Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Identification of domestic water consumption in a house based on fuzzy clustering algorithms

M. A. Corona-Nakamura *, R. Ruelas †, B. Ojeda-Magaña †, D. W. Carr Finch †.
* Departamento de Ciencias Computacionales-CUCEI
Universidad de Guadalajara.
Revolución 1500. C.P. 44840 Guadalajara, Jalisco, México.
†Departamento de Ingeniería de Proyectos-CUCEI
Universidad de Guadalajara.
José Guadalupe Zuno No. 48. C.P. 45101 Zapopan, Jalisco, México.

***Abstract*— This work presents the classification of different types of consumptions of water in a house (sinks, showers, washing machines etc.). This classification takes into account the measured flow and the duration of the flow at a particular point in the water distribution network. The classifier uses the FCM and Gustafson-Kessel algorithms. The data set is called AGUA and it corresponds to real data gathered for a research project in Guadalajara, Mexico. The classifier was trained in an unsupervised way. As such, it learns the patterns for the flow and duration of flow, for each type of consumption. The identified classes are relevant consumption types such as using the sink, using the shower, etc. The results show that the proposed approach gives good results, with 91.6 % of the examples classified correctly, and it could be used in the future as part of a supervisory system in order to make better use of water in households.**

Keywords: Domestic water consumption, clustering algorithms, unsupervised learning, classification.

I. Introduction

Due to climatic changes around the world, it is very important to make better use of water in households. Nowadays, more than ever, we must conserve water as much as possible since there are many places around the world living in critical situations as a consequence of water shortages [1]. These conditions are going to continue and, as population increases, as well as the number of industries, and the climate is changing, a greater quantity of humans will soon be in these precarious conditions for water. It should be noted that other research has been developed looking for some way to regulate water consumption [2], [3], [4].

For this reason, we are proposing that from flow and duration of flow measurements at the input of a household water distribution network, it is possible to automatically identify the place where water was consumed at a particular moment. So, the resulting model could be the basis for the development of a supervisory system in order to achieve water conservation.

The problem of identification of the point of water consumption is that it can vary significantly from one house to another, because it depends on the number of outputs, type of construction, kinds of furniture and devices used. Besides uses, the habits and number of people living in each house, which could have a very significant influence on consumption. So, for the model used the possibility of unsupervised learning is important, in order that it can be adapted to each particular case automatically. This was our motivation to develop a system capable of identifying the point where water is consumed at a particular moment and to qualify the event as a typical or atypical consumption. In the case of atypical consumption, the system must provide recommendations to the user in order to conserve water. This system requires a model such that it can learn consumptions for each point of consumption such as the $Sink$, the $Washing - Machine$, ..., from duration and flow values provided by measurements at the input of the water distribution network of a house. In such a way, the model will allow us to automatically identify the point where water is consumed at a particular moment. As such, the purpose of this work is limited to the development of a classifier with unsupervised learning, such that it could easily be applied to different houses with different conditions; such as the kind and number of points of usage. This paper presents the results of the classifier based on the FCM [5], [6] and the Gustafson-Kessel [6], [7], [8] algorithms applied to the AGUA data set.

In this work we present a study of water consumption in a house. The AGUA data set was obtained during one month for a particular house with three occupants. However, the patterns of consumption are particular to that house, and, for some examples, for those persons. The data set was built according to this house and the available points of usage in this house. Fig. 1 shows distribution of data for seven total points of usage; even if human behavior produces two subclasses in one point of usage and three subclasses in another point of usage for the particular studied case.

The structure of this work is defined in the following way. Section II presents the AGUA data set and the fuzzy clustering algorithms used for cluster identification. Section III contains the results of learning and tests of the model, and

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

Section IV presents the primary conclusions of this work.

II. AGUA DATA SET AND CLUSTERING ALGORITHMS

In this Section we present the AGUA data set and some details about it. Next, we introduce a brief review of unsupervised learning, the FCM algorithm [5], as well as the Gustafson-Kessel algorithm [7], [8]. The results of the classifier are given in Section III.

The clustering algorithms are used in the case of unsupervised learning, that is, they try to identify the correct number and shape of clusters from the characteristics of a data set. All data are classified in one of the clusters, such that every cluster is more or less homogeneous and distinct from the others.

