A self learning algorithm based on data value lifecycle model for the accurate chemical dosing of wastewater treatment

Zhengang Zhai1,*, Bingtao Gao1, Dan Liu1, Tengjun Yao1, Li Zhang1, Zhiyuan Pan1 and Jing Cao1

1The 36th Research Institute of CETC, Jiaxing, Zhejiang, China

*Corresponding author e-mail: wlmqgqjfw@126.com

Abstract. In the process of wastewater treatment (WWT), the instability and hysteresis of the chemical dosing cannot make sure the stability of the water quality. At the same time, chemicals were wasted or overdosed in the process. In this study, we proposed a self learning algorithm, which based on the regression method modified using the data value lifecycle model to compute the accurate quantity of chemical dosing. The mechanisms of discard and cultivation data were established to make it alive in the algorithm. The algorithm is self learning according to the wastewater characteristics as the time goes on using the mechanism. We can make sure the quality of the wastewater is stability and economical by using the artificial intelligence. The experiment approves that the artificial intelligence algorithm was useful and economical to chemical dosing of wastewater treatment.

1. Introduction

In many papers[1,2,3,4,5,7,8], principal components analysis(PCA) and multiple regression were used to real time monitoring of the treatment process and predictive control for improving Moving Bed Biofilm Reactor(MBBR) plant performance. Multiple regression coupled with PCA was used to predict performance [11,12,13,14]of the wastewater treatment. In the paper [19,20], the chemical oxygen demand (COD) and total phosphorus (TP) of the wastewater was monitored using PCA. Multiple regression method was applied to predict influent COD and TP. Model predictive control(MPC) was considered as an advanced control scheme to monitor and optimize process [6,9,10,19,16] of the wastewater treatment.

With the development of the data mining, deep learning and machine learning, more researchers have attempted to find the relationship between the characteristic of influent and effluent wastewater without the hard sensors [17,18,21,22,23]. Some researchers have used several statistical learning and machine learning algorithms as soft sensor for water quality prediction [24,25]. Neural networks were used as the water quality soft sensor to prediction [15]. However, they believed that the data from sensors have the same contribution to chemical dosing in these methods at different time. Actually, the very new data have bigger contribution than the historical data.

In this paper, we proposed a new self learning algorithm that the regression method was modified using the data value lifecycle model. The regression model trained by the latest data is different at different decision time and adaptive to the wastewater quality. We take into account time attributions of the data, and calculate time weight at different time, which represents different contribution to the chemical dosing. As the same time, the discard and cultivation mechanism of data was established to make it alive. The experimental results show this algorithm is very useful.
2. The algorithm
We use the supervised learning method to train decision model, shown in Formula 1. We use the latest historical data to learning the correlation relationship between quantity of the chemical dosing and the wastewater quality.

\[ \sum W \cdot DV + B = 0 \]  \hspace{1cm} (1)

We use the second-order regression function. \( DV \) is the eigenvector, \( DV \in \Omega \), and \( W \) is the weight of data life value calculated based on the time. \( B \) is the offset vector. \( \Omega \) is a set of the feature vectors.

\[ \Omega = (DV_0, DV_1, \cdots, DV_m) \]  \hspace{1cm} (2)

Where \( DV_m \) is the \( m \) dimensional eigenvector of the wastewater quality variables, corresponding to \( DV_m = (X_m, Y_m) \), where \( X_m \) is wastewater quality variables \( X_m \), \( Y_m \) is the correspondence chemical dosing, for example the PH value is \( X_m \), the dosage of the acidic drug is \( Y_m \).

3. Data value lifecycle model
In the model, the data is regarded as a life body, for example an animal or a bird, and is born and died likes a person. So, in the process of data storage and management, we should record the time of producing data as an attribution of the data, which was the most imported factor. We create the time vector set corresponding to the feature vector space of the water quality variables. It is represented by \( T \).

\[ T = (t_0, t_1, \cdots, t_m) \]  \hspace{1cm} (3)

Where \( T \) is a set of \( t_m \), corresponding to the eigenvector \( m \), \( m \in N \). \( N \) is a set of the nature number.

3.1. The cultivation mechanism of data
According to the actual chemical dosing time \( t_p \), we select the most live data \((t_m, t_{p-1})\) in the most recent period as the important data, where \((m < p -1)\), and \((t_m < t_{p-1})\).

We select the historical data during \((t_n, t_m)\) time, where \((t_n < t_m)\) and \((n < m)\), to calculate the average of each dimension value of the wastewater quality variables. We used the formula 4, to calculate and update the data of each dimension value at the \( t_m \) time. By this way, we establish the cultivation mechanism of data.

\[ DV_{(i,t_m)} = \sum_{j=m, k=t_n}^{j=m, k=t_n} DV_{(i,j,k)}/(m-n+1) \]  \hspace{1cm} (4)

Where \( DV_{(i,t_m)} \) is the data value of wastewater quality variables when the \( i \) dimension data value at the \( t_m \) time. \( \sum_{j=n, k=t_n}^{j=n, k=t_n} DV_{(i,j,k)} \) is the sum of these value.

