Validation for Data Classification

HILARIO LÓPEZ and IVÁN MACHÓN and EVA FERNÁNDEZ
Departamento de Ingeniería Eléctrica, Electrónica de Computadores y Sistemas
Universidad de Oviedo
Edificio Departamental 2. Zona Oeste. Campus de Viesques s/n. 33204 Gijón (Asturias)
SPAIN
http://isa.uniovi.es/~hilario

Abstract: - Neural network models must be validated to be used as estimators. In this paper a detailed description for Self-Organizing Map (SOM) validation is proposed to determine the map size and the optimum training data set. This idea was applied in a wastewater biological treatment.

Key-Words: - Self-organizing mapping, validation, data classification, wastewater biological treatment, chemical oxygen demand, sequencing batch reactor.

1 Purpose of the paper

This work is part of the KNOWATER II project "Implementation of a Knowledge Based System for Control of Steelworks Waste Water Treatment Plant", which is sponsored by ECSC and their agreement number is 7210-PR-234. The contractors are Centro Sviluppo Materiali S.p.A., Corus RT&D, Betrieb Forschung Institut (BFI) and Universidad de Oviedo. The main objective of the KNOWATER II project was the development of plant supervision techniques for implementation in wastewater treatment plants.
The present work gives a detailed description of the proposed validation for Self-Organizing Map (SOM) model to be used in a subsequent data classification process. This validation consists in determining the SOM map size and the optimum training data set. These proposed techniques were used to optimize a wastewater biological treatment in a sequencing batch reactor. Finally, the aerobic end-point detection is achieved.

2 Sequencing Batch Reactor

The wastewater is treated biologically in a Sequencing Batch Reactor (SBR). The closed-loop of the oxygen control in the SBR is shown as block diagram in Fig. 1. The dissolved oxygen concentration is controlled by a PID. Air is pumped into the reactor and a valve is regulated. The set-point is between 6 and 5 mgO_2/l. Moreover, a recorder is installed to work as data acquisition interface between the sensors and the developed end-point detection technique located on a PC station. The interface is able to establish a TCP/IP protocol with the developed software. The dissolved oxygen electrode has a temperature sensor to compensate the measurement deviations, so the temperature can be measured.

An initial off-line study of the process was done [1]. Taking into account this previous study, the PID controller output of the oxygen closed-loop was connected and registered as one of the process variables to train the SOM network.

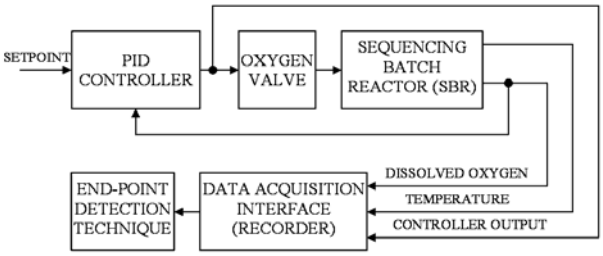

Fig. 1. Connection between the oxygen control closed-loop and the end-point detection technique

3 SOM

Self-Organizing Map (SOM) was used as a first stage to develop a model for data classification. The SOM [2] consists of a regular lattice typically defined in a two dimensional space composed of several neurons placed in the nodes of the lattice. SOM training implies assigning a set of coordinates in the input data space, which are called prototype vectors, to each neuron. Thus, each neuron is represented by a prototype vector and a correspondence is established between the coordinates of each neuron in the input space (data set) and their coordinates in the 2D-lattice or output space.
This study was carried out using the SOM toolbox version 2.0 [3] developed at the HUT (Helsinki University of Technology). The steps taken to analyze the data are outlined in a previous work [4]. Firstly, the most significant process variables are selected. These variables are described in table 1. Secondly, the data were normalized to a zero mean

value and a unitary variance to make SOM treat them in the same way. After normalizing the SOM network was trained with these variables using batch training algorithm. Once the SOM has converged, it stores the most relevant information about the process in its prototype vectors. The final step obtains the best clustering structure that allows the main process zones to be visualized [5].

