Prediction Model of End-point Phosphorus Content in Consteel Electric Furnace Based on PCA-Extra Tree Model

Chao CHEN,1) Nan WANG,1,2,3)* and Min CHEN1)

1) School of Metallurgy, Northeastern University, Shenyang, Liaoning Province, 110819 China.
2) Institute for Frontier Technologies of Low-Carbon Steelmaking, Shenyang, Liaoning Province, 110819 China.
3) Liaoning Province Engineering Research Center for Technologies of Low-Carbon Steelmaking, Shenyang, Liaoning, 110819 China.

(Received on October 5, 2020; accepted on February 9, 2021; J-STAGE Advance published date: March 20, 2021)

According to the actual industrial data from a Consteel electric furnace, a prediction model based on the principal component analysis (PCA) and extremely randomized trees (Extra Tree model) is proposed for end-point phosphorus content. PCA is used to reduce the dimensionality of the input variable affecting the end-point phosphorus content and eliminate the collinearity among the input variables, and then the data transformed by PCA are used as input data for the established Extra tree model. Compared with other feature pre-processing methods, PCA method can greatly improve the regression prediction performance of the Extra Tree model. Finally, the validation by test set indicates that for the PCA-Extra Tree model, the hit rates of end-point phosphorus content are 98%, 96% and 89% with the prediction error range of ±0.005%, ±0.004% and ±0.003%, respectively. The combined PCA-Extra Tree model has achieved the effective prediction for end-point phosphorus content, and provided a good reference for the end-point control and judgment of Consteel electric furnace.

KEY WORDS: Consteel electric furnace; end-point phosphorus content; recursive feature elimination (RFE); principal component analysis (PCA); ensemble learning; Extra Tree model.

1. Introduction

Consteel electric furnace, as an important smelting equipment, is widely used in many iron and steel plants. Similar to BOF (Basic Oxygen Furnace), a series of complicated high-temperature physical and chemical reactions occur in Consteel electric furnace and thus it is of significance to control the smelting process stably, particularly for the end-point precise control. The end-point phosphorus (P) content is one of the main end-point targets in steelmaking process due to its significant influence on steel products.1,2) With the increasing demand for low and ultra-low phosphorus steels, many steel plants apply high-efficiency dephosphorization technology into converter.3,4) However, in view of the fact that most process parameters are correlative with each other, it is difficult to control the end-point phosphorus content in a high accuracy degree, and thus numerous researches have been carried out to realize the precise control of the end-point P content through the mathematical and intelligent models. Mathematical model is very useful for predicting the end-point phosphorus content in BOF;5–7) and nevertheless, mathematical models require a lot of theoretical assumptions and too many parameters are also included in the models, and the prediction accuracy of mathematical model is not always very satisfactory due to the strong nonlinear relationship between the end-point phosphorus content and the affecting factors. Therefore, many studies have proposed to predict the end-point phosphorus content by using intelligent model to increase the prediction precision. For examples, Li et al.8) established a prediction model of end-point phosphorus content for BOF steelmaking process based on Levenberg-Marquardt (LM) algorithm of BP neural network. Wang et al.9) used weighted K-means and group method of data handling (GMDH) neural network to predict the end-point phosphorus content in BOF. Liu et al.10) proposed a prediction method based on computer vision and general regression neural network to predict the end-point phosphorus of BOF. Liang et al.11) employed a two-step case-based reasoning method based on attributes reduction to predict the end-point phosphorus content of BOF. He et al.12) proposed a method by combining the PCA with BP neural network to improve the end-point P content in BOF process. Although the intelligent models mentioned above have achieved high prediction accuracy, and at present, many steelmaking plants are facing severe challenges and need to continuously improve the actual operation to

* Corresponding author: E-mail: wangnan@mail.neu.edu.cn

© 2021 The Iron and Steel Institute of Japan. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs license (https://creativecommons.org/licenses/by-nc-nd/4.0/).
produce the steel product with high quality. Thereby, it is needed to establish the higher precision prediction model of end-point phosphorus (P) content, aiming at providing a better reference for eligible end-point judgment of Consteel electric furnace to increase production efficiency. Over the years, the ensemble learning with superior ability of regression and classification has been widely used in various fields, but the application of end-point P control in steelmaking is rare. In this paper, by comparing the regression prediction ability of various ensemble learnings for the end-point P content in Consteel electric furnace, the Extra Tree model with the optimum prediction effect is selected as the end-point P prediction model. Furthermore, through comparing various feature pre-processing methods, the PCA method is employed to further improve the regression prediction ability for the Extra Tree model. In addition, the validation of the PCA-Extra Tree model is verified by the test set from industrial data.

