Proceedings of the 31st Chinese Control Conference
July 25-27, 2012, Hefei, China

AP-LSSVM Modeling for Water Quality Prediction

LI Yan-jun[1], MING Qian [2]

1. School of Information and Electrical Engineering, Zhejiang University City College, Hangzhou 310015
E-mail: liyanjun@zucc.edu.cn

2. Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027
E-mail: mingqianzju@163.com

Abstract: This paper addresses the problem of water quality predicting based on spectrometry. Spectrometry is a kind of novel, quickly, and green soft measurement technology for predicting water quality such as Total Organic Carbon (TOC) criterion. However the analysis accuracy and robustness of predicting model are greatly affected by training samples in modeling process. For solving such a problem, a suitable and effective clustering method is used to improve the model accuracy as well as the computing process time. Firstly, we propose affinity propagation (AP) clustering method with vector angle cosine similarity based on spectral data of water aiming to choose good training samplers. With the most suitable clusters after AP clustering process, a nonlinear modeling method based on a least squares support vector machine (LSSVM) is then given to evaluate TOCs of water samples. Finally, 100 water samples experiment is applied to the regression mode to assess the effectiveness of AP-LSSVM model. The results indicate that the effectiveness and robustness of our proposed model are better than the single LSSVM model and also superior to the model based on k-means clustering.

Key Words: spectrometry, AP, vector angle, LSSVM

1 Introduction

Water is essential to human life and society development. But pollution of water is becoming more and more serious in our country in recent years with development of economic, growth of population, secondly environment pollution, and so on. Generally, water organic comprehensive index is usually used to measure water pollutants including TOC, BOD (Biochemical Oxygen Demand), COD (Chemical Oxygen Demand) et al. Thus, to deal with water pollution, it is essential to develop fast, effective, and green monitoring methods. Spectroscopy such as fluorescence is a widely investigated and applied technology. Many studies and methods have been developed and proposed on the water quality based on spectrometry, e.g. [1, 2, 3, 4].

Reynolds. D. M, et al first proposed the predicting modeling of BOD based on fluorescence at 1997, [5]. The essence of spectroscopy is to establish a high-precision mapping modeling suitable for many kinds of water samples from different origins. However, the accuracy and robustness of regression modeling is difficult to be guaranteed by using single modeling method with given real-life predicting water samples. Therefore, a hybrid modeling way is proposed in this paper based on clustering and regression analysis to. Affinity propagation (AP) clustering algorithm with vector angle cosine as its similarity is proposed to divide the data points to several appropriate clusters, then least squares support vector machine (LSSVM) algorithm, with combination of AP, also known as AP-LSSVM, is utilized to establishing the regression model based on obtained clusters. Testing experiment results indicate that the proposed method is better than the other two methods used as comparing.

This paper is organized as follows. In section 2, an explanation on basic principle of similarity, AP and LSSVM is presented. In section 3, simulation experiment and result are carried out. Section 4, the conclusions are drawn.

2 Algorithm Principle

Fluorescence provides an effectively, sensitively and exactly way to reflect the components of objects. Generally, the procedure of this kind of method is as follows:

- Detecting spectral data and water TOC index of waste water as training samples.
- Computing and obtaining mapping relationship between spectral data and water TOC index.
- Predicting TOC index of new water samples through putting corresponding spectral data into the proposed mapping function/model.

2.1 Similarity

Clustering is widely used in data mining, pattern recognition et al to group data sets into several clusters and make similarity maximized during intra-clusters and similarity minimized during inter-clusters. For clustering algorithm, similarity is significant. The most commonly used similarity is Euclidean distance, cosine and so on. In water spectral data, the fluorescence shape information reflects the components of water, while the intensity value reflects the amount of components. So vector angel may better reflect the feature information of spectral data. Vector angle cosine is usually used as a similarity rule in clustering algorithms, which is computed as (1).

[*]This work was supported by the National Natural Science Foundation of China for Young Scholars (No. 611041149.) and the Graduate Program Foundation in Zhejiang University City College (XZ2012522030)

$$\cos\theta_{ik} = \frac{X_i \bullet X_k}{\|X_i\| \bullet \|X_k\|} = \frac{\sum_{j=1}^{p} x_{ij} x_{kj}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2}\sqrt{\sum_{j=1}^{p} x_{kj}^2}} \qquad (1)$$

where $X_i = [x_{i1}, x_{i2} \cdots x_{ip}]$ is the spectral data of sample i, θ_{ik} is the vector angle of sample i and k.

