Chinese Journal of Chemical Engineering 23 (2015) 1925–1934

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

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

Supervised local and non-local structure preserving projections with application to just-in-time learning for adaptive soft sensor☆

Weiming Shao, Xuemin Tian *, Ping Wang

College of Information and Control Engineering, China University of Petroleum (Huadong), Qingdao 266580, China

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form 1 July 2015
Accepted 15 July 2015
Available online 11 November 2015

Keywords:
Adaptive soft sensor
Just-in-time learning
Supervised local and non-local structure preserving projections
Locality preserving projections
Database monitoring

ABSTRACT

In soft sensor field, just-in-time learning (JITL) is an effective approach to model nonlinear and time varying processes. However, most similarity criterions in JITL are computed in the input space only while ignoring important output information, which may lead to inaccurate construction of relevant sample set. To solve this problem, we propose a novel supervised feature extraction method suitable for the regression problem called supervised local and non-local structure preserving projections (SLNSPP), in which both input and output information can be easily and effectively incorporated through a newly defined similarity index. The SLNSPP can not only retain the virtue of locality preserving projections but also prevent faraway points from nearing after projection, which endues SLNSPP with powerful discriminating ability. Such two good properties of SLNSPP are desirable for JITL as they are expected to enhance the accuracy of similar sample selection. Consequently, we present a SLNSPP-JITL framework for developing adaptive soft sensor, including a sparse learning strategy to limit the scale and update the frequency of database. Finally, two case studies are conducted with benchmark datasets to evaluate the performance of the proposed schemes. The results demonstrate the effectiveness of LNSPP and SLNSPP.

1. Introduction

Industrial plants are usually equipped with a large number of hardware sensors to deliver data for process monitoring and control [1,2]. However, many important quality related variables are very difficult to obtain, such as product concentration and octane number. Although hardware analyzers provide online measurements of these difficult-to-measure variables, they might not be satisfactory due to their expensiveness, large analysis cycle, and inaccuracy [3]. Some variables, such as the melting index of polypropylene, are mostly available in laboratories only [4]. These technical limitations result in infrequent and inaccurate measurements of these key variables, which may lead to very bad performance of closed loop control system. In some cases, such as sulfur recovery unit [2], the damage of acid gases such as hydrogen sulfide and sulfur dioxide to measuring instruments causes frequent sensor maintenance and replacement with high production cost.

As an alternative, soft sensor is an effective way to solve or alleviate these problems. Although first principal models have the potential for describing the phenomena in a process, the difficulty in modeling complex processes accurately leads to the popularity of data-driven soft sensors [2]. In the past two decades, a variety of algorithms have been applied to develop data-driven soft sensors for industrial processes, such as multivariate statistical regression including principal component regression (PCR) [5] and partial least squares regression (PLSR) [6]. However, both PCR and PLSR are limited to linear processes. One solution to this problem is to develop soft sensors using nonlinear algorithms such as artificial neural networks [7], support vector machines [8,9], and the kernel version of PLS [10]. Although the process nonlinearities are modeled, the performance of soft sensors deteriorates due to the variations of process characteristics as most industrial processes are time-varying [11,12].

Just-in-time learning (JITL) is a commonly used strategy for simultaneously addressing the above two problems, *i.e.*, the process nonlinearity and time-varying. It adopts the philosophy of "divide and rule", where the global model is locally linearized around some sample such that the process nonlinearity can be modeled. Besides, as the historical data set is updated with newly measured samples, JITL can intrinsically cope with the degradation of soft sensor model. Thus soft sensors developed with JITL are called "adaptive soft sensor" [13]. One key step in JITL is to select similar samples of the query sample $\boldsymbol{x}_q$. The more similar the selected samples and $\boldsymbol{x}_q$ are, the better the estimation performance is. Thus our task is to construct the similarity metric. The Euclidean distance is one commonly used similarity criterion [14,15], where the k nearest neighbors (kNN) of $\boldsymbol{x}_q$ are selected as the similar sample set. Besides, k surrounding neighbors [16], k bipartite neighbors [17], and enclosing kNN [18] have also been developed to improve the performance of kNN. In order to further enhance the accuracy of similar

☆ Supported by the National Natural Science Foundation of China (61273160) and the Fundamental Research Funds for the Central Universities (14CX06067A, 13CX05021A).
* Corresponding author.
E-mail address: tianxm@upc.edu.cn (X. Tian).

http://dx.doi.org/10.1016/j.cjche.2015.11.012
1004-9541/

sample set selection, Cheng and Chiu [19] formulated a novel similarity criterion using not only the distance information but also the angle between two samples. The angle between two samples with the query sample as the origin of coordinate can also be used as similarity index for selecting similar sample pairs, which is usually named as correlation coefficient [20,21]. In addition to these distance and/or angle based similarity index, Fujiwara *et al.* [22] proposed to construct similarity criterion by combing the squared prediction error and Hotelling T^2 statistics of PCA model, through which the correlation among process variables can be taken into consideration.

