I&EC research
Industrial & Engineering Chemistry Research

Article

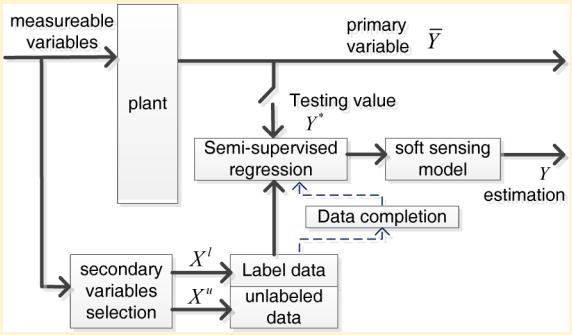
pubs.acs.org/IECR

A Framework and Modeling Method of Data-Driven Soft Sensors Based on Semisupervised Gaussian Regression

Weiwu Yan,* Pengju Guo, Yu Tian, and Jianjun Gao

Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China

ABSTRACT: Soft sensors have been widely used in industrial processes to predict uneasily measured important process variables. The core of data-driven soft sensors is to construct a soft sensor model by using recorded process data. This paper analyzes the geometry and characteristics of soft sensor modeling data and explains that soft sensor modeling is essentially semisupervised regression rather than widely used supervised regression. A framework of data-driven soft sensor modeling based on semisupervised regression is introduced so that information on all recorded data, including both labeled data and unlabeled data, is involved in the soft sensor modeling. A soft sensor modeling method based on a semisupervised Gaussian process regression is then proposed and applied to the estimation of total Kjeldahl nitrogen in a wastewater treatment process. Experimental results show that the proposed method is a promising method for soft sensor modeling.



1. INTRODUCTION

It is well-known that some important process variables in modern industrial processes are difficult or impossible to measure online because of limitations of process technology or measurement techniques. These variables are usually the key indicators of process performance. Soft sensors are widely employed to solve such problems and have been regarded as a valuable alternative to the conventional means for the acquisition of critical process variables.[1]

The core of soft sensors is the soft sensor model. From the aspect of modeling methods, the soft sensor model can be classified into three categories, namely, the first-principles model, the data-driven model, and the mixed model.[1] Because of the complexity of industrial processes and the huge computational burden of the first-principles model, the data-driven model is the most popular for the development of soft sensors. Currently, most modeling methods in data-driven soft sensors are primarily based on statistical or soft computing supervised learning approaches.[2−5]

Although data-driven soft sensors have achieved significant progress in theoretical studies and have successful real applications, there are still some unsolved problems in soft sensor modeling and applications.[1,2,6] The problem of missing values of recorded process data is typical of the crucial problems that modeling methods must address.[7,8] There are two categories of missing data in data-driven soft sensor modeling: missing data of secondary variables (i.e., missing elements of input data) and missing data of primary variables (i.e., missing value of output data). For the secondary variables, the most common causes are the failure, maintenance, or removal of a hardware sensor. Different strategies are available to treat such missing data of identification and process modeling. A very simple approach is to replace the missing values by the mean values of the affected variable. A more efficient approach for handling missing values is to reconstruct the missing values by using the other available variables of the affected samples.[8] Schafer and Graham[9] propose two general approaches to handle missing data based on maximum-likelihood and Bayesian multiple imputations. The Bayesian method for state estimation is also used to treat missing data.[10] Kalman filtering and the data fusion technique are taken to solve the problem of irregular measurements.[11] For the primary variables, missing data are caused by a testing delay in laboratory or uneasy measurement. Such incomplete data, that is, unlabeled data, are hardly used by soft sensors modeling methods based on conventional supervised regression. Fang applied a genetic algorithm to state estimation subject to randomly missing input/output data.[12] Deng et al. discuss identification of nonlinear parameter varying systems with missing output data under the framework of the EM algorithm.[13] Jia proposed a semisupervised recursive weighted kernel regression to update the soft-sensor model online.[14] This paper discusses soft sensor modeling based on semisupervised regression to address unlabeled data.

Semisupervised learning is an active field in machine learning and is routinely used to solve difficult machine learning problems. Methods for semisupervised learning mainly include generative models, graph-based methods, co-training and multiview learning.[15,16] Zhou and Li proposed co-training for semisupervised regression.[17] Brefeld et al. perform multiview semisupervised regression.[18] Cortes and Mohri proposed a

Received: October 31, 2015
Revised: June 1, 2016
Accepted: June 8, 2016
Published: June 9, 2016

DOI: 10.1021/acs.iecr.5b04118
Ind. Eng. Chem. Res. 2016, 55, 7394−7401

simple yet efficient transductive regression model.[19] Hongwei Li presented a semisupervised algorithm to learn a Gaussian process classifier.[20] Vikas Sindhwani proposed a graph-based construction of semisupervised Gaussian process classifiers.[21] This paper focuses on a semisupervised Gaussian process regression to handle missing data in data-driven soft sensor modeling.

The remainder of the paper is organized as follow. Analysis of conventional soft sensor modeling is given in section 2. A framework of soft sensors based on semisupervised regression is discussed in section 3. Section 4 proposes a semisupervised Gaussian process regression (SSGPR) model of soft sensors. An application of the proposed method is given in section 5, and a conclusion is given in section 6.

