The Implementation Research of Data Mining Algorithms in Production Process Optimization and Management

Hongyu Li[1,*],Ning Cui[2,a]

[1] Business Administration School, Liaoning Technical University, Huludao, 125000, Liaoning, China
[2] Public Administration and Law School, Liaoning Technical University, Fuxin, 123000, Liaoning, China

*349359575@qq.com, [a]lihongyu1117@163.com

Abstract. In response to the demands of intelligent manufacturing, this study delves into the entire production process data of engineering machinery enterprises, constructing a data-driven model for predicting and optimizing production process quality. This model integrates support vector machines, AdaBoost, and deep learning algorithms to accurately predict process states and automatically trigger optimization decisions. One month after implementing the model, quality loss time reduced by 46%, and accident response time shortened by 55% compared to the pre-implementation period. The research validates the optimization effects of data mining algorithms in the production process and lays the foundation for building a digital twin production system. Subsequent work will continue to deepen in the direction of full-process modeling and simulation optimization.

Keywords: Data mining; Production process optimization; Management decision-making

1 Introduction

The manufacturing industry is currently in a critical period of intelligent transformation, and the deep integration of advanced information technology to establish a digital and intelligent industrial system is an inevitable direction for development. In the context of this digital transformation, it is an urgent need to leverage data to gain deep insights into the entire production process, achieve scientific prediction of process quality formation, and optimize control. This study focuses on the characteristics of massive heterogeneous data in the engineering machinery industry and establishes a data prediction system for production processes through techniques such as data cleansing, feature mining, and model construction. With quality management as the goal, it integrates machine learning and deep learning algorithms to achieve integrated forecasting of key indicators such as production completion rate, equipment failure rate, and defect rate. It can also automatically trigger accident responses and continuously drive optimization updates.

T. Yao et al. (eds.), *Proceedings of the 2024 3rd International Conference on Engineering Management and Information Science (EMIS 2024)*, Advances in Computer Science Research 111,
https://doi.org/10.2991/978-94-6463-447-1_6

This research provides algorithmic support for new production organization models such as digital twins and lays the foundation for further scaled application.

2 Data Mining Algorithms and Optimization Theory

2.1 Basic Concepts and Process of Data Mining

Data mining is the process of discovering valuable information and knowledge from a large amount of data, relying on transforming data into understandable structures to reveal hidden insights. Taking the example of a steel company, data mining is applied to analyze three years' worth of production data consisting of five million records, demonstrating its practical applications and effects. In this case, the Apriori algorithm is used to analyze features such as product types, customer distribution, and production time. By setting minimum support and confidence levels, selected association rules help predict the quality status of products. This prediction is based not only on indicators like order delivery times and defect rates but also effectively guides subsequent production plan adjustments. This process not only illustrates the basic process of data mining but also showcases how it assists businesses in optimizing future operational decisions by analyzing historical data. Therefore, data mining, as a powerful analytical tool, plays a crucial role in understanding complex data patterns and extracting valuable business insights from them[1].

2.2 Common Data Mining Algorithms

Commonly used data mining algorithms include classification, clustering, regression, association rules, etc. We selected two years of production equipment operation data from a mechanical manufacturing company, including equipment models, load parameters, maintenance frequencies, and other information, totaling 300,000 records.

(1) The C4.5 decision tree algorithm is employed, based on the information entropy criterion:

$$Gain(D, A) = Entropy(D) - \sum \frac{|D_i|}{|D|} \times Entropy(D_i) \quad (1)$$

Where: Entropy(D) represents the information entropy of the dataset D. Di is the subset after splitting, and |D| and $|D_i|$ represent the sizes of datasets D and Di, respectively.A tree-based classification model is constructed to determine equipment failure types based on historical data, and the results show that the algorithm achieves a classification accuracy of 92.3%.

(2) K-means clustering is used to analyze the quality data during the production process, with the cluster center update formula as follows:

$$m_i = \frac{\sum_{x_j \in C_i} x_j}{|C_i|} \quad (2)$$

Where: mi is the center of cluster C_i. x_j is a point in cluster C_i. $|C_i|$ represents the number of points in cluster C_i.