Table 1. Training Variables

Name	Description
OXYGEN	Dissolved oxygen concentration (mgO_2/litre)
CONTROLLER OUTPUT	Output of the PID controller of the oxygen closed loop (0-100)
TEMPERATURE_SBR	Temperature in the SBR (C)

4 SOM validation

According to the properties of the SOM, the trained neural network must achieve the topology preservation of the data. Therefore the neighborhood on the output space and in the input space must be similar. If two prototype vectors close to each other in the input space are mapped wide apart on the grid, this is signaled by the situation where two closest best matching neurons of an input vector are not adjacent neurons. This kind of fold is considered as an indication of the topographic error in the mapping and does not verify the SOM property about training data topology preservation where neighbor neurons of the output space correspond to similar values of the process variables, i. e., regions of the output space represent working zones of the process

The topographic error [6] can be calculated by equation (1) as the proportion of sample vectors for which two best matching neurons are not adjacent. N is the number of samples, x_k is the kth sample of the data set and $u(x_k)$ is equal to 1 if the first and second best matching neurons of x_k are not adjacent neurons, otherwise zero.

$$e_t = \frac{1}{N}\sum_{k=1}^{N} u(x_k) \tag{1}$$

The results of this error measurement are very easy to interpret and are also directly comparable between different models and even mapping of different data sets. Moreover, the prototype vectors approximate to the data set trying to substitute a data vector for a prototype vector of the SOM. A consequence of this approach is the quantization error. Equation (2) is usually used to calculate the average quantization error over the whole data set. N is the number of samples, x_i is the ith data sample and m_b is the prototype vector of the best matching neuron for x_i.

$$e_q = \frac{1}{N}\sum_{i=1}^{N}\|x_i - m_b\| \tag{2}$$

The SOM toolbox uses equations (3) and (4) to determine the output space size. The number of neurons of the output space is determined by equation (3). M is the number of neurons and N is the number of samples of the training data.

$$M = 5\cdot\sqrt{N} \tag{3}$$

On the other hand, the criterion of the utilized toolbox to determine the ratio between the number of rows n_1 and the number of columns n_2 of the 2D grid or output space is calculated according to equation (4). The ratio between sidelengths of the map is the square root of the ratio between the two biggest eigenvalues of the training data. The highest eigenvalue is e_1 and the second highest is e_2.

$$\frac{n_1}{n_2} = \sqrt{\frac{e_1}{e_2}} \tag{4}$$

Five data sets showed in Fig. 2 and Fig. 3, which correspond to the aerobic phase of the SBR, are available to carry out the validation of the model. Each sample is the mean values of the process variables for 8 minutes and 20 seconds. The objective is to find out the model that minimizes the quantization and topographic errors from several neural networks which have been trained using each of these available patterns and, at the same time, for different map sizes. Thus, a specific data set and an optimum map size must be selected. The validation method can be summarized in the following steps [7]:

1) A data set or pattern p_i is chosen to train the network. The data are normalized to a distribution with zero mean value and unitary variance.
2) Batch training is carried out on the SOM map whose sidelengths are calculated by means of equations (3) and (4) using pattern p_i as training data.
3) Once the trained model is obtained, the topographic and quantization errors are calculated for the remaining patterns p_j which have not been used during the training. These patterns must also be previously normalized.

