Research on the COD soft-measuring mechanism based on the clustering approach

Y. Z. Feng[1], G. C. Chen[1], D. Feng[2] & M. Zhuo[1]
*[1]Department of Management and Environmental Engineering,
Logistic Engineering College, China
[2]School of Architecture Liverpool University, UK*

Abstract

By analyzing laws of growth and reproduction of microorganisms in the activated sludge (AS) sewage treatment system, this paper proposes a multi-neural network (NN) COD (chemical oxygen demand) soft-measuring method based on clustering approach for the sewage treatment project. Various reasons which might affect the accuracy of the model are analyzed. Experiments show that the same radiuses of multiple neural network diffusion constant are quite close which means the prediction accuracy is high and the soft-measuring method based on clustering approach is suitable for COD measuring.
Keywords: sewage treatment, clustering approach, neural network, microorganisms.

1   Introduction

Cybenko [1] has proved theoretically that if there are plenty of training data and without restriction of the size of the network, modelling based on NN can always get a satisfactory model structure. But, in the actual industrial process, people often need to face the limited effective process data and due to the real-time requirements, the network structure also cannot be expanded unlimited. Modeling effect normally relies on the good generalization ability of the network. Paper [2] proposed a robust classification method by combining different models which are based on neural networks with fuzzy combination approach. Paper [3] presented a method to improve model prediction accuracy and robustness by adding different models together. Paper [4] discussed Stacked Neural Network Approach for process modeling. The basic idea of these

methods is establishing several independent neural network models and then combining the forecast outputs of different networks together through a certain method. Such a structure with multi-models is conductive to the real-time soft measurement with parameters changing in a wide range and meets the technological requirements for accurate description of the dynamic characteristics of the process.

2   Theoretical basis of NN model in the AS process

Shi *et al.* [5] developed an urban sewage treatment plant operation simulation and forecast expert system which shows that there is a strong correspondence between simulated effluent COD (E-COD) value of the aeration tank and its influent COD (I-COD) value. HRT (hydraulic retention time) is a large time-delay process in the sewage treatment, the correspondence misfit caused by time-delay has great influence on model, while the correspondence between effluent and influent COD value is the basis of modeling based on multiple NN. Therefore, it is very important to research on the causes of delay error, analyze ways of effectively reducing the delay error and determine the theoretical basis of multi-neural network modeling.

2.1 Growth and reproduction laws of microorganisms

Bacterial growth can be divided into at least four phases [6]: 1) Lag phase. After cells are seeded into fresh medium, they are living in a new growth environment and will not immediately divide over a period of time and the number of cells basically maintain constant. 2) Exponential phase. In the exponential phase, cells grow and divide with the greatest rate, the number of cells increases logarithmically, different components in bacteria increase regularly according to the corresponding proportion. 3) Stationary phase. Because of the nutrient consumption, increase of metabolites and pH value changes, the environment is not suitable for bacterial growth anymore, cell growth rate progressively decrease to zero. Bacterial population growth changes from exponential phase to stationary phase. Stationary phase owns the highest number of living bacteria and maintain stable. 4) Death phase. Because of the nutrient depletion, toxic metabolite accumulation and pH value changes, the bacterial death rate gradually increase while the living bacteria and their metabolic activity progressively decrease. Cells begin autolyze and senesce.

2.2 Microorganisms proliferation laws in AS

The proliferation of Microorganisms is the inevitable result of AS reaction and organic matter degradation. The proliferation of microorganisms is actually the growth of activated sludge. Activated sludge is a community of microorganisms composed of multiple bacterial colonies with a variety of physiological functions. All kinds of microorganisms work together to finish the biodegradation process of AS. The proliferation laws are complicated but the general trend can still be expressed as Fig. 1.
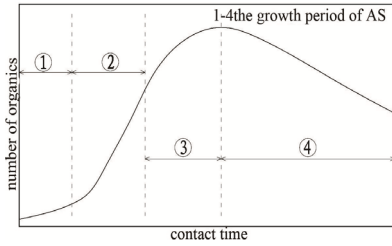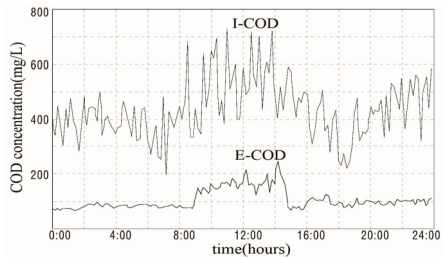
Figure 1:   Growth curve of AS.

