Pol. J. Environ. Stud. Vol. 23, No. 2 (2014), 563-567

Short Communication

Study on Clustering Multi-Model Modeling Method for Activated Sludge Process

Xianjun Du, Ping Yu, Yongwei Ma, Chongyu Ren

College of Electrical and Information Engineering,
Lanzhou University of Technology, Lanzhou 730050, China

Received: 15 May 2013
Accepted: 4 September 2013

Abstract

For a complex process like wastewater treatment, a single model suffers from heavy burden calculation and poor accuracy. A multi-model modeling method based on an improved supervised *k*-means clustering algorithm is proposed. The method introduced the cluster center initialization idea of CCIA algorithm into classical *k*-means clustering algorithm applied to group the data into clusters, and the least squares method is used to construct ARX sub-models. The system model is constructed by weighing all ARX sub-models. The proposed method is used to identify the ammonia concentration model for Benchmark wastewater treatment system and the actual plant process data. Simulation results show that the proposed method can be used to fit nonlinear characteristics of the system with high precision.

Keywords: activated sludge process, *k*-means clustering algorithm, multi-model modeling

Introduction

The activated sludge method is effective for wastewater treatment and has been widely applied and studied. Biological wastewater treatment process systems due to its non-linear uncertainties and other factors make the process a complex mechanism. It is necessary to adopt advanced modeling and control technology for the wastewater treatment process (WWTP) to improve water quality and reduce operating cost. However, it is very difficult to establish a single model for these nonlinear mechanism systems. Researchers often use the input and output data for model identification. For nonlinear systems with a wide range of operating conditions, a single global model is difficult to meet modeling accuracy. However, the multi-model modeling method based on the decomposition principle will improve modeling accuracy, and it received widespread attention in nonlinear system modeling and control [1, 2]. The clustering-based approach is a mature method of data classification, and the use of clustering data classified multi-modeling also has become a subject worthy of study [3-7].

*e-mail: 8843079@qq.com

k-means Clustering

k-means clustering is one of the most simple unsupervised data clustering algorithms with fast convergence, it is suitable for large-scale data set classification [8]. *k*-means clustering uses *k* as the important parameter to divide the data set into *k* clusters.

The algorithm will randomly select the initialization cluster center c_k and the distance of each data points with *k* cluster centers is calculated by equation (1), then determines all the data points belonging to a certain class by equation (2).

$$dist(d_i, c_m) = \sum_{j=1}^{p} (d_{ij} - c_{mj})^2, j \in [1, p] \quad (1)$$

$$dist(d_i, c_j) = \min_{j \in [1,k]} \{dist(d_i, c_1), dist(d_i, c_2), \cdots, dist(d_i, c_k)\} \quad (2)$$

...where *j* represents the data dimension.

When all remaining data points are assigned to the nearest cluster, it will re-calculate the average of each the cluster to re-identify the cluster center by equation (3):

$$c_k = \frac{\sum d_i}{|S_k|}, d_i \in S_k \quad (3)$$

...where, $|S_k|$ represents the number of data points belonging to the k-th cluster.

The above process is repeated until the cluster center no longer sees any change in the end of the clustering operation.

Supervised *k*-means Clustering Multi-Model Modeling Improvement

k-means clustering algorithm is simple and fast convergence, but it has many deficiencies. For example, the clustering result depends on the selection of the *k* value, so the proper value of *k* is difficult to select when we don't have a clear understanding of the data characteristics. The initial value of the cluster centers is randomly selected, it making the clustering process optimal. Unsupervised mechanisms data classification considers just the difference in the input data, but in multi-model modeling process, the final modeling errors cannot be reflected in the classification process, so it will be prone to large modeling errors.

In this paper we improved the initial point selection algorithm and added the supervision mechanism into the classical algorithm to improve its performances in multi-model modeling applications.

It can be found from the above analysis that *k*-means clustering is usually the way to get the initial point randomly selected, and *k*-means algorithm is an iterative algorithm for different initial values that may result in different clustering results, even there is no solution, only when the initial value is close to the final classification results may better clustering results be obtained. Of course, we can also use multiple computing, and select objects as the initial points with a big difference as far as possible to improve the algorithm, but this is obviously not very efficient.

