CLASSIFICATION OF DOMESTIC WATER CONSUMPTION USING AN ANFIS MODEL

M. A. CORONA-NAKAMURA, DCC-CUCEI UNIVERSIDAD DE GUADALAJARA, REVOLUCION #1500. C.P. 44840 GUADALAJARA, JALISCO, MÉXICO, acoronak@yahoo.com.mx
R. RUELAS, DIP-CUCEI UNIVERSIDAD DE GUADALAJARA, JOSÉ GUADALUPE ZUNO # 48. C.P. 45101 ZAPOPAN, JALISCO, MÉXICO, rruelas@newton.dip.udg.mx
B. OJEDA-MAGAÑA, DIP-CUCEI UNIVERSIDAD DE GUADALAJARA, JOSÉ GUADALUPE ZUNO # 48. C.P. 45101 ZAPOPAN, JALISCO, MÉXICO, benojed@hotmail.com
D. ANDINA, DEPTO.SSR UNIVERSIDAD POLITÉCNICA DE MADRID, E.T.S.I INGENIEROS DE TELECOMUNICACIÓN. MADRID ESPAÑA, diego@gc.ssr.upm.es

ABSTRACT
This work presents classification results of different water outputs in a house. Input variables are time and flow measurements in a point of the network distribution, and the identified classes are relevant consumptions as sink consumption, shower consumption, etc. Due to human influence on consumption data, we selected a classifier based on an interpretable model; that allows the incorporation of knowledge provided by users or experts. Thus, this study is based on the well known Anfis model and AGUA (real data taken for a project being developed in Guadalajara, Mexico) the data set corresponding to a supervised case. The result shows that the proposed algorithm works well, with recognition above 91%, and it could be used for a better profit of domestic water management.

KEYWORDS: Residential water consumption, Water supervision, Classification, Anfis, Neuro-fuzzy model.

1. INTRODUCTION
Due to climatic changes around the world, it is very important to get a better profit of water. For this reason, we make a proposal, from flow and time measurements of water consumption at houses, to identify automatically the output where water is consumed at a particular moment. So, this classifier could be the base for the development of a supervision system in order to achieve a better profit of water.

In this work we present a study about the different water consumptions in a house, where three people live. A data set was built according to this house and the available output there. Figure 1 shows distribution of data for seven total outputs; even if human behavior produces two subclasses in one output and three subclasses in another, for the particular studied case. The data set consists of 1000 examples; 100 for each one of five classes and 100 for each one of five subclasses. The classes are: three toilets (t_1, t_2 and t_3), a sink, a shower, a washbowl, and a washing machine. The shower can be divided in three subclasses due to preferences and habits of the human beings, and the washing machine in two subclasses as result of the user intervention at the beginning of the washing cycle.

The classification problem of water consumption is very variable from one house to another, because it depends on the number of outputs, hydraulic installations, kinds of furniture and dispositive used, furthermore on uses and customs of inhabitants of each house, which could have a great influence on consumption. For these reasons, it is necessary to use a model for the classifier that offers easy of interpretation by humans, and the possibility to include knowledge provided by users or experts. These conditions directed us to a neuro-fuzzy model, and specially the Anfis [1] model, as a good alternative for the output classification, because this model is based on a set of conditional "If … Then …" rules.

In order to use the Anfis model, the membership functions must be defined for every variable of the model. The parameters of these functions can be estimated with fuzzy clustering algorithms, such as the Fuzzy c-Means (FCM), if the number of clusters is known, otherwise, the subtractive clustering algorithm can be used, for example. Once the membership functions have been estimated, they can be adjusted with the Anfis model.

We have started this study with a supervised case. However, a classifier installed at each house must require an unsupervised learning, and the participation of the user such that one can identify, or associate labels, to each one of the outputs. Anfis was selected as the classification model and we propose two approaches for the estimation of the initial membership functions values of the model. The selected functions are Gaussians, so we have to estimate their centers and dispersions. In one proposed approach the maximum and minimum data points of each class must be identified in order to estimate the center and dispersion of each membership function. In the other approach we propose to use the mean value and the standard deviation of each class for the calculus of the centers and dispersions respectively. As can be seen in the results section, both proposed approaches for the initial membership functions estimation, in the supervised case, are easy and quick to use, and they allow a reduced number of errors.

The structure of this work is defined as follows. The second section presents neuro-fuzzy models and Anfis model. The third section contains the methodology for the output water classification, according to the data set used. This section also presents the two proposals for the estimation of the initial membership function values. The fourth section presents the classification results, and the fifth section gives the main conclusions about this work.

