Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2018) 26: 3356 – 3366

doi:10.3906/elk-1503-206

Artificial immune system based wastewater parameter estimation

Cengiz SERTKAYA*, Nilüfer YURTAY
Department of Computer Engineering, Faculty of Computer and Information Sciences, Sakarya University, Sakarya, Turkey

Received: 23.03.2015 • **Accepted/Published Online:** 17.12.2015 • **Final Version:** 29.11.2018

Abstract: The basis of a wastewater treatment system is to achieve the desired characteristics of the wastewater treatment process. An estimation of the obtained wastewater treatment characteristics provides the information needed to set up the current process steps, and it is important to have an optimum treatment. In this study, an artificial immune system (AIS) structure is developed to estimate important wastewater output parameters such as pH, DBO, DQO, and SS for the first time. The proposed AIS models are based on the clonal selection principle, and the dataset is provided from the University of California Irvine (UCI) Machine Learning Library. The current dataset is analyzed by principal component analysis (PCA) to obtain maximum system performance. As a result of the simulation, the output parameters are successfully predicted using the AIS model with real data. The classifier's performance ratios are studied separately using the coefficient of determination (R^2) and the mean squared error of prediction (MSEP), and their rates are given in this study.

Key words: Wastewater, principal component analysis, artificial immune system, coefficient of determination, mean squared error, classification

1. Introduction

Water is a basic need of life. It is required for the realization of many vital activities such as economic development, environmental health, food production, and energy production. However, the rapid growth of the human population, increase in environmental pollution, and decrease in the number of water resources have created difficult circumstances. In order to eliminate potential water shortages throughout the world, various precautions should be taken into consideration by researchers.

Because of rapid urbanization, the consequent reduction in green cover makes it difficult to naturally treat environmental pollution [1]. The pollution rate is very high for our world [2]. While the world's population increases rapidly, the water resources in the world are finite. Therefore, these resources must be conserved.

Water sources are mainly polluted by waste. Waste products contain harmful materials that are in solid, liquid, and gaseous states and are produced as a consequence of activities performed by households, industrial enterprises, and agriculture. According to data from the Turkish Statistical Institute (TSI) in 2010, 48.6% of the 3.58 billion cubic meters of wastewater in total collected by the sewerage system in Turkey was released into streams, while the rest was released into seas (41.8%), dam lakes (3.6%), lakes/ponds (2.1%), lands (1%), and other recipient environments (2.8%) [3]. In fact, before this releasing process, the wastewater needs to be cleaned. The wastewater treatment plants, which are built for this purpose, try to minimize its effects on

*Correspondence: cengiz.sertkaya@ogr.sakarya.edu.tr

the environment. While it is done, the wastewater is processed by some physical and chemical treatments [4]. Supervision of the wastewater treatment processes is realized in reference to water purification at an optimum level. The checks are conducted by designating the characteristics of the wastewater. Then the necessary adjustments in the system at each step of the process are realized to obtain the desired purification. Finally, the examination in the plant outlet is performed by health institutions to determine whether the desired water standards have been reached or not [5]. However, water purification involves nonlinear and highly sophisticated processes. In a study, the design and implementation of a nonlinear model-predictive control (NMPC) system are discussed in relation to the wastewater treatment process [6]. Artificial intelligence-based systems are required to control these processes [7].

In the literature, many artificial intelligence system models are defined to ensure control over the wastewater treatment process. In [8], estimations of BOD, TSS, COD, flow rate, nitrogen, and phosphorus parameters were obtained with a determination coefficient of 0.919 using an artificial neural network (ANN) model. In another study, input parameters COD, SS, and temperature were used in an ANN model for the estimation of output parameters COD and SS with a coefficient of 0.98 [9]. In another study, an ANN model was developed to estimate the pH value; as a result of simulation, an error rate of 0.004% was obtained [10].

The aim of the present study was to estimate the characteristic features PH-S, DBO-S, DQO-S, and SS-S of cleaned water using AIS. This paper is organized as follows. Section 2 describes the data-processing steps. Section 3 briefly introduces the concept of an AIS system and presents the proposed methodology. The simulation results are presented in Section 4. Finally, Section 5 presents the conclusion.

