CHEMICAL ENGINEERING TRANSACTIONS
VOL. 61, 2017
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Petar S Varbanov, Rongxin Su, Hon Loong Lam, Xia Liu, Jiří J Klemeš

ISBN 978-88-95608-51-8; **ISSN** 2283-9216
DOI: 10.3303/CET1761172

Just-in-time Modeling with a Combination of Input and Output Similarity Criterions for Soft Sensor Modeling in Fermentation Processes

Congli Mei*, Yao Chen, Hui Jiang, Yuhan Ding, Xu Chen, Guohai Liu

Jiangsu University, Zhenjiang, 212013, Jiangsu, China
clmei@ujs.edu.cn.

Just-in-time learning (JITL) has been used to construct soft sensor models online for its ability of handling strong nonlinearity and changes in processes. The most key procedure in JITL modelling is selecting relevant samples similar to a query sample. However, the common similarity criterions used to select relevant samples do not always function well for only considering the similarity of input data. Large noise or outliers in output data may result in inappropriate predictions of JITL based soft sensors. In this work, a combination of similarity measures, the conventional similarity of input and a novel similarity of output, is proposed for comprehensively understanding and selecting relevant samples. The effectiveness of the proposed soft sensor is demonstrated through an industrial fed-batch Erythromycin fermentation process.

1. Introduction

In process plants, soft sensors become popular to estimate those variables difficult to measure online. Compared to mechanism models, data-driven soft sensors are more popular in recent years (Kadlec et al., 2009), e.g. artificial neural network (ANN) (Pani et al., 2013), principle component regression(PCR) (Ge et al., 2014), partial least-squares(PLS) (Wang et al., 2015), support vector machine(SVM) (Jin et al., 2015b), and Gaussian process regression(GPR) (Jin et al., 2015a). Those soft sensors relied on offline modelling using the recorded historical data. However, in order to guarantee the success of the offline soft sensors, there are several conditions should be fulfilled. Most critically, the historical data should contain all possible future states and conditions of the process. Even if the collected data contains all the required states, another difficulty is the model type, and parameters, in such a way that the model can comprehend all the different conditions. This results in high model complexity, which in turn demand large number of data for the model development, and most processes are existing some kind of time-varying behaviour and that requires a strategy for online adaptation.

Just-in-time learning (JITL) is useful to cope with such kind of situation and have attracted extensive attention in process modelling and soft sensor development (Fan et al., 2014). However, there are still some practical challenges. The first key issue is to establish an appropriate similarity criterion for selecting relevant historical samples. Generally, distance between query sample and historical sample is usually used to design similarity criterions, such as Mahalanobis distance (Nakabayashi et al., 2010) and Euclidian distance (Ito et al., 2004). According to Liu et al. (2012), only utilization of the distance for description of the similarity is not comprehensive, then a new similarity criterion was proposed to select samples adopting the distance and angle between two samples and it showed better performance than only based on distance. However, correlations among variable are neglected in the above mentioned two similarity criterions. Consequently, some good data may not be selected. To this end, correlation based similarity criterion was proposed by Fujiwara et al. (2012). But it was pointed out that it is difficult to obtain optimum parameters of the correlation based criterion (Saptoro, 2014). Besides, it should be noticed that large noise or outliers always exist in process data. Empirically, those in historical input data predefined in the process can be identified easily. However, those in historical output data are difficult to be detected because of complex process nonlinearity.

Please cite this article as: Mei C., Chen Y., Jiang H., Ding Y., Chen X., Liu G., 2017, Just-in-time modeling with a combination of input and output similarity criterions for soft sensor modeling in fermentation processes, Chemical Engineering Transactions, 61, 1045-1050 DOI:10.3303/CET1761172

In a word, the existing similarity criterions do not consider the quality of output data in samples. It means that large noise or outliers in output data would result in inappropriate estimates (Saptoro, 2014).
In this study, a novel combined similarity criterion is proposed for JITL based soft sensor modelling. In the proposed similarity combination of input and output (SCIO), the typical distance and angle based similarity criterion is still used and a new similarity criterion of output is designed based on the idea of membership functions in clustering algorithms (Yong feng et al., 2008). The effectiveness of the proposed SCIO for JITL modelling method through an industrial fed-batch Erythromycin fermentation process.

