ISSN: 2664-9640

Research on Quality Prediction Model based on LightGBM

Lili Meng [1, a], Caihua Wang [1, b] and Lei Zheng [1, c]

[1] School of Mechanical Engineering, North China University of science and technology, Tangshan 063000, China;

[a] mll12@163.com, [b] 1921226657@qq.com, [c] 1677654993@qq.com

Abstract

Product quality forecasting is an important part of product quality control. Quality prediction can monitor product quality in real time during the production process, and identify and correct potential factors affecting product quality in advance. The applicability and advantages and disadvantages of several common prediction models are summarized, and a quality prediction model based on the LightGBM algorithm is developed. The XGBoost algorithm is used for feature selection, and features with large influencing factors and low redundancy are selected to build the LightGBM model for training and prediction. After selecting processing parameters, the inspection results of the workpiece can be predicted in advance, thus laying a solid foundation for controlling and improving product quality. Through example validation, the prediction effect is compared with other models, which proves that the prediction model has better effect in handling large data.

Keywords

Quality prediction; LightGBM; XGBoost.

1. Introduction

In manufacturing, monitoring and control of product quality is critical. As digital transformation advances, more and more data is available for analysis and learning, which in turn enables intelligence in important decision-making and control aspects of the manufacturing process. The research of big data analysis technology also provides new ideas for better product quality control. By analyzing the data in the production process through data mining methods to achieve product quality prediction, product quality information can be monitored in real time during the production process, and potential factors affecting product quality can be identified and corrected in advance. This is of great significance for optimizing production and improving production capacity. Therefore, many scholars have applied machine learning methods in manufacturing quality prediction.

Sun, Lin [1] et al. used support vector machines to achieve intelligent quality prediction of production processes in flexible production mode using a multivariate statistical quality control method to reduce product quality fluctuations due to process perturbations. Donghai [2] et al. used the Dragonfly algorithm to select the product quality prediction features, reduce the dimensionality of their quality characteristics dataset, and establish the XGboost model to deal with the imbalance of the complex mechanical product quality dataset. Jiang Lun [3] et al. proposed a soft measurement model for paper quality based on the gradient-enhanced decision tree (GBDT) algorithm, and the validation showed that the model has good generalization ability for measurement accuracy and has high application value. Jinwen Jiang [4] et al. used XGBoost algorithm for manufacturing quality prediction. and compared with other integrated learning algorithms. Mingze Xia [5] et al. used genetic algorithm to search and optimize the parameters in the support vector machine modeling process, and designed and initially

ISSN: 2664-9640

implemented a product quality prediction system based on support vector machine model optimized by genetic algorithm. 강희종,백준걸 [6] et al. proposed a system for classifying defect types based on features and using unsupervised learning methods such as k-means and SOM (Self-organizedmap) for quality prediction. Zhang Yongbo [7] et al. used principal component analysis to select the feature variables and optimize the BP neural network parameters by genetic algorithm proposed a welding quality improvement prediction model based on the fusion of principal component analysis and GA-BP neural network. Yan Shixuan [8] et al. modified the loss function of the model by setting the category weights and L1 regularization term, and the threshold shift method to improve the LightGBM model.

2. Quality prediction methods based on data models

At present, the more commonly used methods of quality prediction are mainly multiple linear regression analysis, neural network (ANN), support vector machine (SVM) and so on. Multiple linear regression analysis is a kind of regression analysis, which finds out the relationship between independent variables and dependent variables from a large amount of data by statistical methods and establishes suitable regression expressions. For linear problems with known quality influencing factors, satisfactory prediction results can be achieved, but multiple linear regression shows some limitations in the face of large-scale complex production data and cannot handle nonlinear problems. Artificial neural networks use the principle of animal neural network activity to construct nonlinear prediction models, which are more adaptable and also capable of handling linear and nonlinear problems. As an emerging intelligent tool, it is widely used in production process quality prediction and diagnosis. However, it also has disadvantages such as large computational effort, easy to fall into local extreme value points, and weak interpolation ability. Support vector machines are based on the Structural Risk Minimization (SRM) criterion, and their topology is determined by support vectors, which has many advantages over other methods in solving problems with limited number of samples, nonlinearity and high dimensionality.

