RGV Faults Prediction Method Based on Unbalanced Data and Multi-algorithm Fusion

Zikang Lai

(College of Economics and Management, Tiangong University, Tianjin 300387, China)

ABSTRACT: Rail Guided Vehicles (RGV) are the key equipments of modern logistics and warehousing system, and once a fault occurs, it may have a significant impact on the whole logistics system, so the fault management of RGV equipment is very important. In this paper, a real-time faults prediction method is proposed for the characteristics of RGV equipment, which adopts the ideas of multivariate feature engineering, imbalanced data processing, and multi-algorithm fusion, to overcome the difficult including factors affecting the faults of RGVs, the high degree of data imbalance, and the accuracy of prediction. Tests based on real-time data show that the algorithm proposed in this question can achieve better comprehensive performance and meet the demand for real-time prediction of RGV equipment faults, and this method can also be used in faults prediction of other equipment, which provides a feasible means for the intelligent maintenance of logistics equipment.

Keywords: Rail guided vehicle; faults prediction; imbalanced data; model fusion; machine learning

2. Research Methodology

2.1. Research Framework

The overall research framework of the RGV fault factor analysis and prediction model based on machine learning algorithms proposed in this paper is shown in **Figure 2**, which mainly includes data preprocessing, feature engineering, dividing the training set and test set, model training, validation, and evaluating the prediction results. The specific operational details will be explained in the following subsections.