2. Data source and its analysis

The process of the developed model has a number of steps. In the first step, the data are provided for the system. In the second step, the data are preprocessed and analyzed. Then proposed AIS models are developed. In the last step, the developed AIS models are simulated and the obtained results are discussed. The steps of the developed model are schematically shown in Figure 1.

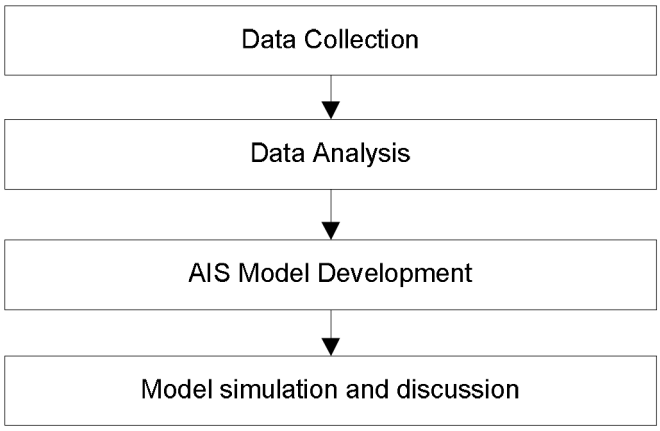

Figure 1. Process steps of the developed model.

The dataset used in this study is obtained from the University of California Irvine (UCI) Machine Learning Library. The daily readings of a municipal wastewater treatment plant make up this dataset. In this study, the input and output features of this plant are used [11].

There are 527 daily data points in the dataset. In addition, 13 features (9 inputs and 4 outputs) are selected. The UCI dataset features used are also shown in Table 1.

Table 1. Water treatment plant dataset attributes.

No.	Parameter	Description	Min	Max	Mean
1	Q-E	input_flow to plant	10050	60081	37168.87
2	ZN-E	input_zinc to plant	0.1	19.1	2.21
3	PH-E	input_pH to plant	7.3	8.7	7.82
4	DBO-E	input_biological demand of oxygen to plant	31	438	189.82
5	DQO-E	input_chemical demand of oxygen to plant	81	941	405.35
6	SS-E	input_suspended solids to plant	98	1228	225.93
7	SSV-E	input_volatile suspended solids to plant	13.2	84.8	60.90
8	SED-E	input_sediments to plant	0.4	36	4.64
9	COND-E	input_conductivity to plant	651	3230	1459.43
10	PH-S	output_pH	7.1	8.1	7.72
11	DBO-S	output_biological demand of oxygen	3	320	19.58
12	DQO-S	output_chemical demand of oxygen	9	350	84.68
13	SS-S	output_suspended solids	6	238	21.56

If the dataset is examined, it is seen that there are some missing values for some days. These days are removed from the current dataset. Following this, the study is continued with the remaining 417 data points.

After the data-cleaning process, principal component analysis (PCA) is performed for each parameter. PCA analysis is known as a variable-reduction procedure [12]. The aim of this process is to select input parameters that have a high correlation in terms of the output parameter. Therefore, this process saves the system from unnecessary input parameters and improves the system performance [13,14].

In this study, the linear correlation method for PCA analysis is used [15]. The correlation equation is given as follows:

$$Corr(X,Y) = \frac{E[(X-\mu_x)\times(Y-\mu_y)]}{\sigma_x\sigma_y}, \tag{1}$$

where E is the expected value of operator, X and Y are two random variables having the expected values μ_x and μ_y, and the terms σ_x and σ_y are the standard deviations. Correlation is a numerical way to quantify the relationship between two variables, e.g., X and Y, and it is denoted by the symbol Corr. This value is always between –1 and 1; thus $-1 < \mathrm{Corr(X,Y)} < 1$. The variables move in the same direction when there is a positive correlation. A positive correlation means that as variable 1 increases, variable 2 increases, and, conversely, as variable 1 decreases, variable 2 decreases. A negative correlation means that as variable 1 increases variable 2 decreases and vice versa.

