ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

Condition Monitoring of Electrical Rotating Parts Based on Mutual Information Feature Selection and Improved KNN Regression Algorithm

Zhiguo Dong[1, a], Jian Feng[2], Wencheng Zhao[3, b], Yu Wang[4], Haiming Niu[5]

[1]Qinghai Electric Power Company, CHN Energy Group, Xining, China

[2]Guoneng Zhishen Control Technology Co., Ltd., Beijing, China

[3]Qinghai Electric Power Company, CHN Energy Group, Xining, China

[4]Qinghai Huanghe Madang Hydropower Development Co., Ltd., CHN Energy Group, Xining, China

[5]Guoneng Zhishen Control Technology Co., Ltd., Beijing, China

[a]12067999@ceic.com, [b]20082511@ceic.com

Abstract

It is necessary to monitor the on - line status of the rotating parts which can cause the shutdown of the electrical equipment. In order to improve the accuracy of the prediction model, the mutual information is used to select the features of the original data. Aiming at the condition monitoring, the distance measurement formula of KNN regression algorithm is improved, and the prediction accuracy is improved by about 80%. Taking the bearing of a wind turbine as an example, the improved KNN regression algorithm can be used for fault warning three weeks in advance.

Keywords

Rotating parts, Mutual information, KNN regression, Condition monitoring.

1. INTRODUCTION

With the improvement of production automation, rotating machinery is more and more widely used in the process of power generation. For example, bearing, gear and other rotating parts are the most common parts in large mechanical equipment. Due to the relatively high speed and poor working conditions of most rotating parts, faults are easy to occur, and will be accompanied by downtime, maintenance difficulties and other problems[1].

Supervisory Control and Data Acquisition (SCADA) system has the advantages of long sampling period, large amount of Data can be recorded, and no additional sensor is required, which is widely used in condition monitoring of wind turbines [2]. At present, a large number of wind turbine condition monitoring methods based on SCADA data have been studied, such as trending analysis, clustering, and normal behavior modeling(NBM) [3].

The NBM method based on SCADA data has attracted the attention of many scholars. Wang used SCADA data to build an integrated NSET model and combined with fuzzy soft clustering to realize on-line monitoring of wind turbine gearbox bearing[4]. After using LARS algorithm to select features of SCADA data, Sun used hidden Markov method to establish the NBM of wind turbine. The experiment proved that the method proposed in the paper had good fault identification effect[5]. Liu used SCADA data to design a fault warning method for wind turbines based on abnormal data reconstruction. The experimental results showed that the bearing

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

faults of wind turbines could be identified at least 3 weeks in advance[6]. There are many algorithms to build NBM. The K Nearest-neighbor (KNN) regression algorithm is a non-parametric modeling method[7]. Due to its advantages such as simple principle and no need to train the model in advance, it is used in this paper to establish the NBM of rotating parts such as gearbox bearing of wind turbine.

SCADA system can record the value of multiple variables of wind turbine. However, if all characteristic variables are used to build a model at the same time, it will lead to "dimensional disaster". Moreover, features unrelated to predictive variables will affect the accuracy of the model. According to whether the data has classification labels or not, feature selection methods can be divided into supervised and unsupervised methods [8]. When the data category is unknown, unsupervised feature selection methods, such as mutual information [9] and spectral analysis [10], should be adopted. Gu proposed a feature selection algorithm based on redundancy analysis and interaction weight[11]. Compared with several feature selection algorithms, the proposed algorithm can obtain better feature selection performance. Mahendra proposed a clustering method based on unsupervised feature selection and cluster center initialization for intrusion detection[12]. Experimental results confirm that the proposed method is suitable for LAN and mobile ad-hoc network, varying data density, and large datasets.

Based on MI feature selection and improved KNN regression algorithm, a condition monitoring method for wind turbine rotating parts is proposed in this paper. Firstly, the MI is used to select the features of the original SCADA data to achieve dimension reduction. Secondly, the NBM of wind turbine rotating parts is established after improving the distance measurement formula of KNN regression algorithm. Finally, the statistical process control (SPC) method is used to set the alarm threshold to realize the condition monitoring of wind turbine rotating parts.

