Research on Soft Sensor Modeling Method and Its Application in Wastewater Treatment

Mingzhu Li*, Shouxing Zhang and Yuanyuan Chen
Haikou University of Economics, Haikou 571127, China

*Corresponding author email: limingzhu110abc@163.com

Abstract. In order to solve the problem of real-time monitoring of water quality index which is difficult to measure in the process of wastewater treatment, a dynamic modeling method of soft sensor based on the VIP (Variable Importance in Projection) and JIT (Just-in-Time) methods was proposed. Firstly, the important variables were selected by the VIP method, and then the data were updated by the JIT method. Finally, the soft sensing model of PLS (partial least squares) was established. The model was used to predict BOD5 which was an important water quality parameter in the process of wastewater treatment. The simulation results indicate that the proposed model achieves better mean square error and correlation coefficient.

Keywords: Soft sensor, variable selection, just in time, wastewater treatment.

1. Introduction
The problem of water pollution has been paid more and more attention by the state and relevant government departments. The state has formulated a series of laws and regulations to restrict the treatment and discharge of wastewater. In the process of wastewater treatment, there are many variables closely related to the safety of the production process, which are difficult to be measured directly online by hardware sensors [1]. In order to improve the qualified rate of effluent, save energy, reduce operating costs and improve system reliability, it is necessary to obtain the important process parameters and water quality parameters of wastewater treatment system in time [2]. Soft sensor is an effective way to solve this problem [3]. The soft sensor has the advantages of rapid response, low cost, relatively simple maintenance, etc. [4], which makes it applied in many industrial processes that are difficult to be measured by hardware sensors.

In this paper, the selection of auxiliary variables and online updating methods are studied to further improve the performance index of the model, so that the important water quality parameters in the process of wastewater treatment can be tracked in time, providing further guarantee for the safe production of sewage, and saving the cost of hardware instruments.

2. VIP Method
In the process of wastewater treatment, there are many measurable variables, which put forward the requirements for variable selection. Choosing appropriate auxiliary variables before establishing soft sensor model can not only reduce the sample data, but also strengthen the explanatory role of variables.

Through the selection of auxiliary variables, the auxiliary variables unrelated to the output variables are eliminated, making the prediction effect of the soft-sensing model more accurate [5]. Many variable selection methods have been proposed, but it is difficult to propose a valid variable selection method for all objects [6]. The common variable selection methods include VIP method[7], genetic algorithm [8], moving window [9], etc. In this thesis, the VIP method will be further studied.
The VIP method is a method for variable selection, and it is based on the PLS method (partial least squares regression). The variable is selected by accumulating its importance in the PLS model. In this method, VIP index can be used to measure the importance of auxiliary variables to target variables, so as to select the auxiliary variables that contribute greatly to target variables. The VIP index can be calculated by formula (1).

\[
VIP_j = \sqrt{p \sum_{k=1}^{l} \left[ \sum_{i=1}^{l} q_{ik}^2 \right] / \| w_k \|^2}
\]

(1)

Where \( p \) is the number of auxiliary variables, \( t_k \) is the \( k \)-th column element of the score vector, and \( q_{ik} \) is the \( k \)-th column element of the regression coefficient. \( w_k \) shows the importance of the variable \( j \). Setting the threshold \( TH \), when \( VIP_j < TH \), the variable \( j \) can be eliminated. It is generally believed that the screening effect is better when \( TH \) is 1. However, some literature [10] believed that \( TH \) between 0.83 and 1.21 could generate more relevant variables.

3. JIT Algorithm

Although soft sensor has achieved great development, but most of the researches are based on static model, that is, global model. With the passage of time, the static model will lead to the change of the operating point of the system, making the soft-sensing model is no longer suitable for the application of working conditions. So the model needs to be updated in time.

The JIT algorithm was proposed by Bontempi G in 1999[11], which has gained attention due to its good adaptability [12]. Based on the principle of "similar input produces similar output", the algorithm memorizes the sample data in the memory, finds the similar data in the sample data according to the input points, and establishes some local models of the system on-line rolling with the changes of the system working points, so as to meet the needs of real-time modeling and control of nonlinear systems. Similar data can be expressed by the Euclidean distance between the samples. If the distance is within the specified range, the two samples are considered to be similar. The Euclidean distance between two points is expressed as formula (2).

\[
d(x_i, x') = \| x_i - x' \|_2
\]

(2)

If the local space composed of \( k \) sample points near the input point \( x' \) is \( \Phi_k \), then the data close to \( x' \) can be expressed as formula (3).

\[
\Phi_k = \{ x_i \mid d(x_i, x') < L \}
\]

(3)

Where \( L \) is the window length and \( i \) ranges from 1 to \( k \).

4. Soft Sensor Modeling Steps

The steps of soft sensor modeling for wastewater treatment are as follows:

Step 1: Requirement analysis. According to the characteristics and actual situation of the wastewater treatment system, the system needs are analyzed and the variables to be measured are defined.

