Multi-model modeling and its application of urban sewage treatment based on clustering analysis

Y U Ping1
College of Electrical and Information Engineering, Lanzhou University of Technology, 730050 Lanzhou, Gansu, China
E-mail: lutyup@163.com

Abstract. For urban sewage treatment processes, a single model suffers from heavy burden calculation and poor accuracy. A modeling method founded on an advanced supervised k-means clustering algorithm is proposed. The method introduced in this paper is the cluster center initialization idea of CCIA algorithm into a classical k-means clustering algorithm is applied to corral the data into clusters or second clustering by judging a pre-set threshold value, and the least squares method is applied to construct the ARX sub-models, and the system model is then constructed by weighing all ARX sub-models. The proposed method is used to determine the ammonia concentration model for wastewater treatment system benchmark and the actual WWTP practice. Simulation and application research results show that the proposed method can be effectively used to fit the nonlinear characteristics of the system with high precision.

1. Introduction
Biological sewage treatment process system has a complex mechanism because of its non-linear, uncertainties characters and other factors. It is very difficult to establish an individual model for these nonlinear mechanism systems. Researchers often use the input and output data from the system to model identification 1. For nonlinear systems with a vast scope of operating conditions, its creation single global model is often difficult to meet the necessary modeling accuracy. However, if we can decompose the nonlinear system into a plurality of partial scope of work conditions, then we would improve the modeling accuracy. This decomposition principles based approach is called multi-model modeling method. It has received widespread attention in nonlinear system modeling and control [1, 2]. A clustered approach is an advanced technique of data classification. Additionally, the use of clustering data classified multi-modeling has also become a subject worthy of study [3-7].

2. Methods
2.1. K-means clustering
K-means clustering is one of the simplest unsupervised data clustering algorithms with fast convergence, and it is suitable for large-scale data set classification [8]. K-means clustering uses k as
the vital parameter to divide the data set into “k” clusters, so that the data points within the higher similarity between clusters of data points in the cluster.

The algorithm randomly selects the initialization cluster center, $c_k$, and the distance of each of the data points with $k$ cluster centers is then calculated by an equation (1). Next, it is determined that every data point belongs to the certain class by equation (2).

$$\text{dist}(d_i, c_j) = \sum_{j=1}^{k} (d_{ij} - c_{ij})^2, \ j \in [1, p]$$  \hspace{1cm} (1)

$$\text{dist}(d_i, c_j) = \min_{j=1, \ldots, k} \{ \text{dist}(d_i, c_1), \text{dist}(d_i, c_2), \ldots, \text{dist}(d_i, c_k) \}$$  \hspace{1cm} (2)

In the equation, $j$ represents the data dimension.

When all remaining data points are assigned to the nearest cluster, the algorithm will re-calculate the average of each cluster in order to identify the cluster center using the equation (3).

$$c_k = \frac{\sum_{d_i \in S_k} d_i}{|S_k|}, \ d_i \in S_k$$  \hspace{1cm} (3)

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

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

2.2. Supervised k-means clustering multi-model modeling improvement

The $k$-means clustering algorithm is a simple and quick convergence, however it contains many deficiencies. For example, the clustering result relies on the selection of $k$’s value; therefore the proper value of $k$ is difficult to select when there is an absence of a clear understanding surrounding the data characteristics. The initial value of the cluster centers is randomly selected, which would make the clustering optimize process ended with local optimal condition or empty cluster. The clustering quality is sensitive to the isolated data points; unsupervised mechanisms data classification only considers the difference between the input data, without taking factors such as output into account. However, in multi-model modeling processes, the final modeling errors cannot be reflected in the classification process, so it will prone to large, long-term corruption.

In this paper, we focus on the initial point selection of the $k$-means clustering algorithm, and discuss how the supervision mechanism will be added into the classical algorithm to improve its performance in multi-model modeling applications.

We can see from the above analysis that $k$-means clustering is primarily the method to obtain the initial point (which is randomly selected), and that the $k$-means algorithm is an iterative algorithm for different initial values that may result in different clustering groups even if there is no complete 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 methods, and select objects as the initial points with a big difference (as far as possible) to improve the algorithm; however the efficiency of this would not be very high.

Khan and Ahmad studied in 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 second for density-based multi-scale data condensation algorithm (DBMSDC) [10]. The CCIA algorithm is an advanced equation based on the fact that the each variable property found in the data set 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 data) can be divided into $k$ clusters. This means each dimension corresponding to the normal distribution curve is divided into $k$ parts with the equal area. Then, by selecting equal diversion points as interval points, we can ensure the differences of each cluster as large as possible. However, because the algorithm is based on the DBMSDC algorithm, the parameters
required to be set by the user are very sensitive to the self-sufficiency (such as the setting of the number of core points). The algorithm has been improved, and is expressed in this paper. We use the main idea of the CCIA algorithm to initialize the cluster center, and then re-cluster the results in accordance with the classical k-means clustering algorithm. This is done with a class threshold value for judgment by performing the cluster again, and examining 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 has been reached, however if the distance does not exceed the threshold value, there is a secondary cluster.

