



A new Feature Selection Algorithm using Modified PSO for an Anaerobic Wastewater Treatment System

¹R.Vijayabhanu, ²V. Radha

Department of Computer Science

Avinashilingam Institute for Homescience and Higher Education for Women

Coimbatore, Tamilnadu

INDIA

Abstract: Wastewater treatment is necessary to preserve the environment and living organisms. COD (Chemical Oxygen Demand) measures the amount of oxygen consumed by the water in the decomposition of organic matter and oxidation of inorganic matter and chemicals. Predicting effluents in the water is a time consuming process. The proposed method uses anaerobic treatment process to utilize anaerobic bacteria (biomass) to convert organic pollutants or COD into biogas in an oxygen free environment. Upflow Anaerobic Filter(UAF) is used as an anaerobic reactor and Cheese whey is used as a wastewater in the UAF. Artificial Neural Networks (ANN) are used as an estimation tool for predicting the performance of filtering treatment process in wastewater treatment plant. Hence the proposed method uses an cost effective techniques such as Dynamic Score Normalization technique with Mahalanobis distance(DSN-M) as a preprocessing step , Modified Particle Swarm Optimization for feature selection process and Adaptive Neuro-Fuzzy Inference System (ANFIS) as a learning algorithm for prediction. The experiment was carried out by comparing dataset without normalization, without feature selection and with ANFIS(Method 1) and dataset with the proposed normalization(DSN-M), feature selection(Modified PSO) and ANFIS (Method 2). The results proved that the proposed method achieves better performance in predicting the COD effluents present in wastewater.

Keywords: Wastewater treatment; Upflow Anaerobic Filter; COD; Dynamic Score Normalization technique with Mahalanobis distance; PSO; ANFIS.

I. Introduction

Water is the main factor for all living organisms. The quality of water gets degraded day by day. Several treatments are required to improve the quality of water. Wastewater can be defined as the stream of used water liberated from homes, industries, businesses, commercial activities and institutions is directed towards treatment plants by a carefully designed and engineered network of pipes. This wastewater can be further classified based on the sources of origin.

Inorganic materials such as heavy metals, acids, bases and salts, organic pollutants from industrial wastewater should be treated before discharging them into the water. Various methods available for treating wastewater such as adsorption, incineration, biological treatment and chemical or electro-chemical oxidation. The treatment is selected based on economics as well as ease of control, efficiency and reliability[1].

Primary, secondary, and tertiary treatment are the centralized wastewater treatment methods. Primary treatment includes physical barriers which removes larger solids from the wastewater. Allow the remaining particulates to be settle down. Secondary treatment includes combination of biological processes which promotes biodegradation by micro-organisms. It includes trickling filters, aerobic stabilization ponds and activated sludge processes. Secondary treatment also includes anaerobic reactors and lagoons. To further purify the wastewater of contaminants, pathogens, and remaining nutrients such as nitrogen and phosphorus compounds , tertiary treatment processes are used. Tertiary treatment can be achieved using one or a combination of processes which includes biological processes, ion exchange, maturation/polishing ponds, carbon adsorption, advanced filtration and disinfection. Sludge is produced in all the three stages of wastewater treatment [2].

The effect on the environment of agro-food production (particularly due to water emissions) is the major issue which has received attention in the recent years. Wastewater from agro-food production contains a relevant concentration of hardly biodegradable compounds phyto-toxic chemicals [3]. The proposed method uses cheese whey as wastewater. Cheese whey is the lactose-rich watery by-product obtained during cheese manufacturing. Cheese whey contains about 5% lactose, which is enough to be used as a substrate for fermentation purposes. [4].

Wastewater treatment methods can be classified into aerobic and anaerobic process. Proposed method uses anaerobic treatment method to predict the COD level in the wastewater. Anaerobic process generates energy in the form of biogas and produces sludge which is considerably lower than aerobic systems. [5] Several anaerobic reactors have been utilized over the past two decades such as anaerobic contact reactor, Anaerobic Filter reactor, anaerobic fluidized bed reactor, Upflow Anaerobic Filter(UAF), Upflow Anaerobic Sludge Blanket(UASB) etc. UAF is used as an anaerobic reactor in the proposed method. The performance of the UAF with cheese-dairy wastewater was regarded as the ANN modeling to predict the effluent COD level in wastewater.

