I. Troch, F. Breitenecker, eds. ISBN 978-3-901608-35-3

PREDICTING COD CONCENTRATION OF ACTIVATED SLUDGE PLANT EFFLUENT USING NEURAL NETWORKS AND GENETIC ALGORITHMS

J. Keskitalo[1], A. Sorsa[1], T. Heikkinen[2], E. Juuso[1]
[1]University of Oulu, Oulu, Finland; [2]UPM-Kymmene corporation, Pietarsaari, Finland

Corresponding Author: J. Keskitalo, University of Oulu, Control Engineering Laboratory
P.O. Box 4300, 90014, Oulu, Finland; jukka.keskitalo@oulu.fi

Abstract. Improving the control of activated sludge process is necessary in reducing discharges to the environment. Process models can be used for predicting the discharges to help plan corrective control actions. A data driven modelling method utilising process data from pulp and paper mill database for predicting the mill activated sludge plant effluent chemical oxygen demand (COD) concentration is presented in this paper. Self-organising maps (SOM) and K-means clustering are used to cluster the process data according to different operating regions of the process. Multilayer perceptron (MLP) artificial neural network models are trained and validated with the data from each cluster. The choice of input variables for each submodel is optimised by genetic algorithm. The results show that the presented method is capable of constructing process models that can predict the effluent COD concentration two days ahead.

1 Introduction

Biological wastewater treatment in an activated sludge process is the most common way of treating pulp and paper mill effluents in Finland. With the annual production of 7 million tons of chemical pulp and 14 million tons of paper and the water consumption of 20-50 m^3 per ton on pulp and 7-15 m^3 per ton of paper produced, proper control of activated sludge processes treating pulp and paper mill wastewater is of great importance. The treatment of forest industry wastewaters has improved a lot since 1970's in Finland while the total production of pulp and paper has increased. Nowadays effluent discharges under normal operation of biological wastewater treatment are not so much an issue as they used to be. However, the microbial population in biological wastewater treatment is sensitive to unusual discharges from the mill and changes in operating conditions. Incidental discharges in the wastewater treatment effluent have become a significant proportion of the total amount of discharges. [14] Due to these reasons, there is a need for better methods of predicting discharges from the treatment plants.

Modelling the activated sludge process and using the model for predicting discharges is a possible method for achieving better control of the process. However, modelling of the activated sludge process is a very challenging task because the reactions of the microorganism are very nonlinear and time varying, and there are variations in flow rates and composition of the incoming wastewater. [11] Both mechanistic white-box and data based black-box modelling have been used with the activated sludge process. The Activated Sludge Model (ASM) No. 1 developed by the International Water Association can be considered as the reference model in white-box modelling. The capabilities of the ASM model family have been extended to also describe biological nitrogen and phosphorus removal with the ASM2 and ASM3 models. However, the ASM models were developed for modelling activated sludge plants treating municipal wastewater and therefore may not be directly applicable to industrial wastewater treatment plants. [3]

Studies using either data from the mill databases or from measurement campaigns for black-box modelling of the activated sludge process have been reported in the literature. In [12] chemical oxygen demand (COD) concentration of a municipal wastewater treatment plant effluent was modelled using multilayer perceptron (MLP) artificial neural network. The architecture and the choice of input variables of the network were optimised with a MATLAB® script which went through different combinations of parameters and calculated the correlation coefficient between the modelled and measured effluent COD. The models were only used for calculating the current value of effluent COD based on other available measurements from the influent and the activated sludge plant and therefore no attempt was made to predict COD discharges. The best MLP model was able to calculate the effluent COD with reasonable accuracy. In [5] MLP models were used for predicting the effluent COD concentration of an activated sludge plant treating pulp and paper mill wastewater. 22 variables from the mill databases were chosen by a process expert and used as the input variables of the MLP model. The model was trained to predict the effluent COD 1-5 days ahead. The prediction of the MLP model was good for one day and moderate for two days ahead. Recently there have also been studies about combining mechanistic and data based approaches into hybrid model structures. For example, in [10] and [11] hybrid model consisting of a modified ASM1 model and artificial neural network has been applied for modelling a full-scale cokes wastewater treatment plant.