2. ANALYSIS OF CONVENTIONAL SOFT SENSOR MODELING

A data-driven soft sensor model is a type of black-box model based only on input−output measurements of an industrial process. The basic structure of conventional soft sensors is shown in Figure 1. In soft sensor modeling, the secondary

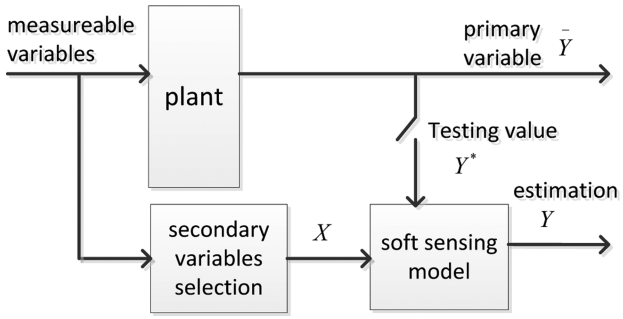


Figure 1. Conventional structure of soft sensor modeling.

variables X are employed to act as the inputs of the soft sensor model, and the calculated value or long time interval sample values of primary variable Y^* are employed to act as the output of the soft sensor model. The soft sensor model seeks a mapping relationship of secondary variables to primary variables; that is, $Y = f(X)$.

However, the conventional structure of soft sensors has some drawbacks. The primary variables, which are difficult or impossible to measure online in industrial processes, are normally determined by offline sample analyses in the laboratory with large time intervals (often several hours). This means that there are no primary variable values in the analysis interval. Therefore, historical data usually contain plenty of secondary variable data without primary variable data and only a few secondary variable data with primary variables. Figure 2 and Figure 3 show the schematic diagram of data characteristics of soft sensor modeling and the schematic diagram of the prediction of soft sensors, respectively.

In Figure 2 and Figure 3, x represents secondary variables, and y represents the primary variable. Green circles represent testing values of primary variables in the laboratory, blue crosses represent missing testing values of primary variables corresponding to secondary variables in the analysis interval, and blue circles represent prediction values of the primary variable by soft sensors. The superscript notation l and u indicate the labeled data and unlabeled data, respectively. The index k is for the labeled data, and indices i and j are for the

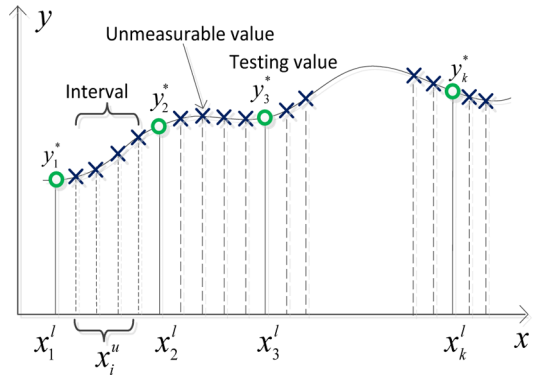


Figure 2. Schematic diagram of data distribution of soft sensor modeling.

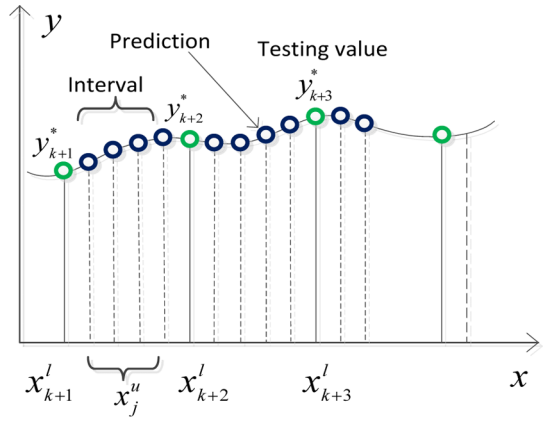


Figure 3. Schematic diagram of prediction of soft sensors.

unlabeled data. (x_k^l, y_k^*) is pair of input−output data with labels, and $(x_i^u, \times)$ is the pair without labels, for example, lack of primary variable data. It is found that the modeling data of data-driven soft sensors are characterized by small amounts of labeled data (x_k^l, y_k^*) and large amounts of unlabeled data $(x_i^u, \times)$. From the aspect of machine learning, soft sensor modeling is essentially a semisupervised regression rather than widely used conventional supervised regression in which (x_k^l, y_k^*) are employed and $(x_i^u, \times)$ are usually discarded. In the framework of semisupervised regression, a large amount of unlabeled data $(x_i^u, \times)$ that significantly affect geometry and properties of the input data are easily able to be involved in soft sensor modeling.

3. FRAMEWORK OF SOFT SENSOR MODELING BASED ON SEMISUPERVISED REGRESSION

Data-driven soft sensor modeling is typical semisupervised regression which makes use of labeled data and unlabeled data for training. Semisupervised regression provides a better suitable framework for constructing a data-driven soft sensor model. Figure 4 shows a framework of data-driven soft sensors based on semisupervised regression, in which two categories of missing data can be addressed. In the framework, the mapping relationship of secondary variables to primary variables is changed from $Y = f(X^l)$ to $Y = f(X^l, X^u)$, in which all secondary variables are involved in the soft sensor modeling. For missing data of secondary variables, it can be also processed in a block of data completion. In the Introduction, there are some methods reviewed to treat such missing data of secondary variables.

DOI: 10.1021/acs.iecr.5b04118
Ind. Eng. Chem. Res. 2016, 55, 7394−7401

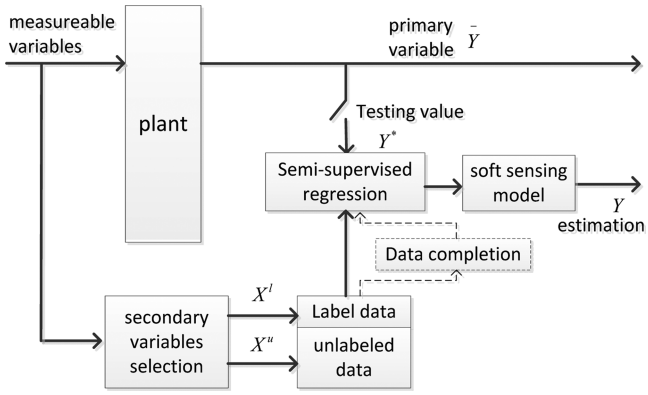


Figure 4. Structure of data-driven soft sensor modeling based on semisupervised regression.

