Engineering Applications of Artificial Intelligence 23 (2010) 1012–1018

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Applying input variables selection technique on input weighted support vector machine modeling for BOF endpoint prediction

Xinzhe Wang [a], Min Han [a,*], Jun Wang [b]

[a] School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023, PR China
[b] Department of Mechanical and Automation Engineering Faculty of Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

ARTICLE INFO

Article history:
Received 4 January 2009
Received in revised form
8 November 2009
Accepted 29 December 2009
Available online 13 May 2010

Keywords:
Basic oxygen furnace
Mutual information
Support vector machine
Variables selection

ABSTRACT

Basic oxygen furnace (BOF) steelmaking is a complex process and dynamic model is very important for endpoint control. It is usually difficult to build a precise BOF endpoint dynamic model because many input variables affect the endpoint carbon content and temperature. For this problem, two effective variables selection steps: mechanism analysis and mutual information calculation are proposed to choose appropriate input variables according to a variable selection algorithm. Then, the selected inputs are weighted on the basis of mutual information values. Finally, two input weighted support vector machine BOF endpoint dynamic models are constructed to predict endpoint carbon content and temperature. Results show that the variable selection for BOF endpoint prediction model is essential and effective. The complexity and precise of two endpoint prediction models are improved.

1. Introduction

BOF is a widely preferred and effective steelmaking method due to its high productivity and considerably low production cost. Therefore, almost 65% of the total crude steel productions in the world are melted by using the BOF method. BOF steelmaking is a very complex chemical physical process. The quality of scrap iron changes from batch to batch. The grades of steel produced vary frequently, and the components of raw materials fluctuate largely. As a result, it is a tough task to implement the operation of BOF production. The main objective of controlling oxygen converter steelmaking is to obtain prescribed parameters for the steel when it is tapped from the furnace, including weight, temperature, and each element content. In practical steelmaking process, the criterion whether the molten steel is acceptable or not is often decided by the endpoint carbon content and temperature. In general, there are two main tasks of BOF: one is to make carbon percentage decrease from approximately 4% in hot metal to less than 0.08% in liquid steel, and the other is to make temperature increase from approximately 1250 °C in hot metal to more than 1650 °C.

Generally, the BOF steelmaking process with sub-lance system can be divided into two stages: static control and dynamic control. Static models include oxygen supplying model, slaging model and bottom blowing model; dynamic models include decarburization speed model, molten steel warming model and the model for the amount of coolant. Dynamic control (Feng et al., 2006) for sub-lance is on the basis of static control.

For converter steelmaking plant, mechanism model, statistical model and neural network model (Blanco and Dixz, 1993) are commonly used in recent years. A solution based on heat and mass balance from the static model (Neto, 1981) has been employed. But unfortunately, a lot of theoretical assumptions and too many parameters are involved in the traditional control methods, such as mechanism models based on heat balance and material balance, or statistic models based on regression analysis. Therefore, these models are often difficult in modeling precisely. With the development of artificial intelligence technology, neural network models and other intelligent methods have been widely applied in BOF steelmaking process. Kubat et al (2004) proposed a fuzzy model for the static control of BOF process. As a result of the application of the fuzzy model, acceptable levels of compatibility were achieved compared to the empirical BOF data in an integrated steel plant in Turkey. Coxa et al. (2002) proposed three BP models for the dynamic control of BOF process, these three models are used to predict the quantity of end-blow oxygen and end-blow coolant and determine whether the coolant is added or not. Xie et al. (2003), Bigeev and Baitman (2006) also adopted different intelligent models to describe the BOF steelmaking process. Though these aforementioned intelligent models make up the deficiencies of the traditional models to some degree, they ignore the influence of input simplifying on the prediction precision. According to Szekely (2003), it is very necessary to

* Corresponding author. Tel.: +86 411 84708719; fax: +86 411 84707847.
E-mail addresses: wxzagm@student.dlut.edu.cn (X. Wang), minhan@dlut.edu.cn (M. Han), jwang@acae.cuhk.edu.hk (J. Wang).

0952-1976/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2009.12.007

simplify the input variables to reduce the complexity and improve the generalized capacity of the industrial model. Han et al. (2008) proposed ICA and Greedy Kernel Components (Han and Huang, 2008) to reduce the dimensions, but ICA and GKC make the inputs inexplicable.

This research focuses on the prediction of BOF endpoint carbon content and temperature. Support vector machine method is used to construct the prediction models. To improve the prediction precision with less dimensional and more effective inputs, a variable selection process which contains mechanism analysis and mutual information is proposed. Furthermore, each selected variable is weighted by the mutual information values. Finally, two simpler and more accurate input weighted support vector machine endpoint prediction models are established.

The main contribution of this paper is that a variable selection method composed of mechanism and mutual information calculation is proposed to select input variables appropriately and effectively. Also, the prediction and hit rate performance of BOF process is improved by endpoint carbon and temperature SVM prediction models.

