Research on prediction classification and compensation for silicon contents in blast furnace based on ridge regression

Ling Sun¹, Linzi Yin¹*, Zhaohui Jiang², Yuyin Guan¹, Xuemei Xu¹

¹School of Physics and Electronics, Central South University, Changsha, China
²School of Automation, Central South University, Changsha, China

*Corresponding author e-mail: yinlinzi@csu.edu.cn

Abstract. In blast furnace, the accurate prediction of silicon contents in hot metal plays an important role in comprehensively judging the furnace states and making correct decisions. However, traditional prediction models are easily affected by multi-collinearity among the independent variables and imbalance sample distribution in hot metal datasets. The prediction values of silicon contents would be difficult to track the real values accurately when the furnace conditions fluctuate dramatically. In this paper, a classification and compensation model for silicon contents is proposed by combining the ridge regression and random forest classification algorithm. Firstly, the ridge regression algorithm is applied to fit and optimize the prediction values outputted by a front-end prediction model. Then, the random forest technology is expected to classify and identify the furnace states. Finally, the predicted values are compensated according to classification results. In this paper, the random forest regression and BP neural network algorithm were used to predict silicon content as front-end models. The related applications show that the proposed classification and compensation model improves the accuracy of predicted values effectively.

1. Introduction

Blast furnace (BF) is a highly complex nonlinear industrial system, the essence of which is to reduce iron from iron-bearing compounds such as iron ore [1]. Silicon content in hot metal, as an important production index, has a linear relationship with the furnace temperature and the quality of hot metal. It is a key parameter to characterize the changing trend of furnace conditions. Therefore, it is necessary to establish an accurate prediction model of silicon contents to improve the control level in BF system.

In recent years, numerous experts and scholars have made lots of efforts on data-driven prediction
for silicon contents. They have put forward abundant prediction models, like Bayesian network model [2], support vector machine model [3-4], neural network model [6-7], random forest model [8], etc. These models have achieved good results in terms of prediction, which confirmed the effectiveness of data-driven models in BF field. However, due to highly complex of BF ironmaking process, the prediction performance of data-driven models is usually restricted by the following two aspects.

(1) The problem of multi-collinearity in hot metal datasets. The high correlation among independent variables would increase the variance of regression models, and cause the serious robustness and reliability of models.

(2) The imbalance of sample distribution because of the fluctuation of furnace states. The sample size of high-silicon and low-silicon are much smaller than that of middle-silicon. Therefore, it is difficult to track and reflect the variation tendency of actual values timely when furnace states fluctuant frequently, which causes a negative impact on prediction performance.

This paper introduces a classification and compensation model for silicon contents. For the first problem, the ridge regression algorithm is applied to optimize the prediction values outputted by a front-end prediction model so that alleviate the multi-collinearity. For the second problem, the random forest technology is suggested to identify furnace states. Finally, the initial prediction values from a front-end model are compensated to improve the prediction accuracy of silicon contents. The related applications show that the proposed model provides more accurate and reliable furnace condition information for BF operators.

2. Prediction problems and solution strategies

2.1 The problem of multi-collinearity

Multi-collinearity means that there is a correlation between certain components and other independent variables in regression models and moreover, they all have a certain degree of influence on the dependent variables [9]. In fact, it is hard to find out a set of irrelevant variables but have a noticeable influence on the dependent variables in a real problem. And multi-collinearity would lead to the deterioration of robustness as well as generalization ability of models.

In statistics, variance inflation factor (VIF) is usually suggested to diagnose if there is a multi-collinearity phenomenon in a dataset and the related definition is as below.

\[
VIF = \frac{1}{1-R^2} \tag{1}
\]

where \(R^2\) is a determinable coefficient of multiple explanatory variables in regression equations. The larger the \(R^2\), the greater the VIF and, the more serious the multi-collinearity among variables. According to experience of experts, it is considered that the VIF less than 10 is acceptable. When the VIF is between 10 and 100, it manifests a weak multi-collinearity between the explanatory variable and others. But when the VIF is greater than 100, some necessary measures should be taken immediately to relieve the multi-collinearity problem.