The acquisition of labeled data requires costs such as offline sample analyses or physical experiment (product quality analyzer), whereas acquisition of unlabeled data is inexpensive. Data-driven soft sensors based on semisupervised regression can be of great practical value in the context of no increasing cost. Under the framework of semisupervised regression, data-driven soft sensor modeling methods based on semisupervised regression can be developed and used to improve the performance of soft sensors. Soft sensor modeling based on semisupervised Gaussian regression is discussed in the next section.

4. SOFT SENSOR MODELING BASED ON SEMISUPERVISED GAUSSIAN REGRESSION

A Gaussian process (GP) is powerful nonparametric machine learning technique for constructing comprehensive probabilistic models.[22] Gaussian process regression (GPR) is a method of interpolation in which the interpolated values are modeled by a Gaussian process.[21,23,24] Gaussian process regression has been used for soft sensor modeling.[25] Standard GPR is deformed into semisupervised Gaussian process regression (SSGPR) by replacing a kernel with a semisupervised kernel in this section.

4.1. Gaussian Process Regression. A Gaussian process (GP) is a collection of random variables that follow a joint Gaussian distribution.[22] A GP is fully described by its mean and covariance functions.

Given a training set $\mathcal{D}$ of n observations, $\mathcal{D} = \{(x_i, y_i)|i = 1,...n\}$, where x denotes an input vector of dimension m, and y denotes output (real value). One can rewrite the training set as $\mathcal{D} = \{X, Y\}$, where X is an $m \times n$ observations matrix, and Y is a real-valued target with dimension n. It is assumed that random variables represent the value of the function $f(x)$ at location x. GP describes a distribution over function: $f(x) \sim GP(m(x), k(x, x'))$ where the mean function is $m(X) = E[f(X)]$ and the covariance function is $K(X, X') = E[(f(X) = m(X))(f(X') = m(X'))]$.

Given a set of testing data x_*, the GPR is to find the predictive output f_* with probabilistic confidence levels. Set $m(X) = 0$ and $k(x, x') = \exp((x - x')^2\ 2\delta^2)$, where δ is the width of the kernel. GP is rewritten as $f(x) \sim GP(0, K(x, x'))$. The joint posterior distribution of the training data and testing data is still a Gaussian distribution:[12]

$$P(f, f_*) \sim N\left(0, \begin{bmatrix} K(x, x') & K(x, x_*) \\ K(x_*, x) & K(x_*, x_*) \end{bmatrix}\right) \quad (1)$$

where $K(x, x')$ denotes the covariance matrix of the training data, $K(x_*, x_*)$ denotes the covariance matrix of the testing data, and

$$K(x_*, x)$$

and

$$K(x, x_*)$$

denote the covariance matrix between the training data and testing data, respectively.

Using the Bayesian inference

$$p(f_*|f, X) = \frac{p(f|f_*)p(f_*|X)}{p(f|X)}$$

one can obtain

$$P(f_*|X_*, X, f) \sim N(K(x_*, x)K(x, x')^{-1}f, K(x_*, x_*) - K(x, x')^{-1}K(x, x_*)) \quad (2)$$

It is assumed that the observed values y differ from the function values $f(x)$ by additive noise ε, which follows an independent, identically distributed Gaussian distribution with zero mean and variance σ^2. The joint posterior distribution of the training data and testing data can then be rewritten as

$$P(y, f_*) \sim N\left(0, \begin{bmatrix} K(x, x') + \sigma^2 I & K(x, x_*) \\ K(x_*, x) & K(x_*, x_*) \end{bmatrix}\right) \quad (3)$$

GP with noise follows the distribution

$$P(f|X, y, x_*) \sim N(\bar{f}_*, \mathrm{cov}(f_*)) \quad (4)$$

where the mean value of GP is

$$\bar{f}_* = E[f_*|X_*, X, f] = K(x_*, x)[K(x, x') + \sigma^2 I]^{-1} f \quad (5)$$

and the covariance of the Gaussian process is

$$\mathrm{cov}(f_*) = K(x_*, x_*) - K(x_*, x)[K(x, x') + \sigma^2 I]^{-1} K(x, x_*) \quad (6)$$

$\bar{f}_*$ is the predicted output of GPR, and $\mathrm{cov}(f_*)$ is the variance, which can be used as the uncertainty estimation for confidence levels.

4.2. Semisupervised Gaussian Process Regression. The kernel function (covariance) of GPR is usually constructed from labeled data. For semi-supervised learning, Vikas Sindhwani proposed a graph-based construction of semi-supervised Gaussian process.[21,23] Semisupervised Gaussian Process regression (SSGPR) in this paper is based on graph-based construction of semisupervised kernels.

Graph-based methods for semisupervised regression use a graph representation of the data with a node for each labeled and unlabeled data point. Given a data set $X_D = \{X_l, X_u\}$ where X_l is the labeled data set associated with labeled data set Y_l, and X_u is the unlabeled data set, the data geometry is modeled as a graph whose vertices are the labeled and unlabeled data and whose edges encode the neighborhood relationships. An

DOI: 10.1021/acs.iecr.5b04118
Ind. Eng. Chem. Res. 2016, 55, 7394−7401

undirected graph $G = \{V, E\}$ is constructed over the data set XD, where the data are represented by the nodes V of the graph and the weights matrix $W = \{W_{ij}\}$ of edge E indicates the distance between the nodes. One then defines the graph Laplacian (Laplacian matrix) on the undirected graph: $\mathcal{L} = D - W$ where W is the weight matrix with elements

$$w_{i,j} = \begin{cases} k(x_i, x_j) & \|x_i, x_j\| \zeta \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

and D is a diagonal matrix with elements

$$D_{i,j} = \begin{cases} \sum_i w_{ij} & if i = j \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Graph Laplacian $\mathcal{L}$ is a symmetric and positive semidefinite matrix.