Although these similarity indexes have been widely used to develop JITL based soft sensors, they are computed in the input space only. In other words, the precious and important output information of labeled samples is wasted, which may lead to inaccurate similar sample selection. However, directly selecting similar samples in the input–output space is almost impossible as the query sample contains only secondary variables. Therefore, we intend to fulfill such a task indirectly in the lower dimensional space by using feature extraction algorithm. Chen *et al.* [14] have conducted a good pioneer work in this respect, in which supervised locality preserving projections (SLPP) are developed for regression problem and relevant sample selection is implemented in the feature space, without knowing the output of the query sample. However, locality preserving projections (LPP) aim to preserve the local structure only and does not concern the relative positions of two disconnected samples. Two faraway samples may become near after the projection. In such a situation, the two samples are mistaken as similar, resulting in inaccurate relevant sample selection and undesirable estimation performance. Such an insufficiency of LPP has already been demonstrated by Yang *et al.* [23] in classification problems.

In order to solve such a problem, in this paper, we propose a novel feature extraction algorithm named local and non-local structure preserving projections (LNSPP), so that the virtue of LPP is preserved with nearby points still close after projection and distant samples still faraway in the feature space. Moreover, the singularity in LPP can be avoided through imposing an orthogonal constraint. This is completed by solving a dual-objective optimization problem. Further, we develop a supervised version of LNSPP for regression problem, *i.e.* SLNSPP, and apply it to JITL (SLNSPP-JITL) for developing adaptive soft sensor. In addition, a sparse learning strategy with adaptive threshold value is presented for database monitoring to reduce the online computational load of neighborhood search in JITL. The proposed schemes are evaluated by two case studies. The first one is used to test the structure preserving ability of LNSPP and the other is employed to examine the performance of SLNSPP-JITL when it is applied to soft sensors.

2. Just-In-Time Learning

Just-in-time learning (JITL) features the following steps.

1) When the estimation need for a query sample $\boldsymbol{x}_q$ arises, database search is implemented to find the most similar k samples of $\boldsymbol{x}_q$ according to some similarity criterions.
2) Build one local model with these similar samples.
3) Estimate the output of $\boldsymbol{x}_q$ by this local model and then discard it.

In general, there are two learning types in JITL. The first one is called locally weighted average (LWA) and the other is named as locally weighted regression (LWR). Let the selected k similar samples be denoted as $\{\boldsymbol{X}_s \in R^{k\times m}, \boldsymbol{Y}_s \in R^{k\times 1}\} = \{(\boldsymbol{x}_1^s, y_1^s), (\boldsymbol{x}_2^s, y_2^s),..., (\boldsymbol{x}_k^s, y_k^s)\}$. Here m is the dimensionality of the input vector and $\boldsymbol{x}_i^s$ with output of y_i^s representing the ith similar sample of $\boldsymbol{x}_q$. In LWA, the estimated output of $\boldsymbol{x}_q$ is formulated as

$$\hat{y}_q = \sum_{i=1}^{k} s_i y_i^s \Big/ \sum_{i=1}^{k} s_i \tag{1}$$

where s_i represents the weight assigned to $\boldsymbol{x}_i^s$, which is usually set as the similarity between $\boldsymbol{x}_q$ and $\boldsymbol{x}_i^s$. Different from LWA, LWR trains a temporary local model f^L using $\{\boldsymbol{X}_s, \boldsymbol{Y}_s\}$ and the predicted value for $\boldsymbol{x}_q$ is computed as

$$\hat{y}_q = f^L(\boldsymbol{x}_q). \tag{2}$$

The local model f^L can be trained by either linear algorithms such as weighted least squares or nonlinear algorithms with more powerful learning ability such as weighted least squares support vector regression. Therefore, LWR is more flexible and complex than LWA, while LWA is easy to implement and usually requires less samples than LWR. In Section 4, the properties of both LWA and LWR will be studied.