Reactions of the microorganisms are time varying and therefore correlations between the model input and output variables will also vary with time. Time varying behaviour of the activated sludge process data was studied in [6] using self-organising maps (SOM). Four years of process data were used as an input in the training of a SOM. The SOM was then clustered using K-means algorithm according to the reference vectors. It was discovered that there is a tremendous variation in the behaviour of the process in different clusters of the process data. Therefore it would be very difficult to train one black-box model to take into account the time varying behaviour of the activated sludge process over a long time period such as several years. In [6] it was also found that some of the variation is seasonal, making it difficult to train one black-box model even with one year of process data.

The approach in the study reported in this paper is to use routinely measured data from the mill databases for black-box modelling of the activated sludge process. COD concentration in the effluent of an activated sludge plant treating pulp and paper mill wastewater is predicted two days ahead using the measured data as model inputs. Due to the time varying behaviour of the process, the process data is first clustered using SOM and K-means in order to separate data from different operating regions of the process into their own clusters. Submodels are trained with the data of each cluster. MLP artificial neural networks are chosen as the model structure for the submodels due to the ability of MLP to describe nonlinear processes. In the available process data there are over hundred variables measured from the influent water, the process itself and the effluent water. It is very difficult to justify the choice of input variables for each submodel if the choice is made manually. Therefore automatic variable selection using genetic algorithms (GA) is implemented. Similar variable selection algorithm has been used in [13] for feature selection from Barkhausen noise data with good results.

2 Description of the process and data

The process data for the study reported in this paper comes from the activated sludge plant of UPM-Kymmene corporation pulp mill in Pietarsaari, Finland. The pulp mill produces 800 000 tons of pulp annually and the activated sludge plant treats an average of 100 000 m^3 wastewater per day. The treatment plant is an ordinary activated sludge plant consisting of primary sedimentation, aeration and secondary sedimentation. Primary sedimentation is used to remove easily settling solids before aeration. In aeration microorganism consume biodegradable organic material for producing energy and cell growth. Activated sludge from the aeration is settled in a secondary clarifier. Effluent of the secondary clarifier is discharged to the sea and settled sludge is returned to the aeration basin to maintain sufficient solids concentration. Simplified diagram of the process is presented in Figure 1.

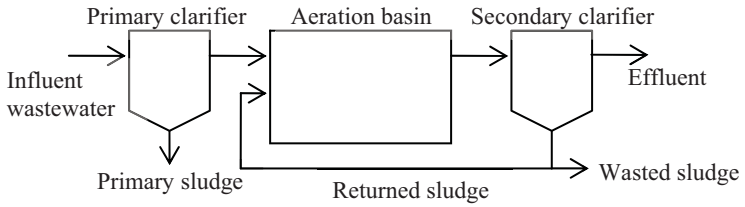

Figure 1. Simplified diagram of the activated sludge process.

Four years of activated sludge plant process data including online and laboratory measurements was extracted from the mill databases. For online measurements the extracted values were daily averages. Laboratory analyses included in the data were usually made once a day excluding weekends. The most recent one year of data was used for the purposes of this study. The data contained some stoppages of production which were marked as missing data. Rows of data which were clearly out of the ordinary range due to measurement errors were also marked as missing. The data also contained missing values due to the infrequency of some laboratory analyses and breakdowns of measurement instruments. A simple linear interpolation algorithm was used for filling the gaps in data. In linear interpolation a straight line is fitted between the values before and after the gap and missing values are calculated using the line equation [8].

Variables which could have an effect on the effluent COD concentration were pre-selected by a process expert in an earlier study [5]. The same selection of variables with some minor changes was used in this study as the basis of further variable selection using GA. Pre-selected variable set is presented in Table 2.

Because process data is used to predict effluent COD concentration two days ahead, the output variable in data was shifted two days forward. Finally, the data was filtered to smooth out the noise in the measurements with exponentially weighted moving average filter with a damping constant value of 0.5.

