ATLANTIS PRESS

Advances in Engineering Research, volume 94

2nd International Conference on Sustainable Development (ICSD 2016)

Biochemical Oxygen Demand Soft Measurement Based on LE-RVM

Long LUO

Mechanical & Electrical Department , Guangzhou Institute of Technology
No. 465, HuanShi Road, Yuexiu Dist., Guangzhou 510075, P. R. China
College of Railway Transportation, Guangzhou Railway Polytechnic
No. 100, Qinglong Road, Baiyun Dist., Guangzhou 510430, P. R. China
e-mail: 276630181@qq.com

***Abstract*—In order to solve the modeling problem of biochemical oxygen demand (BOD) in wastewater treatment process. This paper proposes an online BOD predictive method based on Laplacian Eigenmaps - Relevance Vector Machine (LE-RVM) . First, the easy to obtain the parameters of the wastewater treatment process is acquired, and then the data preprocessing. The preprocessed parameters is processed by LE, and then is applied as input of RVM to build the soft measurement model of BOD. Experiments show that the prediction model is effective with higher convergence speed. The prediction model indicated that the new methods has better recognition effect and higher computation speed.**

Keywords- Wastewater treatment; BOD; Nonlinear dimension reduction; RVM

I. INTRODUCTION

Nowadays, along with the maturity of wastewater treatment technology in China, the focus of wastewater treatment has shifted to the monitoring of effluent water quality, enhancement of operational management and other aspects. Nevertheless, the parameters of effluent water quality of wastewater treatment plant such as the concentration of BOD are ascertained mostly by artificial test, with complicated and tedious test method and comparatively long test period. Obviously, the test results lag far behind the wastewater discharge process, which is liable to cause secondary pollution. For thiskind of situation, on the one hand, we need to improve measuring instrument, including sensor technology; on the other hand, we need to delve into the soft measurement technology, and develop computer software system for water quality soft measurement on this basis, all of these are of great practical significance and application value for realizing the optimization of water discharge [1,2].

In recent years, scholars both at home and abroad have attached importance to the soft measurement technology in the field of wastewater treatment. Among of them, German Michael Hack [3] and others carry out soft measurement study on parameters of municipal wastewater treatment plant by using a three-layered feed forward network, taking the in-fluent water ammonia nitrogen concentration, conductivity and turbidity as auxiliary variables, estimate the in-fluent water COD, and achieve satisfactory results. South Korea Dong jin Choi [4] and others reduce the dimension of 11 in-fluent water parameters (Q, COD, ammonia nitrogen, NO3-N, TSS, SV, C1, P, PH and temperature T) into 5 primary parameters by applying the principal component analysis(PCA) method, and then use neural network to conduct soft measurement study on the total kjeldahl nitrogen of in-fluent wastewater water, to the extent that the research model is simplified, programming is becoming easy, and dynamic control is realized. However, as wastewater treatment process itself is a nonlinear system, and the data is mainly nonlinear in practice, if this linear dimension reduction method, which can not reflect well the relationship between various parameters, is still applied, their effect will lose. For this reason, a kind of classification method based on the combination of Laplacian eigenmap (LE) with Relevance Vector Machine (RVM) is presented in this paper and applied in soft measurement of wastewater treatment process. By utilizing the good generalization ability of RVM, the model of effluent quality BOD, an important indicator in water quality, is constructed in the case of small sample training. Finally, the actual data measured in wastewater treatment plant in Guangzhou are used to construct models, predict and verify the effectiveness of the method proposed in this paper [5].

II. LE-RVM SOFT MEASUREMENT MODEL

A. LE Dimension Reduction

As one of the nonlinear data dimension reduction technologies, LE dimension reduction is similar to general dimension reduction. Meanwhile it also maps the data points $x_{1,} x_2 . . . , x_m \in M$ in higher into lower dimensional space with redistribution. It is assumed that the data points are distributed in a low dimensional sub-manifold M, and the M is embedded into a high dimensional Euclidean space. We need to use manifold features to find the expression in Euclidean space of low dimensional manifold in which the data distributed. In fact, the algorithm mainly bases on the spectrum diagram theory, to solve a Laplacian generalized eigenvalue [6], and it is described as follows:

(1) Construct neighborhood graph. For data sets $X = \{x_{1,}, \ x_2, \ . . . , x_m \in R^n\}$, apply k neighborhood algorithm to calculate the Euclidean distance between each sample point and other points, and take first k points of minimum distance as the neighborhood points, and

then connect the points with edge lines to obtain adjacency matrix G.

