Learning & Education
ISSN：2251-2802

Research on electric vehicle fault prediction method based on Bagging integrated learning

ChuanZhi GE

Sichuan polytechnic university, Sichuan 634000, China

Abstract: To address the problems of poor classification performance and low fault detection rate of machine models caused by the imbalance of electric vehicle fault data samples, this paper proposes a Bagging integrated electric vehicle fault prediction model with LightBDM as the base learner improvement based on the BS_Bagging-cLightGBM model. First, the training set is resampled using the Borderline_SMOTE method in Bagging integrated learning to improve the degree of data imbalance in the training subset and avoid the missing information of small class samples; then, the weight coefficients and regularization terms are embedded in the loss function of the LightGBM base learner to improve the misclassification cost of small class samples in training; finally, we was evaluated and validated, and the experimental results showed that the BS_Bagging-cLightGBM model outperformed the single model in terms of accuracy, recall and F1-score metrics, and showed better prediction capability. The results of the study can provide an important reference for the repair and maintenance of electric vehicles.

Keywords: Bs_Bagging; CLightGBM; Integrated Learning; Electric Vehicle Fault Prediction; Robustness

1. Theoretical foundation

1.1 LightGBM Application Principle

The LightGBM model achieves efficient and accurate training and prediction by using techniques such as decision trees, gradient boosting, histogram algorithms, and a grow-by-leaf strategy. This allows LightGBM to perform well on large-scale and high-dimensional datasets, making it particularly suitable for scenarios in industry that require processing massive amounts of data. LightGBM also provides a variety of parameter tuning methods and flexible interfaces, enabling users to optimize and customize it according to their needs.

（1）Decision Trees: LightGBM uses decision trees as the basic learner to make predictions by recursively dividing the data into subsets and constructing a tree structure.

（2）Gradient Boosting: LightGBM uses the gradient boosting method for training, optimizing the previous results at each iteration. In each iteration, LightGBM uses the prediction results of the current model to calculate the gradient of the loss function, which is then used as input for the next training.

（3）Bucketing: To improve the efficiency of training and prediction, LightGBM uses a technique called bucketing to compress the feature space. Bucketing discretizes continuous features into a fixed number of buckets, and divides discrete features into different buckets. In this way, the number of features that need to be trained and the computational complexity can be greatly reduced.

1.2 Bagging Integration Learning Principle

Bagging (Bootstrap aggregating) is an Ensemble Learning method that improves the accuracy and stability of a model by building multiple independent models and averaging or majoritying their predictions.

1.3 Borderline_SMOTE Principle

Borderline-SMOTE is a synthetic sampling algorithm for oversampling categories with small sample sizes，for a minority class sampleX_i, Using the K-nearest neighbor method to find the$K$ nearest samples to X_i.The distance calculation formula is shown in equation (1):

$$dist(X,Y)=\sqrt{\sum_{i}^{n}(x_i-y_i)^2} \qquad （1）$$

Which，x_i and y_i is two sample points in the n-dimensional space，$dist(X,Y)$ denotes the Euclidean distance of two sample points.

The algorithm will classify all minority class samples into 3 classes, where if more than half of the K nearest neighbor samples belong to the majority class samples, the minority class samples will be called boundary samples, and since boundary samples tend to be more prone to misclassification, only The synthesis of new samples is performed on the randomly selected boundary samples, as shown in equation (2).

$$X_{new} = X_i + \delta \cdot (\hat{X}_i - X_i) \qquad （2）$$

Which，X_i is a sample from one of the few categories to be treated, $\hat{X}_i$ is a minority class sample in the X_i-neighborhood of K，$\delta \in [0,1]$ is a random number.

2. Model Building

This article proposes an improvement approach for fault prediction in an electric vehicle dataset with imbalanced categories, aiming to achieve better classification and prediction for the fault samples (minority class). The approach includes both data-level improvement using ensemble learning and algorithm-level improvement using individual base learners. A BS_Bagging-cLightGBM electric vehicle fault prediction model is proposed, which integrates the improvements in both ensemble learning and individual base learners. Specifically, the BS_Bagging algorithm is used to obtain multiple base learners by randomly selecting samples from the training set with replacement and combining their prediction results through voting to improve the prediction accuracy of the model. Additionally, the Borderline_SMOTE oversampling algorithm is used to process the minority class samples, allowing the model to better identify and improve the classifier's performance for the minority class samples.