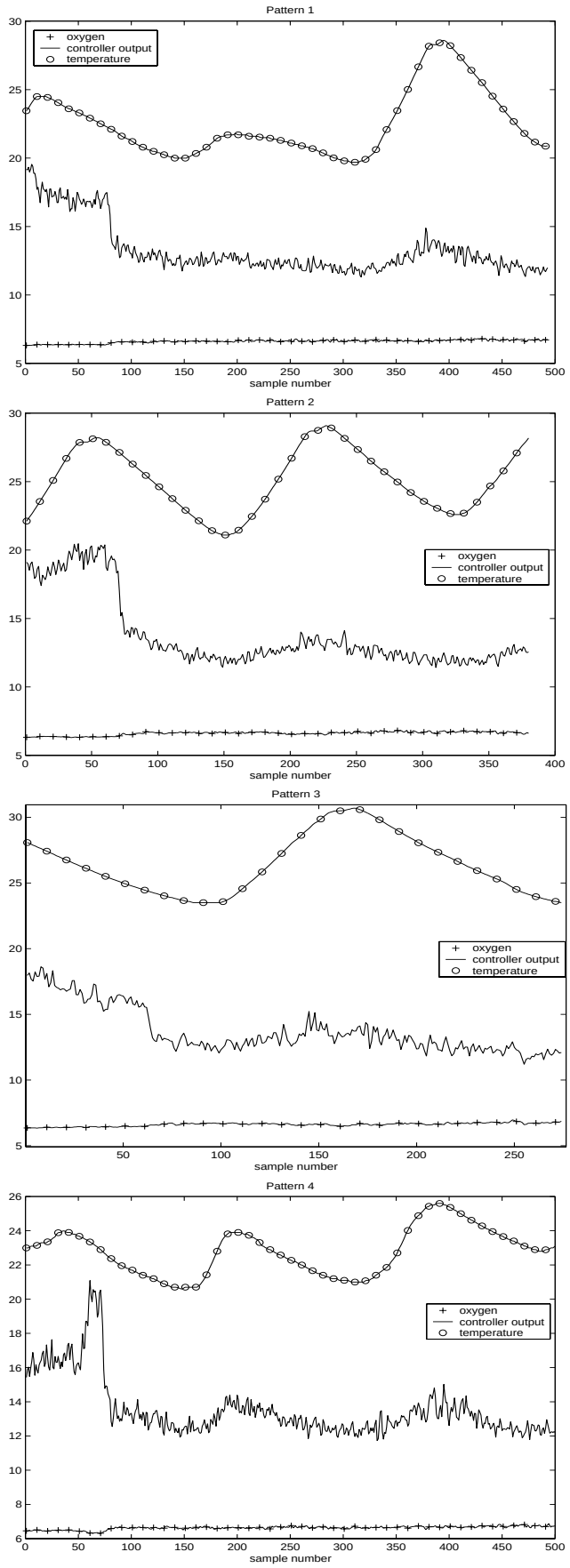

Fig. 2. Pattern 1, 2, 3 and 4

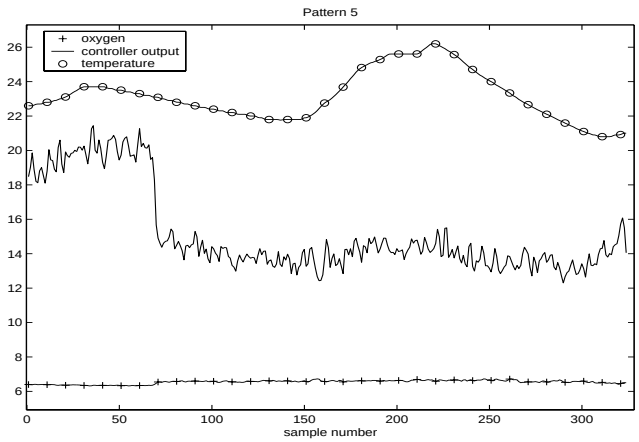

Fig. 3. Pattern 5

4) The size of this trained map is increased and reduced respecting the proportionality of its sidelengths (width and length). Once the size has been modified, the neural network is again trained using pattern p_i.
5) The third and fourth steps are repeated for different map sizes.
6) Steps 1 through 5 are repeated for the remaining patterns p_j, assuming each of these the role of pattern p_i.

Several map sizes have been trained using the five patterns. The mean values of the errors over the available patterns in function of the map number are shown in Fig. 4 and Fig. 5 using each pattern as training data. The sidelengths of the trained maps for each utilized pattern are showed in table 2. It can be seen that the larger the map size the lower the quantization error but the higher the topographic error. This is due to the neural network folds to reduce the quantization error. Moreover, the larger the map size the higher the computational cost. Therefore, there is compromise between the increase of the topographic error and the reduction of the quantization error. A curve, which represents the sum of both errors, has been added to the graphics.

The model whose sidelenghts have been calculated by means of equations (3) and (4) correspond to a horizontal axis value equal to 6 (map number equal to 6). The quantization error has been reduced and the topographic error has been incremented not very much, as can be seen in Fig. 4 and Fig. 5. The results verify that equations (3) and (4) to determine the map sidelengths is a reasonable optimum solution of the compromise between the quantization error, the computational cost and the topographic error. The selected model, which is used as validated pattern, was trained using pattern 1 and the demonstrated criterion, equations (3) and (4), to determine the output space size.

