraction[J]. IEEE Access, 2018:23053-23064.

[10] HUANG Nantian, YANG Xuehang, CAI Guowei, SONG Xing, CHEN Qingzhu, ZHAO Wenguang. Fault depth countermeasure diagnosis of fan main bearing using unbalanced small sample data [J]. Proceedings of the CSEE, 2020,40 (02): 563-574.

[11] LIN W C,TSAI C F,HU Y H,et al.Clustering-based under-sampling in class-imbalanced data[J]. Information Sciences, 2017, 409/410:17-26.

[12] HUANG Jianming, LI Xiaoming, QU Hezuo, ZHANG Lide. Transmission line fault identification method considering wavelet singular information and unbalanced data set [J]. Chinese Journal of electrical engineering, 2017,37 (11): 3099-3107,3365.

[13] HE H,GARCIA E A.Learning from imbalanced data[J].IEEE Transactions on Knowledge & Data Engineering, 2009,21(9):1263-1284.

[14] DI Rui, CHEN Zhengming, LV Jia. Application of xgboost algorithm based on smote in wind turbine blade icing prediction [J]. Information Technology, 2019,43 (12): 81-85,90.

[15] GOODFELLOW I J,POUGET-ABADIE J,MIRZA M,et al.Generative adversarial nets[C]// International Conference on Neural Information Processing Systems.Montreal:MIT Press, 2014: 2672-2680.

[16] Xu huoyao, Chen Lili. Rolling bearing fault diagnosis based on stack sparse self encoder [J]. Machine tool and hydraulic, 2020,48 (14): 190-194

[17] Xue Yan, Zhu Jing, Deng Aidong. Fault diagnosis method of rolling bearing based on entropy feature and stack sparse self encoder [J]. Industrial control computer, 2020,33 (10): 44-46

[18] Geurts P,Ernst D,Wehenkel L.Extremely randomized trees[J].Machine Learning,2006,63(1):3-42.

[19] ARJOVSKY M, BOTTOU, LÉON. Towards Principled Methods for Training Generative Adversarial Networks[J]. Stat, 2017, 1050.

[20] ARJOVSKY M,CHINTALA S,BOTTOU L.Wasserstein generative adversarial networks[C]// International Conference on Machine Learning, Sydney, Australia, 2017: 214-223.

[21] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved Training of Wasserstein GANs[J]. 2017. arXiv:1704.00028 [cs.LG].