J. Mod. Transport. (2019) 27(3):211–221
https://doi.org/10.1007/s40534-019-0191-4

# A fault prediction method for catenary of high-speed rails based on meteorological conditions

Sheng Lin[1] · Qinyang Yu[1] · Zhen Wang[2] · Ding Feng[1] · Shibin Gao[1]

Received: 9 January 2019 / Revised: 6 June 2019 / Accepted: 12 June 2019 / Published online: 29 June 2019

**Abstract** Fault frequency of catenary is related to meteorological conditions. In this work, based on the historical data, catenary fault frequency and weather-related fault rate are introduced to analyse the correlation between catenary faults and meteorological conditions, and further the effect of meteorological conditions on catenary operation. Moreover, machine learning is used for catenary fault prediction. As with the single decision tree, only a small number of training samples can be classified correctly by each weak classifier, the AdaBoost algorithm is adopted to adjust the weights of misclassified samples and weak classifiers, and train multiple weak classifiers. Finally, the weak classifiers are combined to construct a strong classifier, with which the final prediction result is obtained. In order to validate the prediction method, an example is provided based on the historical data from a railway bureau of China. The result shows that the mapping relation between meteorological conditions and catenary faults can be established accurately by AdaBoost algorithm. The AdaBoost algorithm can accurately predict a catenary fault if the meteorological conditions are provided.

**Keywords** High-speed rail · Catenary · Trip · Fault prediction · Data processing · Meteorological conditions

## 1 Introduction

In recent years, the high-speed rails (HSRs) of China have developed rapidly, which means that both scale of operation and catenary expand greatly. The traction power supply system (TPSS) of HSRs requires a very high reliability [1–3]. Catenary is a key component of the TPSS, but there is no standby catenary in TPSS. Meanwhile, the stability and reliability of the catenary system are directly related to the operation state of HSRs. Therefore, an accurate fault prediction of the catenary system and timely warning is crucial to improving the reliability of the entire HSR system.

Zhao et al. [4] established a reliability model of the TPSS based on the Weibull distribution, used the proposed model to predict reliability, and obtained the reliability evolution process. However, this model is applicable only when fault occurrence follows the Poisson distribution, but this is not the case in practice. Moreover, Zhao et al. ignored the influence of meteorological conditions. The catenary system is completely exposed to the external meteorological conditions. The meteorological conditions have a significant influence on the catenary system operation [5]. Recently, Wang et al. [6] studied the influence of

✉ Sheng Lin
slin@swjtu.edu.cn

Qinyang Yu
yuqinyang157@163.com

Zhen Wang
wz1031308410@163.com

Ding Feng
fengding87@hotmail.com

Shibin Gao
gao_shi_bin@163.com

[1] School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China

[2] Laiwu Power Supply Company, State Grid Shandong Electric Power Company, Laiwu 271100, China

external environment on running state of the catenary system, and established the reliability evaluation model in three-state weather.

In power systems, the influence factors such as the external environment on power load forecasting, life prediction, and fault prediction has been highlighted [7, 8]. The power load forecasting methods, which consider the influence of weather conditions, have made significant progress in the weather-sensitive load [9, 10]. In addition, using the real-time electricity price, He et al. [11] proposed a method for forecasting the probability density of the power load. In terms of life prediction, scholars [12–14] used the rough set theory, cross-entropy theory, stochastic process simulation, and other methods to predict the equipment remaining life, and considered the influence of the external service environment on electrical equipment. Andre et al. [15, 16] used the Monte Carlo simulation to develop a model for the prediction of fault rate, fault type, and fault duration of transmission line and bus, and forecasted the annual outage times of the power system. Their model was based on the history of fault data, but the influence of the external environment on transmission lines was ignored. In [17], indexes including the meteorological sensitivity rate, difference of fault number, outage time were introduced to reflect the difference of transmission line risks for different meteorological disasters. In [18, 19], the temporal characteristics of transmission line faults were analysed, the time-varying fault rate simulation model was established, and the fault time distribution was simulated for risk assessment of a transmission line. A fault warning method based on the support vector machine (SVM) and AdaBoost method were proposed in [20]. All the above-mentioned studies consider the influence of external meteorological environment on power system on various levels, which can provide a reference for catenary fault prediction. As there are great improvements in the data acquisition, monitoring, and system management, catenary fault prediction can be supported with comprehensive data. Thus, it is significantly important to consider the overall influence of meteorological conditions on the fault prediction of catenary system.

The main objective of this work is to develop a catenary fault prediction method which can accurately and timely predict the catenary fault based on the external meteorological conditions, and provide decision support for the operation and maintenance of HSRs. In this paper, based on the AdaBoost algorithm, a method is proposed to predict the catenary fault. The proposed method establishes the mapping relation between meteorological conditions and catenary faults. It can predict catenary fault accurately if the meteorological conditions are provided.

The remainder of this paper is organized as follows. Section 2 introduces the influence of meteorological conditions on catenary faults. Section 3 briefly describes the AdaBoost and single decision tree algorithms. Section 4 presents the pre-processing method for historical statistical data and construction of training samples. A case study and the result analysis are provided in Sect. 5, followed by the conclusions in Sect. 6.

## 2 Influence of meteorological conditions on catenary faults

The catenary system is completely exposed to the complex environment. According to field surveys by a railway bureau, the meteorological conditions are one of the influential factors that cause catenary faults. In this work, a trip of the TPSS caused by the catenary system is regarded as a catenary fault, and the influence of meteorological conditions on the catenary fault occurrence is analysed quantitatively.