In this article, we take samples collected from 2650 m³ BF of a domestic steel plant as our research object. And 17 attributes in the dataset are tested by VIF and the related results are shown in the Table 1.
Table 1 Related results of $VIF$ of variables in the hot metal data

| Variables          | Unit    | $VIF$  | Variables          | Unit    | $VIF$  |
|--------------------|---------|--------|--------------------|---------|--------|
| $S_i(n-1)$         | wt%     | 41.83  | $S_i(n-2)$         | wt%     | 42.71  |
| Blast momentum     | kg·m/s  | 32735.09 | Cold air flow     | m³/min  | 552606.55 |
| Oxygen-rich rate   | wt%     | 1598.29 | Permeability Index | m³/min·kpa | 86050.87 |
| Bosh gas volume    | m³/min  | 3240535.59 | Bosh gas index    | m³/min·m² | 3225835.11 |
| Hot air temperature| ℃       | 106244.68 | Theoretical combustion temperature | ℃ | 80674.27 |
| Top pressure       | kpa     | 184922.31 | Oxygen-rich pressure | kpa | 274.62 |
| Cold wind pressure | kpa     | 282156.24 | Full pressure difference | kpa | 125011.53 |
| Standard wind speed| m/s     | 505341.36 | Actual wind speed  | m/s     | 285318.66 |
| Hot air pressure   | kpa     | 581621.15 |                        |         |        |

As can be seen from the Table 1, only two $VIF$ are less than 100, which are 41.83 and 42.71 respectively. Whereas, the other 15 $VIF$ are far greater than 100, which fully demonstrates that there is a critical characteristic of multi-collinearity in the selected dataset.

In this paper, a solution strategy is proposed to minimize the negative effects on prediction models caused by the multi-collinearity. That is, we utilize the ridge regression algorithm to optimize the predicted values from a front-end model, then obtain more accurate and reliable prediction values.

2.2 The problem of unbalanced sample distribution.

The unbalanced distribution is that the numbers of certain classes in datasets are far less than others in a classification problem. The problem means that the categories with less sample size contain too few information to extract related rules for a prediction model. Therefore, the accuracy and robustness would be deteriorated dramatically if the model encounters new unbalanced samples.

During BF ironmaking process, the frequent fluctuations of furnace conditions cause an imbalance distribution of silicon contents in the dataset. We divide silicon contents into three categories named 0, 1 and 2, which represent three furnace states including the low-silicon, middle-silicon and high-silicon contents respectively. 1216 sets of samples are obtained and their distribution of silicon contents is shown in Table 2.

Table 2 Distribution of silicon contents in different intervals of hot metal data

| Interval          | Category | Number | Percentage (%) |
|-------------------|----------|--------|----------------|
| [0,0.32]          | 0        | 132    | 10.86%         |
| (0.32, 0.52]      | 1        | 914    | 75.16%         |
| (0.52, 1]         | 2        | 170    | 13.98%         |

We can see that most of silicon contents are in the middle-silicon segment, and in contrast, only a small part of silicon contents in the high or low-silicon segment. The unbalanced sample distribution would lead to the trained models excessively dependent on the samples in the middle-silicon segment. Consequently, the predicted values of silicon contents in the high-silicon segment are generally lower than the actual values, while the predicted values of the silicon content in the low-silicon are generally higher than the actual values. In general, the unbalanced sample distribution resulting from the fluctuation of furnace conditions would reduce the prediction accuracy and the generalization ability of prediction models.
To this end, it is necessary to identify the furnace states by classifying the predicted results and make some reasonable adjustments. In this paper, we take the random forest classification algorithm to train the prediction values outputted by a front-end prediction model. Then, a classification model is established to achieve the judgment of furnace conditions and facilitate different compensation of silicon contents for three furnace conditions afterwards.

2.3 The solution strategy

For the two problems given in section 2.1 and 2.2, this paper constructs a classification and compensation (CAC) model for the prediction of silicon contents. The flow chart is shown as Figure 1.