The kernel K naturally defines a unique Reproducing Kernel Hilbert Space (RKHS) of functions H. The kernel function K can be a measure of the distance between input vectors. To adapt to the geometry of the data, original RKHS H is deformed to obtain a new RKHS by refining the norm using labeled and unlabeled data. Define $\tilde{H}$ as the space of functions from H with the modified data-dependent inner product: $\langle f,g\rangle_{\tilde{H}} = \langle f,g\rangle_H + f^T Mg$ where f and g are the vectors $\{f(x)|x \in X_D$ and $\{g(x)|x \in X_D\}$, respectively, and M is a symmetric positive semidefinite matrix. The definition of M is based on the construction of a data adjacency graph over the data set and can be derived from the graph Laplacian $\mathcal{L}$. One can set $M = \gamma \mathcal{L}^p$ (or $M = \mathcal{L}$), where p is an integer, and γ is a regularization parameter.

The space $\tilde{H}$ is still an RKHS. Compared to the original function space H, the space $\tilde{H}$ is better suited for semisupervised learning. The form of the new kernel $\tilde{K}$ can be derived in terms of the kernel function K using reproducing properties of an RKHS and orthogonality argument.[21]

$$\tilde{K}(x, x') = K(x, x') - \Sigma_{Dx}^T (I + M\Sigma_{DD})^{-1} M\Sigma_{Dx} \quad (9)$$

where $K(x, x')$ denotes the $l \times l$ covariance matrix of GPR, $\sum_{Dx}$ denotes the column vector $[K(x_1, x),\cdots,K(x_{l+u}, x)]^T$, $\sum_{DD}$ denotes the $(l + u) \times l + u$ covariance matrix of labeled data and unlabeled data, and $\tilde{K}(x, x')$ denotes the $l \times l$ covariance matrix embedded the manifold information.

$\tilde{K}(x, x')$ is the semisupervised kernel that incorporates the geometry information from unlabeled data. It is found that new kernel $\tilde{K}(x, x')$ is built by labeled data and unlabeled data without any target value or function information.

Replacing $K(x, x')$ in eq 5 and 6 by $\tilde{K}(x, x')$, one obtains a new predicted output of SSGPR:

$$\overline{f}_* = E[f_*|X_*, X, f] = K(x_*, x)[\tilde{K}(x, x') + \sigma^2 I]^{-1} f \quad (10)$$

and the uncertain estimations for confidence levels

$$\text{cov}(f_*) = K(x_*, x_*) - K(x_*, x)[\tilde{K}(x, x') + \sigma^2 I]^{-1} K(x, x_*) \quad (11)$$

where $\overline{f}_*$ is the predicted output of SSGPR, and $\text{cov}(f_*)$ is the variance of prediction that can be used as the uncertain estimations for confidence levels.

4.3. Soft-Sensor Modeling Based on Semisupervised Gaussian Regression. On the basis of the analysis given in previous subsections, the soft sensor modeling procedure based on semisupervised Gaussian regression can be illustrated in the following steps:

Step 1: The secondary variables are generally determined according to theoretical analysis and experience of operators.

Step 2: Data are preprocessed. Normalization of data set $X_D = \{X_l, X_u\}$ is a typical data preprocessing method.

Step 3: Construct undirected graph $G = \{V,E\}$ over the data set $X_D = \{X_l, X_u\}$ and compute weights matrix W using eq 7.

Step 4: Compute diagonal matrix D using eq 8.

Step 5: Graph Laplacian is built by $\mathcal{L} = D - W$

Step 6: Compute symmetric positive semidefinite matrix $M = \gamma \mathcal{L}^p$.

Step 7: Compute the covariance matrix of Gaussian process K of the labeled data.

Step 8: Compute semisupervised kernel $\tilde{K}$ using eq 9.

Step 9: Model selection and parameters tuning by cross-validation, that is, tuning of regularization parameter γ and kernel parameter δ.

Step 10: Predict values of the primary variable by the soft sensor based on SSGPR by eqs 10 and 11 online.

Main processes and computations of SSGPR based soft sensor modeling are shown by flowchart in Figure 5.

5. CASE STUDY

To verify the effectiveness of soft sensors based on semisupervised regression, Benchmark Simulation Model No. 1 (BSM1) is used as a case study.[26] BSM1 was developed with an objective to evaluate control strategies in wastewater treatment plants. Wastewater treatment plants are large-scale nonlinear systems subject to large perturbations in influent flow rate and pollutant load, together with uncertainties concerning the composition of the incoming wastewater. Benchmark simulation models consist of predefined plant layout, process models, sensor, and actuator models, influent characteristics, and evaluation criteria. The layout of the BSM1 plant is depicted in Figure 6, which includes the activated sludge system and the secondary clarifier. Sensor and control handles for the biological aeration system and settling tanks are available in the toolbox. Influent data for different weather conditions (dry weather, rain, etc.) are available for a 1-week evaluation.

5.1. Preparation and Preprocessing. In the wastewater treatment process, the total Kjeldahl nitrogen (TKN) is difficult to measure online but is an important indicator in the control system. TKN serves as a primary variable in the experiment. According to process analysis, 8 indices are chosen as secondary variables, which are time, active heterotrophic biomass (X_{BH}), slowly biodegradable substrates (X_S), inert particulate organics (X_I), NH_4^+ + NH_3 nitrogen (S_{NH}), soluble biodegradable organic nitrogen (S_{ND}), particulate biodegradable organic nitrogen (X_{ND}), and flow. Soft sensors for TKN are built on those input−output data.

