Fault value analysis of on board ATP system using big data

Fault value analysis of on board ATP system using big data

Shen Tao[1], Geng Hongliang[2]

[1] Hunan CRRC Times Signal & Communication Co., LTD Beijing Branch

[2] Hunan CRRC Times Signal & Communication Co., LTD Beijing Branch

[1]freud33@163.com [2]genghl@csrzic.com

Abstract. RFM module is widely used for the classification of customer value. The classification results based on big data are often used as portrait of customer loyalty and customer value evaluation. In this paper, RFM model is introduced innovatively to analyse the fault of on board ATP (Automatic Train Protection) system. Through the fault value classification, it can effectively predict the impact of a certain type of fault on operation. For one thing, the operation departments and equipment providers can pre-process the corresponding failures according to the evaluation results, so as to reduce operational losses. For another thing, it can be used as the basis for dealing with big data of ATP faults.

1. Introduction

The concept of customer value was first put forward in 1980s while the raise of customer value study began in 1990s. Bult and Wansbeek first proposed RFM customer value in 1995. Three elements of customer behavior which are recent consumption: R (Recency), consumption frequency: F (Frequency) and consumption monetary: M (Monetary) consist the core components of the customer's potential value.

EN 50126 defines the concept of risk:

- The probability of occurrence of an event or combination of events leading to a hazard, or the frequency of such occurrences;
- The consequence of the hazard.

EN 50126 also makes a detailed classification of the frequency and severity of dangerous incidents.

Combining RFM value analysis module and risk analysis, this paper redefines RFM module in ATP fault value classification:

- R: Time interval between recent faults. The smaller R is , the higher rate of fault re-happening;
- F: Frequency of fault that happened during a period of time. The greater F is, the higher fault rate is;
- M: Fault cost during a period of time. The greater M is, the higher of fault cost.

2. Clustering and fault classification based on RFM model

There are many kinds of operational failures of ATP, the frequency of failures and the impact on operation are also different. By clustering, similar fault categories can be distinguished from many faults. Clustering algorithms which are commonly used includes but not least K-means clustering, hierarchical clustering and spectral clustering.

There're three assumptions of the failure value of the RFM model based on R, F and M dimensions:

- R: Recent faults are more likely to occur than faults that have not occurred recently;
- F: The probability of high frequency fault is greater than that with low frequency;

- M: When fault at high cost re-happens, it is still a high cost fault.

After clustering, the mean value of RFM of each fault category is compared with the total RFM mean value of all faults to classify the fault types and "↑" is used to indicate a higher value while "↓" is used to indicate a lower value. The mean value of "R" is higher than the total mean indicates that it has been a long period of time for the latest failure happened. The mean value of "F" is lower than the total mean indicates that the frequency of failure is lower than the average failure rate. The mean value of "M" is higher than the total mean indicates the fault cost higher than the average failure. This can be summed up for the following 8 types of faults.

Type 1：R↓F↑M↑, this kind of fault occurs with a short interval, high cost of failure and great impact on operation. It is an urgent problem to be solved. It is a key failure;

Type 2：R↓F↓M↑, this kind of fault occurs at very short intervals and are costly, but occur less frequently. From the perspective of long-term operation and maintenance, such failures should be key failures;

Type 3：R↓F↑M↓, this kind of fault occurs at very short intervals and at high frequencies, but at a fraction of the cost. And if the "M" value continues to rise, such failures will have a huge impact on operations. Therefore, such failures should also be followed as a focus on continuous failures;

Type 4：R↓F↓M↓, this kind of fault occurs at very long intervals, with low frequency and cost, and are occasional minor faults. It can't affect the operation in a short time, and there's no need to treat such faults specially. Therefore, such failures are minor failures;

Type 5：R↑F↑M↑, this kind of fault occurs frequently and the cost of the fault is high, but there are situations where there is a long interval. It is possible that such faults have been eliminated in the near future. Once they occur again, such faults should be focused on and can be classified into key faults;

Type 6：R↑F↑M↓, this type of failure occurs more frequently, but at a lower cost and does not occur again for a long time. Similar to type 3 faults, but with low attention and can be treated as general faults;

Type 7：R↑F↓M↓, this type of fault does not occur for a long time, the frequency of occurrence and the cost of failure are low, which is a minor fault and can be temporarily ignored;

Type 8：R↑F↓M↑, this type of fault occurs when the cost is high, but it does not occur for a long time and the frequency of occurrence is low. The value of the fault is low, which can be used as a general fault handling.

3. RFM-based fault evaluation

The traditional method considers that the weights of RFM indicators are equal to the value. However, recent literature studies found that there is a difference in the contribution of each indicator to value. Liu and Shih used the AHP model to obtain RFM prediction weights, classified them by K-means clustering method, and analysed the value of each category by index weights. Research shows that the weight-based RFM method is very effective.