2. Process Description and Data Collection

2.1. Steelmaking Process in Consteel Electric Furnace

The schematic diagram of Consteel electric furnace is shown in Fig. 1. During the smelting process, scrap and hot metal are the main raw materials, and the features of Consteel electric furnace are continuous addition of scrap and scrap preheating with smelting exhaust gas. Consteel electric furnace generally adopts bottom tapping and reserves a certain amount of molten steel. The scrap is added to Consteel electric furnace and mainly melted by molten metal. In addition, in order to enhance the quality of molten steel, Consteel electric furnace adopts submerged arc operation to reduce the direct contact between molten steel and atmosphere.

2.2. Collection of Main Data Parameters

The industrial data of the steelmaking process in Consteel electric furnace, about 2,373 heats, is collected from a steel plant in China. According to the dephosphorization thermodynamics and practical operation in Consteel electric furnace, the end-point phosphorus content is mainly determined by 17 process variables, including the chemical composition of hot metal, hot metal weight, scrap weight, lime weight, dolomite weight, carbon powder weight, smelting cycle, limestone weight, oxygen consumption, natural gas consumption, electricity consumption, end-point C content and end-point temperature. Because the industrial data contains numerous noise data which would disturb the model construction and result in incorrect results, box-plot method is employed in this paper to filter the exceptional data out, and for the preprocessed data, the main statistics information of the process parameters is shown in Table 1. The symbols from $X_1$ to $X_{17}$ present the input process variables, and $Y$ is the output variable. The Pearson correlation coefficients of these variables are shown in Fig. 2. Pearson correlation coefficient reflects the impact degree between two variables in a certain extent. According to Fig. 2, it is found that the Pearson correlation coefficients between scrap weight and hot metal weight, oxygen consumption, electricity consumption and carbon powder weight are $-0.96$, $-0.9$, $0.96$ and $0.69$ respectively, indicating that the addition of scrap weight would increase the electricity consumption and carbon powder weight, but would reduce the oxygen consumption. On the other hand, the Pearson correlation coefficients between hot metal weight and scrap weight, oxygen consumption, electricity consumption and carbon powder weight are $-0.96$, $0.9$, $-0.92$ and $-0.68$.

![Fig. 1. Schematic diagram of Consteel electric furnace. (Online version in color.)](image)