SSGPR- and GPR-based soft sensors are trained, tested, and compared on the same data set. Two weather conditions, rain and dry weather, are considered in the experiment. Both dry data and rain data include 1345 samples, which are divided into labeled data, unlabeled data, and testing data at the respective ratio of 3:6:1. Unlabeled data in the case is actually labeled data whose label is hidden deliberately for the experiment.

DOI: 10.1021/acs.iecr.5b04118
Ind. Eng. Chem. Res. 2016, 55, 7394−7401

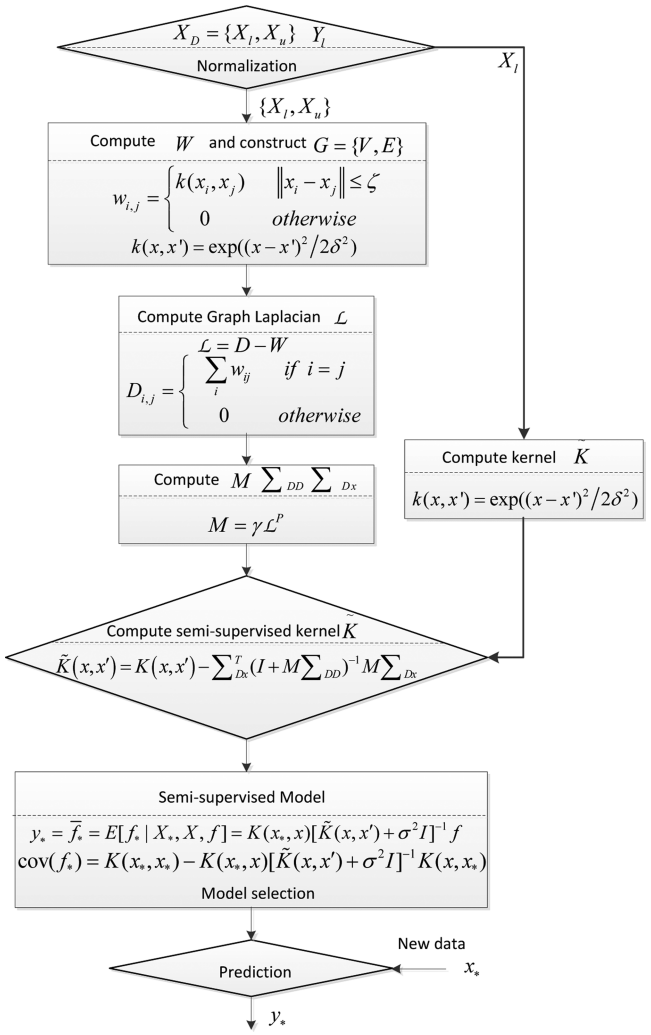


Figure 5. Flowchart of SSGPR based soft sensor modeling.

Variable values are normalized to [0, 1] by the following formula:

$$x = \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{12}$$

5.2. Model Selection. Model selection is an important and challenging issue in the soft sensor field. Model selection of SSGPR-based soft sensors involves choosing the kernel parameters and the noise variance. It is assumed that the noise $\varepsilon \sim (0, 0.36)$ in the experiment. Kernel parameters of semisupervised Gaussian process regression include a graph regularization matrix M and a covariance function K. Gaussian RBF kernel $k(x, x') = \exp(-(x - x')^2/2\delta^2)$ is selected as a covariance function, and $M = \gamma\mathcal{L}$ serves as a semidefinite

Table 1. Mean Squared Errors (MSEs) of Model Selection with the Dry Weather Condition (δ_0 = 0.13)

MSE	$\delta = 0.25\delta_0$	$\delta = 0.5\delta_0$	$\delta = \delta_0$	$\delta = 2\delta_0$	$\delta = 4\delta_0$
$\gamma = 0$	17.0753	9.7721	2.5454	3.9809	71.7526
$\gamma = 1$	15.6792	7.4438	2.1940	3.8772	79.3937
$\gamma = 10$	12.1540	3.9255	1.5214	2.8225	47.5840
$\gamma = 100$	9.2107	2.5602	1.3396	1.7169	24.3611
$\gamma = 1000$	8.5407	2.3715	1.4722	1.3823	9.6000

Table 2. MSEs of Model Selection with the Rain Weather Condition (δ_0 = 125)

MSE	$\delta = 0.25\delta_0$	$\delta = 0.5\delta_0$	$\delta = \delta_0$	$\delta = 2\delta_0$	$\delta = 4\delta_0$
$\gamma = 0$	20.0857	11.8029	3.4800	3.1865	32.2689
$\gamma = 1$	17.8392	8.9526	3.1015	2.8196	30.3903
$\gamma = 10$	13.5209	5.3589	2.3036	2.0997	21.3822
$\gamma = 100$	10.8507	3.5602	1.0444	2.6008	9.7988
$\gamma = 1000$	9.9241	3.1876	2.3564	2.0741	5.1398

Table 3. RMSE of Soft Sensors Based on SSGPR and GPR

	RMSE			
	dry data		rain data	
soft sensor models	test error	train error	test error	train error
SSGPR	0.6696	0.5168	0.8707	0.7256
SVM	0.6859	0.4759	0.9301	0.7021
GPR	1.1248	1.1083	0.9293	0.9514

Table 4. RTVP of Soft Sensors Based on SSGPR and GPR

	RTVP			
	dry data		rain data	
soft sensor models	test error	train error	test error	train error
SSGPR	0.8334	0.8430	0.8222	0.8576
SVM	0.8145	0.8785	0.8001	0.8446
GPR	0.7098	0.5157	0.799	0.7573