However, traditional machine learning methods are often limited when faced with multivariate, big data problems in industry. In contrast, ensemble learning highlights its great advantages. Integrated learning accomplishes learning tasks by building and combining multiple learning machines. Gradient Boosting Decision Tree (GBDT) [9] is an iteration-based decision tree algorithm in integrated learning. GBDT is robust to outliers, and due to its flexible loss function mechanism makes GBDT can handle any data-driven task with fast processing speed and good results, and the model has good explanatory. As a result, it is increasingly used to solve nonlinear, multiparameter estimation and prediction problems, and has an excellent performance in industry. Extreme Gradient Boosting [10] (XGBoost) is developed based on the traditional GBDT model, and its relative to GBDT, XGBoost generates a second-order Taylor expansion for the loss function, which has support for parallel computing, greatly improves the accuracy and speed of the algorithm, adds a regular term to the loss function LightGBM is a framework for implementing the GBDT algorithm, which is mainly used to solve the problems encountered by GBDT on large-scale data processing. Compared with XGBoot, it has faster training efficiency, lower memory usage, and higher accuracy rate. In addition, LightGBM directly supports category feature processing and has better performance improvement. Therefore, based on the comparison of the advantages and disadvantages of the above machine learning algorithms, the LightGBM algorithm is applied to industrial product quality prediction to mine the information of the data generated in the production process and build a prediction model.

ISSN: 2664-9640

3. Experimental verification

3.1. Data Description

A large amount of data is accumulated in the production process of manufacturing industry, including process parameters (such as equipment processing parameters), quality data of workpieces, quality inspection indexes met by the workpieces, etc. Since in actual production, workpieces produced under the same set of process parameter settings will have multiple quality inspection results, it is important to predict the quality of the produced workpieces by process parameters. This paper uses an open dataset from the CCF Big Data and Computational Intelligence Competition for experimental validation.

The dataset has 6000 data with 10 feature variables, containing three types of data A,B,C. The specific information is shown in Table 1.

Table 1: Data description

Field Types	Field Name	Field Explanation
A	Parameter1	Process Parameters 1
A	Parameter2	Process Parameters 2
A	Parameter3	Process Parameters 3
⋮	⋮	⋮
A	Parameter10	Process Parameters 10
B	Attribute1	Workpiece Properties 1
B	Attribute2	Workpiece Properties 2
B	Attribute3	Workpiece Properties 3
⋮	⋮	⋮
B	Attribute10	Workpiece Properties 10
C	Quality_label	Quality control standards met by the workpiece

3.2. Feature Processing

Due to the large range of values for each feature in the dataset, there is a characteristic of data imbalance. For better analysis and modeling, all data are de-biased as well as normalized, and the common means of de-biasing include squaring and taking logarithms. The transformed continuous and partially discrete features are typically normally distributed. Figures 1 and 2 show the distribution histograms of a typical feature (P1) before and after the open-square transformation.