2. NEURO-FUZZY MODELS AND ANFIS

2.1 Neuro-fuzzy models

Neuro-fuzzy models are characterized because we try to use with them the advantages of neural networks and fuzzy logic models. Neural networks provide learning capacity and ability for generalization, on the other side, fuzzy logic provides a logical reasoning based on inference rules. So, in applications where information comes from different sources, such as numerical data and data that present characteristics as imprecision, uncertainty, subjectivity, etc., a neuro-fuzzy model is recommended in order to take the maximum profit of the available information, and to try to incorporate all the possible knowledge related to the problem.

The neuro-fuzzy approach constitutes a way to find, in a heuristic way, the parameters of a fuzzy model, through data processing submitted to a training algorithm. The implementation of a neuro-fuzzy model always must enable the supervision and the interpretation of the learning process and, as with neural networks; the success of the learning process is not guaranteed [2].

The most important reason to combine fuzzy systems and neural networks is the learning capacity of the latter, because such combinations have the ability to learn linguistic rules or membership functions, or to optimize the available ones. Learn; in this case, means to create a rule base or membership functions based on training with a set of data values presented to these models [3, 4]. In order to build a set of fuzzy rules, at least the initial membership functions must be defined. There are two very common approaches [5], one of them consists on the parametric description of the membership functions, parameters that must be optimized during the learning process, and the second proposal where a neural network is used to generate membership values according to the input data. The first option is used most.

2.2 Adaptive Neuro-based Fuzzy Inference System (Anfis)

The neuro-fuzzy selected architecture for the development of this work is an adaptable neuro-fuzzy network, called Anfis, which was developed by Jang in 1993 [1]. This architecture is functionally equivalent to a fuzzy inference system that can be built from the relations between input and output values of a data set. In this inference system, Anfis tunes the membership functions during the training process of the model.

Membership functions and fuzzy rules set that should be adapted to the problem, must be defined before training the Anfis model. For the initial estimation of these parameters it is possible to use a clustering algorithm, such as the Fuzzy c-Means (FCM) [6], the mountain method [7], the subtractive method (SCM) [8], or the expert's knowledge. Fuzzy rules are based on the Takagi-Sugeno inference method, and the conclusions are polynomial functions.

Due to the simplicity of the results interpretation and to its capacity of learning, Anfis is a good candidate for our classification problem. Besides, we are also interested in using the provided information and knowledge, not necessarily expressed in a numerical way, but through logic rules and values that leads to a better understanding of the system behavior.

The data set have the class of each example, reason to do a supervised learning and with disregard of the clustering algorithms for the preliminary estimation of the data set classes. Under such conditions we propose two approaches for the initial estimation of the membership functions. The first approach is related to the characteristic space of each class, and the second approach depends on the mean value and the standard deviation of each one of the classes.

2.3 Subtractive clustering method (SCM)

When there is a data set available where the number of clusters and their centers are unknown, it is very common to use the subtractive method to estimate them [8]. The subtractive clustering method (SCM) is fast and it is based on a similar idea of the mountain clustering method; as both divide the characteristic space of data through a grid, and intersections are a set of candidates that belong to a given cluster. The calculus of the cluster center is based on the data set density [7, 8]. Even if the mountain clustering method is simple and effective, the complexity of calculus grows exponentially with the dimension of the problem, as a result of the density function evaluation over all the points of the grid.

The calculus of a cluster center is based on the density of the data set, and it is estimated by a potential value P_i according to Eq. (1) for data w_i, $i = 1, .., N,$ in a D dimensional space. Each data point has the values of the input and output variables, that means, w_i is equal to the ordered pair (x_i, y_i), where x_i represents the p input variables, and y_i the $(D\text{-}p)$ output variables.

$$P_i = \sum_{j=1}^{N} e^{-\alpha \left\| w_i - w_j \right\|^2} \tag{1}$$

In Eq. (1), α is equal to 4 divided by r squared, where r is the radio that defines the neighborhood of w_i, and the symbol $\|.\|$ represents the Euclidean distance.

2.4 Maximal and minimal method (MM)

A drawback of the subtractive method is that it identifies good centers for the membership functions, but it only identifies one value for all dispersions. In such conditions, profiting from the knowledge of the class for each data point, two approaches are proposed as a way to estimate the dispersion of each individual membership function and to look for better results.

