Based on the Large Data of Complaint Forecasting Models Research

Guilin Li, Xueyan Lan [a], Chunmei Wu

Dalian Jiao tong university, Institute of electrical information, Liaoning Dalian 116028, China

[a]1053017666@qq.com

Abstract

With the diversification of services and the growing number of users, the frequency of complaints of customers has increased dramatically. In order to provide the customers with quality services preferably, constantly expand the storage capabilities of the data, and actively use large data analysis technology in marketing, network operation and maintenance, optimize the network approach to enhance customer service and so on, and achieved some success. In this paper, we first use the large data mining technology to construct a fault - based complaint forecasting model by analyzing the main factors that from user's complaint. We also analyzed the relationship between the causes of complaints and the occurrence of the fault, based on the work orders and fault data collected by the operators. We then use this model to investigate the potential complaints and provide rationalization advice for complaints as well as reducing the amount of complaints and improving customer satisfaction.

Keywords

Complaint prediction model; clustering algorithm; neural network model; complaint data.

1. Introduction

For the time being, with the continuous development of China's communications business, the customers' complaints increased and the information services issues become the primary concern of users because of the particularity of communication services. It is necessary to establish a set of response processes for responding to complaints to customers, and providing customers with satisfactory answers in order to reduce the proportion of complaints and transform the complaints of customers to the satisfaction of customers[1-2].

The data of customers' complaints have an accumulated process. If the typical complaint cases in the complaint history data are extracted, the case database is established and perfected, and the experience of complaint handling can be accumulated, it would be better to predict and deal with the imminent occurrence of the complaint [3]. The improvement of customer service because of optimizing network is becoming a hot spot. In the whole process, we need to deal with a large number of data and use the data mining technology, such as association rules, decision tree, and neural network and support vector machine [4].

2. Data Set Description and Analysis

2.1 Data Preparation

The data is complaints data collected by the operator. It was divided into complaints work order data and fault data. In order to facilitate the modeling process, we unified the data formatted. The formatted data is shown, see Fig. 1.

Structured data format diagram
The single serial number：ID-056-20140930-00569 Complaint time：2014/9/30 22:03 Complaint content：Network reasons, sudden surge in users, major events, etc. Co-occurrence word set：Service class - Network reasons - base communications - base station failure Failure cause：Base station retirement service Model association rule results：Network reasons

Fig.1 A data record graph after the system is structured

2.2 Analysis of Complaints Data Sets

2.2.1 Data Ananlysis

In data analysis, the limitation analysis and spatial metric analysis are involved. First, the article defines a complaint record as TS_complaint, TS_complaint is selected the complaint history data centralized storage of a data record, a complaint history data set is defined as JH_complaint = JH_complaint={$TS\_complaint_1$， $TS\_complaint_2$,$TS\_complaint_3$,……$TS\_complaint_m$}.Then, the text mining of the TS_complaint for the historical data of the complaint is used to determine the spatial metric analysis of the complaint historical data. Assuming that there is a total of M records in the TS_complaint dataset, the main task of text mining is to segment the complaint history data and get the FC_word of the complaint history data. If the typical feature entry is found from the excavated features, it will be labeled with TZ_word. A typical feature term is used to form a typical characteristic term space. The dimension of an entry is determined by the number of words of the term, and also includes the time metric and the spatial mapping, and the vector text that is finally mapped to space [5].The time measurement for the complaint history is shown, see Fig. 2.The spatial metric for the complaint historical data is shown, see Fig. 3.

$$JH\_complaint\_YS=\begin{bmatrix} YS_{11}.....YS_{1n} \\ ... \quad ... \quad ... \\ YS_{m1}.....YS_{mn} \end{bmatrix} \tag{1}$$

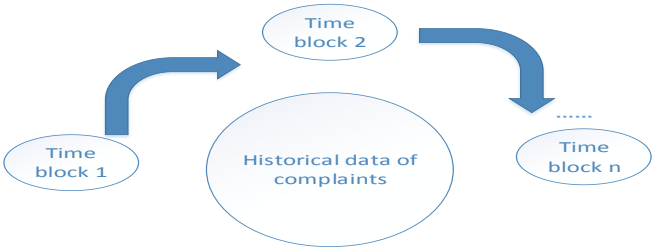

Fig.2 Schematic diagram of the time measurement of complaints historical data

Fig.3 Schematic diagram of the spatial measurement of complaints historical data

2.2.2 Typical Event Induction

Look for typical events in the complaint data to classify and predict the value of events [6].Data after finishing the complaint database, become history data, complaints against complaints work order and fault data are given, through the summary and preliminary qualitative analysis, summarizes the given data set of customer complaint events as shown, see Table 1, typical fault typical event as shown,see Table 2.