2.2 AP Clustering

AP clustering algorithm has been introduced by Frey and Dueck at 2007, [6]. It gives the appropriate number of clusters and samples in each clusters, by searching the exemplars that most represent samples for each cluster through computing representative and available matrix (R and A) between samples. In the computing process, optimization function is as (2). It aims to maximize the similarity between every sample and its exemplar.

$$J_{AP} = \max \sum_{i=1}^{n} s(i, v_i), i = 1, 2 \cdots n \qquad (2)$$

where v_i is the exemplar of sample i, $s(x_j, v_i)$ is the similarity between sample i and its exemplar v_i.

The similarity matrix $S_{n \times n}$ is computed as (3).

$$\begin{cases} i \neq k \\ s(i,k) = \cos\theta_{ik}, k = 1, 2 \cdots n \\ i = k \\ s(k,k) = p \end{cases} \qquad (3)$$

where $s(i,k)$ is the similarity of sample i and k, $s(k,k)$ is taken as input and represents the priori preference that training case k be chosen as an exemplar.

The representative matrix $R_{n \times n}$ is computed:

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t. k' \neq k} \{a(k',i) + s(i,k')\} \qquad (4)$$

where $r(i,k)$ is referred to as the responsibility of cluster k for data point i. $a(i,k)$ is referred to as availability of x_k as a candidate exemplar for x_i, [7].

The availability matrix $A_{n \times n}$ is computed as (5).

$$\begin{aligned} & for \quad i \neq k \\ & a(k,i) \leftarrow \min\{0, r(k,k) + \sum_{i' \notin \{k,i\}} \max\{0, r(i',k)\}\} \\ & for \quad i = k \\ & a(k,k) \leftarrow \sum_{i' s.t. i' \neq k} \max(0, r(i',k)) \end{aligned} \qquad (5)$$

Then using formula (6) can make assignments.

$$v_i \underset{k}{\leftarrow} \arg\max\{a(k,i) + r(i,k)\} \qquad (6)$$

where v_i is the new exemplar through the above iterative calculation.

The program procedure of AP clustering algorithm is as follows:

```
AP Clustering Algorithm
Initialization:
    $p=-0.5$, $r(i,k)=0$, $a(i,k)=0$, for all $i$ and $k$;
Similarity matrix:
    use formula (3) to get S matrix;
Updates:
    use formula (4) to get R matrix;
    use formula (5) to get A matrix;
Results:
    use formula (6) to make assignments.
```

2.3 Least Squares Support Vector Machine

SVM based on statistical learning theory is firstly proposed by Vapnik at 1995, [8]. It is suitable for small and nonlinear training samples with kernel approaches for classification and regression. In order to calculate simplified, Suykens and Vandewalle propose a modification algorithm of standard SVM algorithm as LSSVM at 1999, [9]. The most important difference is that LSSVM uses a set of linear equations for training while SVM uses a quadratic optimization problem. In this paper, LSSVM is used to get regression model. Its optimization function is as (7).

$$\begin{aligned} & \min_{w,e} J(w,e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{k=1}^{n} e_k^2 \\ & s.t. \quad y_k = w^T \varphi(x_k) + b + e_k, k = 1, \ ..., \ n \end{aligned} \qquad (7)$$

where w is the weight vector, γ is the penalty coefficient to balance the structural risk and experience risk, e_k is model error, b is bias parameter, $\varphi(\bullet)$ is non-linear mapping, x_k is input data, y_k is output data. According to duality principle, we get (8) to solve the optimization problem.

$$L(w,b,e,\alpha) = J(w,e) - \sum_{k=1}^{n} \alpha_k \{w^T \varphi(x_k) + b + e_k - y_k\} \qquad (8)$$

where $\alpha_k \in R$ are the Lagrange multipliers which can be positive or negative. Then formula (9) is given through solving formula (8).

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{k=1}^{n} \alpha_k \varphi(x_k) \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^{n} \alpha_k = 0 \\ \dfrac{\partial L}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k \qquad (k = 1, 2 \ \cdots \ \mathrm{n}) \\ \dfrac{\partial L}{\partial \alpha_k} = 0 \rightarrow w^T \varphi(x_k) + b + e_k - y_k = 0 \end{cases} \qquad (9)$$

Then we get the following matrix solution:

$$\begin{bmatrix} 0 & \boldsymbol{1}^T \\ \boldsymbol{1} & \boldsymbol{K}+\dfrac{1}{\gamma}I_{n\times n} \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix}=\begin{bmatrix} 0 \\ y \end{bmatrix} \qquad (10)$$

where $\boldsymbol{K}$ ($n\times n$) is kernel function matrix (RBF kernel function is used in this paper), which is $\left\{ \boldsymbol{K} \middle| K(i,j)=\exp(\frac{-\|x_i-x_j\|}{2\sigma^2}), i=1,2\cdots n, j=1,2\cdots n \right\}$, where n is number of samples, σ^2 is as a bandwidth kernel.

