Modeling and Analysis Based on Industrial Production Data

Kaihan Pang, Ze Wang

***Abstract*— With the continuous development of industrial automation and intelligent technology, the analysis and modeling of industrial production data has become a key means to improve production efficiency and quality. This study explored how to select appropriate models for data processing and optimization in actual production environments by modeling and analyzing the production data of a manufacturing enterprise. Combining relevant domestic and foreign literature, we proposed a comprehensive modeling framework based on machine learning and statistical methods, and designed a highly targeted processing flow, providing a replicable solution for similar enterprises.**

***Index Terms*—Industrial production data, modeling analysis, machine learning, data processing, production optimizationt.**

I. INTRODUCTION

With the advancement of the fourth industrial revolution, the manufacturing industry is gradually moving towards a stage of deep integration of intelligence, digitization and networking. A large amount of real-time data is generated in the industrial production process, including equipment operation data, process parameters, environmental monitoring data, personnel operation records, etc. These data not only reflect the operating status of the production system, but also contain a large amount of potential information that can be used to improve efficiency, reduce costs and ensure quality. Therefore, how to effectively model and analyze these massive, complex, multi-source and heterogeneous industrial production data has become one of the key paths to achieve intelligent manufacturing.

Traditional production management methods mainly rely on manual experience and static statistical methods, which are difficult to cope with the changeable, complex and dynamic operating environment in modern manufacturing systems. In this context, data-driven modeling and analysis methods have emerged and gradually become a research hotspot. Industrial data modeling can not only be used for optimal scheduling of production processes, quality prediction and control, energy consumption analysis, etc., but also assist in the construction of decision support systems and improve the overall operational efficiency of enterprises.

At present, a large number of studies at home and abroad have applied machine learning, data mining and statistical modeling methods to industrial data analysis. For example, linear regression models are often used to evaluate the impact of key production factors on production capacity. Support vector machines and random forests have good performance in scenarios such as fault detection and product quality prediction. With the improvement of computing power, deep neural networks have shown higher flexibility and accuracy in multivariate modeling and nonlinear relationship modeling.

However, industrial data analysis faces some practical challenges: on the one hand, the quality of data varies, and there are problems such as missing, anomaly, and noise. On the other hand, the relationship between data is complex and changeable, and a single model is often difficult to fully reflect the laws of the production system. Therefore, establishing a systematic and generalizable data modeling and analysis process combined with multiple modeling methods is an urgent need in the current industrial scenario.

Based on this, this study is based on the production data of a manufacturing enterprise, combined with actual production needs and data characteristics, and proposes a complete modeling and analysis process covering data preprocessing, feature engineering, model selection and fusion, model evaluation and optimization. The research goal of this paper is to provide industrial enterprises with a practical and general modeling idea and method framework to help them achieve data-driven intelligent transformation in production optimization, quality control, equipment predictive maintenance.

II. RELATED WORKS

In the process of industrial production and manufacturing, data-driven intelligent analysis methods are gradually replacing traditional empirical decision-making methods, and classification models play an important role in quality inspection, fault diagnosis, and production process optimization. Industrial data usually includes static features (such as equipment parameters and process conditions) and time series features (such as trends over time such as temperature, pressure, and vibration). Classification models can be used to deeply explore the inherent patterns of these data and identify and predict different types of industrial events. The application of classification models not only improves production efficiency, but also significantly reduces losses caused by quality problems or equipment failures.