Figure 2:   Curve of I-COD and E-COD.

In the practical wastewater treatment process, the conventional AS process mainly depends on the 3 and 4 stage of bacterial population growth phases, while improved wastewater treatment processes based on AS process can use different time interval of the AS growth curve [6]. There are obvious rising and falling phases in the growth curve, which is conducive to divide the treatment process into several stages so as to facilitate the realization of multi-modeling approach.

Proliferation laws of microorganisms in the wastewater treatment process are the interpretation of the process of growth and reproduction of microorganisms.

Through the detailed analysis of all kinds of factors associated with the growth process, researchers can draw a growth curve of activated sludge and decide the application stages of conventional activated sludge process. The growth curve and its corresponding application are the theoretical basis of composite model modeling approach. According to the actual data distribution of sewage treatment plant, laws of the daily routine of residents also should be considered as one of the parameters of the model. The I-COD and E-COD values in Fig. 2 come from a wastewater treatment plant in Chongqing, China [7], with a sampling period of 10 minutes. These data apparently can be divided into four stages in accordance with time-series, including 0:00-8:00, 8:00-14:00, 14:00-20:00, and 20:00-24:00. The fact that there is only a relatively high effluent COD value in 9:00-14:30 on the effluent COD curve indicates that the COD removal ability of the sewage plant can be divided into three relatively stable sections.

Based on the analysis of the 24 hours of operation data on September 1, with the consideration of measurement error, we believe that the wastewater treating capacity of activated sludge microorganisms can change according to certain rules. This is the basis of multiple neural networks.

3   Data sources and processing

The research results of this paper are verified by the data provided in literature [8].

3.1 Normalization processing

If the number of learning samples of a sub-network is small, the estimated value of similar data to a large extent will depend on the number of learning samples of the network and the accurate estimates cannot be obtained because of the large influence of other sub-networks [9]. From the point of view of mechanism of NN, as the NN input nodes, each process variables in the NN is equally important. So, when the output is so sensitive to a certain value of an input variable that even a small deviation of the input variable can cause a large output error, the value of α should be increased. Conversely, the α in the corresponding formula can be assigned a slightly larger value to improve the sensitivity of output to the secondary input variables. In this paper the α is 0.9.

$$X_i' = \alpha \times (X_i - \min(X)) / (\max(X) - \min(X)) + 1 - \alpha \tag{1}$$

3.2 Untrusted data processing

Human error comes from manual testing stage or data entry stage. According to 3σ principle, the standard deviation is calculated by the Bessel formula: $S = \sigma = \left[\dfrac{1}{n-1} \sum v_i^2 \right]^{1/2}$, data set 7, 39 and 43 are deleted according to the calculating results. Meanwhile, the COD value in data set 22 is 457.5mg/L, which greatly exceeds the corresponding integrated wastewater discharge standard of China (120mg/L); the effluent COD value in data set 22 is greater than its influent COD value, which is obviously not reasonable, so, both data sets are also removed from the learning samples.

3.3 Analysis of probable influencing factors

The above data are analyzed in the form of multi-dimensional clustering analysis, the sample group affecting effluent quality contains multiple characteristics. Data with similar sample characteristics have the small deviation of the sample performance, which is based on the reliability of all sample data. In fact, there are different kinds of measurement error on field sampling data, which will affect the clustering results. There are main factors and secondary factors in sample groups, if there is a big deviation of the secondary factors, the clustering results will deviate the clustering intervals defined by the main factors. The more characteristics exist in the sampling groups; the far more deviation from the real situation is inevitable, especially when all different secondary factors can affect the results of clustering analysis. Even the membership determination method proposed in literature [9] is a better method than that chosen in literature [10], the problem still cannot be fundamentally solved. This paper intends to try one-dimensional clustering method, namely to cluster only the most important factors. If successful, it is apparently helpful to meet the real-time requirement of on-line running. The first step is to verify and choose the main factors from the data relationship between removal ratio of COD and each auxiliary variable.
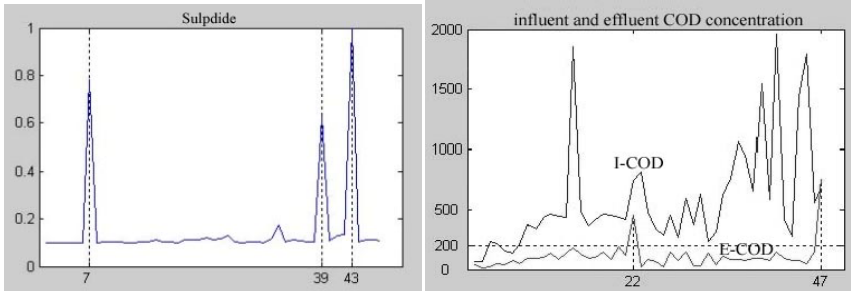
Figure 3:   Normalized value of sulphide.

