Soft-Sensing Modeling Based on GK-LSSVM Method for Online Predictions of BOD in Activated Sludge Process

Fei Luo, Xinghong Qiao, Weihao Liao
Key Laboratory of Autonomous Systems and Networked Control, Ministry of Education
Institute of Automation Science and Engineering, South China University of Technology
Guangzhou 510640, P.R. China
e-mail: aufeiluo@scut.edu.cn, qiaoxinghong@126.com, auliaowh@mail.scut.edu.cn

Abstract—**Five-day biochemical oxygen demand (BOD_5) is one of the key parameters which is widely used to evaluate the biological and chemical evaluation of effluents from wastewater treatment plants. This paper proposes a novel method that can predict online using BOD_5 by synthesizing an online version of Gustafson-Kessel (GK) algorithm and least squares support vector machines (LS-SVM). The clustering algorithm can reduce required number of clusters, form more complex shape clusters, and get better modeling performance. Moreover, an online sparse LSSVM with time window is proposed to reduce the computation time and storage space. The GK-LSSVM of the method of the soft-sensor model was proposed to predict BOD_5 concentration. The results indicate that the proposed method can not only improve prediction accuracy but also efficiently decrease model's update frequency, comparing to the cases using with that of different methods.**

Keywords-biochemical oxygen demand; least square support vector machine; soft sensor; gustafson-kessel algorithm

I. INTRODUCTION

Water quality monitoring has become a focus of environmental departments, which the BOD5 concentration in effluents is an important aspect of quality monitoring. It is widely applied to measure the organic content of wastewater for over 100 years [1]. Currently available method for BOD_5 measurement is very tedious and conventional BOD_5 measurement methods take a long time [2]. Moreover, BOD_5 can be hardly used as feedback data for building control models control of wastewater treatment or as the online monitoring data of effluent water quality index. For example, a feedback control method based on BOD_5 will not be able to compensate influent wastewater disturbances in the intervening period due to the fact that BOD_5 only is able to react to the wastewater quality five days earlier. A rapid and convenient technique for measuring BOD_5 becomes desirable basis for monitoring and controlling wastewater treatment.

A data-derived soft-sensor was proposed to predict BOD_5 under the existing information. In the last decades, the different algorithms such as artificial intelligence (AI) techniques [3] and classic regression methods [4] has become one of the most important topics in BOD_5 measurement. AI techniques produce better results than classic regression methods for developing the software sensors [5]. Under this circumstance, soft-sensor based on AI techniques gradually becomes the hot area for online estimation of BOD_5. Honggui H et al. proposed a dynamic neural network (DNN) model to provide accurate predictions of BOD concentration on-line [6]. In another work by Qiao J et al. K-means clustering algorithm with Takagi–Sugeno fuzzy neural network was introduced to estimate BOD values by the Soft-sensing method [7]. They concluded that the soft-sensing method was capable to accurately measure effluent BOD from the wastewater treatment process. Wag J et al. developed a hybrid TS-SVM model to improve the prediction accuracy of BOD and DO by optimizing the key parameters of SVM [8]. In another work by Noori R et al., the SVM was calibrated to investigate model performance by using different records for many times [9]. Findings indicated that the method had acceptable uncertainty in BOD_5 prediction. Mohammadpour R et al. adapted SVM and feed forward back propagation (FFBP) for online prediction of BOD_5 [10].

As mentioned above, the NN and SVM techniques have become one of the most important topics in BOD_5 measurement. The NN and SVM techniques have been recognized as powerful tools for the case of the non-linear system. However, these algorithms all have their own defects. Their main shortcoming of the techniques is the slowly training speed. To some extent, the defects constrain the application of the algorithm. In this paper, we propose an evolving approach for online prediction of BOD_5 by synthesizing an online version of Gustafson-Kessel (GK) algorithm and least squares support vector machines (LSSVM). The remainder of this paper is organized as follows. In the next section, the main aspects of GK-LSSVM model are briefly illustrated. Then the GK-LSSVM gives reasonable estimates for the BOD prediction and comparison of three different methods in Section 3. The concluding remarks are presented in the last section.