The basic steps of the above-mentioned improved clustering algorithm can be summarized as follows:

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

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

Step 3: Repeat step 2 to obtain pattern-strings ($s$ with $n \times m$-dimensional) for $D$, then place each of the sample data with the same $s$ into one cluster, and calculate the number of the current cluster ($q$), $k \leq q \leq km$.

Step 4: If $q > k$, it symbolizes that the current classification results have a re-cluster possibility, therefore proceed to the next step. Otherwise, output the current clustering result as the final clustering results, and turn to Step 8.

Step 5: Individually calculate the data in each cluster for its cluster center value and the spatial distance between the clusters. Then, analyzing the relationship between the spatial distance and the threshold value (we selected the half of the largest distance between two clusters spatial distance between the clusters as the threshold value). If the value is greater than the threshold, the algorithm will automatically move to the next step. Otherwise, it will retain the current classification, and set $k=q$. If this occurs, turn to Step 8.

Step 6: Re-clustering the two classes with the least amount of difference is calculated from equations (1) and (2), and the merged cluster center is calculated by equation (4), which is shown below:

$$c_p = \frac{n \times c_o + n \times c_j}{N}, p = 1, 2, \cdots, m$$ (4)

In this equation, $c_o$ and $c_j$ represent the two class centers of the clusters about to merge. Additionally, $n_o$ and $n_j$ represent the number of samples contained by the corresponding class; subsequently, $N$ 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 corresponding number of the cluster center get from above as the initial value of the $k$-means clustering, and output the final classification results.

Additionally, data classification based on unsupervised classification purposes only considers the differences between the input data without considering the output factors; it does not reflect the final modeling error found in the classification process, and will inevitably produce significant modeling errors.

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

Ultimately, the process of the new clustering-based supervision multi-model modeling algorithm is as follows:
Step 1: Add the designed appropriate excitation signal to the multi-input and multi-output systems to incentive it sufficiently, then withdraw the input and output data for identification during simulation.

Step 2: Preprocessing the input and output data, for example, by mapping the data to the same scale space by normalization.

Step 3: Select the ARX (auto-regressive exogenous) model, then 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 group the modeling data into k classes.

Step 5: Use the least squares method 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 data points are classified into corresponding categories according to the principle of maximum error. When all data points move neither algorithm, proceed to Step 8.

Step 7: Re-fit the new clustered data to obtain the parameters of the sub-model, then return to Step 6.

Step 8: If all data points move neither, use the current model set as the best multi-model fitting set.

Step 9: Use the test data for the model fitting test, if the test error is in the allowed range, consider this group of models as the best expression of a multi-model, and otherwise return to Step 4.

3. Results and discussion

3.1. Simulation results

The wastewater treatment Benchmark BSM1 (Benchmark Simulation Model No. 1) is developed by the COST (the European Co-operation in the Field of Scientific and Technical Research) group 682, 624, and IWA (International Water Association) [11, 12], and is based on ASM1 (Activated Sludge Model No. 1). ASM1 is focused on carbon and nitrogen removal, and it also includes the system process description, the simulation model, the simulation steps and the evaluation standards. The advent of Benchmark makes up the standard measurement, and the same simulation environment for the performance of different control strategies. It is conducive for the sewage treatment experts and researchers to select the optimal control scheme.

Using the Benchmark BSM1 model built-in literature [12] to simulate and produce large volumes of data, and model the concentration of ammonia nitrogen using this paper’s proposed method. Select 900 sets of data for modeling, and 250 sets of data for verification testing. In this situation, the output variable is the ammonia nitrogen concentration ($S_{\text{NH}}$), and the input variables are the external carbon source flow ($Q_{\text{C}}$). The set point of the dissolved oxygen concentration ($S_{\text{O}_2,\text{set}}$); while the measurable disturbance is the ammonia nitrogen concentration in the source water ($S_{\text{NH},\text{IN}}$). The variables are written in $y$, $u_1$, $u_2$ and $d$.

The second order model structure described as follows:

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + b_1 u_1(k-1) + b_2 u_1(k-2) + b_3 u_2(k-1) + b_4 u_2(k-2) + c_1 d(k-1) + c_2 d(k-2)$$

(5)

The modeling results are then shown below in figure 1.

![Figure 1. Multi-model modeling error (first 200 sets).](image)
Figure 2. Compared results of minimum error of classical \( k \)-means algorithm and method in this paper.

