焊枪故障预测算法说明

Welding Gun Predictive Maintenance Algorithm Instruction

11 组 焊枪挖掘机队

Team 11 : Mining Machine Welding Gun

Chapter One. Demand Analysis and Data Processing

第 1 节 Task description

The competition tasks are divided into three parts:

A. According to the historical data of the welding gun, predict whether the fault occurs within 10 minutes.

B. According to the historical data of the welding gun, predict whether a specific fault (Error Code E016) occurs within 10 minutes.

C. Based on the welding gun 's historical data, predict the remaining working time within five days.

Based on the task of the competition and characteristics the data provide, we can preliminarily analyze actual needs and research objectives.

第 2 节 Data Characteristics and Demand Analysis

For the first two questions, the score is:

$$F2\_score = 5 * precision * recall / (4 * precision + recall)$$

We can obviously see the importance for recall rate is much higher than the accurate rate, considering the actual production circumstance, a small amount of error alarm can simply be fixed by artificial judgment, but in case of omission behavior, tiny error may drop to a major error, making the whole production shutdown, even causing product quality fluctuation and other serious problems.

Based on the analysis above, the preliminary modeling criteria of this competition can be summarized: pay attention to the coverage of the prediction results,

producing some wrong judgments that judge normal conditions as faults is match better than facing the problem of missed judgments. Try to improve the ability of the model on identify fault data.

In addition, in view of the training set and testing set data analysis, we has drew the variable heat map, characteristics distribution figure, box figure, and so on. We can conclude that this research project can be found several difficult points: the first is the large number of data characteristics and characteristics of weak correlation between features with the target variable distribution, at the same time, the second question is the proportion about positive samples to negative samples is close to 1:5000, which adds many obstacles to modeling. The main objectives of the three questions are in line with each other. The training results of the former question will greatly affect the prediction effect of the next question, so it is necessary to pre-process the data and feature markers to optimize the structure of the data set and the training effect.

第 3 节　Data Processing and Feature Engineering

After a simple understanding and analysis of the data and features, the data and feature annotation can be preprocessed. The feature processing method is mainly studied from the business features and cross features. From the perspective of business feature, it can be readily found that there are obviously correlation between multiple feature variables, such as feature 15:In_Electrode_force and feature 21:out_Electrode_force. The combination of characteristics with obvious correlation can optimize the training method and improve the training results. For the feature groups with weak correlation but have great influence on the labeling results, the cross-feature approach is considered for the study. Furthermore, the multi-granularity statistical feature analysis method can also be used to deeply excavate and simplify the feature performance to achieve better results.

In terms of data, this dataset is characterized by large scale, wide distribution and large value. Methods such as missing value filling, data normalization and pre-division are highly recommended to adopted to conduct pre-processing on the dataset. Normalized data were divided into several blocks arranged in time order to facilitate the use of cross validation method to improve the effect of later training.

For the third question, it is necessary to re-label the data before training to convert the label from the error type to the life expectancy of the equipment.

Meanwhile, due to the requirements of the training target, it is necessary to re-screen the data, discard the data which cannot be labeled near the proclitic date, reconstruct the training set, then carry out further training on the task model.

Chapter Two. Model Architecture

In the research of three questions, XGBoost, LightGBM and CatBoost methods are mainly adopted. Model fusion is also used for training. These boosting methods have their own unique advantages, which are mainly embodied in regularization, cross validation, pruning, high-dimension loss function and other features. The regularization method can reduce the variance of the feature quantity, which makes the learning model more simple and clear, and prevent the occurrence of overfitting. Using cross validation method can cut down iteration times and make it optimized to find the optimal boosting process. The loss function with higher derivative can make error propagation and parameter adjustment more accurately by accelerate the learning effect. By integrating the three models can give play to their respective characteristics and improve the generalization and prediction ability of the model.

In the actual processing, the data has the characteristics of sequence in time, and the similarity of recent data on the timeline is higher than that of distant data. Therefore, for questions 1 and 2, the data of the last five days are firstly used to build the basic model by using the 50% fold cross-training method. At the same time, in order to make full use of the rear residual data and improve the quality of the model, the remaining data are used to train the model separately, and the model is grafted onto the master model with the help of the idea of transfer learning, so as to accelerate the training speed and improve the training effect.

Chapter Three. Optimization of loss function

The loss function reflects the direct difference between the predicted value and the target value, which is the key to configure the learning process. Reasonably selecting the loss function can enhance the expression ability of the model, improve

the training speed and increase the prediction effect. In this task, the training of the first two questions mainly adopts AUC loss function, as a result of the positive and negative samples distributing proportion disparity and dense characteristic value distribution. Using the traditional way of evaluating may prone to errors, while AUC is both suitable for a binary classification problem and also has the characteristics of stability, it can effectively reflect the training effect. For the third problem, the combination of MSE loss and MAE loss is adopted. MAE function can effectively avoid the influence of extreme data, while MSE function can accurately reflect the level of intermediate data. Combining their respective advantages to effectively observe and control the training process.