### 2.1 Temporal distribution characteristics of catenary faults

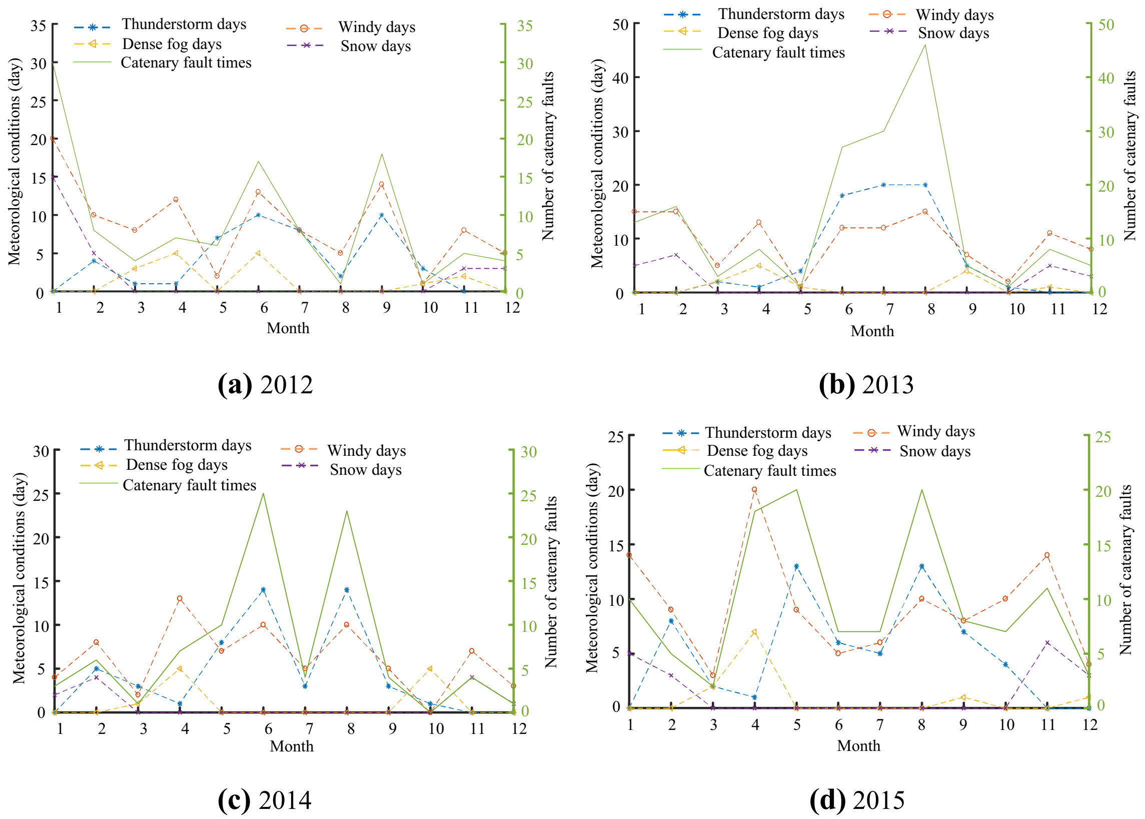
The number of catenary faults and their causes can be collected by field surveys. The results in [21] show that the working state of a catenary system is highly influenced by the external meteorological conditions, such as thunderstorms, gale, snow, and others. The number of catenary faults on a monthly basis under various meteorological conditions was collected by the railway bureau in northwest China from 2012 to 2015, as shown in Fig. 1.

According to Fig. 1, the most influential meteorological conditions in northwest China are, respectively, the gale and dense fog from March to April, the thunderstorm and gale from May to October, and the snow and gale from November to February. Meanwhile, when the days of the most influential meteorological condition increase or decrease, the number of catenary faults changes correspondingly. Therefore, there is a strong correlation between the meteorological conditions and the number of catenary faults.

### 2.2 Spatial distribution characteristics of catenary faults

In order to depict the spatial distribution characteristics of catenary faults, the catenary fault frequency (CFF) is introduced and defined as

$$C_{\mathrm{FF}} = \frac{\sum_{i=1}^{z} o_i}{\sum_{i=1}^{z} l_i}, \qquad (1)$$



**(a)** 2012

**(b)** 2013

**(c)** 2014

**(d)** 2015

**Fig. 1** Number of catenary faults and different meteorological days for 2012–2015

where $C_{\mathrm{FF}}$ indicates the catenary fault frequency in a year per kilometre, $l_i$ the length of line $i$, $o_i$ the number of catenary faults in a year, and $z$ the number of lines.

According to the data for central China in the period of 2012–2015, the corresponding CFF for each power supply section of Wuhan Bureau is shown in Fig. 2, which is calculated by Eq. (1).

As can be seen in Fig. 2, catenary fault frequency is diverse across regions. Namely, the CFF of Wuchang region is the largest, reaching the maximum of 0.85 times/km in 2012, and then followed by those of the Hanyang and Huangzhou regions with the CFF of more than 0.5 times/km in three statistical years. There was no catenary fault in the Jingzhou region during 2013–2015 and in the Wuxue region in 2012 and 2013. Meanwhile, the CFF of the Huangpo region is the lowest within the whole statistical period. Therefore, it can be concluded that CFF is strongly correlated to the geographical locations.

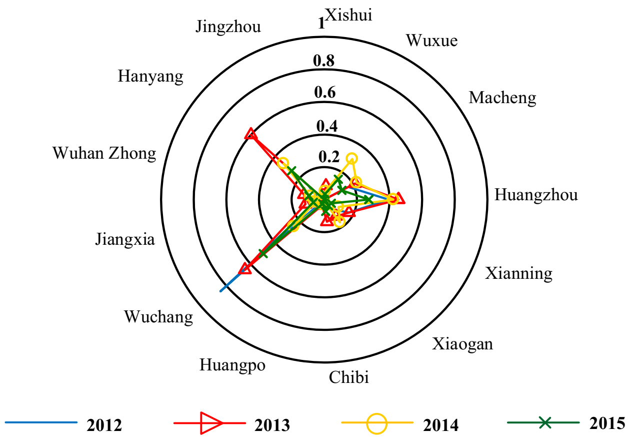
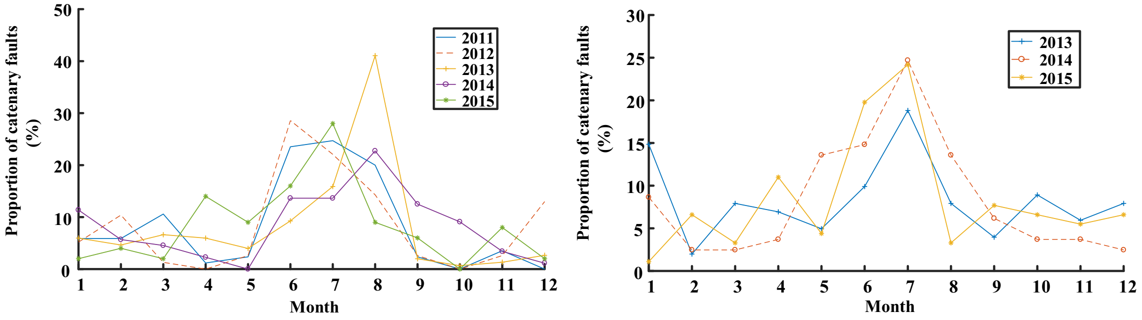


**Fig. 2** CFF in central China regions in the period of 2012–2015

In order to reveal the temporal and geographical correlation between the meteorological conditions and number of catenary faults, the fault data from the railway bureaux in northwest and central China are statistically analysed on a monthly basis, and the results are shown in Fig. 3.