### Table 1. Statistics information of process variables in Consteel electric furnace.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Symbol</th>
<th>Unit</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot metal [Si] content</td>
<td>$X_1$</td>
<td>%</td>
<td>0.342</td>
<td>0.140</td>
<td>0.590</td>
<td>0.073</td>
</tr>
<tr>
<td>hot metal [Mn] content</td>
<td>$X_2$</td>
<td>%</td>
<td>0.126</td>
<td>0.080</td>
<td>0.190</td>
<td>0.022</td>
</tr>
<tr>
<td>hot metal [P] content</td>
<td>$X_3$</td>
<td>%</td>
<td>0.123</td>
<td>0.061</td>
<td>0.170</td>
<td>0.020</td>
</tr>
<tr>
<td>hot metal [S] content</td>
<td>$X_4$</td>
<td>%</td>
<td>0.024</td>
<td>0.006</td>
<td>0.048</td>
<td>0.007</td>
</tr>
<tr>
<td>hot metal [C] content</td>
<td>$X_5$</td>
<td>%</td>
<td>5.40</td>
<td>3.98</td>
<td>8.64</td>
<td>0.95</td>
</tr>
<tr>
<td>hot metal weight</td>
<td>$X_6$</td>
<td>t</td>
<td>66</td>
<td>33</td>
<td>131</td>
<td>23</td>
</tr>
<tr>
<td>scrap weight</td>
<td>$X_7$</td>
<td>t</td>
<td>44</td>
<td>11</td>
<td>75</td>
<td>22</td>
</tr>
<tr>
<td>lime weight</td>
<td>$X_8$</td>
<td>kg</td>
<td>4272</td>
<td>2295</td>
<td>6673</td>
<td>754</td>
</tr>
<tr>
<td>dolomite weight</td>
<td>$X_9$</td>
<td>kg</td>
<td>523</td>
<td>0</td>
<td>2958</td>
<td>569</td>
</tr>
<tr>
<td>carbon powder weight</td>
<td>$X_{10}$</td>
<td>kg</td>
<td>309</td>
<td>0</td>
<td>3703</td>
<td>417</td>
</tr>
<tr>
<td>smelting cycle</td>
<td>$X_{11}$</td>
<td>min</td>
<td>55</td>
<td>36</td>
<td>86</td>
<td>7</td>
</tr>
<tr>
<td>limestone weight</td>
<td>$X_{12}$</td>
<td>kg</td>
<td>202</td>
<td>0</td>
<td>4069</td>
<td>523</td>
</tr>
<tr>
<td>oxygen consumption</td>
<td>$X_{13}$</td>
<td>Nm³</td>
<td>3717</td>
<td>1608</td>
<td>6182</td>
<td>948</td>
</tr>
<tr>
<td>natural gas consumption</td>
<td>$X_{14}$</td>
<td>Nm³</td>
<td>179</td>
<td>39</td>
<td>359</td>
<td>60</td>
</tr>
<tr>
<td>electricity consumption</td>
<td>$X_{15}$</td>
<td>MWh</td>
<td>12</td>
<td>0</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>end-point C content</td>
<td>$X_{16}$</td>
<td>%</td>
<td>0.39</td>
<td>0.03</td>
<td>0.72</td>
<td>0.16</td>
</tr>
<tr>
<td>end-point temperature</td>
<td>$X_{17}$</td>
<td>°C</td>
<td>1619</td>
<td>1585</td>
<td>1652</td>
<td>11</td>
</tr>
<tr>
<td>end-point P content</td>
<td>$Y$</td>
<td>%</td>
<td>0.012</td>
<td>0.004</td>
<td>0.018</td>
<td>0.003</td>
</tr>
</tbody>
</table>
respectively, indicating that the addition of hot metal weight would reduce the electricity consumption and carbon powder weight, but would increase the oxygen consumption. Based on the Pearson correlation coefficient, it can also be noted that the effects of scrap weight and hot metal weight on the oxygen consumption, electricity consumption and carbon powder weight are opposite. Obviously, the variation of scrap ratio (ratio of scrap weight and hot metal weight) would greatly cause the complex change of other smelting parameters, which would increase the smelting burden. Meanwhile, there also exists a complex nonlinear relationship between the input and output variables, which interferes with the establishment of prediction model.

3. Prediction by Various Regression Models

According to the investigation on the above industrial data, it can be concluded that the industrial data of Consteel electric furnace shows highly nonlinear relationships between the input and output variables, and meanwhile the complex relationship including linear and nonlinear also exists among the input variables. Therefore, more complex prediction models are required to fit the relationship between input and output variables. In order to obtain an optimum prediction result of end-point P content of Consteel electric furnace, various regression models are selected in this paper to determine the most appropriate prediction model, including multiple linear regression (MLR), Decision Tree (DT), Random Forest (RF), AdaBoost (AB), XGBoost (XB) and Extra Tree (ET).\(^{16-21}\) Among these regression models, Random Forest, AdaBoost, XGBoost and Extra Tree models are the ensemble learning models, and one of the most popular machine learning ideas improves the model stability and accuracy by combining multiple weak classifiers into a strong classifier.\(^ {22,23}\) Furthermore, all the machine learning models adopted in this paper are dealt with by the scikit-learn of python library, which is an open machine learning package under the Berkeley Software Distribution (BSD) license.\(^ {24}\) In order to better compare the regression capability among various models, the performances of regression models are evaluated respectively by the determination coefficient (R\(^2\)) defined in Eq. (1), mean absolute error (MAE) defined in Eq. (2), mean square error (MSE) defined in Eq. (3) and root mean square error (RMSE) defined in Eq. (4). The value of R\(^2\) is closer to the value of 1, the better model regression performance is, whereas the smaller the values of MAE, MSE and RMSE are, the better model regression performance is. In addition, in order to obtain a better result, the method of 5-fold cross-validation is employed and the results are shown in Table 2.

**Determination coefficient**

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

**Mean absolute error**

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

**Mean square error**

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)
\]

**Root mean square error**

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (4)
\]

where, \(y_i\) is the actual data, \(\hat{y}_i\) is the predicted value, \(\bar{y}\) is the mean of actual data.