3. JITL Using Supervised Local and Non-local Structure Preserving Projections for Adaptive Soft Sensor

3.1. Local and non-local structure preserving projections

Assume that we have n training samples $\{\boldsymbol{X} \in R^{n\times m}, \boldsymbol{Y} \in R^{n\times 1}\} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_n, y_n)\}$. The LPP seek for several projection vectors $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_l] \in R^{m\times l}$, such that the local structure can be preserved. That is, two nearby points in original space are expected to be still close in projected space. It maps $\boldsymbol{X} = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, ..., \boldsymbol{x}_n^T]^T$ to $\boldsymbol{Z} = [\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, ..., \boldsymbol{z}_n^T]^T$, where $\boldsymbol{z}_i = \boldsymbol{x}_i\boldsymbol{A} \in R^{1\times l}$. Take $l = 1$ for example, the objective function of LPP is defined as [24]

$$\begin{aligned}\min\ J^L &= \frac{1}{2}\sum_{i,j}(z_i - z_j)^2 W_{ij}^L \\ &= \frac{1}{2}\sum_{i,j}(\boldsymbol{x}_i\boldsymbol{a} - \boldsymbol{x}_j\boldsymbol{a})^2 W_{ij}^L \\ &= \sum_i \boldsymbol{a}^T\boldsymbol{x}_i^T D_{ii}^L \boldsymbol{x}_i\boldsymbol{a} - \sum_{i,j}\boldsymbol{a}^T\boldsymbol{x}_i^T W_{ij}^L \boldsymbol{x}_j\boldsymbol{a} \\ &= \boldsymbol{a}^T\boldsymbol{X}^T\left(\boldsymbol{D}^L - \boldsymbol{W}^L\right)\boldsymbol{X}\boldsymbol{a} \\ &= \boldsymbol{a}^T\boldsymbol{S}^L\boldsymbol{a}\end{aligned} \tag{3}$$

where $\boldsymbol{W}^L$ represents the adjacency matrix and $\boldsymbol{D}^L$ is a diagonal matrix whose entries are the row (or column, as $\boldsymbol{W}^L$ is symmetric) sum of $\boldsymbol{W}^L$, that is, $D_{ii}^L = \sum_j W_{ij}^L$. Here $\boldsymbol{S}^L = \boldsymbol{X}^T(\boldsymbol{D}^L - \boldsymbol{W}^L)\boldsymbol{X}$ is called "local scatter matrix". The elements of $\boldsymbol{W}^L$ are defined as

$$W_{ij}^L = \begin{cases} S_{ij}, & \boldsymbol{x}_i \ \&\ \boldsymbol{x}_j \text{ connected} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $0 \le S_{ij} \le 1$ represents the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, usually defined by heat kernel,

$$S_{ij} = \exp\left(-d_{ij}^2/2\sigma^2\right) \tag{5}$$

where d_{ij} represents the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. By analyzing Eq. (4), one can readily find that two faraway samples may probably be nearby after the projection, as the connected weight between them is usually 0. That is, LPP considers only the local structure but ignores the non-local structure. Similar to Eq. (3), the preserving of non-local structure can be accomplished by [25]

$$\begin{aligned}\max\ J^N &= \frac{1}{2}\sum_{i,j}(z_i - z_j)^2 W_{ij}^N \\ &= \frac{1}{2}\sum_{i,j}(\boldsymbol{x}_i\boldsymbol{a} - \boldsymbol{x}_j\boldsymbol{a})^2 W_{ij}^N \\ &= \sum_i \boldsymbol{a}^T\boldsymbol{x}_i^T D_{ii}^N \boldsymbol{x}_i\boldsymbol{a} - \sum_{i,j}\boldsymbol{a}^T\boldsymbol{x}_i^T W_{ij}^N \boldsymbol{x}_j\boldsymbol{a} \\ &= \boldsymbol{a}^T\boldsymbol{X}^T\left(\boldsymbol{D}^N - \boldsymbol{W}^N\right)\boldsymbol{X}\boldsymbol{a} \\ &= \boldsymbol{a}^T\boldsymbol{S}^N\boldsymbol{a}\end{aligned} \tag{6}$$