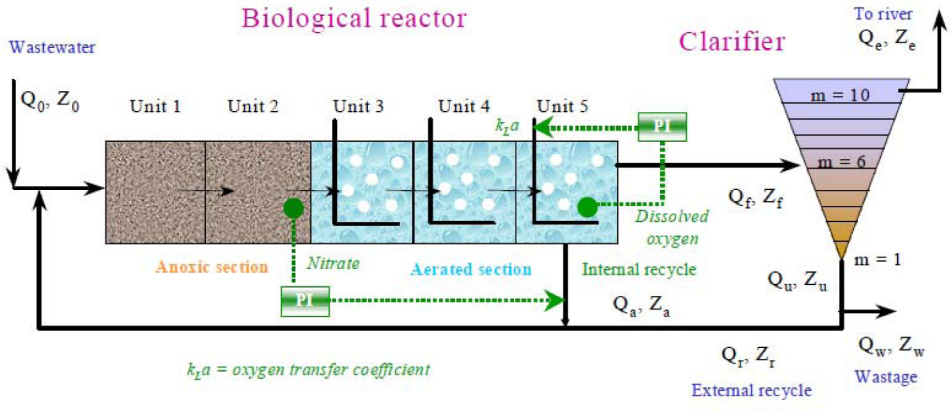


Figure 6. General overview of the BSM1 plant.

DOI: 10.1021/acs.iecr.5b04118
Ind. Eng. Chem. Res. 2016, 55, 7394−7401

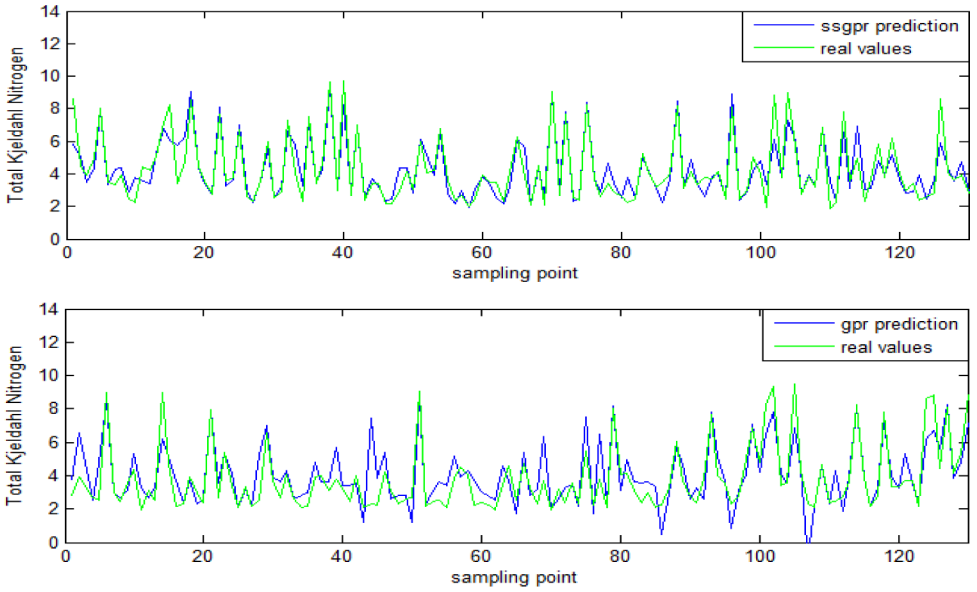


Figure 7. TKN prediction of soft sensors based on SSGPR and GPR on dry data set (δ = 0.13 and γ = 100).

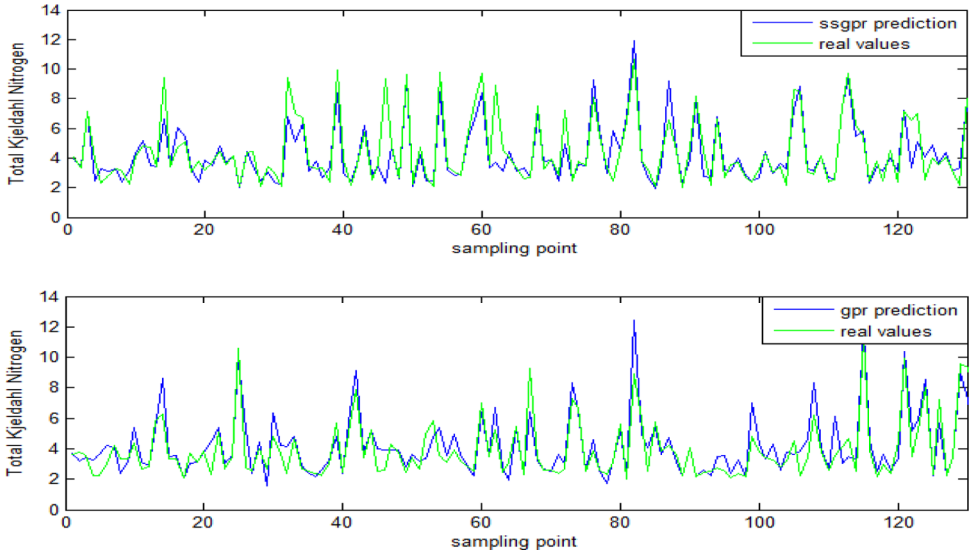


Figure 8. TKN prediction of soft sensors based on SSGPR and GPR on rain data set (δ = 0.125 and γ = 100).

