APPLICATION OF LARGE SCALE DATABASE-BASED ONLINE MODELLING ON BLAST FURNACE OPERATION

Masahiro Ito*, Shinroku Matsuzaki*, Harutoshi Ogai**
Kenichi Mori, Kenko Uchida**, Shinichi Saito*** and Nozomi Sasaki***

** Nippon Steel Corporation*
*** Waseda University, *** Nittetsu Elex Corporation*

Abstract: A practical method of JIT modelling for a blast furnace process data whose characteristics are very complicated physical phenomena and stiff nonlinear process are studied. The proposed method is composed of severely selecting process data variables by stepwise method and searching the past similar data in quantized topological space constructed by the selected variables. The effectiveness of this method is confirmed. It is also confirmed that it is sufficiently possible to apply this method to online use because the calculation time is quite short for searching the past similar data and estimating the future.

Keywords: Database, Dynamic modelling, Nonlinear models, On-line control, Regression estimates, Steel manufacture, Variables, Validity

1. INTRODUCTION

In recent years, the new local modelling methods are attracted a great deal of considerable attention, which are called "Just-In-Time modelling", after this JIT modelling for short, see (Zheng and Kimura, 2001a and 2001b) or "Lazy Learning", see (Atkeson, *et al.*, 1997; Bontempi, *et al.*, 1999). That is based on the background that it have become to be able to store large scale process data and to search data quickly as technology advances of computer hardware and database system. These modelling methods are first characterized to store the observational process data directly in database among the wide process operation range beyond the designed static specific operation conditions. They are next characterized to search the relevant data to the system input, called "query", as the "neighboring" data whenever it becomes required to estimate (or predict, control) the system state with the database. Then the system output correspond to the query are got by a local model that complements the outputs of the neighboring data. These modelling methods are also characterized to discard the local model whenever the system has been estimated and then to cope with the further storage of the observational process data. Case-Based Reasoning (CBR) in the field of qualitative Reasoning, see (Tsutsui, *et al.*, 1997), is also based on the same idea. A problem to be solved in JIT modelling is having to execute a huge amount of computation to calculate the distances between all the observational process data and the query and having to put all the data in the proper order for the purpose of searching the neighboring data close to the query. For example, blast furnace is a complicated and stiff nonlinear process (or system) in its physical phenomena. Then various sensors have been provided on the blast furnace body. Therefore, it becomes hardly to apply JIT modelling on the real situation of online plant operation, because of a huge amount of computing load when a large scale database would be constructed to deal with a topological space which is composed of the observational process data variables and of their time delay variables, or topological variables.

This paper describes a practical method for a large scale database that avoids the above-mentioned problem when JIT modelling would be applied on the real situation of online plant operation. This practical method is composed of the following steps.

Step 1 Positively eliminating noisy variables from all the observational process data variables including their topological ones by applying "stepwise method" for the purpose of selecting the only effective variables that contribute to estimate the system output.

Step 2 Storing observational process data sets that belong to a multidimensional topological space composed of the selected variables in quantized topological space for searching.

Step 3 Searching neighboring data close to query on the quantized space quantum by quantum for increasing the efficiency of searching and drastic reducing a huge amount of computing load.

Step 4 Estimating system output that correspond to query by using local model which complements outputs of searched and selected neighboring data.

This practical method is characterized in the same way as JIT modelling in the terms of discarding local model whenever system have been estimated and then coping with process's characteristic change in time by storing a further storage of observational process data. In this paper, the practical method is named "**L**arge scale database-based **O**nline **M**odelling; LOM", and a case study is described about applying LOM to blast furnace operation for the purpose of verifying its validity.