After the correlation process is performed, the relationships between the input and output parameters are found and given in Table 2.

The correlation results summarize the linear relationship between input and output parameters.

- The output parameter PH-S has a negative correlation with input parameters ZN-E, DQO-E, and SSV-E and has a positive correlation with others.
- The output parameter DBO-S has a negative correlation with input parameters ZN-E and SSV-E and has a positive correlation with others.

Table 2. Correlation results.

	Variables	Outputs			
		PH-S	DBO-S	DQO-S	SS-S
Inputs	Q-E	0.08	0.01	–0.06	–0.01
	ZN-E	–0.14	–0.03	0.03	–0.04
	PH-E	0.35	0.01	0.00	–0.06
	DBO-E	0.00	0.16	0.27	0.15
	DQO-E	–0.01	0.09	0.30	0.07
	SS-E	0.10	0.02	–0.01	0.02
	SSV-E	–0.11	–0.01	0.14	–0.02
	SED-E	0.02	0.03	0.06	0.00
	COND-E	0.10	0.02	0.17	–0.01

- The output parameter DQO-S has a negative correlation with input parameters QE and SS-E and has a positive correlation with others.
- The output parameter SS-S has a negative correlation with input parameters QE, ZN-E, PD-E, SSV-E, and COND-E and has a positive correlation with others.

The parameters having a correlation over 0.05 are chosen for use in the AIS structure. AIS models for each output parameter are determined in reference to the obtained results (Table 3).

Table 3. Selected input parameters which absolute correlation values are larger than 0.05.

	Variables	Outputs			
		PH-S	DBO-S	DQO-S	SS-S
Inputs	Q-E	X		X	
	ZN-E	X			
	PH-E	X			X
	DBO-E		X	X	X
	DQO-E		X	X	X
	SS-E	X			
	SSV-E	X		X	
	SED-E			X	
	COND-E	X		X	

After the examination of data, we find that data variability ranges are different for all parameters. Therefore, the data normalization process is performed between 0 and 1. The formula used for the normalization process is given as follows:

$$N(x_i) = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{2}$$

where x_i is an original value of the related attribute, $N(x_i)$ is the normalized value of the related attribute, and $x_{\max}$ and $x_{\min}$ are maximum and minimum values of the related attribute.

The required training and test dataset creation process is done by using the k-folds cross validation technique. The whole data are randomly partitioned into three equal-size subsample datasets. For each fold, one fold is used for test and the other two are used for training AIS.

3. The proposed AIS structure

The first study related to AIS was the article entitled "The immune system, adaptation and machine learning" published in 1986. The process of recognizing germs and exterminating them is known as the natural immune system in the human body. AIS realizes some natural immune system approaches such as learning and recognition capabilities in a computer environment [16].

In the literature, AIS is applied in many areas such as pattern recognition [17,18], computer security [19,20], network security [21], dynamic business programming [22], disease diagnosis [23,24], dynamic control of group elevator systems [25], and optimization of job shop scheduling problems [26], and successful results are obtained.

AIS structures have many techniques. These techniques include negative selection, positive selection, and clonal selection [27]. In this study, the developed AIS structure is based on the clonal selection principle. A flowchart of the proposed AIS estimation model is shown in Figure 2.