Table 1. Historical Data Set Customer Complaints Typical Event Statistics

No.	Type of complaint	Main content of complaints
1	Network class	Slow speed, can not open the page
2	Phone signal class	The phone has a signal but can not be used
3	Network coverage class	No signal on the phone
4	Wireless network class	The wireless network signal is strong but can not be used
5	Traffic class	Excess flow, but no reminder

Table 2. Typical types of events for the types of faults given to the data set

No.	Fault type	The main content of the fault
1	Base station retreat	Equipment damage, exit service
2	Derived base station retreat	Equipment damage, send a single deal
3	Warning of exit services	The device has failed and a warning has occurred

The events that will be mined from the history of the complaint are called a Typical Event, and the higher the frequency of typicality, the more practical it is[7].The complaint historical data contains a number of typical atypical events that are characteristic of typical atypical events. The difference between typical and atypical events is shown,see Table 3.

Table 3. Comparison of typical events and atypical events

Discrimination scale	Typical event	Atypical event
Cause	The cause is clear	The cause is not clear
Similarity	High	Low
The probability of occurrence	Table and higher	Unstable

3. Theoretical and Technical Support

3.1 Key Technologies for Data Processing

Data mining is the key to data processing. In the article, the weight of words to represent text, complaints historical data collection of text T, set complaints text $t \subset T$, so $d=\{c_1:q_1,c_2:q_2,\ldots\ldots c_m;q_m\}$,of which the c_m is the complaint history text word, q_m is the weight of the word, the general use of TD - IDE to calculate the weight of word.weighting formula is:

$$q_{i,j} = \mathrm{tf}_{ij} * \log \frac{N}{n_i} \tag{2}$$

Among them,qij represents the weight of ci in text tj,fij is the frequency of ci in text tj,ni is the total number of text ci in the complaint history data,N is the total number of complaints.N is the total number of complaints historical data.The complaint historical data is represented by the vector, you can calculate the distance between the two text, this distance is the text of the similarity measure of

the distance formula with Euclidean distance[8]. di,dj stands for the different historical data vectors, qki and qkj respectively for the Kth weight of text di,dj.The formula for the continental distance between text is:

$$DIS(d_i, d_j) = \left[\sum_{k=1}^{m} (q_{ki} - q_{kj})^2 \right]^{\frac{1}{2}} \quad (3)$$

3.2 K-Means Clustering Algorithm

We clustered the similar documents or hyperlinks and the results of the corresponding preprocessing can be analyzed based on K-means clustering. It contains two major groups of clustering analysis of complaint clustering mining and fault type [9-10].

The basic steps of the k-means algorithm are as follows:

Step 1, select k objects from n data objects as the initial cluster center;

Step 2, according to the mean of each cluster object (center object);

Step 3, recompute the average (center object) of each (changed) cluster class.

Step 4, back to step 2.

Here's a clear illustration of this, see Fig. 4:

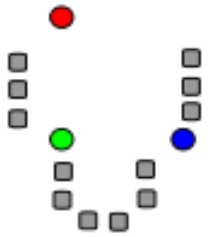

Fig.4 (a) initialize the cluster center (k = 3)

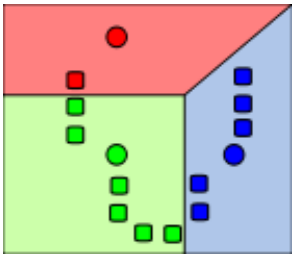

Fig.4 (b) the distance of sample to the center. k clusters are calculated

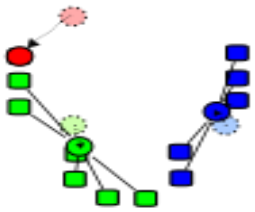

Fig.4 (c) Generate a new center for each cluster

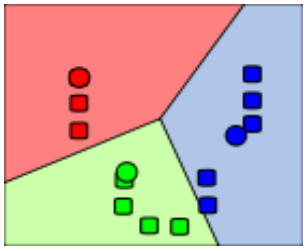

Fig.4 (d) Repeat STEP2 and STEP3 until the termination condition is met

4. 4.Based on the Model That Is Given to the Data

4.1 The Establishment of the Complaint Forecasting Model

First, you need to preprocess the data, then improve the efficiency of the data mining by reducing the dimension of the vector space according to the typical time feature.The data preprocessing diagram is shown,see Fig. 5.