Table 5. RMSE of SSGPR-Based Soft Sensors with Different Sized Unlabeled Data

	dry data		rain data	
sizes of unlabeled data	test error	train error	test error	train error
fully unlabeled data	0.6696	0.5168	0.8707	0.7256
half unlabeled data	1.0146	0.5874	0.9021	0.6409
no unlabeled data	1.1248	1.1083	0.9293	0.9514

matrix in this paper. The optimal model can be obtained by selecting optimal regularization parameter γ and kernel parameters δ. Cross-validation is employed for optimal model selection of soft sensors based on SSGPR.

Model selections of SSGPR-based soft sensors are investigated over a range of values for kernel width δ and scale γ. For each data set, one computes the mean weights of weight matrix (δ_0) in the training set and probed δ in the range [$0.25\delta_0$ $0.5\delta_0$ δ_0 $2\delta_0$ $4\delta_0$]; the range for γ is [0, 1, 10, 100, 1000]. SSGPR recedes into GPR when γ = 0. A 10-fold cross-validation is implemented on the labeled data set for model selection. The results of the model selections of SSGPR-based soft sensors are given in Table 1 and Table 2.

From the results of cross-validation in Table 1 and Table 2, it is found that parameters δ = 0.13 and γ = 100 best fit the model with the dry weather condition, and parameters δ = 0.125 and γ = 100 achieve the best performance for the model with the rain weather condition.

5.3. Result and Discussion. Two performance indicators, root of mean square error (RMSE) and relative variance tracking precision (RVTP), are used to evaluate the performance of soft sensors:

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right]^{1/2} \tag{13}$$

where $\hat{y}_i$ is the estimation of soft sensors and

$$\mathrm{RVTP} = 1 - \sigma_{\mathrm{error}}{}^2/\sigma_{\mathrm{property}}{}^2 \tag{14}$$

DOI: 10.1021/acs.iecr.5b04118
Ind. Eng. Chem. Res. 2016, 55, 7394−7401

where σ_{error}^2 is the covariance between the estimation and real values, and $\sigma_{property}^2$ is the variance of real measurements. RVTP reflects the performance that the estimation of soft sensors follows with a varying trend of real values. If RVTP is closer to 1, the soft sensor model has better tracking performance.

On the basis of the optimal model parameters, SSGPR-based soft sensors are trained on the labeled data and unlabeled data and tested on testing data. GPR-based soft sensors are trained on the labeled data and tested on testing data. The prediction of TKN by soft sensors based on SSGPR and GPR are implemented on the dry data set and rain data set, respectively. The experimental results are shown in Table 3 and Table 4. Estimated outputs of a soft sensor based on SSGPR and GPR as well as real values of TKN on dry data are shown in Figure 7. Estimated outputs of a soft sensor based on SSGPR and GPR as well as real values of TKN on rain data are shown in Figure 8.

From Table 3 and Table 4, the results demonstrate that the SSGPR-based soft sensor achieves significant performance improvements over the soft sensor model based on standard GP on testing data sets. SSGPR-based soft sensors have higher accuracy and better tracking performance than GPR-based soft sensors by incorporating the information on unlabeled data with labeled data. The SSGPR-based soft sensor also compares with the SVM-based soft sensor, which usually has good performance in most cases. It is found that although the SVM-based soft sensor has better performance than the GPR-based soft sensor, the SSGPR-based soft sensor has better performance over the SVM-based soft sensor by virtue of the information on unlabeled data.

From Figure 7 and Figure 8, it is found that soft sensors based on SSPGR achieve good performance in the estimation of TKN of wastewater treatment plants. Estimated outputs of soft sensors based on SSPGR match real values of the TKN and follow the varying trend of the TKN very well.

To further investigate the affection of unlabeled data on prediction accuracy, unlabeled data sets of different sizes are considered in the experiments while the labeled data and testing data stay the same. Three scales of unlabeled data—fully unlabeled data, half unlabeled data, and no unlabeled data—are chosen for SSGPR soft sensor modeling. The prediction results of TKN by SSGPR-based soft sensors using different scales of unlabeled data are shown in Table 5. SSGPR-based soft sensor without unlabeled data is actually equivalent to GPR-based soft sensor. From the comparison of soft sensors with different sized unlabeled data, the prediction performance of the soft sensor improves with increasing unlabeled data size and significantly confirms the effectiveness of the proposed semisupervised soft sensor.

The experimental results and theory analysis show that semisupervised regression provide a more reasonable framework for data-driven soft sensors, in which information on all recorded data including both labeled and unlabeled data are involved in the soft sensor modeling. Plenty of unlabeled data, which provide a wealth of essential process information and significantly affect the geometry and characteristics of the input data, are very useful to improve the performance of data-driven soft sensors. Soft sensors based on semisupervised regression achieve significant performance improvements by incorporating the information on unlabeled data together with labeled data. The performance of soft sensors is improved constantly with increasing unlabeled data. In the framework, information behind the data can be fully exploited to build soft sensors with good performance. Soft sensor modeling based on semisupervised regression is a promising framework for data-driven soft sensor modeling.

6. CONCLUSION

From the aspect of characteristic of soft sensor modeling data, soft sensor modeling is essentially semisupervised regression rather than widely used supervised regression. This paper introduces a framework of data-driven sensors based on semisupervised regression and proposes a soft sensor modeling method based on SSGPR. In the framework of semisupervised regression, all recorded data including both labeled data and unlabeled data are involved in the soft sensor modeling. SSGPR-based soft sensors achieve performance improvements by incorporating the unlabeled data together with labeled data. The results of a case study demonstrate that semisupervised regression builds better soft sensors by virtue of the information on unlabeled data. Semisupervised regression provides a reasonable and promising framework for soft sensor modeling. Many semisupervised regression algorithms can be developed and used to improve performance of soft sensors in the future.

