Prediction of mechanical properties of hot-rolled strip steel based on PCA-GBDT method

Tianqi Cheng¹,², Guici Chen¹,²
¹College of Science, Wuhan University of Science and Technology, Wuhan 430065, China
²Hubei Province Key Laboratory of Systems Science in Metallurgical Process, Wuhan University of Science and technology, Wuhan 430065, China

Corresponding author and e-mail: Guici Chen, chenguici@wust.edu.cn

Abstract. Improving the quality of rolled steel products is the primary task of the entire steel industry. As the main step of rolled steel production, hot rolling has received extensive attention. As far as the hot rolling process is concerned, the chemical composition of steel and related process parameters are the most direct factors affecting the quality of hot rolled steel sheets. Based on Principal Component Analysis (PCA) and Gradient Lifting Decision Tree (GBDT) methods, this paper takes the tensile strength as the research object and constructs a prediction model for the mechanical properties of hot rolled strip steel. Through principal component analysis of characteristic data, 28 variables were reduced into 8 new indicators to be used for GBDT regression analysis. 2489 pieces of data were divided into a training set and a test set at a ratio of 7:3, to be used in the training set to build a regression model, and get a root mean square error of 16.7393. The data in the test set was used to predict the tensile strength value, with the root mean square error reaching 18.2650.

1. Introduction
The development of strip steel products is the same as my country's intensive economic growth mode, and the essence is to improve product quality and economic benefits. As one of the main types of steel products, hot-rolled strip steel accounts for the largest proportion of the total steel output and occupies a dominant position in the production of rolled steel.

With the rapid development of the global economy, hot-rolled strip producers are increasingly concerned about the quality stability of their products and the flexible production capacity of their production lines. The prediction system of steel mechanical properties has become one of the research directions that metallurgists around the world pay the most attention to. Due to the complexity of steel rolling production data, it cannot adapt to the actual process of hot rolling production. In recent years, with the development of big data, the improvement of computer performance, the emergence of artificial intelligence and the proposal and development of neural networks have provided effective technical support for the study of mechanical properties of hot-rolled products, which has become the mainstream of research [1-6]. Due to the complexity of the rolling production process and the severe nonlinearity, time-varying, large lag, strong coupling and multi-parameter characteristics of rolling data, neural networks have certain shortcomings. Based on this, we propose a steel rolling mechanical performance prediction model based on principal component analysis-gradient boosting decision tree (PCA-GBDT). GBDT algorithm is one of the best performance methods in machine learning and has been applied in
many fields, such as Zhang Weiwei (2017) used the measured vehicle travel time data to propose a principal component analysis-gradient boosting decision tree (PCA-GBDT) method for urban road travel time prediction. The results show that, compared with the traditional KNN method, time series ARIMA method, and SVM method, the PCA-GBDT method has higher prediction accuracy and algorithm stability [7].

2. Research methods

2.1. Principal component analysis

Among the \( p \) observation variables \( x_{i1}, x_{i2}, \ldots x_{ip} \) of \( N \) observation objects, the observation matrix \( X_{N \times p} \) is obtained. Assuming that the observation matrix is standardized, the \( p \) observation variables are combined and called a comprehensive variable. The core of PCA is to take a linear combination in each set:

\[
f_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \sum_{j=1}^{p} \beta_j x_{ij} = \beta' x_i, \quad i = 1, 2, \ldots, N
\]

(1)

Where \( \beta = (\beta_1, \beta_2, \ldots \beta_p)' \), \( x_i = (x_{i1}, x_{i2}, \ldots x_{ip})' \), \( \beta_j \) is the weight coefficient of the \( j \)-th variable of the observation value \( x_{ij} \), \( f_i \) is the principal of the \( i \)-th observation object Ingredient score. In this paper, \( N=2489 \) (the number of hot-rolled strip steel), \( P=28 \) (chemical composition and process parameters), because different weighting coefficient \( \beta \) can get different comprehensive variables, so there can be multiple comprehensive variables for \( p \) variables. The variance of a comprehensive variable is a scale that can comprehensively reflect the degree of change of the observed variable.