Khan and Ahmad studied the initial value selection of the *k*-means clustering algorithm, and proposed the CCIA (cluster center initialization algorithm) algorithm [9]. The CCIA algorithm consists of two parts: one for the initialization of the cluster centers, and the other a density-based multi-scale data condensation algorithm (DBMSDC) [10]. CCIA algorithm is a novel algorithm proposed based on the fact that each variable property of the data will affect the spatial distribution of the sample data. It assumes that each dimension of the variable properties is in line with the normal distribution and that all data can be divided into k clusters, that is, each dimension corresponding to the normal distribution curve, the data is divided into *k* parts with the equal area. Then, through selecting the equal diversion points as the interval points, we can ensure that the differences of each cluster are as large as possible. However, because the algorithm is based on DBMSDC algorithm, so the parameters are very sensitive to self-sufficiency (such as the setting of the number of core point). In this paper, we learn the cluster center initialization thinking of the CCIA algorithm, and re-cluster the clustering results with a class threshold value. The threshold value is to judgment whether doing cluster again and the termination conditions for the clustering process. If the spatial distance between clusters is larger than the threshold value, then the end of the clustering is the secondary cluster.

The basic steps of the improved clustering algorithm based on the above ideas can be summarized as follows:

Step 1: Set the number of categories of k, create a set of sample data D ($n \times m$-dimensional).

Step 2: For sample data D, initialize the cluster center c_j through selecting the equal diversion points as the interval points and get pattern-strings (s_j) of the each sample data, where j represents the data attributes, $j \in [1, m]$.

Step 3: Repeat Step 2 to get pattern-strings (s with $n \times m$-dimensional) for D, then put each sample data with the same s into the same cluster, and calculate the number of all the current clusters (q), $k \le q \le k^m$.

Step 4: If $q > k$, it represents the current classification results and has re-cluster possibility, with the algorithm going to the next step; otherwise, output the current clustering result as the final clustering results and go to Step8.

Step 5: Calculate separately the data in each cluster for its cluster center value and the spatial distance between the clusters. Analyzing the relationship between the spatial distance and the threshold value (we selected half of the biggest distance between two clusters' spatial distance as the threshold value). If the value is greater than the threshold algorithm, go to next step; otherwise, retain the current classification, set $k = q$, and go to Step8.

Step 6: Re-clustering the two classes with the smallest difference calculated from equations (1) and (2), and the merged cluster center is calculated by equation (4):

$$c_p^{'} = \frac{n_i \times c_{ip} + n_j \times c_{jp}}{N}, p = 1, 2, \cdots, m \quad (4)$$

...where, c_i and c_j represent the two class center of the two clusters going to merge; n_i and n_j represent the number of samples contained in the corresponding class; and N represents the number of all samples in these two classes.

Step 7: $q = q - 1$, return to Step 4.

Step 8: Use the *k*-means clustering algorithm to classify all data samples with the number of the cluster center from above as the initial value of the *k*-means clustering, and output the final classification results.

In addition, data classification based on unsupervised purposes only considers the difference between the input data, without considering the output factors. It does not reflect the final modeling error in the classification process, and it will inevitably produce large modeling errors.

Here, an improved supervised multi-model modeling method is proposed. The basic idea is for the clustering of data points in each category. If the error is too large for the corresponding parameters of the model, we put this point to the other cluster, which will make the model error smaller; then, re-identify the model parameter after this operation.

Simulation Results

Wastewater treatment Benchmark BSM1 (Benchmark Simulation Model No. 1) is developed by the COST group 682/624 and IWA (International Water Association) [11, 12]. It is based on ASM1 (activated sludge model No. 1), and focused on carbon and nitrogen removal. It is covered with the process of sewage treatment system, the simulation model, the simulation steps, and the evaluation standards.

We used the Benchmark BSM1 model built in literature [12] for simulation to get large amounts of data, and model the concentration of ammonia nitrogen by the proposed method in this paper. Here, 900 sets of data selected for modeling and 250 sets for verification test. The output variable is the ammonia nitrogen concentration (S_{NH}); the input variables are the external carbon source flow (Q_C) and the set point of the dissolved oxygen concentration ($S_{O,set}$); the measurable disturbance is the ammonia nitrogen concentration in the source water ($S_{NH,IN}$). The entire variables are written in y, u_1, u_2, and d.

The second-order model structure is described as follows:

$$\begin{aligned} y(k) = {} & a_1 y(k-1) + a_2 y(k-2) + b_{11} u_1(k-1) + b_{12} u_1(k-2) \\ & + b_{21} u_2(k-1) + b_{22} u_2(k-2) + c_1 d(k-1) + c_2 d(k-2) \end{aligned} \quad (5)$$

The modeling error results are shown in Fig. 1.

