Send Orders for Reprints to reprints@benthamscience.ae

Feature Selection and Long-Term Modeling for the Blast Furnace Iron-making Process Based on Random Forests

W.H. Wang[1,*], J.Q. Liu[2] and X.Y. Liu[3]

[1]*Basic Department, Zhejiang University of Water Resources and Electric Power, Hangzhou, China*

[2]*Zhejiang Gongshang University, Hangzhou, China*

[3]*China Jiliang University, Hangzhou, China*

Abstract: So far, the accurate modeling and control of blast furnace iron-making process (BFIP) is still an open problem due to its excessive complexity. Aiming at the issue of long time-delay and strong cross-coupling characteristics of BFIP, the random forests (RF) algorithm is introduced for predicting the silicon content in hot metal, which is the most key indicator of inner state of blast furnace. In the proposed model, both short and long-term BFIP features are adopted as inputs, without variable pre-selecting, to modeling the long-term dynamics of BFIP. Simulation results show that the RF algorithm can successfully identify the importance of different features (the latest silicon content in hot metal obtains the largest value of importance), can effectively decrease the effect of the redundancy and cross-coupling among variables. The RF model also can achieve similar or better prediction performance compared with support vector machines (SVM), which indicates that it is potential to modeling such BFIP-type complex industrial process using RF algorithm.

Keywords: Decision tree, Predict, Silicon content in hot metal, Random forests, Support vector machine.

1. INTRODUCTION

Blast furnace iron-making process (BFIP) is an extremely complicated nonlinear industrial process with intense stochastic character. There are 108 principal chemical reactions happening in the blast furnace simultaneously under high temperature and pressure. Because of the difficult measurement conditions, the silicon content in hot metal ([Si]), which has significantly positive correlation with the furnace temperature, is often considered as the alternative indictor of the thermal state of blast furnace (BF). To better control the iron-making process and produce iron with higher quality, the silicon content in hot metal should be controlled strictly at a certain degree. For this purpose, to predict the silicon content in hot metal has been taken as one of the most challenging issues in all the operational problems of BF. However, the inner reaction mechanisms of iron-making blast furnace are too complicated for the human being to understand. The BFIP owns most of the features of complex industrial process, such as the large scale time delay, variable-coupling, high temperature and pressure, multiphase simultaneous momentum, and so on. Up to now, the revealed mechanism of BFIP is too inadequate to construct a mechanism model to control the BF successfully. In the past decades, just because of this, many data-driven-algorithm based models have been contributed to it, such as auto-regression [1], neural networks [2, 3], fuzzy logic [4], partial least squares [5-6], grey system theory [7-8], support vector machines [9], *et al*., and have shown good performance to some degree. Many tools above serve as universal nonlinear approximators. Hence these data-driven models can provide sufficient potential to learn the data selected from BFIP. Nevertheless, their prediction performance in practice can not meet the requirements of real applications. Herein the main reason may lie in the long time-delay, strong cross-coupling, and redundancy characteristics of BFIP.

Random forests (RF) algorithm, proposed by Breiman L [10] in 2001, belongs to the nonparametric tree-based ensemble learning methods and has been widely applied to data mining and machine learning fields. RF has the advantages of both the adaptive nearest neighbors and the bagging algorithms [11] for effective data adaptive inference, and has been used extensively in different applications, including identification of DNA-binding proteins [12], prediction [13], recognition of handwritten digits [14], and many others. More specifically, the greedy one-step-at-a-time node splitting method introduces into RF algorithm the regularization for effective analysis in the so-called "large p, small n" or "high-dimension, low-sample-size" problems and the "grouping property" of trees [15] brings to RF the capability of adeptly handling the correlation and cross-coupling among variables. Furthermore, the variable importance measures that are obtained by RF algorithm provide a good criterion for ranking and selecting variables. So, all of the above properties of RF make it an potential tool for modeling BFIP with all sufficiently long-term BFIP features regardless of the correlation and interaction among them. In this paper, RF algorithm is applied to evaluating the im-

*Address correspondence to this author at the Basic Department, Zhejiang University of Water Resources and Electric Power, Hangzhou, China; E-mail: zjuwangwh@163.com