The remainder of this paper is organized as follows. Section 2 presents the proposed BOF steelmaking endpoint dynamic model system. Section 3 and Section 4 describe the variable selection algorithm and support vector machine model, respectively. Experiment results are presented in Section 5 and concluding remarks are in Section 6.

2. BOF steelmaking endpoint dynamic model system

2.1. Process description

The BOF comprises a vertical solid-bottom crucible with a vertical water-cooled oxygen lance entering the vessel from above. General view of BOF is given in Fig. 1. The vessel is tiltable for charging and tapping. The charge is normally made up of 85% molten pig iron (hot metal) and 15% scrap before blowing the oxygen. After that, the converter is rotated to a vertical position and then oxygen is turned on. During the blowing period, high-purity oxygen is blown onto the top of the hot metal at a speed of 16,000 cubic feet per minute. Assistant materials such as burnt-lime, dolomite and iron ore are added into vessel by two or three batches during the blowing time. BOF process aims to lower the levels of impurities which are carbon, silicon, manganese and phosphorus and raise the temperature from about 1350 to 1680 °C. The heat is from the oxidation reaction and no external heat is required. Carbon is oxidized to carbon monoxide and carbon dioxide, are taken out from the exhaust gas hood. Silicon, manganese and phosphorus are oxidized and combined with the assistant materials to form the slag. If the carbon level and temperature meet the requirements, the liquid steel is tapped from the tapping hole into the steel ladle. Slag is on the top of steel and be left in the converter.

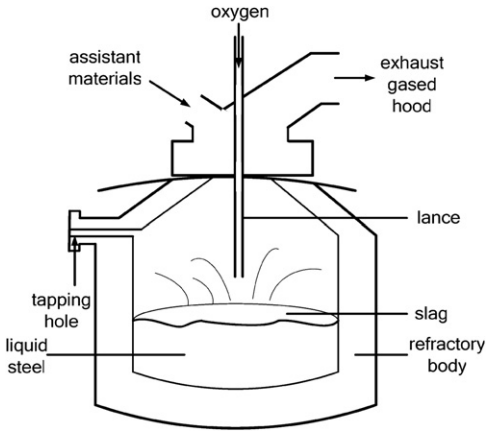

Fig. 1. General view of BOF.

The objective is to produce a desired amount of steel, which consists of specified chemical composition, at the proper tapping temperature. Control is difficult because the entire process takes only half an hour and it is hard to sample and analyze during this time. Generally, the real production adopts dynamic control practice which is shown as Fig. 2. Before blowing, the static model estimates the oxygen consumption value of the total steelmaking process. When the blowing reaches about 85% of the total value, the sub-lance (i.e. a probe is dipped into bath) goes down into the liquid steel to detect carbon content and temperature. Based upon the on-line measured carbon content and temperature, the oxygen volume and the added coolant weight in second blow period are determined. The second blow period is also called dynamic period or end-blow period. Correspondingly, in metallurgy, the model of this period is called dynamic model. If the carbon content or temperature of second measurement does not hit the target, a re-blow step is needed until liquid steel carbon content and temperature are acceptable.

Carbon content and temperature of liquid steel are two criterions to determine whether the steel can be tapped. Dynamic period is designed to enhance the hit rate of carbon content and temperature at the end of the blow. In dynamic period, chemical reactions tend to equilibrium and the main reaction occurs between oxygen and carbon. Other impurities, such as manganese, phosphor and so on, have achieved dynamic equilibrium values basically. In this period, Oxygen affects molten steel's carbon content through the reaction with carbon. At the same time, the caloric from the oxidation reaction makes the temperature rise. Coolants added into converter mainly affect the molten steel's temperature by physical effect and reduction reaction. During the dynamic process, decarbonization rate is described as formula (1) and formula (2)

$$-\frac{dw_{[C]}}{dt} = kw_{[C]} \tag{1}$$

$$t = \frac{Q}{I} \tag{2}$$

where $w_{[C]}$ is bath carbon content, k is a coefficient, t is blowing time, Q is blown oxygen volume and I is blow velocity of oxygen. The integration of formula (3) is given by

$$\ln w_{[C]} = \ln w_{[C]_0} - kt \tag{3}$$

where $w_{[C]_0}$ is the carbon content measured by sub-lance. Take exponent to formula (3) on both sides, carbon content of bath in dynamic period can be calculated as

$$w_{[C]} = w_{[C]_0} e^{-kt} \tag{4}$$

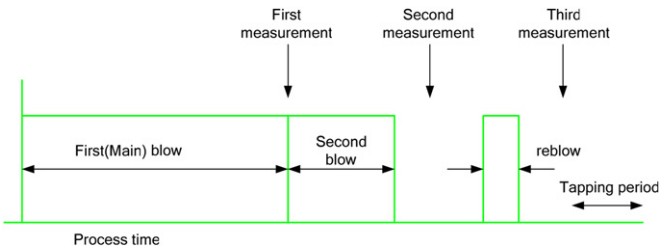

Fig. 2. Dynamic control practice.