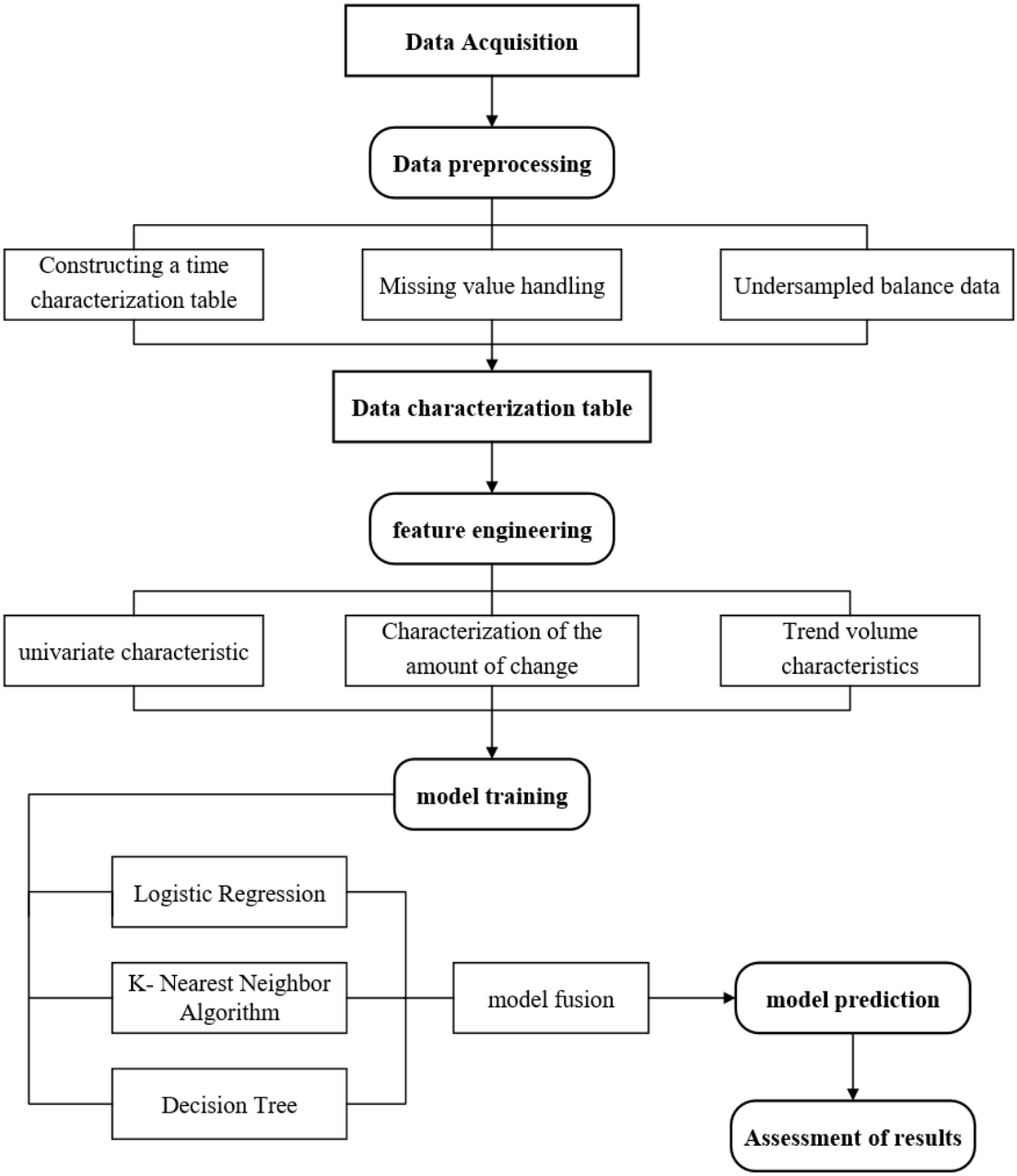

Figure 2. Overall research framework of RGV failure factor analysis and prediction modeling

2.2. Algorithms for Imbalanced Data Processing

For the acquired raw data, this paper organizes them into the form of data that can be used to train the model through data preprocessing. Considering that different detectors have different time points and time intervals for recording the faults of RGV operation and the factors that may affect the fault current and voltage of RGV, this paper divides the time interval from the occurrence time of the first fault to the last fault recorded by the fault information statistics into time points with a uniform time interval of 10 seconds, constructs a time feature table, and uses this table as a benchmark for other feature data

Processing.

For missing value processing, data rows with missing data for all four features were directly removed. Other data with missing data for individual feature labels are filled in by using feature column averages for missing values. Then continue to add fault labels in the table, matching the fault record data in the fault information statistics of the RGV with the time of the time feature table, determining whether the time nodes in the time feature table are in the time period in which the RGV fault occurs, and if they are in then add labels after the corresponding time nodes, labeling the corresponding fault code, and if they are not in then set the label to 0, which represents that there is no RGV fault at this time node, in order to The new time characterization table is obtained.

Since the acquired dataset is highly imbalanced , if this table is used for predictive model construction and testing, it will lead to relatively large result Faults.NearMiss is a widely used undersampling technique[10] , when instances of two different classes are very close to each other, the space between the two classes to be classified can be increased by deleting most of the similar instances, and class distribution can be balanced by randomly eliminating most of the class examples, to improve the accuracy of data classification. to improve the accuracy of data classification. In order to prevent the problem of information loss when deleting part of the faultless data, the Nearest Neighbor method is chosen when using NearMiss undersampling technique, and the specific steps are as follows:

1. Step 1: First calculate the distance between the fault-free data instances of the majority class and the faulty data instances of the minority class, where the fault-free data of the majority class will be undersampled.
2. Step 2: Select n instances of fault-free data of the majority class with the smallest distance from the faulty data of the minority class.
3. Step 3: If there are k instances of RGV fault data in the minority class, k*n instances of fault-free data in the majority class can be obtained using the nearest neighbor method.

2.3. Feature Engineering

Based on the original features such as fault time, voltage, current, temperature and humidity in the feature table of RGV data obtained from data preprocessing, this paper constructs other new features from the perspectives of univariate features, change volume features and trend volume features respectively for model training and testing.

Univariate characteristics of the construction: mechanical power is a physical quantity that indicates the speed of mechanical work, is an important factor to measure the performance of the machinery itself. Considering that the working power of RGV may be related to its faults, this paper takes the two factors of current and voltage as the univariate characteristics of constructing power (power=current*voltage).

Constructing the feature of changing quantity: in the process of RGV working, its detecting values such as voltage and current are changing in real time. Considering that the detected quantities may change significantly when a fault occurs, this paper calculates the rate of change of its features based on the trend of the individual monitoring quantities over time, respectively, and uses this rate of change as a new feature as a feature column for machine learning.

Trend volume feature construction: In order to investigate the relationship between the trend volume features of four variables, namely voltage, current, temperature and humidity, and their occurrence of faults during the operation of RGVs, the coefficients of variation, kurtosis and skewness of the four variables mentioned above are computed in this paper, and they are added as new features in the RGV data feature table.

2.4. Predictive Algorithms

Because of the high real-time requirements, this paper does not use popular deep learning methods, but chooses Logistic Regression, K-Nearest Neighbors Search (KNN) and Decision Tree three algorithms, which have the characteristics of fast running speed and strong robustness, and are more suitable for faults prediction of this kind of scenarios.