The first proposal takes into account the characteristic space where data are defined, and it proposes to use the maximum and minimum data points in order to calculate the centers and the dispersion of the membership functions, as shown by Eq. (2) and Eq. (3) respectively, where w_i represents the data of the *i* class.

$$c_i = \left[\min(w_i) + \max(w_i)\right]/2 \tag{2}$$

$$\sigma_i = \left[\max(w_i) - \min(w_i)\right]/2 \tag{3}$$

2.5 Statistical method (SM)

This approach is based on the calculus of the mean value and standard deviation of each class, as the initial estimation of the membership function parameters of the Anfis model. These values are calculated according to Eq. (4) and Eq. (5) for the centers c_i and standard deviation σ_i for a set of *n* data of the class *i*.

$$c_i = \frac{1}{n}\sum_{j=1}^{n} w_j \tag{4}$$

$$\sigma_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(w_j - c_i\right)^2} \tag{5}$$

2.6 AGUA data set

As a previous step to learn how to classify the different water outputs, it is necessary to dispose of a data set. A study in a house was then required. This was done in a house with three inhabitants. The collected data are in a data set called AGUA, which consists of seven classes, corresponding to a shower, a sink, a washbowl, a washing machine and three toilets; even if users intervention produces subclasses in two classes (the washing machine and the shower), leading to a total of ten classes in this case. Under these conditions the data set can be considered with seven classes or a data set with ten classes. Each data was obtained with measurement with a maximum precision of ±125 ml.

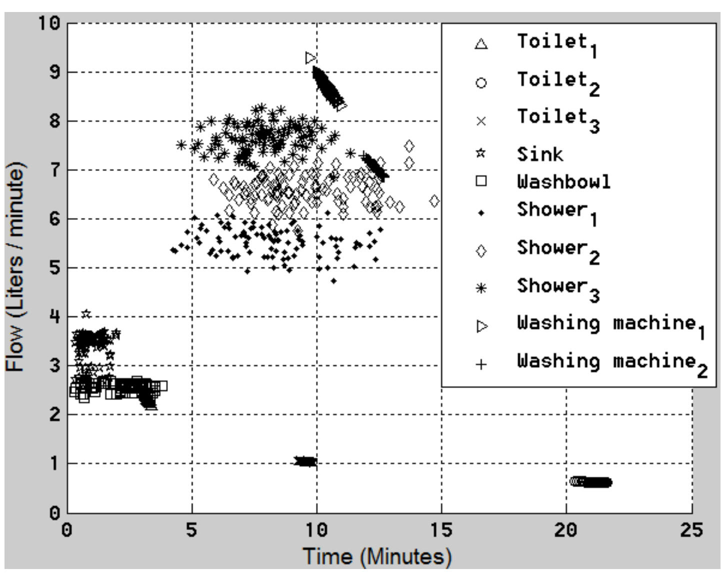


Figure 1. AGUA data set.

The data set is limited, because it only contains data about an individual house. However, we had tried to maintain conditions such that consumptions remain representative of the consumption habits of the family, and those ones of medium level families, as these are the kinds of families prevalent in our city. One important observation about data is that when consumption is fixed, as in the case of toilet and washing machine, the classes are more compact. This situation is immediately affected when consumption depends on humans, as in the shower, the most remarkable case, and the washbowl and the sink.

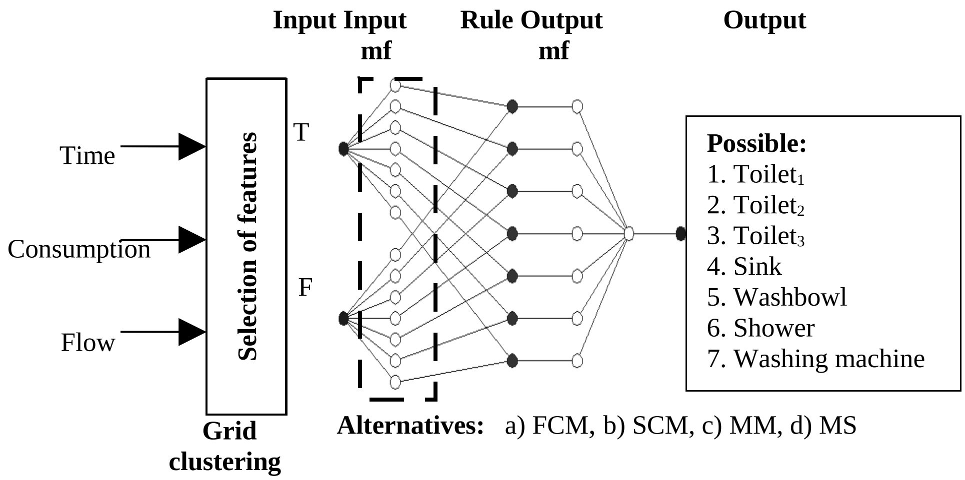
So, the most important in this work is to identify the consumption characteristics in houses, as the conditions and people change from one house to another. For example, the output pressure of water, the state of the hydraulic installations, the use of dispositive to control water, the consumption habits, etc. Nevertheless, even if great differences can be found, the trained classifier for each particular case is able to correctly identify the output according to the consumption variables (flow and time).