Figure 3 indicates that the catenary faults in these two regions are mainly concentrated in June, July, and August. However, in December and January, the proportion of catenary faults in northwest China is higher than in central China. In view of the meteorological characteristics of the two regions, the main reasons for such results may be concluded as follows. Both in central and northwest China, there is the maximum amount of thunderstorm, gale, rain and high temperature in June, July, and August. Besides, snow and low temperature mainly occur in December and January. In central China, the summer lasts for a long time,



**(a)** Central China **(b)** Northwest China

**Fig. 3** Proportion of catenary faults in different regions in China

and the weather conditions do not fluctuate drastically during winter. In addition, the catenary system is almost unaffected by icing due to fewer snow and low temperature. Therefore, the fault distribution of the catenary system in central China can be approximated by a “single-peak” model. In contrast, the northwest region has a longer winter with snow and ice. Therefore, the fault distribution of catenary system in the northwest region can be approximated by a “peak-valley” interlaced model.

## 2.3 Analysis of meteorological conditions influence on catenary faults

The influence of meteorological conditions on catenary faults is always reflected in factors such as precipitation of rainstorm, heavy rain, moderate rain, thunderstorm, shower and light rain, wind speed, and temperature [21, 22].

1. Influence of precipitation. On the one hand, precipitation affects air humidity and insulation performance, and causes flashover because of the damp. Moreover, the water flow on the equipment surface can easily cause a short circuit. On the other hand, if there is lightning in rainy days, the lightning may lead to overvoltage and insulation damages; moreover, the overvoltage may invade the substation and cause trip.
2. Influence of wind speed. First, high wind speeds lead to catenary wire tension. Second, the gale causes the vibration of catenary wire and affects the current collection performance of the pantograph. Most importantly, the branches, plastics, and other foreign bodies blew by the gale may hang from the catenary, resulting in the short circuit.
3. Influence of temperature. The high temperature leads to the large tension of contact wires and short insulation distance, resulting in the short circuit. Meanwhile, under the low temperature, ice accumulates on a wire, which interrupts the current flow from contact wire to the pantograph.

## 2.4 Statistical analysis on influential factors of catenary faults

The influential factors are analysed using the actual data of the Beijing–Shanghai HSR (with a length of 1318 km) collected in the period of 2012–2015. The statistical results are shown in Table 1. Moreover, weather-related fault rate (WRFR) is introduced to represent the correlation between various meteorological conditions and the number of catenary faults. It indicates the frequency of catenary faults under a particular meteorological condition:

$$W_{\mathrm{RFR}} = \frac{\sum_{i=1}^{z} q_i}{t_{\mathrm{WB}} \cdot \sum_{i=1}^{z} l_i}, \tag{2}$$

where, $q_i$ denotes the number of catenary faults on line $i$ under the particular meteorological condition, $l_i$ denotes the length of line $i$, $t_{\mathrm{WB}}$ is the statistical time of a certain weather condition, and $z$ is the number of lines.

**Table 1** Statistical results of the data form Beijing–Shanghai HSR in the period 2012–2015

| Weather condition | Gale | Dense fog | Thunder storm | Moderate rain | Heavy rain | Normal weather |
|---|---|---|---|---|---|---|
| Statistical time (day) | 6 | 7 | 58 | 1 | 2 | 26 |
| Number of faults | 27 | 30 | 261 | 5 | 11 | 65 |

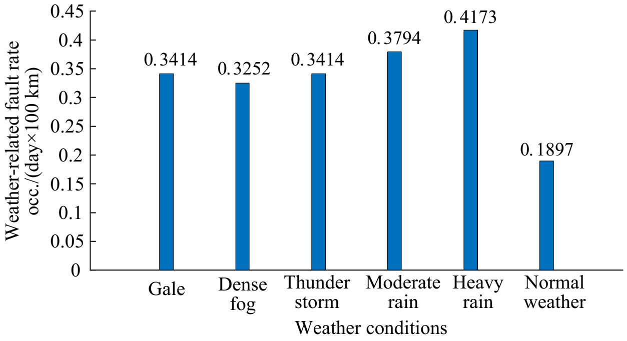


**Fig. 4** FRWB under different weather conditions

Using the statistical data given in Table 1 and the Eq. (2), the WRFR can be calculated as shown in Fig. 4.

As can be seen in Fig. 4, the WRFR under the gale, dense fog, and rain is higher than that under the normal weather. The highest fault rate is under the heavy rain condition. In general, the worse the weather is, the greater the possibility of a fault is. The influence of multiple uncertain factors makes it difficult to build an accurate mathematical model for catenary faults. In fact, there is a coupling relationship between various meteorology conditions. The catenary faults prediction is to determine whether the system could work healthily in the next period of operation with the current system state. It is often based on the massive multi-source data provided by the monitoring system. The fault prediction can be viewed as a classification prediction problem with supervised learning. In most cases, the learner accuracy is significantly influenced by training data and its distribution, and it is hard to build accurate classifiers directly. However, it is easier to generate a relatively accurate weak classifier. The AdaBoost algorithm is one of the most widely used machine learning methods for training different weak classifiers using the same training set. After training, the weak classifiers can be combined into a strong classifier. Namely, by combining the attributes of weak classifiers, the resultant classifier can possess a stronger generalization ability.

## 3 AdaBoost algorithm

### 3.1 Basic theory of AdaBoost algorithm

The AdaBoost algorithm is an important characteristic classification algorithm for machine learning, and it is widely applied to the power system fault warning [20], wind speed prediction [23], and other fields [24, 25]. Zhang et al. [26] compared the prediction accuracy of SVM, BP neural network, and AdaBoost, and indicated the superiority of AdaBoost algorithm.