1874-4443/15 2015 Bentham Open

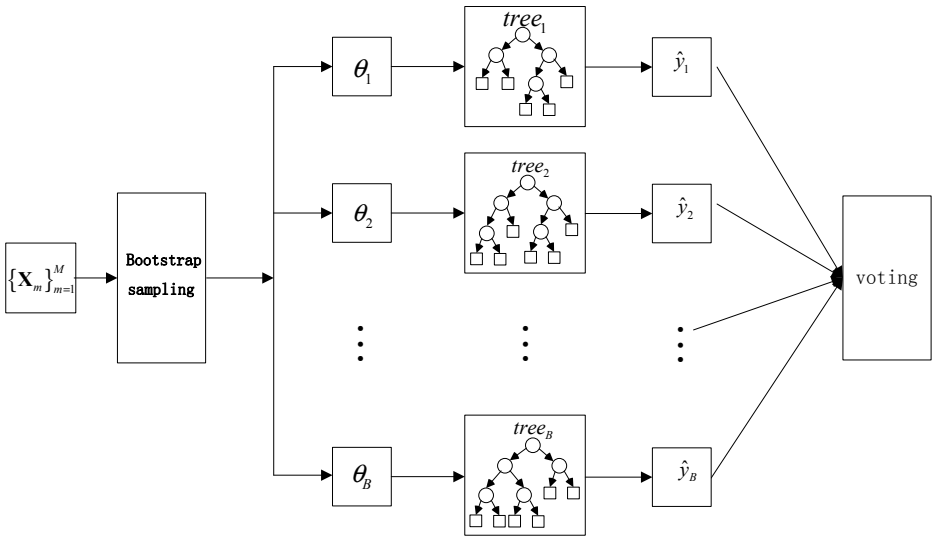

Fig. (1). A general architecture of random forests.

portance of input features and modeling the long-term dynamic of BFIP.

The rest of this paper is organized as follows. In section 2, the classification trees and RF techniques are reviewed briefly. Section 3 describes the data used and the experiment setup. The simulation results and comparison analysis are given in section 4, and section 5 concludes the paper.

2. RANDOM FORESTS

Given a set of training data points $\mathcal{X}_t = \{(\mathbf{X}_m, y_m), m = 1,2,\cdots,M\}$, where $\mathbf{X}_m$ is an input observation and y_m is a predictor output [16]. A weak learner $f(\mathbf{X},\mathrm{X}_t)$ with low bias and high variance can be created using the training set X_t. By randomly sampling from the set X_t, a collection of weak learners $f(\mathbf{X},\mathcal{X}_t,\theta_k)$ can be created, where $f(\mathbf{X},\mathcal{X}_t,\theta_k)$ is the *k*th weak learner and θ_k are independent identically distributed random vectors generated by applying bootstrap sampling. RF can be taken as an ensemble of unpruned regression or classification trees. Fig. (**1**) presents the structure of an RF model, where *B* denotes the number of trees. It can be shown that as the number of trees *B* increases, the out-of-bag (OOB) data set error rates converge, which indicates that over-fitting phenomena can be avoided in large RF [10]. It is crucial for good modeling performance to ensure the low bias and low correlation properties in RF, which can be guaranteed by growing the trees to maximum depth and applying the randomization strategy as follows [16]:

(1). Each tree is grown on the bootstrap sample set which includes about two-thirds of the original training data.

(2). At each node of the tree, *n* variables are randomly selected out of all the *N* variables. Here *n* is often initialized by $n = \log_2(N)+1$ or $n = \sqrt{N}$ and then is adjusted to reach the minimum error for the left training data (OOB data).

(3). At each node of the tree, only the unique variable that provides the best split performance is used out of the *n* candidate ones.

The estimation of variable importance, which is a very attractive feature offered by RF, can also be conducted based on the OOB data. Variable importance measure in RF is defined as the average decrease in classification accuracy on the OOB data. Specifically, when estimating the importance of the variable x_j, the number of correct classifications R^{oob} is counted for every tree based on the OOB data set, and then this number is recounted after permuting the values of variable x_j randomly. Therefore, the importance measure $\overline{D_j}$ for variable x_j is given by the average of these two numbers of all the trees in the RF model.

Researches show that random forests algorithm is one of the most accurate learning algorithms. It can run efficiently on large data sets, handle thousands of inputs without feature extraction and feature selection, and estimate the variable importance. And it is an effective method to tackle with outliers and missing data. The main motivation of this paper is to take the above advantages of random forests to model the complex dynamic of BFIP.

