Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 13 (17): 3528-3532, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.3528.3532

Multi-model Modeling for Activated Sludge Process Based on Clustering Analysis under Benchmark

Wang Qiang, Du Xianjun, Yu Ping and Ma Yongwei
College of Electrical and Information Engineering,
Lanzhou University of Technology, 730050, Lanzhou, Gansu, China

Abstract: For wastewater treatment processes, a single model suffers from heavy burden calculation and bad accuracy. A modeling method based on ARX (auto-regressive exogenous) multi-model using 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 or second clustering by judging a preset threshold value. It will improve the clustering results to make better services for the subsequent modeling work. And the least squares method is used to construct ARX sub-models. The system model is constructed by weighting all ARX sub-models. The proposed method is used to identify the ammonia concentration model for wastewater treatment system Benchmark. Simulation results show that the proposed method can be used to fit nonlinear characteristics of the system with high precision.

Key words: Activated sludge process, k-means clustering algorithm, ARX model, multi-model modeling

INTRODUCTION

Biological wastewater treatment process systems due to its non-linear, uncertainties and other factors make the process a complex mechanism. It is very difficult to establish a single model of these nonlinear mechanism systems. Researchers often use the input and output data of the system to model identification to model. For nonlinear systems with a wide range of operating conditions, its creation single global model is difficult to meet the modeling accuracy. However, multi-model modeling method based on decomposition principle by decomposed the nonlinear system into a plurality of partial scope of work modeling conditions on local conditions. It will improve the modeling accuracy and it received widespread attention in nonlinear system modeling and control (Li *et al*., 2010; Huang and Zhang, 2010). Clustering-based approach is a mature method of data classification, the use of clustering data classified multi-modeling has also become a subject worthy of study (Cong *et al*., 2010; Xu *et al*., 2009; Li *et al*., 2008; Zhang *et al*., 2012; Frey and Dueck, 2007).

In this study, the research goal is to use an improved k-means clustering algorithm to obtain the classification description of the process data and then create a classification results based multi-model description of the system. For actual use, the system model is weighted synthesized by the adaptive strategy depending on the working conditions.

MATERIALS AND METHODS

k-means clustering: k-means clustering is one of the most simple unsupervised data clustering algorithm with fast convergence, it is suitable for large-scale data set classification (Likas *et al*., 2003). k-means clustering use k as the important parameter to divide the data set into k clusters, such that the data points within the a similarity higher and lower similarity between clusters of data points in the cluster.

The algorithm randomly select the initialization cluster center c_k and the distance of each of the data points with k cluster centers is calculated by Eq. 1, then determined the every data point belongs to the certain class by Eq. 2:

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

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

where, j represents the data dimension.

When all remaining data point is assigned to the nearest cluster, it will re-calculate the average of each cluster to identified cluster center by Eq. 3:

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

Corresponding Author: Du Xianjun, College of Electrical and Information Engineering, Lanzhou University of Technology, 730050, Lanzhou, Gansu, China

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

The above process is repeated until the cluster center is no longer 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 in the absence of a clear understanding of data characteristics. The initial value of the cluster centers is randomly selected, it making the clustering process may be local optimal ended, at the same time, there may appear the sample data set is empty and can not be effectively update. The clustering quality is sensitive to those isolated data points. Unsupervised mechanisms data classification is just consider the difference between the input data, without considering factors such as output, however, in multi-model modeling process, the final modeling errors can not be reflected in the classification process, so it will prone to large modeling errors.

In this study, the initial point selection algorithm and the supervision mechanism will be added into the classical k-means algorithm to improve its performances in multi-model modeling application.

It can be seen from the above analysis:

- k-means clustering is usually the way to get the initial point randomly selected
- k-means algorithm is an iterative algorithm for different initial values may result in different clustering results even there is no solution
- Only when the initial value is close to the final classification results may be better clustering results

Of course, you 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 high efficiency.

Khan and Ahmad study on the initial value selection of the k-means clustering algorithm and proposed the CCIA (cluster center initialization algorithm) algorithm (Khan and Ahmad, 2004). The CCIA algorithm consists of two parts, one is for the initialization of the cluster centers and the other is density-based multi-scale data condensation algorithm (DBMSDC) (Mitra *et al*., 2002). CCIA algorithm is a novel algorithm proposed based on the fact that the each variable property of the data will affect the spatial distribution of the sample data. It assumes that each dimension of the variable properties in line with the normal distribution and 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, it can ensure that the differences of each cluster as large as possible. However, because the algorithm is based on DBMSDC algorithm, so the parameters are required the user to set is very sensitive to the self-sufficiency (such as the setting of the number of core point). The cluster center initialization thinking of CCIA algorithm is introduced into the new method to improve the algorithm. Re-cluster the clustering results in accordance with the classical k-means clustering algorithm with a class threshold value for judgment whether doing cluster again. If the spatial distance between clusters is larger than the threshold value, then the end of the clustering, otherwise 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 the 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 (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 cluster (q), $k = q = k^m$
- **Step 4:** If q>k,it represents the current classification results has the re-cluster possibility, algorithm go to the next step; Otherwise, output the current clustering result as the final clustering results, turn to Step 8
- **Step 5:** Calculate separately of 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 (select the half of the biggest distance between two clusters spatial distance between the clusters as the threshold value); If the value is greater than the threshold, algorithm go to next step; Otherwise retain the current classification and set k = q, turn to Step 8
- **Step 6:** Re-clustering the two classes with the smallest difference calculated from Eq. 1 and 2 and the merged cluster center is calculated by the Eq. 4:

$$c'_p = \frac{n_i \times c_{ip} + n_j \times c_{jp}}{N}, p = 1, 2, \cdots, m \tag{4}$$

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

- **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 get from above as the initial value of the k-means clustering and output the final classification results

In addition, data classification based on unsupervised classification purposes only consider 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, put this point to the other cluster which to make the model error smaller; then, re-identification of the model parameter after this operation.

Ultimately, the steps of the new clustering-based supervision multi-model modeling algorithm are as follows:

- **Step 1:** Add the designed appropriate excitation signal to the multi-input multi-output system to incentive it sufficiently, then get out the input and output data for recognizing through simulation
- **Step 2:** Preprocessing the input and output data, such as map these data to the same scale space by normalization
- **Step 3:** Select the ARX (auto-regressive exogenous) model order and divide the data into two types of modeling data and test data
- **Step 4:** Use the improved k-means clustering algorithm mentioned above to cluster the modeling data into k classes
- **Step 5:** Use the method of least squares for the ARX model fitting to obtain the initial parameters of the models for each type of data
- **Step 6:** Calculate the error of data points for each model, the corresponding data points are classified into corresponding categories based on the principle of maximum error; when all data points move neither, algorithm turn to Step 8
- **Step 7:** Re-fitting the new clustered data to get the parameters of the sub-model and then back to Step 6
- **Step 8:** If all data points move neither, use the currently model set as the best multi-model fitting set
- **Step 9:** Use the test data for model fitting test, if the test error is in the allowed range, considering this group of model as the best expression of multi-model; Otherwise to Step 4

SIMULATION RESULTS

Wastewater treatment Benchmark BSM1 (Benchmark Simulation Model No.1) is developed by the COST group 682/624 and IWA (Alex *et al*., 2008; Du *et al*., 2011) based on ASM1 model. It focused on the carbon and the nitrogen removal process. It covered with the process of sewage treatment system, the simulation model, the simulation steps, the water data under a variety of weather source and the evaluation standards. Researchers can purchase the COST official business WWTP simulation software package; also can be self-developed in Matlab/Simulink, C/C++ language. Researchers can do evaluation study on the control strategy and the control performance by simulation the software package of sewage treatment system. The advent of Benchmark makes up the standard measurement and the same simulation environment for the performance of different control strategies and it is conducive to the sewage treatment experts and researchers to select the optimal control scheme.

Using the Benchmark BSM1 model built in literature (Du *et al*., 2011) to simulation to produce large amounts of data and model the concentration of ammonia nitrogen by the proposed method in this study. Select 900 sets of data for modeling and 250 sets of data for verification test. Here, 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 described as follows:

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

The modeling results are shown in Fig 1.

As it can seen from the figure, ultimately, the model number is k = 3.

To further validate the effectiveness of the strategy in this section, the compared results of the minimum error and the verify error are shown in Fig. 2. Obviously, in

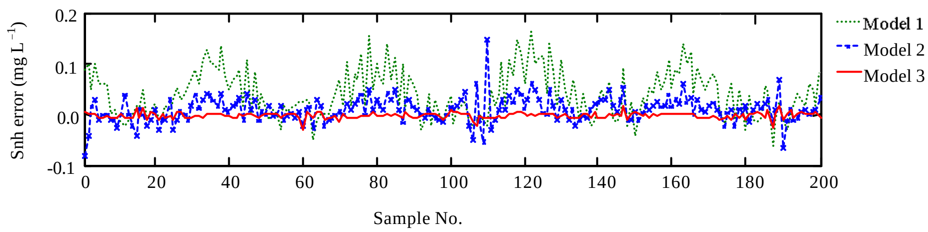

Fig. 1: Multi-model modeling error (first 200 sets)