Therefore, the principal component can be defined in terms of variance:

\[
\text{Var}(f) = \frac{1}{N-1} \sum_{i=1}^{N} \left( \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .
\]

(2)

The first principal component satisfies

\[
\begin{align*}
& \max \beta_1' \text{Var} \beta_1 \\
& \text{s.t. } \beta_1' \beta_1 = 1
\end{align*}
\]

(3)

Where \( V \) is the sample covariance matrix.

The \( k \)-th principal component satisfies

\[
\begin{align*}
& \max \beta_k' \text{Var} \beta_k \\
& \text{s.t. } \beta_k' \beta_k = 1, \beta_k' \beta_m = 1, m < k, \beta_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{pk})
\end{align*}
\]

(4)

By determining the normalized weight coefficient, the variance of the principal component \( f_i \) is maximized. The largest variance is interpreted as the first principal component, the second largest variance is the second principal component, and so on. Several principal components obtained by PCA are a linear combination of original variables and are not related to each other, which can represent the most important characteristic component of the original data.

2.2. Gradient boosting decision tree

The Gradient Boosting Decision Tree (GBDT) algorithm was proposed by Jerome Friedman in 2001 and can be used for classification and regression. It is a combination of gradient boosting algorithm and decision tree.

Assuming the training set is \( T = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\} \), in the optimization process, GBDT adopts forward segmentation regression, by continuously adding a new decision tree to Reduce the error function value without changing the parameters of the existing decision tree. The loss function is calculated as follows:

\[
L(f) = \sum_{i=1}^{n} L(y_i, f(x_i)).
\]

(5)

For continuous variables, the classic form of the loss function is the sum of squared errors:

\[
L(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2.
\]

(6)
Assuming that in the m-th round of learning, the strong learner is \( f_m(X) \) and the loss function is \( L(y, f_m(X)) \), then the maximum loss function of the i-th sample in the \( m + 1 \) round decreases. The direction is its gradient direction:

\[
\begin{align*}
  r_{i,m+1} &= -\left[ \frac{\partial L(y, f(X))}{\partial f(x_i)} \right]_{f(x_i)=f_m(x_i)}.
\end{align*}
\]

(7)

Using \((x_i, r_{i,m+1}), i = 1,2, ... n\), a decision tree and the \( m + 1 \)th regression tree can be fitted. Assuming that the number of leaf nodes is \( J \), the corresponding leaf node area is \( R_{m+1,j} (j = 1, 2, ..., J) \).

For each sample in the leaf node, find the minimum loss function, which is to fit the optimal output value of the leaf node,

\[
  c_{m+1,j} = \arg \min_c \sum_{x_i \in R_{m+1,j}} L(y_i, f_m(x_i)) + c.
\]

(8)

The decision tree of the \( m + 1 \) round can be obtained, and its fitting function is:

\[
  h_{m+1}(x) = \sum_{j=1}^{J} c_{m+1,j} I(x \in R_{m+1,j}).
\]

(9)

Therefore, the final expression of the strong learner in the \( m + 1 \) round is as follows:

\[
  f_{m+1}(x) = f_m(x) + \sum_{j=1}^{J} c_{m+1,j} I(x \in R_{m+1,j}).
\]

(10)

This paper comprehensively uses these two methods, considering the characteristics of non-linearity, time-varying, large lag and other characteristics of rolled steel production data, and proposes a mechanical performance prediction method based on PCA-GBDT, and verifies the model's performance based on the actual data detected. Forecast reliability and algorithm stability.

3. Empirical analysis

3.1. Data sources

In this paper, the production data of 2,489 hot-rolled steel strips produced by domestic large-scale hot-rolling mills are used to discuss the influence of the chemical composition of steel and related process parameters on the mechanical properties of hot-rolled steel strips.

3.2. PCA-based pivot selection

Using python software, input the original data to form a sample set, and perform dimensionality reduction processing on the original data, but the feature representation obtained after the principal component dimensionality reduction is not clear, so we first sort the original features by importance.

As can be seen from Figure 1, the six characteristic variables of CT, FT, FET, RT, FDH and FDT have a greater impact on the compressive strength of hot-rolled strip steel, and play an important role as
an external factor to control the quality of strip steel products. They are all external forces that can be controlled artificially.