3. CASE STUDY

3.1. Key Parameter Selecting

As a very complicated and highly coupled nonlinear system, BFIP contains many factors influencing silicon content in BF, which can be categorized as either control parameters or state parameters. Control parameters mainly include charging materials properties, air, oxygen rich, and so on. State parameters mainly contain feeding speed, composition and slag, hot air index *et al*. For the current study, based on the actual production situations of No.1 BF at Laiwu Iron and Steel Group Co., 6 key parameters are selected as input

variables, which contain three control parameters: coal injection (CJ), blast temperature (BT), blast quantity (BQ), and three state parameters: difference between theoretical value, actual value of iron output (MFe), gas permeability (GP), coal rate (CR). Additionally, noting that BF is a large inertial system, the latest silicon content in hot metal (denoted as $[Si]_{n-1}$) has also been adopted as another key parameter. In total, 7 different key parameters are selected for the following BFIP modeling, which can be denoted as

$$X^0 = ([Si]_{n-1}, CJ, BT, BQ, MFe, GP, CR)$$

3.2. Long-Term Feature Introduction for the Large Time-Delay in BFIP

The current thermal status of BF is determined not only by the melting parameters acquired currently but also by the ones selected quite long time before. In fact, large time delay is a ubiquitous phenomenon for iron-making parameters. But how to estimate the accurate values of time delays of different parameters has still not been solved till now [17]. To address this problem, we will introduce long-term features into the BFIP modeling process.

BF is a huge industrial reactor, in which complex physical and chemical reactions would happen. The raw materials of BF are consisted of iron ore, coke, fluxes, and so on. They are filled into the top of BF by layers. At the same time, the tuyeres blow in the preheated compressed air and auxiliary fuels from the bottom of BF. As various materials move down through BF, the hot ascending gases will heat them progressively and then the carbon monoxide in hot gases transformed them to molten hot metal and slag, which would continuously trickle down into the hearth and gather at the bottom and top of the hearth respectively due to the different densities. Finally, the liquid hot metal and slag are tapped at regular intervals through tapholes. In general the whole iron-making blast furnace process will last 6-8 h.

During the whole BFIP process (6-8 h), the 7 key parame ters selected above will take several different values as the ti me changes. However every value may be crucial in determi ning hot metal quality. So the long-term modeling strategy is necessary for modeling BFIP here. Note that the tapping int erval of the BF concerned in this paper is about 2 h. Consequ ently, all the parameter information during the latest 4 succes sive tapping process will be adopted simultaneously as the m odel inputs. As shown later in the paper, the redundancy and correlation of all the involved variables can be effectively de alt with by RF algorithm. Consequently, there are totally 28 input variables which are denoted as

$$\begin{aligned} X :&= ([Si]_{n-1}, CJ, BT, BQ, MFe, GP, CR, \cdots, \\ &[Si]_{n-4}, CJ_{n-3}, BT_{n-3}, BQ_{n-3}, MFe_{n-3}, GP_{n-3}, CR_{n-3}) \\ &= (x_1, x_2, x_3, x_4, x_5, x_6, x_7, \cdots, x_{22}, x_{23}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}) \end{aligned} \quad (1)$$

The new input feature set X contains necessary long-term attributes about the corresponding iron-making process, based on which the essential long-term dynamic model of BFIP could be explored. Although there may exist correlation and interaction among these 28 variables, RF algorithm provides an effective method for tackling these problems.

Noting the huge diversity of different variables in order of magnitude, a preprocessing of the original dimensional variables should be undertaken before the model is developed. In this paper, both the input and output variables were normalized by the following equation:

$$x_{mj}^* = \frac{x_{mj} - \overline{x_j}}{\delta(x_j)}, \quad m = 1,2,\cdots,M, \quad j = 1,2,\cdots,N \quad (2)$$

where $\overline{x_j} = \frac{1}{M}\sum_{m=1}^{M} x_{mj}$; $\delta(x_j) = \frac{1}{M-1}\sum_{m=1}^{M}(x_{mj} - \overline{x_j})^2$ are mean and standard deviation of all data for the jth variable. After standardization treatment $X^* = (x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*, x_7^*, \cdots, x_{22}^*, x_{23}^*, x_{24}^*, x_{25}^*, x_{26}^*, x_{27}^*, x_{28}^*)$ are used as input vectors of RF.

3.3. Model Criteria

To verify the performance of the proposed models comprehensively, four criteria are considered in this paper, namely, Hit-rate (the rate of hitting the target), RMSE (root mean square error), MAE (mean-absolute error) and CC (correlation of coefficient) to evaluate the accuracy of the model in metallurgical field. The Hit-rate is defined as follows:

$$\text{Hit-rate} = \frac{1}{m}\left(\sum_{k=1}^{m} H_k\right) \times 100\%, \quad (3)$$

$$H_k = \begin{cases} 1 & |y(k) - \hat{y}(k)| < 0.1 \\ 0 & otherwise \end{cases}, \quad (4)$$

where m is the size of the testing samples, $y(k)$ is the observed value at instance k, and $\hat{y}(k)$ is the corresponding predicted value.

The RMSE is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(y(k) - \hat{y}(k))^2} \quad (5)$$

The smaller RMSE, the better predictive accuracy.