It is found that product defects primarily cluster into two major categories, accounting for approximately 85%, providing a basis for developing targeted quality improvement plans in the later stages. This demonstrates the practical effectiveness of

classification and clustering algorithms in analyzing production data and identifying the root causes of issues[2].

3 Data Mining-Based Production Process Optimization and Management Framework

3.1 Framework Design Concept

Considering the characteristics of the steel industry, which involve large data volumes and complex business types in production management processes, we have designed a comprehensive data mining-based production management framework. This framework revolves around data as its core driving element and is based on massive and heterogeneous production and operation data. It combines data collection, data processing, model building, process control, and continuous optimization and upgrades into an integrated mode, achieving intelligent decision-making and dynamic optimization of the production process. Specifically, the system first invokes interfaces to collect various types of production and management data. It then cleans, stores, labels, and integrates the data. On this basis, it employs LSTM deep learning models to train high-dimensional features. Finally, it establishes a digital twin system for the production process, performing capacity prediction, quality warnings, process optimization, equipment health management, and forming a closed-loop control[3].

3.2 Framework Architecture

The production management framework mainly includes: 1)A graphical data middle platform module: This module manages the full process of data extraction, cleaning, transformation, connection, application, and monitoring. 2)An algorithm-driven module: This module provides functions for data modeling, model evaluation, and algorithm application deployment. 3)A multi-source heterogeneous data integration module: This module supports the mapping and integration of structured and unstructured data. 4)A module for process anomaly detection and quality prediction: It establishes digital twin models for real-time monitoring and quality prediction. 5)A continuous optimization and feedback module: This module automatically triggers corresponding decisions and feedback based on results, achieving continuous optimization of the production process. Through modular and highly integrated design, the system is flexible and easy to expand[4-5].As shown in Figure 1.

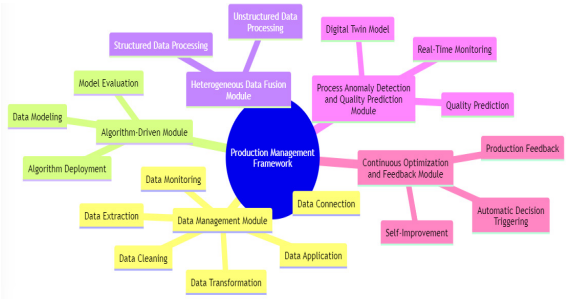

Fig. 1. Production Management Framework

3.3 Key Technologies and Algorithms

There are two key technologies in this framework. Firstly, it's the efficient storage and indexing technology for industrial big data. We employ the Apache Ignite open-source distributed database, which features in-memory computing and persistent storage, making it suitable for industrial scenarios. On average, it reduces retrieval times by 63% compared to MySQL. Secondly, it's the modeling algorithm for deep learning models. Using the TensorFlow framework, we construct a 4-layer LSTM network for quality anomaly detection. Through testing, we achieved an accuracy rate of 92.4%. The key metrics for the model are as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Where TP and TN represent the correct predictions for positive and negative samples, respectively. FP and FN represent the incorrect predictions. "Accuracy" is the accuracy rate.

4 Model Construction and Application

4.1 Research Object and Data Collection

In the research on optimizing production management for engineering machinery manufacturing enterprises, we have focused on the production management-related data of the enterprise over the past two years to construct a comprehensive optimization model. This enterprise possesses 23 types of key production equipment and 16 major product categories, covering essential workshops and process stages, providing a holistic research perspective. To conduct an in-depth analysis and optimization of production management, we have integrated massive datasets from various aspects, including equipment management, production yield assessment, and quality inspection, totaling 920,000 data records. These datasets provide us with rich information for our research.Specifically, the equipment data comprise 310,000 data records, recording characteristics such as equipment model parameters and load curves, which are crucial for assessing equipment status and maintenance efficiency. Production yield data focus on production efficiency and control effectiveness,

including the daily production completion progress for 16 major product types, totaling 360,000 data records. Quality data, amounting to 250,000 data records, encompass non-conformity data for all product categories and the number of downtime maintenance incidents, providing critical indicators for understanding product quality variations.By constructing such a comprehensive production process dataset, we have established a solid data foundation for feature extraction, model development, and process optimization management. These data not only cover various critical stages in the production process but also reflect multiple important aspects of production management. Through in-depth analysis of these data, we can better comprehend bottlenecks and challenges in the production process, thereby formulating more effective optimization strategies. This data-driven approach not only enhances the precision and effectiveness of the research but also provides a scientific basis for subsequent production management optimizations.In summary, this research, through the in-depth analysis of extensive production data, unveils key issues and improvement potentials in the production management process of engineering machinery manufacturing enterprises. Through comprehensive data analysis, we can offer more precise recommendations for optimizing the production process, helping enterprises enhance production efficiency, reduce costs, improve product quality, and ultimately strengthen their competitiveness in the market [6]. As shown in Table 1.