Salau[1] et al. proposed a network traffic classification method based on software defined network (SDN). Combining multiple machine learning algorithms, the method achieved efficient DNS, Telnet, Ping and Voice traffic classification by extracting and analyzing traffic features, and achieved superior performance in real-time and offline network traffic classification. Shan[2] proposed a

Manuscript received April 25, 2025

Kaihan Pang, Software engineering, Tianjin Polytechnic University, China, Tianjin

Ze Wang, School of Software, Tiangong University, Tianjin, China

random feature mapping method based on AdaBoost algorithm and result fusion, combined with sparse autoencoder and feature subset generation, and improved the classification performance on twenty classic data sets by enhancing classifier stability and weighted probability selection method. The KNN classifier improved the classification accuracy by more than 20% on the Balance data set. Xu[3] et al. proposed a personalized recommendation system combining BERT model and nearest neighbor algorithm. By deeply understanding the semantic information of product titles, the matching ability between products and users is enhanced, and the classification accuracy of product recommendations is effectively improved on the eBay e-commerce platform. Fan[4] proposed a concrete vibration quality assessment method based on machine learning, combining a multi-parameter data acquisition system with an improved classification algorithm. By accurately classifying the concrete vibration quality level, it achieved a recognition accuracy of 93.75% in the visualization of the concrete vibration process and the implementation of autonomous robot vibration. Zhang[5] proposed a credit risk assessment modeling framework based on multiple classifiers, combining traditional classifiers with intelligent classifiers, and achieved significant results in credit risk assessment through the fusion of classification algorithms and data features. Hoque[6] et al. proposed a breast cancer classification model based on XGBoost, combining feature importance evaluation with an improved gradient boosting algorithm, and achieved a classification accuracy of 94.74% in breast cancer detection by effectively processing breast cell image features..

In practical applications, classification models provide effective technical support for improving the level of intelligent manufacturing. Classification models can determine whether tool quality meets the standards and analyze key features through relevant methods, thereby optimizing the workpiece process and improving product consistency. They can also detect process parameters that lead to defects early and provide data support for production process optimization [8].

In summary, the application prospects of classification models in industrial data analysis are broad. Through effective classification and combined with gain value, SHAP and other methods for feature analysis, not only can the prediction performance of the model be improved, but also strong support can be provided for the optimization of industrial processes and fault warning. This study is committed to improving the accuracy and efficiency of industrial data analysis by combining classification models and feature importance analysis, and contributing to the development of intelligent manufacturing and Industry 4.0.

III. Model selection and process design

In the process of industrial production data modeling and analysis, model selection is a crucial step. Since production data has time series, multi-dimensionality and complex nonlinear relationships, a single modeling method is often difficult to fully capture the inherent laws of the data. Therefore, this study combines a variety of machine learning and statistical methods in model selection, aiming to improve prediction accuracy and stability through the idea of ensemble learning.

First, considering the characteristics of industrial data, this study selected four common and mature modeling methods: regression analysis, support vector machine (SVM), random forest, and neural network. Regression analysis was selected as the benchmark model, mainly used to study the linear relationship between key factors and production efficiency in the production process. In industrial data, many production factors show a clear linear relationship with production efficiency. Regression analysis can quickly and effectively capture these relationships, thus providing guidance for preliminary analysis. However, due to the strong nonlinear characteristics that often exist in industrial data, a single regression model does not perform well when dealing with complex problems. Therefore, we subsequently introduced two nonlinear modeling methods, support vector machine (SVM) and random forest.

Support vector machine (SVM) is a powerful classification and regression method that can effectively process high-dimensional and nonlinear data. In industrial data, the relationship between many factors that affect production is not linear, but presents a more complex pattern. SVM maps the data by constructing a high-dimensional feature space, so as to better capture these complex nonlinear relationships. On this basis, random forest, as an integrated learning method, can effectively reduce the overfitting risk of a single model and improve the generalization ability of the model through the training and voting mechanism of multiple decision trees. Random forest can not only process large-scale data sets, but also has good robustness and strong feature selection capabilities. It shows excellent performance when facing complex industrial data.

However, with the rise of deep learning, neural networks, especially deep neural networks (DNNs), have become increasingly prominent in time series data modeling and nonlinear pattern recognition. In industrial production, the complexity and high dimensionality of many data make it difficult for traditional methods to achieve the maximum effect, while deep neural networks can automatically extract high-level features of data through multi-level nonlinear transformations, thereby providing strong support for modeling complex production processes. Therefore, in the final stage of model design, we use neural networks as a deep optimization model to further improve the accuracy and stability of predictions.