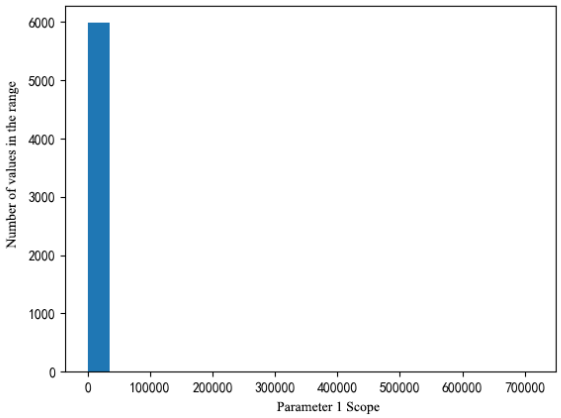

Figure 1: histogram of distribution before square root transformation

ISSN: 2664-9640

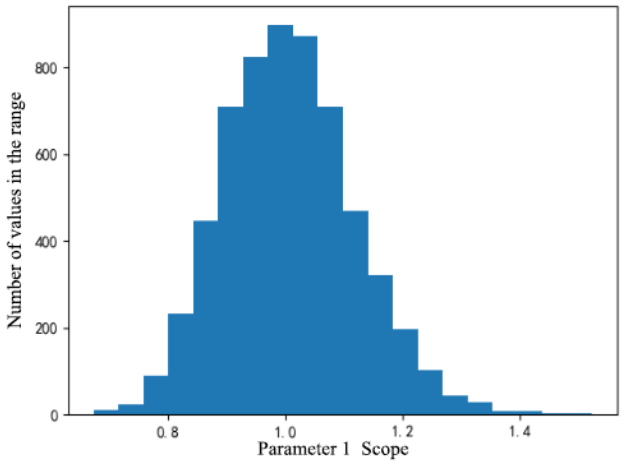

Figure 2 Distribution histogram after square root transformation

The degree of influence of each feature variable relative to quality was analyzed using the XGBoost algorithm, and the results are shown in Figure 3. It can be seen that Attribute4, Parameter2, Attribute6 and other feature variables have a relatively significant impact on the prediction of product quality. The top 15 features with a high degree of influence on product quality were selected for modeling

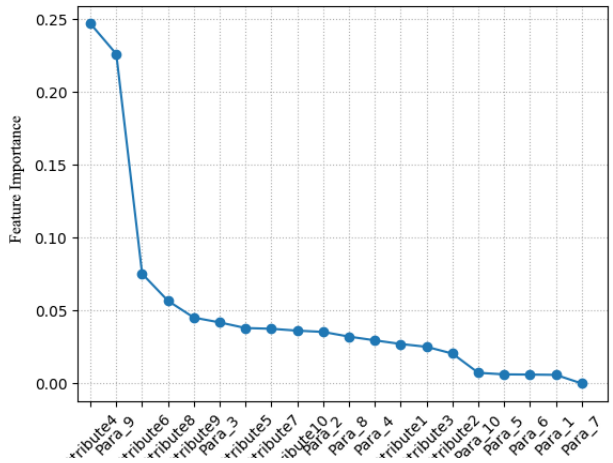

Figure 3 importance of characteristic variables

3.3. Construction of the model

For the feature-selected dataset, it is divided into two mutually exclusive training set and test set, with 80% of the training set and 20% of the test set. The training set is used to train the model parameters and the test set is used to test the accuracy of the trained model, and stratified random sampling is used to reduce the sampling error. The quality prediction model is built based on the LightGBM algorithm for the feature-engineered training set and compared with the prediction results of support vector machines, neural networks, XGBoost, and GBDT. The model input data is divided into process parameter data and workpiece attribute data, and the output is the quality inspection standard that the workpiece conforms to.

For LightGBM model parameters, the main parameters are learning_rate, num_iterations, max_depth,num_leaves,min_data_in_leaf,min_sum_in_leaf, and min_sum_in_leaf. hessian_in_leaf, feature_fraction, bagging_fraction, bagging_freq, reg_alpha,reg_lambda. The final determination of the optimal parameters of the model using the grid search method is shown in Table 2.

ISSN: 2664-9640

Table 2 Data description

Model Parameters	Parameter values
learning_rate	0.1
num_iterations	708
max_depth	8
num_leaves	50
min_data_in_leaf	19
feature_fraction	0.9
bagging_fraction	1
bagging_freq	5
reg_alpha	0.03
reg_lambda	0.01

3.4. Experimental results

The accuracy and model processing time of each model for product quality prediction are shown in Figures 4 and 5. It can be seen from the figures that the traditional machine learning methods are less effective in prediction when dealing with scenarios with large data volumes, and neural networks have higher prediction accuracy compared to support vector machines, but have more processing time. Relatively speaking, the integrated learning algorithm has better results in terms of prediction accuracy and processing time. In terms of processing time, LightBGM algorithm has a significant improvement compared with GBDT and XGBoost.