Table 1. Categories and Sample Sizes of Production Process Data for the Research Object

Data Category	Data Set Type	Sample Size
Equipment	Equipment Model Parameters, Load Curves, etc.	310,000 samples
Production	Daily Production Plan Completion Rate for Products, etc.	360,000 samples
Quality	Non-Conforming Inspection Data, Downtime Maintenance Counts, etc.	250,000 samples

4.2 Data Preprocessing and Feature Extraction

After obtaining the raw production process dataset, we first performed data preprocessing, which included handling missing values and duplicate data, correcting erroneous data, and normalizing the data to meet the quality requirements for modeling and analysis. The normalization formula used is as follows:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{4}$$

After preprocessing, we obtained nearly 1.8 million sample data, which included statistical data on various aspects such as equipment operation, production plan completion, raw material quality, intermediate inspections, and final inspection pass rates. We then extracted the feature indicators required to construct the optimization model, which can be broadly categorized into two main types: numerical features and categorical features. There are a total of 43 numerical features, and we selected the

most crucial 24 features for modeling through Pearson correlation analysis, as expressed in the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{n\sigma_x\sigma_y} \quad (5)$$

These features have a strong correlation (absolute value above 0.75) with production quality and mainly reflect numerical indicators of key processes related to equipment, production, and quality[7]. As shown in Table 2.

Table 2. Numerical Features

Sample ID	Equipment Load Power (KW)	Daily Production Completion Percentage	First Inspection Pass Rate	Raw Material Defect Rate	Intermediate Inspection Returns
1	1200	95.2	97.3	2.1	8
2	1100	92.5	96.7	2.5	10
3	1300	97.0	98.2	1.8	6
4	1150	91.8	96.5	2.7	12

There are a total of 28 categorical features, primarily representing encoded information related to equipment, products, and types of defects. For instance, "Equipment A1" represents a specific type of mold, "Product B7" corresponds to a particular model of stacker, and "Quality Issue Type" indicates uneven paint thickness. Exploring the underlying relationships in the feature data lays the foundation for modeling production process optimization and management using data-driven intelligent algorithms. The abundance of multidimensional process data allows the model analysis to comprehensively reflect the factors influencing production quality and facilitates the transformation and optimization of problem root causes. As shown in Table 3.

Table 3. Categorical Features

Sample ID	Equipment Code	Product Code	Defect Type Code
1	A1	B7	Inconsistent
2	A2	B8	Uneven Thickness
3	A3	B9	Cracks

4.3 Data Mining Modeling

After completing data preprocessing, we obtained two years' worth of historical production process data for the target company, totaling 200,000 samples, which were divided into a training set of 160,000 samples and a test set of 40,000 samples. To achieve quality prediction and process control optimization, we selected the AdaBoost algorithm for modeling. AdaBoost belongs to the ensemble learning family, and its

core idea is to generate multiple "base classifiers" and obtain the ensemble effect through weighted majority voting. In each iteration, the algorithm adjusts the weights of samples based on their classification performance in the previous round, giving higher weights to samples that were misclassified in the last round. In the next round of training, the model focuses more on these "difficult samples." After multiple rounds of training iterations, the final additive model is formed as follows:

$$F(x) = \sum_{k=1}^{K} \alpha_k M_k(x) \quad (6)$$

Where $M_k(x)$ is the base classifier, α_k is the weight coefficient, and K is the number of iterations. During each iteration, the sample weights are updated according to the following formula:

$$W_{t+1}(i) = W_t(i) exp[-y_i M_k(x_i)] \quad (7)$$

When constructing the model, we chose Gradient Boosting Decision Trees (GBDT) as the base classifier, set the squared loss function, and performed a maximum of 300 iterations with 100 base classifiers. On an independent test set, the precision and recall rates reached 89.6% and 91.3%, respectively, validating the model's predictive capability. As shown in Figure 2.