![Flow chart of prediction of silicon contents](image)

Fig. 1 The flow chart of prediction of silicon contents.

where \( S_{\text{firs-pre}} \) are the initial prediction values of silicon contents from a front-end model, \( S_{\text{sec-pre}} \) are the second prediction values after optimizing \( S_{\text{firs-pre}} \) by ridge regression model, and \( S_{\text{class}} \) are the classification results (0, 1 or 2) by the random forest classification model.

Firstly, we establish the front-end prediction model (In the section 4, the random forest regression and the BP neural network algorithm are applied respectively.) to acquire initial predicted values of silicon contents \( S_{\text{firs-pre}} \). Then, a global optimization model is built based on the ridge regression algorithm to obtain \( S_{\text{sec-pre}} \), which are much closer to the true values than \( S_{\text{firs-pre}} \). Meanwhile, by applying the random forest classification algorithm, \( S_{\text{firs-pre}} \) are classified into three categories to represent three furnace conditions. The final predicted values of silicon contents are produced by an integrated compensation model based on \( S_{\text{firs-pre}}, S_{\text{sec-pre}} \) and \( S_{\text{class}} \).

The proposed model involves three machine learning algorithms including a front-end prediction algorithm, the ridge regression algorithm and the random forest classification algorithm. The ridge regression effectively alleviates multi-collinearity in the dataset, and the random forest classification reduces negative impacts on models’ performance due to fluctuation conditions. Thus, the final prediction values are more accurate than \( S_{\text{firs-pre}} \) through classification and compensation.

3. The classification and compensation model for prediction of silicon contents

3.1 Optimization model based on the ridge regression algorithm

The ridge regression analysis\(^9\) is a complementary regression to the least square method (LSM). It is not an unbiased estimate compared with LSM but explores regression coefficients closer to the truth
by abandoning part of the iterative precision. It controls the complexity of a linear model with L2
norm penalty term to the loss function so that makes the regression coefficients more robust. Generally,
it is one of the most popular linear regression regularization methods and the related definition is as
below.

$$\sum_{i=1}^{n}(y_i - \sum_{j=0}^{p} w_j x_{ij})^2 + k \sum_{j=0}^{p} w_j^2$$  \hspace{1cm} (2)

With a proper $k$ value, the regression model is balanced between variance and bias.

Deriving $w$ in the equation 2, $W$ can be worked out.

$$W = (X^T X + kI)^{-1} X^T Y$$  \hspace{1cm} (3)

where $I$ is the unit matrix.

In the ridge regression algorithm, different regression coefficients can be obtained by different $k$
values, and the variables are determined by ridge trace to reduce the correlation among variables. This
paper set $k=0.75$ in the related experiments.

For the multi-collinearity among variables in the hot metal dataset, the initial prediction values of
silicon contents are optimized based on the ridge regression algorithm. After ridge trace of $Si_{f_{ir-pre}}$ and 17 variables in Table 1, three attributes including $Si_{f_{ir-pre}}$, $Si(n-1)$ and $Si(n-2)$ are selected as the features of the model. At last, a regression model is established as below.

$$Y = b + w_1 * Si(n-1) + w_2 * Si(n-2) + w_3 * Si_{f_{ir-pre}}$$  \hspace{1cm} (4)

where $b$, $w_1$, $w_2$, $w_3$ are regression coefficients of the ridge regression model.

3.2 Classification model for furnace conditions based on random forest algorithm

The random forest algorithm (RF) consists of plentiful decision trees. For each decision tree, its
training set is a subset by randomly sampling with repeated return. Furthermore, only a part of features
is used to train this tree. The two strategies greatly improve the diversity of trees, which is effective to
prevent the over-fitting. RF includes random forest classification (RFC) and random forest regression
(RFR).

The main idea of RF is the combination of bagging and the decision tree algorithm. (CART
decision tree is applied in this article). Bagging, also known as self-sampling, is a technique for
repeatedly sampling from data with return. That is, bagging trains a base learner on basis of a
self-sampling set.

