Chemometrics and Intelligent Laboratory Systems 151 (2016) 78–88

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Soft sensor development for the key variables of complex chemical processes using a novel robust bagging nonlinear model integrating improved extreme learning machine with partial least square

YanLin He, ZhiQiang Geng, QunXiong Zhu *

College of Information Science & Technology, Beijing University of Chemical Technology, Beijing 100029, China
Engineering Research Center of Intelligent PSE, Ministry of Education of China, Beijing 100029, China

ARTICLE INFO

Article history:
Received 5 September 2015
Received in revised form 7 December 2015
Accepted 14 December 2015
Available online 28 December 2015

Keywords:
Partial least square
Extreme learning machine
Bagging
Soft sensor
Tennessee Eastman process
Purified Terephthalic Acid Process

ABSTRACT

Some key variables in the complex chemical processes are very difficult to measure due to the nonlinearity, the disturbances, and the technological limitations. In order to accurately predict the difficult-to-measure variables, soft sensor based on a novel robust bagging nonlinear model integrating improved extreme learning machine with partial least square (RB-PLSIELM) is developed. Motivated by the ensemble ideas, the proposed RB-PLSIELM model is based on the bagging ensemble scheme to combine some individual nonlinear models integrating improved extreme learning machine with partial least square (PLSIELM). The sub-data for building the individual PLSIELM model are re-sampled from the original training data using the bagging tool. The problem of over-training in the PLSELM model can be avoided by using the bagging re-sampling technology. The proposed RB-PLSIELM model was demonstrated by applying it to predicting the key variables of the Tennessee Eastman Process (TEP) and the Purified Terephthalic Acid Process (PTAP). The simulation results obtained by RB-PLSIELM are compared with those obtained by the individual PLSELM model, the ELM model, and the partial least square regression (PLSR) model. Compared with the other models, the RB-PLSIELM can achieve higher prediction accuracy and stability.

1. Introduction

Advanced control systems play a more and more important role in the modern industries. However, the completion of advanced control systems is not very easy since there are many complicated interactions in complex industrial processes [1–3]. The modeling methods are quite important for designing advanced control systems. Until now, many efforts have been made to develop accurate models for predicting the key variables in complex processes [4–6]. Developing an accurate model for a certain system has been an active topic in control field for over many years. There are two well-known kinds of modeling methods: the first-principles mechanistic or physical method and the data-driven method. The mechanistic or physical models are established by analyzing the detailed knowledge of industrial processes. However, it takes many efforts and much time to establish a mechanistic or physical model due to the increasing complexity of the industrial processes [7]. The knowledge of the complex processes is more and more difficult to obtain. In addition, when one establishes the mechanistic or physical model, many difficult differential and algebraic equations with unknown parameters need to be taken into consideration. Thus, it is more and more complicated to develop a mechanistic or physical model for complex industrial processes.

Compared with the mechanistic or physical models, the data-driven models are much easier to develop. The data-driven method has attracted more and more attention from researchers [8–10]. On one hand, the knowledge of the complex processes is not necessary for developing the data-driven models. Only the related data of the complex processes are needed when constructing the data-driven models. That is to say, the difficulties in obtaining the knowledge of the complex processes during developing the mechanistic or physical models are avoided in building the data-driven models. On the other hand, with the development of the Distributed Control System (DCS), the process data is much easier to obtain for developing the data-driven models [11]. Hence, it is more and more practical to develop data-driven models than mechanistic or physical models. Among the data-driven methods, there are two widely applied categories: the statistical regression (SR) based data-driven methods and neural networks (NNs) based data-driven methods. Partial least square (PLS) method, as one of SR data-driven methods, has been widely and successfully adopted to develop soft sensors for industrial processes [12,13]. Compared with the SR methods, the NNs have a better ability to deal with the high nonlinear dataset. Thus, the NNs based methods can be adopted as a promising tool for developing data-driven models for complex processes. NNs have some salient features: the learning ability, the parallel computing,

* Corresponding author. Tel.: +86 10 64426960; fax: +86 10 64437805.
E-mail address: zhuqx_buctielab@163.com (Q. Zhu).

http://dx.doi.org/10.1016/j.chemolab.2015.12.010
0169-7439/

and the universal functional approximation [14]. So NNs have been successfully applied to building models for many complex and highly nonlinear processes [15–17]. There are many kinds of NNs, such as feedforward NNs, recurrent NNs, functional link NNs, and so on. Among all the NNs, the single hidden layer feedforward NNs have been widely used due to their easiness in the structure and good ability in the generalization performance [18].

Recently, a novel single hidden layer NN named extreme learning machine (ELM) has been proposed by Huang et al. in 2004 [19–22]. ELM has a brand new learning method that is different from the well-known error back propagation (EBP) method. In ELM, a random and mathematical method is adopted to train the ELM. The weights between the input layer nodes and the hidden layer nodes are generated using a random manner; however, the weights between the hidden layer nodes and the output layer nodes are calculated out using the Moore-Penrose generalized inverse method. Compared with the learning speed of the EBP based NNs, ELM has an extremely fast learning speed. Many parameters like the learning rate and the learning epochs in the EBP method have been avoided in the ELM model. And the generalization performance of ELM has been proved to be better than many other NNs [22]. That is to say, ELM can be adopted to build a data-drive model with simple structures, fast learning speed, and well generalization performance. Therefore, ELM has been widely and successfully applied as an effective tool in many fields, such as classification [20], modeling [23], prediction [24], control [25], and so on.