As we can see from the figure, ultimately the model number is k=3. The error is smaller in some areas and in each sub-model represents the current corresponding operation condition.

To further validate the effectiveness of the strategy in this section we compared the basic k-means clustering method (k=3) with the improved strategy in this section for clustering analysis. The compared results are shown in Fig. 2. Obviously, the minimum error of the improved strategy is smaller than the classical k-means clustering algorithm.

Using the verify data to test the multi-model modeling error of the above two methods. The parity error is shown in Fig. 3.

Also, we can use the standard deviation formula (equation 6) and the maximum absolute error formula (equation 7) to calculate the standard deviation and the maximum absolute error in the modeling and verification process. The results are shown in Table 1, where, N represents the number of data points:

$$\sigma = \sqrt{\frac{\sum_{k=1}^{N} (y(k) - \hat{y}(k))^2}{N}} \quad (6)$$

$$MAXE = \max_{k \in [1,N]} |y(k) - \hat{y}(k)| \quad (7)$$

The data in Table 1 further illustrate that the proposed method in modeling and verification error is smaller than the classical k-means method. The proposed method has higher precision and better fitting performance of the systems non-linear characteristics.

Off-line multi-model modeling and verifying the plant process use the proposed method in this paper. All the 1,200 actual data sets were collected from the Gansu Pingliang sewage treatment plant in China in January 2011. Here, 1,000 sets data for modeling and 200 sets for verifying. Fig. 4 shows the ammonia nitrogen concentration outputs of the model and the actual measured values (first 200 sets). It can be seen that the error between the actual system output and

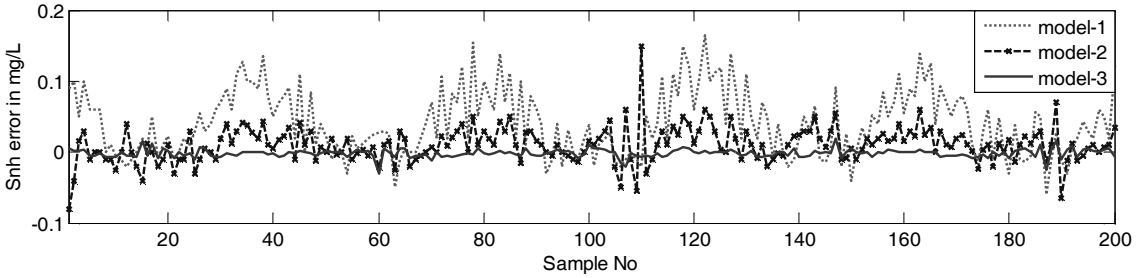

Fig. 1. Multi-model modeling error (first 200 sets).

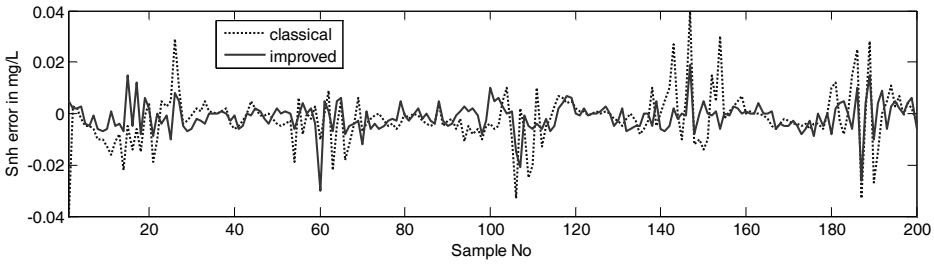

Fig. 2. Compared results of minimum error of classical k-means algorithm and method in this paper.

Table 1. Compared results of these two methods.

Error	Classical *k*-means algorithm	Method in this paper
Standard deviation of the modeling error	0.010927	0.005027
Standard deviation of the verification error	0.0067942	0.0050152
Maximum absolute error of the modeling process	0.04151	0.02705
Maximum absolute error of the verification process	0.03024	0.02432

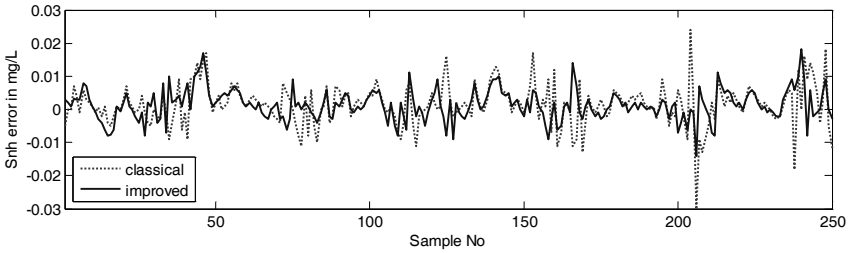

Fig. 3. Compared results of verify error of classical *k*-means algorithm and method in this paper.