2. JUST-IN-TIME MODELLING

JIT modelling, that is a basic concept of "Large scale database-based Online Modelling; LOM", is explained as follows. It's assumed that an object process is a nonlinear dynamic system, and that its characteristics, or dynamic behaviours, are given a regression model as shown the equation (1).

$$\begin{aligned}\mathbf{y}(t+p) = f\{&\mathbf{y}(t), \mathbf{y}(t-1), \cdots, \mathbf{y}(t-n_y), \\ &\mathbf{u}(t-d), \mathbf{u}(t-d-1), \cdots, \mathbf{u}(t-d-n_u)\}\end{aligned} \quad (1)$$

Where,

$\mathbf{u}(t)$ is the control input vector of system at time t,

$\mathbf{y}(t)$ is the observational output vector of system at time t,

n_u is the order of control input vector,

n_y is the order of observational output vector,

p is the estimate time (or the predict time),

d is the time delay,

f is the unknown nonlinear function.

Now, the system input vector $\mathbf{x}^k$ and the system output vector $\mathbf{y}^k$ are redefined as the following equations (2) and (3).

$$\mathbf{y}^k = \mathbf{y}(k+p) \quad (2)$$

$$\begin{aligned}\mathbf{x}^k = \{&\mathbf{y}(k), \mathbf{y}(k-1), \cdots, \mathbf{y}(k-n_y), \\ &\mathbf{u}(k-d), \mathbf{u}(k-d-1), \cdots, \mathbf{u}(k-d-n_u)\}\end{aligned} \quad (3)$$

As the time progressing, a large number of data, which is composed of the system input vector $\mathbf{x}^k$ and the system output vector $\mathbf{y}^k$, for example $(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2), \cdots$, have being stored from the system as a data set $\{(\mathbf{x}^k, \mathbf{y}^k)\}, (k = 1, 2, \cdots,)$, where k is the discrete time. Then, JIT modelling is equal to finding out the nonlinear function f from the stored data set $\{(\mathbf{x}^k, \mathbf{y}^k)\}, (k = 1, 2, \cdots,)$ whenever it becomes required to estimate (or predict, control). The basic concept of JIT modelling is the follows.

For example, when it becomes necessary to estimate a system state, the present system state $\{(\mathbf{x}^{k_q}, \mathbf{y}^{k_q})\}$ is defined as the query. In the case that a neighboring data which is similar to the query exists in the database as the observed process data in the past, so that the present nonlinear function f^{k_q}, which describes the transient behaviour of the system from the present, is regard to be similar to the past linear function f^{k_i}. If the observed process data in the past exist plural, the system output vector $\mathbf{y}^{k_q}$ are estimated by using local model which complements the outputs of the selected neighboring data sets $\mathbf{y}^{k_i}$. Whenever the system state have been estimated, the observational process data are newly storing and renovating the data set $\{(\mathbf{x}^k, \mathbf{y}^k)\}$, so that the transient change of the process characteristic could be reflected at the next time that the system will be estimated later.

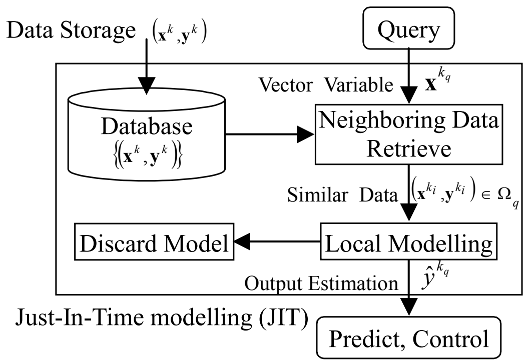

Fig.1. Just-In-Time modelling.