In the entire modeling process, data preprocessing is an indispensable link, especially for missing values, outliers and noise data that are common in industrial data. In order to ensure the accuracy and stability of the model, we have comprehensively cleaned and processed the data. First, the missing values in the data are filled by interpolation and mean filling to avoid modeling deviation caused by data loss. Secondly, for the detected outliers, the Z-score method and box plot method based on statistical analysis are used for outlier detection to ensure that the model will not be affected by extreme data. Then, the data is normalized by standardization and normalization to ensure that the scale differences between different features will not have too much impact on the model. Finally, in order to further improve the

efficiency of the model and reduce the computational complexity, we use dimensionality reduction methods such as principal component analysis (PCA) to remove redundant features, so that the model can converge faster and avoid the "dimensionality disaster" caused by high-dimensional data.

The design of model selection and data processing flow is not only the application of a single algorithm, but also the integration of a comprehensive method. In the process of data processing, a series of operations such as cleaning, normalization and feature selection are used to ensure the quality and availability of data; in terms of model selection, a comprehensive modeling framework with strong adaptability and robustness is constructed to meet the needs of different problems through the combination of regression analysis, support vector machine, random forest and neural network. Such design ideas and methods enable the entire analysis framework to more flexibly cope with various complex situations in industrial production, and provide reliable prediction support for subsequent production optimization, quality control and other issues.

IV. Conclusion

Aiming at the modeling and analysis of industrial production data, this study proposed a comprehensive set of data preprocessing, feature engineering and modeling processes, combining multiple modeling methods such as regression analysis, support vector machine, random forest and deep neural network, and successfully applied them to production process optimization and quality control. Through comprehensive cleaning and preprocessing of data, the interference of missing values, outliers and noise on the modeling results is eliminated, providing high-quality input data for subsequent model training. At the same time, in terms of model selection, considering the diversity and complexity of industrial data, this study adopted the ideas of ensemble learning and deep learning, so that the modeling results can more accurately reflect the underlying laws of the data.

Overall, this study provides an effective process and method framework for data modeling and analysis in industrial production, which has strong practicality and promotion value. With the continuous advancement of data acquisition technology and the improvement of computing power, data-driven production optimization and quality control will be more intelligent and automated in the future. The results of this study lay the foundation for the development of intelligent manufacturing and industrial Internet, and provide reference and reference for other industrial enterprises.

References

[1] Salau A O, Beyene M M. Software defined networking based network traffic classification using machine learning techniques[J]. Scientific Reports, 2024, 14(1): 20060.
[2] Shan W, Li D, Liu S, et al. A random feature mapping method based on the AdaBoost algorithm and results fusion for enhancing classification performance[J]. Expert Systems with Applications, 2024, 256: 124902.
[3] Xu K, Zhou H, Zheng H, et al. Intelligent Classification and Personalized Recommendation of E-commerce Products Based on Machine Learning[J]. arXiv preprint arXiv:2403.19345, 2024.
[4] Fan S, He T, Li W, et al. Machine learning-based classification of quality grades for concrete vibration behaviour[J]. Automation in Construction, 2024, 167: 105694.
[5] Zhang X, Yu L. Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods[J]. Expert Systems with Applications, 2024, 237: 121484.
[6] Hoque R, Das S, Hoque M, et al. Breast Cancer Classification using XGBoost[J]. World Journal of Advanced Research and Reviews, 2024, 21(2): 1985-1994.
[7] Li C, Chen Y, Shang Y. A review of industrial big data for decision making in intelligent manufacturing[J]. Engineering Science and Technology, an International Journal, 2022, 29: 101021.
[8] Zhou Y, Yi H, Yue X, et al. Method for loose bolt positioning and prediction of bolt axial force in bolt group[J]. Measurement, 2024, 227: 114316.