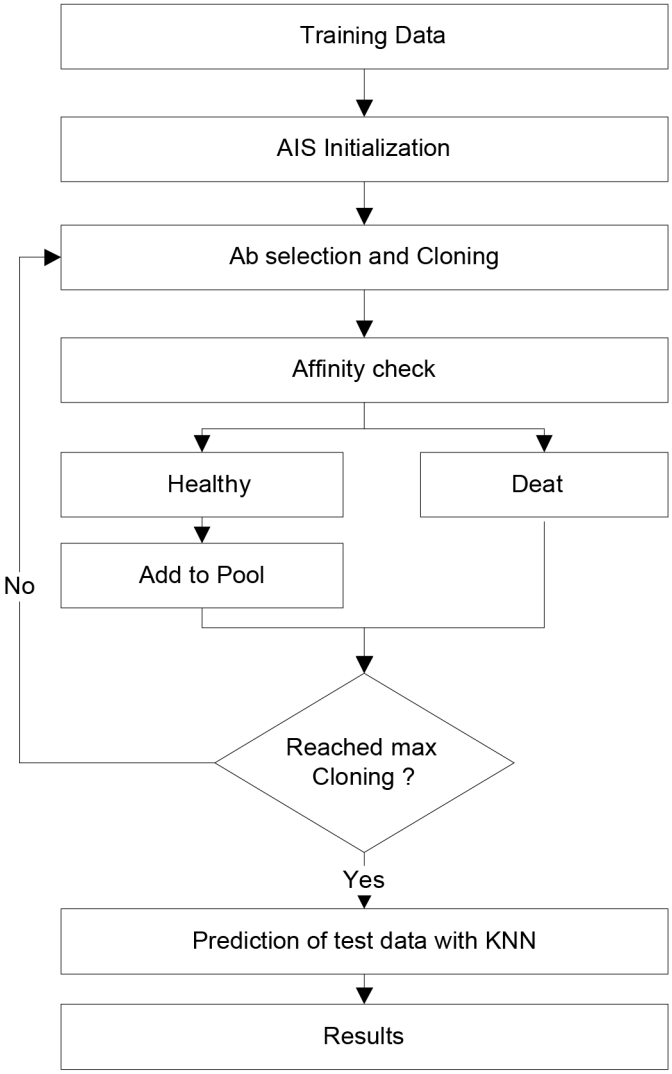

Figure 2. Flowchart of the proposed AIS estimation model.

The following processes are performed:

Step 1. The training data of 277 samples are randomly selected from the initial dataset, and are applied to the AIS structure. These data constitute the system's permanent data pool.

Step 2. The required parameters of the AIS structure are determined. These parameters are threshold, affinity (measure of being healthy), and the maximum number of clonings. In this study, the threshold value is determined to be 0.5, and the maximum number of clonings is determined to be 20. A Euclidian equation is used for affinity measure. Because of various experiments, the best results were obtained from these parameter values. Equations are given as follows:

$$\text{Euclidean} \quad (Ab, Ag) = \sqrt{\sum_{i=1}^{n} (Ab_i - Ag_i)^2} \tag{3}$$

$$Affinity(A) = 1 - Euclidean\,(Ab, Ag) = \sqrt{\sum_{i=1}^{n} (Ab_i - Ag_i)^2}, \tag{4}$$

where n is the number of attributes, Ab is antibody, and Ag is a new antigen vector. Ag can be represented by a single dimensional array of attributes.

$$Ag = \{a_1, a_2, a_3, a_4, ..., a_n\}$$

Antibody (Ab) is the representation of solution candidates. Ab can be represented by a single dimensional array of attributes.

$$Ab = \{a_1, a_2, a_3, a_4, ..., a_n\}$$

In Ag and Ab, a is attribute value and n is attribute count in one single row of data.

Step 3. Two antibodies in the same class are randomly selected for cloning. Two attributes of these two clones are randomly selected, and these attribute values are placing from one to another. This newly created sample is called an antigen. A simple cloning procedure is shown below in Figure 3.

	Attributes					
	Q-E	ZN-E	PH-E	SS-E	SSV-E	COND-E
First Sample	0.50	0.18	0.43	0.08	0.73	0.69
Second Sample	0.54	0.07	0.50	0.07	0.72	0.57
New Sample	0.54	0.18	0.50	0.07	0.72	0.69

Figure 3. Cloning process.

Step 4. The cloned sample antigen is checked for affinity. Affinity describes whether this antigen is in the same class as its cloned antibodies. If the antigen is identified as healthy, it is accepted as an antibody and is added to the permanent antibody pool. If not, it is destroyed.

Step 5. The cloning process is repeated until it reaches the maximum number of clonings. The training process of the AIS model is completed when the maximum number of clonings is reached.

Step 6. The test data are estimated by the trained AIS model by performing a k-nearest neighbors (k-NN) algorithm.