Figure 4:   I-COD and E-COD value.

It can be seen from Figs. 5(a) to (d) that the variation trends of T, S, pH, Q and MLSS do not have any correspondence with the removal ratio of COD or its time delay. While from Fig. 5(d), there is an obvious correspondence between COD removal ratio and I-COD values. Furthermore, from the definition of COD removal ratio ($\eta = (COD_{in} - COD_{out})/COD_{in}$), it is quite reasonable to choose I-COD value as the main factor, which is also consistent with Zhang's research results of orthogonal test[11].
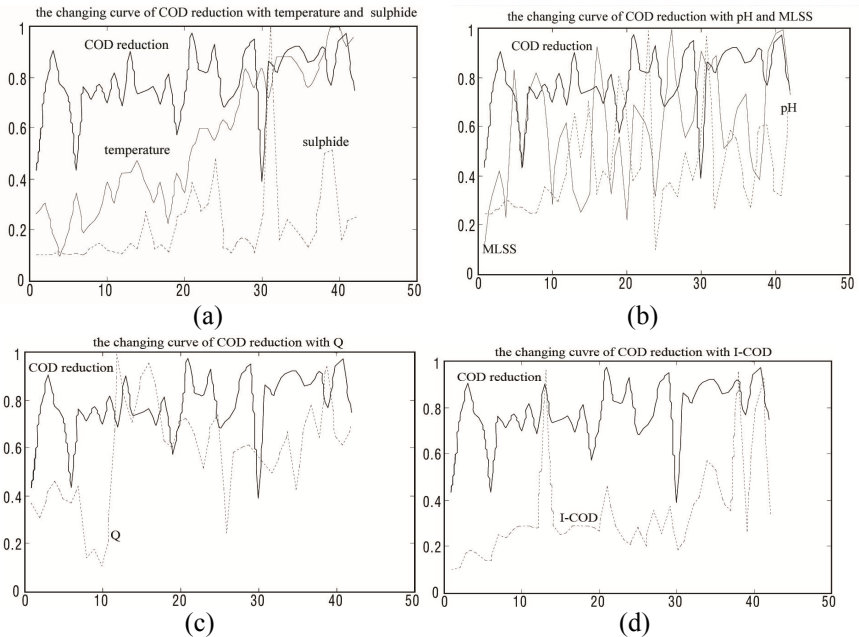

Figure 5:   The changing curve of COD reduction with (a) temperature and sulphide, (b) pH and MLSS, (c) Q and (d) I-COD.

3.4 Clustering analysis of historical data

After determining the I-COD value as the main factor, the normalized data of COD data are clustered. The sample data distribution is shown in Fig. 6 in accordance with the order from small to large. In order to improve generalization performance, it is better to resolve the classification problem within the range of 0.1-0.5. The cluster radius r lies in [0, 0.2]. From the convenience of numerical rounded perspective, the value of r can be given as 0.05, 0.1 and 0.2, and the subnet numbers can be 4, 2 and 1. From the perspective of network applications, r=0.05 is a better choice.

Table 1: Data distribution statistics.

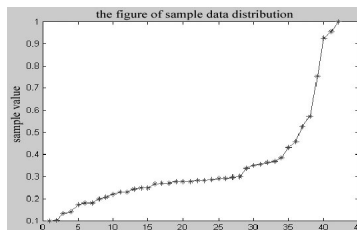| Clustering distribution	Data distribution SN of sample group	Sample size
0.1-0.2	1,2,3,4,5,6,28,32 8,9,10,11,12,13,15	8
0.2-0.3	16,17,18,19,20,21,2 4, 25,26,27,30,33,42	20
0.3-0.5	29,31,34,38,40,46, 23,25	8
0.5-1.0	14,36,37,41,44,45	6

Figure 6: Data distribution.