And the MAE and CC are as follows respectively:

$$\text{MAE} = \frac{1}{m}\sum_{k=1}^{m}\left|y(k) - \hat{y}(k)\right| \quad (6)$$

$$\text{CC} = \frac{\frac{1}{m}\sum_{k=1}^{m}(y(k) - \bar{y})(\hat{y}(k) - \bar{\hat{y}})}{\sigma(y)\sigma(\hat{y})} \quad (7)$$

where $\bar{y}, \bar{\hat{y}}$ are the mean of the observed values and the forecasted values respectively, and $\sigma(y), \sigma(\hat{y})$ represent their standard deviation.

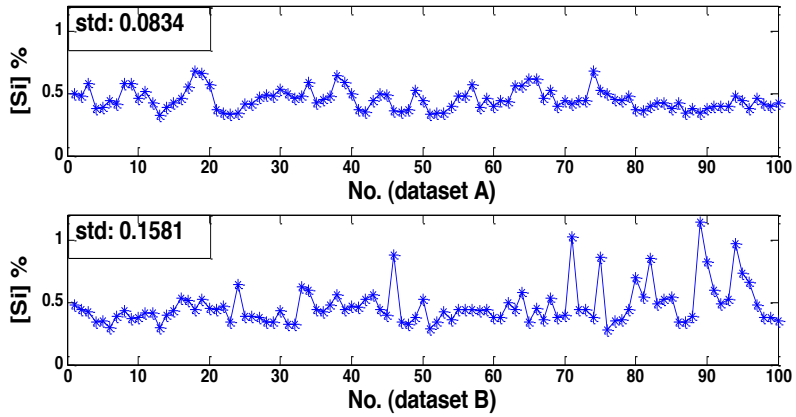

Fig. (2). Two groups of test sequences of silicon content in hot metal.

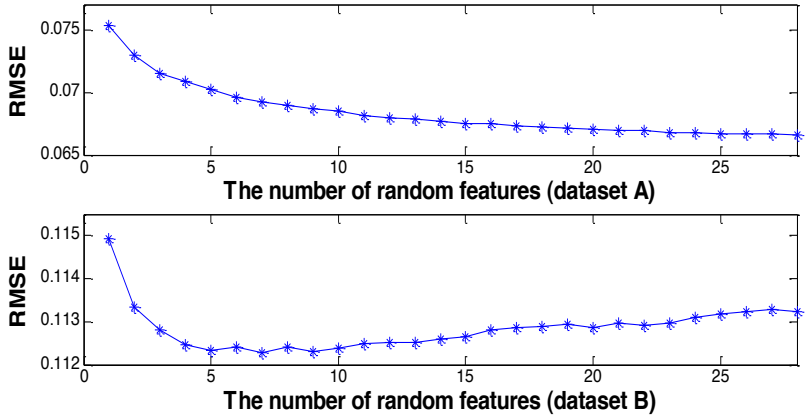

Fig. (3). The RMSE of two training data with different n .

4. RESULTS AND DISCUSSION

4.1. Determination of the RF Structure

There are two essential parameters in RF, namely B (the number of decision tree) and n (the number of random-selected variables to split the node of the tree). B is generally simple to select, since it is only needed to select the one that is big enough to ensure the convergence of RF. In this paper, B is selected as 1000. n is the only parameter which should be selected experimentally to improve the performance of RF algorithm. According to the number of the input variables, the maximum of n should be 28. Since the appropriate number of n is not known, our RF model was trained independently by adding the number of n from 1 to 28 to estimate the variable importance and predict the silicon content.

To evaluate the performance of RF more comprehensively, support vector machine (SVM), which is considered as one of the state-of-the-art machine learning methods so far, is also used to model the BFIP dynamic here. In this study, the kernel function used in SVM is the radial basis function (RBF) $K(\mathbf{x},\mathbf{x}') = \exp\left(-\left\|\mathbf{x}-\mathbf{x}'\right\|^2 / \sigma\right)$. As proposed by Hsu and Lin [18], the optimal values of γ and σ are obtained by 10-folder cross-validation and grid search using the different combinations of $\gamma = [2^{-2}, 2^{-1}, \cdots, 2^{12}]$, $\sigma = [2^{-4}, 2^{-3}, \cdots, 2^{10}]$.

4.2. Prediction Results Comparisons

To make the experimental results of the RF algorithm more convincing, two different groups of BFIP data (denoted by dataset A and dataset B), which own different statistical characteristics from each other, are used for establishing the RF models. The sizes of the two groups of samples are both 300, and herein the first 200 samples are used as training dataset and the last 100 samples as test dataset. Fig. (**2**) shows the corresponding time series of silicon content of the two groups of test samples. As can be seen, the distribution of first silicon content series is relatively steady, while the other series is more volatile. More specifically, the standard deviation of the second silicon content series is 0.1581, which is almost as twice as that of the first sequence (0.0834).