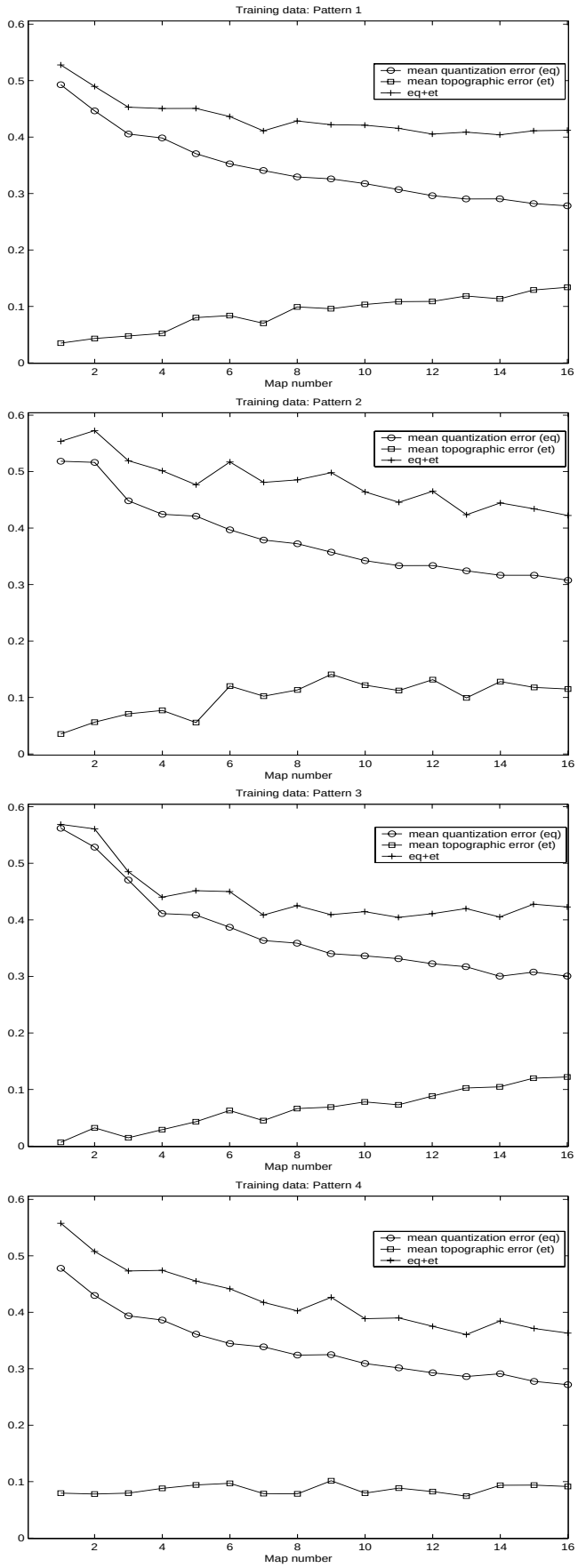

Fig. 4 Mean Value of Errors using Pattern 1, 2, 3 and 4 as Training Data

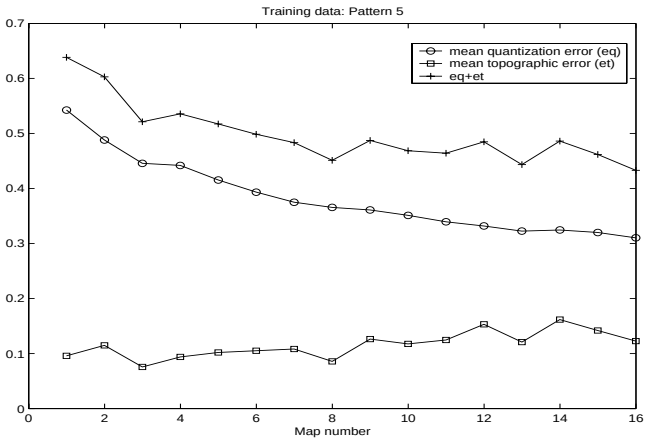

Fig. 5. Mean Value of Errors using Pattern 5 as Training Data

5 Data classification

The data classification process consists of a two-stage procedure [8]. Firstly, the prototype vectors are obtained training the data of the aerobic phase using a SOM algorithm and then clustering them using a K-means algorithm (Fig. 6), see [9]. Ten clustering structures were obtained varying the predefined number of clusters.

Finally, the best clustering structure between the ten structures, which have been obtained from the K-means algorithm, is selected using the Davies-Bouldin index [10]. This index searches the model that minimizes the within-cluster distance and maximizes the between-clusters distance. The Davies-Bouldin index is suitable for evaluation of K-means partitioning, because it gives low values indicating good clustering results for spherical clusters. The best clustering structure is determined by the lowest value of the Davies-Bouldin index.