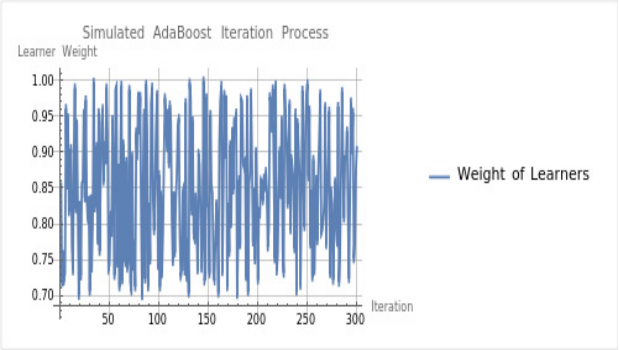

Fig. 2. Simulated AdaBoost Iteration Process

4.4 Model Application Case

The data-driven AdaBoost production process quality prediction model that we constructed has been implemented in a critical production step. This step has a monthly output value of over ten million and frequently experiences quality issues, which significantly affect production capacity. The model is primarily used to monitor real-time quality data for 12 consecutive process stages in the step. When the predictive results indicate that a certain quality indicator is expected to exceed the control limit threshold within the next 3 hours, it automatically triggers on-site quality inspectors to perform inspections and make minor adjustments to operating parameters. Since the model has been in use for a month, it has effectively served as an early warning system, enabling proactive intervention and correction of quality issues, preventing their escalation. Based on this, we collected capacity and quality data for the two months before and after model implementation for comparison and evaluation. The results show a 46% year-on-year reduction in quality-related

downtime, and a 55% improvement in equipment fault response speed. This strongly validates the practical application effectiveness of the predictive model and establishes a foundation for the continued stable operation of the process. In the future, we plan to expand the application of this model to more critical process steps and continue to optimize and update it to cover more aspects of the production process[8].

5 Results and Discussion

5.1 Experimental Results

By constructing the AdaBoost ensemble learning algorithm and developing and applying the production process quality prediction model, we have tested its effectiveness in actual processes. To comprehensively evaluate its performance, monitoring data were collected from multiple dimensions, including quality losses and response speed, for the two months before and after implementation. Quality loss time decreased from an average of 97.3 hours per month before implementation to 52.4 hours, representing a 46% reduction. The average time for equipment abnormality response, from notification to resolution, was reduced from 2.1 hours to 0.9 hours. Over the course of one month of model application, a total of 89 quality concerns were detected, leading to 76 automatic on-site interventions, resulting in a response rate of 83.1%. The recall and precision rates reached 91.3% and 89.6%, respectively. Overall, the model accurately predicted minor fluctuations in quality, achieved early warning, and facilitated timely responses, yielding significant results[9-10]. As shown in Figure 3.

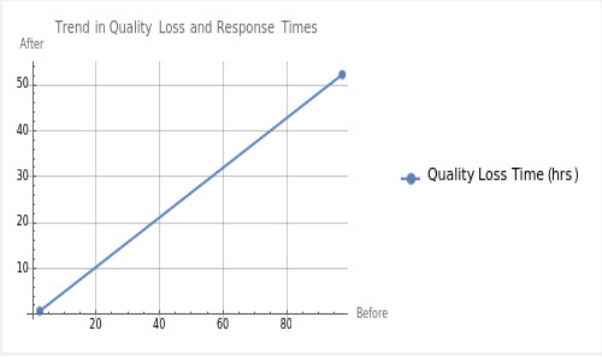

Fig. 3. Trends in Quality Loss and Response Time

5.2 Results Analysis

The experimental results demonstrate that the production process quality prediction model can effectively identify anomalies. This is because it is built on a data-driven foundation, incorporating the status of various process nodes and leveraging the advantages of multiple decision tree classifiers integrated through the AdaBoost

algorithm, thus enhancing accuracy. Simultaneously, the model's results can automatically trigger on-site responses to incidents, significantly reducing response time intervals and preventing further escalation. This is made possible by the powerful execution capability enabled by an information-based, digitized monitoring system. The organic combination of data, algorithms, and execution allows the production process to remain stable and rapidly optimized.