2.1 Improved Bagging Integration Learning based on Borderline_SMOTE

The proposed BS_Bagging-cLightGBM model in this article uses the Bagging ensemble learning method with an improvement based on Borderline_SMOTE. The Bagging algorithm is an ensemble learning method that generates multiple subsets from the original training set by random sampling, trains a base classifier on each subset, and combines the results of all base classifiers through voting or averaging to obtain the final classification result. This method can reduce the variance of the model and improve its generalization ability.

In the Bagging algorithm, we use the BS (Bootstrap) method for random sampling. The BS method is a random sampling method with replacement that ensures that each subset has the same size and that each sample has an equal probability of being selected. In this article, we combine the BS method with the Borderline_SMOTE algorithm to construct the BS_Bagging-cLightGBM model.

2.2 Improving the LightGBM model

To meet the requirement of fast training for high-dimensional and large-scale electric vehicle datasets, this paper selects the LightGBM model as the base learner for fault prediction. As a gradient boosting algorithm based on decision trees, LightGBM algorithm reduces the difficulty of the algorithm by using the histogram algorithm and leaf-wise growth, providing significant speed advantages for large sample classification. For imbalanced datasets, the loss calculation for normal samples is given higher priority in the LightGBM model's loss function, resulting in a bias towards learning normal samples during training while neglecting faulty samples. Therefore, this paper adds class weights to the LightGBM model's loss function, which increases the loss calculation for faulty samples and introduces an L1 regularization term to prevent overfitting.

The loss function in the model reflects the difference between the real and trained classification results, and the smaller the value of the loss function, the better the practice effect. Therefore, the setting of the loss function affects the model training results. In the standard LightGBM model, for a data set with n samples $T = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$， y_i is the training sample x_i， The corresponding operating state (normal or fault), $y_i = \begin{cases} 1, & x_i \text{ It is a sample of failure} \\ 0, & x_i \text{ It is a normal sample} \end{cases}$, The loss function is shown in equation (3)：

$$L(\varphi)=\sum_i l(\hat{y}^{(t)},y^{(t)}) \qquad (3)$$

Which，$\hat{y}_i$ Corresponding to the class prediction of a single decision tree for sample x_i .

To address the issue of imbalance in the dataset and prevent overfitting, this paper introduces a weight coefficient a into the loss function of the LightGBM model, and also adds a regularization term to constrain the complexity of the decision tree. The regularization term includes L1 regularization and L2 regularization, among which L1 regularization can generate a sparse weight matrix, which is beneficial for feature selection and suitable for the car data in this paper. The improved loss function is shown in Equation (4)：

$$L(\varphi)=\sum_i a_i \times l(\hat{y}^{(t)},y^{(t)})+\delta\|\omega\|_1 \qquad (4)$$

Which, a_i is the category weight, the corresponding use terms in the model adjust the weights of each category of data in the parameter search process, emphasizing the cost of misclassification of faulty samples. $\delta\|\omega\|_1$ is the L1 regularization term, feature parameter δ is obtained automatically from the decision tree and regularization parameter ω is selected by random search.

2.3 BS_Bagging-cLightGBM Electric Vehicle Failure Prediction Model

In order to improve the classification performance of the integrated classifier for imbalanced electric vehicle data, the previous two sections of this article respectively proposed a serial optimization approach from the aspects of data diversity, effectiveness, and the performance of individual base learners. At the data level, the Borderline_SMOTE method was used to collect training subsets for multiple base learners, avoiding information loss, improving the degree of class imbalance, and ensuring differentiation between data subsets. At the algorithm level, the class weights were incorporated into the loss function of the LightGBM model to increase the attention to small class samples during model training. Finally, a Bagging ensemble learning fault prediction model was constructed using the cLightGBM model as the base learner and Borderline_SMOTE as the sampling method. The improved model is referred to as the BS_Bagging-cLightGBM model. The technical roadmap for predicting electric vehicle faults based on the BS_Bagging-cLightGBM model is shown in Fig. 1.

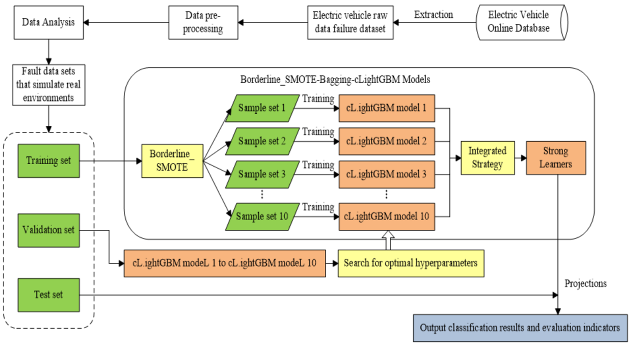

Fig. 1 Model workflow

3. Model Validation and Analysis

3.1 Data pre-processing

The data of this study comes from the online database of pure electric vehicles on the platform of New Energy Vehicle China National Big Data Alliance, and the vehicle operation data are all from all kinds of electric vehicles running in Guangzhou area, and the data collection time is from July to August 2021, and the sampling period of vehicle uploading data platform is 10 s/time, totaling 3690580 items, containing 52 characteristic fields, consisting of vehicle sensor data, fault code information, and location information, and the specific meanings and examples are shown in Table 1.