3.2. The discard mechanism of data
In this mechanism, it is very simple by setting a very litter value or zero value to eliminate the contribution of the chemical dosing decision. We can select the historical data before \( t_m \) time to discard. We set the value of each feature vector to very small value or zero.
3.3. The data value lifecycle model

We take the time information of the feature vector into the data value analysis model to calculate the time weight of the data. The time weight represents the contribution of data to the chemical dosing. The size of the weight indicates the contribution of the data on the WWT.

Based on the difference between the chemical dosing time $t_p$ and the data production time $t_x$, which is within the range $x \in (m_p - 1)$, we can compute the contribution of the data at the time $t_x$ to the chemical dosing. We believe that live data is more valuable than historical data. The weight presents the different contribution, which is evaluated and calculated according to formula 5.

$$W_{(i,t_x)} = e^{rac{1}{\text{abs}(t_p - t_x)}}$$  \hspace{1cm} (5)

Where $x \in (m_p - 1)$, $W_{(i,t_x)}$ is the value of the $i$ eigenvector at the time $t_x$ relative to the time of the decision time $t_p$. A large value means that the impact on quantity of the chemical dosing is large, otherwise it is small.

To the cultivation data, we also update the time weight of the $m$ wastewater quality variables at the time $t_m$ by the formula 6.

$$W_{(i,t_m)} = \frac{\sum_{j=m,k=t_x}^{m-n,k=t_m} W_{(i,j,k)}}{(m-n+1)}$$ \hspace{1cm} (6)

Where $W_{(i,t_x)}$ is the time weight of the $i$ wastewater quality variables relative to the decision $t_p$ time, at the $t_m$ time. $\sum_{j=m,k=t_x}^{m-n,k=t_m} W_{(i,j,k)}$ is the sum of all time weights in the selected time period $(t_x, t_m)$.

To the discard data, we take the time weight of the data value before the time $t_m$ as $\lambda$ by the formula 7.

$$W_{(i,t_x)} = \lambda$$ \hspace{1cm} (7)

Where $W_{(i,t_x)}$ is that the time weight of the data in $l_i$ ($t_x < t_m$), and the value of the $i$ dimension wastewater quality variables.

4. Experiment

In the experiment, we use the wastewater quality variables PH of a wastewater treatment plant. We use the second-order regression function as the train model. When the decision time $t_p$ is $t_{12}$, we select the data of the $(t_x, t_{11})$ period as the important and alive data, and the data of $(t_{11}, t_1)$ period as the cultivation data, the data of before time $t_1$ as the discard data. We set the value of the wastewater quality variables as zero, time weight $\lambda = 0.001$. When $t_{12}$ is 20 and $x_{12}$ is 11.4, we get the result $y_{12} = 3.1794$. We reduce the quantity of the chemical dosing from 6 to 3.1794, shown in the figure1. The red asterisk represents the result by the algorithm. The blue line represents the result by old method. We can see the algorithm is very useful to the chemical dosing in the process of WWT.

**Table 1.** The value of different parameters in the experiment.

| $N_0$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|-------|---|---|---|---|---|---|---|---|---|---|----|----|
| $\Omega$ | $X$ | 9.8 | 10 | 10.2 | 10.4 | 10.7 | 11 | 11.3 | 11.6 | 11.8 | 12 | 12.5 | 13 |
| | $Y$ | 2.6 | 2.7 | 2.8 | 2.9 | 3.3 | 3.5 | 3.7 | 4 | 4.6 | 5 | 5.5 | 6 |
| $T$ | $t$ | 1 | 4 | 5 | 8 | 9 | 10 | 12 | 13 | 14 | 15 | 17 | 18 |
5. Conclusion

In this paper, we propose a self learning algorithm base on the data value lifecycle model, which time weight is dynamic and adaptive to the decision time. We establish the mechanism of the cultivation and discard data. The experiment has proved that the algorithm is useful. Next we can research other artificial intelligence methods to make the chemical treatment process more economical and better performance.

6. Acknowledgments

In this paper, the research was sponsored by the Major R & D Plan of Science and Technology of Zhejing (Project No. 2019C03099).

References

[1] Amaral, A.L., Mesquita, D.P., Ferreira, E.C.. Automatic identification of activated sludge disturbances and assessment of operational parameters. Chemosphere 91, 2011. 705-710.

[2] Avella, A.C., Gorner, T., Yvon, J., Chappe, P., Guinot-Thomas, P., de Donato, P.., 2011. A combined approach for a better understanding of wastewater treatment plants operation: statistical analysis of monitoring database and sludge physico-chemical characterization. Water Res. 45(2011), 981-992.

[3] Åmand, L., Olsson, G., Carlsson, B. Aeration control - a review. Water Sci. Technol. 67(2013), 2374-2398.

[4] Cecil, D., Kozlowska, M. Software sensors are a real alternative to true sensors. Environ. Model. Softw. 25(2010), 622-625.