4 Establishment of multi-NN soft-measuring model

Since the NN constructed in this article is also a distributed architecture, using current research results for reference, this paper takes fuzzy classifier shown in Fig.8 as a tool for network synthesis. The fuzzy clustering unit shown in Fig.7, named as PCA-DRBF, can realize the classification function of input space, in which sample characteristics X is translated as T through a PCA dimension reduction technique, after that, eigenvectors reduced from M dimensions to N dimensions are distributed to N mutually exclusive subspaces. Different from the PCA-DRBF method, this paper proposes an RBF clustering unit shown as Fig. 8, which directly chooses a sample eigenvector, I-COD value, as the classification indicator. From the specific operation concerned, there are many differences

between the two methods. For a new process input data, the fuzzy clustering unit of PCA-DRBF method firstly identifies its membership for each sub-network, and according to the different memberships, outputs of all sub-networks are integrated as the total output of the entire distributed network.

Because the cluster radius r equals 0.05 in this paper, the cluster centers are also fixed as 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85 and 0.95. According to the characteristics of samples, the cluster centers in this paper are limited to 0.15, 0.25, 0.4 and 0.75. The respective subordinate network of each new input is determined by Euclidean distance. In addition, to reflect the fuzziness, using the formulation of kernel radius presented for reference [9], this paper defines the kernel radius as 0.005, which means if an input data lies in the interval of [-0.005, 0.005], it can be viewed as belonging to two sub-networks and do not be classified strictly. In the training process, it is determined by the comparison of output values of both sub-networks with test value that which sub-network is more suitable for such an input data. Such a sub-network is retained as retention program. In the forecasting process, the calculating process is a little complicated. Firstly, the correlation degree of input value which lies in the interval [-0.005, 0.005] and the corresponding sample in the retention program will be calculated [12], then the correlation degree will be compared with a set value to determine which subnet will be chosen. Through the above fuzzy approach, the complexity of the wastewater treatment process has been taken into fully consideration.

Assuming that there are N sub-networks $\{RBF_i, i = 1,\ldots,N\}$ in the distributed network shown in Fig. 8, the membership degree set of new input variable x (influent COD value) for each sub-network is $\{\mu_i, i = 1, \ldots, N\}$, then the output of the whole neural network Y is:

$$Y = \sum_{i=1}^{N} \mu_i \bullet f_{RBF_i}(x) \quad (\mu_i = 0 \quad OR \quad \mu_i = 1) \tag{2}$$
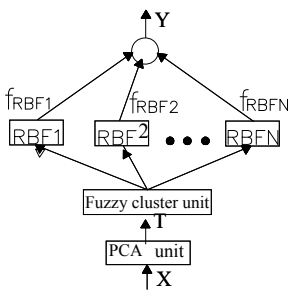

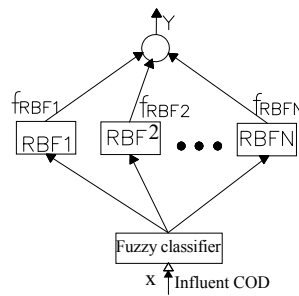Figure 7:   PCA-DRBF architecture.   Figure 8:   RBF architecture.

The membership degree of input x to each sub-network is obtained by eqn (3) and eqn (4):

$$d_i = \left\| \mathrm{x} - \mathrm{x}_{center_i} \right\| \tag{3}$$

$$\mu_i = 1, \eta_{j \neq i} = 0, if : 0 \leq d_i \leq 0.005, i = 1, \cdots, N \tag{4}$$

If the input data x lies in the interval [-0.005, 0.005], then, assuming there are two sample groups X_1 and X_2, corresponding to the subnet i and subnet i+1 respectively, based on the calculation of $d = \left\| \mathrm{X} - \mathrm{X}_1 \right\|$ and $d = \left\| \mathrm{X} - \mathrm{X}_2 \right\|$, the finally subnet is chosen by the comparison of the relative size of the two membership degree.