Step 2: Build model. The principle, structure and characteristics of the soft sensing method are analyzed, and the dynamic model of soft sensing is established.

Step 3: Simulation experiment. After the singular value and missing data are processed, 70% of the data are used to train the model and 30% to verify the model. The system simulation is carried out by MATLAB to further analyze the prediction effect of the algorithm in the application of wastewater treatment.

5. Case Analysis

In this paper, the important water quality parameter BOD₅ which is difficult to measure in wastewater treatment is taken as the research object. The auxiliary variables are selected through the VIP method, combined with the on-line update method of JIT, a systematic PLS soft sensing dynamic model is established, in order to further improve the prediction effect of the algorithm.

Three models are established to predict BOD₅. The models are PLS model, VIP-PLS model (the model combining VIP variable selection), and JIT-VIP-PLS model (the model combining VIP variable
selection and JIT method). The predictive effect is displayed in Figure 1, Figure 2 and Figure 3, respectively.

**Figure 1.** The prediction of PLS model.

**Figure 2.** The prediction of VIP-PLS model.

**Figure 3.** The prediction of JIT-VIP-PLS model.
The mean square errors (MSE) and the correlation coefficients (CC) of the three models are shown in Table 1.

| Model        | MSE   | CC     |
|--------------|-------|--------|
| PLS          | 9.2257| 0.8676 |
| VIP-PLS      | 6.7137| 0.9064 |
| JIT-VIP-PLS  | 5.3142| 0.9253 |

From figure 1-3 and table 1, it can be seen that the prediction accuracy of the system is improved after variable selection. At the same time, the prediction accuracy is further improved after online updating model, which illustrates the effectiveness of the proposed algorithm.

6. Conclusion
In this paper, the dynamic model of PLS soft sensing model is established by combining the VIP method and the online updating method of JIT algorithm, which provides an effective solution for the real-time measurement of BOD\textsubscript{5} in wastewater treatment. The research is helpful to improve the prediction accuracy of the model, make the important water quality index can be tracked in time, provide the basis for advanced control and optimization of production, and provide further guarantee for the safe production, and save the cost of hardware instrument. It has a certain engineering application value.

Acknowledgments
This research was supported by Hainan Provincial Natural Science Foundation of China (project number: 618QN254), Hainan Provincial Natural Science Foundation of China (project number: 620QN285), Haikou University of Economics Scientific Research Project (HJKY (ZD) 20-10), Haikou University of Economics Scientific Research Project (HJKY (ZD) 21-01).

References
[1] Chen Zhaobo, Ren Yueming. Wastewater Treatment Plant, Measurement, Automatic Control and Fault Diagnosis [M]. Beijing: Chemical Industry Press, 2008:1-20.
[2] Guan Qiu. Research on Soft-sensing Method Based on Neural Network for Wastewater Treatment [D]. Zhejiang University of Technology, 2005.
[3] Cao Pengfei, Luo Xionglin. Modeling of soft sensor for chemical process[J].CIESC Journal, 2013, 64(3): 788-800.
[4] Lai Huige, Zhu Xuejun,Yu Jinshou. Research on Intelligent Algorithm-based Soft-sensing Technology for Sewage Treatment Process [J]. Control and Instruments in Chemical Industry, 2011, 38(03): 241-244+278.
[5] Wang Jian, Yang Yaoquan, Ma Gaowei. Research on Method of Soft Measurement Secondary Variables Selection [J].Electric Power Science and Engineering, 2011, 27(7):37-40.
[6] Mehmood T, Liland K H, Snipen L, et al. A review of variable selection methods in Partial Least Squares Regression[J]. Chemometrics & Intelligent Laboratory Systems, 2012, 118(none): 62-69.
[7] Gosselin R, Rodrigue D, Duchesne C. A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications[J]. Chemometrics and Intelligent Laboratory Systems, 2010, 100(1):12-21.
[8] Hasegawa K, Miyashita Y, Funatsu K. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists[J]. Journal of Chemical Information & Computer Sciences, 1997, 37(2):306-310.
[9] Ypda D, Yzl B, Jhj C, et al. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares[J]. Analytica Chimica Acta, 2004, 501(2):183-191.
[10] CHONG I G, JUN C H. Performance of some variable selection methods when multicollinearity is present [J]. CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 2005, 78(1): 103-112.

[11] Bontempi G, Birattari M, Bersini H. Recursive lazy learning for modeling and control [C] // European Conference on Machine Learning. Springer-Verlag, 1999.

[12] Kaneko H, Funatsu K. Moving Window and Just-In-Time Soft Sensor Model Based on Time Differences Considering a Small Number of Measurements [J]. Industrial & Engineering Chemistry Research, 2015, 54(2): 150104102633007.