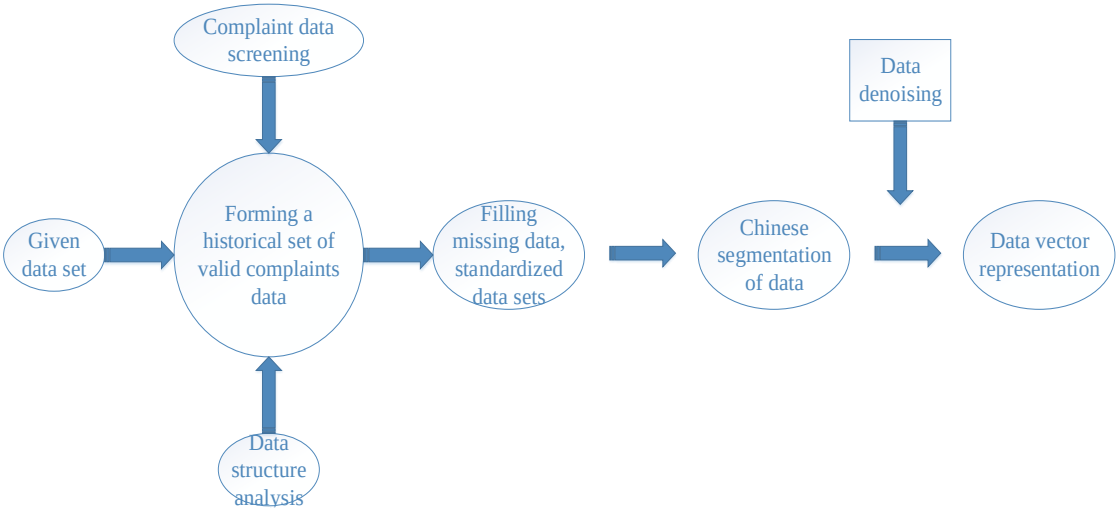

Fig.5 Schematic diagram of data preprocessing

After processing data, a typical case of a given data set needs to be mined[11].This definition of the complaint text is GX_words, and a preprocessed collection of historical data sets YCL_words can reflect the semantic relevance of the complaint text..The general extraction process for GX_words is shown,,see Fig. 6.

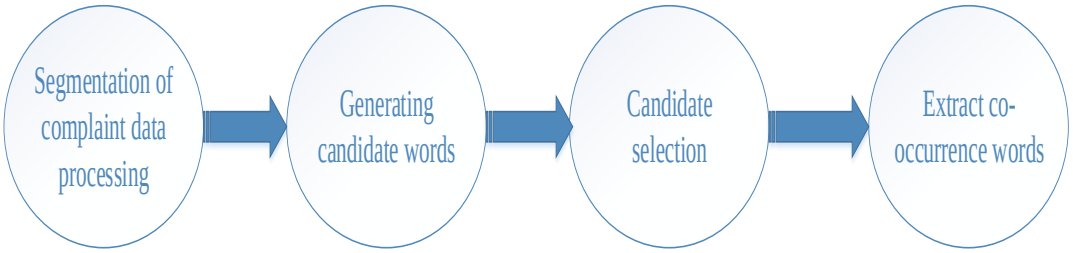

Fig.6 GX_words's general extraction process diagram

Based on the GX_words defined in the above definition,the commonality between the data is derived by using the association rules for modeling, which provides the basic criterion for predicting the system.The association rule mining is the relationship of each complaint to the natural text.

