DBScan and SVM for Fault Diagnosis of Wind Turbine Based on SCADA Data

Feifei Yin[1, a], Yu Gong[2, b]

[1]Network and information Office, North China electric power University, Baoding 071000, China;

[2]School of Computer, North China electric power University, Baoding 071000, China.

[a]yff1021@163.com, [b]gongyu0217@163.com

Abstract

As a kind of clean energy, wind turbine is widely used in our country. With the increasing intelligence of wind turbine, its components are more and more complex, and its maintenance cost is higher and higher. Therefore, we must find and diagnose the location and type of fault in time. A new fusion diagnosis method based on SCADA data is proposed in our paper, which uses a series of methods to clean data to ensure the reliability and validity of data, DBSCAN method is used to divide majority into subsets, and threshold value is added to control the number of classes to achieve the purpose of data balance. Finally, SVM is used to diagnose and identify fault types. After experimental verification, this method has a certain improvement in accuracy compared with the existing data-driven method.

Keywords

DBScan, Data balance, SVM, Fusion diagnosis method.

1. Introduction

China has made great achievements in the field of wind power generation. Compared with other energy renewal methods, wind power generation is better [1].Traditional thermal power generation has been gradually replaced by wind power generation due to its low efficiency and environmental pollution. Tidal power generation has recently developed, due to the high demands for natural environment and immature technology, so there is no breakthrough in the field of power generation. Based on the above advantages, wind power generation has become a mainstream way of power generation, but, meanwhile, it also has varity problems. For instance, a lot of the wind turbines are installed in the plain with high altitude and in the area with poor natural environment, so the installation cost and maintenance cost are very high, which makes us carry out the maintenance in time and accurately, otherwise once there is a fault, it will cause great loss[2].

Therefore, in the first place, we must find and diagnose faults to minimize the cost[3].At present, there are many fault diagnosis methods for wind turbine. CMS can diagnose faults further to determine specific fault types and help to reduce maintenance time and cost. CMS is an advanced method proposed at present.[4]. But CMS method has not been widely used because it needs to install a series of sensors, which will increase many additional maintenance costs and installation costs. Jing studied SVM and PCA for fault classification in 2015 [5].

Regretfully, more and more current fault diagnosis methods are came up and based on the idea that number of majority data and minority data is relatively balanced, with great errors, and the diagnosis results will be biased to most types of data. It shows good diagnostic performance when it is balanced, but under the condition of real data imbalance, the performance will become very poor. Cause a series of losses. In general, we can regard the normal data as the majority and the fault data as the minority. In a normal environment, the fault data is very few and far less than the number of normal data[6]. There is a huge data imbalance, which has a great impact on our training of classifiers.

Varity of methods recently have been put forward to deal with the problem of data imbalance[7].It is mainly divided into data-based and method-based methods. Based on the data, including over-sampling, under-sampling, over-sampling refers to increasing the number of samples in a few classes, and under-sampling refers to randomly reducing the number of samples which belongs to the majority.

For over-sampling, a SMOTE method is also proposed, which is an improvement of over- sampling method. Reduce the possibility of data over fitting by controlling the quality of new samples[8].Data imbalance methods based on algorithm level includes cost sensitive, ensemble learning and single classifier. Cost sensitivity means that each training sample can be weighted or sensitive factors can be introduced into the algorithm, that is, multiple classifiers can be used to get the results by voting or combining[9].

It can also be divided into homomorphic ensemble learning method (combination of the same classifier) and heteromorphic ensemble learning method (combination of multiple classifiers). Liu proposed an under-sampling-bagging algorithm in [5].And Over-sampling [10] is proposed by Xie. The first two methods are based on data.

Although there are many ways to deal with imbalance problem, most of them have their own shortcomings, such as the quality problem of the generated samples of smote, the generated samples may exist at the boundary of the classifier[8].So such samples will be more fuzzy boundary judgment, and will cause bad results in fault diagnosis.

For the under-sampling method, because it randomly select the part of data. The number of copies of data may lead to the occurrence of under fitting. In order to solve the above problems, in this paper, DBScan and SVM are proposed for fault diagnosis. By DBScan method, we first cluster most classes and divide them into several majority subsets according to the imbalance proportion.

When we generate most class subsets, we limit the number of samples generated, here we introduce a threshold T, which is selected according to the imbalance proportion of data. When we divide most class subsets into several majority class subsets, we compare them with a few classes to train classifier.