(2) Construct weight matrix. If the bright points on the neighborhood graph are connected, its weight is 1, otherwise 0. And then use the essence of kernel function to establish the estimation formula of edge weights for matrix G, namely, use kernel function to define the weight of edge connecting point i and point j.

$$w_{i,j} = k(x_i, x_j)$$
$$k(x_i, x_j) = \exp(-\| x_i - x_j \|^2 / 2\sigma^2) \quad (1)$$

(3) Calculate the eigenvalue of weight matrix. Based on the eigen decomposition of Laplacian L=D-W, we can obtain eigenvector v0,v1,…,vd that correspond to d+1 minimum eigenvalues, and then use the eigenvector v0, v1,..., vd as embedded space coordinates.

Obtain the overall optimal solution through solving the formula above, and take out the corresponding d eigenvectors $\{x^*_1, x^*_2, \ldots, x^*_d\}$ from small to large according to eigenvalue to constitute processed samples, which can be easily imputed into the support vector machine (SVM) in the next step to construct wastewater treatment soft prediction model. Compared with other nonlinear dimension reduction algorithm, LE dimension reduction algorithm is of high computational efficiency and without partial minimum problem in optimization process.

B. Relevance Vector Machine (RVM)

Relevance vector machine (RVM), proposed by M. E. Tipping, is a sparse probabilistic model similar to support vector machine (SVM). It is based on the active decision theory, to obtain sparse model by removing irrelevant points from the structure of prior parameters under the framework of Bayesian. Compared with support vector machine (SVM), relevance vector machine(RVM) has the following advantages: (1) subjective setting error parameters is avoided; (2) the relevant vector is less than the SVM; (3) kernel function need not meet the Mercer condition, resulting in a greater range of options7.

For a givendata set $\{x_i, t_i\}_{i=1}^{l}$, $x_i \in R^d$, $t_i \in R$, d is the number of dimension of vector data set. In the experiment, the target value is affected by noise, so it is defined as $t = y(x, w) + \varepsilon$, in which, ε is the noise that obeys Gaussian distributions with 0 mean error and variance σ^2, and its probability is $p(\boldsymbol{t}_i \mid \mathrm{y}(x_i), \boldsymbol{\sigma}^2) = \mathrm{N}(\mathrm{y}(x_i, w_i), \boldsymbol{\sigma}^2)$. It is defined that $\mathrm{y}(x, w) = \sum_{i=1}^{l} w_i \varphi(x, x_i) + w_0$, in which, l is the size of data set, $t = (t_1, t_2 ..., t_l)$, $x = (x_1, x_2 ..., x_l)$, w is weight vector, $\varphi(x, x_i)$ is kernel function. In this paper, common radial basis kernel function is used. Assume independent identically distributed, the likelihood function of the whole data set can be expressed as:

$$p(\boldsymbol{t} \mid w, \sigma^2) = (2\pi\sigma^2)^{-\frac{l}{2}} \exp\left\{-\frac{1}{2\sigma^2}\|t - \boldsymbol{\Phi} w\|^2\right\} \quad (2)$$

Where, $w = (w_0, w_1, ..., w_l)$ $\boldsymbol{\Phi}$ is inputted kernel functionmapping, $\boldsymbol{\Phi} = [\boldsymbol{\varphi}(\boldsymbol{x}, x_1) \quad \boldsymbol{\varphi}(\boldsymbol{x}, x_2) \quad ... \quad \boldsymbol{\varphi}(\boldsymbol{x}, x_l)]^{\mathrm{T}}$. In order to improve the generalization ability of the model under the bayesian framework, maximum likelihood method is used to train model weights w. It is defined that each weight obeys Gaussian prior probability distribution , and its expression is:

$$p(w \mid \alpha) = \prod_{i=1}^{l} \mathrm{N}(w_i \mid 0, \alpha_i^{-1}) \quad (3)$$

in this formula: $\alpha = (\alpha_1, \alpha_2, ..., \alpha_l)$ is the prior distribution function of weight w. Based on the combination of formula (3) with formula (4), and according to the rule of Bayesian, the posterior probability distribution of the weight is calculated as:

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\sigma}) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (4)$$

where, $\boldsymbol{\Sigma} = (\boldsymbol{\alpha}^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \boldsymbol{A})^{-1}$, $\boldsymbol{\mu} = \boldsymbol{\alpha}^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{t}$, $\boldsymbol{A} = \mathrm{D}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_l)$, among of which, Dia is the matrix composed of eigenvalue $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_l)$. The formula (4) indicates that the posterior distribution of weight is determined by mean value $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. To estimate weight model, the hyper-function optimal value $\boldsymbol{\alpha}$ should be first estimated and determined. Under bayesian framework, the likelihood distribution of hyper-function could be calculated through the following formula:

$$p(\boldsymbol{t} \mid \boldsymbol{\alpha}, \sigma^2) = (2\pi)^{-\frac{l}{2}} \left|\sigma^2 \mathrm{I} + \boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\right|^{-\frac{l}{2}} \bullet \exp\left\{-\frac{1}{2}\boldsymbol{t}^{\mathrm{T}}(\sigma^2 \mathrm{I} + \boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\boldsymbol{t}\right\} \quad (5)$$

By solving the maximum likelihood distribution, hyper-function optimal value $\boldsymbol{\alpha}_{MP}$ and σ_{MP} can be obtained. So far, the model of target value t is constructed. For input value x_*, the probability distribution of its corresponding output is:

$$p(\boldsymbol{t}_*, \mid \boldsymbol{t}, a_{\mathrm{MP}}, \sigma_{\mathrm{MP}}{}^2) = \mathrm{N}(\boldsymbol{t}_* \mid \mathrm{y}(\boldsymbol{x}_*, \boldsymbol{w}), \sigma_*^2) \quad (6)$$

where, the vector $\boldsymbol{t}_*$ is the predicted value of $\boldsymbol{x}_*$, and its mean value is $y(\boldsymbol{x}_*, \boldsymbol{w}) = \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{x}_*)$, variance shows its uncertainty, and its formula is $\sigma_*^2 = \sigma_{\mathrm{MP}}^2 + \boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_*) \boldsymbol{\Sigma} \boldsymbol{\varphi}(\boldsymbol{x}_*)$.

C. Soft Measurement Modeland Data Processing

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this template and download the file for A4 paper format .

Soft measurement model based on LE-RVM is built as follows: (1) collect various auxiliary variables of wastewater processing, pre-process the data, eliminate abnormal data and normalize the data. (2) set the variables of wastewater processing, including auxiliary variable, controllable variable and disturbance variable as the sample data. (3) reduce non-linear dimension of sample data, get rid of the correlation between variables, and get the sample data after dimension reduction processing. (4) Input the sample data after dimension reduction into RVM, and build the mathematical model of variable that to be measured.

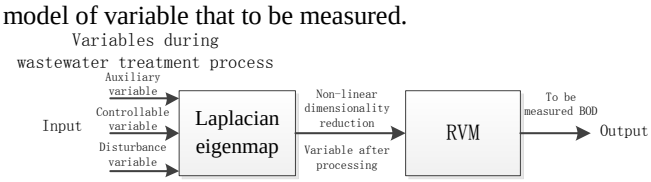


Figure 1. Soft measurement chart based LE-RVM

III. SIMULATION DISCUSSION

In this paper, select parameters such as TN, TP, NH4+-N, DO in aeration basin, T, PH, ORP, MLSS, No3-N and electric conductivity of aeration basin k as the auxiliary variables. Use the 1000 groups of data collected in a wastewater treatment plant in Guangzhou to build soft measurement model, select 350 samples, and get 322 samples after pre-processing through 3σcriterion, from which we choose 200 samples as the training set and the rest 100 samples as the test sample to test the generalization ability of the model. Wastewater soft measurement method based on LE-RVM mainly includes two processes: firstly, get the training data after dimension reduction through LE algorithm. Secondly, build predictor model of BOD through RVM.

The process to reduce dimension of raw data is as follows:

(1) Collect 11 parameters such as the en fluent water flow of original wastewater treatment plant, dissolved oxygen concentration in aeration basin, TN, TP and so on to constitute the signal sample set.

(2) Eliminate abnormal data through 3σcriterion and normalize the data.

(3) Adopt Euclidean distance method to calculate the distance between point A and other points in the sample set. According to the performance of the neighborhood point, we set the number of neighbors K=8, thus the 8 neighbors closest to data points are defined as the data neighborhood, forming data set neighborhood graph G.

(4) Choose the weight value and calculate the local reconstruction weight matrix W according to the neighborhood point of each sample point.

(5) Characteristically map and get the eigenvector matrix through calculating d dimension embedding.