3 Methodology

3.1 Self-organising map

Self-organising maps (SOM) are a class of artificial neural networks based on competitive learning. Neurons compete with each other and only one winning neuron can be activated at a time. In SOM the neurons are placed in one- or two-dimensional array. During the competitive learning process the locations of the neurons in the array become arranged to represent intrinsic statistical features of the input data. [4]

The SOM defines a mapping from n-dimensional input data into one- or two-dimensional array of neurons. Each neuron is associated with an n-dimensional reference vector $m_i=[\mu_{i1}, \mu_{i2},\ldots,\mu_{in}]^T$. During learning each n-dimensional data vector $x = [\varepsilon_1, \varepsilon_1,\ldots, \varepsilon_n]^T$ is compared with all the reference vectors in some metric such as Euclidean distance. The parameters of the reference vector of the winning neuron and its neighbourhood are updated to match the data vector better. [9]

3.2 K-means

K-means is a partitional clustering algorithm where the data is clustered into a predefined number of *k* clusters. The algorithm begins with picking randomly *k* cluster centres. Each point in data is assigned to its closest cluster centre. Cluster centres are recalculated using the current cluster memberships. If a convergence criterion of no or minimal reassignment of patterns to new cluster centres or minimal decrease in a sum of squared errors of data point and cluster centre locations is not met, then cluster centres are recalculated iteratively. [7]

3.3 Multilayer perceptron

Multilayer perceptrons (MLP) belong to the feedforward class of artificial neural networks. MLP consists of different layers of computation nodes or neurons: the input layer, one or more hidden layers and an output layer. The input signal propagates through each layer in forward direction. An example of the connections of the neurons is presented in Figure 2. Each neuron calculates its output as a function of a weighted sum of its inputs. The activation function is usually sigmoidal nonlinear function. An important feature of an MLP with single hidden layer is its ability to compute a uniform approximation of an arbitrary continuous function. [4]

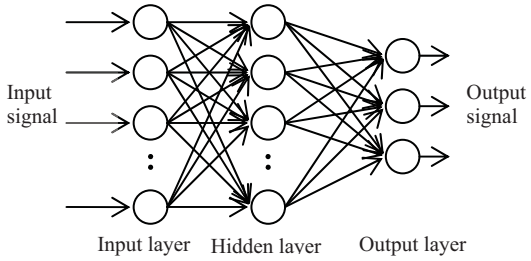

Figure 2. An example of the structure of an MLP with an input layer, one hidden layer and an output layer.

MLPs are trained by adjusting their synaptic weights for each set of data in a supervised manner with an error backpropagation algorithm. Error backpropagation consists of two passes through the network. In the forward pass, an input vector is applied to the input layer of the network. The signal propagates through the network and produces the output signal. The output signal of the network is subtracted from the desired target signal to calculate an error signal. In the backward pass, this error signal is back-propagated through the network and the synaptic weights are adjusted in order move the response of the network closer to the desired response. The purpose of training the network is to teach the network to generalise. The network generalises well when the input-output mapping of the network is also correct for test data not used in training the network. [4]

3.4 Genetic algorithms

Genetic algorithms are optimisation methods where a population of possible solutions is improved over several generations using genetic operators inspired by biological evolution. Population consists of individual solutions which are encoded into chromosomes. The most commonly used encoding is the binary alphabet but other encodings such as real valued encoding can be used. The information in the chromosomes about the variables in the solution is decoded, and the fitness of the solution is evaluated by an objective function. The chromosome without knowledge of the encoding provides no information about the optimisation problem. The optimisation process, however, operates on the encoded chromosomes using genetic operators. [1]

Individuals are chosen for reproduction with a probability related to their fitness values. Recombination operators are used to combine genetic information from the selected parents to produce offspring. Mutation operator is applied to the new population in order to guard against losing good genetic material in the selection and crossover operations. The new population is then evaluated using the objective function. The GA is usually terminated after predefined number of generations. [1]

4 Combined algorithm

Artificial neural networks, K-means clustering and genetic algorithms have been combined in pre-processing, tuning and modelling.