2. FEATURE SELCTION BASED ON MI ALGORITHM

There are multiple features in the data collected by SCADA system. The more the number of features, the more comprehensive and specific the performance of wind turbine operation will be reflected. Therefore, in general, the classification performance of high-dimensional data is better. However, with the increase of feature dimension, there are often some repetitive features, namely "redundant features", or features that have little or no correlation with the target features. The large number of these two features will not only increase the operation cost, but also reduce the accuracy of subsequent models. Feature selection is an effective data processing method. By screening existing features, a feature subset is selected from the original feature set to achieve the best prediction effect.

Generally, feature selection algorithms can be divided into three categories: filter, wrapper and hybrid methods[]. Compared with the others, filter methods have simple procedures and relatively high computational efficiency, which are suitable for large-scale data. The informatin theory-based algorithm is one of the most widely used filter methods. Next, the concept of mutual information (MI) of information theory and the feature selection algorithm based on MI are introduced.

The MI $I(X;Y)$ is a measure of the independence between two random variables X and Y. It can also be regarded as a measure of the information contained in one random variable about another random variable.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

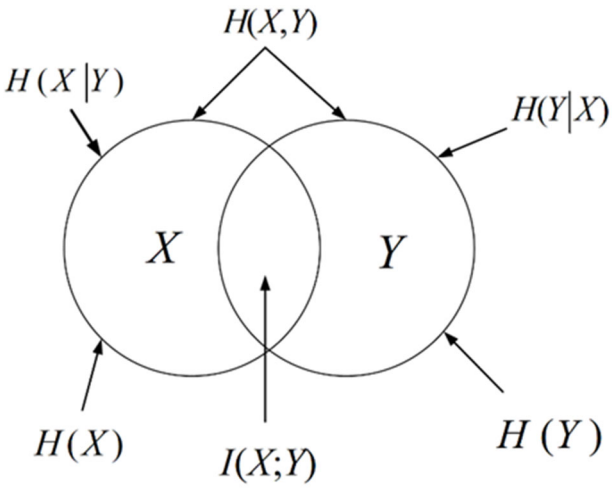

Figure 1. Concept diagram of mutual information correlation

The formula of $I(X;Y)$ is as follows:

$$\begin{aligned} I(X;Y) &= \sum_{(x,y)\in(X,Y)} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= H(X)-H(X|Y)=H(X)+H(Y)-H(X,Y) \end{aligned} \tag{1}$$

where $H(X)$ is information entropy, $H(X|Y)$ is conditional entropy, $H(X,Y)$ is joint entropy. The formular of $H(X)$ and $H(X,Y)$ is as follows:

$$H(X)=-\sum_{x\in X} p(x)\log p(x) \tag{2}$$

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \tag{3}$$

where $p(x)$ is the probability distribution, $p(x,y)$ is the joint probability distribution function.

For feature selection, MI can be regarded as a materiality criterion of the correlation between the two features. The MI feature selection algorithm is introduced below:

S1: Determine the target feature Y and the raw feature set $X=\{X_1,X_2,...,X_m\}$.

S2: The information entropy $H(Y)$, $H(X_i)$ and the joint entropy $H(X_i,Y)$ are calculated.

S3: Based on Eq.(1), the MI sequence $[I(X_1;Y),I(X_2;Y),...,I(X_m;Y)]^T$ is calculated. The top n features which have the maximun MI are selected as the feature subset $X_s=\{X_{s1},X_{s2},...,X_{sn}\}$.