2. Similarity combination for relevant sample selection and soft sensor modelling

In the JITL modelling, it is crucial to construct similarity criterion for selecting suitable training samples matching the query sample. It is nature to use distance to describe similarity between two data points. The study of Saptoro (2014) gives few variants of distance based similarity criterions: Eucledian distance based, weighted Eucledian distance based and Mahalanobis distance based. The needed JITL based soft sensor is constructed with the relevant data selected by using similarity criterion. However, it was pointed out that the distance-based similarity criterion, commonly utilized in JITL is not comprehensive because of only considering the distance of input. The distance and angle based similarity criterion was proposed to describe similarity of input comprehensively. Now, it is popular in the field of JITL modelling.
The distance and angle based similarity criterion between the query data point and historical data point is defined as follows(Liu et al. 2012)

$$S_{qi} = \omega \exp\left(-d_{qi}\right) + (1-\omega)\cos\left(\theta_{qi}\right) \tag{1}$$

For $\cos\left(\theta_{qi}\right) \geq 0, \quad i = 1,\ldots\ldots k$

$$d_{qi} = \left\| x_i - x_q \right\|_2 \tag{2}$$

$$\cos\left(\theta_{qi}\right) = \left\langle x_i, x_q \right\rangle \Big/ \left(\left\| x_i \right\|_2 \left\| x_q \right\|_2 \right) \tag{3}$$

Where d_{qi} and $\cos\left(\theta_{qi}\right)$ are the distance similarity and the angle similarity between x_q and x_i respectively. $\omega\left(0 \leq \omega \leq 1\right)$ is a weight parameter, and only distance similarity (or angle similarity) is adopted when $\omega = 1$ (or $\omega = 0$). The value of S_{qi} is bounded between 0 and 1, and when S_{qi} approaches 1, x_q closely resembles x_i. It should be noticed that Eq. (1) cannot be used to compute the similarity S_{qi} between x_q and x_i if $\cos\left(\theta_{qi}\right)$ is negative.
It can be seen that only similarity of input is considered in the distance and angle based similarity criterion. It was pointed out that those existing similarity criterions only considering similarity of input can be badly influenced by large noise or outliers (Saptoro, 2014). In fact, input is always predefined in a process, in which large noise and outliers are easy to be identified. However, the quality of output cannot be judged easily for lacking predefined references. With large noise or outliers in output, selected relevant samples used for JITL modelling with conventional similarity criterions may result in inappropriate predictions.
For selected relevant samples by using the distance and angle based similarity criterion, it can be assumed that most of them are appropriate and close to each other. It means few samples with large noise or outliers are far to most of the selected samples. A distance based index was designed to evaluate the quality of output, which is described as follows

$$S'_{yi} = \exp(-d'_{yi}) \tag{4}$$

$$d'_{yi} = \sum_{k=1, k\neq i}^{L'} \sqrt{\left(y_i - y_k\right)^2} \tag{5}$$

Where y_i and y_k are outputs of the ith and the kth selected relevant samples with a conventional similarity criterion respectively. It is obvious that the values of S'_{yi} corresponding to the samples with large noise or outliers must be much smaller than those normal selected relevant samples.

To combine similarity of input and output and avoid inappropriate selected samples, a similarity combination can be defined as

$$h(i) = S_{qi}(i) S'_{yi}(i) \tag{6}$$

The Eq.6) can be interpreted that only those samples with similar inputs to the query sample and without larger noise or outliers in outputs can achieve high values of $h(i)$ The detailed steps of JITL soft sensor modeling based on the proposed combination Eq(6) are list as follows:

(1) Set the value of ω and relevant sample size L'.

(2) For a new query sample, use Eq(1) -(3) to calculate the similarity S_{qi}.

(3) Sort S_{qi} in descending order and choose the first L' relevant samples.

(4) For the L' relevant samples, use Eq(4) - (6) to calculate $h(i)$.

(5) Sort $h(i)$ in descending order, then select first $L(< L')$ relevant samples for JITL modelling.

(6) After outputting predictions related to the query samples, discard the JITL model.

3. Case study

In this study, GPR is used to construct JITL based soft sensors for its advantages of less parameters, easily optimizing and modelling uncertainty (Rasmussen and Williams, 2006). For comparisons, two above-mentioned conventional similarity criterions, i.e. the Mahalanobis distance (MD) based similarity criterion and the distance and angle (DA) based similarity criterion, are also studied. To evaluate different methods, the estimation error between predictions and real values (Error) and the root-square error (RMSE) are used and defined as follows

$$Error = \hat{y}_i - y_i \tag{7}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{8}$$

Where $\hat{y}_i$ and y_i are the predicted and observed values respectively, n denotes the number of query samples, and $i = 1,2,\ldots,n$ .The RMSE values are commonly used to indicate prediction accuracy of soft sensors.