4.1 SOM and clustering

The pre-processed data was clustered with SOM and K-means in order to separate data from different operating regions of the process. Three important variables describing operating region of the process were chosen as the inputs to the SOM: sludge load, DSVI and aeration basin temperature. SOM with 25 x 25 neurons was constructed and trained with batch training algorithm for 50 epochs. The reference vectors of the SOM were then clustered into six clusters with K-means algorithm. The training data were assigned to the clusters by finding the best matching neuron for each row of the data. The states of the process associated with each cluster are presented in Table 1.

Cluster	Aeration basin temperature	DSVI	Sludge load
1	4	2-3	1
2	4-5	4-5	3-4
3	1	2	2
4	5	1	3-5
5	3-4	1	3-4
6	3-4	1-2	1-2

Table 1. The states of the process associated with each cluster. Numbers indicate the levels of variables: 1 – low, 2 – slightly low, 3 – medium, 4 – slightly high and 5 – high.

4.2 Genetic algorithm

The choice of input variables for submodel of each cluster was optimised with GA. The information was binary coded: the variable is selected if the bit is 1 and not selected if the bit is 0. There were 23 possible input variables for each of the six submodels making the size of the chromosome 138 bits. The initial population of 200 chromosomes was generated with the probability of a bit being 1 was 0.5. The parents for the recombination operator were chosen with tournament selection. In tournament selection five candidates are randomly chosen for each tournament, and the candidate with the lowest value of the objective function to be minimised is chosen as a parent. If a randomly generated number is higher than a crossing probability of 0.9, the chosen parents are then used to create offspring by single-point recombination with a randomly chosen splitting point. Otherwise the parents are directly added to the new population. The parent selection and recombination is carried on until a new population with the size of 200 chromosomes has been created. Each bit of every population member is subjected to mutation with a probability of 0.02. If mutation happens for a certain bit, the value of the bit is changed. Finally, elitism is applied by replacing the worst chromosome of the new population with the best chromosome of the previous population.

The objective function used in evaluating the fitness of solutions utilises correlation coefficients between the effluent COD predicted by MLP submodels and measured values. Correlation coefficient isn't a perfect measure of prediction accuracy. Very good correlation coefficient may result from actually a very lousy prediction. Therefore correlation coefficients between the outputs of multiple linear regression (MLR) models and measured values were also utilised in the objective function. The reason for utilising the correlation coefficients of the MLR models is that if some choice of input variables results in a good prediction with two different model structures, there is a higher chance that the good correlation coefficient is actually due to good prediction. To avoid including insignificant input variables in the model, a penalty term calculated from the number of input variables is added to the objective function. The objective function used in this study is

$$J = -\left(\frac{\sum_{j=1}^{n} \left(R_{j\,MLP} - 0.002 \cdot m_j{}^2 \right)}{n} + \frac{\sum_{j=1}^{n} \left(R_{j\,LR} - 0.002 \cdot m_j{}^2 \right)}{n} \right) \tag{1}$$

where R_{jMLP} is the correlation coefficient between the effluent COD values predicted by the MLP submodel and the measured values of the validation data of the jth cluster, R_{jLR} is the correlation coefficient between the effluent COD values predicted by the linear regression model and the measured values of the validation data of the jth cluster, m_j is the number of input variables in the submodel of the jth cluster and n is the number of clusters.

4.3 MLP submodels

MLP submodels predicting effluent COD two days ahead were created and trained using input variables decoded from the chromosomes. Data from each cluster was split into training and validation data. Three fifths of the data was used for training the MLP models and two fifths for stopping the training and validating the models. Each submodel has one hidden layer with seven neurons. Neurons in the hidden layer have hyperbolic tangent sigmoid transfer functions and the output layer neuron has linear transfer function. Determining the required number of hidden layer neurons for submodels is difficult because the clusters have different lengths of training data. Therefore the networks were trained with Levenberg-Marquardt algorithm with Bayesian regularisation. The regularisation prevents overfitting the data by making sure that the effective number of parameters in the network remains the same regardless of the total number of parameters [2].