3. IMPROVED KNN REGRESSION ALGORITHM

K Nearest-neighbor (KNN) regression algorithm is a kind of learning method based on the instance, the basic idea is through some distance measure, found the training set and test point closest to the K neighbors, using the K nearest points' information forecast test points, that is, with the output of this K neighbor points as the prediction results, the average of its steps are as follows:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

S1: Calculate the distance d between a point $X_t = (x_1^t, x_2^t, ..., x_n^t, y_t)$ in the training set and the test point $X = (x_1, x_2, ..., x_n, y)$. The formula of distance calculating is as follows:

$$d = \|X - X_t\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - x_i^t)^2} \tag{4}$$

S2: Based on Eq(4), the distance sequence $[d(1), d(2), ..., d(m)]^T$ between all points in the training set and the test points is calculated.

S3: The top K points in the training set closest to the test point are selected as the subset $X_t = \{X_{t1}, X_{t2}, ..., X_{tk}\}$.

S4: The formula for calculating the predicted output of test point $X = (x_1, x_2, ..., x_n, y)$ is as follows:

$$\hat{y} = \sum_{j=1}^{k} y_{tj}^{'} / k \tag{5}$$

The purpose of KNN regression algorithm is to obtain the predicted value, that is, the output of the test point is unknown, so the output value does not participate in the distance calculation. However, in the problem of wind turbine component state monitoring, the actual output of the test point can be measured by the SCADA system. The core idea of the normal behavior modeling method is to use the historical normal data to learn the model of normal operation state, and then use the real-time data and the residual of model output to judge whether the research object deviates from the normal state at this time.

According to the characteristics of condition monitoring, the improvement of KNN regression algorithm is as follows:

Calculate the distance d between a point $X_t = (x_1^t, x_2^t, ..., x_n^t, y_t)$ in the training set and the test point

$X = (x_1, x_2, ..., x_n, y)$. The improved distance measurement formula is as follows:

$$d = \|X - X_t\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - x_i^t)^2 + (y - y_t)^2} \tag{6}$$

4. CONDITION MONITORING OF ELECTRICAL ROTATING PARTS BASED ON NBM

Rotating mechanical parts are common and important parts in electrical equipment. Due to its long-term high speed rotation state and harsh working environment, once the fault occurs, it will not only affect the production process, but also have high maintenance difficulty and high maintenance cost. Therefore, the condition monitoring method based on normal behavior modeling is of great significance for improving the reliability of rotating mechanical parts.

The basic idea of applying the normal behavior modeling method based on SCADA data to the condition monitoring of rotating mechanical parts is using the historical data under normal state to establish the model related to predictive characteristics, and get the output value of the model. The residual difference between the output value of the model and the actual output value is calculated to determine whether the rotating part deviates from the normal state. The

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

main steps of rotating part state monitoring based on mutual information feature selection and improved KNN regression algorithm are as follows:

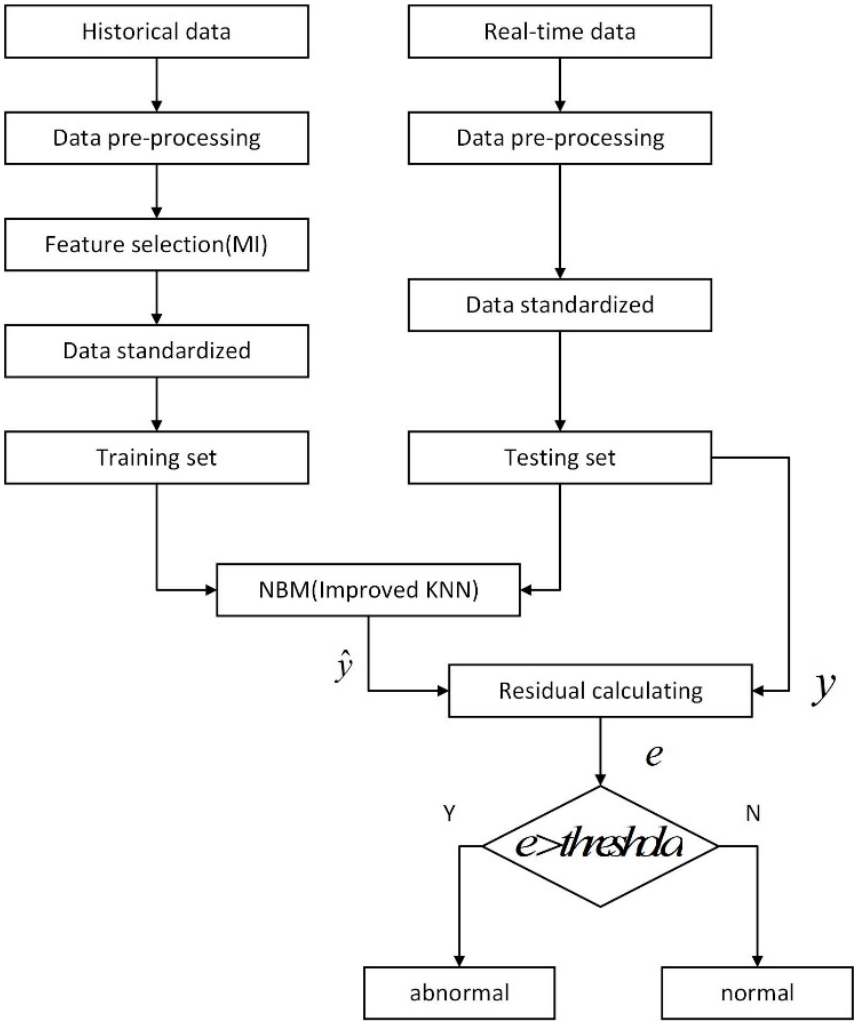

Figure 2. Schematic diagram of water level control

(1) Offline Processing: The main content of this stage is to pre-processing the historical data and feature selection is used to obtain the training set.

Derive some historical data from the SCADA system. Select the normal data from the historical data based on the pre-processing rules and then standardize the normal data.

S1: Delete the samples with missing values or incomplete records and the samples with the active power less than or equal to zero.

S2: Based on the status code and related technical documents, delete the sample with the status code recorded as not normal, such as shutdown, power limitation, fault, and maintenance.

S3: The remaining samples are standardized. The formula of the Z-score standardized is as follows:

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{7}$$

where x is the raw data, $\bar{x}$ is the mean of the corresponding parameter, σ is the standard deviation, x^* is the standardized data.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

Based on the MI algorithm feature selection is used to the normal data and the training set is formed.

(2) Online Estimation: The main content of this stage is to online calculate the output value based on improved KNN algorithm.

Collect the real-time data and determine whether it is the normal data based on the pre-processing rules. If the real-time data is not the normal data, it means that the current condition of research object is obviously not normal. If it is the normal data, the normal data is retained as the input of KNN.

(3) Condition Monitoring: The main content of this stage is using SPC technology to set alarm threshold to infer the current operational condition of research object based on the real-time residuals. If the residuals exceeds the limit for a long time, then the research object is considered to have a significant failure at this time.

5. EXPERIMENT

5.1. Research Object and Data Pre-processing

The operational data used in the following case are from an onshore WTs at a wind farm in Hebei province, China. The number of the wind turbine is D31. The main characteristic parameters are as follows: rated power is 1.5 MW, rated wind speed is 12 m/s, cut-in wind speed is 3 m/s, cut-off wind speed is 25 m/s, data sampling interval of the SCADA system is 5 min. A gearbox overheating fault occurred in D31 from 2017/11/17 8:30 to 2017/11/18 14:31. There are about 70000 samples with the time span of 2017/4/1 to 2017/12/31 were derived from the SCADA system.