As just now, the internal chemical factors are extremely important to the mechanical properties of the product. Among them, C, Mn, and Si are the factors that play a major role in the mechanical properties of the product, which can improve the strength, hardness and wear resistance of the steel. S and P are elements harmful to steel grades, the lower the better. Nb can refine the grain of steel, reduce the overheating sensitivity and temper brittleness of steel; improve the welding performance of steel, and increase the strength and corrosion resistance of heat-resistant steel. Ni can improve the strength, toughness, heat resistance, corrosion resistance, acid resistance, and magnetic permeability of steel. Increase the hardenability and hardness of steel.

![Figure 2. Crushed stone chart.](image)

| ingredient | total   | variance percentage | accumulation |
|------------|---------|---------------------|--------------|
| 1          | 8.060   | 28.787              | 28.787       |
| 2          | 3.097   | 11.060              | 39.848       |
| ...        | ...     | ...                 | ...          |
| 8          | 1.108   | 3.956               | 75.508       |
| ...        | ...     | ...                 | ...          |
| 28         | 4.614E-13 | 1.648E-12           | 100.00       |

The main idea of PCA is to reduce dimensionality and use as few factors as possible to explain more variation. Through SPSS gravel graphs, it is found that there are 8 factors with characteristic values greater than 1, and the top 8 can be found through the contribution rate of the factors to the variable explanation. A factor can express the cumulative contribution rate of the variable to 75.508%, which can represent most of the sample information to a certain extent. Therefore, the number of pivots is determined to be 8, and the feature vectors corresponding to the first 8 pivots are output to form a feature matrix, and the 8-dimensional features are used as the input of the predictive performance index.

3.3. Model rating indicators
In order to evaluate the prediction performance of the mechanical performance prediction model, the experimental simulation uses root mean square error (RMSE) to measure the accuracy of the prediction model. The smaller the RMSE value, the more accurate the prediction.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}
\]

Where \( n \) is the number of samples, \( \hat{y}_i \) is the predicted value of the \( i \)-th sample, and \( y_i \) is the true value of the \( i \)-th sample.

### 3.4. Modeling process and result analysis (based on PCA-GBDT to establish regression prediction model)

In order to ensure the generalization of the algorithm and the portability of the experiment, the data set is divided into training set and test set according to the ratio of 7:3, and they are put into the regression model for training, and the root mean square of the regression model of the data in the training set is found. The error is 16.7393. Then the data in the test set is used to predict the compressive strength, and compared with the real value in the test set, the root mean square error is 18.2650, which shows that the predicted compressive strength is more accurate and can guarantee the accuracy of the prediction.

### 4. Conclusions

This paper proposes a mechanical performance prediction model based on PCA-GBDT, and empirically verifies the effectiveness of the method. Aiming at the problems of multi-dimensional hot-rolled strip data, the existence of data noise, and the redundancy between attributes, PCA screening is used to obtain low-dimensional features to reduce the dimensionality of a subset of model input attributes, reduce training time, and improve model stability. Comparison with other machine learning models shows that the GBDT model has superior performance in terms of prediction accuracy and model interpretation ability, thus can well identify complex nonlinear relationships, and is suitable for predicting the mechanical properties of hot rolled strip steel.

### References

[1] Ghaisari J., Jannesari H. and Vatani M. 2011 J. Artificial neural network predictors for mechanical properties of cold rolling products. Advances in Engineering Software. chapter 1.

[2] Xiaomei X. BP neural network and its application in predicting mechanical properties of hot rolled strip 2014 C. the Chinese Society for Metals. P 501.

[3] Yang Y, Xin T and Wei S 2014 J. Discussion on prediction of mechanical properties of hot rolled strip based on genetic algorithm and BP artificial neural network. Industrial control computer. chapter 6 pp 71-74.

[4] Yong M, Xiaopeng W and Shasha M 2019 J. Prediction of mechanical properties of hot rolled strip steel based on BP neural network under the framework of Keras deep learning. Metallurgical Automation. chapter 2 pp 6-10.

[5] Shixiong H, Weigang L and Wei Y 2018 J. Prediction of mechanical properties of hot rolled strip steel based on convolutional neural network. Journal of Wuhan University of Science and Technology. chapter 5 pp 338-344.

[6] Shixiong H 2019 D. Prediction of mechanical properties of hot rolled strip steel based on deep learning. Wuhan University of Science and Technology.

[7] Weiwei Z, Ruimin L and Zhongjiao X 2017 J. Prediction method of urban road travel time based on PCA-GBDT. Road construction. Chapter 6 pp 6-11.