For an Erythromycin fermentation process, biomass concentration plays a decisive role in the final product (Erythromycin) concentration. So, the primary way of ensuring product quality of Erythromycin is to control biomass concentration which can be affected by many process factors. In this example, 182 samples are selected as the query data, and 1274 samples are used as the database. The relevant samples to a query sample are selected for JITL modelling from the database. Every sample contains fifteen input variables and one output variable.

After variable selection through the principal component analysis based method described by Shakil et al. (2009), five input variables, i.e. DO saturation, pH, Temperature, Agitator power, Aeration rate, are selected as secondary variables, and the output variable is biomass concentration. The data characteristics of the secondary variables and the primary variable are shown in Figure 1. From the figure, it can be easily seen that the process has strong nonlinear characteristics. The predicted outputs with different relevant sample selection methods are shown in Figure 2. Also, prediction errors of three methods for this case are shown in Figure 3. In here a query sample comes, L (=15) training samples are selected from the database for JITL based soft sensor modelling. From the figure, it can be easily seen that SCIO performs the best.

The three criterions were studied for soft sensor modelling with varying L . In this case, $L' = L + 5$. The RMSEs for query samples with different similarity criterions and varying L are shown in Figure 4. These results suggest that MD and DA based similarity criterions do not function very well, and the proposed SCIO can track the nonstationary behaviour of the process data more closely. In fact, inputs of a plant are always known in advance. However, the outputs of that depend on the performance of signal transfer. Therefore, it is easy to find out those incorrect inputs. But, great errors in outputs are difficult to be identified. It is necessary to consider the quality of output data for improving prediction performance of JITL based soft sensors. In our method, the similarity index is used to overcome the effects of those outputs with great errors or noises.

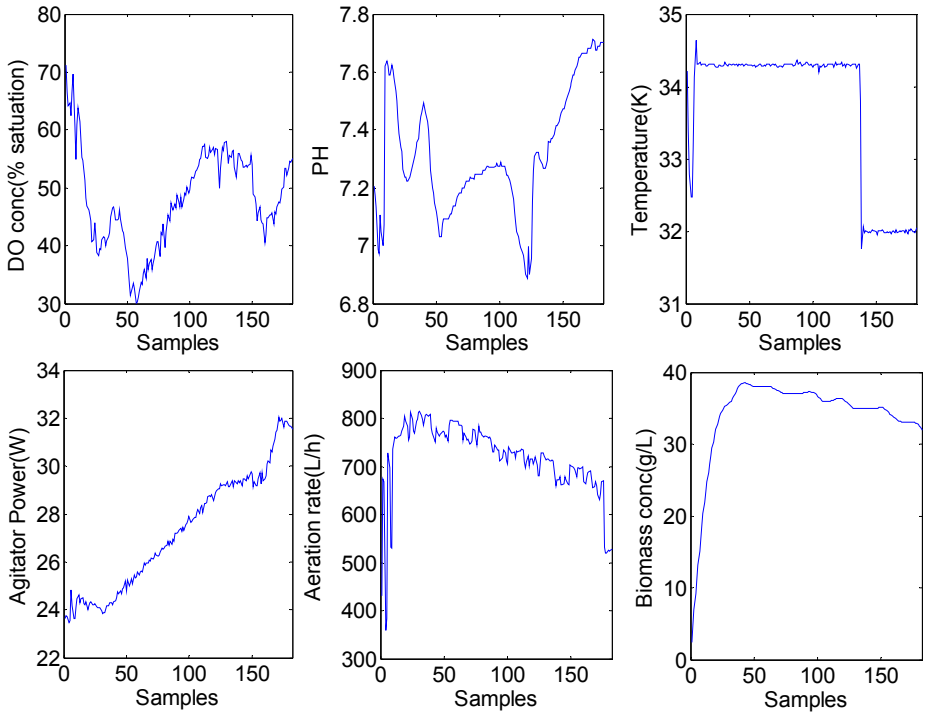

Figure 1: Data characteristics of the input and output variable in the fed-batch Erythromycin fermentation process.