II. ONLINE GK CLUSTERING ALGORITHM

A. GK Clustering Algorithm

The fuzzy c-means algorithm (FCM) is one of widely used clustering algorithms [11]. However, the traditional FCM algorithm is large degree dependent on the initialization values and liable to be easily trapped in a local minimum.

978-1-5386-3995-5/17/$31.00 ©2017 IEEE

Gustafson and Kessel applied an adaptive norm to extend the Euclidian distances of the standard FCM. The GK algorithm is given as following.

Let $X = \left(x_1, x_2, \cdots, x_N\right)$ be a set of N data objects. The distance norm D_{ij}^2 between the jth value of the data $\left(x_j\right)$ and the center of the ith cluster $\left(v_i\right)$ can account as following:

$$D_{ij}^2 = \left\| x_j - v_i \right\|_{M_i}^2 = \left(x_j - v_i\right)^T M_i \left(x_j - v_i\right) \tag{1}$$

where the norm matrix M_i is a positive definite symmetric matrix. M_i can be obtained from Eq. (2) and Eq. (3).

$$M_i = \det\left(F_i\right)^{\frac{1}{n}} F_i^{-1} \tag{2}$$

$$F_i = \frac{\sum_{j=1}^{N} \mu_{ij}^m \left(x_j - v_i\right)\left(x_j - v_i\right)^T}{\sum_{j=1}^{N} \mu_{ij}^m} \tag{3}$$

In this minimization problem, the center v_i and the membership degrees μ_{ij} are updated according to the expressions given below.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{M} \left(D_{ij} / D_{kj}\right)^{\frac{2}{m-1}}} \tag{4}$$

$$v_i = \frac{\sum_{j=1}^{N} \left(\mu_{ij}\right)^m x_j}{\sum_{j=1}^{N} \left(\mu_{ij}\right)^m} \tag{5}$$

Objective function based fuzzy clustering algorithms minimize an objective function of the type

$$J\left(X, V, U\right) = \sum_{i=1}^{M} \sum_{j=1}^{N} \left(\mu_{ij}\right)^m D_{ij}^2 \tag{6}$$

where $V = \left(v_1, v_1, \cdots, v_M\right)$ is a C-tuple of cluster prototypes, and $m \in \left[1 \quad \infty\right)$ is a weighting exponent which determines the fuzziness of the clusters. The objective function is, then, minimized under the following constraints:

$$\sum_{i=1}^{M} \mu_{ij} = 1 \, , \mu_{ij} \in \left[0,1\right] , 1 \le j \le N \tag{7}$$

B. Basic Least Squares Support Vector Machines

Support Vector Machines (SVM) motivated by statistical learning theory is a powerful methodology for providing better prediction results [12]. However the quadratic programming (QP) with linear constrain of SVM formulation of the learning problem leads to need huge memory and consume much time. Least Squares Support Vector Machines (LS-SVM) are reformulations to standard SVM, which involves a linear Karush–Kuhn–Tucker system instead of inequality constraints and works with a quadratic programming problem [13]. The basic formulation of LS-SVM is described as follows. Consider a given regression data set $S = \left\{\left(x_i, y_i\right)\right\} \; i = 1, 2, \cdots n$, $x_i \in R^n$ and $y_i \in R$ represent input and output data, respectively. The regression model is taken the form:

$$y\left(x\right) = \omega^T \varphi\left(x\right) + b \quad \omega \in \mathrm{H} \, , \; b \in R \tag{8}$$

where the nonlinear mapping function $\varphi\left(\cdot\right)$ introduced by a kernel function maps the input space in to a higher dimensional feature space, b is the bias term and ω is the weight vector. In the least-squares version of the SVM algorithm, the cost function J is minimized by formulating the following optimization problem.