The basic idea of the AdaBoost algorithm is to integrate a large number of weak classifiers that have a general classification ability to form a classifier with a strong classification ability. The specific steps of the AdaBoost algorithm are as follows.

1. Select a weak learning algorithm $C$ based on a single decision tree, and construct a training set $\boldsymbol{G}$ which is expressed as $\boldsymbol{G} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p), \ldots, (x_m, y_m)\}$, where $m$ denotes the number of samples.
2. Assume that the sample weight distribution $\boldsymbol{V}_n$ represents the weight of a sample in the $n$th iteration. Initialize the sample weights, $\boldsymbol{V}_1 = (v_1, v_2, \ldots, v_m) = (1, 1, \ldots, 1)/m$, $n = 1, 2, \ldots, N$, where $N$ denotes the number of iterations.
3. When $n = 1, 2, \ldots, N$, train a weak classifier $C_n(\boldsymbol{X})$ by the single decision tree method and classify the original training set $\boldsymbol{X}$ by $C_n(\boldsymbol{X})$; the classification result is expressed as $C_n\ (\boldsymbol{\alpha}_j)$, $\boldsymbol{X} = (x_1, x_2, \ldots, x_p, \ldots, x_m)$.
4. Calculate the classification error rate of $C_n(\boldsymbol{X})$ by

$$\varepsilon_n = \sum_{i=1}^{m} V_n(p) \cdot I\left(C_n\left(\boldsymbol{\alpha}_j\right) \neq y_p\right), \tag{3}$$

where $I\left(C_n\left(\boldsymbol{\alpha}_j\right) \neq y_p\right)$ is equal to 1 when $C_n\left(\boldsymbol{\alpha}_j\right) \neq y_p$; otherwise, $I\left(C_n\left(\boldsymbol{\alpha}_j\right) \neq y_p\right)$ is equal to 0.

5. Calculate the weight of $C_n(\boldsymbol{X})$ by

$$a_n = \frac{1}{2}\ln\left(\frac{1-\varepsilon_n}{\varepsilon_n}\right), \tag{4}$$

6. Update sample weight distribution:

$$\begin{aligned} V_{n+1}(p) &= \frac{V_n(p)}{Z_n} \times \begin{cases} \mathrm{e}^{-a_n}, C_n\left(\boldsymbol{\alpha}_j\right) = y_p \\ \mathrm{e}^{a_n}, C_n\left(\boldsymbol{\alpha}_j\right) \neq y_p \end{cases} \\ &= \frac{V_n(p)}{Z_n} \cdot \mathrm{e}^{-a_n y_p C_n\left(\alpha_j\right)}, \end{aligned} \tag{5}$$

where $Z_n = \sum_{p=1}^{m} V_n(p) \cdot \mathrm{e}^{-a_n y_p C_n\left(\boldsymbol{\alpha}_j\right)}$ denotes the normalization factor, such that $\sum_{p=1}^{m} V_{n+1}(p) = 1$.

7. Repeat Steps 3–6 for $N$ times to obtain $N$ different weak classifiers.
8. Combine all the trained weak classifiers into one strong classifier which is defined by

$$y = C(\boldsymbol{X}) = \operatorname{sgn}\left[\sum_{n=1}^{N} a_n C_n\left(\boldsymbol{\alpha}_j\right)\right]. \tag{6}$$

### 3.2 Construction of weak classifiers

In this work, the single decision tree [27, 28] is chosen to construct weak classifiers. The decision tree makes a decision by using the threshold division method for a single feature vector. This method has the following advantages: short computation time, fast calculation, and certain accuracy. In addition, this method can be well adapted to the AdaBoost algorithm. The specific steps of the single decision tree are as follows.

1. Assume a weight vector $\boldsymbol{V}_n$= ($v_1$, $v_2$, …, $v_p$, …, $v_m$), where $m$ is the number of samples.
2. Extract the characteristic values of each column in matrix $\boldsymbol{G}$ to form a new vector $\boldsymbol{\alpha}_{\boldsymbol{j}} = \left(x_{1-j}, x_{2-j}, \ldots, x_{p-j}, \ldots, x_{m-j}\right)^{\mathrm{T}}$, $j$ = 1, 2, …, $s$, where $s$ is the number of the characteristics.
3. Determine the threshold $H_k$ according to the data size of vector $\boldsymbol{\alpha}_j$:

$$\begin{cases} H_k = \min\left(\boldsymbol{\alpha}_{\boldsymbol{j}}\right) + (k-1) \cdot H_{\text{step}} \\ H_{\text{step}} = \left(\max\left(\boldsymbol{\alpha}_{\boldsymbol{j}}\right) - \min\left(\boldsymbol{\alpha}_{\boldsymbol{j}}\right)\right)/K \end{cases}, \tag{7}$$

where $k$ = 0, 1, 2, …, $K$, $k$ is the number of steps; $H_{\text{step}}$ is the step length; $\max\left(\boldsymbol{\alpha}_{\boldsymbol{j}}\right)$ and $\min\left(\boldsymbol{\alpha}_{\boldsymbol{j}}\right)$ are the maximum and minimum values in the vector.

4. Initialize an $m \times 1$ column vector $\boldsymbol{\beta}_j$, and classify each element of vector $\boldsymbol{\alpha}_j$ by mode 0 and mode 1 to obtain the classifications $\beta_j^0$ and $\beta_j^1$, respectively.
5. Initialize an $m \times 1$ column vector $\boldsymbol{e}$, and compare the corresponding elements of $\boldsymbol{\beta}_j^0$, $\boldsymbol{\beta}_j^1$ and $\boldsymbol{Y} = \left(y_1, y_2, \ldots, y_m\right)^{\mathrm{T}}$. If the obtained values are the same, the elements in the respective location of $\boldsymbol{e}$ are modified to 0, and the modified vectors are denoted as $\boldsymbol{e}_K^0$ and $\boldsymbol{e}_K^1$. Then, use Eq. (8) to calculate the error rate of the two classification methods by

$$E_K^r = \begin{cases} \boldsymbol{V}_n \boldsymbol{e}_K^0, & \text{model } 0 \\ \boldsymbol{V}_n \boldsymbol{e}_K^1, & \text{model } 1 \end{cases}, \tag{8}$$

where $r$ is equal to 0 or 1, and it expresses the classification method.