In Table 2, it is obvious that the values of R\(^2\), MAE, MSE and RMSE for RF, AB, XB and ET models are higher than those of MLR and DT, indicating that the ensemble learning models have a higher comprehensive regression ability than those of MLR and Decision Tree established by the industrial data in Consteel electric furnace. Among the ensemble learning models, the R\(^2\) values of RF and ET are higher than those of AB and XB, suggesting that RF and ET have the better fitting ability than AB and XB, and meanwhile the value of MAE, MSE and RMSE of RF and ET are also higher than those of AB and XB, indicating that the deviation between the predicted value and the real value is reduced effectively by RF and ET. Although RF, AB, XB and ET belong to the ensemble learning models, the basic principle of these algorithms is different, which is the main reason for the different regression performances. In brief, AB and XB models are based on the boosting tree. For the all generated weak models, the structures of these models are similar to series structure, and every weak model is to learn the prediction residual from the previous weak model. Thus, every weak model generated by AB and XB is relevant. Furthermore, RF and ET are based on the bootstrap aggregating (bagging), and their model structures are similar to parallel structure, namely every weak model generated by AB and XB is irrelevant. The final result of RF and ET is the combination of the results from every weak model. In addition, the Extra tree model possesses the better regression performance than Random Forest, shown in Table 2. The Extra Tree model was proposed first by P. Geurts, D. Ernst and L. Wehenkel.\(^ {25}\) Compared with Random Forest, the main characteristic of Extra Tree model is more random selection of input variable and more random selection of
variable split point, which is one of the main reasons why the regression ability of Extra Tree model is better than Random Forest. The main parameters of Extra Tree model from scikit-learn selected in this paper are presented in Table 3. Among all the parameters of Extra Tree, n_estimators is crucially important, which is directly related to the model accuracy. Therefore, it is need to appropriately adjust the values of n_estimators to ensure the accuracy of the Extra Tree model. Obviously, the advantages of Extra tree model are concise parameters and higher prediction performance.

4. Model Improvement by Feature Pre-processing

The actual industrial data contains a large amount of redundant, collinear and overlapping information, and thus using all the industrial data to build a model is complex and time-consuming. Meanwhile, the noise and useless information also affect the accuracy and robustness of the model. In the above section, it is obvious that the Extra tree model possesses concise parameters and higher prediction performance, but up to now, the method of adjusting the n_estimators to improve the prediction performance of Extra tree model is insufficient. In order to further improve model accuracy, two methods of feature pre-processing included RFE (recursive feature elimination) and PCA (principal component analysis) are employed to improve the performance of Extra Tree in this section.

4.1. Model Improvement by RFE

To further improve the regression effect of Extra Tree model, the method of feature selection is employed. In the field of machine learning, removing some features which are redundant or irrelevant features can improve the model effect effectively. Therefore, proper feature selection method is very important. Feature selection methods can be divided into three categories, including filter, wrapped, and embedded. The filter feature selection method is to use statistical method to evaluate each variable and then to eliminate the unqualified variables from the input variables directly, which mainly involves removing features with low variance, univariate feature selection, chi-square test, Pearson correlation and mutual information. The wrapped feature selection method is to select a base model firstly, and then continuously to increase or decrease the number of input variables, according to the fitting effect of the selected model. The recursive feature elimination (RFE) and stability selection are the main wrapped feature selection methods. The embedded feature selection method is a feature selection which generates the important input variable during the modeling process, mainly including randomized sparse models, L1-based feature selection and L2-based feature selection.

In section 2, the industrial data investigation by Pearson correlation coefficient and data statistics is carried out. It is obvious that there is a strong linear relationship among the input variables of scrap weight, hot metal weight, oxygen consumption, electricity consumption and carbon powder weight. Therefore, according to the characteristics of industrial data in Consteel electric furnace, the wrapped feature selection method is employed to select the optimum feature, particularly the specific method is the combination of the RFE and various tree-model to select the important feature, and the validity of feature selection is verified by cross-validations method. 5-fold cross-validation and R² are used as the evaluation criterion. In Fig. 3, it can be noted that the combination of RFE and XGBoost (XB-RFE) possesses the better improvement ability for the Extra Tree when the number of feature is set as 16.