The best clustering corresponds to a number of two clusters and has been projected onto the SOM. It is displayed in Fig. 6. Cluster 1 corresponds to the first hours of the aerobic treatment where the values of the controller output are high due to the high chemical oxygen demand (COD). During this period the biological activity is high and the toxic substances are eliminated by means of the cellular metabolism, whereas cluster 2 represents the data collected after this high biological activity where the values of the controller output are lower because the COD has decreased. If this is the state of the treatment plant, the biological treatment of the aerobic stages can be finished improving the capacity of the plant.

Table 2. Trained Map Sizes

Map number	Map size (training data: Pattern 1)	Map size (training data: Pattern 2)	Map size (training data: Pattern 3)	Map size (training data: Pattern 4)	Map size (training data: Pattern 5)
No 1	7 x 4	7 x 4	6 x 4	7 x 4	6 x 4
No 2	8 x 4	8 x 5	7 x 4	8 x 5	7 x 5
No 3	10 x 5	10 x 6	8 x 5	10 x 6	8 x 6
No 4	11 x 6	11 x 6	10 x 6	11 x 6	9 x 6
No 5	13 x 6	13 x 7	11 x 6	13 x 7	10 x 7
No 6	14 x 7	14 x 8	12 x 7	14 x 8	11 x 8
No 7	15 x 8	15 x 9	13 x 8	15 x 9	12 x 9
No 8	17 x 8	17 x 10	14 x 8	17 x 10	13 x 10
No 9	18 x 9	18 x 10	16 x 9	18 x 10	14 x 10
No 10	20 x 10	20 x 11	17 x 10	20 x 11	15 x 11
No 11	21 x 11	21 x 12	18 x 11	21 x 12	17 x 12
No 12	22 x 11	22 x 13	19 x 11	22 x 13	18 x 13
No 13	24 x 12	24 x 14	20 x 12	24 x 14	19 x 14
No 14	25 x 13	25 x 14	22 x 13	25 x 14	20 x 14
No 15	27 x 13	27 x 15	23 x 13	27 x 15	21 x 15
No 16	28 x 14	28 x 16	24 x 14	28 x 16	22 x 16

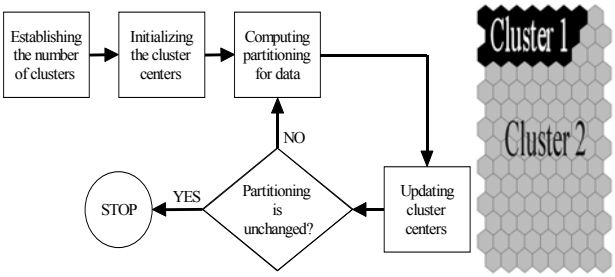

Fig. 6. K-means Algorithm and Best Clustering Structure

6 Results

The proposed AI techniques achieve the data classification. The process monitoring is obtained estimating the current process state by means of using the validated pattern (pattern 1).

The training data set must only contain the samples of the aerobic stage and is determined by the mean value of the controller output because this signal can be considered as a key variable to estimate the states of the treatment, see [1] and [11].

As mentioned above, the best clustering structure is composed of 2 clusters and is calculated by means of the Davies-Bouldin index. A cluster corresponds to HIGH COD and the other is the LOW COD.

The cycles of the biological treatment at the sequencing batch reactor can be clearly observed in Fig. 7 and Fig. 8. The higher values correspond to the anoxic stage when the controller output is saturated and equal to 100%. The rest of the data corresponds to the aerobic stage (including sedimentation).