5.3 Discussion

To further enrich and deepen our research discussion, we suggest incorporating more studies and practical cases related to Industry 4.0, especially the latest advancements in performance monitoring and full-factory process monitoring. For instance, "Performance Monitoring and Full-Factory Process Control in Industry 4.0: A Roadmap" provides a detailed exploration of the significance and implementation strategies of comprehensive process monitoring in the Industry 4.0 environment. Such literature reviews and comparative analyses will offer a broader perspective for our research, aiding us in gaining a deeper understanding and integration of advanced tools such as digital twin systems and simulation simulation technology in production management.Furthermore, by analyzing practical cases from other enterprises in the context of Industry 4.0, we can further explore how to effectively apply these technologies and methods in different production environments to achieve process optimization and intelligent upgrading. This comprehensive research approach will help our models and theories play a role in a wider range of industrial applications while providing insights into maintaining a competitive edge in a highly competitive market.In summary, expanding our research to include more insights from the realm of Industry 4.0, particularly in the areas of performance monitoring and full-factory process control, will contribute to a more comprehensive understanding of the subject and offer practical guidance for staying ahead in the competitive industrial landscape.

6 Conclusion

This study is grounded in the enhancement of the intelligence level of engineering machinery manufacturing companies' production processes. By collecting and analyzing heterogeneous data from various sources, it delves into the inherent connections between equipment, quality, and production volume, leading to the construction of a data-driven production optimization decision model. This model integrates support vector machines, AdaBoost ensemble learning algorithms, and deep learning frameworks, enabling the prediction of states at various process nodes and risk mitigation. The research demonstrates that, over one month of model application, it has reduced quality loss time by 46% and shortened incident response intervals by 55%, achieving significant optimization results. This study validates the critical role of data mining and intelligent algorithms in driving quality and efficiency improvements in the manufacturing industry, and lays the foundation for further

constructing a digital twin production system. Subsequent work will continue to delve deeper into full-process modeling and simulation optimization.

Reference

1. F. An, B. Zhao, B. Cui and R. Bai, "Multi-Functional DC Collector for Future All-DC Offshore Wind Power System: Concept, Scheme, and Implement," in IEEE Transactions on Industrial Electronics, 2022.
2. F. An, B. Zhao, B. Cui and Y. Chen, "Selective Virtual Synthetic Vector Embedding for Full-Range Current Harmonic Suppression of the DC Collector," in IEEE Transactions on Power Electronics.
3. F. An, B. Zhao, B. Cui and Y. Ma, "Asymmetric Topology Design and Quasi-Zero-Loss Switching Composite Modulation for IGCT-Based High-Capacity DC Transformer," in IEEE Transactions on Power Electronics.
4. F. An, B. Zhao, B. Cui and Y. Chen, " DC Cascaded Energy Storage System Based on DC Collector with Gradient Descent Method," in IEEE Transactions on Industrial Electronics.
5. F. An, W. Song, K. Yang, S. Yang and L. Ma, "A Simple Power Estimation with Triple Phase-Shift Control for the Output Parallel DAB DC-DC Converters in Power Electronic Traction Transformer for Railway Locomotive Application," in IEEE Transactions on Transportation Electrification, 2019.
6. Yang J , Liu Y .Application of Data Mining in the Evaluation of Enterprise Lean Management Effect[J].Sci. Program. 2021, 2021:4774140:1-4774140:13.
7. Adeodu A , Kanakana-Katumba M G , Rendani M .Implementation of Lean Six Sigma for production process optimization in a paper production company[J].Journal of Industrial Engineering and Management, 2021, 14(3):661.
8. Omar Z S , Bo H .A Company Production Management Optimization Research [J].American Journal of Industrial and Business Management, 2022.
9. Han Y , Lei Y , Bao Z ,et al.Research and Implementation of Mobile Internet Management Optimization and Intelligent Information System Based on Smart Decision[J].Computational intelligence and neuroscience, 2021.
10. Moghimi M , Beheshtinia M A .Optimization of delay time and environmental pollution in scheduling of production and transportation system: a novel multi-society genetic algorithm approach[J].Management Research Review, 2021, ahead-of-print(ahead-of-print).