4.1 Evaluation of the model adaptability

Through the clustering analysis, there are only 8 data in the first class. Taking 7 data among the 8 as training sample data and another one as input data, researchers can always find the best fit point because of the convergence performance of RBF network. However, because the sample size is too small, the expansion constant distribution of the network does not show any regularity and cannot be used for accurate forecast. Similar problems lie in the third class and fifth class. Taking relative error (RE) and recognition rate (RR) as criteria of assessing the forecast accuracy of RBF network, the forecast analysis is performed according to the data of second class.

$$RE = (Y - P\_test) / P\_test \times 100\% \tag{5}$$

$$RR = m / n \times 100\% \tag{6}$$

where, m is the number of actual outputs which are satisfied with the condition $\left| (Y - P\_test) / P\_test \right| \times 100\% < \varepsilon$, ε is the forecast precision, $\varepsilon = 10\%$ in this paper. n is the total output number. There are two understandings about parameter Y and P_test. One explanation is that the two parameters represent the desired output and actual output respectively, the other is that the two parameters represent the anti-normalized values of the desired and actual output. This paper adopts the second understanding. Comparing with reference [8], the differences include: (1) calculation accuracy. The relative error of the model proposed in this paper ranges from 0.06% to 23.01%, average relative error is 7.66%, recognition rate is 83.3% ($\varepsilon = 10\%$). The relative error of the original model ranges from 0.9% to 9.2% and the average relative error is 3.4%, recognition rate is 92.3% ($\varepsilon = 5\%$). (2) details of models. There are eight hidden layer neurons in the original BP neural network model, training times are not less than 10,000, excitation function of the hidden layer is tansig and learning sample size is 30. When the SSE of the BP model is less than 0.001, the training stops and the predict process begins when the test data are input. The hidden layer of this RBF network provided in this paper are defined in the calculation process, after only 13 times of iteration, the SSE drops to lower than 10^{-28}, the learning

sample size is only 13 groups. The above analysis shows that even the effectiveness of the proposed new model is not as good as the original model, but it is obviously that convergence of the RBF neural network is much better than the original. So, the new neural network is the best choice for online operation.

4.2 Analysis of the results of model prediction

Theoretically, when the diffusion constant of network is 0.25, the prediction results will be highly accurate; while the actual constant is 0.265, with a small deviation from the ideal value, which verify the correctness of the theory analysis of this paper. The new neural model can greatly reduce search range in selecting the optimal value. From the above technical indicator comparison, the forecast precision of the new model is not high enough because of various reasons. Through analysis of these causes, this paper put forward several issues that should be paid attention to in future research.

4.2.1 The lack of sample size

To achieve the desired generalization ability (prediction error is less than a given value ε) requires a large number of samples, which in turn will increase the amount of network calculation. So, there is contradiction between the generalization ability and the amount of calculation of learning. Literature [13] provides an experiential formula: $m \geq d_{vc} \div \varepsilon$, where d_{vc} is the *VC dimension* of the network. It represents the capacity of function class and can be approximate evaluated by the number of independent parameters. According to this experiential formula, the sample size m should equal 50 ($\varepsilon = 0.1$, $d_{vc} = 5$), while there are only 20 groups used in this case actually. The lack of sample size is one of the reasons causing the lower prediction precision. From the common ideas of researching an issue, we rerun the model by taking cluster radius r as 0.2 so that the sample size increases to 30. The distribution of training and testing samples are shown in Table 2. Data in Table 3 and Fig. 3 shows that the relative error of the model prediction results ranges from 0.16% to 13.27%, the average relative error is 4.67% and the recognition rate is 83.3%. Compared with the initial results above, network performance has been greatly improved. The calculation results indicate that with the increase of sample size, the prediction ability of the model is also improved apparently. Meanwhile, if the cluster radius is 0.2, the diffusion constant needs to be taken 0.371 to ensure the best prediction results. Comparing with the theoretical best value of 0.3, the deviation is increasing, which means the value of cluster radius r (0.2) is not as appropriate as the original value (0.05). In addition, test results also confirm that with the sample size increasing, the number of hidden layer neurons increases synchronously. For BP neural network, the convergence performance declines sharply with the increasing of sample size. Even the operation time of RBF network also increase with the increasing of sample size, its iterations (10^1) is much less than that of BP neural network (10^4), which proves that RBF network has obvious advantages when the sample size is large.