3. LARGE SCALE DATABASE-BASED ONLINE MODELLING (LOM)

Large scale database-based online modelling (LOM) is studied for the purpose of applying JIT modelling online to the multidimensional topological space composed of a real process data. LOM is organized the following components:

1. Reducing the number of topological space by using stepwise method
2. Effectively searching neighboring data on quantized topological space
3. A local model that generated and discarded whenever a process state becomes required to be estimated

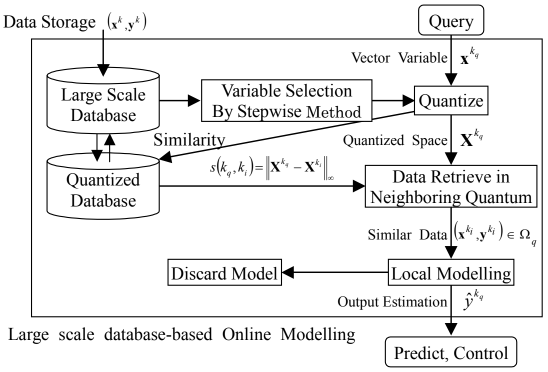

Fig.2. Large scale database-based online modelling.

3.1 Reducing the number of topological space by using stepwise method

For practical effect to estimate with a regression model, stepwise method is a technique that increases or decreases input variables based on a examination standard to aim at lessening a residual error, or a sum of squares due to error between the observational value and the estimated one. That is, an input variable is added when the input value is added to a regression model and "variable proportion F", or the value that variable quantity of the sum of squares due to error normalized by of the residual error variance, becomes bigger than the beforehand examination standard. The other side, an input variable is removed when "variable proportion F" becomes smaller than the standard. The procedure is done in turn that from the input variable that has the biggest single correlation coefficient to the input variable that has the smaller one, so that most proper regression model is acquired as the final remained regression model when there is no input variable that is added or removed.

For blast furnace, much number multidimensional space has to be treated because the number of observational data item is extremely huge and its topological variables have to be also considered. So that LOM is studied as a practical modelling method with using stepwise method under the severe examination standard for reducing the number of multidimensional space.

3.2 Quantizing topological space and searching the "neighbor"

For the purpose of realizing to search a large number and scale of online process data much quickly with effectively searching and drastically reducing the computational load, LOM stores observational process data in the beforehand database for searching, and searches the neighboring data close to the query in the quantized topological space of the database quantum by quantum. At first, a quantized space $\mathbf{X}^k$ is defined by quantizing the topological space that a vector variable $\mathbf{x}^k$ to which belongs.

$$\mathbf{X}^k = Z\left(\mathbf{x}^{k_i}\right), (i = 1,2,\cdots,n) \quad (4)$$

Where, $Z(\cdot)$ is the quantizing operator, n is the number of the data that belongs to the same quantized space $\mathbf{X}^k$. In the case that a vector variable $\mathbf{x}^k$ has one dimension, a quantized space $\mathbf{X}^k$ is an interval, and in the case that a vector variable $\mathbf{x}^k$ has two dimensions, a quantized space $\mathbf{X}^k$ is a rectangle, and generally a quantized space $\mathbf{X}^k$ is multidimensional prism. Secondarily, a similarity $s(k_i, k_j)$ is defined between the quantized space $\mathbf{X}^{k_i}$ and $\mathbf{X}^{k_j}$. For example, a similarity is referred as an infinite norm of the quantized space's reciprocal relation, see (Ito, *et al.*, 2004).

$$s(k_i, k_j) = \left\|\mathbf{X}^{k_i} - \mathbf{X}^{k_j}\right\|_\infty \quad (5)$$

Then, the quantized space that contains the query variable vector $\mathbf{x}^{k_q}$ is named as $\mathbf{X}^{k_q}$, the neighoring space Ω_q is defined as follows.

$$\Omega_q = \left\{\mathbf{X}^{k_p} \mid s(k_q, k_p) = \min_{\mathbf{X}^{k_p} \in T} s(k_q, k_p)\right\} \quad (6)$$

Where, T is a set of topological space. By quantizing, a similarity s is defined and treated as a discrete value, so that searching for the neighboring data becomes simple and efficient in the quantized database with inspecting at first the same quantum, secondly the next quanta, …, .

The several ways of determining the quantum's width are proposed. In this paper, a uniform equalized way, or the simplest way, is adopted as the first step of the application for blast furnace process data.