Then the output function of LSSVM regression is obtained as:

$$y(x)=\sum_{k=1}^{n}\alpha_k K(x_k,x)+b \qquad (11)$$

where x is the input data (spectral data of predicting sample) as predicting. $y(x)$ is the model output (TOC index value in this paper). α_k, b are the coefficients getting from formula (10).

```
LSSVM Algorithm
Input:
    $\gamma$, $\sigma$
Get modulus $\alpha$, $b$
    use formula (10) to get modulus
Output :
    Use formula (11) to get new sample TOC value.
```

3 Simulation

3.1 Sample

It is more suitable using real-life water samples than titrating solution or dilute solution as our testing experiment samples. That can reflect the diversity and complexity of water in nature.

In this paper, the water samples were collected from 100 different places in different time and locations, including urban domestic sewage and surface water, specifically including river, lake, seawater, waste water from car washer, et al. The effective samples are 79 in urban domestic sewage, and 21 in surface water. Fluorescence data of water was obtained using portable analyzer developed by Zhejiang University. And TOC index was obtained using TOC-VCSH of SHIMADZU Inc, which are based on principle of combustion catalytic oxidation.

Before the training processing, pre-processing is often performed. The characteristics of fluorescence data are high -dimensional, spectral overlap, nonlinear, et al. To solve those problems, we run a blank assay to eliminate the interference of spectral data such as ambient noise and temperature drift. The preprocessed spectral data is shown in figure 1.

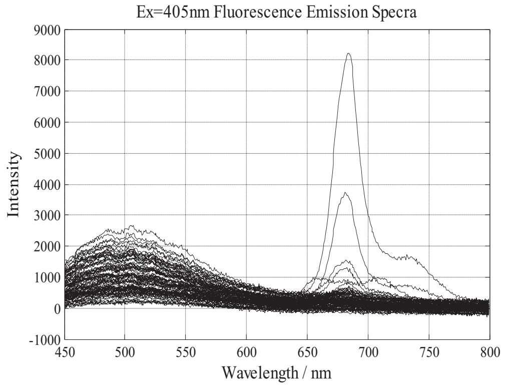

Fig. 1: Preprocessed fluorescence spectral data

3.2 Modeling

For comparison purpose, three kinds of modeling methods for spectral water quality analysis are conducted in this paper to validate the effectiveness of our method. The other two kinds of methods are LSSVM modeling method with k-means clustering and single LSSVM model. K-means algorithm is one of the most commonly used clustering algorithms.

In the proposed modeling methods, there are two main operations process: clustering and regression. In the first stage, the data points are divided into several groups using AP or k-means algorithm. Each clusters contains similar training samples. After the clustering process, LSSVM algorithm is used to establish the predicting regression model between spectral data and TOC index by TOC-VCSH of training samples for each cluster. After that, the final results can be given for a new sample and its vector angle cosine as a criterion is used to judge the appropriate cluster and regression model in order to get analysis value. In this paper, 89 samples are chosen to obtain proposed modeling and parameter optimization as training set, others as test samples to evaluate the effectiveness of three models.

For 89 samples, several clusters are given through the above mentioned k-means and AP clustering methods. Then the most suitable cluster is chosen to develop a regression model using above-mentioned LSSVM algorithm through the cosine similarity between predicting sample and cluster exemplar. Then predicting TOC values are got. To compare the validity of proposed model, RMSE (root mean square error), MAE (mean absolute error), and MAPE (mean absolute percentage error) are used as criteria to show the results of the predicting TOC using three methods.

RMSE, MAE, MAPE are evaluation index, which are calculated as following:

$$RMSE=\sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_j-\hat{y}_j)^2} \qquad (12)$$

$$MAE=\frac{1}{m}\sum_{j=1}^{m}\left|y_j-\hat{y}_j\right| \qquad (13)$$

$$MAPE=\frac{1}{m}\sum_{j=1}^{m}\left|\frac{(y_j-\hat{y}_j)}{y_j}\right| \qquad (14)$$

where y_j is the TOC value of water through TOC-VCSH, and $\hat{y}_j$ is the predicting value through classified modeling of the same water sample j. m is the number of the testing samplers.

3.3 Simulation Results

In our test, the results of k-means and AP clustering methods are shown in Table 1 considering with computing time as reference. And the evaluation index under three modeling methods are showed in Table 2.