To further test the effect of RFE, we also employ the combination of RFE and various tree-models to improve the model performance, such as MLR, DT, RF and AB and XB, and the results are shown in Figs. 4(a), 4(b), 4(c), 4(d) and 4(e) respectively. As shown in Fig. 4, it is obvious that for the DT, the combination of RFE and Adaboost (AB-RFE) possesses the better ability when the number of feature is set as 5; and for the RF, the combination of RFE and Decision Tree (DT-RFE) possesses the better ability when the number of feature is set as 15; and for the AB, the combination of RFE and Decision Tree (DT-RFE) possesses the better ability when the number of feature is set as 15; and for the XB, the combination of RFE and Decision Tree (DT-RFE) possesses the better ability when the number of feature is set as 14. In brief, for the optimum combination of RFE and tree-model, the specific statistical information is shown in Table 4 (ini.
represents the initial regression ability of model and rfe represents the regression ability of model processed by RFE). It is obvious that the models of DT, AB and ET can be effectively improved by RFE, and particularly, the combination of XGBoost and RFE (XB-RFE) can improve the values of $R^2$, MAE, MSE and RMSE of Extra Tree from $0.61 \times 10^{-2}$, $0.04 \times 10^{-4}$ and $0.19 \times 10^{-2}$ to $0.63 \times 10^{-2}$, $0.03 \times 10^{-4}$ and $0.18 \times 10^{-2}$ respectively.

4.2. Model Improvement by PCA

In section 2, Pearson correlation coefficient is employed to investigate the relationships among the input variables. It is obvious that there is a collinearity problem among the hot metal weight, scrap weight, oxygen consumption, natural gas consumption, electricity consumption and carbon powder. For the collinearity problem, the main effect is to make the regression coefficients unreliable and increase the variance of regression prediction. Therefore, it is important and significant to deal with the collinearity problem among the input variables in Consteel electric furnace, and the principal component analysis (PCA) is employed to reduce the impact of collinearity problem. Principal component analysis is a technique of data dimension reduction. It replaces a large number of original interrelated variables by a smaller number of uncorrelated principal components (namely derived variables), while retaining as much as possible of internal information present in data set of original variables\(^{29,30}\) as shown in Eq. (5).

$$Y_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \ldots + a_{1m}X_m$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \ldots + a_{2m}X_m$$

$$\ldots$$

$$Y_n = a_{n1}X_1 + a_{n2}X_2 + a_{n3}X_3 + \ldots + a_{nm}X_m$$

where, $Y_1$, $Y_2$, ..., $Y_n$ are the transformed data by PCA; $X_1$, $X_2$, $X_3$, ..., $X_m$ are the original input variables, $a$ is the transformed coefficient, $m$ is the dimension of original data, subscript $n$ is the dimension of transformed data, namely number of principle component set by human.

According to Eq. (5), it is found that the original data is linearly transformed to the new data through PCA. Meanwhile, a group of input variables from the original data with high linear relationship is transformed to another group of variables with non-linear relationship. Then the transformed variable is input to the Extra Tree model to predict the end-point P content. Setting the optimum number of principle component is the key mission of the PCA-Extra Tree model. In order to detect the optimal
In order to better verify the validation of PCA, the test by other regression models including MLR, DT, RF, AB and XB are also carried out. The different test results of these models with different number of principal component is shown in Fig. 6, and the optimum number of principal component of MLR, DT, RF, AB and XB is set as 16, 7, 5, 16 and 14, respectively. Under the optimum number setting of principal component, the specific information is shown in Table 5 (ini. represents the initial regression ability of model and pca represent the regression ability of model after PCA). It is obvious that the models of DT, RF, XB and ET can be effectively improved by PCA. In particular, the $R^2$ value of ET is improved from 0.61 to 0.68, which is better than that of Extra Tree improved by XB-RFE. Meanwhile, the results of MAE, MSE and RMSE of PCA-Extra Tree are also better than the corresponding evaluation standard of Extra Tree improved by XB-RFE. That is to say, the feature pre-processing method of RFE and PCA model can all improve the performance of Extra Tree, but by contrast, the improvement effect of RFE is not as significant as PCA for the Extra Tree. Thus, the combination of PCA and Extra Tree is employed as the final prediction model for the endpoint P content from the Consteel electric furnace.

4.3. Modelling Verification

Through the above discussion, it is obvious that PCA method can better improve the regression ability of Extra Tree model compared with the feature selection method of RFE. Therefore, according to the data characteristics in Consteel electric furnace, the principal component analysis (PCA) is an effective method of feature pre-processing to improve the regression prediction accuracy of Extra Tree model. In order to test the validation of PCA-Extra Tree, the whole data is randomly segmented to training set and test set with ratio of 7:3. The training set is used to establish the PCA-Extra Tree model, and then the test set is input into the established model to verify its regression ability. Meanwhile, the models of MRL, DT, RF, AB, XB, Extra Tree and Extra Tree improved by XB-RFE (XB-RFE-Extra Tree) are also tested to make a comparison. The specific comparison is shown in Table 6.