A k-NN algorithm is a type of instance-based learning algorithm and is a nonparametric method used for classification. Nonparametric refers to the fact that a k-NN algorithm does not make any assumptions on the underlying data distribution. The output of a k-NN classification is a class membership, and classification is performed by a majority vote of neighbors [28].

For a k-NN algorithm, classifying the data requires only a few parameters. These are k and the distance metric parameters. To choose the optimal value of k, the algorithm was made to run various values between 5 and 50. The best k-value was found to be 20. A k-NN algorithm makes predictions by measuring the distance between the point and its k neighbors. In the literature, one of the most popular choices to measure this distance is known as Euclidean [29]. In this study a Euclidean equation is used for the distance metric. The Euclidean equation is given in Eq. (5).

$$\text{Euclidean } (Ab, Ag) = \sqrt{\sum_{i=1}^{n} (Ab_i - Ag_i)^2} \tag{5}$$

Here n is the number of attributes, Ab is one of the nearest neighbors of Ag, and Ag is the test data from our test dataset. Ab comes from our solution candidates after the training of AIS. The closest antibody's class is chosen for determining the class of Ag.

4. Simulation results

As a result of data analysis, four AIS models have been developed for the estimation of four output parameters. The input and output parameters of these models are shown in Figure 4.

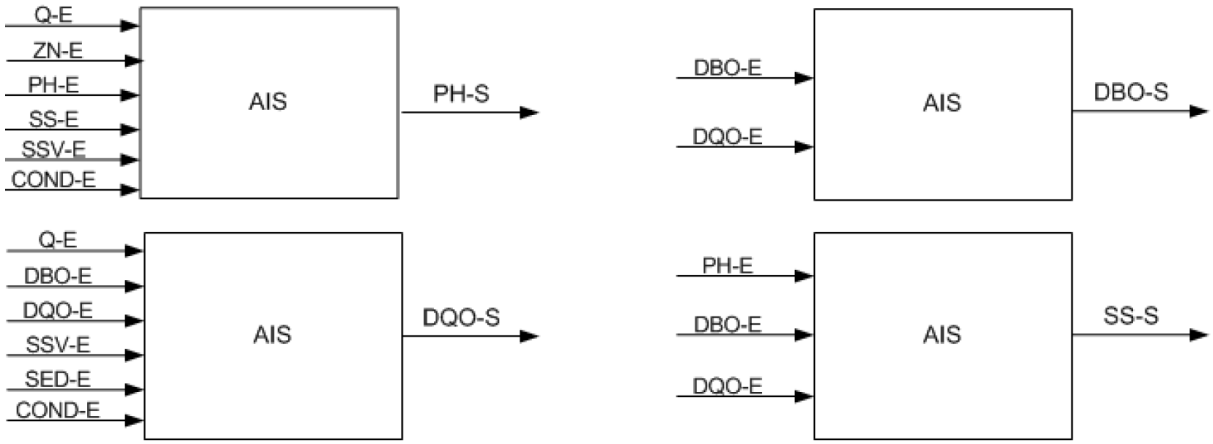

Figure 4. Input and output parameters of AIS estimation model.

The proposed AIS structure is simulated three times, and in every simulation two folds of data are used for training and the other fold is used for testing the AIS models. The generated Ab candidates in the training phases are shown in Table 4.

Table 4. Generated antibodies in training for AIS models.

Training dataset	PH-S	DBO-S	DQO-S	SS-S
Fold 1+2	206	278	248	210
Fold 1+3	211	246	219	213
Fold 2+3	200	254	210	277

The AIS models' performances are measured by the coefficient of determination (R^2) or the mean squared error of prediction (MSEP). This coefficient of determination is a measure of the accuracy of prediction of the trained AIS models. Higher R^2 values indicate better prediction. The MSEP may also be used to measure the accuracy of prediction. Lower MSEP values indicate better prediction. The R^2 and MSEP equations are given as follows:

$$R^2 = \{\frac{1}{r} \times \sum_{i=1}^{r} [(x_i - \bar{x}) \times (y_i - \bar{y})]/(\sigma_x \times \sigma_y)\}^2, \tag{6}$$

where r is the number of predictions, x_i is the real value, $\bar{x}$ is the mean x-value, y_i is the predicted value, $\bar{y}$ is the mean y-value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

$$MSEP = \frac{1}{r} \times \sum_{i=1}^{r} (x_i - y_i)^2, \tag{7}$$

where r is the number of predictions, x_i is real, and y_i is predicted result.