AUTHOR INFORMATION

Corresponding Author

*E-mail: yanwwsjtu@sjtu.edu.cn.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by the Natural Science Foundation of China (Grant 60974119).

REFERENCES

(1) Fortuna, L.; Graziani, S.; Rizzo, A.; Xibilia, G. M. *Soft Sensors for Monitoring and Control of Industrial Process*; Springer-Verlag: London, 2007.

(2) Kadlec, P.; Gabrys, B.; Strandt, S. Data-driven Soft Sensors in the process industry. *Comput. Chem. Eng.* **2009**, *33* (4), 795−814.

(3) Grbic, R.; Sliskovic, D.; Kadlec, P. Adaptive soft sensor for online prediction and process monitoring based on a mixture of Gaussian process models. *Comput. Chem. Eng.* **2013**, *58* (11), 84−97.

(4) Yan, W.; Shao, H.; Wang, X. Soft sensing modeling based on support vector machine and Bayesian model selection. *Comput. Chem. Eng.* **2004**, *28* (8), 1489−1498.

(5) Lin, B.; Recke, B.; Knudsen, J. K. H.; Jorgensen, S. B. A systematic approach for soft sensor development. *Comput. Chem. Eng.* **2007**, *31* (5−6), 419−425.

(6) Jin, X.; Wang, S.; Huang, B.; Forbes, J. F. Multiple model based LPV soft sensor development with irregular/missing process output measurement. *Control Eng. Pract.* **2012**, *20* (2), 165−172.

(7) Walczak, B.; Massart, D. L. Dealing with missing data. Part I. *Chemom. Intell. Lab. Syst.* **2001**, *58* (1), 15−27.

(8) Walczak, B.; Massart, D. L. Dealing with missing data. Part II. *Chemom. Intell. Lab. Syst.* **2001**, *58* (1), 29−42.

(9) Schafer, J. L.; Graham, J. W. Missing data: Our view of the state of the art. *Psychological Methods* **2002**, *7* (2), 147−177.

(10) Zhao, Z.; Huang, B.; Liu, F. Bayesian method for state estimation of batch process with missing data. *Comput. Chem. Eng.* **2013**, *53* (11), 14−24.

(11) Guo, y.; Zhao, y.; Huang, B. Development of soft sensor by incorporating the delayed infrequent and irregular measurements. *J. Process Control* **2014**, *24* (11), 1733−1739.

DOI: 10.1021/acs.iecr.5b04118
Ind. Eng. Chem. Res. 2016, 55, 7394−7401

(12) Fang, H.; Wu, J.; Shi, Y. Genetic adaptive state estimation with missing input/output data. *Proc. IMechE, Part I: J. Systems and Control Eng.* **2010**, *224* (5), 611−617.
(13) Deng, J.; Huang, B. Identification of nonlinear parameter varying systems with missing output data. *AIChE J.* **2012**, *58* (11), 3454−3467.
(14) Ji, J.; Wang, H.; Chen, K.; Liu, Y.; Zhang, N.; Yan, J. Recursive weighted kernel regression for semi-supervised soft-sensing modeling of fed-batch processes. *J. Taiwan Inst. Chem. Eng.* **2012**, *43* (1), 67−76.
(15) Zhu, X.; Goldberg, A. B. *Introduction to Semi-Supervised Learning*; Morgan and Claypool Publishers: San Rafael, CA, 2009, 1−130.
(16) Belkin, M.; Niyogi, P.; Sindhwani, V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *J. Mach. Learn. Res.* **2006**, *7*, 1−48.
(17) Zhou, Z. H.; Li, M. *Semi-Supervised Regression with Co-Training*; In Proceedings of the 19th International Joint Conference on Artificial Intelligence; IJCAI: Edinburgh, Scotland, 2005.
(18) Brefeld, U.; Gartner, T.; Scheffer, T.; Wrobel, S. *Efficient co-regularized least squares regression*. In Proceedings of the 23rd International Conference on Machine Learning; ACM: Pittsburgh, PA, 2006, 137−144.
(19) Cortes, C.; Mohri, M. On Transductive Regression, In *Advances in Neural Information Processing Systems 19*; MIT Press: Vancouver, Canada, 2006.
(20) Li, H.; Li, Y.; Lu, H. *Semi-supervised learning with Gaussian Processes*. In Proceedings of the 2008 Chinese Conference on Pattern Recognition; IEEE: Beijing, China, 2008; pp 1−5.
(21) Sindhwani, V.; Chu, W.; Keerthi, S. S. *Semi-Supervised Gaussian Process Classifiers*. In Proceedings of the 20th International Joint Conference on Artificial Intelligence; Morgan Kaufmann Publishers Inc: Hyderabad, India, 2007; pp 1059−1064.
(22) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, 2006.
(23) Sindhwani, V.; Niyogi, P.; Belkin, M. *Beyond the Point Cloud: From Transductive to Semi-supervised Learning*; In Proceedings of the 22nd international conference on Machine learning; ACM: Bonn, Germany, 2005; pp 824−831.
(24) Wu, Q.; Law, R.; Xu, X. A sparse Gaussian process regression model for tourism demand forecasting in Hong Kong. *Expert Syst. Appl.* **2012**, *39* (5), 4769−4774.
(25) Grbić, R.; Slišković, D.; Kadlec, P. Adaptive soft sensor for online prediction and process monitoring based on a mixture of Gaussian process models. *Comput. Chem. Eng.* **2013**, *58*, 84−97.
(26) Alex, J.; Benedetti, L.; Copp, J. *Benchmark Simulation Model no. 1 (BSM1)*; Technical Report; Lund University: 2008

DOI: 10.1021/acs.iecr.5b04118
Ind. Eng. Chem. Res. 2016, 55, 7394−7401