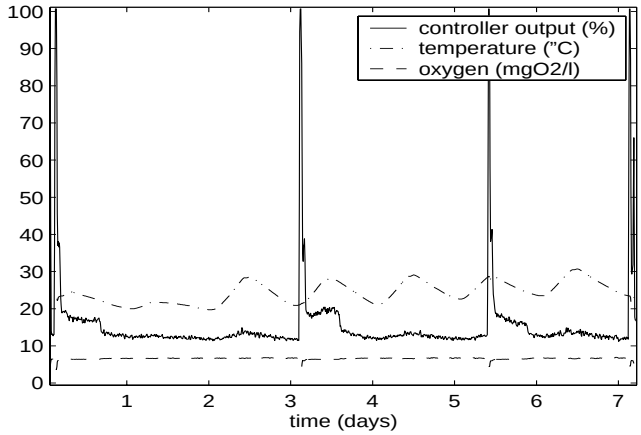

Fig. 7. Process Values of one year ago

An important aspect appears: the end-point of the aerobic reaction. This end-point detection can be used to finalize the aerobic stage and in this way the duration of the cycle is shorter increasing the operating capacity of the plant. The estimation of the time of the main activity of the treatment (aerobic phase end-point) achieves operating cost savings and increases the plant performance; see [12]. The duration of the cycle was initially 48-72 hours one year ago, see Fig. 7, and it has been reduced to 24 hours as is showed in Fig. 8. In this way the operating capacity of the plant has been increased by reducing the retention time. The process state is estimated projecting the current values onto a SOM network by means of standing out the best matching neuron from the rest of the neurons. This SOM network is used as a pattern and is previously stored and validated using the validation method explained above (pattern 1). Thus, the end of the main biological activity can be identified.

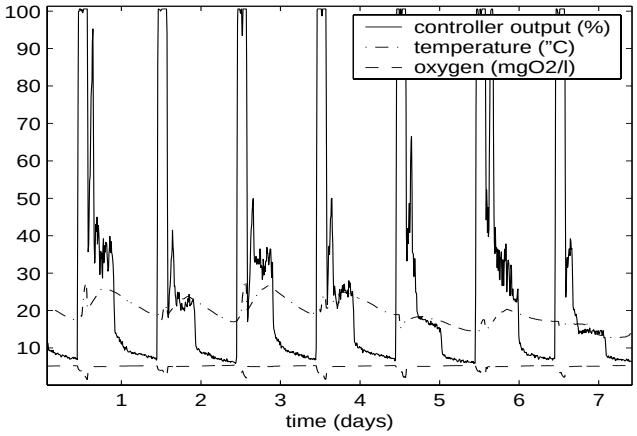

Fig. 8. Current Process Values

7 Conclusions

The data classification is obtained training a SOM network using the data of the aerobic stage as training set with a subsequent K-means algorithm and making use of Davies-Bouldin index for partition validation. Also a procedure is outlined to determine the optimum map size and the training data set for SOM validation. This method is the end-point detection technique that allows the identification of the end of the aerobic reaction achieving operating cost savings and increasing the plant performance. In this way, total retention time was reduced from 48-72 hours to 24 hours.

References:

[1] López H. and I. Machón. 2004a. "Biological wastewater treatment analysis using som and clustering algorithms," in *Proc. 12th Mediterranean Conference on Control and Automation*, Kusadasi.

[2] Kohonen T. 2001. *Self-Organizing Maps*. New York: Springer-Verlag.

[3] Vesanto J.; E. Alhoniemi; J. Himberg; K. Kiviluoto and J. Parviainen. 1999. "Self-organizing map for data mining in matlab: the som toolbox," *Simulation News Europe*, pp. 25–54.

[4] López H.; I. Machón and S. Roces. 2003. "Waste treatment monitoring using self-organizing map and condition achievement maps," in *Proc. IFAC 5th Symposium on Intelligent Components and Instruments for Control Applications*, Aveiro.

[5] López H. and I. Machón. 2004b. "Self-organizing map and clustering for wastewater treatment monitoring," *Engineering Applications of Artificial Intelligence*, vol. 17, no. 3, pp. 215–225.

[6] Kiviluoto K. 1996. "Topology preservation in self-organizing maps," in *IEEE International Conference on Neural Networks*, vol. 1, pp. 294–299.

[7] Machón I. and H. López. 2004. "An application for on-line control of a sequencing batch reactor," in *Proc. IFAC Workshop on Modelling and Control for Participatory Planning and Managing Water Systems*, Venice.

[8] Vesanto J. and E. Alhoniemi. 2000 "Clustering of the self-organizing map," *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 586–600.

[9] McQueen J. 1967. "Some methods for classification and analysis of multivariate observations," in *5-th Berkeley Symposium on mathematics, Statistics and Probability*, no. 1, pp. 281–298.

[10] Davies D. and D. Bouldin. 1979. "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 1, no. 2, pp. 224–227.

[11] López H. and I. Machón. 2004c. "An introduction to biological wastewater treatment explained by som and clustering algorithms," in *Proc. IEEE International Symposium on Industrial Electronics*, Ajaccio.

[12] Andreottola G.; P. Foladori and M. Ragazzi. 2001. "On-line control of a sbr system for nitrogen removal from industrial wastewater," *Water Science and Technology*, vol. 43, no. 3, pp. 93–100.