Table 1 Selected operating data of new energy electric vehicles

Field names	Meaning	Example			
t_volt	Total voltage/V	377.9	377.0	…	342.5
t_current	Total current/A	7.82	14.40	…	12.55
max_cell_volt	Battery maximum individual voltage/V	3.92	3.25	…	3.54
min_cell_volt	Battery minimum individual voltage/V	3.85	3.40	…	3.51
max_temp	Maximum battery temperature value/℃	42	42	…	39
min_temp	Minimum battery temperature value/℃	35	38	…	35
SOC	Battery charge state/%	80	80	…	10
mileage	Accumulated mileage/km	42578	42578	…	42766
max_alarm_lvl	Maximum alarm level for faults	0	0	…	2

3.2 Data Analysis

To analyze the data distribution of feature fields, this article conducted statistical analysis and selected a vehicle dataset with sufficient samples and weak class imbalance. Frequency histograms and correlation diagrams were plotted for the field features including speed, mileage, t_volt, t_current, SOC, isulate_r, max_cell_volt, min_cell_volt, max_temp, min_temp, and max_alarm_lvl. As shown in Fig. 2, the histograms reveal the distribution patterns of feature values in the fault dataset, such as the vehicle speed being concentrated in the slow to medium range, which is consistent with real-world conditions. The distribution of the highest alarm level indicates that the dataset has a significant degree of class imbalance. In addition, most of the field features in the figure have some blue dots falling outside the red lines, indicating that their distributions are not strictly normal.

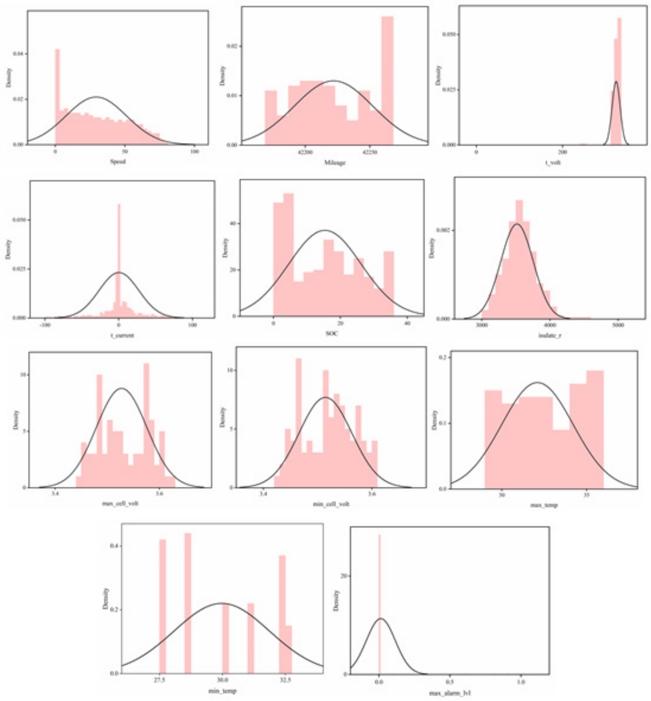

Fig. 2 Distribution of feature data in the dataset

The Spearman correlation coefficient heatmap in Fig. 3 shows that the Spearman correlation coefficient between max_alarm_lvl and speed is only 0.16. This indicates that the data itself contains too little information to explain the correlation, or that there is a class imbalance problem, i.e., there is a huge difference in the number of samples between the classes. Through exploration of the fault dataset, it was found that although there are 58 feature fields in the original dataset, there are more than 10 irrelevant features and 12 valuable features with complete information missing. Only 14 feature fields with research value remain. Therefore, this article proposes automated feature engineering, which aims to automatically construct a large number of features from the dataset to simulate high-dimensional real-world vehicle fault datasets and improve the performance and efficiency of the model training.