Figure 1 shows the data set. The corresponding classes are: t_1, t_2, t_3, sink, washbowl, S_1, S_2, S_3, WM_1 and WM_2, where S_1 to S_3 represents the same class but different habit of consumption, whereas WM_1 and WM_2 are the washing machine subclasses produced by user intervention at the beginning of the washing cycle.

3. OUTPUT CLASSIFICATION OF WATER

3.1 Classification

Due to great differences in consumption in different locations and activities, the characteristic space presents big regions without data. This conducts us to select not bounded membership functions, such that all the space is covered and the classifier has the possibility to recognize data points beyond the frontier of the available data. So, a Gaussian function was selected to represent the fuzzy sets of the model, and it has the advantage that it only needs the calculation of the center and dispersion for each membership function and it facilitates the analytical analysis of the model when necessary. The diagram of the proposed classifier is appreciated in Figure 2.



mf: membership function

Figure 2. Classification system for water consumption.

3.2 Input membership functions

There are a lot of works that propose a solution to the problem of finding input membership functions; as can be seen in Figure 2, where 4 alternatives are considered. One of them is based on the FCM algorithm [9, 10]. This method generates fuzzy inference system structure with radii of 0.5 to find the number of clusters, but results greatly depend on the initial state. With the MM and SM alternatives both, individual centers and dispersions for each membership function can be calculated, as was earlier explained.

Once the classifier is trained, some membership functions of the Anfis model must be changed in order to have better recognition of new data, especially when they are between classes. In this work, a Gaussian is changed by a bounded Gaussian (Gaussian2), a bell or a triangular function. Figure 3.b shows the approximation of a Gaussian by another function.

3.3 Fuzzy rules

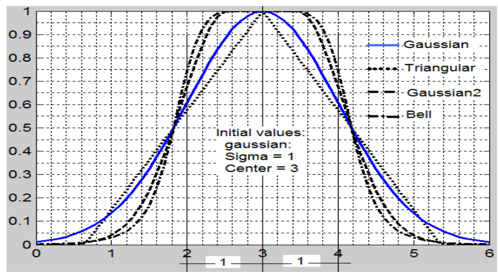
Building fuzzy models from classification or clustering results generally results in a quantity of rules equal to the number of classes or clusters. Examples of fuzzy rules describing relations between inputs and output are given in Figure 3.a, where t_i and F_i represents the membership functions of the inputs, and $class_i$ the class or the cluster of the conclusion of rule *i*.

1. If time is t_1 and flow is Fa_1 Then Output is $class_1$
2. If time is t_2 and flow is Fa_2 Then Output is $class_2$

...

n. If time is t_n and flow is Fa_n Then Output is $class_n$

(a)



(b)

Figure 3. a) Fuzzy rules. b)Approximation of a Gaussian by another kind of function.

3.4 Input membership functions optimization

The methods used for membership optimization were the backpropagation and a hybrid algorithm; with a gradient method for the antecedent parameters estimation, and a least squares method for the conclusion estimation. Both methods were applied to the initial values provided by the SCM, the MM or the SM methods. So, once the model was trained, the RMSE criterion and the total test errors of the classifiers were registered. This last criterion is used most.

4. RESULTS

This section presents the classification results of the AGUA data set, using 7 and 10 classes. Results are based on total errors, and the RMSE is used only when two or more classifiers give the same result and we need to select one of them. Table 1 shows test results, indicating data used for training and testing of the classifiers. The interest in using both criteria, total of errors and RMSE, is that they show a different point of view of the models, and they allow a better selection of the classifier and, particularly, the selection of the model that better generalizes.