6. Repeat Step 3–5 $K$ times, and record the error rates of classifiers with the corresponding thresholds and classification models.
7. Repeat Step 2–6 $s$ times, and select the eigenvector $\boldsymbol{\alpha}_j$, whose threshold equal to $H_K$ and classification models correspond to the minimum error rate. Finally, calculate the classification function of a weak classifier by

$$C_{\mathrm{t}}\left(\boldsymbol{\alpha}_{\boldsymbol{j}}\right) = \begin{cases} -1, & x_{p-j} > H_K \\ 1, & x_{p-j} \leq H_K \end{cases} \text{ model } 0 \quad \text{or} \quad \begin{cases} -1, & x_{p-j} \leq H_K \\ 1, & x_{p-j} > H_K \end{cases} \text{ model } 1 \, . \tag{9}$$

## 4 Fault prediction on catenary system

### 4.1 Statistic and process input data for AdaBoost

As the field data contains much complex information, it is difficult to predict the catenary faults directly. Namely, the data should be first screened for validity. The required data can be divided into two types: historical running-state data and meteorological data. It also includes the catenary operating states, catenary fault types, protection information, catenary outage time, operation conditions, and weather information during the predicted period. The data types and sources are presented in Table 2. The meteorological data should be standardized and transformed into a mathematical form by attribute construction and discretization.

#### *4.1.1 Attribute construction*

The attribute sets of meteorological conditions include the precipitation grades, mean temperature grades, and wind scales during daytime and night.

#### *4.1.2 Discretization of meteorological data*

1. According to the rainfall intensity, the precipitation is divided into seven grades as shown in Table 3.
2. Use the equal-width division method to discretize the continuous temperature variables:

**Table 2** Data types and sources

| Data type | Specific data | Data sources |
|---|---|---|
| Historical running-state data | Catenary operating state<br>Catenary fault type<br>Protection action information<br>Outage time of TPSS<br>Fault time | Faults and maintenance information recorded by the railway bureau |
| Historical meteorological data | Meteorological information during three successive days, including the precipitation, wind speed, and temperature | Faults record from the railway bureau<br>Meteorological monitoring system<br>Meteorological information system |
| Meteorological data during the predicted period | Meteorological information during the predicted period including the precipitation, wind speed, and temperature | Meteorological information system<br>Weather forecasting |

**Table 3** Precipitation grade classification

| Precipitation event | Rainstorm | Heavy rain | Moderate rain | Thunderstorm | Shower | Light rain | No rain |
|---|---|---|---|---|---|---|---|
| Grade | 6 | 5 | 4 | 3 | 2 | 1 | 0 |

$$P_f = \left[T_{\min} + (f-1)\frac{T_{\max} - T_{\min}}{F}, T_{\min} + f\frac{T_{\max} - T_{\min}}{F}\right], \tag{10}$$

where $P_f$ refers to the range of the temperature level, $f = 1, 2, \ldots, F$, $F$ is the number of divisions, and $T_{\max}$ and $T_{\min}$ denote the maximum and minimum temperatures in the statistical time, respectively.

3. Classify the wind power into 0–12 grades according to the standard of China Meteorological Administration.

## 4.2 Construction of sample set

The catenary fault may be caused by impact effect of weather conditions. For example, lightning or strong wind leads to short-circuit trip of the TPSS. On the other hand, it may be a product of cumulative effects from external meteorological conditions, such as short circuit due to low sag of contact line over long time of high temperatures and flashover of the insulation device caused by continuous rainfall.

The external meteorological conditions are considered as a characteristic vector $\boldsymbol{X}$ that affects the catenary fault occurrence, and $\boldsymbol{Y}$ that denotes whether there is a fault on catenary. The sample set is constructed according to Sect. 4.1. Suppose that there are $m$ data samples; then, the constructed sample set can be expressed as matrix $\boldsymbol{G}$, where $p = 1, 2, \ldots, m$, $j = 1, 2, \ldots, s$, and $s$ is the number of characters that could be taken into account, and the matrix $\boldsymbol{G}$ is expressed as

$$\boldsymbol{G} = \left\{ \begin{matrix} x_{1-1} & x_{1-2} & \cdots & \cdots & \cdots & x_{1-j} & \cdots & \cdots & \cdots & x_{1-s} & y_1 \\ x_{2-1} & x_{2-2} & \cdots & \cdots & \cdots & x_{2-j} & \cdots & \cdots & \cdots & x_{2-s} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{p-1} & x_{p-2} & \cdots & \cdots & \cdots & x_{p-j} & \cdots & \cdots & \cdots & x_{p-s} & y_p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{m-1} & x_{m-2} & \cdots & \cdots & \cdots & x_{m-j} & \cdots & \cdots & \cdots & x_{m-s} & y_m \end{matrix} \right\}, \tag{11}$$

where $x_{p\text{-}j}$ denotes a set of influential factors such as precipitation, temperature, and wind scale on sample $p$; $y_{\mathrm{p}} = (-1 \vee 1)$, the value of $-1$ means no catenary fault, and the value of 1 a catenary fault.

## 4.3 Catenary fault prediction based on AdaBoost

The catenary fault prediction based on the AdaBoost algorithm includes the following steps.

1. Input the training data, including the catenary fault data and meteorological data.
2. Set the initial weight $\boldsymbol{V}_1$ and iteration number $N$, and initialize the AdaBoost algorithm.
3. Update the weights through the iterative computation. Train the optimal decision tree by different weights of $\boldsymbol{V}_n$. Construct multiple weak classifiers, and combine them with the weights to generate a strong classifier.
4. Use the future meteorological data provided by the Weather Forecast as an input data for fault prediction, and obtain the final prediction result using the trained strong classifier.

The specific calculation flow chart is shown in Fig. 5.