Word support:

$$\sup(c_i, c_j) = p(c_i, c_j) \tag{4}$$

The word's confidence:

$$Ide(w_i, w_j) = \frac{1}{2}\left(\frac{p(w_i, w_j)}{p(w_i)} + \frac{p(w_i, w_j)}{p(w_j)}\right) \tag{5}$$

Define the data mining space for the complaint history data:

$$\begin{aligned} &Data\_spa = (Com\_words, All\_words, \mathrm{Re}l) \\ &Com\_words = \{t\_word_1, t\_word_2 \cdots t\_word_n\}, t\_word_1 \in Com\_words\} \\ &Data\_spa \quad \mathrm{Re}l = \{r_1, r_2 \cdots r_k\} \\ &r = (t_x, t_y), t_x, t_y \in All\_words, Common\_words, \mu = \{\mu, \rho\}, \mu, \rho, \\ &t\_word_n, t\_word_m \end{aligned} \tag{6}$$

Among them, μ indicates the given degree of support, ρ Indicates a given confidence level.In space, the c file complaint text and the z word set form the matrix of c * z,which c represents a complaint text data in the complaint text collection, z represents the collection of words for the collection of complaints.So the support for the given data set is:

$$\sup(t\_word_x, t\_word_y) = \sum_{k=1}^{n}(t\_word_{kx} \times t\_word_{ky}) \quad (7)$$

$$Ide(t\_word_x, t\_word_y) = \frac{1}{2}\left(\frac{\sum_{k=1}^{n}(t\_word_{kx} \times t\_word_{ky})}{\sum t\_word_x} + \frac{\sum_{k=1}^{n}(t\_word_{ks} \times t\_word_{ky})}{\sum t\_word_y}\right) \quad (8)$$

According to the features of the complaint history data, this paper uses the improved condensed hierarchical clustering algorithm, which is all the complaints and text bottom-up merging into a tree and division level algorithm in the opposite direction.In this case, enter the collection of complaints historical data collection c and the co-existing word set GX_words[12-14].

4.2 Implementation of Typical Data Mining

The article mainly focuses on the case of complaint and malfunction type[15-16].First, you need to cluster the information about the pre-processed information: the first group is grouped according to the type of the malfunction, the location of the site, and the withdrawal situation of the base station; The second group was based on the number of customer complaints.In terms of its similarity, the greater the similarity, the closer the result is, whereas the opposite is also true.The result of the clustering algorithm based on the data is shown,see Fig. 7.

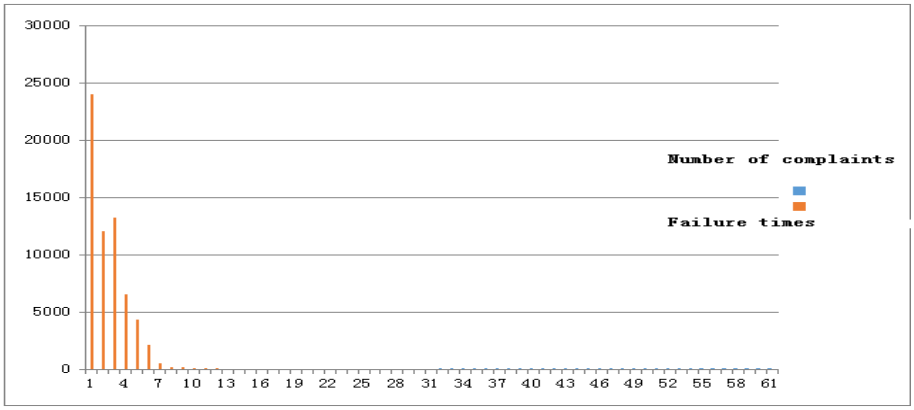

Fig.7 Cluster diagram based on historical data

As shown in the figure, for analyzing the corresponding icon within 17 times, for more than 17 complaints from customers set it to 0, and experimental verification when the number of clusters of 7 corresponding relationship is the most reasonable.Using the k-means algorithm to excavate 2364 corresponding complaint information, then classify it and divide it into 7 parts, and the data in each section is shown,see Table 4:

Table 4. the relevant complaints information table based on the historical data

Data section	1	2	3	4	5	6	7
Mean	1.465	0.841	31	202	265	87	921
variance	0.5	16.998	14.762	8.807	6.4	11.931	3.87

The reference utilization rate of the mean is approximately 38%, and the malfunction warning utilization is very low.