### Table 6. Comparison of model accuracy within different error range.

<table>
<thead>
<tr>
<th>Model</th>
<th>Prediction error range (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>±0.0005</td>
</tr>
<tr>
<td>MLR</td>
<td>94%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>92%</td>
</tr>
<tr>
<td>RandomForest</td>
<td>98%</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>96%</td>
</tr>
<tr>
<td>XGBoost</td>
<td>97%</td>
</tr>
<tr>
<td>Extra Tree</td>
<td>98%</td>
</tr>
<tr>
<td>XB-RFE-Extra Tree</td>
<td>98%</td>
</tr>
<tr>
<td>PCA-Extra Tree</td>
<td>98%</td>
</tr>
</tbody>
</table>

### Table 5. Comparison of various regression models processed by PCA.

<table>
<thead>
<tr>
<th>Evaluation standard</th>
<th>MLR</th>
<th>DT</th>
<th>RF</th>
<th>AB</th>
<th>XB</th>
<th>ET</th>
</tr>
</thead>
<tbody>
<tr>
<td>ini. pca</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pca</td>
<td>0.21</td>
<td>0.22</td>
<td>0.15</td>
<td>0.21</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ini. pca</td>
<td>0.23</td>
<td>0.24</td>
<td>0.51</td>
<td>0.53</td>
<td>0.61</td>
<td>0.68</td>
</tr>
<tr>
<td>MAE (10⁻³)</td>
<td>0.22</td>
<td>0.21</td>
<td>0.22</td>
<td>0.21</td>
<td>0.16</td>
<td>0.14</td>
</tr>
<tr>
<td>MSE (10⁻⁴)</td>
<td>0.07</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>RMSE (10⁻²)</td>
<td>0.27</td>
<td>0.26</td>
<td>0.28</td>
<td>0.27</td>
<td>0.21</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Fig. 7. Comparison of prediction value and actual values by PCA-Extra Tree. (Online version in color.)

and it is obvious that the PCA-Extra Tree model possesses higher prediction accuracy than the other models. In particular, in the error ranges of ±0.004%, ±0.003%, ±0.002% and ±0.001%, the prediction accuracy of PCA-Extra Tree model achieves 96%, 89%, 82% and 60% respectively. As illustrated in Fig. 7, the prediction values of end-point phosphorus content possesses a good consistent with the actual values of end-point phosphorus content, which proves that the PCA-Extra Tree model is a regression prediction model with high accuracy for the end-point P content of Consteel electric furnace.

5. Conclusion

Through the collection and analysis of actual industrial data in Consteel electric furnace, a regression model combined PCA and Extra Tree model is established. The following conclusions can be drawn.

1) By using Pearson correlation coefficient to analyze the industrial data in Consteel electric furnace, the relationship between input and output variables is found to be highly non-linear, and a complex relationship also exists among the input variables. In particular, there is collinearity problem among the hot metal weight, scrap weight, oxygen consumption, natural gas consumption, electricity consumption and carbon powder, whereas there exists a non-linear relationship among the other input variables.

2) Through comparing various ensemble learning models, it is obvious that the Random Forest and Extra Tree models possess better prediction performance than Adaboost and XGBoost for the end-point P content of Consteel electric furnace. Meanwhile, the Extra Tree model possesses the optimum regression prediction ability, and the evaluation standard including $R^2$, MAE, MSE and RMSE achieve 0.61, $0.14 \times 10^{-2}$, $0.04 \times 10^{-4}$ and $0.19 \times 10^{-2}$ respectively, indicating that the Extra Tree model possesses better fitting ability for the production data in Consteel electric furnace.

3) By combining RFE algorithm and tree-based algorithm for feature selection, it is found that the combination of XGBoost and RFE can improve the values of $R^2$, MAE and RMSE of Extra Tree from $0.61$, $0.14 \times 10^{-2}$ and $0.19 \times 10^{-2}$ to $0.63$, $0.13 \times 10^{-2}$ and $0.18 \times 10^{-2}$, respectively.

4) Through the prediction error statistic of test set, the hit rates of end-point P content predicted by PCA-Extra Tree model are 98%, 96% and 89% respectively with the prediction errors ranges of ±0.005%, ±0.004% and ±0.003%, which greatly improve the prediction accuracy of low error range than the Extra Tree model. In actual application, the PCA-Extra Tree model could achieve accurate prediction for the end-point P content and provide a good reference for end-point judgment in Consteel electric furnace.

Acknowledgements

This work was supported by the National Key R&D Program of China [Grant numbers 2017YFB0304201, 2017YFB0304203 and 2016YFB0300602].

REFERENCES