Research on Strip Steel Process Optimization Based on Data-Driven Approach and Genetic Algorithm

Yiliang Huang *, Yang Liu, Wenjian Wei

School Of Information Science and Engineering, Dalian Polytechnic University, Dalian, China, 116034

* Corresponding Author Email: tydxfl@163.com

Abstract. In the context of modern manufacturing, ensuring consistent product quality while optimizing production processes is a significant challenge. This paper presents a comprehensive approach to improving the quality inspection and process optimization of steel strip production using machine learning models and genetic algorithms. Historical production data were pre-processed through cleaning, feature selection, and z-score normalization to ensure accuracy in the model's performance. Two classification models, Logistic Regression and K-Nearest Neighbors (KNN), were evaluated for their ability to predict product quality. The Logistic Regression model outperformed KNN with an accuracy of 0.36, making it the preferred choice for the online inspection system. To optimize the steel strip production process and achieve the target hardness value, a Genetic Algorithm (GA) was applied to determine the best combination of process parameters. The GA simulation considered key factors such as strip thickness, width, carbon and silicon content, and furnace temperatures. The results demonstrate that the integration of a predictive model and optimization algorithm provides a data-driven solution for enhancing product quality and operational stability in steel strip manufacturing.

Keywords: Logistic Regression, K-Nearest Neighbors, Genetic Algorithm, Process Optimization.

1. Introduction

In modern manufacturing industries, ensuring product quality while optimizing production processes has become a critical challenge [1]. Steel strip production, as a core component of many industrial applications, requires precise control over various process parameters to achieve consistent product quality [2]. However, traditional methods of setting process parameters largely rely on the experience of field operators, which can lead to variability in product quality [3, 4]. Moreover, the growing demand for higher quality standards and cost-efficient production necessitates a more data-driven approach to managing and optimizing manufacturing processes [5].

In response to this challenge, machine learning techniques, such as logistic regression [6] and K-nearest neighbors (KNN) [7], have been widely explored for predictive modeling. These methods enable manufacturers to analyze large volumes of production data, identify key factors influencing product quality, and build models capable of making accurate predictions [8]. In addition to predictive modeling, optimization algorithms like genetic algorithms (GA) have shown significant promise in refining process parameters [9]. By simulating the mechanisms of natural evolution, GA can efficiently search for the optimal combination of process parameters, further enhancing product quality and production efficiency [10].

This research addresses the need for a reliable online inspection model and an optimized production process for steel strip products. By integrating machine learning models with optimization techniques, it aims to develop a comprehensive solution that not only predicts product quality but also optimizes the production process to achieve target specifications.

2. Data Processing and Model Selection

The data in this paper originates from https://cxcy.upln.cn/. To build an effective online inspection model, we begin by thoroughly pre-processing the historical data. This involves cleaning the data to handle any missing values and outliers, ensuring that the dataset is both accurate and consistent for model training. We then focus on feature selection, initially including all variables related to hardness, as identified in the earlier analysis. As the model evolves, we refine these features by removing or adding variables to improve the model's predictive accuracy. After feature selection, we apply z-score normalization, which transforms the data using its mean and standard deviation, standardizing the feature set to ensure that differences in the scale of the variables don't negatively impact model performance.

Once the data is prepared, we explore two different models for training and evaluation. Logistic Regression, a linear classification model, uses the sigmoid function to handle binary classification problems. It's favored for its simplicity, ease of implementation, and high computational efficiency, though it can sometimes suffer from overfitting when too many features are present. On the other hand, K-Nearest Neighbors (KNN) is a non-parametric model that classifies data points based on their proximity to other labeled examples in the dataset. While intuitive and simple to use, KNN requires careful tuning of the parameter K and can be computationally expensive when dealing with large datasets. Both models are well-suited to our problem, and their strengths will be assessed to determine which performs best in predicting outcomes based on the pre-processed data.

First, this study divides the dataset into a training set and a test set, with the test set accounting for 20%. Furthermore, the feature data is standardized, as shown in Eq1.

$$z = \frac{\mathrm{x}-\mu}{\delta} \tag{1}$$

Furthermore, in training the logistic regression classifier, this study sets the maximum number of iterations to 2000. The relevant formula is shown in Eq2.

$$\mathrm{p} = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 \mathrm{x}_1+\beta_2 \mathrm{x}_2+\cdots+\beta_n \mathrm{x}_n)}} \tag{2}$$

In the equation, n=12, mainly because there are 12 feature values. The labels for the test data are predicted, resulting in the predicted class labels for each test sample.

Regarding accuracy calculation, it is primarily determined by calculating the proportion of correctly predicted samples to the total number of test samples.

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

Furthermore, the performance evaluation report is obtained using the following formula, where precision is calculated as:

$$precison = \frac{TP}{TP+FP} \tag{4}$$

Recall:

$$call = \frac{TP}{TP+FN} \tag{5}$$

F1 Score:

$$all = 2 \times \frac{precision \times call}{precision + call} \tag{6}$$

For K-nearest neighbor (KNN) training, after defining and initializing the model, it is set to consider the 5 nearest neighbors for each test sample during prediction. The training labels are predicted using the Euclidean distance formula. Similarly,

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i^2 - p_i^2)} \tag{7}$$

The evaluation of model accuracy is consistent with that of the logistic regression classifier and will not be elaborated on here due to space constraints. The prediction results of the two algorithms are shown in Figures 1 and 2.