The calculated minimum confidence level is 20%, the minimum support is 10%, and the maximum number of complaints is 10.The probability of two kinds of complaints in the complaint and

malfunction association rules is 43.25% for both the I2 and the K8.Analysis can get customer complaint there are three main factors: network coverage problem, have signal but cannot be used (actual to cover blind spot problems), slow network (actual is too far away from the base station, belong to the weak coverage).By association rule, the main reason for customer complaints is the withdrawal of the base station, which is more than 77%.

5. Test Results and Analysis

(1) The main factors that cause customer complaints

Through the structural analysis of the data presented, the GX_words extraction and pretreatment of the data using the predictive model,The collected 6104 complaints data record and keyword extraction by co-occurrence words that have signal in complaints work order but unable to use, network coverage problem, slow network problems such three factors as keywords, and respectively marked as t_wordk1、t_wordk2、t_wordk3 and plug in the formula,so get the weight of three factors to network coverage problems over have signal but unable to use slow network problems.Through the normalized processing of specific weight is: network covering problem (35%), a signal but can't use accounted for 30%, slow speed problem (25%).

(2) The relationship between the cause and the failure of the customer's complaint

Through the analysis of the 28524 malfunction data information given out the main cause of customer complaint, there are three main of which base station take back (including base station exit service and MME derivative base station exits service) accounted for about 94% of the total malfunction.The relationship between the three factors of complaint and the reserving of the base station exit service was analyzed by the malfunction model.Using the established malfunction model,Lower dimensional processing for the data set, and then the standard transformation of the sample matrix elements is normalized matrix Z.

Calculated to obtain the regression equation of various parameters can be concluded that partial regression coefficient absolute value relatively large principal component index expression for 8th main component U8,21th main component U21,22th main component U22.The above model and linear regression are available, the cumulative contribution rate of the interest rate is 85%, and the weighted sum of the main components of m is obtained by the final index value combination,In other words, we have more than 86% of the reason that the withdrawal of the base station is the three main types of customer complaints(Among them,through the model calculation: network coverage and base station take correlation was 88%, but cannot be used with the base station signal back take correlation is 84%, slow speed and base station take correlation was 86%, the three average correlation of 86%).

(3) The relevance and validity of the failure model and prediction model

The 160 complaint data and 160 fault data are selected to check the relevance and validity of the fault model and the forecast model.see Fig. 8.

It is knowable that the performance index value is relatively stable when the equipment is out of service. When the number of station equipment has been returned from 5 to 10 times, the performance index has fluctuated slightly. When the number of retreats of the station equipment is greater than 10 times, the performance index is gradually fluctuated by a large shaft. Therefore, this kind of index belongs to the performance indicator of the central axis, and it can set up a two-threshold stage warning according to the tolerance level of engineering practice.

The central axis wave schematic of the fault based complaint prediction model based on the selected data is shown, see Fig. 9 and Fig. 10.

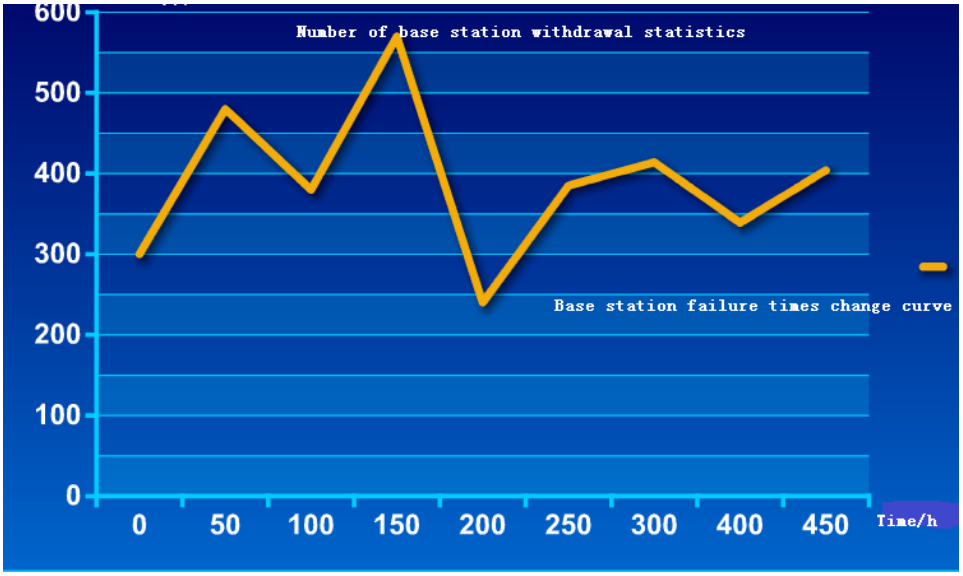

Figure 8. The numerical distribution chart based on the historical data of the number of base station