Finally, SVM classifiers map the results back to two or several categories that we have already divided in advance, and get the actual results. Compared with the current over-sampling and under-sampling which proposed method does not delete or add samples, so it avoids the problem of important information loss and over fitting. Greatly improved accuracy and effectiveness.

The method proposed in this paper has three main contributions: first, we compare many classification frameworks and choose an optimal classification framework based on experiments, and implement the framework .Second, Generally, SCADA data has a high dimension, which will cause large redundancy in the calculation, and there may be some errors in the original SCADA data, so we propose a series of data cleaning methods which can improve the performance of the experiment through have a pre-process step.

The rest of this article is presented as follows: In Section 2 not only analyzes the SCADA data set that we use in this experiment but also realize the data pre-processing process. In Section 3, A solution for integrating wind turbine diagnosis model is provided. In Section 4 our detailed experimental results are given. Finally, We summarize the work of this paper and look forward to the future work in Section 5.

2. Data preparation and Pre-processing

2.1 Data collection

2.1.1 Data Distribution

As shown in Table 1, the number of instances in the normal category is greatly different from that in the fault category. The purpose of our fault diagnosis system is to highlight these few kinds of data, and the hypothesis based on the balance of most and a few kinds of data leads to the imbalance of the data set, which makes the diagnosis system incline to the side of most kinds, resulting in the inaccuracy of the results.

Table 1 Dataset distribution

Wind Turbine	Majority instances	Minority instaces
2	62,456	56
10	65,789	13
23	59,213	54

We have two wind farms during 12 months SCADA data used. Specifically, the SCADA data used in this paper contains 29 parameters for each turbine, and the data collection time is one second apart. Table 2 shows the parameters of wind power plants. [4].

Table 2 Paraments for wind turbines in the wind power plants

Parameters	Description/Value
Rated power	15000kW
Cut_in speed	3m/s
Cut_out speed Blade’s rotation	20m/s 9.7 rpm-19 rpm

2.1.2 SCADA Data

The data we use is collected once a second. Each SCADA data contains 29 parameters, which are divided into four categories. These data can are grouped into some categories ,which has different value.

1. Some parameters determine the power output of the wind turbine.

2. Some parameters can help to analyze the health of wind turbines..

3. Some parameters are used to measure the operation performance of wind turbine.

2.1.3 Status Data

In order to understand the current operation status of wind turbine, the parameters we collect mainly include the following parameters: the wind turbine (WT) number, failure cause, failure maintenance activity, maintenance start date and maintenance end date. Table 3 shows two true examples when it has faults.

Table 3 Examples of status data collected by supervisory control

Wind Turbine	Fault type	Casuse	Maintenance
2	Blades	Blade 2 fault	Replacement
10	Generator	Generator fault	Replacement

2.2 Data Cleaning.

Generally, the SCADA data we collect contains a lot of errors. If you want to improve the stability and accuracy that we presented in this paper, we must clean the data in advance. The common errors in SCADA data include duplicate value, missing value, invalid value, etc. Some factors may cause model diagnosis errors, so the data must be cleaned before training the model.

2.3 Feature Selection.

Since we only need part of the features and can judge the fault types, First, we need to make sure that the information characteristics collected from the wind turbine[4].

2.4 Feature Reduction.

For SCADA data, we usually get high-dimensional data, but high-dimensional data is very inconvenient in calculation and feature extraction. If we do not reduce data, it will waste computing space and cause data redundancy. Our common data reduction methods include Fuzzy-rough instance selection (FRIS) method. FRIS was proposed by Jensen et al[11]. The data can be used later. Through experiments, we can see that the feature reduction by rough set fuzzy set theory can reach 97%. The operation efficiency is greatly improved.

2.5 Data Balance.

Generally speaking, there are two main ways to deal with unbalanced data. Under-sampling is to achieve data balance by randomly selecting a part of most classes. Over- sampling is to generate a few kinds of data according to certain rules. Although both methods can achieve the effect of data balance to a certain extent, the balanced sample points have more or less side effects on the production of the model, sometimes it will greatly affect the performance of the model[11].

Based on the above proposed problems, this paper proposes a clustering method, which generates several subsets of most classes according to the unbalanced proportion, combines the subsets of most classes with the subsets of few classes, and then trains a Bayesian classifier.