3.3 Local model

When the query variable vector $\mathbf{x}^{k_q}$ is given, the system output estimation is executed with applying a local model to the neighboring data sets. In JIT modelling, representative local models are locally weighed averaging (LWA) and locally weighted regression (LWR). In this paper, the simplest averaging way is adopted as the first step of the application for blast furnace process data. That is, the estimated system output vector $\hat{y}^{k_q}$ is calculated as follows.

$$\hat{y}^{k_q} = F\left(\mathbf{x}^{k_q}\right) = \frac{1}{M} \sum_{\mathbf{y}^k : (\mathbf{x}^k, \mathbf{y}^k) \in \Omega_q} \mathbf{y}^k \quad (7)$$

Where, M is the number of the system output vector $\mathbf{y}^k$ that belongs to the neighboring space Ω_q.

4. LARGE SCALE DATABASE-BASED ONLINE MODELLING ON BLAST FURNACE OPERATION

In blast furnace, a large number of sensors that detect various physical amounts, for example, temperature, pressure and gas composition, are set up. Each observational data are transmitted to a process computer, and stored in some kind of recording device in the process computer. Furthermore, some kind of the blast furnace operation indicators are computed and stored successively in the process computer by combining their observational data, for example, “K-value” as the gas permeability inside furnace indicator and “Heat Load” as the indicator of the chemical reaction inside furnace and furnace body cooling. A large scale database is constructed by handling these a large number of blast furnace process data which stored in a process computer.

In this chapter, by applying LOM mentioned in the third chapter to the blast furnace process data, the future estimation executes based on the searching results for the past similar process data, and then the validity and efficiency of LOM are verified.

4.1 Large scale database of blast furnace operation data

In this paper, the large scale database is constructed by the process data of the No.3 blast furnace in Nagoya works, Nippon Steel Corp. The number of data items is 235, the data sampling time is 1 hour and the data stored period is from 1st January 2004 to 31th January 2005, so that the number of all the data is 9528.

4.2 Equating regression model

In this paper, a blast furnace process is treated as a multi-input, multi-output and nonlinear dynamic system. Because a blast furnace process consists of much kind of complicated physical phenomena simultaneously, it is quite difficult to separate accurately beforehand “system input variable as a cause” and “system output variable as a result”. So in this paper, in the equation (1), system output variables $\mathbf{y}$ are treated equally to system input variables $\mathbf{u}$, and the assumption is set that both variables are described as gathered variables form a blast furnace $\mathbf{y}$. That is, a dynamic behaviour of a blast furnace after p hours from a present time is expressed with a regression model as the equation (8).

$$\mathbf{y}(t+p)=f\{\mathbf{y}(t),\mathbf{y}(t-1),\mathbf{y}(t-2),\cdots,\mathbf{y}(t-n_y)\} \quad (8)$$

Then, the equation (8) is redefined like the equation (2) and (3).

$$\mathbf{y}^k=\mathbf{y}(k+p) \quad (9)$$

$$\mathbf{x}^k=\{\mathbf{y}(k),\mathbf{y}(k-1),\mathbf{y}(k-2),\cdots,\mathbf{y}(k-n_y)\} \quad (10)$$

The equation (9) and (10) express that a large number of data sets gathered form a blast furnace, or $\{(\mathbf{x}^k,\mathbf{y}^k)\},(k=1,2,3,\cdots)$, are successively stored in a database.

4.3 Reducing variables by applying stepwise method

When the number of variables gathered from a blast furnace is N and the first variable is treated as the output of a regression model, the equation (8) becomes the equation (11) by describing each element of the variable vectors.

$$y_1(t+p)=f\begin{Bmatrix} y_1(t), & y_1(t-1), & \cdots, & y_1(t-n_1), \\ y_2(t), & y_2(t-1), & \cdots, & y_2(t-n_2), \\ \cdots, & \cdots, & \cdots, & \cdots, \\ y_N(t), & y_N(t-1), & \cdots, & y_N(t-n_N) \end{Bmatrix} \quad (11)$$

The equation (11) shows that the value of the first variable after p hours from a present time $y_1(t+p)$ is described with a regression model whose elements number is $\sum_{i=1}^{N}(n_i+1)$.