Table 2:   Training set and testing set.

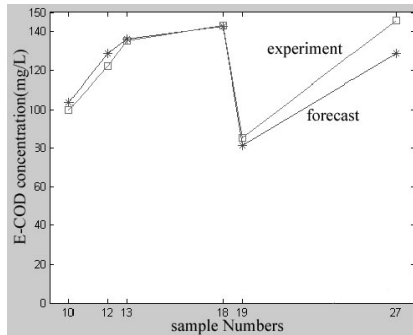| Test sets	10, 12, 13, 18, 19, 27
Train sets	1, 2, 3,4 ,5, 6, 8, 9,11,15,16,17,20, 21,23,24,25,26,28,29,30,31,32,33, 34,35,38,40,42,46

Figure 9:   Testing and forecast value.

Table 3:   Comparisons of test and prediction results of RBF network (mg/L).

SN	Test results	Prediction results	RE (%)	Desired output	Actual output	RE (%)
10	99.8	103.9	3.95	0.541	0.563	3.84
12	122	128.8	5.26	0.658	0.694	5.14
13	135.5	136.1	0.46	0.729	0.732	0.45
18	143	142.8	0.17	0.768	0.767	0.16
19	85.2	81.2	4.93	0.464	0.443	4.75
27	146	128.9	13.27	0.784	0.694	12.97

4.2.2 Unsatisfactory choice of secondary variables

Through the initial mechanism analysis, the number of auxiliary variables requires 8, which can be reduced by principle component analysis, by the way. However, in order to adapt to the special case of limited sample data and because of the traditional conception that as long as the oxygen supply is not excessive, there is no longer need to take DO into consideration, so, DO, an important control variable, was excluded from the secondary variables. The first deficiency to do so is the loss of the meaning of soft measurement (saving energy consumption), it is inappropriate to exclude the DO factor out of auxiliary variables set because the mechanism analysis shows that the value of DO does have a certain influence on the effluent COD value.

4.2.3 Selection scheme of samples of training set and test set

In the modelling process discussed in this paper, the reason the samples we chose from secondary variables are not maximum or minimum is mainly because the effect of interpolation is better than that of the generalization. But in reality, it is obvious that no guarantee can be provided that the value of input variable is not the maximum or minimum value of current sample source.

4.2.4 Choice of clustering center point and secondary important factors

Two modelling processes are completed in this paper. The deviation between clustering center point and diffusion constant is very small in the first modeling process (0.25 and 0.265) and large in the second modeling process (0.3 and 0.371). The reasons of the existence and change of the deviation include the following aspects. Firstly, the distribution of samples does not fluctuate with uniform amplitude around the clustering center. Secondly, finding the classification formed by the main vector from the five feature vectors will certainly be influenced by the other four vectors. The author assumes that with the sample size increasing, the fluctuation caused by the first reason will decrease and the fluctuation caused by the second reason will increase. Because the more samples, the greater the impact to the classification based on the main feature vector from other feature vectors will be. Choosing the most important feature (i.e. secondary important factor) from the rest feature vectors to balance this effect may solve the problem. This mode of thinking is quite similar with hierarchical clustering method. After the main and secondary factors are chosen from the feature vector set, the selection of sub-network for input variables can be done by fuzzy classifier which is based on the rules of fuzzy control.

4.2.5 Dynamic tracking of the actual process

There are many cases on the research and application of soft measurement, but only a few cases combine prediction model with time series [14, 15]. This is a common problem of current soft measurement technique research and application. As far as wastewater treatment process is concerned, there is no corresponding relationship between the value of input variable detected by operator at a particular moment and the effluent quality of the sewage plant at that time. A time delay which does not be mentioned in many literatures is closely related with HRT in engineering application. The reasons that not many reference literatures point this out are not only because of the sampling period (normally one day) of the data, but also because there is no correlation with time, which means time parameter does not be looked upon as a variable. This adverse effect is more apparent in a distributed multi-network model, because those data with far different time series might be clustered into the same class in the modeling process. Due to the obvious ignorance of time variable, the credibility and reliability of the predicting outcomes provided by such kind of clustering method are questionable. It is obvious that with the pre-treatment of the coarse and fine screen and primary settling tank, the index data of T, pH must have changed after mixing with the original material in the reactor. Such a change process is a complex function of time: the initial change is faster, and then it gradually slows down, accompanying by the potential influences of reaction