Given a training set W, it is assumed that the normal class X_1 is the majority class and the fault class is the minority class. It can be seen that there is little difference between the minority classes. Therefore, we can define the ratio of the minority classes of the majority classes as N, which is expressed as follows:

$$n = \frac{n_1}{n_2} \approx \frac{n_1}{n_3} \approx \frac{n_1}{n_{c+1}} \tag{1}$$

First, we divide majority X_1 into k sub-classes by DBScan as X_1=[A_1;A_2···A_k]. But because there are so many kinds, But most of them will form new imbalances after this kind of balance. Therefore, we must introduce a threshold T to control the number of majority samples of each new generated subset. We illustrate this conclusion with a real case, as demonstrated in Fig. 1. As shown in Fig. 1(a) The result of case we can see in Fig. 1(b), The meanings of the symbols in the figure have been given.

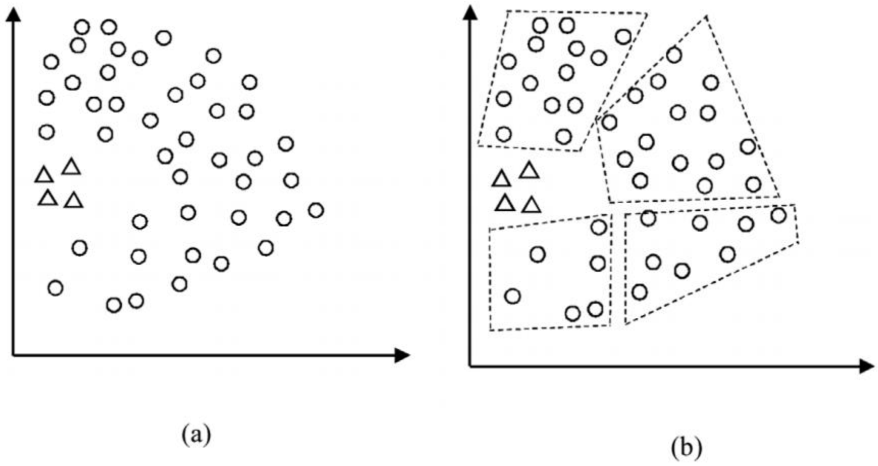

Fig. 1 DBScan(a) a true problem; (b) a binary classification

After that, we construct a classifier. The selection principle of the classifier will be introduced later. The corresponding labels and data set labels are given in the formula. As a result, the labels of the minority classes are updated to [k + 1, k + 2, · · ·, C + k].

3. Diagnosis Solution

Our purpose is to determine the state of the generator by judging the category of the data through the SCADA information. The general process is to process the text information in the form of data and inject it into the classifier, which classifies it and returns the state information of the user.

3.1 Fault Classsfication.

We have a training data set,$W_1 = [X_1; X_2; \cdots X_{t+1}]$,it has t kinds of minority classes and one majority class.X_i=[X_1;X_2;X_3⋯X_{t+1}], i = 1, 2,..., t + 1 which stands for the data of the i th class. We use these symbols to refer to various types. One is a majority class and the other is a minority class. We can use the following formula to calculate the posterior probability.

$$p(c|x) = \frac{p(c)p(x|c)}{p(x)} \tag{2}$$

We can know P (c) is prior probability of class, the conditional probability of sample x belonging to a certain majority or minority class is p(c│x), P (x) is presented as normalization. In order for this formula to work, we need to assume that each variable is independent and that formula (2) can be calculated as, Eq. (2) can be computed as:

$$p(c|x) = \frac{p(c)p(x|c)}{p(x)} = \frac{p(c)}{p(x)} \prod_{i=1}^{m} p(a_i|c) \tag{3}$$

where a_i is the value of a variable in the training set, we can know P (a_i|c) is a conditional probability in the training set.