Assume that the RFM weights are w_R, w_F and w_M respectively, and bring them into Equation (1) to calculate the faulty integrated RFM value.

$$S_I^j = w_R \times S_R^j + w_F \times S_F^j + w_M \times S_M^j \tag{1}$$

Among them, S_I^j refers to the comprehensive RFM value of the fault of j. w_R, w_F, w_M are calculated based on the AHP method, and the weights of R, F, and M are respectively calculated. S_R^j, S_F^j, S_M^j respectively refer to the standardization of the fault R, F, M values. The specific calculation equation (2) is as follows:

$$S_R^j = \frac{R_{max}-R}{R_{max}-R_{min}} \qquad S_F^j = \frac{F-F_{min}}{F_{max}-F_{min}} \qquad S_R^j = \frac{M-M_{min}}{M_{max}-M_{min}} \tag{2}$$

Where R_{max} and R_{min} represent the maximum and minimum values of R, F_{max} and F_{min} represent the maximum and minimum values of F, and M_{max} and M_{min} represent the maximum and minimum values of M. Since the two variables F and M are larger, the impact of the fault is greater, that is, the impact on the fault value is positive, so the forward normalization equation is used. The effect of

R on the value of the fault is negative, that is, the smaller the R, the greater the impact of the fault, so the equation is calculated by reverse normalization.

Finally, the calculated S_I^j is used as the value of the fault impact during the period of the fault, as a measure of the value of the ATP fault.

4. Dynamic RFM model based on Markov chain

In the period of ATP failure, the fault value may change with changes in operational requirements, operating line changes, software updates, hardware updates, and seasonal changes. It is not a static value. Observing the change of the value of the fault is of great significance to fault assessment and operational support.

Statistics show that changes in the value of failure often have greater volatility with software and hardware updates. Assuming that the frequency of ATP software updates is once a year, a dynamic stochastic process probability model is established on the change of failure value in years.

The Markov chain property is a system state transition law. The process is based on the current state and does not depend on the past state, analyses and predicts future development trends, and obtains decision information. The Markov model can be expressed as:

$$X(n) = X(0)\,P^n \tag{3}$$

In the equation, $X(n)$ represents the state probability vector at the moment n; the state probability vector representing the initial moment $X(0)$; and P^n representing the state transition probability matrix. Equation (3) represents on the basis P^n and $X(0)$ the prediction step n. Define the state transition matrix P as Equation (4).

$$P\{X_{n+1} = j / X_n = i,\ X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\} = P\left\{X_{n+1} = \frac{j}{X_n} = i\right\} = P_{ij} \tag{4}$$

Among it, $i_0, i_1, \dots, i_{n-1}, i, j$ represent the state at different time points, and the above transfer matrix equation is expressed as a matrix in the form of:

Table 1. ATP fault value transfer matrix.

	A	B	C	⋯	H
A	P_{AA}^n	P_{AB}^n	P_{AC}^n	⋯	P_{AH}^n
B	P_{BA}^n	P_{BB}^n	P_{BC}^n	⋯	P_{BH}^n
C	P_{CA}^n	P_{CB}^n	P_{CC}^n	⋯	P_{CH}^n
⋮	⋮	⋮	⋮	⋱	⋮
H	P_{HA}^n	P_{HB}^n	P_{HC}^n	⋯	P_{HH}^n

As shown in Table 1, the fault group reaches equilibrium at a specific time t, assuming that at this balanced time point, the fault value group is divided into A, B ... H. The dynamic fault classification model can be defined as a Markov chain: if the fault belongs to the fault group A at time $t = n\text{-}1$, at $t = n$, the probability of belonging to the fault group B is P_{AB}^n, and the probability of belonging to the fault group C is P_{AC}^n... Equation (5) can be derived, that is, each row element of the state transition matrix is 1:

$$P_{AA}^n + P_{AB}^n + P_{AC}^n + \cdots + P_{AH}^n = 1 \tag{5}$$

According to the clustering method in chapter 2, fault groups can be grouped into eight categories. The fault group transition situation in each subsequent year is calculated. When the transition matrix

approximates equilibrium, the value distribution of the future fault group can be predicted by multiplying the obtained transition matrix by the original fault value classification group vector.

5. Conclusion

This paper introduces the RFM model into the ATP fault value classification as the basis for fault processing big data research. K-means clustering method is used to cluster RFM and the fault value classification results are obtained. On this basis, the Markov chain algorithm is applied to the dynamic fault value classification model, so that the ATP operator can effectively predict and evaluate the fault value and pre-process the corresponding fault according to the evaluation result to reduce the operational loss.

References

[1] Railway applications - The specification and demonstration of reliability, availability, maintainability and safety (RAMS). *CENELEC EN 50126 (IEC 62278).* 2002
[2] Yan Kewu, Zhang Lei and Sun Qiang. Application of ID3 Algorithm in Decision Tree Classification in Customer Segmentation of Aviation Market. *Business Research.* 2008
[3] Zhai Jintao and Zhao Wei. Customer Lifetime Value Segmentation and Customer Relationship Strategy[J]. *Journal of Xi'an University of Finance and Economics*. 2005
[4] Liu Chaohua. Research on customer classification model based on customer value. *Doctoral thesis of Huazhong University of Science and Technology.* 2008
[5] D Sculley and Web Scal. K-Means Clustering Proceedings. *19th international conference on World wide web.* 2010
[6] L Kaufman and P Rousseeuw. Finding Groups in Data: an introduction to cluster analysis *Wiley Series in Probability and Mathematical Statistics*. 1990