Bath temperature can be described as

$$T = T_0 + T_r - T_c \quad (5)$$

where T_0 is the temperature measured by sub-lance, T_r is the raised temperature by blowing the oxygen and T_c the decreased temperature by adding the coolant.

2.2. Solution strategy

The BOF endpoint prediction model here is used to predict endpoint carbon content and temperature of liquid steel. They are concerned with carbon content and temperature measured by sub-lance, second blow oxygen volume and added coolant such as lime, sheet iron, ore and dolomite.

High dimension is common in industry field and it is disadvantageous for the improvement of model precision. With the dimension increasing, the density of sample data always reduces by exponent and makes the data insufficient. While the dimension increases from one to three, the data density decreases from 50% to 12.5% and is illustrated in Fig. 3. As we know,

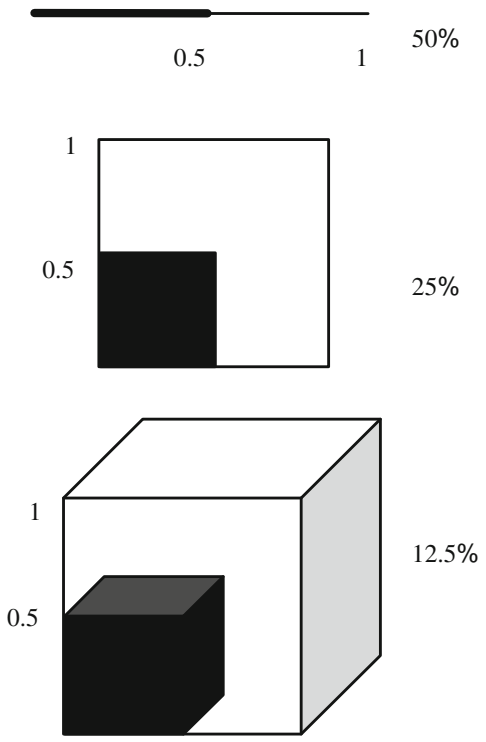

Fig. 3. The situation of sample data density with dimension increasing.

machine learning methods have the ability of self-learning, but they usually need enough data information to study adequately. The low density of training sample will limit the study of intelligent model.

In this study, corresponding variable selection techniques are adopted to choose more important input variables and decrease the number of inputs. Two variables selection steps to simplify the endpoint carbon content and temperature prediction models are described as follows:

- First, analyze the mechanism of reaction and determine the importance of each variable based on the physical meaning to reduce the inputs as few as possible.
- Second, calculate the mutual information between input and output variables. Add or remove input variables based on mutual information value and mechanism analysis.

After variable selection, the input weighted support vector machine is used to construct the endpoint carbon content and temperature prediction model. The weights are determined according to the mutual information values, to distinguish the importance of each input and strengthen the effect of more important ones to the output. Two separate models are employed to predict endpoint carbon content and temperature. BOF steelmaking dynamic model system framework is shown in Fig. 4.

3. Variables selection

As described above, the available inputs for prediction models are sub-lance measured carbon content and temperature, second blow oxygen volume and quantities of coolants contain lime, sheet iron, ore and dolomite. Through Section 2.1, sub-lance measured carbon content and second blow oxygen volume are selected as the input variables of endpoint carbon content prediction model. All available input variables are chosen for endpoint temperature prediction.

3.1. Mutual information

A number of selection criteria, such as correlation coefficient and least square regression error, can be used to select the input variables. Mutual information (MI) is chosen here because MI is capable of estimating a general dependence between two variables (Reyhani et al., 2005; Cover and Thomas, 1990).

Between two random variables X and Y, mutual information value can be interpreted as a quantity that measures the knowledge on Y provided by X. Therefore, it can Y.

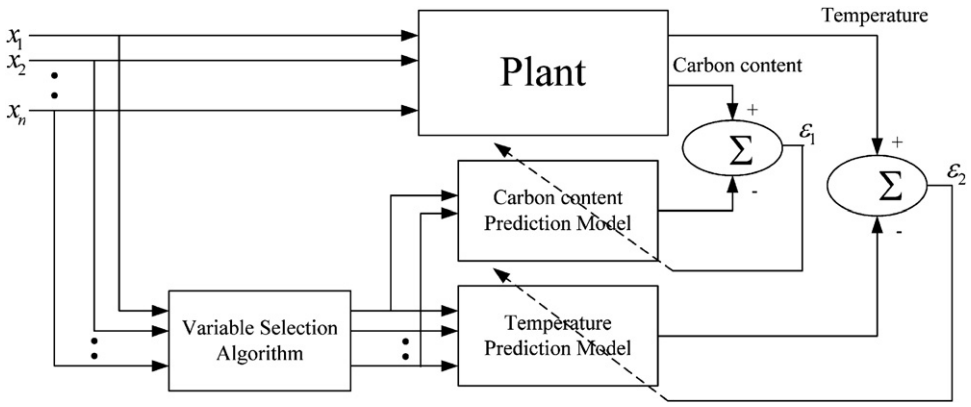

Fig. 4. Structure of BOF steelmaking dynamic model system.