process. Neglecting or ignoring the impacts of these factors and other oversimplifications of the process may compromise the credibility of model predictions. The best and feasible way to solve these problems is the implementation of online algorithm based on a small sampling period (such as $\phi = 10\,m$) of input variables. As a result, effluent quality at time T is a function of feed water quality conditions at time (T-t) and other procedure parameters, which reflects the real process of sewage treatment. Otherwise, soft measurement technique transforms into just some kind of data-based prediction research. If the importance of mechanism analysis to understand the reaction process has been ignored in the modeling, the neural network is obviously not a successful soft instrument. The clustering method needed by multiple network models can use the time-based clustering analysis technique proposed in 2.2 of this paper. If there are historical data of the same time frame of last cycle, they can also be clustered into the same class.

5   The conclusion and prospects

This paper analyzes the microbial basis of multiple neural network modeling based on clustering method in the sewage treatment works. Case analysis shows that the proposed clustering method not only has diffusion constant close to the cluster radius, but also has better prediction accuracy. This paper also analyzes a variety of reasons which may affect the accuracy of model and point out that online real-time algorithm is the future research direction and the focus of subsequent research.

Acknowledgements

References

[1]   Cybenko G. Approximation by superpositions of a sigmoidal function. Math [J]. Control Signal System, 2, pp. 303-314, 1989.
[2]   Cho, S. B. & Kim, J.H. Combining multiple neural networks by fuzzy integral for robust classification. *IEEE Trans on System, Man and Cybernetics*, 25(2), pp. 380-384, 1995.
[3]   Bates, J.M. & Granger C.W.J. The combination of forecasts. *Operations Research Quarterly*, 20(1), pp. 319-325, 1969.
[4]   Sirdhar, D.V., Seagrave, R.C & Bartlett, E.B., Process modeling using stacked neural networks. *AICHE J*, 42(9), pp. 2529-2539, 1996.
[5]   Shi Han chang, Diao Huifang, Liu Heng *et al*. Application of treatment plant operation simulation and forecast software. *China Water and Wastewater*, 17(10), pp. 61-63, 2001.

[6]   Li Jun, Yang Xiushan, Peng Yongzhen. *Microorganism and water treatment engineering*, Chemical industry press: Beijing, pp.74-98, 2002.

[7]   Feng Yuzhao. *Study on hybrid intelligent control for variable rate anoxic biological filter wastewater treatment* [D]. Chongqing: Chongqing University Ph.D. Thesis, pp.5-37, 2004.

[8]   Gao Ping ping, *The Study Based on Neural Network of Water Quality Prediction in Sewage Disposal* [D].Chengdu : Southwest Jiaotong University, pp. 11-49, 2005.

[9]   Chang Yuqing, Lin Tian, Wang Fuli. Online Learning of Distributed Networks and Its Application in Soft Sensing. *Chinese Journal of Scientific Instrument,* 23(3), pp. 295-297, 2002.

[10] Luo Rongfu, Shao Huihe. Local learning methodology for distributed networks and its application in inferential control. *Acta Automatic Sinica*, 20(6), pp.739-741, 1994.

[11] Zhang Wenyi, Zhong Meiying, Cai Jian-an. Modeling Study of Activated Sludge Process Based on Artificial Neural Network. *Water & Wastewater Engineering*, 28 (6), pp.12-15, 2002.

[12] Wang Xudong, Shao Huihe, Luo Rongfu. Distributed RBF neural networks and its application in soft sensing. *Control Theory and Applications*, 15(4), pp.558-563, 1998.

[13] Yan Pingfan, Zhang Changshui. *Artificial neural network and simulated evolutionary computation*. Tsinghua University Press: Beijing, pp. 32-37, 2002.

[14] Luo Jianxu, Shao Huihe. Developing dynamic soft sensors using multiple neural networks. *Journal of Chemical Industry and Engineering*, 54(12), pp.1770-1773, 2003.

[15] Ran Weili, Qiao Junfei. BOD Soft-Measuring Approach Based on PCA Time-Delay Neural Network. *Transaction of China Electro technical Society*，19(12), pp.78-82, 2004.