The Prediction of COD level using ANN in the proposed work can be structured as follows: Data collection and preprocessing preceded by normalization technique using Dynamic Score Normalization technique with Mahalanobis distance, which is followed by Feature selection algorithm such as particle swarm optimization (PSO) with an effort to conclude the prediction of COD using ANFIS by the model when applied to the description of Upflow Anaerobic Filter (UAF) treating agro-food industrial wastewaters such as cheese-dairy. The experiments are carried out for the two methods. In Method 1, experiment was done without normalization, without feature selection and with ANFIS. In Method 2, experiment was carried out using normalization(DSN-M), feature selection(Modified PSO) and with ANFIS.

The paper is organized as follows: Section II describes related works, Section III describes methodology used in the proposed work, Section IV deals with the experimental results and Section V describes conclusion of the proposed work.

II. Related Work

Shuizhou Ke *et al.* [6] reviewed several applications and studied about two-phase anaerobic degradation for wastewater treatment. They also revealed the performance of application to treat the wastes from coffee, distillery, cheese whey and dairy, sludge, pulp and paper, food and solid, etc., and summarizes environmental and operational conditions, reactor configurations, and comparisons of two-phase anaerobic digestion with other anaerobic reactors.

Bahaa Mohamed Khalil *et al.* [7] examined the potential of the Artificial Neural Network (ANN) on simulating interrelation between water quality parameters. Several ANN structures, inputs and training possibilities are assessed and the best ANN model and modeling procedure is selected. They compared the prediction capabilities of the ANN with the linear regression models and auto correlated residuals. They also concluded that the ANN models are more accurate than the linear regression models having the same inputs and output.

Areerachakul [8] compared the predictive ability of the Adaptive Neuro-Fuzzy Inference System (ANFIS) model and Artificial Neural Network (ANN) model to estimate the Biochemical Oxygen Demand (BOD) on data from 11 sampling sites. The data is obtained from the Department of Drainage and Sewerage, Metropolitan Administration. The five parameters of water quality namely Chemical Oxygen Demand (COD), Ammonia Nitrogen (NH₃N), Dissolved Oxygen (DO), Nitrate Nitrogen (NO₃N), and Total Coliform bacteria (T-coliform) are used as the input of the models. These water quality indices affect the biochemical oxygen demand. The test is used to measure the amount of oxygen consumed by the organisms during a specified period of time (usually 5 days at 20°C). This result may be applied to automate BOD estimations.

Rajinikanth *et al.* [9] carried out research to determine the suitability and stability of an Upflow Anaerobic Filter (UAF) with small buoying polyethylene supports for the treatment of wastewater discharged from various small-scale agro-food industries to evaluate the maximum treatment capacity according to the various substrates, hydrodynamics inside the reactor and the specific biomass activity. Slobodan Ribaric *et al.* [10] tested normalization techniques such as Bayes-based normalization and several heuristic techniques (min-max, double-sigmoid, tanh, z-score, median-MAD, and piecewise-linear).

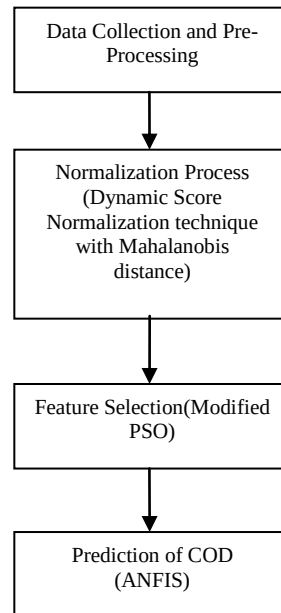
Seydi Ghomsheh *et al.* [11] introduced a new approach for training the adaptive network based fuzzy inference system (ANFIS). They applied one of the swarm intelligent branches namely Particle Swarm Optimization (PSO) with some modification in it to the training of all parameters of ANFIS structure. They also inspired these modifications by natural evolutions. Finally this method is applied to the identification of nonlinear dynamical system and is compared with basic PSO and showed quite satisfactory results.

III. Methodology

A. Anaerobic Wastewater Treatment System

The steps involved in the prediction of COD level present in wastewater treatment are shown in figure 1.