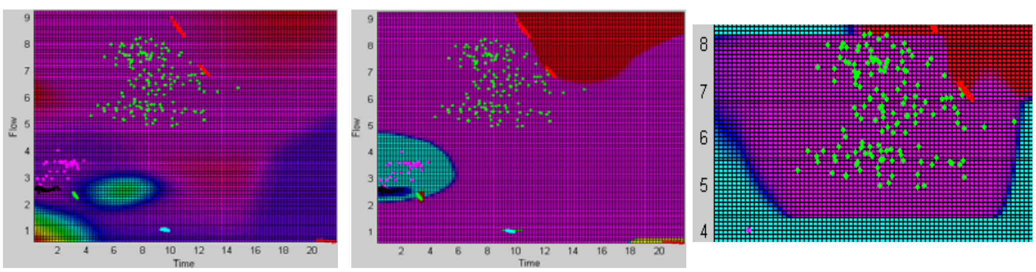
AGUA data set	Training-test data	Output of the Anfis model	Initial estimation of the MPs	Optimization method	Iterations	MPs	RMSE	Total of errors
7 classes	Odd-even	Lineal	SCM	H	112	g	0.193	9
		Lineal	MM	H	58	g2	0.127	8
		Constant	SM	R	385	g2	0.107	5
	Even-odd	Lineal	SCM	H	663	g	0.124	7
		Constant	MM	R	1000	g2	0.116	7
		Constant	SM	R	1000	g2	0.108	7

10 classes	Odd-even	Lineal	SCM	H	101	g	0.431	24
		Lineal	MM	H	50	g	0.311	13
		Constant	SM	R	18	g2	0.267	12
	Even-odd	Constant	SCM	H	84	g2	0.407	44
		Constant	MM	R	353	g2	0.252	10
		Constant	SM	R	1105	g2	0.244	8

Table 1. Total test errors.

According to Table 1, the best results correspond to the Anfis model with constant output, initial estimation of membership functions with the SM approach, optimization with the retropropagation method, and Gaussian2 membership functions. Table 1 also presents the number of iterations needed for training each classifier. However, it is necessary to make an evaluation in order to find a good balance between total errors and learning speed, as the better classifier can not have sufficient time for an application in real time.

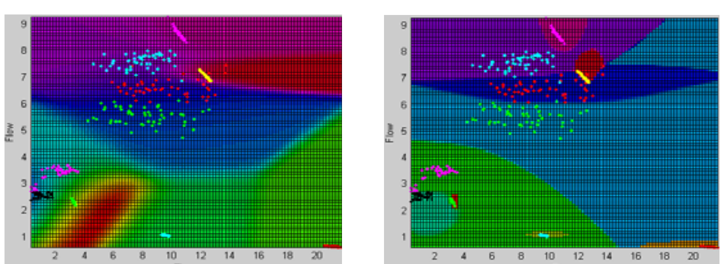
In Figure 4, for seven classes, and in Figure 5, for ten classes, are the results of two Anfis models, one with linear conclusions (L) in the rules, and the other with constant (K) conclusions, but initialized with the subtractive and the SM methods. Figure 4.e shows how the shower class was limited, as this class defines all the space without data. This change on the classifier behavior results from the change of membership functions.



(a) (b) (c)

Figure 4. Classification zones of: a) $Anfis_L$SCM_H_11. b) $Anfis_L$SM_R_901. c) Modified shower class.

As can be seen in Figure 4.d, all classes are well delimited for 7 classes except the shower class, which defines its class and the rest of the characteristic space where there are no data. A possible solution is to limit this class, which can be done by changing the kind of membership functions representing the shower and sink classes. For example, using a bell function instead of a Gaussian2 function, except for the shower class in the flow variable where the change was for a triangular function. The result is shown in Figure 4.e. For the AGUA data set with 10 classes, there is no need to modify the membership functions, as the whole characteristic space is better divided. This can be seen in Figure 5.



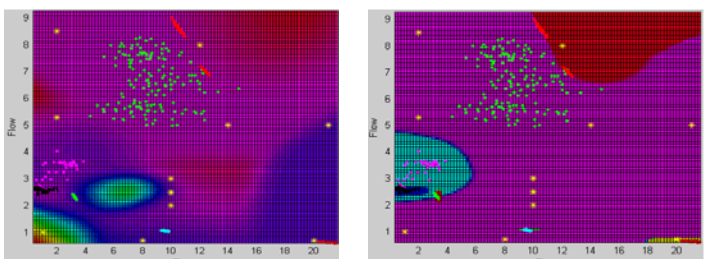
(a) (b)

Figure 5. Classification zones of: a) $Anfis_K$SCM_H_84. b) $Anfis_K$SM_R_1105.