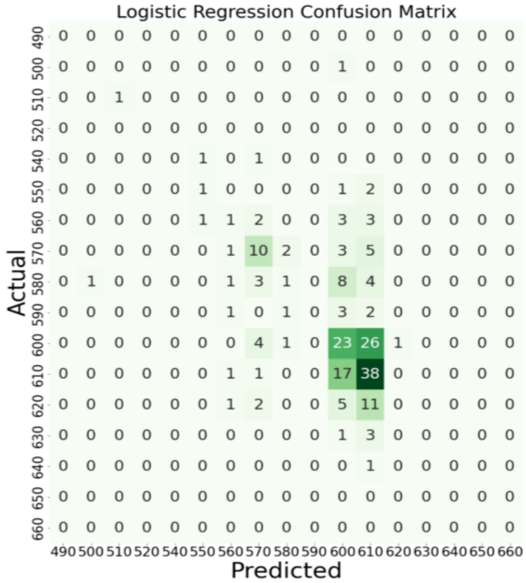

Figure 1. Logistic regression analysis

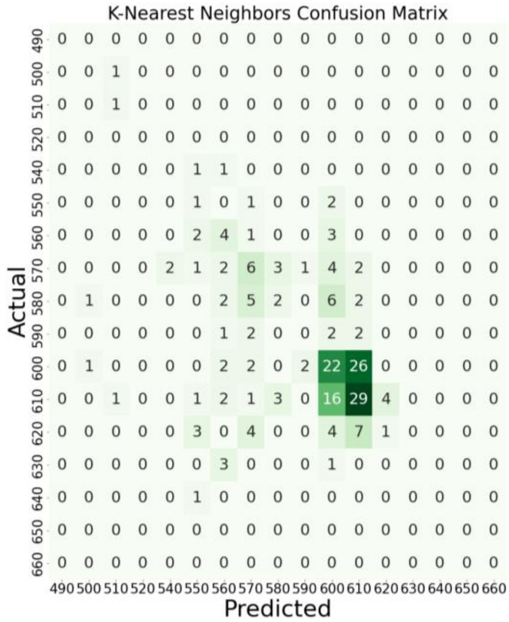

Figure 2. k Nearest Neighbour Analysis

The results show that the accuracy of logistic regression is 0.36 as shown in Fig. 3 the accuracy of K nearest neighbour classifier is 0.31 as shown in Fig. 3 and Fig. 2. i.e. logistic regression performs better and hence logistic regression model is selected for the model. The partial results of the test set for partial testing using logistic regression are shown in Figure 3 below:

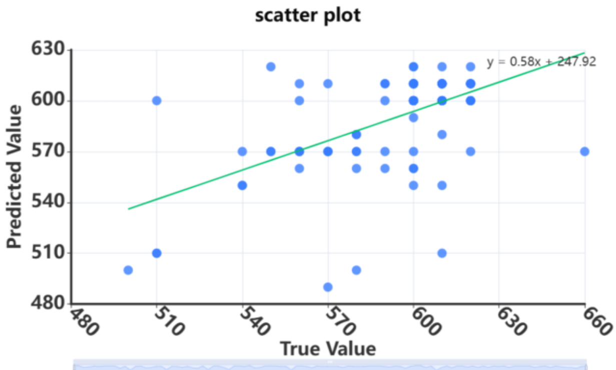

Figure 3. Algorithm Performance Visualization

In summary, the logistic regression model is selected as the online inspection model of strip steel product quality.

3. Process Optimisation via Genetic Algorithm

A data-driven process parameter optimisation solution needs to be established in order to solve the problem, as field operators rely on their personal experience to set the process parameters of their products, which can easily lead to unstable product quality. Among them, the specific optimisation objectives are as follows:

(1): Achievement of target hardness value.

(2): Ensure that the process parameters are set reasonably and operationally.

Based on previous section, this study chose a logistic regression model for prediction.

This optimisation algorithm will use genetic algorithm in order to search for the optimal process combination to achieve the target hardness.

Genetic Algorithm (GA) is an algorithm that simulates the mechanism of natural selection and genetics. The algorithm draws on the evolutionary law of "survival of the fittest and survival of the fittest" in the theory of biological evolution, and solves optimisation problems by simulating the mechanisms of selection, crossover (hybridisation) and mutation in the natural evolutionary process. Genetic algorithms are usually used to solve search and optimisation problems, where the search space is represented as a "population" of candidate solutions, and each candidate solution is encoded as a "chromosome" or "individual". Each candidate solution is encoded as a "chromosome" or "individual". The three processes of selection, crossover and mutation will be utilised in this work.