Figure 1 Process involved in the proposed method for Prediction of COD



B. Data collection and Preprocessing

The data must be evaluated to determine the validity and quality of data consisting of variable measurements. It is due to the endless sources of disturbances such as Electromagnetic interference, defective installation, insufficient maintenance etc.,. Hence data preprocessing is necessary for exact measurement. The proposed method used 100 records of cheese whey effluent to predict the COD level in wastewater. It was analyzed to detect and eliminate the missing values and outliers (screen the data) using K-Nearest Neighbor algorithm (K-NN).

C. Normalization

Generally normalized score is obtained by using normalization function. Normalization techniques are used to improve the prediction accuracy of Adaptive Network Based Fuzzy Inference System (ANFIS) model in predicting the COD level.

D. Dynamic Score Normalization with Mahalanobis Distance

The DSN-M technique is discussed in this section. DSN-M uses Dynamic Time Warping (DTW) to equal the cumulative recognition rate of a classifier against a standard curve that is identical for all the classifiers [12]. DSN-M considers the accumulated Gaussian probability function as standard curve ($\mu = 0, \sigma = 1$). Now, consider

$$R = (r_1, r_2, \dots, r_L) \quad \text{----- (1)}$$

is the L-point discrete version of the accumulated recognition function and

$$G = (g_1, g_2, \dots, g_L) \quad \text{----- (2)}$$

is the L-point discrete version of the accumulated probability function of the standard Gaussian distribution. A warping functions between R and G is any series of couples of indexes recognizing points of R and G to be joined [13]:

$$W(R, G) = c_1, c_2, \dots, c_T \quad \text{----- (3)}$$

in which $c_t = (i_t, j_t)$, being T, i_t, j_t integers, $1 \leq t \leq T$, $1 \leq i_t \leq L$, $1 \leq j_t \leq L$.

Now, if a distance measure $d(\cdot, \cdot)$ is considered between points of R and G ,

$$d(c_t) = d(r_{i_t}, g_{j_t}) \quad \text{----- (4)}$$

$d(c_t)$ can be combined to $W(R, G)$ the dissimilarity measure

$$D_{W(R, G)} = \sum_{t=1}^T d(c_t) \quad \text{----- (5)}$$

DSN-M uses Mahalanobis distance. The elastic matching procedure identifies the optimal warping function $W * (R, G) = c_1^*, c_2^*, \dots, c_T^*$ which satisfies the conditions of

- monotonicity,
- continuity,
- boundary conditions,

and for which it results [13]:

$$D_{W*(R,G)} = \min_{W(R,G)} D_{W(R,G)} \quad \text{---- (6)}$$

Now consider $W * (R, G)$ be the optimal warping function for the matching between R and G , given a matching score a_i , $i = 1, 2, \dots, n$, its normalized value is given as:

$$a'_i = \frac{1}{\text{card}(G_i)} \cdot \sum_{g_q \in G_i} g_q \quad \text{---- (7)}$$

where G_i is the collection of samples of G related with the sample r_i of the accumulated recognition function is given by,

$$G_i = \{g_q | (q, i) \in W * (R, G)\} \quad \text{---- (8)}$$

The normalized value of a score a_i is specified by the average value of the points of the accumulated Gaussian function which is equal to the sample r_i , being r_i the sample of the accumulated recognition function that corresponds to a_i . As a result, DTW is exploited to map the value of the accumulated recognition function r_i on the standard curve. The average of the values on the standard curve that corresponds to r_i is taken as the normalized value a'_i corresponding to a_i .

E. Feature Selection Algorithm for the Prediction of COD

Feature selection allows to reduce the feature space and helps in improving the accuracy by reducing training time. A Feature Selection Algorithm (FSA) is a computational solution which can be motivated by the definition of relevance. The main goal of FSA is to determine the relevant features of the dataset based on relevance. In Feature selection process, subsets features available from the data are selected to apply the learning algorithm. The best subset selection of data has the least number of dimensions which contributes to accuracy. This is the main stage of preprocessing the data [14]. The proposed method uses modified PSO algorithm for feature selection.

F. Modified PSO Algorithm

The Particle Swarm Optimization (PSO) algorithm is one of the population-based search algorithms which depends on the simulation of the social behavior of birds within a flock [15]. In PSO, each individual flies in the search space with velocity is dynamically adjusted based on the flying experience. Each individual flies in the PSO is a point in the D- dimensional search space[16].