5 Results and discussion

The optimisation of the choice of input variables with the algorithm described in Section 4 was run ten times to avoid the effect of randomly generated initial population. In each run of the algorithm, the number of generations in the genetic algorithm was 50. The best set of submodels from these results was chosen by looking at the measured effluent COD and the outputs of the submodels, because the best objective function value doesn't necessarily correspond with the best model. Measured and two days ahead prediction of effluent COD are presented in Figure 3. Data in Figure 3 is grouped by clusters and it is not presented in chronological order.

From the Figure 3 it can be seen that the predicted effluent COD follows the measured values quite closely. The exact values are not always predicted perfectly, but the submodels are able to predict the direction of change. Clusters represent different operating regions of the process as can be seen in Table 1. Clusters 4-6 represent the normal operation of the process with low DSVI values. These clusters contain most of the process data. Even though the DSVI stays at low constant value, there is significant variation in the effluent COD due to changes in the influent COD concentration, biodegradability of the organic load and the microbial population. The submodels are apparently able to take into account the effect of these changes to predict the resulting change in effluent COD concentration.

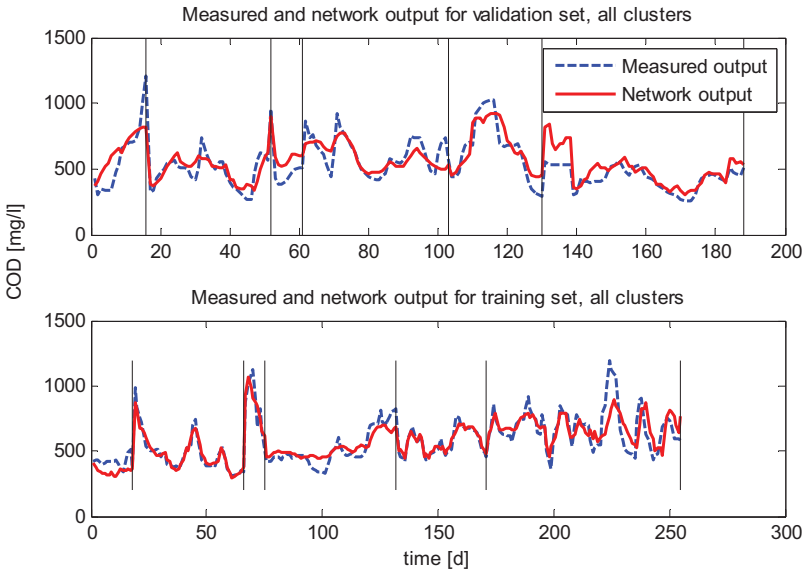

Figure 3. Measured effluent COD and the outputs of the submodels for each cluster. Validation data is in the upper picture and network training data in the lower picture. Black vertical lines indicate cluster breaks. Data before first vertical line belongs to cluster one, data before second vertical line belongs to cluster two and so on.

The choice of input variables for submodel of each cluster is presented in Table 2. The number of input variables in submodels is quite high, indicating that the penalty term calculated from the number of input variables didn't have enough effect and there are still redundant input variables left. The previous value of the predicted variable, effluent COD concentration wasn't included as an input variable in any of the submodels even though it has good correlation with the predicted variable. Therefore other input variables provided enough information of the future effluent COD, and the previous value would have been a redundant variable. Sludge indices and nutrient concentrations in the effluent were included in many of the submodels. Sludge indices provide information of the state of the biomass and effluent nutrient concentrations provide information of the excess soluble nutrients which are required for biomass growth and reducing the organic load in the wastewater.