There are about 60 operational parameters available in the SCADA system. Considering that the case studies the gearbox fault, 13 related operational parameters are initially selected as Tab. I:

TABLE I. Related Optional Parameters

Gearbox-related parameters	Other important parameters
Gearbox oil inlet pressure	Active power
Gearbox oil filter-front pressure	Nacelle temp
Gearbox no-drive bearing temp	Ambient temp
Gearbox drive bearing temp	Wind direction
Gearbox oil temp	Wind speed
	Main shaft speed
	Wind turbine status code
	Main bearing temp

For the raw samples, the data pre-processing was performed first to select the normal data accordind to the pre-processing rule in Part IV. After the pre-processing, the number of samples is 51,000. The first 29,000 samples with time span of 2017/4/1 to 2017/9/11 were z-score standardized as the control group. The remaining 22,000 samples served as experimental group and were also standardized based on Eq(7).

The step after data pre-processing is feature selection. The target feature in this case is the gearbox oil temp and the raw feature set contains all other parameters shown in Tab.1. The top 5 paramters are gearbox oil temp, active power, wind speed, nacelle temp and ambient temp.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

TABLE II. Selected Parameters

Parameters	MI
Gearbox oil temp	1
Active power	0.7665
Wind speed	0.7634
Nacelle temp	0.6883
Ambient temp	0.6542

The control group was divided into two parts, with the first 18,000 samples as the training set and the second 11,000 samples as the test set. The experimental group was divided into three parts, with samples 1-11,000 as the test set, samples 1-500 as the reference set for subsequent threshold setting, and samples 11001-22,000 as the training set.

5.2. NBM

First, experiments were carried out on the control group to verify the effect of the improved KNN regression algorithm on NBM. The prediction accuracy of the unimproved and improved algorithms was compared. Root mean square error (RMSE) is used as indicator to evaluate the performance of them. The formulas of RMSE is as follows:

$$\sigma_{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where n is the number of samples, y_i is the real-time output, $\hat{y}_i$ is the output of the model.

The performance of the two algorithm are shown in TABLE II.

TABLE III. RMSE

	Unimproved	Improved
RMSE	0.7174	0.1447
Computation time/s	3.0886	3.1118

According to Tab.III, compared with the unimproved KNN regression algorithm, the improved KNN regression algorithm has no significant improvement in computing time, but in terms of prediction accuracy, the improved KNN regression algorithm is about 80% higher than the previous one, which indicates that the improved KNN regression algorithm has a better fitting effect on the gearbox normal state model.

5.3. Condition Monitoring

In the actual production, the staff judge the running state of the electrical equipment by whether the monitored characteristic quantity exceeds the limit. In this paper, the statistical process control (SPC) technology[13] is used to set the alarm threshold to realize the state monitoring and fault warning of wind turbine gearbox bearing.

SPC technology is mainly used for real-time monitoring and early warning of the production process. Since the failure of the gearbox bearing is manifested as an increase in the temperature of the oil pool, the upper alarm threshold is set here. The formulas are as follows:

(1) According to the normal distribution correlation theory, the probability of the random variable $X \sim N(\mu,\sigma^2)$ falling in the interval $(-\infty, \mu + 2.326\sigma]$ is:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

$$P(-\infty < X \le \mu + 2.326\sigma) \approx 0.99$$

If the value of X exceeds the above formula for a long time, it can be considered that the production process is affected by abnormal factors, resulting in significant changes in the probability distribution of the random variable. In practical application, the sample mean $\bar{X}$ and sample standard deviation S are used to replace the mean and variance of normal distribution to set the threshold. The formula is as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{8}$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{X})^2} \tag{9}$$

where e_i is the residuals, n is the number of samples.

The formula of alarm threshold T is as follows:

$$T = \bar{X} + 2.326S \tag{10}$$

Experiment was carried out on the experimental group, and its residual error was shown in the Fig 3. According to Eq(10), the alarm threshold calculated by reference set is 0.4675.