[5] Di Trapani, D., Christensso, M., Ødegaard, H. Hybrid activated sludge/biofilm process for the treatment of municipal wastewater in a cold climate region: a case study. Water Sci. Technol. 63(2011), 1121-1129.

[6] Gutierrez, G., Ricardez-Sandoval, L.A., Budman, H., Prada, C. An MPC-based control structure selection approach for simultaneous process and control design. Comput. Chem. Eng. 70(2014), 11-21.
[7] Haimi, H., Corona, F., Mulas, M., Sundell, L., Heinonen, M., Vahala, R. Shall we use hardware sensor measurements or soft-sensor estimates? Case study in a full-scale WWTP. Environ. Model. Softw. 72(2015), 215-229.

[8] Huang, M., Ma, Y., Wan, J., Modeling a paper-making wastewater treatment process by means of an adaptive network-based Fuzzy inference system and principal component analysis. Ind. Eng. Chem. Res. 51(2012), 6166-6174.

[9] Jing, L., Chen, B., Zhang, B., Li, P. Process simulation and dynamic control for marine oily wastewater treatment using UV irradiation. Water Res. 81(2015), 101-112.

[10] Kim, H., Lim, H., Wie, J., Lee, I., Colosimo, M.F, Optimization of modified ABA2 process using linearized ASM2 for saving aeration energy. Chem. Eng. J. 251(2014), 337-342.

[11] Liu, Y., Pan, Y., Sun, Z., Huang, D. Statistical monitoring of wastewater treatment plants using variational Bayesian PCA. Ind. Eng. Chem. Res. 53(2014), 3272-3282.

[12] Lou, I., Zhao, Y. Sludge bulking prediction using principle component regression and artificial neural network. Math. Probl. Eng. (2012) http://dx.doi.org/10.1155/2012/237693.

[13] Mannina, G., Trapani, D., Di Viviani, G., Ødegaard, H., 2011. Modelling and dynamic simulation of hybrid moving bed biofilm reactors: model concepts and application to a pilot plant. Biochem. Eng. J. 56(2011), 23-36.

[14] Martín de la Vega, P.T., Martinez de Salazar, E., Jaramillo, M., a, Cros, J.. New contributions to the ORP & DO time profile characterization to improve biological nutrient removal. Bioresour. Technol. 114(2012), 160-167.

[15] Nasr, M.S., Moustafa, M. a E., Seif, H. a E., El Kobrosy, G. Application of artificial neural network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-Egypt. Alex. Eng. J. 51(2012), 37-43.

[16] O'Brien, M., Mack, J., Lennox, B., Lovett, D., Wall, A. Model predictive control of an activated sludge process: a case study. Control Eng. Pract. 19(2011), 54-61.

[17] Olsson, G., Carlsson, B., Comas, J., Copp, J., Gernaey, K.V., Ingildsen, P., Jeppsson, U., Kim, C., Rieger, L., Rodriguez-Roda, I., Steyer, J.-P., Takacs, I., Vanrolleghem, P. a, Vargas, a, Yuan, Z., Åmånd, L. Instrumentation, control and automation in wastewater from London 1973 to Narbonne 2013. Water Sci. Technol. 69(2014),1373e1385.

[18] Plattes, M., Henry, E., Schosseler, P.M., Weidenhaupt, A., Modelling and dynamic simulation of a moving bed bioreactor for the treatment of municipal wastewater. Biochem. Eng. J. 2006.32, 61-68.

[19] Vega, P., Revollar, S., Francisco, M., Martín, J.M.. Integration of set point optimization techniques into nonlinear MPC for improving the operation of WWTPs. Comput. Chem. Eng. 2014.68, 78-95.

[20] Xiaodong Wang, Harsha Ratnaweera, Johan Abdullah Holm, Vibeke Olsbu. Statistical monitoring and dynamic simulation of a wastewater treatment plant: A combined approach to achieve model predictive control. Journal of Environmental Management 193 (2017) 1-7.

[21] W. Yan, P. Guo, Y. Tian, J. Gao, A framework and modeling method of data-driven soft sensors based on semisupervised Gaussian regression, Ind. Eng. Chem. Res. 55 (2016) 7394–7401.

[22] J. Shiri, A. Keshavarzi, O. Kisi, S. Karimi, Using soil easily measured parameters for estimating soil water capacity: soft computing approaches, Comput. Electron. Agric. 141 (2017) 327–339, http://dx.doi.org/10.1016/j.compag.2017.08.012.

[23] H. Seshan, M.K. Goyal, M.W. Falk, S. Wuertz, Support vector regression model of wastewater bioreactor performance using microbial community diversity indices: Effect of stress and bioaugmentation, Water Res. 53 (2014) 282–296.

[24] S. Akilandeswari, B. Kavitha, Comparison of ANFIS and statistical modeling for estimation of chemical oxygen demand parameter in textile effluent, DerChem. Sin. 4 (2013) 96–99.

[25] M. Ay, O. Kisi, Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques, J. Hydrol. 511 (2014) 279–289.