Firstly, the experiments are conducted on the first 200 samples to train the RF models by adding the number of n from 1 to 28, and then determine the optimal number of n and calculate the importance measure for variables accordingly. In this section, the results of the RF algorithms are the average ones computed from 20 trials.

Fig. (**3**) shows the RMSE of two training datasets with different value of n respectively. From Fig. (**3**), we can see that the optimal value of n corresponding to the minimum RMSE of the first training samples is 28 and the optimal value of n of the other training samples is 7. Therefore, we obtain the optimal RF configuration of two groups training

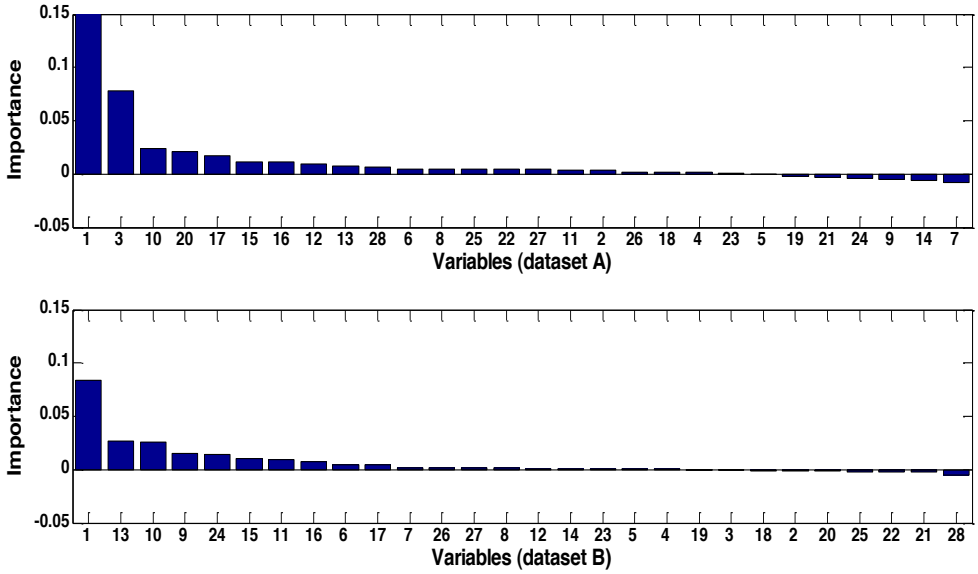

Fig. (4). The average value of variables importance for two training data.

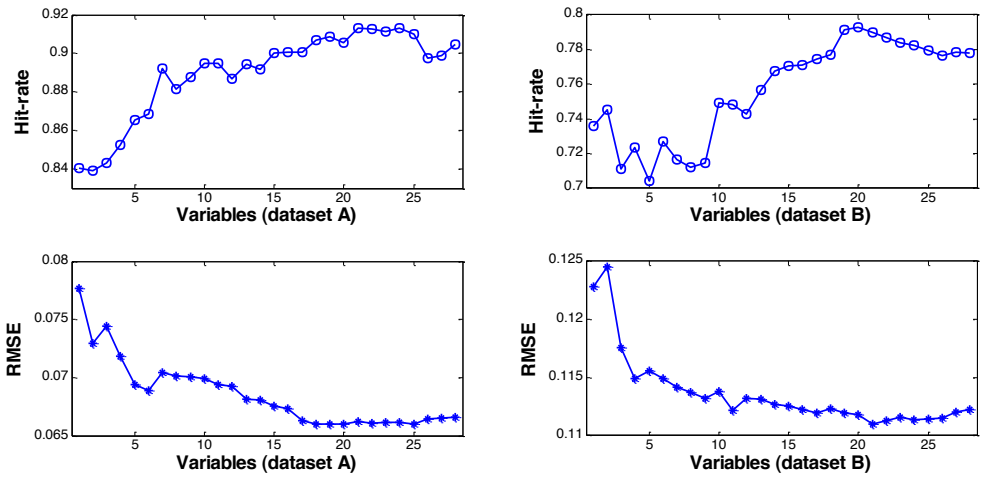

Fig. (5). The Hit-rate and RMSE obtained by adding the inputs one by one according to their importance ranking results.

data, based on which the variable importance can be estimated.