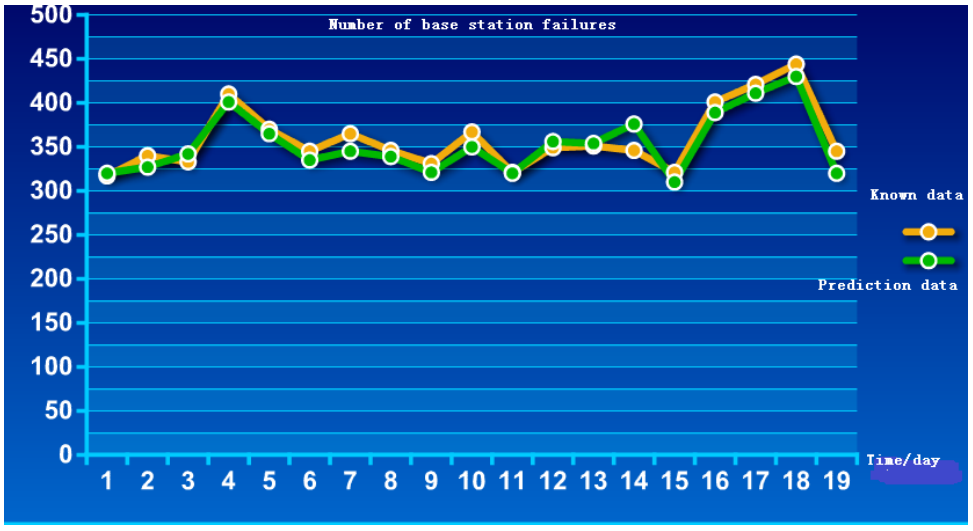

Figure 9. Schematic diagram of the axial fluctuation of the complaint prediction model (short term)

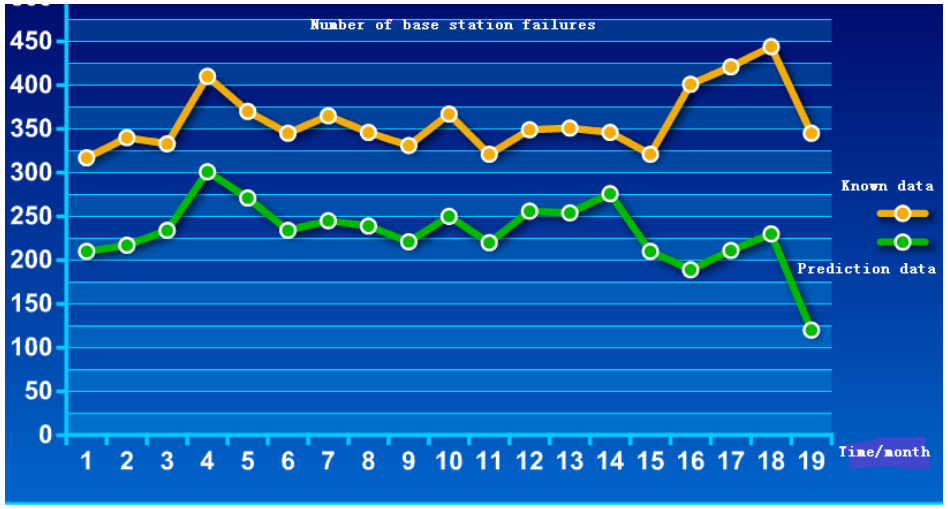

Figure 10. Schematic diagram of the axial fluctuation of the complaint prediction model (medium and long term)

As you can see from the image above, the paper based on the malfunction model based complaint forecasting system can be better suited to the known data in the short term.According to the 160

selected data, in the short term (no more than 30 days), the accuracy rate can be estimated to be about 87 percent. In the medium and long term(No more than 5 years), the forecast accuracy rate can be about 47%.What has been discussed above, the article established complaints prediction system based on malfunction model in the short term forecast is more accurate, But for the medium to long term, the predictions are general.