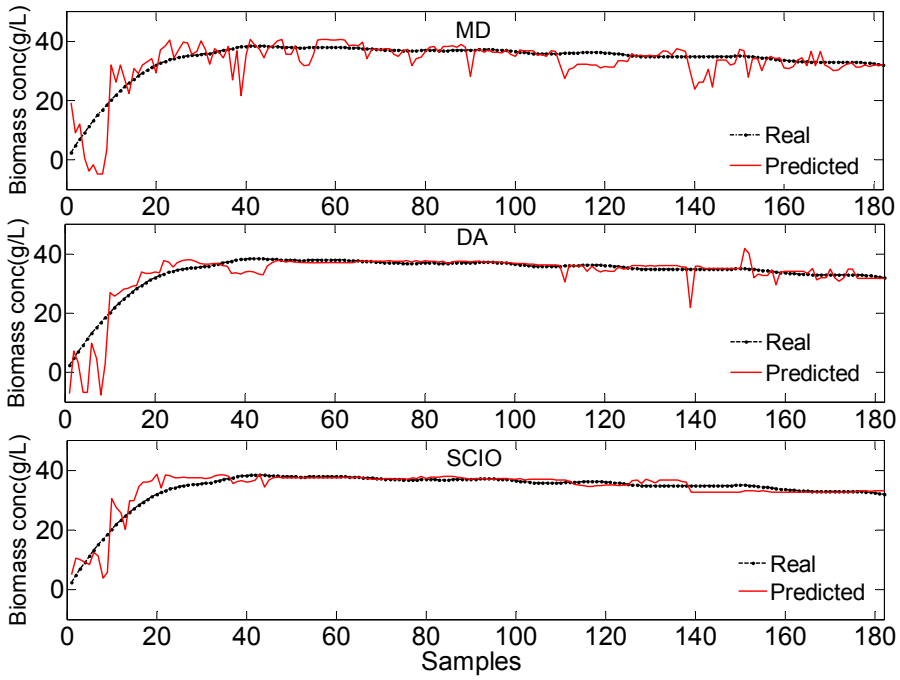

Figure 2: Prediction results with different similarity criterions in the fed-batch Erythromycin fermentation process.

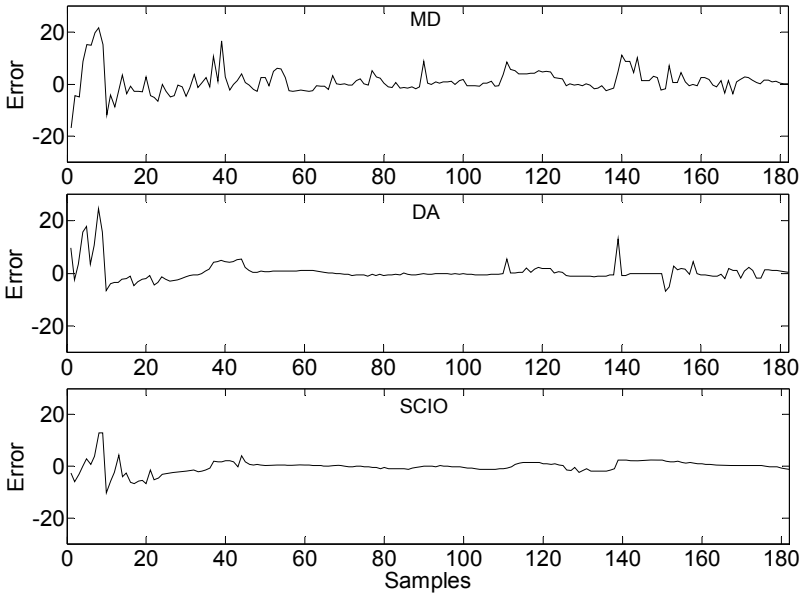

Figure 3: Prediction errors with different similarity criterions in the fed-batch Erythromycin fermentation process.

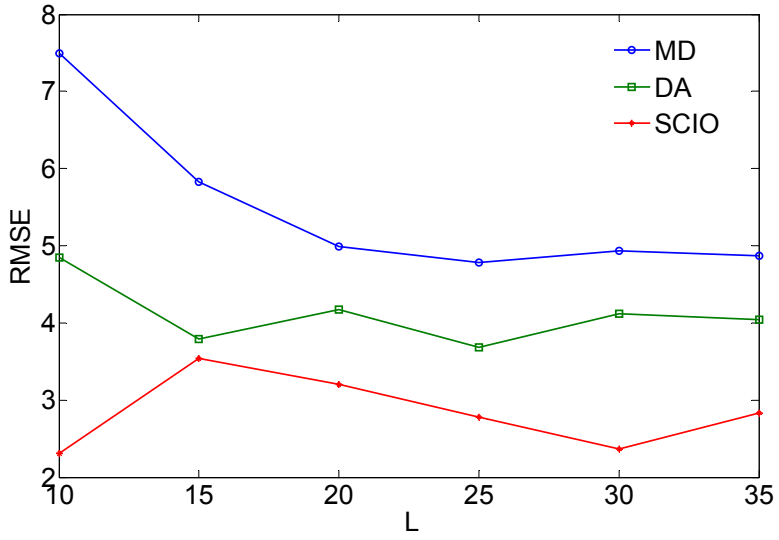

Figure 4: RMSEs with varying L in fed-batch Erythromycin fermentation process.