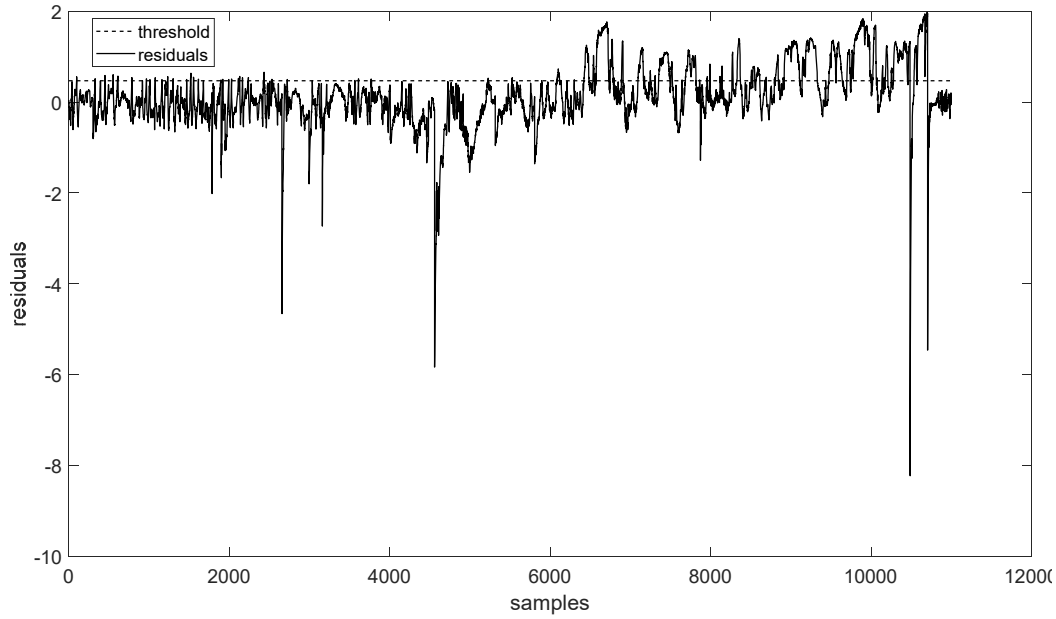

Figure 3. The residuals of NBM

It can be seen from the Fig. 3 that there is a short-time over-limit residual error at samples No. 1-3000.

At the samples No. 3000-6000, the fluctuation range of residual error gradually increases, and it can be considered that the gearbox bearing is in the early failure state at this time.

At samples No. 6000-10500, the residual is shown to be significantly exceeding the threshold for a long period of time. Then, it is considered that the gear box has obvious failure and needs to be stopped for maintenance.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

At data No. 10500-11000, the residual error is far below the threshold and there is no obvious fluctuation, indicating that the gearbox is in the normal working condition after maintenance.

In the actual production, if the residual error is directly taken as the alarm basis, it is easy to cause false alarm or frequent alarm. The system alarm rule is set here: if the system detects that the residual error exceeds the threshold for 30 consecutive minutes, the gearbox bearing will be considered to have an obvious fault and trigger the alarm. The system alarm situation is shown in the Fig. 4:

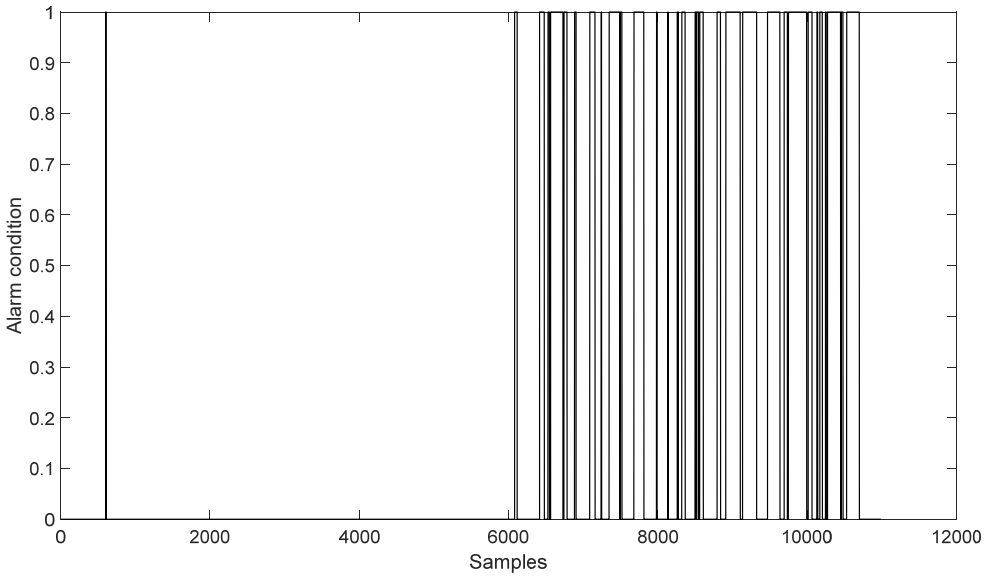

Figure 4. System alarm diagram

As can be seen from the above figure, the system triggered an alarm in the early failure. After sample point No. 6000, the system was triggered by an alarm several times at this time of 2017/10/26, which can achieve an early warning three weeks ahead of the SCADA system.

6. CONCLUSION

In this paper, the mutual information algorithm is used to select the features of the original data collected, the improved KNN regression algorithm is used to model the normal behavior of the electrical rotating parts, and the alarm threshold is set in combination with the SPC technology to realize the condition monitoring.

(1) The KNN regression algorithm with improved distance measurement formula has better modeling effect than that without improvement;

(2) The mutual information algorithm can effectively reduce the dimension of data;

(3) On-line condition monitoring of rotating parts of electrical equipment can realize early warning of faults.

REFERENCES

[1] WANG Rui, CHEN Zhigang.Syudy on centrifugal gas compressor bearing fault character obtaining based on data mining[J]. Oil Field Equipment,2006(05):23-26.

[2] XIONG Zhongjie,QU Yingning,FENG Yanhui,et al. Fault analysis of wind turbine pitch system based on machine learning[J]. Acta Energiae Solaris Sinica,2020,41(05):85-90.

[3] Tautz-Weinert J , Watson S J . Using SCADA data for wind turbine condition monitoring – a review[J]. Iet Renewable Power Generation, 2017, 11(4):382-394.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202412_10(12).0006

[4] WANG Ziqi, LIU Changliang, Liu Shuai.Condition monitoring of wind turbine gearbox based on ensemble nonlinear state estimation technique and soft fuzzy clustering [J]. Chinese Journal of Scientific Instrument,2019,40(07):138-146.

[5] SUN Qunli, ZHOU Ying, LIU Changliang.Research on fault diagnosis of wind turbine based on LARS feature selection[J]. Renewable Energy Resources,2020,38(10):1349-1354.

[6] LIU Shuai, LIU Changliang, ZHEN Chenggang.Fault warning method for wind turbine based on classified data reconstuction [J]. Chinese Journal of Scientific Instrument,2019,40(08):1-11.

[7] LI Jianguo, HE Yunpeng, LI Bowen. Short-term traffic forecast of stereo garage based on improved KNN algorithm[J]. Measurement & Control Technology, 2020,39(06):115-120.

[8] XU Junling, ZHOU Yuming, CHEN Lin,et al.An unsupervised feature selection approach based on mutual information [J]. Journal of Computer Research and Development,2012(02):158-168.

[9] Peng H , Long F , Ding C . Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(8):1226-1238.

[10] PAN Feng, WANG Jiandong, NIU Ben.Unsupervised feature selection approach based on spectral analysis [J]. Journal of Computer Applications, 2011,31(08):2108-2110+2114.

[11] Xiangyuan Gu, Jichang Guo, Chongyi Li, et al. A feature selection algorithm based on redundancy analysis and interaction weight. 2020, :1-15.

[12] A M P , A S T , B K D . Unsupervised feature selection and cluster center initialization based arbitrary shaped clusters for intrusion detection[J]. Computers & Security, 2020, 99.

[13] SHEN Haoxin, LV Chengde, ZHANG Rui. Application research on SPC in quality stability control of safety command receiver-decoder [J]. Quality and Reliability, 2020(02):51-55.