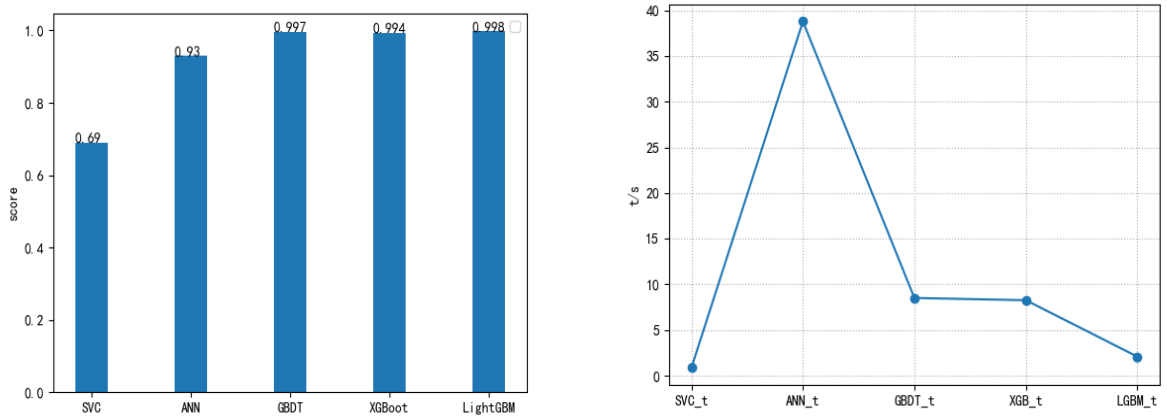

Figure 4 Comparison of model prediction scores and times

4. Conclusion

In this paper, a quality prediction model based on the public dataset of CCF Big Data and Computational Intelligence Competition is established by LightGBM algorithm, and feature selection and special importance analysis are performed by XGBoost algorithm. And the prediction results are compared with neural networks, support vector machines, GDBT and XGBoost models. It is demonstrated that the LightGBM algorithm shows greater advantages of the model in terms of processing time and prediction accuracy and produces more accurate prediction results when dealing with work tasks with large data. This is of great practical significance for improving product quality.

ISSN: 2664-9640

Acknowledgements

Municipal self-project (research on quality control system based on big data and internet of things, 19140201F).

References

[1] Sun, Lin, Yang, S. Yuan. SVM-based intelligent prediction of production process quality in flexible production model[J]. Systems Engineering Theory and Practice,2009,29(06):139-146.
[2] Dong, H., Tian, S.. Product quality prediction of complex machinery based on DA-XGboost algorithm[J]. Combined machine tools and automated machining technology,2021(03):53-56.
[3] Jiang L, Man Yi, Li Jigeng, Hong Mona, Meng Zwei, Zhu Xiaolin. Soft measurement model of paper quality based on gradient-enhanced decision tree algorithm[J]. China Paper, 2020,39(05):37-42.
[4] Jiang JW,Liu WG.Application of XGBoost algorithm in manufacturing quality prediction[J]. Intelligent Computers and Applications,2017,7(06):58-60.
[5] Xia M.Z.. Research on product quality prediction system based on improved support vector machine[D]. General Research Institute of Mechanical Sciences,2020.
[6] 강희종,백준걸. Improved Quality Prediction Method by Clustering Data in Semiconductor Manufacturing Process[J]. Journal of the Korean Institute of Industrial Engineers,2020,46(2).
[7] Zhang YB, Zhu YT. Strip steel welding quality prediction based on principal component analysis and GA-BP neural network[J]. Thermal Processing Technology,2020,49(17):128-132. doi:10.14158/j.cnki.1001-3814.20183987.
[8] Yan Shixuan,Zhu Ping,Liu Zhao. Research on automotive fault prediction method based on improved LightGBM model[J]. Automotive Engineering,2020,42(06):815-819+825.
[9] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. 2016.
[10] Friedman J H. Greedy Function Approximation: a Gradient Boosting.