The following is the design process:

3.1. Initialising populations

$$p_{population} = \{x_{i1} + x_{i2} + \ldots + x_{ipop}\} \tag{8}$$

Here x_{i1}, x_{i2},..., x_{ipop} are all individuals in the population, and in this model the population size is set to bepop = 100.

3.2. Define the fitness function

Virtuous

$$f_{(x)} = finess(x) \tag{9}$$

Using the fitness function $f_{(x)}$ to evaluate the x_{ipop} strengths and weaknesses of individual populations.

3.3. Option

The probability of selecting into the next generation of individuals through the selection of fitness is

$$p = \frac{f_{(x_{select})}}{\sum_{j=1}^{pop} f_{(x_j)}} \tag{10}$$

3.4. Cross-cutting

According to the crossover formula

$$c1 = x_{p1}[1: c], x_{p2}[c + 1: length] \tag{11}$$

$$c2 = \left[x_{p2}[1: c], x_{p1}[c + 1: length]\right] \tag{12}$$

Perform simulation processes for biological reproduction

3.5. Variation

$$x'_{select} = x_{select} + \Delta x_{select} \tag{13}$$

Where Δx_{select} is a randomly varying quantity with a small value.

3.6. Optimisation

Firstly, the target hardness value is set to 600, The number of iterations is set to 50 and the training set is optimised to obtain the optimal solution and optimal fitness 20.

Based on the standardised formula

$$\frac{x-\mu}{\delta} \tag{14}$$

Obtainable

$$x = z \times \delta + \mu \tag{15}$$

Conversion from inverse normalisation to the original data scale results in an optimal solution for the original process parameters. That is, the original optimal solution is shown in the Table 1:

Table 1. Table of original optimal solutions

Process parameters	(be) worth
Strip thickness	8818.683996
Strip width	214.4598286
carbon content	404.2284208
silicon content	5.55962872
Strip speed	629.1652759
Furnace temperature	737.407258
Furnace temperature	624.6095088
Retarder temperature	627.2060628
Over-aging furnace temperature	334.9561633
Fast cooling furnace temperature	68.6737079
Hardening furnace temperature	27.58978065
Levelling machine tension	2365.404784

4. Conclusion

This study developed an effective online inspection model for steel strip product quality using logistic regression and K-Nearest Neighbors (KNN). After pre-processing the historical data, including cleaning, feature selection, and standardization, the performance of both models was tested. The results demonstrated that the logistic regression model outperformed KNN with a prediction accuracy of 0.36 compared to KNN's 0.31. Thus, logistic regression was selected as the final model for this task.

To further optimize the steel strip production process, a genetic algorithm (GA) was employed to find the best process parameter settings that achieve the target hardness value. The genetic algorithm's iterative search for an optimal combination of parameters, including strip thickness, width, carbon content, silicon content, and various furnace temperatures, led to a stable and effective process solution. This optimization ensures not only the achievement of the target hardness but also operational feasibility, providing a practical data-driven approach to process parameter optimization in the production of steel strips.

In conclusion, the combination of logistic regression for quality prediction and genetic algorithms for process optimization provides a robust framework for improving product quality and process stability in modern manufacturing. This approach can be extended to other similar industrial applications where both prediction and optimization are critical.

References

[1] AMMAR M, HALEEM A, JAVAID M, et al. Improving material quality management and manufacturing organizations system through Industry 4.0 technologies [J]. Materials Today: Proceedings, 2021, 45: 5089-5096.

[2] VAFADAR A, GUZZOMI F, RASSAU A, et al. Advances in metal additive manufacturing: a review of common processes, industrial applications, and current challenges [J]. Applied Sciences, 2021, 11 (3): 1213.

[3] LI X, LIU X, YUE C, et al. Systematic review on tool breakage monitoring techniques in machining operations [J]. International Journal of Machine Tools and Manufacture, 2022, 176: 103882.

[4] JI C, SUN W. A review on data-driven process monitoring methods: Characterization and mining of industrial data [J]. Processes, 2022, 10 (2): 335.

[5] ZAFARZADEH M, WIKTORSSON M, BAALSRUD HAUGE J. A systematic review on technologies for data-driven production logistics: Their role from a holistic and value creation perspective [J]. Logistics, 2021, 5 (2): 24.

[6] van den GOORBERGH R, van SMEDEN M, TIMMERMAN D, et al. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression [J]. Journal of the American Medical Informatics Association, 2022, 29 (9): 1525-1534.

[7] KOHLI S, GODWIN G T, UROLAGIN S. Sales prediction using linear and KNN regression [M] // Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019. Springer, 2020: 321-329.

[8] WANG J, XU C, ZHANG J, et al. Big data analytics for intelligent manufacturing systems: A review [J]. Journal of Manufacturing Systems, 2022, 62: 738-752.

[9] ALAM T, QAMAR S, DIXIT A, et al. Genetic algorithm: Reviews, implementations, and applications [J]. arXiv preprint arXiv: 2007.12673, 2020.

[10] ZHOU Y, WANG Y, WANG K, et al. Hybrid genetic algorithm method for efficient and robust evaluation of remaining useful life of supercapacitors [J]. Applied Energy, 2020, 260: 114169.