In a blast furnace process data, the number of the observational variables N is quite a large, so that it is necessary to set beforehand the order of observational output vector n_y on relatively large, because many variables whose dynamic characteristic shows from the short time period change to the long one simultaneously exist. As a result, the number of variables that construct a regression model becomes quite large.

Now, for example, p is set to 1, and the estimated output variable is set to a molten iron temperature. Under the assumption that molten iron temperature after 1 hour from a present time is expressed with using observational data within past 12 hours including a present data, or $n_1=n_2=\cdots=n_{235}=12$, the molten iron temperature after 1 hour from a present time $y_1(t+1)$ is expressed as a regression model whose variable number is 3055 like the equation (12).

$$y_1(t+1)=f\begin{Bmatrix} y_1(t), & y_1(t-1), & \cdots, & y_1(t-12), \\ y_2(t), & y_2(t-1), & \cdots, & y_2(t-12), \\ \cdots, & \cdots, & \cdots, & \cdots, \\ y_{235}(t) & y_{235}(t-1) & \cdots, & y_{235}(t-12) \end{Bmatrix} \quad (12)$$

With applying LOM, variables are selected by calculating “ variable proportion F” against the molten iron temperature after 1 hour from a present time $y_1(t+1)$ with stepwise method. At first, adopting a general test criterion on stepwise method, or $F_{in}=F_{out}=2.0$, 415 variables are selected, but many noise variables still exist in the selected ones, so that the estimation accuracy for the molten iron temperature is not enough. Therefore adopting a more severe test criterion, or $F_{in}=F_{out}=20.0$, 35 variables shown in the Table.1 are selected.

Table.1. Selected variables for molten iron temperature after 1 hour from a present time

Number	Selected variables by stepwise method Content	F value
1	Molten iron temperature (present)	28879
2	Tapping velocity (present)	706
3	Molten iron Ti concentration (present)	416
4	Molten iron temperature (before 4 hours)	222
5	Flame temperature (before 3 hours)	160
6	Tapping velocity (before 1 hour)	134
7	Slag Al2O2 (present)	100
8	Pulverzied Coal Ratio (present)	94.6
9	Heat Load (present)	90.6
10	Slag TiO2 (present)	80.1
…	…	…
35	#10 tuyere blast volume (present)	20.6

4.4 Quantizing topological space and verifying estimation accuracy by LOM

With quantizing the each 35 variables, including topological variables, selected by stepwise method, the quantized 35 dimensional topological space is constructed. Several guidelines of setting the number of quantizing are known, in this paper, the number of quantizing, or $N_s = 20$, is decided as the number that leads the best accuracy for a molten iron temperature estimation by referring Sturges's formula as follows,

$$N_s \cong 1 + \log_2 n_{all} = 1 + \log_2 9528 \cong 14 \quad (13)$$

and Leave-one-out Cross Validation and so on.

The accuracy of the molten iron temperature estimation is evaluated by the correlation coefficient between the estimated molten iron temperature by LOM $\hat{y}_1^{k_q+1}$ and the actual molten iron temperature $y_1^{k_q+1}$ in the random picked up 200 data as the query $\mathbf{x}^{k_q}$ from all the 9528 data sets. The correlation coefficient ρ is 0.788, and Fig.3 shows that it is sufficiently possible to estimate the molten iron temperature after 1 hour from a present time by searching the past similar process data sets.