In addition, RF hardly requires input preparation since it can automatically complete the selection
of hidden features and provides a reliable selection index of feature importance. RF is insensitive to
multivariate linearity, and the related results of prediction are less affected by unbalanced datasets.
Therefore, it is suitable for analysis of the hot metal dataset with unbalanced samples.

In this paper, a random forest classification algorithm (RFC) is constructed as a classification
model based on the hot metal dataset. The 17 features in Table 1 and $Si_{f_{ir-pre}}$ serve as 18 input
variables to achieve the identification of furnace states. Specific steps are shown as follows.

Step 1. Sample randomly $k$ ($k=300$) self-service subsets from the hot metal dataset with return.

Step 2. For each self-service sample set, an independent decision tree is trained and grows down in
turn recursively. At each node of a tree, $m$ features ($m \leq 18$) are randomly chosen from 18 features, and
an optimal feature is selected from the $m$ features based on the Gini index for branch growth.
Step 3. Repeat the above steps for \( k \) times.

Step 4. The \( k \) decision trees are integrated into a model, named random forest, to classify the initial prediction values of silicon contents and predict the furnace conditions.

### 3.3 A compensation model

Due to imbalance of the sample distribution in the hot metal dataset, the predicted silicon contents in the low-silicon segment are generally larger than actual values and those in the high-silicon segment are usually smaller than true values.

To improve the prediction accuracy, we present a compensation model according to \( S_{f_{ir-pre}} \) as well as \( S_{f_{ir-pre}} \) and \( S_{secp_{-pre}} \).

\[
\begin{align*}
S_i &= \frac{S_{f_{ir-pre}} + S_{secp_{-pre}}}{2 + (0.32 - S_{f_{ir-pre}})} \quad , S_{class} = 0 \\
S_i &= 0.5* S_{f_{ir-pre}} + 0.5* S_{secp_{-pre}} \quad , S_{class} = 1 \\
S_i &= \frac{S_{f_{ir-pre}} + S_{secp_{-pre}}}{2 + (0.52 - S_{f_{ir-pre}})} \quad , S_{class} = 2
\end{align*}
\]

(5)

\( S_i \) are the final prediction values of silicon contents after compensation, 0.32 is the maximum endpoint value of the low-silicon segment and 0.52 is the minimum endpoint value of the high-silicon segment (the partition is based on artificial experience).

The compensation strategy pulls down the predicted values of the low-silicon segment and drives up those of high-silicon segment within a reasonable range. For instance, if \( S_{class} \) is 0, which is a low-silicon fluctuation, \( S_{f_{ir-pre}} \) should be appropriately reduced in compensation model, and if \( S_{class} \) is 1, which is a high-silicon fluctuation, \( S_{f_{ir-pre}} \) would be properly raised up.

The strategy combines outputs of three algorithms skillfully to optimize and compensate \( S_{f_{ir-pre}} \). It improves the prediction hit rate of silicon contents availably.

### 4. Experiments and analysis

In order to analyze the performance of the CAC model proposed in this paper, we utilize random forest regression (RFR) and BP neural network algorithm (BP) as the front-end model respectively. They are trained based on the same dataset and 17 features in Table 1 to generate \( S_{f_{ir-pre}} \). The testing samples are 244 sets, accounted for 20% of 1216 sets of samples.

According to actual experience of the BF system, we consider three parameters as evaluation indicators to analyze the prediction performance of the CAC model: \( a < 0.03 (H_{rate} (0.03/\%)) \) and mean square error (MSE).

1. Predicted hit rate \( (H_{rate}) \)

\[
H_{rate} = \frac{1}{n} \sum_{t=1}^{n} H_t \times 100\%
\]

(6)

where \( n \) represents the number of testing samples, \( H_t \) is the Heaviside function, which is defined as follows (here \( a \) is 0.03).

\[
H_t = \begin{cases} 
1 & , |y(t) - \hat{y}(t)| \leq a \\
0 & , else
\end{cases}
\]

(7)

2. Mean square error (MSE)
\[
MSE = \frac{1}{n} \sum_{t=1}^{n} (y(t) - \hat{y}(t))^2
\]  
(8)

where \(y(t)\) and \(\hat{y}(t)\) are the actual and predicted values of silicon contents respectively.