Table 5 shows R^2 values of the AIS models for different thresholds.

Table 5. Coefficient of determinations for AIS models.

Threshold = 0.3					Threshold = 0.4				
Folds	PH-S	DBO-S	DQO-S	SS-S	Folds	PH-S	DBO-S	DQO-S	SS-S
Fold 1	0.65	0.7	0.73	0.65	Fold 1	0.82	0.89	0.88	0.8
Fold 2	0.66	0.53	0.45	0.47	Fold 2	0.88	0.76	0.58	0.6
Fold 3	0.54	0.47	0.52	0.83	Fold 3	0.72	0.66	0.76	0.93
Average	0.62	0.57	0.57	0.65	Average	0.81	0.77	0.74	0.78
Threshold = 0.5					Threshold = 0.6				
Folds	PH-S	DBO-S	DQO-S	SS-S	Folds	PH-S	DBO-S	DQO-S	SS-S
Fold 1	0.99	0.99	0.94	0.93	Fold 1	0.95	0.97	0.92	0.92
Fold 2	1	0.92	0.78	0.73	Fold 2	0.98	0.93	0.78	0.7
Fold 3	0.81	0.83	0.94	1	Fold 3	0.84	0.8	0.9	0.96
Average	0.93	0.91	0.89	0.89	Average	0.92	0.90	0.87	0.86

A graphical representation of obtained R^2 results is given in Figure 5.

Table 6 shows MSEP values of the AIS models.

A graphical representation of obtained MSEP results are given in Figure 6.

The best result has been achieved with threshold 0.5. As seen from the best results, the proposed AIS structure successfully estimates the desired results. The second fold for the prediction of DQO-S and SS-S has the lowest prediction rates. It may be caused by an inhomogeneous splitting of the dataset or the complex nature of the wastewater parameters. These deviations can be also realized when the values in the test data take actual input values that are not present in the training data or when there is an inadequate amount of

data. The proposed method gives good estimation accuracies on average. The obtained test results prove that the developed AIS structure can learn the problem and is good at achieving the desired outcomes.

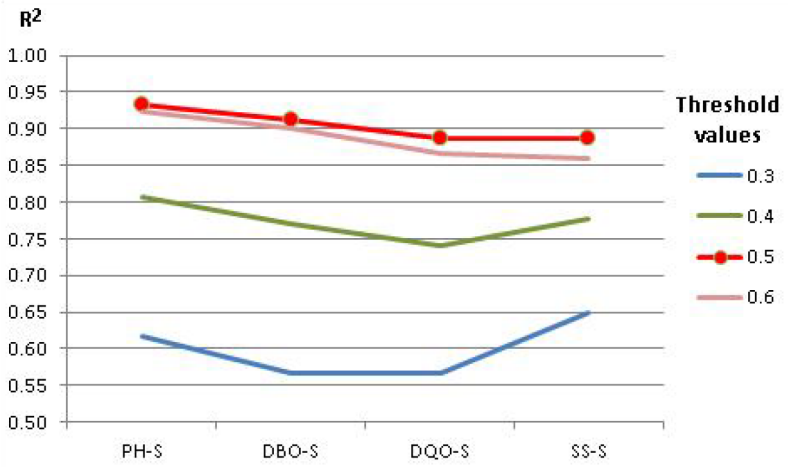

Figure 5. R2 results for different threshold values.

Table 6. Mean squared errors for AIS models.