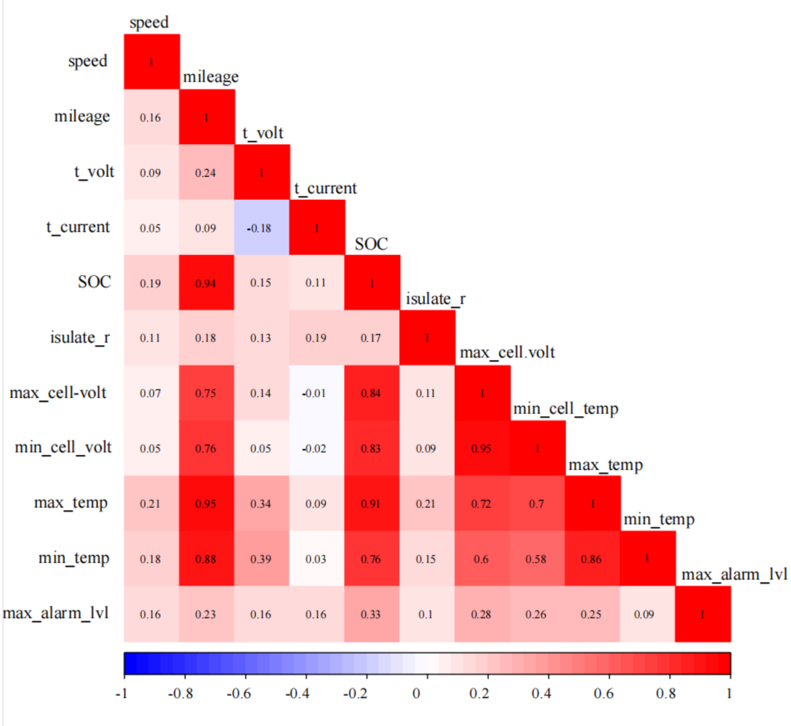

Fig. 3 Spearman correlation coefficient heat map

3.3 Model evaluation indexes

3.3.1 Mixing Matrix

The evaluation metrics are built on a confusion matrix, which is listed in Table 2.

Table 2 Confusion matrix for two classes of problems

	Actual positive category	Actual negative category
Classified as positive category	TP	FP
Classified as negative category	FN	TN

3.3.2 Accuracy

Accuracy (AC) is the ratio of the number of samples correctly predicted as positive classes by the model to the total number of samples predicted as positive classes, and is a measure of the accuracy of the model for positive class prediction. The calculation formula is shown in equation (5).

$$AC=\frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$$

3.3.3 Precision

Precision (Pre) is the percentage of samples for which the model is truly positive among all samples predicted to be positive classes. Precision is applicable to the case of balanced sample categories because it focuses on the recognition accuracy of the model for positive classes and can better reflect the classification performance of the model. It is an index to measure the accuracy of the model for positive class recognition. The formula is shown in equation (6)：

$$Pre=\frac{TP}{TP+FP} \qquad (6)$$

3.3.4 Recall Rate

Recall (RE) is the percentage of samples that are correctly predicted as positive classes by the model out of all samples that are actually positive classes. It is a measure of the model’s ability for positive class identification. The formula is shown in equation (7)：

$$RE = \frac{TP}{TP + FN} \qquad (7)$$

3.3.5 F1-score

F1-score is a metric that considers both precision and recall, and is a measure of the overall performance of the model. The F1-score is the summed average of precision and recall, and the F1-score is in the range of [0, 1], the closer to 1 means the better performance of the model. F1-score takes into account both precision and recall, and can evaluate the classification performance of the model comprehensively, especially for the case of unbalanced sample categories.

3.4 Results and Analysis

After data preprocessing and feature engineering, the dimensionality of the simulated real-world car fault dataset was increased from 14 to 314 dimensions. The training set contained a total of 21,599 samples, including 203 faulty samples, and the test set contained 16,233 samples, including 303 faulty samples and 15,930 non-faulty samples.

The following steps were taken to compare and validate the BS_Bagging-cLightGBM model with the single LightGBM model and cLightGBM model：

（1） Using the same dataset and feature processing methods, the LightGBM model, cLightGBM model, and BS_Bagging-cLightGBM model were trained separately and then used to predict on the test set.

（2） The model's accuracy, precision, recall, F1-score, and other evaluation metrics were calculated, and a comparative analysis was performed.

（3） The training time and prediction time of the models were compared, and the models' operating efficiency was analyzed.

（4） Cross-validation was used to validate the models and compare their generalization ability.

Through the above comparative analysis, the performance and generalization ability of the models can be comprehensively evaluated, and the optimal model can be selected for practical applications. The experimental results are shown in Table 3, indicating that the BS_Bagging-cLightGBM model outperforms the single models LightGBM and cLightGBM in all indicators. The precision, recall, and F1-score of the BS_Bagging-cLightGBM model are 8%, 6%, and 7% higher than those of the LightGBM model, and 5%, 1%, and 3% higher than those of the cLightGBM model, respectively. This suggests that the BS_Bagging-cLightGBM model performs better in predicting electric vehicle faults and can more accurately predict the fault status of electric vehicles.