Given two discrete variables $X=[x_1,x_2,\ldots,x_n]$ and $Y=[y_1,y_2,\ldots,y_n]$, the MI between X and Y is traditionally defined as

$$I(X,Y)=-\sum_{i,j}p_{xy}(x_i,y_j)\log\left(\frac{p_{xy}(x_i,y_j)}{p_x(x_i)p_y(y_j)}\right) \tag{6}$$

where x_i is a sample of variable X and y_j is a sample of variable Y; p_x and p_y are marginal probability distributions of X and Y, respectively; p_{xy} is a joint probability distribution between X and Y.

Mutual information is related to the information theoretic notion of entropy by the following equations:

$$I(X,Y)=H(X)+H(Y)-H(X,Y)=H(X)-H(X/Y)=H(Y)-H(Y/X) \tag{7}$$

where $H(X)$ and $H(Y)$ are the entropy of X and Y, respectively, $H(X,Y)$ is their joint entropy, $H(X/Y)$ and $H(Y,X)$ are the conditional entropies, respectively.

$H(X)$, $H(X/Y)$ and $H(X,Y)$ are defined as

$$H(X)=-\sum_i p_x(x_i)\log(p_x(x_i)) \tag{8}$$

$$H(X/Y)=-\sum_i p_{x,y}(x_i)\log(p_{x/y}(x_i/y_i)) \tag{9}$$

$$H(X,Y)=-\sum_{i,j}p_{xy}(x_i,y_j)\log(p_{xy}(x_i,y_j)) \tag{10}$$

where $p_{x/y}(x_i)$ is the conditional probability of X given Y. The entropy $H(X)$ is known as a measure of the amount of uncertainty about the random variable X, while $H(X/Y)$ is the amount of uncertainty left in X when knowing Y.

Hence, from Eq. (7), $I(X,Y)$ is the reduction in the uncertainty of the random variable X by the knowledge of another random variable Y, or, equivalently, the amount of information that Y contains about X. the relationships can be illustrated as Fig. 5.

If X and Y are independent, $p_{xy}(x,y)=p_x(x)p_y(y)$ and $I(X,Y)=0$, while if X and Y are related, $I(X,Y)=H(X)=H(Y)$. It can be shown $I(X,Y)\geq 0$. Thus, mutual information can play the same role with the correlation coefficient method while it reveals nonlinear correlation between variables.

3.2. Variable selection algorithm

In this section, a proposed input variables selection algorithm is presented. The algorithm is based on the mechanism analysis and mutual information calculation.

Through the mechanism analysis, the carbon content which is measured by sub-lance and the quantity of oxygen which is blown after the sub-lance measurement are more important markedly than other input variables to endpoint carbon content. In addition, all the input variables have effect on endpoint temperature. In this step, for an output variable, if an input variable x_i is selected, marked as $m^1_{X_i}=1$ and if ignored, marked as $m^1_{X_i}=0$.

In mutual information step, calculate the MI values between each input variable and output variable. The MI values are represented as $m^2_{X_i}$. Then, the variable selection algorithm is performed by the following procedure:

1) Initialization: Set F to the input variable set, S to the empty set;
2) $\forall X_i\in F$, analyze the importance to output variable and establish the $m^1_{X_i}$ value;
3) $\forall X_iF$, compute the MI value $m^2_{X_i}=I(X_i,Y)$;
4) Compute the threshold value $T=\frac{\beta\sum X_i\in Fm^1_{X_i}m^2_{X_i}}{\sum_{X_i\in F}m^1_{X_i}}$;
5) $\forall x_iF$, if $m^2_{x_i}>T$, then x_i is selected as an input variable into the set S.

The parameter β controls the threshold, can be established by cross-validation.

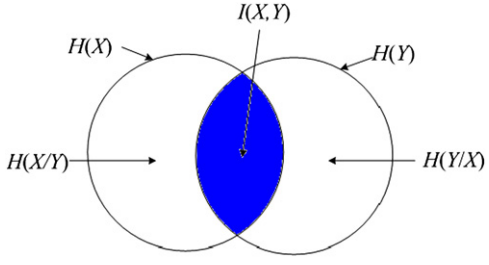

Fig. 5. Mutual information between two variables.

4. Support vector machine (SVM) model

After the appropriate input variables are selected, support vector machine method is used to construct BOF endpoint prediction model. SVM has been received increasing attention because of its remarkable characteristics, including a strong theoretical foundation, good generalization performance, the absence of local minima, and sparse representation of solution (Vapnik, 1995).