2.4.1. Logistic Regression

Logistic Regression is one of the commonly used classification models, mainly used to solve binary classification problems[11] . In Logistic Regression, the mapping function is applied as a sigmoid function with the function formula (1):

$$h(x)=\frac{1}{1+e^{-x}} \qquad (1)$$

According to the sigmoid function the probability that the data sample is classified as 1 (positive sample) can be found as (2):

$$P(y=1|x;\theta)=h_\theta(x;\theta)=\frac{1}{1+e^{-\theta^T x}} \qquad (2)$$

The probability of being 0 (negative sample) is (3):

$$P(y=0|x;\theta)=1-h_\theta(x;\theta)=\frac{e^{-\theta^T x}}{1+e^{-\theta^T x}} \qquad (3)$$

Since Logistic Regression is a probabilistic model, there are many ways to solve for the parameters. The above formulas can be used to solve for the parameters by means of great likelihood estimation, and in practice, the gradient descent method is also commonly used to solve for the parameters in an iterative manner.

2.4.2. K- Nearest Neighbor (KNN) Algorithm

KNN, also known as K- Nearest Neighbor algorithm, mainly classifies different feature values by measuring the distance between them, and its core idea is that in a feature space, when most of the K nearest neighbor samples of a sample belong to a certain same category, the sample will also be classified into that category[12] .

In the KNN algorithm, the distance between individual samples is generally used as an indicator of similarity between samples by calculating the distance between the samples, and there are two calculation methods for distance calculation: the Euclidean distance and the Manhattan distance, and in this paper, we use the Euclidean distance, which is commonly used in the research, for the calculation of (4):

$$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2} \qquad (4)$$

The general flow of the KNN algorithm is:

Calculate the Euclidean distance from the test sample point (that is, the point to be classified) to each of the other sample points;

Sort the above Euclidean distances and select the K points with the smallest distance;

The categories to which these K points belong are compared, and the attribute with the highest frequency to which they belong is returned as the attribute of the test sample point.

2.4.3. Decision Tree Algorithm

Decision Tree is a tree structure classification algorithm, with the continuous improvement of the algorithm, Decision Tree algorithm has a variety of algorithms such as ID3, C5.0, CART and so on. This paper focuses on the binary classification problem of whether the RGV is faulty or not, so the CART algorithm in Decision Tree algorithm is chosen for model construction.

The CART algorithm is a dichotomous recursive segmentation technique, which mainly consists of Decision Tree generation and Decision Tree pruning. Decision Tree generation is based on the training data to generate Decision Tree, and then the validation dataset is used to validate the generated Decision Tree using the minimum loss function as a criterion to form a prediction model. Decision Tree is pruned to form a prediction model[13] .

The CART classification tree uses the Gini index to select the optimal features, and for the binary

$$Gini(p)=\sum_{k=1}^{2}p_k(1-p_k)=2p(1-p) \qquad (5)$$

classification problem, if the probability that a sample point belongs to 1 (positive sample) is p, the Gini index of the probability distribution is (5):

2.4.4. Stacking-based algorithmic fusion model

Stacking is modeling by stacking the original data fitted by the model, it first learns the original data through the first layer of base model, then stacks the output data of the base model shape in columns

and fits the new data through the second layer of model[14] . Integrating multiple models improves the generalization ability of the model as the immunity of a single model is relatively low .[15]

In order to further improve the prediction performance of the model, this paper adopts Stacking to fuse Logistic Regression, KNN Decision Tree model. A 5-fold cross-validation is used on the training set data, and the Logistic Regression and KNN models are used to learn the prediction of the training set data, respectively, and the predicted results are input into the second layer Decision Tree model training as the new feature 1 and the new feature 2, and then the trained second-layer model is applied to test on the test set, to obtain the final fused model prediction results, the whole process is shown in Figure 3.