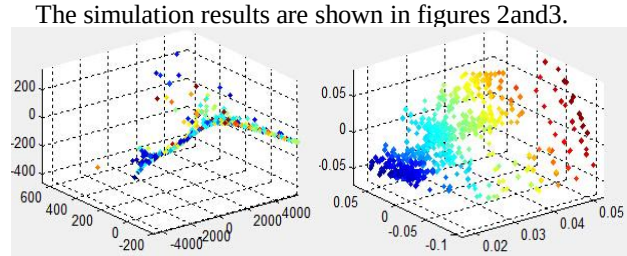
The procedure of predictor modeling:

(1)Input the eigenvector matrix after dimension reduction into relevance vector machine to be trained and choose the auxiliary variable and predictor variable of the model.

(2)Train and predict regression of the model.

(3) Generalize the model to test the predicted effects.

The simulation results are shown in figures 2and3.



(a)PCA (b) LE

Figure 2. dimension reductioneffects comparison through PCA and LE algorithm

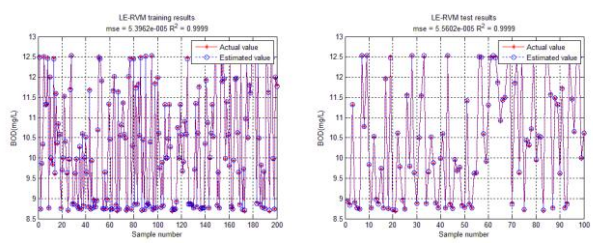


Figure 3. LE-RVM training results

Figure 2(a) shows the effects after dimension reduction through PCA and figure 2(b) represents the results after dimension reduction through LE. Judging from the figures, after dimension reduction through PCA, along with diverse data overlap lest it's difficult to classify correctly. Whereas, LE can map successfully multidimensional data into three-dimensional space. The data sample distributing in three-dimensional space desn't overlap and the signal sample can maintain its relative independence in three-dimensional space which proves that dimension reduction through LE is effective.

Figures 3 and 4 reveals that the error of mean square of effluent water BOD model training set and measurement set based on laplacian eigenmap and relevance vector machine is 0.00005396 and 0.0000556, and the coefficient of determination is up to 0.9999 and 0.9999, which show that the method provided in the paper has a high degree of fitting precision and that dimension reduction through laplacian eigenmap and RVM prediction method has higher degree of prediction precision and generalization ability in small sample.

IV. CONCLUSION

According to the characteristics of wastewater biochemical treatment, we select the important parameter BOD in effluent water quality that is difficult to be measured online in wastewater treatment plant as the predictor index, and the 11 parameters such as easily measured in-fluent water flow, TN, TP,NH4+-N, DO in aeration basin, T, PH, ORP, MLSS, NO3-N and electric conductivity of aeration basin k as the auxiliary variables, build soft measurement model of BOD combining with LE-RVM, eliminate abnormal data and normalize data through 3σcriterion, then reduce non-linear dimension through LE, finally predict through RVM. The result of the experiment validates that soft measurement model combining - laplacian eigenmap with relevance vector machine can tackle the relevant problems effectively between variables with higher degree of prediction process and faster rate of convergence, which provides application value for automatic real-time control of wastewater.

ACKNOWLEDGMENT

This work is supported by project of Guangzhou science and innovation commission (201604010032), project of Guangdong province higher vocational education brand professional construction. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] WC Wang,et al.Analyze Soft Measurement Technology Based Intelligent Algorithm in Sewage Treatment, [J] Applied Mechanics & Materials, 2014:3164-3167

[2] R Marques,et al.Assessment of online monitoring strategies for measuring N2O emissions from full-scale wastewater treatment systems. [J]Water Research, 2016:171-179

[3] M.Hack,M.Konhe. Estimation of waterwater process parameters using neural networks.[J]water.sci.technol,2005:101-105

[4] Dong jin choi,Heek yung Park. A hybid artificial neural network as a software sensor for optimal control of a waterwater treatment process.[J]water.res,2001:3959-3967

[5] Ge, Zhiqiang ,et al.Nonlinear Soft Sensor Development Based on Relevance Vector Machine,.[J] Industrial & Engineering Chemistry Research ,2010:8685-8693

[6] James C. Sanders,et al. Fully Automated Data-Driven Respiratory Signal Extraction from SPECT Images Using Laplacian Eigenmaps.[J] IEEE Transactions on Medical Imaging,2016:1-12

[7] R Bachour,et al. Wavelet-multivariate relevance vector machine hybrid model for forecasting daily evapotranspiration.[J] Stochastic Environmental Research & Risk Assessment,2015:1-15