Given a set of input and output variable couples $\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_N,y_N)\}$, where $\mathbf{x}_i\in\mathbf{R}^m$, $y_i\in\mathbf{R}$, SVM is initially used to map the input data $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ into a high-dimensional feature space $\boldsymbol{F}$ by using a function column vector $\Phi(\cdot)$, and a linear regression is then performed in this feature space so that

$$f(\mathbf{X})=\mathbf{w}^T\cdot\Phi(\mathbf{X})+b \tag{11}$$

where $\mathbf{w}\in\mathbf{R}$, b is the threshold.

To obtain the two unknowns variables $(\mathbf{w},b)$ in Eq. (11), the following function should be minimized

$$C\sum_{n=1}^{N}E_\varepsilon(f(\mathbf{x}_n)-y)+\frac{1}{2}\|\mathbf{w}\|^2 \tag{12}$$

where $\|\cdot\|$ denotes the Euclidean distance, C is the regularized factor which can be ascertain by cross-validation, $E_\varepsilon(\cdot)$ is the loss function.

An ε-insensitive error loss function will be used in the present study

$$E_\varepsilon(f(\mathbf{x})-y)=\begin{cases}0, & \text{if } |f(\mathbf{x})-y|<\varepsilon\\ |f(\mathbf{x})-y|-\varepsilon, & \text{otherwise}\end{cases} \tag{13}$$

Re-express the optimization problem by introducing slack variables ξ_n and $\hat{\xi}_n$ for each data point.

$$C\sum_{n=1}^{N}(\xi_n+\hat{\xi}_n)+\frac{1}{2}\|\mathbf{w}\|^2 \tag{14}$$

$$\text{subject to}\quad f(\mathbf{x}_i)-y_i\leq\xi_i+\varepsilon\quad i=1,\ldots N$$

$$y_i-f(\mathbf{x}_i)\leq\hat{\xi}_i+\varepsilon\ \ i=1,\ldots N\xi_i,$$

$$\hat{\xi}_i\geq 0\quad i=1,\ldots N$$

The optimization problem expressed in Eq. (13) can be solved by introducing Lagrange multipliers $a_n\geq 0$, $\hat{a}_n\geq 0$, $\mu_n\geq 0$ and

$\hat{\mu}_n \geq 0$ and optimizing the Lagrangian

$$L = C\sum_{n=1}^{N}(\xi_n+\hat{\xi}_n)+\frac{1}{2}\|\mathbf{w}\|^2-\sum_{n=1}^{N}(\mu_n\xi_n+\hat{\mu}_n\hat{\xi}_n) - \sum_{n=1}^{N}a_n(\varepsilon+\xi_n+y_n-t_n)-\sum_{n=1}^{N}\hat{a}_n(\varepsilon+\hat{\xi}_n-y_n+t_n) \quad (15)$$

Setting the derivatives of the Lagrangian with respect to $\mathbf{w}$, b, ξ_n and $\hat{\xi}_n$ to zero gives

$$\mathbf{w} = \sum_{n=1}^{N}(a_n-\hat{a}_n)\phi(\mathbf{x}_n) \quad (16)$$

$$\sum_{n=1}^{N}(a_n-\hat{a}_n) = 0 \quad (17)$$

$$a_n+\mu_n = C \quad (18)$$

$$\hat{a}_n+\hat{\mu}_n = C \quad (19)$$

Using these results to eliminate the corresponding variables from the Lagrangian, the dual problem involves maximizing

$$\tilde{L}(\mathbf{a},\tilde{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n-\hat{a}_n)(a_m-\hat{a}_m)k(\mathbf{x}_n,\mathbf{x}_m) -\varepsilon\sum_{n=1}^{N}(a_n+\hat{a}_n)+\sum_{n=1}^{N}(a_n-\hat{a}_n)t_n \quad (20)$$

The predictions for new inputs can be made using

$$y(\mathbf{x}) = \sum_{n=1}^{N}(a_n-\hat{a}_n)k(\mathbf{x},\mathbf{x}_n)+b \quad (21)$$

The parameter b can be found by considering the Karush–Kuhn–Tucker (KKT) conditions

$$b = t_n-\varepsilon-\sum_{m=1}^{N}(a_m-\hat{a}_m)k(\mathbf{x}_n,\mathbf{x}_m) \quad (22)$$

5. Results

The paper constructs the endpoint carbon content and temperature prediction models according to the approach described above with industrial production data set of 60 data points. The available input variables are carbon content of molten steel which is measured by sub-lance x_1, quantity of re-blown oxygen x_2; temperature of molten steel which is measured by sub-lance x_3; weight of lime x_4; weight of mixed material x_5; weight of sheet iron x_6; weight of ore x_7; weight of dolomite x_8. Two output variables are endpoint carbon content y_1 and endpoint temperature y_2.

5.1. Pre-treatment and criteria

Before modeling, the input and output data were normalized in the range of 0–1 by the follow transform:

$$x_i'(k) = \frac{x_i(k)-x_{i(\min)}}{x_{i(\max)}-x_{i(\min)}} \quad (23)$$

where $x_i'(k)$ is the normalized value of $x_i(k)$, and $x_{i(\max)}$ and $x_{i(\min)}$ are the minimum and maximum of the variable x_i.