Threshold = 0.3					Threshold = 0.4				
Folds	PH-S	DBO-S	DQO-S	SS-S	Folds	PH-S	DBO-S	DQO-S	SS-S
Fold 1	0.32	0.28	0.29	0.34	Fold 1	0.16	0.12	0.13	0.21
Fold 2	0.34	0.43	0.47	0.49	Fold 2	0.15	0.21	0.38	0.35
Fold 3	0.4	0.43	0.41	0.24	Fold 3	0.23	0.35	0.22	0.09
Average	0.35	0.38	0.39	0.36	Average	0.18	0.23	0.24	0.22
Threshold = 0.5					Threshold = 0.6				
Folds	PH-S	DBO-S	DQO-S	SS-S	Folds	PH-S	DBO-S	DQO-S	SS-S
Fold 1	0.01	0.01	0.05	0.10	Fold 1	0.07	0.01	0.06	0.09
Fold 2	0	0.06	0.22	0.25	Fold 2	0.01	0.06	0.20	0.29
Fold 3	0.14	0.12	0.04	0	Fold 3	0.11	0.17	0.08	0.04
Average	0.05	0.06	0.10	0.12	Average	0.06	0.08	0.11	0.14

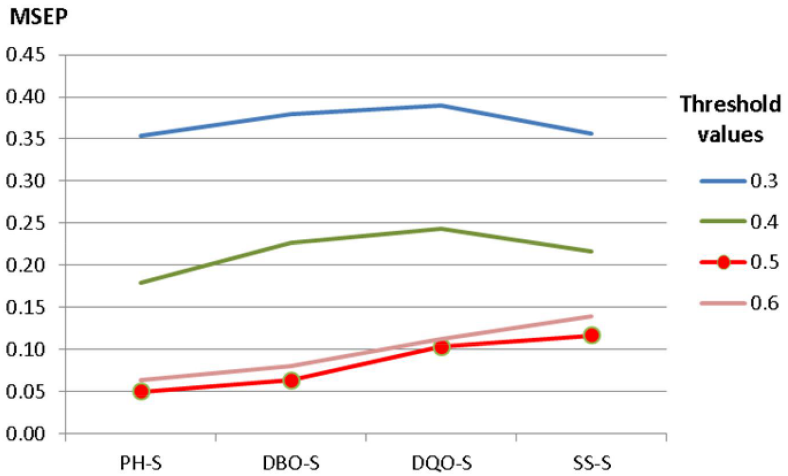

Figure 6. MSEP results for different threshold values.

5. Discussion and conclusions

In this study, a new AIS-based model is developed for wastewater treatment plant parameter estimation for the first time. The stages of the developed model are described in detail. The proposed structure has many advantages. First, for nonlinear situations, such as in wastewater parameter estimation problem, an AIS structure is expected. Second, this approach can be adapted and used in real-time wastewater plant management. In this way, this system can provide decision support for administrators. Finally, this system can be directly adapted to the wastewater plant's mechanical system to manage it automatically. A drawback of the proposed system is the data preprocessing requirement at the beginning of the process.

In future studies, an attribute-weighting procedure can be applied to the proposed AIS structure. In this manner, the performance of these algorithms can be improved.

Acknowledgement

This work was funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) (Project Number = 2233).

References

[1] Raha D. Sustainable Management of Wastewater Treatment Plant Using ISO 14001:2004 EMS and Neural Network Modeling. Indian Chem Engr 2005; 47: 182-188.

[2] IUCN/UNEP/WWF. Caring for the earth a strategy for sustainable living. Gland Switzerland, 1991.

[3] Sertkaya C. Atıksu arıtma tesisi kontrolünde yapay bağışıklık sisteminin kullanılması, PhD, Sakarya University, Sakarya, Turkey, 2016 (in Turkish).

[4] Tchobanoglous G, Burton FL, Stensel HD. Wastewater Engineering: Treatment and Reuse. 4th ed. New York, NY, USA: McGraw-Hill, 2002.

[5] Samsunlu A. Atık Suların Arıtılması. İstanbul, Turkey: Birsen Yayınevi, 2011 (in Turkish).

[6] Han H, Qiao J. Nonlinear model-predictive control for industrial processes: an application to wastewater treatment process. IEEE T Ind Electron 2014; 61: 1970-1982.

[7] Liang J, Luo F, Yu R, Xu Y. Wastewater effluent prediction based on fuzzy-rough sets RBF neural networks. In: 2010 Networking, Sensing and Control (ICNSC) Conference; 11–13 April 2010; Chicago USA. pp. 393-397.