In general, individual best and global best are the major PSO algorithms. In individual best, each individual compares position to its own best position. The global best drives the movement of particles includes the position of the best particle from the entire swarm. This type of algorithms can be computed as follows:[11]

- 1) The swarm, $P(t)$ of the particles should be initialized such that the position $\vec{x}_i(t)$ of each particle $P_i \in P(t)$ is random within the hyperspace, with $t = 0$.

- 2) The performance F of each particle is evaluated using its current position $\vec{x}_i(t)$.

- 3) Compare the performance of each individual to its best performance. If $F(\vec{x}_i(t)) < pbest_i$, then

$$pbest_i = F(\vec{x}_i(t)) \quad \text{---- (9)}$$

$$\vec{x}_{pbest_i} = \vec{x}_i(t) \quad \text{---- (10)}$$

- 4) Compare the performance of each individual to global best particle. If $F(\vec{x}_i(t)) < gbest$, then

$$gbest = F(\vec{x}_i(t)) \quad \text{---- (11)}$$

$$\vec{x}_{gbest} = \vec{x}_i(t) \quad \text{---- (12)}$$

- 5) Change the velocity vector for each:

$$\vec{v}_i(t) = \vec{v}_i(t-1) + \rho_1(\vec{x}_{pbest_i} - \vec{x}_i(t)) + \rho_2(\vec{x}_{gbest} - \vec{x}_i(t)) \quad \text{---- (13)}$$

Where ρ_1 and ρ_2 are random variables. The second term is referred to as the cognitive component, while the last term is the social component.

- 6) Move each particle to a new position:

$$\vec{x}_i(t) = \vec{x}_i(t-1) + \vec{v}_i(t) \quad \text{---- (14)}$$

$$t = t + 1 \quad \text{---- (15)}$$

- 7) Go to step 2, and repeat until convergence.

The random variables ρ_1 and ρ_2 are defined as $\rho_1 = r_1 C_1$ and $\rho_2 = r_2 C_2$ with $r_1, r_2 \sim U(0,1)$ and C_1 and C_2 are positive acceleration constant.

Remove the worst particle in population and replace it by new particle. It is to be noted that how to determine worst particle and how to generate new particle for current population. Select the particles which have worst local best value at generation. Then randomly choose two particles from the population and use crossover operator to generate the two new particles. Finally, select the best one from two newly generated particles and replace the worst particle with newly selected particle. In other hand, combine GA operator and PSO algorithm to make these modifications. This modification make this algorithm to be faster than basic algorithm.

G. Prediction of COD using ANFIS

The amount of oxygen required to chemically oxidize organic compounds in water. The chemical oxygen demand (COD) test is a common water quality test used to indirectly measure the total amount of organic compounds in a water sample using a strong oxidizing agent such as potassium dichromate. A high COD value indicates a high concentration of organic matter in the water sample. It can be predicted with high accuracy using proposed ANFIS

H. ANFIS

ANFIS is a method based on the input-output data of the system under consideration. Success in obtaining a reliable and robust ANFIS network depends heavily on the choice of process variables involved as well as the available data set and the domain used for training purposes [17]. Basically, a fuzzy inference system is composed of five function blocks [18]:

- (i) Rule base containing a number of fuzzy if-then rules.
- (ii) A database which defines the membership function of the fuzzy sets are used in the fuzzy rules.
- (iii) The decision-making unit which perform the inference operation on the rules.
- (iv) Fuzzification inference transforms the crisp inputs into degree of match with linguistic values.
- (v) Defuzzification inference which transforms the fuzzy results of the inference into a crisp output.

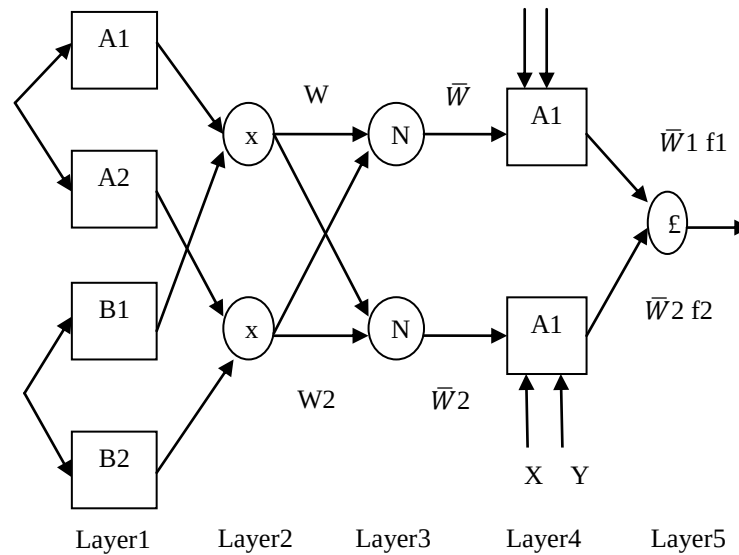
For simplicity, a fuzzy inference system with two inputs x and y , and one output is assumed., a common rule set with two fuzzy if-then rules is defined for the first-order Sugeno fuzzy model is as follows:

$$\begin{aligned} \text{Rules 1: If } x \text{ is } A_1 \text{ and } y \text{ is } B_1, \\ \text{then } f_1 = p_1x + q_1y + r_1 \end{aligned} \quad \text{---- (16)}$$

$$\begin{aligned} \text{Rules 2: If } x \text{ is } A_2 \text{ and } y \text{ is } B_2, \\ \text{then } f_2 = p_2x + q_2y + r_2 \end{aligned} \quad \text{---- (17)}$$

Here type-3 fuzzy inference system proposed by Takagi and Sugeno [18] is used. In this inference system the output of each rule is a linear combination of input variables added by a constant term. Average weight of each rule's output is the final output.

Figure 2 A typical ANFIS architecture



Each node in layer 1 is an adaptive node with a node function which may be a Gaussian membership function or any membership functions. Each node in layer 2 is a fixed node labeled Π which indicates the firing strength of each rule. All nodes in layer 3 is a fixed node which is labeled as N represents the normalized firing strength of each rule. In Layer 4 every node is an adaptive node with a node function. In layer 5, single node is a fixed node labeled Σ , represents the overall output (Z). It is defined as the summation of all incoming signals [19].

IV. Experimental Results

Treating cheese whey as an agro-food wastewater in Upflow Anaerobic Reactor (UAF) for the duration of 6 months were obtained. Each and every database have some erroneous or missing values in the dataset. These missing values and outliers can be eliminated using K-Nearest Neighbor algorithm(K-NN). The dataset then undergoes normalization, feature selection and prediction.

The controlling parameters of Upflow Anaerobic Filter and influent COD levels were considered as the input parameters. To recognize these input parameters, the model was trained in order to predict the corresponding effluent level of COD in wastewater as output parameters. The model should be trained with the data in the training subset with influent COD levels, flow rate and OLR to optimize the network weights so as to minimize the appropriate error function. After preprocessing, n number of features were selected based on modified PSO algorithm and it is given as input.

In order to evaluate the performance of the proposed model two statistical indices are used: Mean Square Error (MSE) and Regression coefficient (R^2) values that are derived in statistical calculation of observation in model output predictions can be defined as:

$$MSE = \sum_{i=1}^N \frac{(x-y)^2}{N} \quad (15)$$

$$R^2 = 1 - \frac{\sum (x-y)^2}{\sum x^2 - \frac{\sum y^2}{n}} \quad (16)$$

where w_0 is the actual values of W_{t+1} with $\{i = 1, 2, \dots, n\}$ observations, n is the total observation number and w_p is the predicted W_{t+1} value.

The following table 1 represents the values obtained during experiment. Method 1 shows the output of dataset without normalization, without feature selection and with ANFIS and Method 2 illustrates the output of dataset with the proposed normalization(DSN-M), feature selection(Modified PSO) and with ANFIS.

Table 1. Comparing the parameters obtained from Method 1 and Method 2

Prediction of COD based on the parameters	R^2	MSE	Training Time(secs)	Testing time(secs)
Method 1	0.77	0.0051	169.64	0.352
Method 2	0.89	0.0032	32.23	0.262

