Chemometrics and Intelligent Laboratory Systems 151 (2016) 181–189

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Active probabilistic sample selection for intelligent soft sensing of industrial processes

Zhiqiang Ge *

State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, PR China

ARTICLE INFO

Article history:
Received 28 September 2015
Received in revised form 30 December 2015
Accepted 6 January 2016
Available online 13 January 2016

Keywords:
Soft sensor
Limited labeled samples
Active learning strategy
Gaussian process regression
Probabilistic modeling

ABSTRACT

This paper proposes a new active learning strategy based soft sensor upon the Gaussian process regression (GPR) model, in order to improve the prediction performance under a limited number of labeled data samples. The main objective of the new soft sensor is to opportunely label data samples in such a way as to maximize the soft sensing performance while minimizing the number of samples used, and thus to reduce the costs related to human efforts. By taking advantage of the GPR model, the information of prediction uncertainty is used to make a new probabilistic sample selection strategy, upon which the active learning GPR model is formulated for soft sensing. Detained analyses and comparative studies are carried out between the active learning strategy driven GPR model and random selection strategy driven GPR model through an industrial case study.

1. Introduction

In the past years, soft sensing of key variables in industrial processes has become more and more important, due to the increased requirements of product quality measurement [1]. Different from conventional process variables such as temperature, pressure or stream flow which can be measured frequently, some other variables are difficult to obtain through online measurement, e.g. product quality-related indices, element concentrations in stream or gas flows, or viscosity and melt variables. In modern process industries, it is highly desired that those important difficult-to-measure variables be obtained online, based on which a feedback quality control system can be generated. Typically, principle model-based soft sensors/virtual sensors have been constructed for online estimation of key variables in the process industry. However, due to the complexity of modern process industries, it is very difficult to obtain a detailed and reliable principle model.

In contrast, with the wide use of distributed control systems in modern process industries, a large amount of data have been collected, based on which various data-based soft sensing approaches have been developed [2,3]. Compared to those, which are mainly based on the knowledge of the process, the data-based soft sensing methods are much more flexible, and thus have been used more widely in the process industry [4,5]. To date, plenty of data-based soft sensing or quality estimation methods have been proposed, such as principal component regression (PCR) and partial least squares (PLS) for linear processes, artificial neural network (ANN) and support vector machine (SVM) for nonlinear processes, etc. [6–20]. Recently, a new probabilistic nonlinear modeling method namely Gaussian process regression (GPR) has caught much attention in this area. It is demonstrated that a large class of neural network based nonlinear regression models will converge to an approximate Gaussian process regression model. Several comparative studies have shown that the GPR model performs better than other nonlinear modeling approaches [21–26]. Another advantage of the GPR model is that it can provide a probabilistic estimation result, which means the estimation uncertainty of the key variable can be obtained. Therefore, the GPR model has been considered as an attractive nonlinear modeling method for soft sensor development.

For data-based soft sensor development, conventionally, samples of both ordinary process variables and target variables (those variables to be estimated) are required. Nevertheless, while the data for process variables can be easily recorded, the acquirement of target variables is much more difficult which always incorporates expensive instruments, laboratory analyses, or significant human efforts. As a result, we may only have a very limited number of data samples for the target variables, and have a large amount of datasets for other process variables. A solution of this problem is given by semi-supervised learning methods, in which the regression model has been designed to incorporate both labeled and unlabeled data samples. Here, we refer to the data sample with both process and target variables as the labeled sample and to the one that only contains ordinary process variables as the unlabeled sample. Traditional semi-supervised learning methods include self-training based methods, probabilistic generative model based methods, co-training methods, graph-based methods and etc. [27,28]. However, while the semi-supervised learning method does improve the regression performance, it may also introduce significant computational effort

* Tel.: +86 87951442.
E-mail address: gezhiqiang@zju.edu.cn.