On the other hand, the Bagging ensemble method can improve the performance and robustness of the model. In this case, using the BS_Bagging-cLightGBM model can obtain more accurate predictions of electric vehicle faults. At the same time, the cLightGBM model also has better performance than the LightGBM model, so it can be preferred in practical applications.

Table 3 Performance comparison of different models

Models	AC	Pre	RE	F1-score
LightGBM	0.90	0.80	0.87	0.83
cLightGBM	0.92	0.83	0.92	0.87
BS_Bagging-cLightGBM	0.94	0.88	0.93	0.90

4. Conclusion and Prospects

4.1 Conclusion

In this paper, a Bagging ensemble learning-based method for electric vehicle fault prediction was studied. Through data analysis and feature engineering, a BS_Bagging-cLightGBM model was constructed, and it was compared with the single models LightGBM and cLightGBM.

According to the experimental results analysis, the BS_Bagging-cLightGBM model performs better in accuracy, recall, and F1-score than the single models, demonstrating superior predictive ability. Furthermore, compared to single models, the BS_Bagging-cLightGBM model exhibits higher stability and robustness, effectively reducing the risk of model overfitting. Additionally, the experimental results also indicate that the cLightGBM model outperforms the LightGBM model in all indicators, suggesting that the LightGBM implemented in C++ has higher operational efficiency and performance advantages, suitable for large-scale data processing and analysis tasks.

In summary, the BS_Bagging-cLightGBM model proposed in this paper has high predictive capability and stability in electric vehicle fault prediction, which is of practical value for improving the accuracy and efficiency of electric vehicle fault prediction.

4.2 Prospects

Although the proposed BS_Bagging-cLightGBM model shows good prediction ability and stability in electric vehicle fault prediction, there are still some directions that need further improvement and exploration：

（1）Data acquisition: The dataset used in this article is relatively small and may not fully reflect the complexity and diversity of electric vehicle faults. Therefore, future work can consider expanding the dataset size and collecting more data samples to improve the model’s generalization ability and prediction performance.

（2）Feature selection: Although multiple feature engineering methods were used in this article, the features were not deeply analyzed and screened. Therefore, in the future, it is possible to combine domain experts’ experience and knowledge to conduct more comprehensive and in-depth feature mining and analysis to improve the model’s prediction ability and efficiency.

（3）Model optimization: The proposed BS_Bagging-cLightGBM model has shown good prediction ability and stability, but there are still some directions for model optimization, such as further adjusting model parameters, increasing regularization constraints, optimizing the loss function, etc., to improve the model’s performance and generalization ability.

In summary, future research can combine the above directions to further improve the accuracy and efficiency of electric vehicle fault prediction to meet practical application needs.

References

[1] Yang Z S. Product failure prediction model based on FTRL and XGBoost algorithm[J].Computer System Applications,2019,28(03):179-184.

[2] BAUDER R A, KHOSHGOFTAAR T M, HASANIN T. Data sampling approaches with severely imbalanced big data for medicare fraud detection[C]//2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI). IEEE, 2018: 137-142.

[3] GUO J, WAN X, LIN H, et al. An active learning method based on mistake sampling for large scale imbalanced classification[C]//2017 International Conference on Service Systems and Service Management. IEEE, 2017: 1-6.

[4] BITEUS J, LINDGREN T. Planning flexible maintenance for heavy trucks using machine learning models, constraint programming, and route optimization[J]. SAE International Journal of Materials and Manufacturing, 2017, 10(3): 306-315.

[5] Zhang T S, Zhi H Y. Automatic fault diagnosis method for asynchronous motors based on improved LightGBM algorithm[J].Automation Applications, 2022(06):29-31+36.

[6] Xiao Q, Mu Y F, Jiao Z P, et al. Online Prediction of Remaining Battery Life for Electric Vehicles Based on Improved LightGBM[J].Journal of Electrical Engineering Technology,2022,37(17):4517-4527.

[7] COSTA C F, NASCIMENTO M A. Ida 2016 industrial challenge: Using machine learning for predicting failures[C]//International Symposium on Intelligent Data Analysis. Springer, Cham, 2016: 381-386.

Fund Projects:

This work was supported by the open fund of Major Hazard Measurement and Control Key Laboratory of Sichuan Province (KFKT-2021-01), the open fund of Sichuan Provincial Key Lab of Process Equipment and Control (GK202207)