Fig. (**4**) gives the average measurements of variable importance in decreasing order for two groups of datasets based on the optimal RF configuration, from which the variable importance ranking can be obtained. Here 1000 trees are used to estimate the variable importance in the forest and the number of variables used to split a node is 28, 7 respectively. As one can see, among all the parameters $[Si]_{n-1}$ (variable 1) obtains the largest value of importance on both datasets. Besides, variables 10, 15, 16, 13, 17, *et al.* also have the similar importance on both datasets. On the other hand, there also exists difference in the evaluation of variable importance on two samples. The reason may lie in that variable importance ranking may, to some extent, depend on the training dataset and the value of the key parameter n. The results obtained here can be considered as an illustration for the time-varying characteristic of BFIP dynamic from the perspective of melting parameter importance.

To investigate the capability of the variable selection of RF, we establish RF models with the number of inputs increased one by one according to the variable importance ranking results obtained above, and then use the models to predict the silicon content. Fig. (**5**) depicts how the prediction accuracy (Hit-rate) and the root mean square error (RMSE) vary on the 100 test samples when the number of inputs increase according to the variable ranking results.

As can be seen in Fig. (**5**), generally speaking, Hit-rate first increases and then reduces when the number of variables varies from 1 to 28, and there exists an optimal variable

Table 1. Results obtained by different models for two test datasets.

Dataset	Model	RMSE	MAE	CC	Hit-rate	#inputs
A	RF	0.0663	0.0521	0.6021	91.30%	21
	SVM	0.0668	0.0535	0.5966	88.00%	10
B	RF	0.1122	0.0793	0.4054	79.30%	20
	SVM	0.1196	0.0818	0.3035	77.00%	7

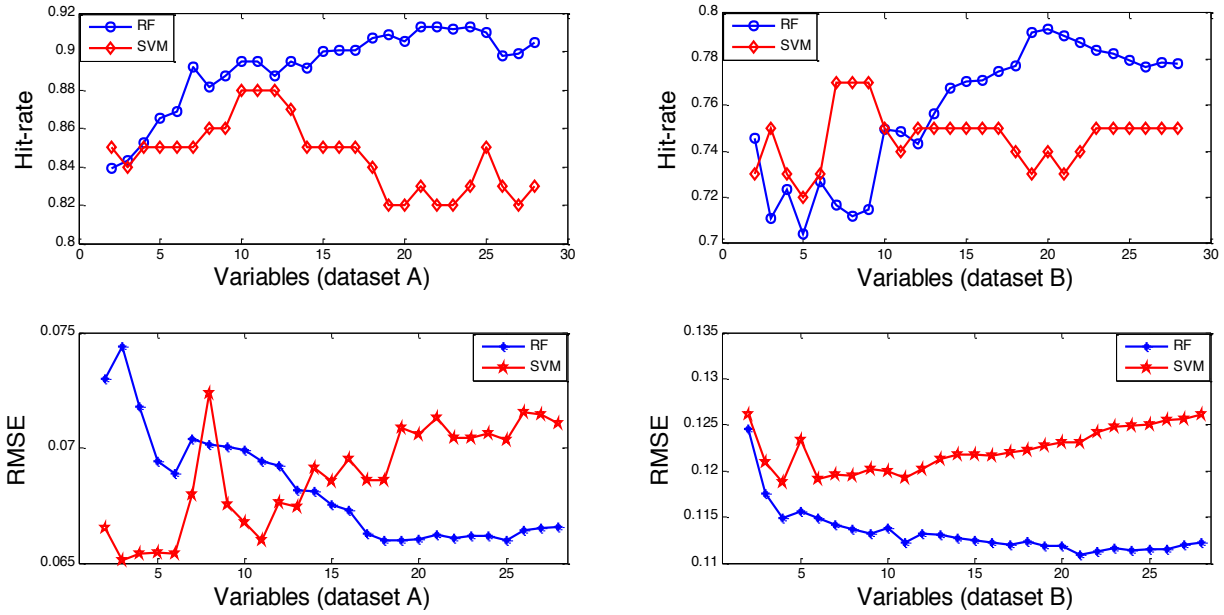

Fig. (6). The dependency of the Hit-rate and RMSE for RF and SVM on the number of variables.

number. The same trend can be found with respect to RMSE. The results show that the RF algorithm here can also serve as an effective feature selection method. On the other hand, we also note that due to the dissimilarity between the standard deviations of two datasets, RF gets much better prediction performance on the first dataset (the corresponding Hit-rate and RMSE are 91.30% and 0.066 respectively, while those of the second dataset are 79.30% and 0.1110). Additionally, when the number of inputs is larger than the optimal one, the performance of the RF models is rather steady (just becomes worse slightly), which tells that the silicon content prediction model based on RF has potential to handle all the short and long-term features in BFIP.