[8] Yılmaz EC, Doğan E. Modelling of wastewater treatment plant performance using adaptive neuro fuzzy inference systems. Elec Lett Sci Eng 2008; 4: 1-9.

[9] Subaşı H. Modeling wastewater treatment performance using artificial neural networks. MSc, Adana, Turkey, 2010.

[10] Meenakhsipriya B, Saravanan K, Sathiyavathi S. Neural based pH system in effluent treatment process. Mod Appl Sci 2009; 3: 166.

[11] Riano D. Learning rules within the framework of environmental sciences. ECAI 98-W7 (BESAI98) 1998; 151-165.

[12] MacGregor JF, Kourti T. Statistical process control of multivariable processes. Control Eng Pract 1995; 3: 403-414.

[13] Oliveira-Esquerre KP, Mori M, Bruns RE. Simulation of an industrial wastewater treatment using artificial neural networks and principal components analysis. Braz J Chem Eng 2002; 19: 365-370.

[14] Civelekoğlu G. The modeling of treatment processes with artificial intelligence and multistatistical methods. PhD, Süleyman Demirel University, Isparta, Turkey, 2006.

[15] Everitt BS, Dunn G. Applied Multivariate Data Analysis. 2nd ed. London, UK: Wiley, 2001.

[16] Garrett SM. How do we evaluate artificial immune systems? Evolutionary Computation 2005; 13: 145-178.

[17] Wang H, Peng D, Wang W, Sharif H, Wegiel J, Nguyen D, Bowne R, Backhaus C. Artificial Immune System based image pattern recognition in energy efficient wireless multimedia sensor networks. In: Military Communications Conference 2008; San Diego, USA. pp. 1-7.

[18] Neagoe VE, Neghina CE. An artificial immune system approach for unsupervised pattern recognition in multispectral remote-sensing imagery. In: 2011 Proceedings of the 13th IASME/WSEAS International Conference on Mathematical Methods and Computational Techniques in Electrical Engineering conference on Applied Computing. pp. 228-233.

[19] Harmer PK, Williams PD, Gunsch GH, Lamont GB. An artificial immune system architecture for computer security applications. IEEE T Evolut Comput 2002; 6: 252-280.

[20] Oil CM, Wang YT, Ou CR. Intrusion detection systems adapted from agent-based artificial immune systems. In: IEEE International Conference on Fuzzy Systems; 2011; Taipei, Taiwan. pp. 115-122.

[21] Ou CM. Host-based intrusion detection systems adapted from agent-based artificial immune systems. Neurocomputing 2012; 88: 78-86.

[22] Engin O, Doyen A. A new approach to solve hybrid flow shop scheduling problems by artificial immune system. Future Gener Comp Sy 2004; 20: 1083-1095.

[23] Özşen S, Güneş S. Performance evolution of a newly developed general-use hybrid AIS-ANN system: AaA-response. Turk J Elec Eng & Comp Sci 2013; 21: 1703-1719.

[24] Mousavi M, Bakar AA, Zainudin S, Long ZA, Sahani, M, Vakilian M. Negative selection algorithm for dengue outbreak detection. Turk J Elec Eng & Comp Sci 2013; 21: 2345-2356.

[25] Baygin M, Karakose M. Immunity-based optimal estimation approach for a new real time group elevator dynamic control application for energy and time saving. Sci World J 2013; 2013: 1-12.

[26] Atay Y, Kodaz H. Optimization of job shop scheduling problems using modified clonal selection algorithm. Turk J Elec Eng & Comp Sci 2014; 22: 1528-1539.

[27] Sertkaya C. Immune System in Computer Security. MSc, Institute of Natural Sciences, Sakarya University, Turkey, 2009.

[28] Lefkovits S, Lefkovits L. Distance based k-nn classification of gabor jet local descriptors. Procedia Tech 2015; 19: 780-785.

[29] Saini I, Singh D, Khosla A. QRS detection using k-nearest neighbor algorithm (KNN) and evaluation on standard ECG databases. J Adv Res 2013; 4: 331-344.