In neural networks based models, how to select the optimal hidden layer nodes number is a very important problem that needs to be solved. Too many hidden layer nodes may make the neural networks be over-fitting. However, too few hidden layer nodes may weaken the nonlinear mapping performance of the neural networks. So some researchers have proposed several methods to select the proper number of the hidden layer nodes. He et al. adopted the trial-and-error method to determine the proper number of the hidden layer nodes in ELM [18]. However, the trial-and-error method consumes much time. The proper number of the hidden layer nodes can be determined by using pruning algorithm [26]. However, the nodes number that adding or deleting in the pruning algorithm is very difficult to determine. In addition, the output weights matrix of ELM tends to be ill-conditional when the number of the hidden layer nodes exceeds the number of the samples. What is more, the least square method used in ELM cannot well handle the strong multicollinearity among the outputs of the hidden layer nodes. Zhao et al. proposed an ELM based on partial least square (PLSELM) model to estimate the effluent quality [27]. In the PLSELM model, the weights between the hidden layer nodes and the output layer nodes are calculated out by the partial least square method instead of the least square method. The partial least square can well deal with the multicollinearity among the outputs of the hidden layer nodes when a large number of hidden layer nodes are assigned. Additionally, the over-fitting problem also can be avoided in PLSELM. When a large number of hidden layer nodes are assigned in ELM, the latent variables selected in the partial least square can prevent the model from being over-fitting.

Although the over-fitting and the multicollinearity problems have been solved in the PLSELM model, the performance of PLSELM is still limited. The training of PLSELM model may suffer from over-training. The over-training has a bad effect on the performance. Under this situation, PLSELM model can learn the training data with acceptable errors but it cannot guarantee that the generalization performance on the testing dataset is good. The performance of individual model, as well as the generalization performance of the individual model can be enhanced by using an ensemble of those individual models [28–30]. There are two popular ensemble machine-learning methods: boosting and bagging [31,32]. The ensemble method combines the output of some individual models to enhance the overall ensemble model results. Some researchers have used the ensemble model to enhance the performance of the individual model. Niu et al. adopted bagging based ensemble neural network model to improve the accuracy of the individual neural network model [33]. Tian and Meng used bagging based ensemble ELM to predict the day-ahead electricity price [34].

In this paper, soft sensor development using a novel robust bagging nonlinear model integrating improved extreme learning machine with partial least square (RB-PLSIELM) is proposed. On one hand, the RB-PLSIELM model can avoid the over-training problem by using the bagging re-sampling technique. And then the RB-PLSIELM model can enhance the performance of the individual model. On the other hand, an improved ELM with the double parallel structure is adopted. In the double parallel structure, there are connections between the input layer nodes and the output layer nodes [35]; however, in the traditional structure, there are not those connections. Thus, the double parallel structure can make PLS be able to extract the latent variables not only from the outputs of the hidden layer nodes but also from the original inputs, which can further enhance the performance of the individual model. Lastly, the weights between the input layer nodes and the hidden layer nodes are randomly generated in individual models. Then the bad influence of the random input weights on the performance can be alleviated by using the ensemble technology. In the RB-PLSIELM model, the individual model is based on the nonlinear model integrating improved extreme learning machine with partial least square (PLSIELM), and the bagging tool is used to generate the sub-data by re-sampling from the original data. Then the sub-data are adopted to build the individual PLSIELM models. In order to test the performance of the proposed RB-PLSIELM model, RB-PLSIELM was applied to developing soft sensor for predicting the key variables of the Tennessee Eastman Process (TEP) and the Purified Terephthalic Acid Process (PTAP). The simulation results obtained by RB-PLSIELM were compared with those obtained by the individual PLSELM model, the ELM model, and the partial least square regression (PLSR) model. The results demonstrated that the proposed RB-PLSIELM model achieved higher prediction accuracy and more robust generalization ability.

The remaining parts of this paper are organized as follows: in Section 2, the bagging, the PLS algorithm and the basic ELM algorithm are briefly introduced; the proposed RB-PLSIELM model is described in

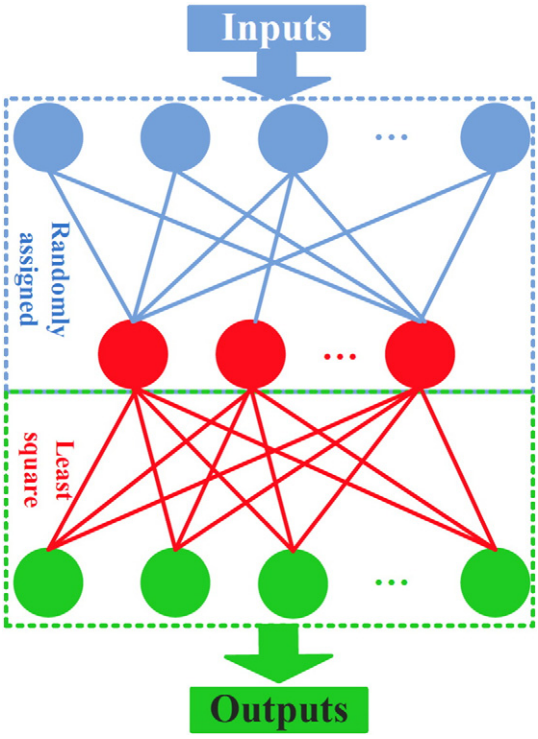

Fig. 1. Structure of the traditional ELM.