Additional validation of the effectiveness of this strategy is the comparison of the basic \( k \)-means clustering method \((k=3)\) with the improved strategy in this section for clustering analysis. The compared results are displayed above in figure 2. Obviously, the smallest error of the improved strategy is less than the classical \( k \)-means clustering algorithm.

By using the verified data to test the above two method’s multi-model modeling errors, the parity error is shown in figure 3.

Figure 3. Compared results of verify error of classical \( k \)-means algorithm and method in this paper.

Also, we can use the standard deviation formula (found in equation (6) and the maximum absolute error formula (found in equation (7) to calculate the standard deviation and the maximum absolute error in the modeling and verification process. These results are shown below in table 1. In the equations, \( N \) represents the number of data points.

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

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

| error                                  | classical \( k \)-means algorithm | method in this paper |
|----------------------------------------|-----------------------------------|----------------------|
| standard deviation of the modeling error | 0.010927                          | 0.005027             |
| standard deviation of the verify error  | 0.0067942                          | 0.0050152            |
| maximum absolute error of modeling process | 0.04151                           | 0.02705              |
| maximum absolute error of verify process | 0.03024                           | 0.02432              |
The data seen in table 1 further illustrates that the proposed method in modeling and verify error is smaller than the classical $k$-means method. The proposed method contains a higher level of precision, and offers better fitting performance of the system of non-linear characteristics.

3.2. Application results
The Sewage Treatment Plant of Pingliang City in Gansu Province was built in 2006. The designed processing capacity is 50,000 m$^3$ per day. It is designed for nitrogen and phosphorus removal by the Carrousel oxidation ditch process of live sewage, and a minute amount of industrial waste water, which benefits a population of about 220,000 people. In order to secure a better effect from the biological nitrogen and phosphorus removal process, modification to the original design is begun at the start of 2011. The initial oxidation ditch will be transformed to aerobic section with a bottom aeration of A$_2$/O process. It is put into operation at the end of 2012.

![Figure 4](image.png)

**Figure 4.** Ammonia concentration model outputs and measured values.

![Figure 5](image.png)

**Figure 5.** Verify error of ammonia concentration.

Off-line multi-model modeling and verification of the plant processes is conducted with 1,200 sets actual data obtained from the sewage treatment plant in January 2011, by means of the proposed method in this paper. 1,000 sets data for modeling, and 200 sets for verifying was used to conduct the study. Figure 4 shows the ammonia nitrogen concentration outputs of the model and the actual measured values, as well as the error between the model output and actual measurement values, which are shown in figure 5.

4. Conclusions
In this paper, an improved clustering algorithm is proposed and applied in the multi-model modeling process of the urban sewage activated sludge process. This study is conducted under Benchmark simulation and actual plant application. The initial value selection method and supervised mechanism are improved sufficiently throughout the study. Additionally, we offer the specific implementation steps for the algorithm. This method can be used to determine the initial model parameters of the model adaptive control. However, the multi-model obtained by this method is the high precision of its best matching sub-model output error. Therefore, the method of selecting the best matched sub-model in the control process (multi-model switching or weighted) is need to continue research in the future.
Acknowledgments
This study is supported by the Excellent Young Teachers Program of Lanzhou University of Technology (No. Q201408), and the Natural Science Foundation of Gansu Province (No. 1212RJYA031).

References
[1] Li Q L, Lei H M, Shao L, et al. 2010 Control and Decision 12 1866-7
[2] Huang Y R and Zhang S D 2012 Automation & Instrumentation 2 10-11
[3] Cong Q M, Zhao L J and Chai T Y 2010 J. Northeastern Univ. (Natural Science) 9 1221-2
[4] Xu H X, Liu G H, Zhou D W, et al. 2009 Chinese Journal of Scientific Instrument 10 2226-7
[5] Li W, Yang Y P and Wang N 2008 Control and Decision 5 560
[6] Zhang Y, Liu G H, Wei H F, et al. 2012 Control and Decision 7 1117
[7] Frey B J and Dueck D 2007 Science 5814 972-5
[8] Likas A, Vlassis N and Verbeek J J 2003 Pattern Recognition 2 451
[9] Khan S S and Ahmad A 2004 Pattern Recognition Lecture 11 1293
[10] Mitra P, Murthy C A and Pal S K 2002 IEEE Transaction on Pattern Analysis and Machine Intelligence 6 734-6
[11] Alex J, Benedetti L, Copp J, et al. 2008 Benchmark Simulation Model No. 1 (BSMI) ( London: IWA Publishing House)
[12] Du X J, Hao X H, Li H J, et al. 2011 Asian Journal of Chemistry 10 4457-9