4. Conclusions

Conventional similarity criterions can only be used to describe the similarity between inputs of a query sample and a historical sample. However, large noise or outliers are often in outputs of historical samples. It may result in selecting inappropriate relevant samples by using conventional similarity criterions for JITL based soft sensor modelling. In this paper, a novel similarity combination is proposed for JITL based soft sensor modelling. In the criterion, conventional distance and angle based similarity criterion was combined with a novel quality criterion of output. After using the presented method, inappropriate samples with large noise or outliers can be avoided to be selected for JITL modelling. An industrial case is used to verify the proposed method. Results show that the proposed similarity combination performs better than the MD based similarity criterion and the distance and angle based similarity criterion.

Acknowledgments

The authors gratefully acknowledge the financial support provided by Natural Science Foundation of Jiangsu Province of China (Grant No. BK20130531, BK20140538), the Priority Academic Program Development of Jiangsu Higher Education Institutions (Grant No. PAPD 6), the Graduate practical innovation Foundation of Jiangsu province (Grant No. SJLX16_0441).

References

Fan M., Ge Z., Song Z., 2014, Adaptive Gaussian Mixture Model-Based Relevant Sample Selection for JITL Soft Sensor Development, Industrial & Engineering Chemistry Research, 53, 19979-19986.

Fujiwara K., Kano M., Hasebe S., 2012, Development of correlation-based pattern recognition algorithm and adaptive soft-sensor design, Control Engineering Practice, 20, 371-378.

Ge Z., Huang B., Song Z., 2014, Nonlinear semisupervised principal component regression for soft sensor modeling and its mixture form, Journal of Chemometrics, 28, 793-804

Ito M., Matsuzaki S., Odate N., Uchida K., Ogai H., Akizuki K., 2004. Large scale database online modeling for blast furnace, Control Applications, 2004, Proceedings of the 2004 IEEE International Conference on, IEEE, pp. 906-911.

Jin H., Chen X., Wang L., Yang K., Wu L., 2015a, Adaptive soft sensor development based on online ensemble Gaussian process regression for nonlinear time-varying batch processes, Industrial & Engineering Chemistry Research, 54.

Jin H., Chen X., Yang J., Zhang H., Wang L., Wu L., 2015b, Multi-model adaptive soft sensor modeling method using local learning and online support vector regression for nonlinear time-variant batch processes, Chemical Engineering Science, 131, 282-303.

Kadlec P., Gabrys B., Strandt S., 2009, Data-driven soft sensors in the process industry, Computers & Chemical Engineering, 33, 795-814.

Liu Y., Gao Z., Li P., Wang H., 2012, Just-in-Time Kernel Learning with Adaptive Parameter Selection for Soft Sensor Modeling of Batch Processes, Ind.eng.chem.res , 51(11), 4313–4327.

Nakabayashi A., Nakaya M., Ohtani T., Chen D., Wang D., Li X., 2010, A process simulator based on hybrid model of physical model and just-in-time model, Proceedings of SICE Annual Conference, pp. 97-100.

Pani A.K., Vadlamudi V.K., Mohanta H.K., 2013, Development and comparison of neural network based soft sensors for online estimation of cement clinker quality, ISA transactions, 52, 19-29.

Rasmussen C.E., Williams C.K.I., 2006, Gaussian processes for machine learning, MIT Press, 14(481), 69-106

Saptoro A., 2014, State of the art in the development of adaptive soft sensors based on just-in-time models. Procedia Chemistry, 9, 226–234.

Shakil M., Elshafei M., Habib M.A., Maleki F.A., 2009, Soft sensor for NOx and O2 using dynamic neural networks. Computers & Electrical Engineering, 35, 578–586.

Wang Z.X., He Q.P., Wang J., 2015, Comparison of variable selection methods for PLS-based soft sensor modeling, Journal of Process Control, 26, 56-72.

Yongfeng F.U., Hongye S.U., Zhang Y., Chu J., 2008, Adaptive soft-sensor modeling algorithm based on FCMISVM and its application in PX adsorption separation process, Chinese Journal of Chemical Engineering, 16(5), 746-751.