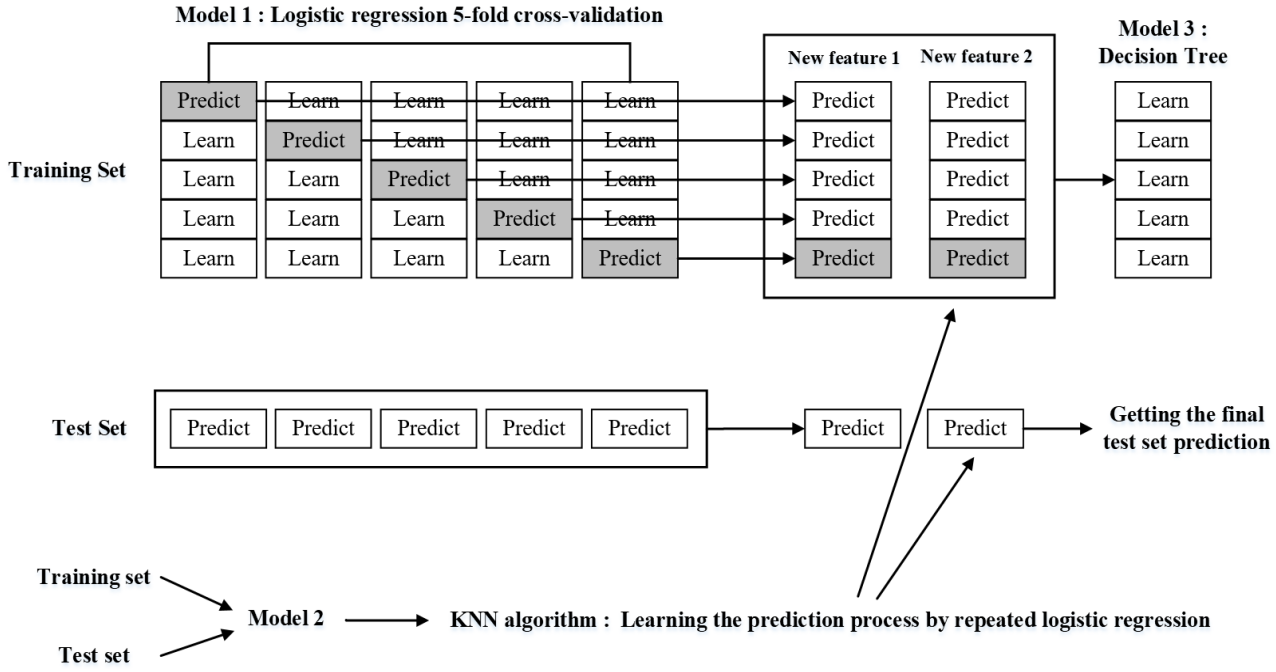

Figure 3. Algorithm fusion model based on Stacking

2.5. Evaluation indicators

In order to be able to objectively evaluate the performance of the RGV prediction model built by applying the machine learning algorithm, this paper selects the commonly used two-dimensional confusion matrix of the classification model evaluation model with Accuracy, Recall, Precision (Precision), F1-score and AUC values to evaluate the prediction model, and the representation of the confusion matrix is shown in **Table 1**:

Table 1 Confusion matrix representation

	Actual malfunction	Practically trouble-free
Faulty prediction	TP	FP
Prediction of no failures	FN	TN

1. Accuracy: indicates the ratio of the number of samples with correct classification results to the total number of test samples, calculated as (6):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \qquad (6)$$

2. Recall: indicates the ratio of the number of samples with faulty RGVs and correct classification results to the total number of samples with faults in the actual RGVs, calculated as (7):

$$\text{Re}\,call = \frac{TP}{TP + FN} \qquad (7)$$

3. Precision: indicates the proportion of the number of samples with faulty RGVs and correct classification results to the total number of samples predicted to have faulty results, calculated as (8):

$$\text{Pr}\,ecision = \frac{TP}{TP + FP} \qquad (8)$$

4. F1-score: is based on the reconciled average of Recall and Precision, calculated as (9):

$$\text{F1-score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \qquad (9)$$

5. AUC value: it is the average performance value for evaluating the binary classification model, the higher value represents the better performance of the model.

3. Experimental Design and Analysis of Results

3.1. Data

As the basis data for prediction are the real-time data from various types of sensors installed in the RGV, including the voltage and current conditions of the main motor as well as its internal temperature and humidity data, etc., in the form of the data shown in **Table 3.**

Table 3. Characterization of selected RGV data

timing	Voltage/V	Current/A	Temperature/°C	Humidity/RH	fault condition
2021-08-25 12:57:40	235.11	0.3847	29.2	56.1	0
2021-08-25 12:58:00	235.11	0.5811	29.3	56.2	0
2021-08-25 12:58:10	235.21	0.4804	29.3	56.1	0
2021-08-25 12:59:10	234.94	0.1788	29.3	56.2	0
2021-08-25 12:59:20	234.62	0.17	29.3	56.1	0
2021-08-25 13:00:00	234.51	0.1771	29.3	56	0
2021-08-25 13:01:30	234.99	0.1778	29.3	56.2	0
2021-08-25 13:01:40	234.77	0.17735	29.3	56.2	0
2021-08-25 13:03:20	234.85	0.1781	29.4	56.2	0
2021-08-25 13:05:10	234.78	0.7023	29.4	56	1