After the normalization, all the inputs own the same importance to the output variable. But based on the mutual information calculation, the selected input variables have different weightiness to the output variable, some inputs are more important for output and some are less important. To distinguish the different weightiness and enlarge the information in more important variables relatively, two input weighted support vectors machine models are used to predict the endpoint carbon content and temperature. The weights are added onto each input variable and established by consulting the mutual information values, as formula (24) shows

$$w_{x_{ij}} = \frac{m^2_{x_{ij}}}{(\sum_{x_{k\in S_j}} m^2_{x_{kj}})/l} \quad (24)$$

where $w_{x_{ij}}$ is the weight of the variable x_i to the output variable y_i, $m^2_{x_{ij}}$ is the mutual information value between the variable x_i and output variable y_i, s_j is the set of input variables for output variable y_j, l is the number of inputs.

Two criteria are employed to evaluate the results: hit ratio and root mean square error (RMSE). Hit is used to judge whether the prediction result satisfies the request of practical production or not. If the absolute value of distance$|\Delta T|$ between temperature prediction model output and practical temperature satisfies $|\Delta T| < 15\,^{\circ}\mathrm{C}$, define that temperature hits the target. As the same, if the absolute value of distance $|\Delta T|$ between carbon content prediction model output and practical carbon content satisfies $|\Delta C| < 0.05\%$, define that carbon content hits the target. If carbon content and temperature hit the target at the same time, define that simultaneously hits the target. Hit ratio is the ratio of hit heats in all prediction heats that can be calculated as formula (25)

$$Hr = \frac{N_{\mathrm{hit}}}{N_{\mathrm{heats}}} \quad (25)$$

where Hr is the hit ratio, N_{hit} is the number of heats that hit the target, N_{heats} is the number of the predicted samples.

Root mean square error (MSE) is used to evaluate the precision of the model which can be calculated as follows:

$$\mathrm{rmse} = \sqrt{\frac{\sum_{i=1}^{n}(p_i-r_i)^2}{n}} \quad (26)$$

where p_i is the prediction value, r_i is the practical value, n is the number of the predicted samples.

5.2. Variable selection for BOF endpoint prediction models

Based on the mechanism analysis, carbon content which is measured by sub-lance and the quantity of oxygen which is blown after the sub-lance measurement are selected as input variables for endpoint carbon content prediction model. And for endpoint temperature prediction model, all the input variables are selected. Hence, a matrix that can represent the result of mechanism analysis variable selection is

$$M_1 = [m^1_{x_{ij}}]_{8\times 2} = \begin{bmatrix} 11000000 \\ 11111111 \end{bmatrix}^{\mathrm{T}}$$

where $m^1_{x_{ij}}$ is the marker that denotes whether the ith input variable is selected for the jth output.

Calculate the mutual information valves $m^2_{x_{ij}}$ between relevant inputs and outputs, the results are shown in Table 1.

Table 1
Mutual information values between input variables and output variables.

MI value	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y_1	2.6	2.7	3.2	2.2	1.5	1.2	2.2	2.0
y_2	3.0	3.2	3.4	3.4	1.5	1.3	3.2	3.2

Calculate the threshold values for the input variables selection of these two prediction models, using the formula (18)

$$T_{y_j} = \frac{\beta \sum_{i=1}^{8} m^1_{x_{ij}} m^2_{x_{ij}}}{\sum_{i=1}^{8} m^1_{x_{ij}}} \quad j = 1, 2 \tag{27}$$

The parameter β and SVM parameters which are the width of radial basis function b and regulation coefficient C are established by cross-validation. The parameters values are shown in Table 2 with β=0.9.

According to the variable selection criterion $m^2_{x_{ij}} > T_{y_j}$, select variables x_1, x_2, x_3 into the set S_1 as the inputs of endpoint carbon content prediction model and variables $x_1, x_2, x_3, x_4, x_7, x_8$ into the set S_2 as the inputs of endpoint temperature prediction model. The weighted value for each input variable is calculated as formula (24) and the final input to SVM are $w_i x_i$.

5.3. Tests

The support vector machine method is more proper for small sample data than neural network, so the support vector machine method is chosen here. Radial basis function is selected as kernel function. Thirty-five data points are used to modeling and the remaining 25 data points are used to test the model.

Based upon the variable selection, the selected input variables for endpoint carbon content and temperature prediction models are shown in Table 3.

Endpoint carbon content and temperature prediction curves are shown in Figs. 6 and 7, where solid line represents the practical value and the dashed represents the prediction value. It shows a good performance on prediction and achieves a satisfying result.

The result that compared using all variables is shown in Table 4. Here hit ratio (C) denotes the hit ratio of carbon content prediction; RMSE (C) denotes the RMSE of carbon content prediction. Hit Ratio (T) and RMSE (T) are concerned with temperature prediction. SVM means support vector model with all available input variables. MI_IWSVM means support vector model based on mutual information variable selection and input weighted. When endpoint carbon content prediction error satisfies $|\Delta C| < 0.05\%$, the hit ratio raises from 84% to 92%

Table 2
The parameter values of T, b and C.