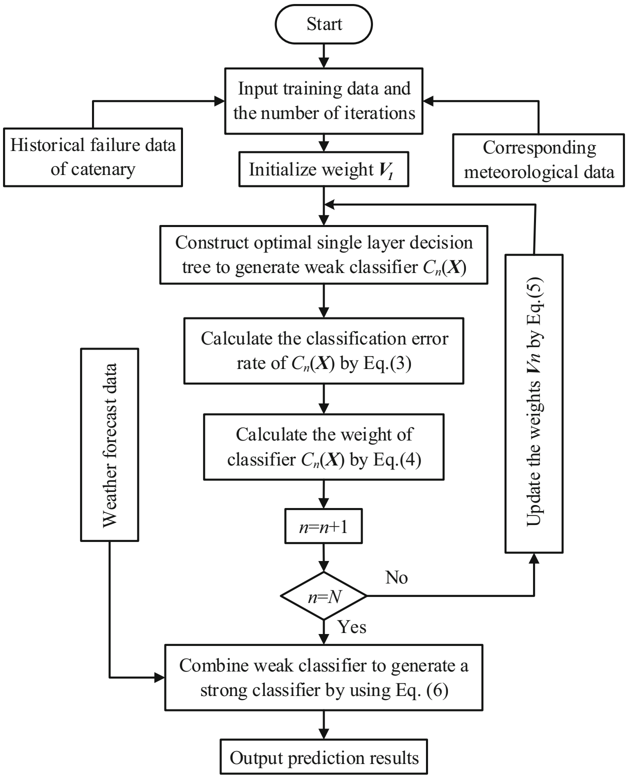

Start
Input training data and the number of iterations
Historical failure data of catenary
Corresponding meteorological data
Initialize weight $V_1$
Construct optimal single layer decision tree to generate weak classifier $C_n(X)$
Calculate the classification error rate of $C_n(X)$ by Eq.(3)
Calculate the weight of classifier $C_n(X)$ by Eq.(4)
$n$=$n$+1
$n$=$N$
No
Yes
Weather forecast data
Update the weights $Vn$ by Eq.(5)
Combine weak classifier to generate a strong classifier by using Eq. (6)
Output prediction results

**Fig. 5** Flow chart of the catenary fault prediction

**Table 4** Temperature classification

| Temperature (°C) | [17–20) | [20,23) | [23,25) | [25,28) | [28,31) | [31–33] |
|---|---|---|---|---|---|---|
| Grade | 1 | 2 | 3 | 4 | 5 | 6 |

## 5 Case study

### 5.1 Data selection and standardization

During data pre-processing, we found that the real-time meteorological data records in 2014 were not complete, as they could not match the catenary operation records. Therefore, we selected the real-time meteorological and catenary operation data in 2011, 2012, 2013, and 2015 from the railway bureau. The statistical data collected in June 2011, June 2012, and June 2013 were selected as the training data, and the data collected in June 2015 was selected as the test data. The training data consisted of 43 events, including 10 fault events and 33 normal events. The test data consisted of 18 events, including 2 fault events and 16 normal events.

The historical data was pre-processed by the steps introduced in the previous chapter. The field data analysis revealed that in the selected samples, there is no fog-related fault. At the same time, the Meteorological Information System showed that there was no foggy day in the seasons of study. Therefore, fog was not considered in the training and test data. In the selected samples, the lowest temperature was 17 °C, and the highest temperature was 33 °C. The detailed temperature classification calculated by Eq. (10) is given in Table 4.

The data samples include the recording time, precipitation grade, temperature grade, wind scale, and catenary state. Through data pre-processing, the training sample set and test sample set are presented in Tables 5 and 6.

### 5.2 Construction of strong classifier

The training data was divided into two categories. One category only shows the influence of precipitation, and the other one shows the joint influence of precipitation, wind scale, and temperature. For simplicity, we only take the influence of precipitation grade as an example to illustrate the processes of constructing the weak classifiers based on the single decision tree and training the weak classifier based on the AdaBoost.

The representation matrix of training data about precipitation grade was as follows:

$$\boldsymbol{G} = \left\{ \begin{matrix} x_{1-1} & x_{1-2} & x_{1-3} & x_{1-4} & y_1 \\ x_{2-1} & x_{2-2} & x_{2-3} & x_{2-4} & y_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{p-1} & x_{p-2} & x_{p-3} & x_{p-4} & y_p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{m-1} & x_{m-2} & x_{m-3} & x_{m-4} & y_m \end{matrix} \right\} = \left\{ \begin{matrix} R_{1\mathrm{td}} & R_{1\mathrm{tn}} & R_{1\mathrm{y}} & R_{1\mathrm{b}} & y_1 \\ R_{2\mathrm{td}} & R_{2\mathrm{tn}} & R_{2\mathrm{y}} & R_{2\mathrm{b}} & y_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_{p\mathrm{td}} & R_{p\mathrm{tn}} & R_{p\mathrm{y}} & R_{p\mathrm{b}} & y_p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_{\mathrm{mtd}} & R_{\mathrm{mtn}} & R_{\mathrm{my}} & R_{\mathrm{mb}} & y_{\mathrm{m}} \end{matrix} \right\}, \tag{12}$$

where, $R_{\mathrm{ptd}}$, $R_{\mathrm{ptn}}$, $R_{\mathrm{py}}$, and $R_{\mathrm{pb}}$ indicate the precipitation grades in the current daytime, current night, the average precipitation grade on the previous day, and the average precipitation grade for 2 days before the current day with respect to sample $p$, respectively.

Then, the weights were initialized as $\boldsymbol{V}_1 = (1, 1, \ldots, 1)/43$. Following the weak classifier calculation process, the optimal decision feature vector of the first weak classifier was obtained as $\boldsymbol{\alpha}_2 = (x_{1\text{-}2}, x_{2\text{-}2}, \ldots, x_{\mathrm{p}\text{-}2}, \ldots, x_{43\text{-}2})^{\mathrm{T}}$, and the classification function was given as:

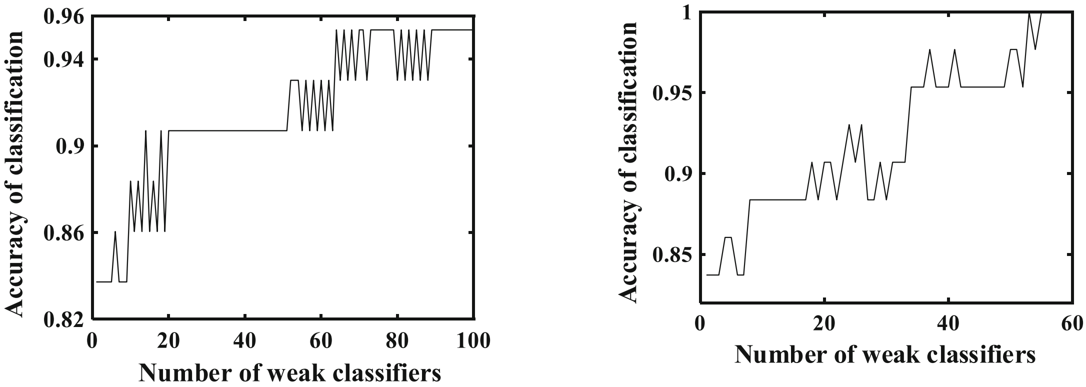
$$C_1(\boldsymbol{\alpha}_2) = \begin{cases} -1, & x_{p-2} \leq 5.4 \\ 1, & x_{p-2} > 5.4 \end{cases}, \tag{13}$$