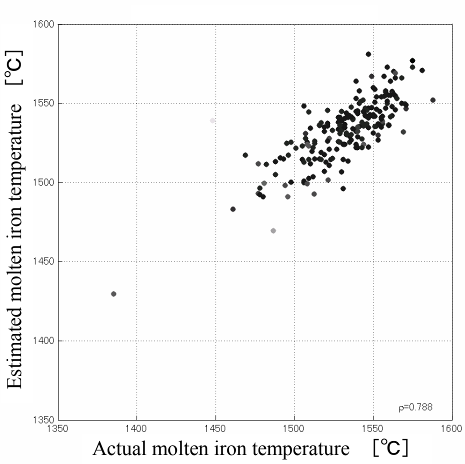

Fig.3 Correlation between actual molten iron temperature after 1 hour and estimated molten iron temperature

4.5 Searching past similar process data and estimating future process data by LOM

An any data set is selected in the large scale database as the query $\mathbf{x}^{k_q}$. And data sets are searched which exist in the neighboring quanta close to the query quantum belongs on the quantized multidimensional topological space. So it becomes possible to search the past similar process data to the query.

For example, the data set of 29th January 2005 is selected as the query in the all data sets from 1st January 2004 to 31th January 2005. Some data sets exist in the neighboring quanta, so that 5 process data sets are found out within the neighboring quanta whose similarity is 2. It means that 5 cases that are similar to the process state of 29th January 2005 exist in the past, see Fig.4.(b). In this case, in the similarity s is 0 or 1, that is, no similar data set exists in the same quantum and the next quanta. The local model shown by the equation (7) is applied to the output vector $\mathbf{y}^k$ of the 5 searched neighboring data sets, and then the molten iron temperature after 1 hour is estimated. The results are shown in Fig.4.(a). Fig.4.(a) shows that the molten iron temperature after 1 hour, which is defined as the output variable when the variable proportion F is calculated with stepwise method, are quite accurately estimated.

Furthermore, the dynamic behaviour of molten iron temperature not only after 1 hour but also until 12 hours from the present time is calculated by using the equation (7). Fig.4.(a) shows that the dynamic behaviour of a molten iron temperature until 12 hour later from 06:00 29th January 2005 are also well accurately estimated. That is, paying attention to the variable proportion F about the molten iron temperature after 1 hour, variables that compose the topological space are selected, in the case of proper searching the past similar process data, it is confirmed that the estimation for the molten iron temperature over after 1 hour are available.

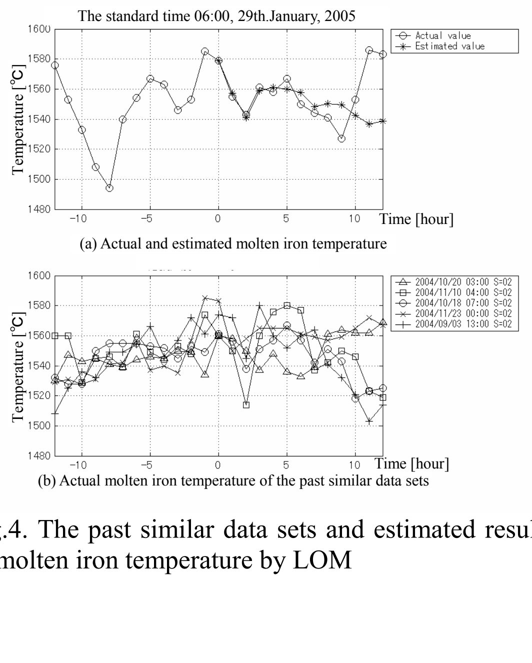

Fig.4. The past similar data sets and estimated result of molten iron temperature by LOM

Furthermore, when blast furnace is considered as a multi-input, multi-output system, the other elements of the estimated system output vector $\hat{y}^{k_q}$, for example, the molten iron Si concentration or the total K-value, are calculated, and then it is confirmed that the estimation for them after 1 hour and over after 1 hour are available as well as the molten iron temperature. (Fig.5., Fig.6.)