4.1 RFR as the front-end prediction model

In this section, we take random forest regression algorithm as the front-end model and use the CAC model to compensate \(S_{firs-pre}\). The procedures are as follows.

Step 1. Based on the principle of bagging, a decision tree is trained.

Step 2. RFR consists of \(k\) (\(k\) is generally greater than or equal to 200) CART decision trees and produces the \(S_{firs-pre}\).

Step 3. The ridge regression model is expected to optimize \(S_{firs-pre}\) and output \(S_{sec-pre}\). The regression coefficients are: \(b = 0.03\), \(w_1 = 0.058\), \(w_2 = 0.089\), \(w_3 = 0.793\), and the regression equation is shown as below.

\[
Y = 0.03 + 0.058*S_i(n - 1) + 0.089*S_i(n - 2) + 0.793*S_{firs-pre}
\]  
(9)

Step 4. RFC is applied to identify the furnace conditions and output \(S_{class}\).

Step 5. Based on \(S_{firs-pre}\) and \(S_{sec-pre}\) and \(S_{class}\), the final prediction values are outputted.

Fig.2.a depicts the comparisons of actual values with \(S_{firs-pre}\) and actual values with \(S_{sec-pre}\). Fig.2.b shows the errors between the truth and \(S_{firs-pre}\), as well as, the errors between the truth and \(S_{sec-pre}\).

Fig.2.a Comparisons of the true values and predicted values

Fig.2.b Errors of silicon contents between pre-compensated and compensated
As can be seen from Fig.2.a and Fig.2.b, the actual values of silicon contents show a sharp randomness and volatility. The prediction ability of the front-end prediction model grows down heavily when the furnace conditions fluctuate suddenly. After optimizing the silicon contents by the CAC model, the errors between the final predictions and actual values are significantly reduced, especially for those fluctuating furnace conditions.

Table 3 gives the prediction results in different furnace conditions before and after compensation. For the middle-silicon segment, $H_{rate}$ (0.03/%) after compensation is raised by nearly 10.73% and MSE is reduced by 3.85%. For the high-silicon and low-silicon segments, the improvements of prediction accuracy after compensation are more obvious than that in the middle-silicon segment. $H_{rate}$ (0.03/%) is raised up by 12.50% and MSE is declined by 3.77%. In general, $H_{rate}$ (0.03/%) after compensation is higher about 10.93% than before and MSE is improved by 3.45%.

| Interval       | Before compensation | After compensation | Before compensation | After compensation |
|---------------|---------------------|--------------------|---------------------|--------------------|
| (0.32,0.52)   | 50.90%              | 56.36%             | 0.0026              | 0.0025             |
| [0.0,0.32] $\cup$ (0.52,1] | 46.15%              | 51.92%             | 0.0053              | 0.0051             |
| [0,1]         | 48.77%              | 54.10%             | 0.0029              | 0.0028             |

### 4.2 BP neural network algorithm as the front-end prediction model

In this section, BP neural network algorithm (BP) is chosen as the front-end model to predict silicon contents and the overall process is like section 4.1.

Fig.3.a describes the comparisons of actual values with $S_{fir-pre}$ and actual values with $S_{sec-pre}$ when BP is used for the front-end model. Fig.3.b shows the errors between the actual values and $S_{fir-pre}$, as well as, the errors between the actual values and $S_{sec-pre}$.

According to Fig.3.a and Fig.3.b, it is concluded that the prediction results produced by the CAC model are more accurate and reliable than that of the front-end model.