**Table 5** Training sample set

| Time | Weather | | Average precipitation | | Temperature | | Wind scale | | Running state |
|---|---|---|---|---|---|---|---|---|---|
| | In day | In night | Previous day | Two days before | In day | In night | In day | In night | |
| 2011.06.03 | 2 | 4 | 0 | 0 | 4 | 1 | 2 | 2 | − 1 |
| 2011.06.05 | 0 | 1 | 3 | 3 | 4 | 2 | 5 | 4 | 1 |
| – | – | – | – | – | – | – | – | – | – |
| 2013.06.30 | 2 | 0 | 1 | 1 | 5 | 4 | 2 | 2 | − 1 |

**Table 6** Test sample set

| Time | Weather | | Average precipitation | | Temperature | | Wind scale | | Running state |
|---|---|---|---|---|---|---|---|---|---|
| | In day | In night | Previous day | Two days before | In day | In night | In day | In night | |
| 2015.06.01 | 4 | 6 | 0 | 0 | 6 | 3 | 5 | 2 | 1 |
| 2015.06.02 | 4 | 1 | 5 | 0 | 4 | 3 | 2 | 2 | − 1 |
| – | – | – | – | – | – | – | – | – | – |
| 2015.06.30 | 4 | 4 | 0 | 0 | 5 | 2 | 2 | 2 | − 1 |



**(a)** Influence of precipitation

**(b)** Joint influence of precipitation, wind and temperature

**Fig. 6** Classification accuracy under different meteorological conditions

where $x_{p\text{-}2}$ represents the eigenvalues of an eigenvector $\boldsymbol{\alpha}_2$ in a line $p$, and 5.4 is the threshold value calculated by Eq. (7).

Finally, the error rate of each classifier was calculated and the weights were adjusted to obtain a strong classifier by the AdaBoost algorithm. Using the two above-mentioned categories, two different training sets were obtained, respectively. Then, the accuracy on each training set was calculated, as shown in Fig. 6.

In Fig. 6, the accuracy on both training sets increases with the number of weak classifiers. In Fig. 6a, the maximum accuracy is 0.9535, and the curve tends to become stable when the number of classifiers reaches the value of 64. In Fig. 6b, the maximum accuracy of 1 is achieved when the number of classifiers reaches the value of 53. Thus, in the case of joint influence of precipitation, wind, and temperature, the accuracy of classification is higher and less number of weak classifiers is required compared with the case of a single influence of precipitation.

By comparison, it is observed that the results of the first training set have more oscillations and lower accuracy. Thus, we select the precipitation, wind, and temperature as influential factors to construct weak classifiers.

### 5.3 Results of catenary faults prediction

The proposed fault prediction method was evaluated through a comparison with the decision tree and BP neural network algorithm on the test data, and the obtained results are shown in Table 7. And the bold number in Table 7 indicates the inaccurate prediction result.

According to the results presented in Table 7, the prediction accuracy of the AdaBoost was 88.89%, and almost all the catenary faults were correctly predicted except for two errors. The first one was the data on 02 June 2015, and the second one was the data on 26 June 2015. The AdaBoost algorithm predicted that there was a high fault probability on catenary under current meteorological conditions, which is a false alarm. With more sample data, the prediction accuracy of the AdaBoost algorithm can gradually stabilize at about 90% [24, 26].

The prediction accuracy of the decision tree is 77.8% and the BP algorithm is 83.3%, which were lower than that of the AdaBoost algorithm. In this paper, the single decision tree algorithm is the weak classification algorithm to construct the strong algorithm. Therefore, the prediction accuracy will be significantly lower than the AdaBoost algorithm. For the BP neural networks, although the

**Table 7** Fault prediction results on the test set

| Sample data | 2015.06.01 | 2015.06.02 | 2015.06.03 | 2015.06.04 | 2015.06.06 | 2015.06.07 |
|---|---|---|---|---|---|---|
| Actual state | 1 | − 1 | − 1 | − 1 | − 1 | − 1 |
| Prediction result | | | | | | |
| AdaBoost | 1 | **1** | − 1 | − 1 | − 1 | − 1 |
| Decision tree | 1 | − 1 | **1** | − 1 | − 1 | − 1 |
| BP neural network | 1 | − 1 | − 1 | − 1 | − 1 | **1** |
| Sample data | 2015.06.13 | 2015.06.14 | 2015.06.15 | 2015.06.16 | 2015.06.17 | 2015.06.20 |
| Actual state | − 1 | − 1 | − 1 | 1 | − 1 | − 1 |
| Prediction result | | | | | | |
| AdaBoost | − 1 | − 1 | − 1 | 1 | − 1 | − 1 |
| Decision tree | − 1 | − 1 | − 1 | 1 | **1** | **1** |
| BP neural network | − 1 | − 1 | − 1 | **− 1** | **1** | − 1 |
| Sample data | 2015.06.21 | 2015.06.23 | 2015.06.24 | 2015.06.25 | 2015.06.26 | 2015.06.30 |
| Actual state | − 1 | − 1 | − 1 | − 1 | − 1 | − 1 |
| Prediction result | | | | | | |
| AdaBoost | − 1 | − 1 | − 1 | − 1 | **1** | − 1 |
| Decision tree | − 1 | − 1 | − 1 | − 1 | **1** | − 1 |
| BP neural network | − 1 | − 1 | − 1 | − 1 | − 1 | − 1 |

training accuracy can reach 100%, the generalization effect is worse than the AdaBoost algorithm. Moreover, because of randomness in the learning phase, the BP algorithm may converge to local minima. In conclusion, the strong classifier constructed by the AdaBoost algorithm had a stronger generalization ability than the single decision tree and BP neural network.