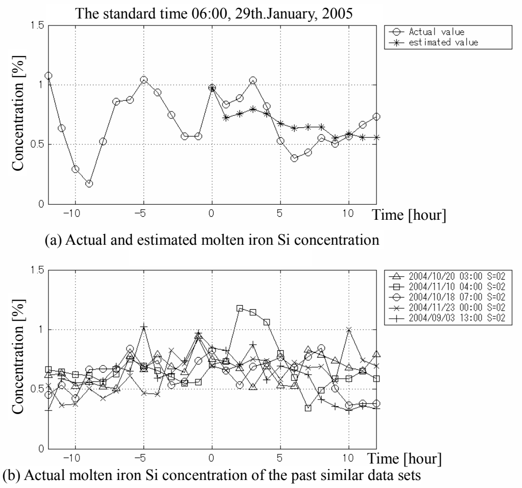

Fig.5. The past similar data sets and estimated result of molten iron Si concentration by LOM.

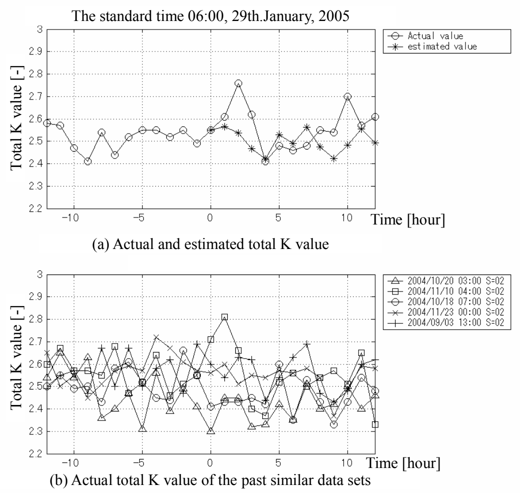

Fig.6. The past similar data sets and estimated result of total K-value by LOM.

5. CONCLUSION

In this paper, the practical method of JIT modelling for a real blast furnace process data whose characteristics are very complicated physical phenomena and stiff nonlinear process are studied, based on the background that it have become to be able to store large scale process data and to search data quickly as technology advances of computer hardware and database system. Then, the effectiveness of the proposed practical method is confirmed. It is also confirmed that it is sufficiently possible to apply this proposed method to online use because the calculation time is quite short for searching the past similar process data and estimating the future process data.

Now the new blast furnace process data is successively stored and the database is renewed, so that it is expected to estimate the blast furnace state among the broad operation range by storing the observational process data beyond the designed static specific operation conditions.

In this paper, as the first step of the application for blast furnace process data, the simplest uniform equalized way is adopted for deciding the width of the quantum, and the simplest averaging way is adopted as the local model. As the next step, it is expected to improve the estimate accuracy by studying the width deciding way from a viewpoint for a data spatial density and the locally weighted model mentioned in JIT modelling.

REFERENCES

Atkeson, C.G., A.W. Moore, and H. Schaal (1997). Locally weighted lerning. *Artificial Intelligence Review*, **11**, 11-73.

Bontempi, G., M. Birattari and H. Bersini (1999). Lazy learning for local modelling and control design. *Int.J.Control*, **72**, 643-658.

Tsutsui, H., A. Kurosaki and T. Sato (1997). Nonlinear modelling technique using historical data for case. *Trans. Soc. Instrum. Control Eng.*, **33**, 947-954.

Ito, M, S. Matsuzaki, H. Ogai, N. Odate, K. Uchida, S. Saito and N.Sasaki (2004). Large scale database-based online modelling on blast furnace operation. *Tetsu-to-Hagane*, **90**.

Zheng, Q. and H.Kimura (2001a). A new Just-In-Time modelling and its applications to rolling set-up modelling. *Trans. Soc. Instrum. Control Eng.*, **37**, 640-646.

Zheng, Q. and H.Kimura (2001b). Just-In-Time modelling for function prediction and its applications. *Asian Journal of Control*, **3**, 35-44.