Figure 3 Comparing regression coefficient values of two methods

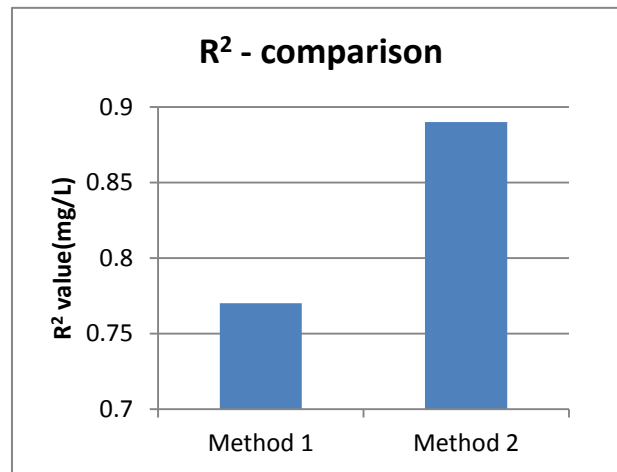


Figure 3 reveals the comparison of the two methods. It is clear from the figure that the regression coefficient values obtained using method 2 gives better results than the method 1.

Figure 4 Comparing Mean Square Error values of two methods

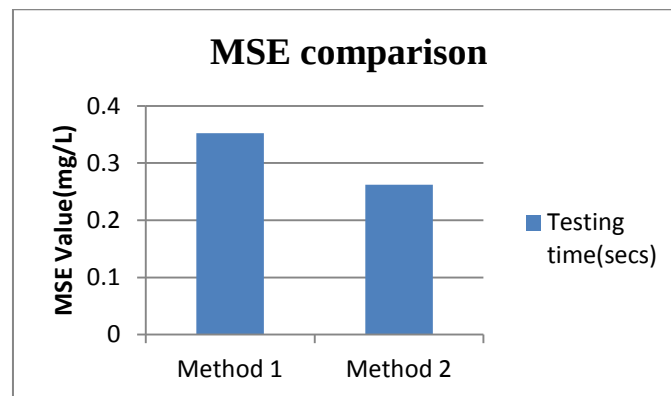


Figure 4 illustrates the men square error values obtained using the two methods. The method 2 has lower mean square error than method 1. It shows the proposed method achieves good results in predicting COD effluent in wastewater.

Figure 5 Comparing Execution of Testing time of two methods

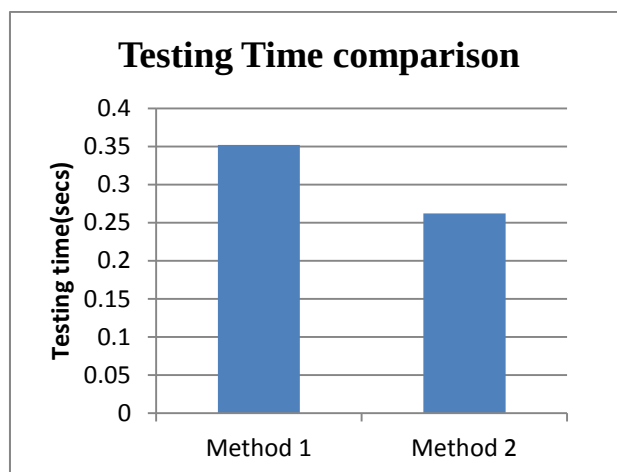


Figure 5 represents the time taken to complete the average execution. The method 2 takes less testing time to complete the overall execution than the method 1. Thus the proposed method predicts the COD effluent level in less time.

The experimental results shows that the proposed method achieves better results in terms of regression coefficient, Mean square error and average execution time. Thus the proposed method achieves better results for predicting the COD level.

V. Conclusion

Designing Wastewater treatment plant is difficult task due to the high nonlinearity of the plant and the non-uniformity and variability of the crude supply as well as the nature of the biological treatment. The proposed method uses raw data from cheese whey effluent treated in a Upflow Anaerobic Filter(UAF) during a period of 6 months has been analyzed to recognize the datasets. The missing values and outliers in the dataset has been eliminated using K-Nearest Neighbor algorithm (K-NN). The ANN model was trained and tested in MATLAB using Dynamic Score Normalization technique with Mahalanobis distance as normalization process followed by Modified Particle Swarm optimization (PSO) as a feature selection to maximize prediction accuracy and by using ANFIS to predict the target COD Levels. Average prediction Accuracy is 81% for Method 1 and 90% for Method 2. The ANFIS training algorithm provides good estimates of COD levels covering a range of data for training and testing purposes.