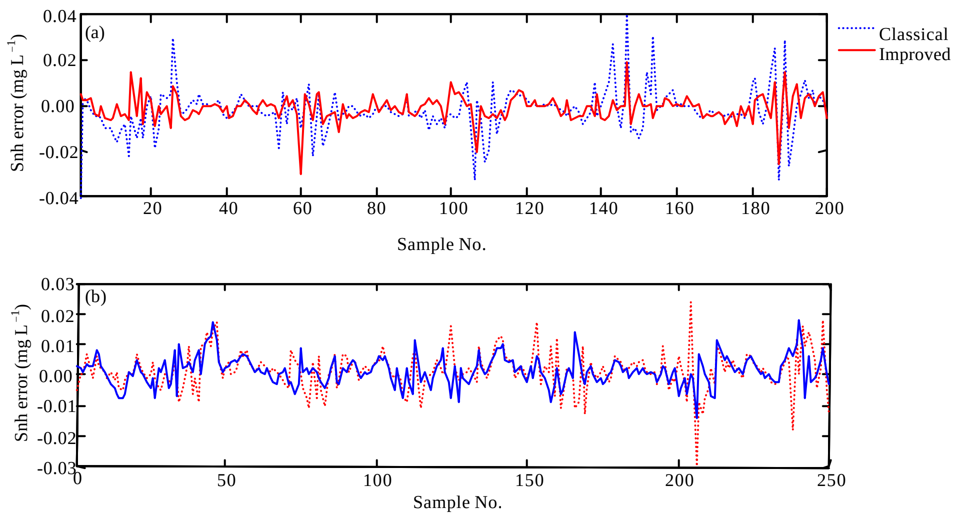

Fig. 2(a-b): Compared results in error of the classical k-means algorithm and method in this study (a) (a) Minimum error and (b) Verify error

Table 1: Error check results of these two methods

Error	Classical k-means algorithm	Method in this study
Standard deviation of the modeling error	0.0109270	0.0050270
Standard deviation of the verify error	0.0067942	0.0050152
Maximum absolute error of modeling process	0.0415100	0.0270500
Maximum absolute error of verify process	0.0302400	0.0243200

Fig. 2a, 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 and the parity error is shown in Fig. 2b.

The standard deviation formula (Eq. 6) and the maximum absolute error formula (Eq. 7) are to calculate the standard deviation and the maximum absolute error in modeling and verify process. The results are shown in Table 1. Where, N represents the number of data points:

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

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

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

CONCLUSION

An improved clustering algorithm is proposed in this study and applied in the multi-model modeling process of the wastewater activated sludge process under Benchmark. The initial value selection method and supervised mechanism are improved sufficiently. The specific implementation steps of the algorithm are also given out. This method can be used to determine the initial model parameters of the model adaptive control. However, multi-model 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 need to continue research in the future.

ACKNOWLEDGMENTS

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

REFERENCES

Alex, J., L. Benedetti, J. Copp, K.V. Gernaey and U. Jeppsson et al., 2008. Benchmark simulation model No. 1 (BSM1). Prepared by the IWA Taskgroup on Benchmarking of Control Strategies for WWTPs. https://www.iea.lth.se/publications/Reports/LTH-IEA-7229.pdf.

Cong, Q.M., L.J. Zhao and T.U. Chai, 2010. A multi-model softsensing method of water quality in wastewater treatment process. J. Northeastern Univ. (Nat. Sci.), 31: 1221-1225.

Du, X.J., X.H. Hao, H.J. Li and Y.W. Ma, 2011. Study on modelling and simulation of wastewater biochemical treatment activated sludge process. Asian J. Chem., 23: 4457-4460.

Frey, B.J. and D. Dueck, 2007. Clustering by passing messages between data points. Science, 315: 972-976.

Huang, Y. and S. Zhang, 2010. Multi-model LSSVM inverse control system based on nearest neighbor clustering algorithm. Autom. Instrum., 2: 10-13.

Khan, S.S. and A. Ahmad, 2004. Cluster center initialization algorithm for K-Means clustering. Pattern Recognition Lett., 25: 1293-1302.

Li, Q.L., H.M. Lei, L. Shao and Z.X. Chen, 2010. Multiple-model modeling method based on differential evolution algorithm. Control Decis., 12: 1866-1869.

Li, W., Y. Yang and N. Wang, 2008. Multi-model LSSVM regression modeling based on kernel fuzzy clustering. Control Decis., 5: 560-562.

Likas, A., N. Vlassis and J.J. Verbeek, 2003. The global k-means clustering algorithm. Pattern Recognit., 36: 451-461.

Mitra, P., C.A. Murthy and S.K. Pal, 2002. Density-based multiscale data condensation. IEEE Trans. Pattern Anal. Machine Intell., 24: 734-747.

Xu, H., G. Liu, D. Zhou and C. Mei, 2009. Soft sensor modeling based on modified kernel fuzzy clustering algorithm. Chin. J. Scient. Instrum., 10: 2226-2231.

Zhang, Y., G.H. Liu, H.F. Wei and W.X. Zhao, 2012. Multi-model LSSVM modeling for nonlinear systems based on twice affinity propagation clustering. Control Decis., 7: 1117-1120.