4. Conclusion

RGV is more and more widely used in modern warehousing and logistics, and is the key to the operation of automated three-dimensional warehouses. Once its failure occurs, it will cause extensive paralysis of the storage warehouse, resulting in corporate losses. In this paper, we utilize the original data of RGV's operation to perform data preprocessing operations such as time feature table construction, missing value processing, new feature construction, etc., and construct new feature variables with power as a single feature variable, the amount of change in the measured value, and the amount of trend in the measured value (coefficient of variation, kurtosis, and skewness), respectively, and utilize Logistic Regression, KNN, and Decision Tree algorithms were used to train and test the data to construct the RGV faults prediction model, and the accuracy and AUC values were used as the main evaluation indexes of the model to evaluate the model effect. The results show that the new features constructed with the trend quantities of RGV voltage, current, temperature and humidity, the model of RGV faults prediction constructed by Decision Tree algorithm has the highest prediction accuracy and the best stability of the model, and it can do the warning of RGV faults in advance in the practical application, which reduces the damage of the warehouse paralysis and the loss of the enterprise due to the RGV faults. warehouse

paralysis and enterprise loss due to RGV failures.

Reference

[1] YU Yongjiang,QU Yannan,LIU Pretty. Design of shuttle system and its application in logistics system[J]. Logistics Technology and Application,2007,12(8):86-89.

[2] MIN Dingyong,LIU Qiang,MA Lixin,ZHONG Yanni. Application of high-flow transfer RGV system in large-scale e-commerce warehouse logistics[J]. Logistics Technology and Application,2020,25(12):149-154.

[3] GONG Zhongwei, LI Lu, CHEN Jingsheng, WANG Shanshan, MO Wenjun. Algorithmic study on dynamic scheduling model of RGV with no faults in two processes[J]. Journal of Capital Normal University (Natural Science Edition),2020,41(03):20-23.

[4] Zeng Qingxuan,Feng Qi. Dynamic scheduling strategy of RGV based on computer simulated faults[J]. Engineering Construction and Design,2019,(19):165-166.

[5] CHEN Hua,SUN Qiyuan. Scheduling study of linear reciprocating 2-RGV system based on TS algorithm[J]. Industrial Engineering and Management,2015,20(05):80-88.

[6] Xia Wenhui,Xie Fei. Bayesian decision criterion for fault diagnosis of manufacturing production logistics equipment system[J]. Statistics and Decision Making,2008(08):54-56.

[7] GUO Xi,PU Yun,ZHENG Bin. Reliability analysis of cold chain logistics system based on faulty Bayesian network[J]. Control and Decision Making,2015,30(05):911-916.

[8] ZHANG Yijue,LIU Ye. Simulation and fault analysis of logistics conveyor system[J]. Manufacturing Automation,2022,44(02):1-4+9.

[9] ZHANG Guiqin,ZHANG Yangsen. Intelligent scheduling algorithm for linear reciprocating rail automated guided vehicle[J]. Computer Engineering,2009,35(15):176-178+181.

[10] Jiang JY. Product quality prediction based on PCA ____ NearMiss and XGBoost[J]. Internal Combustion Engines and Accessories,2021,(01):122-123.

[11] Liu Chengxing.Research on the Application of Logistic Regression Model in Risky User Detection in Banking Financial Institutions[J]. FinTech Times,2022(09):71-73.

[12] Gu, W.-H. Stock risk prediction based on improved KNN algorithm[J]. Modern Business,2022,(18):157-160.

[13] QIAO Jian,ZHU Jiahui,YAN Kanghuan. A telecom customer churn prediction model based on random forest CART feature selection improvement algorithm[J]. Telecommunications Engineering Technology and Standardization,2022,35(03):78-82.

[14] Bao Haibo,Wu Yangchen,Zhang Guoying,Li Jiangwei,Guo Xiaoxuan,Lai Jinghua. A net load forecasting method based on feature-weighted Stacking integrated learning[J]. Electric Power Construction,2022,43(09):104-116.

[15] ZHANG Zhan,HAN Hua,CUI Xiaoyu,FAN Yuqiang. Fault diagnosis of refrigeration system based on multiple models integrated learning[J]. Thermal Power Engineering,2020,35(05):153-162.