$$\min_{\omega, e} J\left(\omega, e\right) = \frac{1}{2} \omega^T \omega + \frac{\gamma}{2} \sum_{i=1} e_i^2 \qquad \gamma > 0 \tag{9}$$

$$\text{subject to } \; y_i = \omega^T \varphi\left(x_i\right) + b + e_i \qquad i = 1, 2, \cdots N \tag{10}$$

where the parameter γ is the penalty weight controlling the overfitting phenomenon and the model complexity; $e_i = y_i - \tilde{y}_i$ is the regression error, $e = \left[e_1, \cdots, e_N\right]$ is the learning residual vector. The first and second part of Eq. (9) present the weight decay and the evaluation of accuracy of LSSVM model for all training data. Eq. (10) gives the definition of the regression error. To solve this convex optimization problem, the following Lagrangian function is formed as [14, 15]:

$$L\left(\omega, b, e; \alpha\right) = J\left(\omega, e\right) - \sum_{i=1}^{N} \alpha_k \{\omega^T \varphi\left(x_i\right) + b + e_i - y_i\} \tag{11}$$

where $\alpha_i \in R\left(i = 1, 2, \cdots N\right)$ with Lagrange multipliers that can be positive or negative in the LSSVM formulation. The optimum solution for solution of Eq. (11) can be determined as the following set of partially differentiating equations,

$$\frac{\partial L}{\partial \omega} = 0 \Rightarrow \omega = \sum_{i=1}^{N} \alpha_i \varphi\left(x_i\right) \tag{12}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i = 0 \tag{13}$$

$$\frac{\partial L}{\partial e_i} = 0 \Rightarrow \alpha_i = \gamma e_i \qquad i = 1, 2 \cdots, N \tag{14}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow \omega^T \varphi\left(x_i\right) + b + e_i - y_i = 0 \tag{15}$$

Note that sparseness is lost which is clear from the condition $\alpha_i = \gamma e_i$. After eliminating of the variables ω and e_i , the solution can be obtained as the following linear equations.

$$\Phi \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{16}$$

where Φ is a square matrix given by:

$$\Phi = \begin{bmatrix} 0 & 1_n^T \\ 1_n & \kappa + \gamma^{-1} \mathrm{I} \end{bmatrix} \tag{17}$$

where $y = [y_1 \quad y_2 \quad \cdots \quad y_N]^T$, $\alpha = [\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_N]^T$, $1_n = [1 \quad 1 \quad \cdots \quad 1]^T$, I is the $N \times N$ identity matrix, κ denotes the kernel matrix with ijth element.

$$\kappa(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) = \exp\left(-\frac{\|x_i - x_j\|}{2\delta^2} \right) \quad i, j = 1, 2, \cdots, N \tag{18}$$

where σ is the kernel size.

The parameters α and b can be calculated by solving Eq. (19) as follows:

$$\begin{bmatrix} b \\ \alpha \end{bmatrix} = \Phi^{-1} \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{19}$$

The resulting LSSVM model for function estimation can get the following form:

$$y(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b \tag{20}$$

C. GK-LSSVM Clustering Algorithm

Least Squares Support Vector Machines (LS-SVM) is a powerful tool to identify and control for nonlinear systems. And fuzzy clustering method is a new data analysis tool to deal with ambiguity and uncertainty of knowledge; it has been successfully applied to many areas of classification. A new soft sensor method based on fuzzy clustering and LSSVM is proposed. The structure diagram for the control design is shown in Figure 1. The output of this simple usage can be obtained by using Eq. (5) and Eq. (20).

$$y = \sum_{i=1}^{M} \psi_i y_i \tag{21}$$

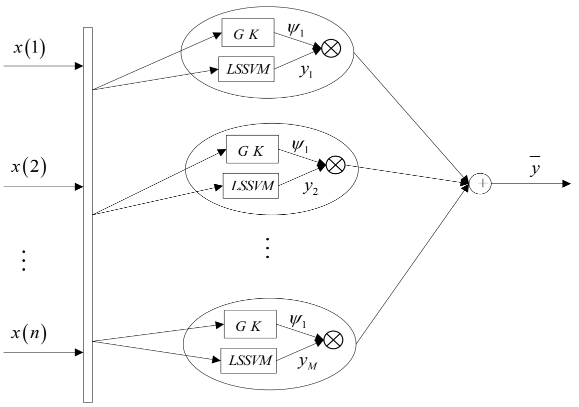

Figure 1. The structure of GK-LSSVM method

III. Experiments and Results Analysis

The GK-LSSVM algorithm mentioned above is put in use for an online prediction of BOD_5 values in wastewater treatment plant effluent. The dataset used in this study is supplied by the University of California Irvine (UCI) Machine Learning Library [16]. A total of 527 daily data records in the dataset have been used, each consisting of 38 process variables. A schema of the plant with different measurement points and variables is indicated in Figure 2. After cleaning some missing values, principal component analysis (PCA) is performed for the remaining data points. The aim of performing PCA is to select the only relevant variables.