Table 1: Comparison of Two Clustering Algorithms

Clustering Algorithm	Samples of Each Cluster	Run Time (s)	Parameter
K-means	{39,37,13}	271.24	k=3
AP	{39,19,15,9,4,2,1}	6.78	p=-0.5

Table 2: Comparison of Three Modeling Methods

Method	RMSE	MAE	MAPE	Parameter

LSSVM	1.9644	1.3346	0.1906	σ=3
				γ=70

K-means-LSSVM	1.1648	1.3072	0.1737	σ=0.9
				γ=200

AP-LSSVM	1.1412	1.0984	0.1591	σ=0.9
				γ=200

From the above results, it shows that AP clustering algorithm runs faster than k-means, and it need not to determine the number of clusters before clustering. According to the prediction effectiveness, three performance criterions such as RMSE, MAE, MAPE are improved 40.7%, 2.1%, 8.9% and 41.9%, 17.7%, 16.5%, respectively, by method based on clustering (k-means and AP, respectively) than single directing LSSVM regression method.

Figure 2 shows the results of three modeling methods. Obviously, the best modeling approach is based on AP clustering, with RMSE, MAE and MAPE improved 2.0% , 16.0% and 8.4%, respectively than k-means-LSSVM. The reason is that: k-means clustering method generates exemplars randomly, while AP method considers all data points as potential exemplars and makes assignments by neighbor information. So AP clustering is more reasonable and stable. Figure 2 also shows the effectiveness and superior performance of our proposed modeling method.

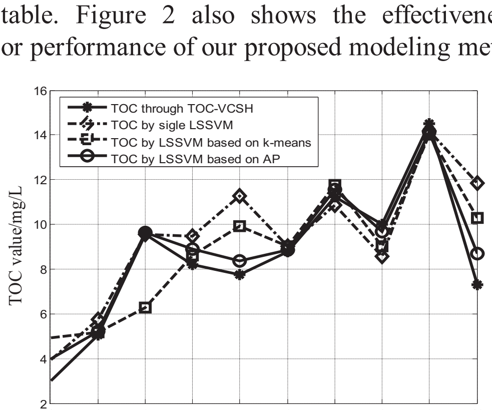

Fig. 2: Results of three modeling methods

4 Conclusion

This paper studies the hybrid model AP-LSSVM for kinds of real-life water samples. We give a mapping relationship between water fluorescence data and TOC index value. From testing results in this paper, the proposed methods can give a predicting model with higher accuracy and robustness than the single modeling method. The results also show that the accuracy of TOC value using proposed methods is significantly improved. From our study, it can be seen that AP clustering performs better than k-means in speediness, accurateness et al, especially AP implements the number of cluster automatically according to characteristics of data while k-means algorithm determines the number of clusters by people. In the future, it is easy to obtain high-dimensional, high-precision spectral data with the development of hardware instrumentation to generate a lot of training samples; thus, the classified modeling method based on spectral data proposed in this paper provides a good way to deal with the samples for modeling. Therefore, there are large spaces for development and improvement for our work on classification modeling approaches in future.

References

[1] Du Shu-xin, DU Yang-feng, WU Xiao-li, Detection of Dissolved Organic Matter based on Tree-dimensional First-order Derivative Fluorescence Spectrometry, Spectroscopy and Spectral Analysis, 30(12): 3268–3271, 2010.

[2] Zhuo Jian-fu, Guo Wei-dong, Deng Xun, Zhang Zhi-ying, Xu Jing, Huang Ling-feng, Fluorescence Excitation- Emission Matrix Spectroscopy of CDOM from Yundang Lagoon and Its Indication for Organic Pollution, Spectroscopy and Spectral Analysis, 30(6): 1539–1544, 2010.

[3] Wu Yuan-qing, DU Shu-xin, YAN Yun, Ultraviolet Spectrum Analysis Methods for Detecting the Concentration of Organic Pollutions in Water, Spectroscopy and Spectral Analysis, 31(1):233–237,2011.

[4] S. A. Baghoth, S.K. Sharma, G.L. Amy, Tracking Natural Organic Matter (NOM) in a Drinking Water Treatment Plant Using Fluorescence Excitation-Emission Matrices and PARAFAC, Water Research, 45: 797–809, 2011.

[5] D. M. Reynolds, S. R. Ahmad, Rapid and direct determination of wastewater BOD values using a fluorescence technique, Water Research, 31(8): 2012-2018, 1997.

[6] Frey B.J., Dueck D. Clustering by Passing Message Between Data Points, Science, 315(5814): 972-976, 2007.

[7] Delbert Dueck, Brendan J.Frey, Non-metric affinity propagation for unsupervised image categorization, IEEE Press, 2007.

[8] Vapnik V, The nature of statistical learning theory. New York: Springer Verlay, 1995.

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