Variables a_i are all Gaussian distribution P (a_i|c)~$N(\mu_{c,i}, \sigma_{c,i}^2)$, the mean and variance of the i th variable is presented with $\mu_{c,i}$ and $\sigma_{c,i}^2$,therefore we can know how to calculate the P (a_i|c),as follows:

$$P(a_i|c) = \frac{1}{\sqrt{2\pi}\sigma^{c,i}} exo(-\frac{(a_i-\mu_{c,i})^2}{2\sigma_{c,i}^2}) \tag{4}$$

Finally, we can use the follow equation to compute the diagnosis classification.

$$h(x) = arg \max_{c \in Y} p(c) \prod_{i=1}^{m} P(a_i|c) \tag{5}$$

The different variable of each class are calculated as follows acc-ording to the following equation.

$$Mean_{pq} = \frac{1}{n_p} \sum_{t_p=1}^{n_p} a_{tq} \tag{6}$$

$$Var_{pq} = \frac{1}{n_p} \sqrt{\sum_{t=1}^{n_p} \left(a_{t_{pq}} - Mean_{pq}\right)^2} \tag{7}$$

$$P_p = \frac{n_p}{\sum_{t=1}^{C+N} n_{tp}} \tag{8}$$

We can classify the test sample X by using Eq. (5), and we also can have the real label of the test sample y_r mapping to sample X:

$$y_r = \begin{cases} 1, y_0 \in [1, k] \\ y_0 - k + 1, y_0 \in [k+1, C+k] \end{cases} \tag{9}$$

4. Results of Experiments

In this Chapter, our proposed method is validated through varity of experiments. For the classification problem, we usually use ACC, ROC curve to fit the performance of the experiment. In this paper, TP, TN, FP, FN and other relevant parameters are used to obtain the value of ACC according to the relevant rules. The Bayes classifier proposed in this paper is compared with other common classifiers, and then through a series of comparative experiments, the experimental results show the effectiveness of the proposed method.

4.1 Diagnosis Performence Metrics.

TP, TN, FP, FN are proposed and quoted in this paper. These parameters can be used to calculate a series of parameters about classification accuracy, such as accuracy and precision.[12]. Through these parameters we can verify the validity of the diagnosis model, we usually propose three indicators, accuracy, recall and precision.Equation(10),Equation(11), Equation (12) defines the classification accuracy, recall ,precision respectively.Proportion of correct data (TP + TN) determined by the model to the total data.

$$ACC = \frac{TP+TN}{TP+TN+FN+FP} \tag{10}$$

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

For all positive cases (TP + FN) in the data set, the proportion of positive cases (TP) correctly judged by the model to all positive cases in the data set[13].

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

Through a series of data, we can know that our ideal model is to classify all the relevant parameters such as accuracy, precision, precision, AOC curve and so on. And the higher the above value is, the better.

4.2 Over Performance.

This part includes some different experiments. In the first experiment, we do not process the data, but directly inject four classifications to observe the performance index. Then a series of experiments and data are used to discuss the performance of the proposed centralized classifier to determine whether they are suitable for the situation in this paper. In the next experiment, after data preprocessing operations proposed. We put the data into different classifiers, we process the data in combination with the methods of processing imbalance and data preprocessing proposed in this paper, and compare the results of three times.

In order to verify whether the model is over fitted, we divide the data into test set and training set[4], and randomly select 10 wind turbine ground data as test set and training set[7].

Table 6 shows the accuracy of various classifiers proposed in this paper. In Table 6, the top classifier for each metric is marked . As shown in table 6, in general, there is no optimal classifier that can achieve the best of all indicators. This is mainly because the classifiers used in different environments are different.

After using the threshold method, the results are shown in Table 7. We can see the performance of different classifiers. In Table 7, We can see that both Bayes and KNN have maximum values on different indexes. In summary, we are still unable to determine the best classifier for this paper.

For a class imbalance problem, from the results of above two experiments, we can find that after the data balance is injected into the classifier, different classifiers have a certain amount of improvement, among which Bayes method is the fastest.

Combining all the results, SVM classifier can get good experimental results in each experiment. The best classifier of this diagnosis system is selected by observing AUC curve. By observing AUC curve, we can see that in AUC, most indexes of SVM are better than all other classifiers. Therefore, in the future similar experiments, SVM classifier can be our first choice.

In the last experiment, we use the DBSCAN Bayes method to process the training data, Figure 8 shows the corresponding ROC curve.

Experimental results verify the effectiveness of the method proposed in this paper. The performance of the model is improved by applying the method presented in this paper. The validity and correctness of the model are also improved.