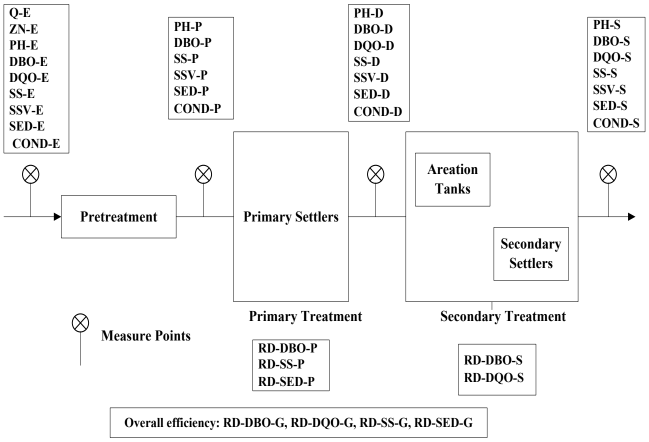

Figure 2. Layout of the WWTP with the points of measurement and the names of measured variables

The input variables X of the soft-sensor model consist of 19 process variables namely biological oxygen demand, chemical demand of oxygen and suspended solids et al. The he model's output variable y is the concentration of BOD_5. The list of identifiers for the measured and computed variables in this WWTP is given in Table I.

The performance of the proposed approach is evaluated by using root mean square error (RMSE) method.

$$RMSE = \sqrt{\sum_{i=1}^{N} \left(\bar{d}_i - d_i \right)^2 / N} \tag{22}$$

70 input data are used to train model and the other 330 input data are used to test model. The simulation results of BOD_5 values based on GK-LSSVM method are shown in Figure 3~6. Figure 3 shows real index values (solid line) and estimated output (dashed line) via GK-LSSVM algorithm for training data; Figure 4 is the corresponding training error. Figures 5 compare BOD predictions real index values (solid line) and estimated output (dashed line) in the testing process. Figure 6 is the corresponding testing error. The Table II indicate the comparison of SVM, RVM and Fast-RVM based soft-sensor models [17, 18]. The RMSE of GK-LSSVM is 10.754, which is smaller compared to that of the other three soft-sensor models. The GK-LSSVM algorithm has a better approximation performance in prediction of BOD_5 values than the other three methods and a less computation time. The experimental results demonstrate the effectiveness and reliability of the proposed method.

TABLE I. AUXILIARY LIST OF EXPERIMENT

Code	Parameter	Code	Parameter
RD-SED-G	Global Performance Input Sediments /mg·L^{-1}	PH-S	Output PH /mg·L^{-1}
RD-SS-G	Global Performance Input Suspended Solids /mg·L^{-1}	DBO-E	Input Biological Demand of Oxygen To Plant /mg·L^{-1}
RD-DBO-G	Global Performance Input Biological Demand Of Oxygen /mg·L^{-1}	DQO-D	Input Chemical Demand of Oxygen To Secondary Settler /mg·L^{-1}
RD-DQO-G	Global Performance Input Chemical Demand Of Oxygen /mg·L^{-1}	SED-D	Input Sediments to Secondary Settler /mg·L^{-1}
SS-D	Input Suspended Solids to Secondary Settler /mg·L^{-1}	DQO-S	Output Chemical Demand of Oxygen /mg·L^{-1}
RD-SS-P	Performance Input Suspended Solids to Primary Settler /mg·L^{-1}	SED-S	Output Sediments /mg·L^{-1}
DQO-E	Input Chemical Demand Of Oxygen to Plant /mg·L^{-1}	SS-S	Output Suspended Solids /mg·L^{-1}
DBO-D	Input Biological Demand Of Oxygen to Secondary Settler /mg·L^{-1}	PH-D	Input Ph to Secondary Settler /mg·L^{-1}
RD-DQO-S	Performance Input Chemical Demand of Oxygen to Secondary Settler /mg·L^{-1}	RD-DBO-P	Performance Input Biological Demand of Oxygen in Primary Settler /mg·L^{-1}
RD-DBO-S	Performance Input Biological Demand Of Oxygen to Secondary Settler /mg·L^{-1}		