However, the method of machine learning needs to be improved in the following aspects. First, the AdaBoost algorithm uses the single decision tree for weak classifiers construction in this work. Since only the decision tree is used in the training process, the accuracy of prediction results with decision tree is not high enough, which further decreases and limits prediction accuracy of the strong classifier. This problem may be solved by using better classification methods such as support vector machine (SVM). Furthermore, the AdaBoost algorithm constructs a strong classifier by updating the weights of different weak classifiers, but it pays more attention to the misclassified samples in the training process. Thus, the weights of samples that are easily misclassified will gradually increase with the number of iterations. This leads to the imbalance of samples and causes the decrease in classification accuracy. This problem can be solved by optimizing the weights updating process of the classifiers.

## 6 Conclusions

The external meteorological conditions, including the precipitation, wind speed, and temperature, have a significant impact on catenary fault. In this paper, the relationship between the catenary fault and meteorological conditions is analysed. The cumulative effect of meteorological conditions on the catenary system is taken into account in catenary fault prediction, and the AdaBoost algorithm is utilized to construct a strong classifier to predict the catenary fault by using the historical meteorological data. The obtained prediction results demonstrate that the AdaBoost algorithm could provide prediction for the catenary faults with an accuracy of 88.89% by considering the external meteorological conditions.

**Acknowledgements** This work was supported by the Scientific and Technological Research and Development Program of China Railway Corporation under Grant N2018G023, and by the Science and Technology Projects of Sichuan Province under Grants 2018RZ0075.

## References

1. Feng D, Lin S, Sun XJ et al (2017) A technical framework of PHM and active maintenance for modern high-speed railway traction power supply systems. Int J Rail Trans 1(5):1–25
2. Feng D, He ZY, Lin S et al (2017) Risk index system for catenary lines of high-speed railway considering the characteristics of time-space differences. IEEE Trans Transp Electr 3(3):739–749
3. Wang Z, Feng D, Lin S, et al (2016) Research on reliability evaluation method of catenary of high-speed railway considering weather condition. In: International conference on probabilistic methods applied to power systems, Beijing, China
4. Zhao F, Zhang QP, Wang SH (2015) Research of reliability prediction and modelling for traction power supply system. J Railw Sci Eng 12(3):678–684
5. Cheng HB, He ZY, Wang Q (2015) Analysis method for meteorological factor associated accident model of high-speed railway. Electr Power Autom 35(9):49–54
6. Wang Z, Lin S, Feng D et al (2018) Reliability evaluation method of catenary system considering the weather condition. J China Railw Soc 40(10):49–56
7. Chen P-C, Kezunovic M (2016) Fuzzy logic approach to predictive risk analysis in distribution outage management. IEEE Trans Smart Grid 7(6):2827–2836
8. Park C-H, Jang G, Thomas RJ (2008) The influence of generator scheduling and time-varying fault rates on voltage sag prediction. IEEE Trans Power Del 23(2):1243–1250
9. Tucci M, Crisostomi E, Giunta G et al (2016) A multi-objective method for short-term load forecasting in European countries. IEEE Trans Power Syst 31(5):3537–3547
10. Gao YJ, Sun YJ, Yang WH et al (2017) Weather-sensitive load's short-term forecasting research based on new human body amenity indicator. Proc Chin Soc Electr Eng 37(7):1946–1954
11. He YY, Liu Y, Han AY et al (2017) Short-term power load probability density forecasting method based on real time price and support vector quantile regression. Proc Chin Soc Electr Eng 37(3):768–778
12. Li LL, Zhang SN, Li ZG (2016) The life prediction method of relay based on rough set theory and relay's initial life information. Trans Chi Electr Soc 31(18):46–53
13. Li R, Liu HL, Lu Y et al (2014) A combination method for distribution transformer life prediction based on cross entropy theory. Power Syst Prot Control 42(2):97–103
14. Lei YG, Li NP, Lin J (2016) A new method based on stochastic process models for machine remaining useful life prediction. IEEE Trans Power Deliv 65(12):2671–2684
15. Dos Santos A, Barros MTCD (2015) Stochastic modelling of power system faults. Electr Power Syst Res 126:29–37
16. Dos Santos A, de Barros MTC (2016) Predicting equipment outages due to voltage sags. IEEE Trans Power Deliv 31(4):1683–1691
17. Wang J, Xiong XF, Liang Y et al (2016) Geographical and meteorological factor related trans-mission line risk difference assessment method and indexes. Proc Chin Soc Electr Eng 36(5):1252–1259
18. Wang J, Xiong XF, Liang Y et al (2016) The distribution of weather-related transmission line failure and its fitting. Electr Power Autom Equ 36(3):109–115
19. Wang J, Xiong XF, Liang Y et al (2016) Time-varying failure rate simulation model of transmission line and its application in power system risk assessment considering seasonal alternating meteorological disasters. IET Gener Transm Distrib 10(7):1582–1588
20. Wang J, Xiong XF, Liang Y et al (2016) Early warning method for transmission line galloping based on SVM and AdaBoost bilevel classifiers. IET Gener Transm Distrib 10(14):3499–3507
21. He ZY, Feng D, Lin S (2016) Research on security risk assessment for traction power supply. J Southwest Jiaotong Univ 51(3):418–429
22. Gu SQ, Feng WX, Zhao C (2015) Method of lighting hazard risk evaluation for traction electric network of high-speed railway. High Volt Eng 41(5):1526–1535
23. Wu JL, Zhang BH, Wang K (2012) Application of AdaBoost-based BP neural network for short-term wind speed forecast. Power Syst Technol 36(9):221–225
24. Li PF, Yan YD, Zhang KB et al (2018) Influence of training data on engine fault diagnosis based on AdaBoost. J Xi'an Polytech Univ 32(06):670–677
25. Cao YL, Gao S, Kan YX (2018) Influence of training data on engine fault diagnosis based on AdaBoost. J Civ Aviat Univ China 36(06):16–20
26. Zhang W, Sheng WX, Liu KY et al (2018) A prediction method of fault risk level for distribution network considering correlation of weather factors. Power Syst Technol 42(08):2391–2398
27. Zhang XF, Chen DQ, Yang YS et al (2018) Remote sensing inversion of highway land use based on decision tree. Highway 63(09):191–199
28. Hao JW (2017) Research in aircraft maintenance and tracking system based in data classification and prediction technology. North China Electric Power University