5. Conclusions

This paper presents a new wind power fault diagnosis system based on multi-method fusion of SCADA data. In the data preprocessing part of the method, many methods are added to improve the data quality effectively and gives a lot of convenience to the later construction models. By using T-threshold DBScan algorithm handles data imbalance, uses Bayes method for fault diagnosis, calculates the accuracy rate and other related information by using ROC curve and ACC criteria. The experimental results show that, while proving the effectiveness of this method, it also improves the performance of the model and the accuracy of diagnosis to some extent. After a series of transfer learning, this method is also applicable to other fields, such as medical diagnosis and have a far-reaching impact on other areas of research. However, further efforts are needed to extract the features of fault types. This paper does not consider the impact of the real environment on the generator, including the impact of some unexpected events on the generator. The fault tolerance rate is low. In the future work, we will explore more fault features and take them into full account in our model by combining with the actual situation in the field.

Acknowledgements

The authors acknowledge the Fundamental Research Funds for the Central Universities（Grant:2018 MS078）.

References

[1] A. Chen, H. Zhou, J. Jiao, and T. Gao, “SVD and statistic theory based modified TPLS,” 7th International Conference on Intelligent Control and Information Processing, ICICIP 2016 - Proceedings, pp. 49–54, 2017, doi: 10.1109/ICICIP.2016.7885914.

[2] P. Borkar and L. G. Malik, “Review on Vehicular Speed, Density Estimation and Classification Using Acoustic Signal,” International Journal for Traffic and Transport Engineering, vol. 3, no. 3, pp. 331–343, 2013, doi: 10.7708/ijtte.2013.3(3).08.

[3] G. P. Goodwin and P. N. Johnson-Laird, “Diagnosis of Ambiguous Faults in Simple Networks,” Proceedings of the twenty-Seventh Annual Conference of Cognitive Science Society, pp. 791–796, 2005, [Online]. Available:

[4] Y. Zhao, D. Li, A. Dong, D. Kang, Q. Lv, and L. Shang, “Fault prediction and diagnosis of wind turbine generators using SCADA data,” Energies, vol. 10, no. 8, pp. 1–17, 2017, doi: 10.3390/en10081210.

[5] G. Chen, Y. Liu, and Z. Ge, “K-means Bayes algorithm for imbalanced fault classification and big data application,” Journal of Process Control, vol. 81, pp. 54–64, 2019, doi: 10.1016/j.jprocont.2019.06.011.

[6] L. Duan, M. Xie, T. Bai, and J. Wang, “A new support vector data description method for machinery fault diagnosis with unbalanced datasets,” Expert Systems with Applications, vol. 64, pp. 239–246, 2016, doi: 10.1016/j.eswa.2016.07.039.

[7] T. Jin, X. Qiu, D. Hu, and C. Ding, “Channel Error Estimation Methods Comparison under Different Conditions for Multichannel HRWS SAR Systems,” Journal of Computer and Communications, vol. 04, no. 03, pp. 88–94, 2016, doi: 10.4236/jcc.2016.43014.

[8] J. M. Martínez-García, C. P. Suárez-Araujo, and P. G. Báez, “SNEOM: A sanger network based extended over-sampling method. Application to imbalanced biomedical datasets,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7666 LNCS, no. PART 4, pp. 584–592, 2012, doi: 10.1007/978-3-642-34478-7_71.

[9] H. Y. Lo, J. C. Wang, H. M. Wang, and S. De Lin, “Cost-sensitive stacking for audio tag annotation and retrieval,” ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, no. May 2011, pp. 2308–2311, 2011, doi: 10.1109/ICASSP.2011.5946944.

[10] J. Xie and Z. Qiu, “The effect of imbalanced data sets on LDA: A theoretical and empirical analysis,” Pattern Recognition, vol. 40, no. 2, pp. 557–562, 2007, doi: 10.1016/j.patcog.2006.01.009.

[11] Y. Qian, Q. Wang, H. Cheng, J. Liang, and C. Dang, “Fuzzy-rough feature selection accelerator,” Fuzzy Sets and Systems, vol. 258, pp. 61–78, Jan. 2015, doi: 10.1016/j.fss.2014.04.029.

[12] W. Wang, M. Zhang, D. Wang, and Y. Jiang, “Kernel PCA feature extraction and the SVM classification algorithm for multiple-status, through-wall, human being detection,” Eurasip Journal on Wireless Communications and Networking, vol. 2017, no. 1, 2017, doi: 10.1186/s13638-017-0931-2.