Experimental comparisons have been made between RF and SVM algorithms, since SVM is considered to be one of the state-of-the-art methods for BFIP modeling. Table **1** shows the forecasting results of RF and SVM models for two silicon content series. The results of both algorithms presented here are obtained based on feature selection, that is, with the optimal input variable set. And the optimal testing results are shown in boldface under the corresponding criteria.

For dataset A, the best Hit-rate obtained by RF algorithm is 91.30% while that of SVM is only 88.00%. Meanwhile, the Hit-rate of RF model is also 2.30% higher than that of SVM for dataset B. Additionally, the similar advantages can be found when all other 3 criteria are concerned. Consequently, all the results in Table **1** show that the RF algorithm can achieve higher prediction accuracy than SVM for both datasets.

We also note that the input number of the optimal RF model is 21, while that of SVM is only 10 on the first dataset and those of the second dataset are 20 and 7 respectively. Fig. (**6**) illustrates the reason for this: with the increase of the input number, and hence the increase of complexity of the relationship between all the model inputs, RF can effectively deal with the correlation and interaction among variables, but SVM fails to do so. By restraining the effect of the correlation and interaction, RF can successfully mine the essential model between all the inputs. Fig. (**6**) depicts this point in detail, in which we can see that compared with SVM algorithm, RF algorithm can obtain better prediction Hit-rates and RMSEs when the number of inputs becomes saturated.

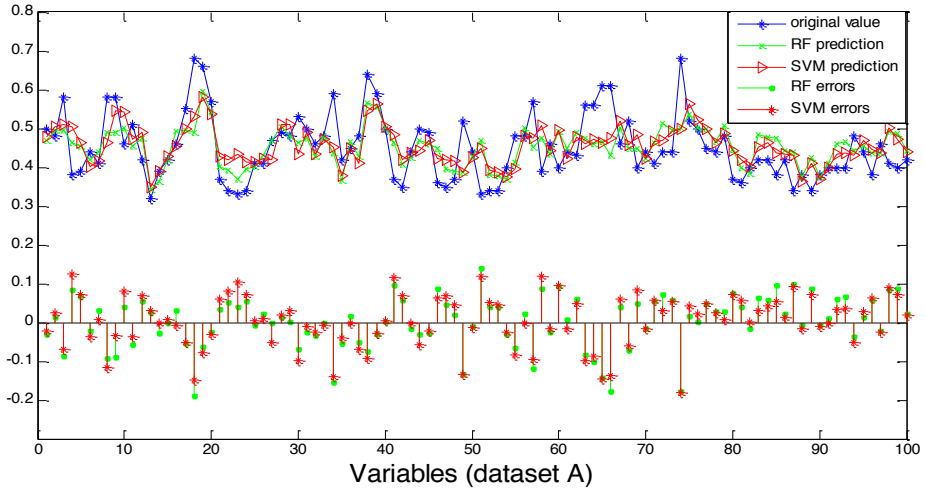

Fig. (7). Prediction results and errors comparison between the RF and SVM models for dataset A.

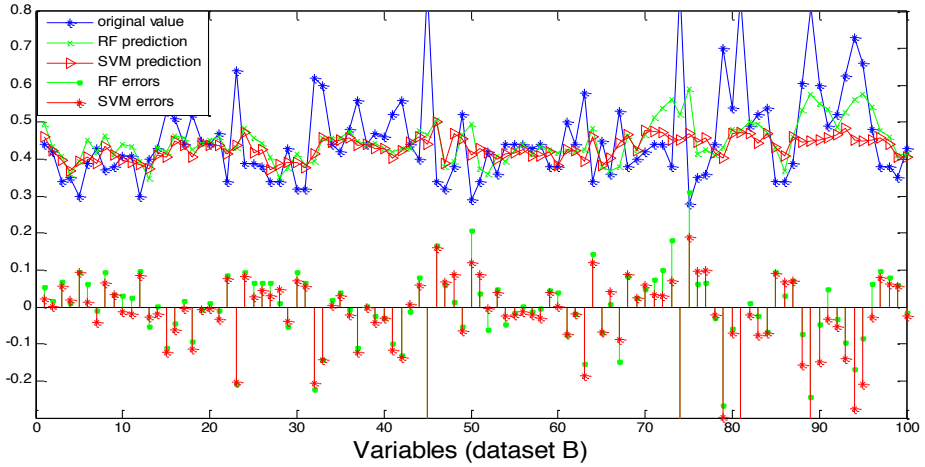

Fig. (8). Prediction results and errors comparison between the RF and SVM models for dataset B.