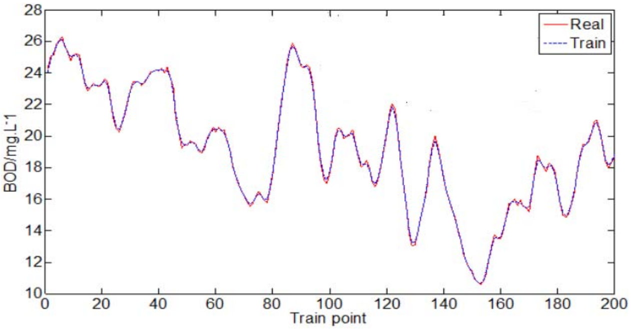

Figure 3. The real and estimated output for training data

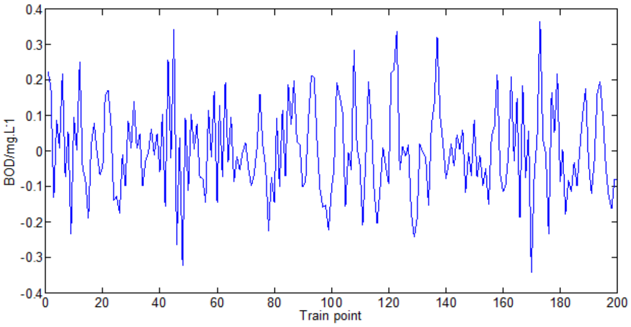

Figure 4. Error diagram

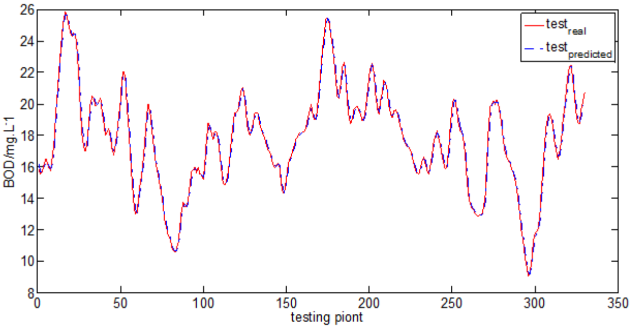

Figure 5. The model output of testing data and the actual output

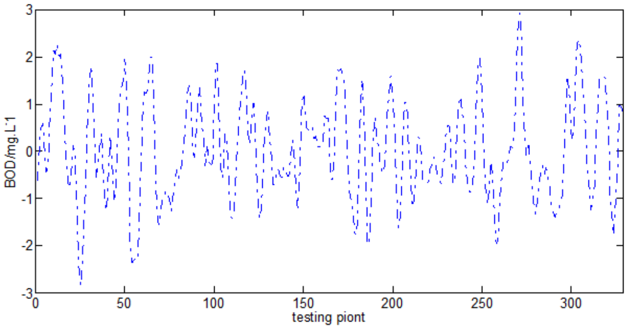

Figure 6. Error diagram

TABLE II. THE PERFORMANCES OF THE FOUR ALGORITHMS

Algorithm	RMSE	Runtime (s)
SVM [17, 18]	0.1562	more than half an hour
RVM [17, 18]	0.1315	91.168
Fast-RVM [17, 18]	0.1814	23.575
GK-LSSVM	0.1088	10.754