6. Conclusions and decision making recommendations

By analyzing the data collected from the operator, the model is established,the use of data mining algorithm and association rules algorithm, it is concluded that the base station take back (including MME derivative base station take back) is the main reason for the failure.then,after the linear regression algorithm and the index reduction algorithm were processed, it was concluded that the regressive of the base station (including the MME derived base station withdrawal) was about 94% of the total fault.

The prediction system based on the failure model based on the fault model is predicted to be accurate in the short term, but in the medium to long term, the forecast effect is general.By establishing the model, the article provides data comparison analysis, simplifies the barrier operation and improves the maintenance efficiency of the equipment. At the same time, according to the prediction result, it is reasonable to schedule the relocation maintenance to improve the efficiency, and make the complaint response information in time.

References

[1] Tian HongWei,Lin Quan,Li HongJun, etc. A fault fan power loss estimation method based on correlation analysis of research [J]. Science, technology and engineering, 2016, 16 (11) : 185-188.
[2] Gu HongXun,Yang Ke. The mobile user behavior analysis system based on big data and application cases [J]. Journal of telecom science, 2016, 32 (3) : 139-146.
[3] Zhang YingZhi,Liu JinTong,Shen GuiXiang,etc. Based on the failure correlation analysis of nc machine tools system reliability modeling [J]. Journal of jilin university: engineering science, 2017, 47 (1) : 169-173.
[4] Huang RongShun,Xie ShiNa, Luo Xiao, etc. Based on the regression analysis problem of airport congestion prediction research [J]. Journal of mathematics practice and understanding, 2015 (2) : 89-96.
[5] Zhang XiaoLei,Chen Shan,Ma XiaoLi. Power failure analysis system based on big data research [J]. Power technology, 2016, 40 (11) : 2245-2246.
[6] Yang XiangRong,Wang XiWu,Wang YongXin. Based on the characteristic values of nominal data correlation analysis [J]. Computer and digital engineering, 2016, 44 (5) : 822-824.
[7] Song Zhu,Qin ZhiGuang, Luo JiaQing, etc. The telecom user behavior characteristics in the survey and data analysis [J]. Journal of university of electronic science and technology, 2015 (6) : 934-939.
[8] Sun BingXiang,RuanHaiJun, Xu WenZhong, etc. Based on the static non-cooperative game electric vehicle charging price influence factors of quantitative analysis [J]. Journal of electrotechnics, 2016, 31 (21) : 75-85.
[9] Dong Bin,YanDi,Wang Zheng, etc. The flow calculation of data technology in the application of operator in real-time signal processing [J]. Journal of telecom science, 2015, 31 (10) : 165-171.
[10] Xue LiHong,Liang XiaoJiang. Facing the operators of mobile network traffic safety audit key technology [J]. Journal of telecom science, 2015, 31 (12) : 111-116.
[11] Yang ShiHai,Li Tao, Chen MingMing, etc. The smart grid online fault diagnosis based on data mining and analysis [J]. Journal of electronic design engineering, 2017 (1) : 136-139.
[12] Lu HuaPu,Sun ZhiYuan,Qu WenCong. Traffic flow failure data correction method based on the space-time model [J]. Journal of transportation engineering, 2015, 15 (6) : 92-100.

[13] Cong Wei, Hu MingHua,Wang YanJun. Based on the analysis of historical data controllers communication behavior characteristics [J]. Journal of university of electronic science and technology, 2015 (4) : 617-622.
[14] Tao Yuan,S. Joe Qin.Root cause diagnosis of plant-wide oscillations using Granger causality[J].Journal of Process Control.2013:23-25.
[15] Liu MouHai, Fang Tao,Jiang Yun, etc. Based on the transient frequency component of the correlation analysis of fault line selection method [J]. Power system protection and control, 2016, 44 (2) : 74-79.
[16] Xu Shun. Research and application of fault analysis algorithm for MVB network based on the MVB network. [J]. Computer measurement and control, 2016, 24 (10) : 64-67.