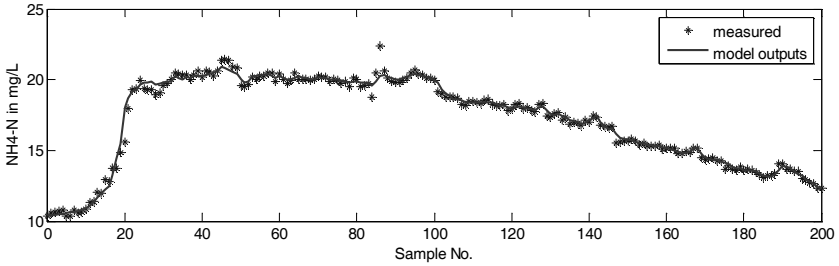

Fig. 4. Ammonia concentration model outputs and measured values.

the model prediction output is very small. Therefore, this model can be used to predict the actual operation of the system.

Conclusions

An improved clustering algorithm is proposed in this paper and applied in the multi-model modeling of the wastewater activated sludge process under Benchmark and the actual plant process data. The initial value selection method and supervised mechanism are improved sufficiently. We also give out the specific implementation steps of the algorithm. This method can be used to determine the initial model parameters of the model's adaptive control. However, multi-modeling obtained by this method is the high precision of its best matching sub-model output error. Hence, how to select the best matched sub-model in control process (multi-model switching or weighted) is needed to continue research in the future.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No. 61064003, No. 61263008) and the Natural Science Foundation of Gansu Province (No. 1212RJYA031).

References

1. LI Q., LEI H., SHAO L., CHEN Z. Multiple-model modeling method based on differential evolution algorithm. Control and Decision, **25**, (12), 1866, **2010**.
2. HUANG Y., ZHANG S. A multi-model LSSVM inverse control system based on nearest neighbor clustering algorithm. Automation & Instrumentation, **2**, 10, **2012**.
3. CONG Q.M., ZHAO L.J., CHAI T.Y. A Multi-model Softsensing Method of Water Quality in Wastewater Treatment Process. Journal of Northeastern University (Natural Science), **31**, (9), 1221, **2010**.
4. XU HAIXIA, LIU GUOHAI, ZHOU DAWEI, MEI CONGLI. Soft sensor modeling based on modified kernel fuzzy clustering algorithm. Chinese Journal of Scientific Instrument, **30**, (10), 2226, **2009**.
5. LI W., YANG Y.P., WANG N. Multi-model LSSVM regression modeling based on kernel fuzzy clustering. Control and Decision, **23**, (5), 560, **2008**.
6. ZHANG Y., LIU G., WEI H., ZHAO W. Multi-model LSSVM modeling for nonlinear systems based on twice affinity propagation clustering. Control and Decision, **27**, (7), 1117, **2012**.
7. FREY B. J., DUECK D. Clustering by passing message between data points. Science, **315**, (5814), 972, **2007**.
8. LIKAS A., VLASSIS N., VERBEEK J.J. The global *k*-means clustering algorithm. Pattern Recogn., **36**, (2), 451, **2003**.
9. KHAN S.S., AHMAD A. Cluster center initialization algorithm for *k*-means clustering. Pattern Recogn. Lett., **25**, (11), 1293, **2004**.

10. MITRA P., MURTHY C.A., PAL S.K. Density-Based Multiscale Data Condensation. IEEE T. Pattern Anal., **24**, (6), 734, **2002**.
11. ALEX J., BENEDETTI L., COPP J., GERNAEY K.V., JEPPSSON U., NOPENS I., PONS M.-N., RIEGER L., ROSEN C., STEYER J.P., VANROLLEGHEM P., WINKLER S. Benchmark Simulation Model No. 1 (BSM1). Prepared by the IWA Taskgroup on Benchmarking of Control Strategies for WWTPs, **2008**.
12. DU X. J., HAO X. H., LI H. J., MA Y. W. Study on Modeling and Simulation of Wastewater Biochemical Treatment Activated Sludge Process. Asian J. Chem., **23**, (10), 4457, **2011**.