http://dx.doi.org/10.1016/j.chemolab.2016.01.003
0169-7439/

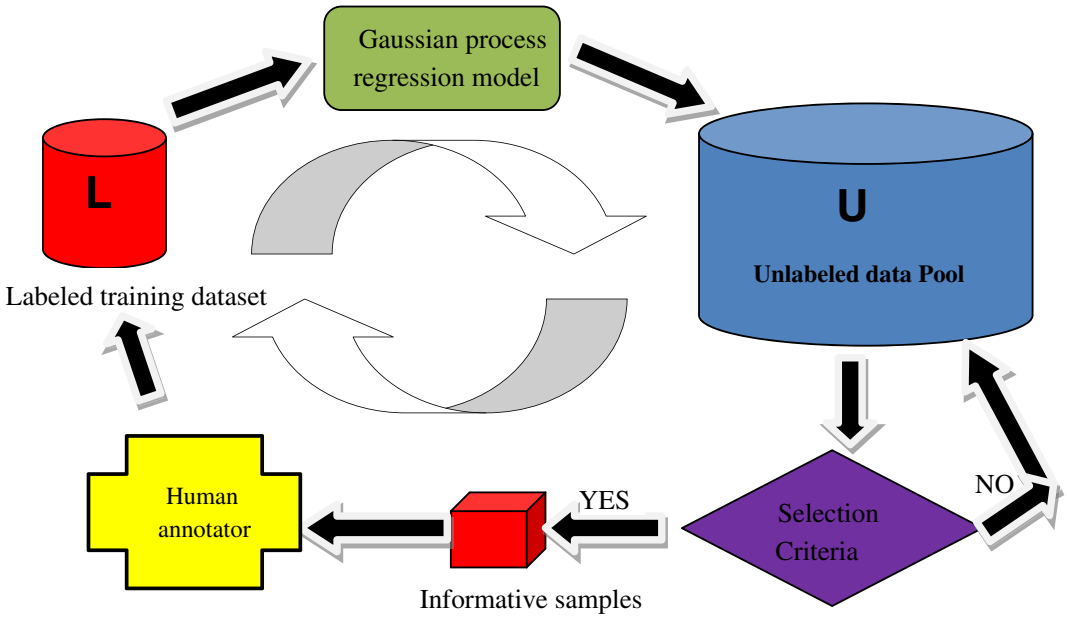

Fig. 1. Active learning modeling procedure based on the GPR model.

to the system. Besides, the extent of improvement often depends on the designed structure of the semi-supervised model.

Alternatively, active learning is also an attractive approach which can successfully handle the regression problem under limited labeled data samples [29]. The main idea of the active learning method is to opportunely collect labeled data samples in such a way as to maximize the estimation performance of the regression model while minimizing the number of labeled samples, in order to reduce the efforts related to sample labeling. Therefore, based on the limited number of labeled data samples, an important task of the active learning method is how to select significant and informative samples from the unlabeled dataset for labeling. Then, those newly labeled data samples are added to the training dataset for soft sensing model development. Through this process, the inner structure of the regression model will not be changed, and the choice of the model structure is left completely open, which could be a simple algorithm such as PCR or PLS, or a very complicated one such as neural network and the kernel-learning method [29].

In the present paper, we intend to incorporate the active learning method into the Gaussian process regression model. Based on the probabilistic structure of the GPR model, a new active learning strategy is proposed for nonlinear soft sensor modeling under limited labeled data samples. In this new active learning GPR model, the information of estimation uncertainty is used as the evaluation index for probabilistic selection of unlabeled data samples from the unlabeled dataset. By labeling those informative samples, the performance of the GPR model can be boosted through the most efficient direction, while the number of labeled data samples is simultaneously minimized.

The rest of this paper is structured as follows. In Section 2, the principle of the Gaussian process regression model is briefly introduced, followed by the detailed methodology of the active learning strategy

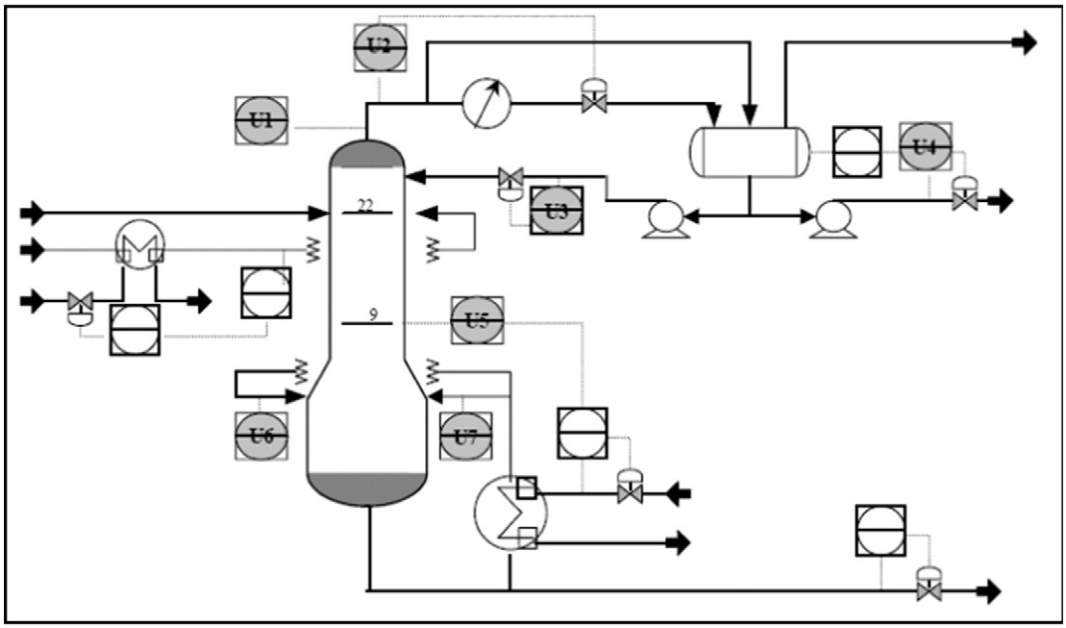

Fig. 2. The flowchart of the debutanizer column [1].