![Fig.3.a Comparisons of the true values and predicted values](image)
The related results of the predicted silicon contents are displayed in Table 4. Firstly, for the middle-silicon segment, $H_{rate}$ (0.03/%) is increased by 10.34% and $MSE$ is improved by 3.77% after compensation. Secondly, for the low-silicon and high-silicon segments, $H_{rate}$ (0.03/%) after compensation is raised by 15.34% and $MSE$ is reduced by 6.90%. Thirdly, as a whole, $H_{rate}$ (0.03/%) rises around 14.58% and $MSE$ declines or so 2.63% after compensation.

| Interval          | $H_{rate}$ (0.03/%) | $MSE$    |
|-------------------|--------------------|----------|
| Before compensation |                   | After compensation | Before compensation | After compensation |
| [0.32,0.52]       | 45.90%             | 52.94%   | 0.0029          | 0.0027             |
| [0,0.32] U (0.52,1] | 19.64%             | 21.67%   | 0.0053          | 0.0051             |
| [0,1]             | 39.51%             | 45.27%   | 0.0038          | 0.0037             |

4.3 Result analysis

All the experimental results demonstrate that after compensation by CAC model, the predicted silicon contents are much closer to the truths, especially for the high and low-silicon segments. Though sometimes, the predicted errors of a small part of samples has a slight increase, the overall performance is greatly improved.

In addition, we find that the CAC model performs better with BP. For example, its $H_{rate}$ (0.03/%), as a whole, is higher 3.65% than those of RFR. The reason is that RFR, in terms of itself, is superior to BP in prediction for silicon contents.

5. Conclusion

In this paper, we put forward a classification and compensation model for silicon contents in BF hot metal. The related experimental results demonstrate that the classification and compensation model is effective in improving the stability, prediction accuracy and generalization performance. It provides a feasible way to establish a faithful prediction model for the BF ironmaking process.

Whereas, it is worth noting that the final curve of prediction values always closely follows the trend of the initial prediction curve, which restricts the optimization space. Next, we will insist to explore the fusion of this model with other front-end models, and facilitate the practicability and universality.
6. Acknowledgments

This work was financially supported by National Natural Science Foundation of China (Grant No. 61773406 and No. 61502538).

References

[1] C.H.Yan, L.Jian, J.M.Chen, et al. Data-driven Modeling and Predictive Algorithm for Complex Blast Furnace Ironmaking Process [J]. Acta Automatica Sinica, 2009, 35(6):725-730.
[2] X.Y.Liu, X.G.Liu, W.H.Wang, Application of Bayesian Network to Predicting Silicon Content in Hot Metal [J]. Iron and Steel, 2005, 40(3):17-20.
[3] Jian L, Gao C H, Li L, et al, Application of least squares support vector machines to predict the silicon content in blast furnace hot metal. ISIJ international, 2008, 48(11):1659–1661.
[4] P.Zhou, R.F.Li, D.W.Guo, et al.. Multi-output support vector regression modeling for multivariate molten iron quality indices in blast furnace ironmaking process [J]. Control Theory & Applications, 2016, 33(6):727-734.
[5] Z.H.Jiang, M.L.Dong, W.H.Gui, et al. Two-dimensional Prediction for Silicon Content of Hot Metal of Blast Furnace Based on Bootstrap [J]. Acta Automatica Sinica, 2016, 42(5):715-723.
[6] Khosravi A, Nahavandi S, Creighton D, Atiya A F. Comprehensive review of neural network-based prediction intervals and new advances[J]. IEEE Transactions on Neural Networks, 2011, 22(9): 1341–1356.
[7] Chen W, Wang B X, Han H L, Prediction and control for silicon content in pig iron of blast furnace by integrating artificial neural network with genetic algorithm. Ironmaking & Steelmaking, 2010, 37(6): 458–463.
[8] W.H.Wang, X.G.Liu, X.Y.Liu, Prediction model of silicon content in blast furnace hot metal based on random forests algorithm [J]. Metallurgical Industry Automation, 2014(5):33-38.
[9] N.Yang, The unique role of ridge regression analysis in solving multicollinearity problems [J]. Statistics and Decision, 2004(3):14-15.