Figs. (**7** and **8**) give the prediction results and the error sequences obtained by the RF and SVM models. All these results also tell that the RF model can obtain better prediction results compared with SVM for both datasets.

CONCLUSION

The modeling and the prediction of the key indicator of blast furnace [Si] has not been solved till now, due to the complexity of BFIP. There exist the phenomena of long time-delay and strong cross-coupling in BFIP which play a crucial role in preventing current data-driven models from good modeling performance. This paper investigates the applicability of the RF algorithm for modeling BFIP. The simulation results obtained on two different datasets show that the RF algorithm can effectively handle the correlation and interaction between different control and state parameters of BFIP, and can evaluate the importance of these parameters. All these results demonstrate the rationality of establishing the long-term BFIP model by introducing the long-term BFIP parameters into RF algorithms. The results also show that it has potential to model such BFIP-type complex industrial process using RF algorithm.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work was supported by Zhejiang Provincial Natural Science Foundation of China (Grant No. LY14F030020) and the Foundation of Zhejiang educational committee (Grant No.Y201329298).

REFERENCES

[1] H. Saxén, "Short-term prediction of silicon content in pig iron," *Can. Metall. Q.*, vol. 33, no. 3, pp. 319-326 1994.

[2] R. Radhakrishnan, and A. R. Mohamed, "Neural networks for the identification and control of blast furnace hot metal quality," *J. Process. Control*, vol. 10, pp. 509-524, 2000.

[3] J. Chen, "A predictive system for blast furnaces by integrating a neural network with qualitative analysis," *Eng. Appl. Artif. Intell.*, vol. 14, pp. 77-85, 2001.

[4] R. D. Martin, F. Obeso, J. Mochon, R. Barea, and J. Jimenez, "Hot metal temperature prediction in blast furnace using advanced model based on fuzzy logic tools," *Ironmaking Steelmaking*, vol. 34, no. 3, pp. 241-247, 2007.

[5] T. Bhattacharya, "Prediction of silicon content in blast furnace hot metal using partial least squares (PLS)," *ISIJ Int.*, vol. 45, no. 12, pp. 1943-1945, 2005.

[6] X. J. Hao, F. M. Shen, G. Du, Y. S. Shen, and Z. Xie, "A blast furnace prediction model combing neural network with partial least square regression," *Steel Res. Int.*, vol. 76, pp. 694-699, 2005.

[7] X. Y. Liu, and W. H. Wang, "Using multivariate grey model and principal component analysis to modeling the blast furnace," In: *Proc. 2011 Multimedia Technology Conf.*, Hangzhou, CHN, pp. 2789-2792, 2011.

[8] X. Y. Liu, and W. H. Wang, "Dynamic grey model for forecasting silicon content in blast furnace hot metal," In: *Proc. Control and Decision Conf.*, Yantai, CHN, pp. 1900-1904, 2008.

[9] L. Jian, C. Gao, L. Li, and J. Zen, "Application of Least Squares Support Vector Machines to Predict the Silicon Content in Blast Furnace Hot Metal", *ISIJ Int.,* vol. 48, no. 11, pp. 1659-1661, 2008.

[10] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[11] T. K. Ho, "The random subspace method for constructing decision forests," IEEE *Trans. on Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832-844, 1998.

[12] G. Nimrod, A. Szilagyi, C. Leslie, N. Ben-Tal, "Identification of DNA-binding proteins using structural, electrostatic and evolutionary features," *J. Mol. Biol.*, vol. 387, no. 4, pp. 1040-1053, 2009.

[13] L. Guo, Y. Ma, B. Cukic, and H. Singh, "Robust prediction of fault-proneness by random forests," In: *Proc. Symposium on Software Reliability Engineering conf., Saint-Malo*, Bretagne, pp. 417-428, 2004.

[14] S. Bernard, L. Heutte, and S. Adam, "Using random forests for handwritten digit recognition," In: *Document Analysis and Recognition Conf., Parana*, pp. 1043-1047, 2007.

[15] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, Hemant A. J. Minn, and M. S. Lauer, "High-dimensional variable selection for survival data, *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 205-217, 2010.

[16] Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: *A survey and results of new tests*," Patter Recogn., vol. 44, no.2, pp. 330-349, 2011.

[17] C. Gao, J. Chen, J. Zeng, X. Liu, and Y. Sun, "A chaos-based iterated multistep predictor for blast furnace ironmaking process," *Aiche Journal*, vol. 55, no, 4, pp. 947-962, 2009.

[18] W. Hsu, and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Network*, vol. 13, no. 2, pp. 415-425, 2002.

Received: May 26, 2015 Revised: July 14, 2015 Accepted: August 10, 2015