	T	b	C
Carbon content prediction model	2.385	5	80
temperature prediction model	2.775	3	9

Table 3
Input and output variables of endpoint carbon content and temperature prediction models.

Input variables	Output
x_1 carbon content of molten steel measured by sub-lance (10^{-2}%) x_2 quantity of reblow oxygen (m^3) x_3 temperature of molten steel measured by sub-lance (°C)	y_1 endpoint carbon content (10^{-2}%)
x_1 carbon content of molten steel measured by sub-lance (10^{-2}%) x_2 quantity of reblow oxygen (m^3) x_3 temperature of molten steel measured by sub-lance (°C) x_4 quantity of lime (t) x_5 quantity of ore (t) x_6 quantity of dolomite (t)	y_2 endpoint temperature (°C)

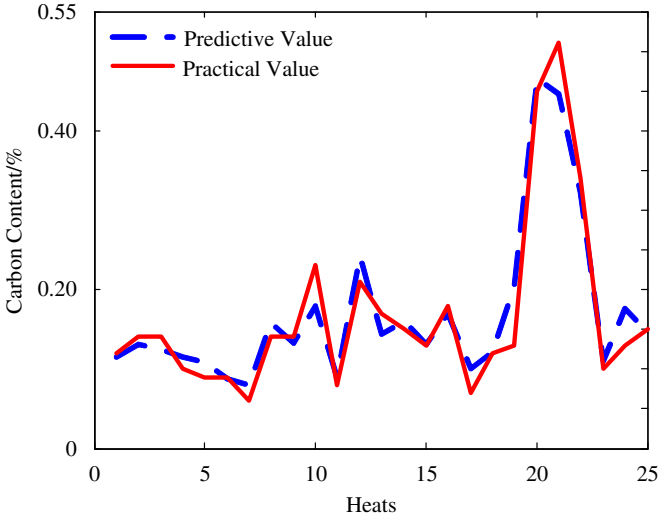

Fig. 6. Curves of endpoint carbon content prediction.

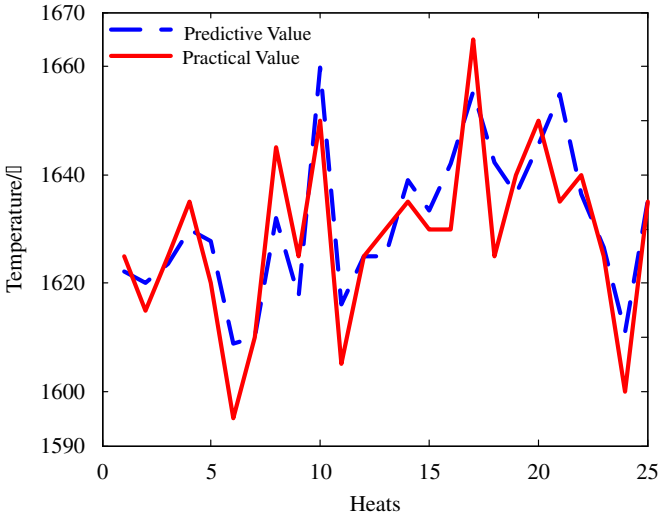

Fig. 7. Curves of endpoint temperature prediction.

Table 4
Comparison of the models before simplification to after of endpoint carbon content prediction.

	Hit ratio (C) (%)	RMSE (C)	Hit ratio (T) (%)	RMSE (T)
SVM	84	3.9989	88	9.0921
MI_IWSVM	92	2.7828	92	8.7539

Table 5
Statistical analysis of the error.

	MI_IWSVM (C)	SVM (C)	MI_IWSVM (T)	SVM (T)
Mean (x)	2.0421	2.7210	6.9399	7.3689
Standard deviation (s)	1.9293	2.1246	5.4457	5.5297
Variance (s^2)	3.7224	5.7630	29.6557	30.5773

Table 6
Comparison between two methods in references and the method in this paper.

	G_NN	ICA_NN	MI_IWSVM
Hit ratio of carbon content	88%	92%	92%
Hit ratio of temperature	88%	92%	92%
Hit ratio of simultaneity	76%	88%	88%
RMSE of carbon content	3.5964	3.0754	2.7828
RMSE of temperature	11.6064	8.7645	8.7539

and root mean square error (RMSE) decrease from 3.9989 to 2.7828. When endpoint temperature prediction error satisfies $|\Delta T| < 15\,^{\circ}\mathrm{C}$, the hit ratio raises from 88% to 92% and root mean square error decreases from 9.0921 to 8.7539.

From the comparison in Table 4, the hit ratio and RMSE are improved through the proper variable selection and input weighted. The results indicate that variable selection algorithm is essential and effective.