Table 2 contains the classification results for 12 new data with the AnfisSCM and AnfisSM models. Figure 6 shows the relative position of these examples to the classes of each classifier. As can be seen, these new examples can be considered as noisy data as none of them is included in the cloud of points of the data set. Besides, for the test results both classifiers give few errors as shown in Table 1. So, in Table 2 it is given the estimated class for each example, and the calculated class with the classifiers. The classifiers identify some examples well but most of the characteristic space is without data, and there are some errors too. According to these results, it is interesting to apply a method in order to identify normal cases from data that represents different behaviors or noise. A possibility is to use fuzzy logic and GKPFCM clustering algorithm, as proposed in [11], such that well known consumption behaviors can be differentiated from no common data. This study will be done in a future work.

Examples	Inputs values		Class		
	Time (min)	Flow (L/min)	Estimated	Calculated	
				$AnfisSCM_7$	$AnfisSM_7$
1	12	8	7	7	7
2	8	0.7	5	2	6
3	21	5	6	3	6
4	1	1	5	1	6
5	2	5.3	6	7	6
6	2	8.5	6	7	6
7	10	1.2	3	4	6
8	10	2	5	5	3
9	10	2.5	5	4	6
10	10	3	4	6	6
11	14	5	6	6	6
12	20	0.7	2	2	2

Table 2. Classification results of new data with the $AnfisSCM_7$ and $AnfisSM_7$ models.



(a) (b)

Figure 6. Relative position of new data with: a) AnfisSCM. b) AnfisSM.

5. CONCLUSIONS

In this work we have proposed and applied a classification system to water consumption identification in houses. The outputs of the classifiers are three toilets, a sink, a shower, a washbowl, and a washing machine. As it can be seen, the selected and trained classifiers achieve a good recognition percentage that in the worst case was above 91%, or 44 errors from 500 data. So, they can be considered a good solution for this kind of applications. In this way, such a classifier could be of great help for a supervising system, able to help in finding the output where water is consumed at every moment. An integral system could be developed looking for more efficient use of water.

6. REFERENCES

[1] J. S. R. Jang, "Adaptive-Network-Based Fuzzy Inference Systems", *IEEE Trans, on Systems, Man and Cybernetics*, Vol. 23, No. 3, 1993, pp. 665-685.

[2] L. A. Zadeh and J. Kaeprzyc, editors. (1999) *Computing with words in Information/Intelligent Systems 1 (Foundations).* Volume 33 of Studies in Fuzziness and Softcomputing. Phisyca-Verlag

[3] D. Nauck and R. Kruse, "Designing Neuro-Fuzzy Systems Through Backpropagation," W. Pedrycz (ed.), *Modelling: Paradigms and Practice*, Kluwer, 1996, pp. 203-228.

[4] D. Nauck and R. Kruse, "Neuro-Fuzzy Systems Research and Applications outside of Japan," M. Umano, I. Hayashi and T. Furuhashi (eds.), *Fuzzy-Neural Networks, Soft Computing Series*, 1996, pp. 108-134.

[5] D. Nauck, F. Klawonn and R. Kruse, *Foundations of Neuro-fuzzy Systems*, Wiley, 1997.

[6] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[7] R. Yager, D. Filev, "Generation of Fuzzy rules by. Mountain-Clustering", J*ournal of Intelligent and Fuzzy. Systems*, vol. 2, no. 3, pp. 209-219, 1994.

[8] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation", *J. of Intelligent & Fuzzy Systems*, Vol. 2, No. 3, 1994.

[9] S. H. Park, K. J. Lim, S. H. Kang, J.M. Seo and Y. G. Kim, "Comparison of Classification Rate Between BP and ANFIS with FCM Clustering Method on Off-line PD Model of Stator Coil," *KIEE Int. Trans. on Electrophysics and Applications*, Vol. 5-C, No. 3, 2005, pp. 138-142.

[10] K. Tutuncu, N. Allahverdi, "Reverse modeling of a diesel engine performance by FCM and ANFIS", *ACM International Conference Proceeding Series*; Vol. 285. Proceedings of the 2007 International Conference on Computer Systems and Technologies. Bulgaria, 2007.

[11] B. Ojeda Magaña, R. Ruelas, F. Buendía Buendía, D. Andina. "Greatter knowledge extraction based on fuzzy logia and GKPFCM clustering algorithm", *Proc. Of the 6th Int. Conf. on Computational Intelligence, Man-Machine Systems and Cybernetics*, WSEAS, Tenerife, Spain, December 14-16, 2007, pp. 48-52.