Variable \ Cluster	1	2	3	4	5	6
COD concentration, 3						
Flow rate, 4			x		x	
Flow rate, 3						
Solids concentration, 1		x		x		
Solids concentration, 2	x				x	
Solids concentration, 3	x				x	
Solids concentration, 4	x	x				
COD concentration, 1		x				x
pH, 1					x	x
pH, 4						x
Conductivity, 1			x			
Total nitrogen concentration, 1						
Total nitrogen concentration, 3	x	x		x		
Total phosphorus concentration, 1		x			x	
Total phosphorus concentration, 3		x	x	x		x
Dissolved oxygen, 2		x			x	
Temperature, 1			x			
Temperature, 2					x	x
Settling of sludge, 2			x			
DSVI, 2	x	x	x		x	
SVI, 2	x	x		x		
Sludge retention time	x	x		x		
Sludge load	x				x	

Table 2. Pre-selected set of variables and their use as input variables of the submodels. Numbers after variable name indicate the measurement location. 1 - before aeration, 2 - after aeration, 3 - after secondary clarification and 4 - from return sludge.

6 Conclusions

Reducing incidental discharges from activated sludge plants treating pulp and paper mill wastewater is important in order to reduce the total amount of discharges to the environment. Modelling the activated sludge process and using the model to improve process control is a possible method for reducing discharges. The results presented in this paper show that using SOM and K-means for finding different operating regions of the process and training MLP submodels for the operating regions with input variable selection by genetic algorithms is an efficient way to predict COD concentration in the activated sludge plant effluent.

7 Acknowledgements

This research was supported by the Finnish Funding Agency for Technology and Innovation (Tekes), Kemira Oyj, Stora-Enso Oyj and UPM-Kymmene Oyj.

8 References

[1] Chipperfield, A.: *Introduction to genetic algorithms.* In: Genetic algorithms in engineering systems, (Eds. Zalzala, A. M. S. and Fleming, P. J.), IEE, London, 1997, 1 – 45.

[2] Demuth, H., Beale, M. and Hagan, M.: *MATLAB® Neural Network Toolbox™ 6, User's Guide.* The Mathworks, 2008.

[3] Gernaey, K. V., Loosdrecht, M. C. M., Henze, M., Lind, M. and Jørgensen, B.: *Activated sludge wastewater treatment plant modelling and simulation: state of the art.* Environmental Modelling & Software, 19 (2004), 763 – 783.

[4] Haykin, S.: *Neural networks: a comprehensive foundation (2nd edition)*. Prentice Hall, New Jersey, 1999.

[5] Heikkinen, M., Heikkinen, T., and Hiltunen, Y.: *Modelling of activated sludge treatment process in a pulp mill using neural networks.* The 6th International Conference on Computing, Communications and Control Technologies, 2008, Orlando, USA.

[6] Heikkinen, M., Heikkinen, T. and Hiltunen, Y.: Process states and their submodels using self-organizing maps in an activated sludge treatment plant. The 49th SIMS Conference on Simulation and Modelling, 2008, Oslo, Norway.

[7] Jain, A. K., Murty, M. N. and Flynn, P. J.: Data clustering: a review. ACM Computing Surveys, 31, (1999), 264 – 323.

[8] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M.: *Methods for imputation of missing values in air quality data sets*. Athmospheric Environment, 38, (2004), 2895 – 2907.

[9] Kohonen, T.: *Self-organizing maps (2nd edition).* Springer-Verlag, Berlin Heidenberg, 1997.

[10] Lee, D. S., Vanrolleghem, P. A. and Park, J. M.: *Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant.* Journal of Biotechnology, 115, (2005), 317 - 328.

[11] Lee, D. S., Jeon, C. O., Park, J. M. and Chang, K. S.: *Hybrid neural network modeling of a full-scale industrial wastewater treatment process.* Biotechnology and Bioengineering, 78, (2002), 670 – 682.

[12] Moral, H., Aksoy, A. and Gokcay, C. F.: *Modeling of the activated sludge process by using artificial neural networks with automated architecture screening.* Computers and Chemical Engineering, 32, (2008), 2471 – 2478.

[13] Sorsa, A. and Leiviskä, K.: *Feature selection from Barkhausen noise data using genetic algorithms with cross-validation.* To be published in International Conference on Adaptive and Natural Computing Algorithms, 2009, Kuopio, Finland.

[14] Ukkonen, M.: *Metsäteollisuuden jätevesien häiriöpäästöt ja niihin varautuminen*. Publications of the Southeast Finland Regional Environment Centre, 2005.