Take the prediction error $|\Delta C|$ and $|\Delta T|$ as statistical variable x. Table 5 details the statistical analysis of the error.

5.4. Comparison

Xie et al. (2001) combined gray model with a neural network method (G_NN) to predict the endpoint carbon content and temperature of steel with the same data as this paper used and chose all the data available as the model inputs. When the requirement of prediction error is $|\Delta T| < 15\,^{\circ}\mathrm{C}$ and $|\Delta C|$ which is also the same as this paper, endpoint temperature hit ratio is 88%, endpoint carbon content hit ratio is 88%, and the simultaneous hit ratio is 76%.The root mean square error of temperature prediction is 11.6064 and the mean square error of the carbon content is 3.5964.

Han et al. (2008) employed mechanism analysis and independence component analysis to reduce the input dimension, and then construct RBF neural network models (ICA_NN) to predict the endpoint carbon content and temperature of steel with the same data as this paper used. When the requirement of prediction error is $|\Delta T| < 15\,^{\circ}\mathrm{C}$ and $|\Delta C| < 5$ which is also the same as this paper, endpoint temperature hit ratio is 92%, endpoint carbon content hit ratio is 92%, and the simultaneous hit ratio is 88%. The root mean square error of temperature prediction is 8.7645 and the mean square error of the carbon content is 3.0754.

The comparison between these two methods and method in this paper (MI_WSVM) is shown in Table 6, all criteria of the models in this paper are better than the model in references. For hit ratios, ICA_NN and MI_WSVM are the same. For root mean square error of temperature prediction, MI_WSVM is a slightly better than ICA_NN. But for carbon content prediction, MI_WSVM makes a considerable progress as the temperature measured by sub-lance is selected by mutual information calculation.

6. Conclusions

This research presents an effective variable selection and modeling method for converter steelmaking endpoint carbon content and temperature prediction. Variables selection process contains two steps of mechanism analysis and mutual information calculation. The selected variables are used to construct the prediction models. After the pre-treatment, input dimension of endpoint carbon content prediction model reduces from 8 to 3 and endpoint temperature prediction model's inputs reduces from 8 to 6. The model precision and hit ratio achieve a remarkable improvement comparing with using all the inputs. The results of this method are better than existed methods. Dimension reduction pre-treatment for modeling is necessary and valid.

Acknowledgments

This research is supported by the project (2007AA04Z158) of the National High Technology Research and Development Program of China (863 Program), the project (60674073) of the National Nature Science Foundation of China, the project (2006BAB14B05) of the National Key Technology R&D Program of China and the project (2006CB403405) of the National Basic Research Program of China (973 Program). All of these supports are appreciated.

References

Bigeev, A., Baitman, V., 2006. Adapting a mathematical model of the end of the blow of a converter heat to existing conditions in the oxygen-converter shop at the Magnitogorsk Metallurgical Combine. Metallurgist 50 (9), 469–472.
Blanco, C., Dixz, M., 1993. Model of mixed control for carbon and silicon in a steel converter. ISIJ International 33 (7), 757–763.
Feng, J., Zhang, J., et al., 2006. Application and research of converter math model. Journal of Hebei University of Science and Technology 27 (1), 65–69.
Coxa, I.J., Lewis, R.W., Ransing, R.S., Laszczewski, H., Berni, G., 2002. Application of neural computing in basic oxygen steelmaking. Journal of Materials Processing Technology 120, 310–315.
Kubat, C., Taskin, H., et al., 2004. Bofy-fuzzy logic control for the basic oxygen furnace (BOF). Robotics and Autonomous Systems 49, 193–205.
Han, M., Huang, X., 2008. Greedy kernel components acting on ANFIS to predict BOF steelmaking endpoint. In: Proceedings of the 17th World Congress the International Federation of Automatic Control, pp. 11007–11012.
Han, M., Wang, X., Wang, Y., 2008. Applying ICA on neural network to simplify BOF endpiont predicting model. In: IEEE World Congress on Computational Intelligence, pp. 772–777.
Neto, L.C., 1981. End-blow model for control of LD converters and statistic analysis of its performance. Master's Thesis, Federal University of Minas Gerais, Brazil.
Reyhani, N., Hao, J., Ji, Y., Lendasse, A., 2005. Mutual information and gamma test for input selection. In: European Symposium on Artificial Neural Networks. Bruge, Belgium.
Szekely, N., 2003. Simplifying the model of a complex industrial process using input variable selection. Periodica Polytechnica Electrical Engineering 47 (1-2), 141–147.
Cover, T., Thomas, J., 1990. Elements of Information Theory. John Wiley.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
Xie, Shuming., Chai, Tianyou., Tao, Jun., 2001. A New Method of Converter Steelmaking Dynamic Endpoint Prediction. Acta Automatica Sinica (1), 136–139.
Xie, S.M., Tao, J., et al., 2003. BOF steelmaking endpoint control based on neural network. Control Theory and Application 20 (6), 903–907.