IV. CONCLUSIONS

In this work, we propose a hybrid evolving fuzzy algorithm based on GK cluster and LSSVM approach. The clustering algorithm can reduce required number of clusters, form more complex shape clusters, and get better modeling performance. Moreover, at each sampling period the algorithm recursively modifies the model by adding a new data pair and deleting the least important one out of the consideration on real-time property. The data pair deleted is determined by the absolute value of Lagrange multiplier from last sampling period. To reduce the computation time and storage space of online LSSVM with time window, an online sparse LSSVM with time window is proposed. It only takes samples ranking at partial moments among sliding time window as training samples set. The GK-LSSVM is applied to predict BOD_5 values in organic matter in effluents from wastewater treatment plants by the soft-sensor model. The results demonstrate that the proposed method can not only improve prediction accuracy but also efficiently decrease model's update frequency, comparing to the cases using SVM, RVM and Fast-RVM.

ACKNOWLEDGMENT

This work is supported by Science and Technology Planning Project of Guangzhou (201604010032), and Science and Technology Planning Project of Guangdong Province (2016B090918028).

REFERENCES

[1] Meneses C G R, Saraiva L B, Melo H N S, et al. Variations in BOD, algal biomass and organic matter biodegradation constants in a wind-mixed tropical facultative waste stabilization pond[J]. Water Science and Technology, 2005, 51(12): 183-190.

[2] Singh K P, Basant N, Malik A, et al. Multi-way modeling of wastewater data for performance evaluation of sewage treatment plant—A case study[J]. Chemometrics and Intelligent Laboratory Systems, 2009, 95(1): 18-30.

[3] Dogan E, Ates A, Yilmaz E C, et al. Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand[J]. Environmental progress, 2008, 27(4): 439-446.

[4] Oliveira-Esquerre K P, Seborg D E, Bruns R E, et al. Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill: Part I. Linear approaches[J]. Chemical Engineering Journal, 2004, 104(1): 73-81.

[5] Singh K P, Gupta S, Kumar A, et al. Linear and nonlinear modeling approaches for urban air quality prediction[J]. Science of the Total Environment, 2012, 426: 244-255.

[6] Honggui H, Junfei Q. Biological oxygen demand (BOD) soft measuring based on dynamic neural network (DNN): A simulation study[C]//Asian Control Conference, 2009. ASCC 2009. 7th. IEEE, 2009: 757-762.

[7] Qiao J, Li W, Han H. Soft Computing of Biochemical Oxygen Demand Using an Improved T–S Fuzzy Neural Network[J]. Chinese Journal of Chemical Engineering, 2014, 22(11): 1254-1259.

[8] Wang J, Liu Y, Li Z, et al. BOD-DO Concentration Forecasting Model Based on Hybrid TS-SVM[J]. Asian Journal of Chemistry, 2013, 25(14): 7853.

[9] Noori R, Yeh H D, Abbasi M, et al. Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand[J]. Journal of Hydrology, 2015, 527: 833-843.

[10] Mohammadpour R, Shaharuddin S, Chang C K, et al. Prediction of water quality index in constructed wetlands using support vector machine[J]. Environmental Science and Pollution Research, 2015, 22(8): 6208-6219.

[11] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques[J]. Journal of intelligent information systems, 2001, 17(2): 107-145.

[12] Gencoglu M T, Uyar M. Prediction of flashover voltage of insulators using least squares support vector machines[J]. Expert Systems with Applications, 2009, 36(7): 10789-10798.

[13] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.

[14] Wei G, Li G, Wu Y, et al. Application of Least Squares-Support Vector Machine in system-level temperature compensation of ring laser gyroscope[J]. Measurement, 2011, 44(10): 1898-1903.

[15] Kisi O. Modeling discharge-suspended sediment relationship using least square support vector machine[J]. Journal of hydrology, 2012, 456: 110-120.

[16] Blake C, Merz C J. {UCI} Repository of machine learning databases[J]. 1998.

[17] Cao Tao. Study of soft sensor modeling for Wastewater Treatment Process Base on Relevance Vector Machine [D]. South China University of Technology, 2015.

[18] XU Yuge, LIU Li, CAO Tao. On-line soft measuring model based on Fast-RVM[J]. CIESC Journal, 2015, 66(11):4540-4545.