References

- [1] Ch. Cominallis, A. Nerini, "Anodic oxidation of phenol in the presence of NaCl for wastewater treatment" *Journal of Applied Electrochemistry* vol. 25, pp. 23-28, 1995.
- [2] Michiel R.J. Doorn, Sirintornthep Towprayoon, Sonia Maria Manso Vieira, "Wastewater Treatment and Discharge", *IPCC Guidelines for National Greenhouse Gas Inventories* vol. 5, Chapter 6, 2006.
- [3] Simona Caudo, Chiara Genovese, Siglinda Perathoner, "Copper-pillared clays (Cu-PILC) for agro-food wastewater purification with H₂O₂", Elsevier Inc, 2007, doi:10.1016/j.micromeso.2007.05.011
- [4] Mongi Ferchichi, Edward Crabbe, Gwang-Hoon Gil, "Influence of initial pH on hydrogen production from cheese whey" *Journal of Biotechnology* vol. 120, pp. 402–409, 2005.
- [5] Renato Carraha Leita, Adrianus Cornelius van Haandel, Grietje Zeeman, Gatzte Lettinga, "The effects of operational and environmental variations on anaerobic wastewater treatment systems: A review", *Bio-resource Technology*, vol.97, pp. 1105–1118, 2006.
- [6] Shuizhou Ke, Zhou Shi, Herbert H.P. Fang, "Applications of two-phase anaerobic degradation in industrial wastewater treatment", *International Journal on Environment and Pollution*, vol. 23, no. 1, 2005.
- [7] Bahaa Mohamed Khalil, Ayman Georges Awadallah, Hussein Karaman, Ashraf El-Sayed, "Application of Artificial Neural Networks for the Prediction of Water Quality Variables in the Nile Delta", *Journal of Water Resource and Protection*, vol. 4, pp. 388-394, 2012.
- [8] S. Areerachakul, "Comparison of ANFIS and ANN for Estimation of Biochemical Oxygen Demand Parameter in Surface Water", *International Journal of Chemical and Biological Engineering*, vol.6, 2012.

- [9] Rajagopal Rajinikanth, Rengaraj Ganesh, Pradeep Kumar, Joseph V. Thanikald, "High rate anaerobic filter with floating supports for the treatment of effluents from small-scale agro-food industries", *Desalination and Water Treatment* vol. 4, pp. 183–190, 2009.
- [10] Slobodan Ribaric and Ivan Fratric, "Experimental Evaluation of Matching-Score Normalization Techniques on Different Multimodal Biometric Systems", *IEEE Electro technical conference*, pp. 498- 501, 2006.
- [11] V.Seydi Ghomsheh , M. Aliyari Shoorehdeli , M. Teshnehlab, "Training ANFIS Structure with Modified PSO Algorithm", *Proceedings of the 15th Mediterranean Conference on control 7 Automation*, July 2007.
- [12] R. Modugno, G. Pirlo and D. Impedovo, "Score normalization by Dynamic Time Warping", *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSAP)*, Pp. 82-85, 2010.
- [13] D. Impedovo, G. Pirlo, C. Trullo and E. Stasolla, "Combination of Measurement-Level Classifiers: Output Normalization by Dynamic Time Warping", *Proc. 10th Int. Conf. Document Analysis and Recognition (ICDAR-10)*, Barcelona, Spain, IEEE Comp. Soc. Press, Pp. 416-420, 2009.
- [14] L.Ladha, T.Deepa, "Feature Selection Methods and Algorithms", *International Journal on Computer Science and Engineering (IJCSE)* , vol. 3, no. 5, May 2011. ISSN : 0975-3397
- [15] J. Kennedy, R.C. Eberhart,, "particle swarm optimization", *Proceedings of the IEEE International Conference on neural networks*, vol. 4, pp. 1942 –1948, 1995.
- [16] Andries P.Engelbrecht, "Computational Intelligence An introduction", John Wiley & Sons Limited, 2002.
- [17] M. Buragohain and C. Mahanta "A novel approach for ANFIS modelling based on full factorial design" , *Applied Soft Computing* , pp. 609–625, 2008.
- [18] J.S.R. Jang, "ANFIS: Adaptive-network-based fuzzy inference systems", *IEEE Trans. Syst. Man Cybern.*, vol. 23, pp.665-685,1993.
- [19] M.T. Dastorani, A. Moghadamnia, J. Piri and M. Rico-Ramirez, "Application of ANN and ANFIS models for reconstructing missing flow data", *Environmental Monitoring Assess*, 2009.

Acknowledgments

The heading of the Acknowledgment section and the References section must not be numbered.