Each cluster is represented by one point called prototype or group center. So, based on a distance measure, the nearest points to a prototype are the elements that identify each cluster. The Euclidean measure is the most popular, though the shapes of the clusters identified using this measure are not well adapted to the natural distribution of data, because it only allows spherical forms. Other distance measures have been proposed, such as the Mahalanobis distance, trying to provide more flexibility to the clustering algorithm. The Mahalanobis distance, for example, gives the possibility of finding ellipsoidal shapes for the clusters.

Some of the best known clustering algorithms such as the *c-medoids* [9], the *k-means* (hard c-means) [10], the *Fuzzy c-Means* (FCM) [5], the *Possibilistic c-Means* (PCM) [11], *Gustafson-Kessel* (GK) [7], and the *Gath-Geva* (GG) [12] are iterative algorithms. More about clustering algorithms can be found in [13]. The GK algorithm is based on an adaptive distance norm, such that the clusters are well adapted to the natural distribution of data. In this work the FCM and an improvement of the GK algorithms are used with the AGUA data set, which are presented in the next two subsections.

A. AGUA data set

The AGUA data set was obtained in a house inhabited by three persons. This house has seven points of water consumption. These points are: three $Toilets$, one $Sink$, one $Shower$, one $Washbowl$, and one $Washing - Machine$. However, human behavior has produced three subclasses in the $Shower$ and two subclasses in the $Washing - Machine$ outputs, so data set can be considered with a total of 10 classes; the $Shower$ can be divided in three subclasses due to preferences and habits of the human beings, and the $Washing - Machine$ in two sub-classes as result of the user intervention at the beginning of the washing cycle. Fig. 1 shows the total distribution of data. The data set consists of 1000 examples; 100 for each one of five classes and 100 for each one of five subclasses [14].

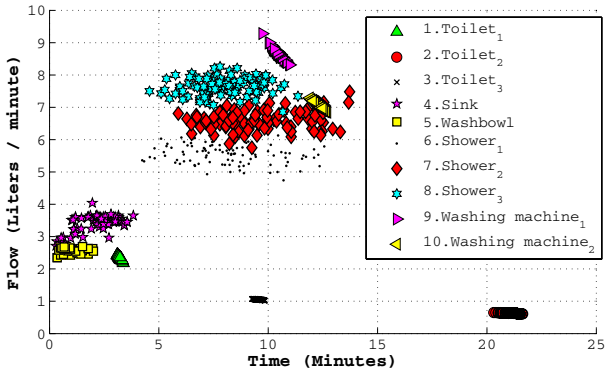

Fig. 1. AGUA data set

Data set is limited, because it only contains data about an individual house. However, we have tried to maintain conditions such that consumptions remain representative of the consumption habits of the persons living in this region of Mexico, and families with a median level of income, as these are the kinds of families prevalent in this area. One important observation about data is that when consumption is fixed, as in the case of $Toilets$ and $Washing - Machine$ (WM), the classes are more compact. This situation is immediately different when consumption depends on humans, as in the shower, the most remarkable case, and the washbowl and the sink. Each datum was obtained by measurement with a maximum precision of ± 125 ml.

The AGUA data set has two input characteristics (duration and flow) and the corresponding points of usage which are: $Toilet_1$, $Toilet_2$, $Toilet_3$, $Sink$, $Washbowl$, $Shower_1$, $Shower_2$, $Shower_3$, WM_1 and WM_2, where $Shower_1$ to $Shower_3$ represents the same output but a 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.

B. Unsupervised learning

Fig. 1 shows the classes for each point of consumption or behavior. So, the next step is to decide what kind of model and learning approach we want to use, in order to build a model that help us to automatically identify the point of consumption and behavior when water is consumed. For this application a supervised learning approach was applied in a previous work with an *Adaptive Neuro-Fuzzy Inference System* (ANFIS) model [15], [16]. However, we are interested in unsupervised learning as this greatly simplifies learning for new houses and for new families. Unsupervised learning only uses the input characteristics and tries to identify some groups, characterized by their prototypes and their shapes. The performance of any fuzzy clustering method is improved when the number of clusters is known a priori [17]. Under these conditions, the output available for each consumption

in the AGUA data set, is only used in the test stage of the classifier.

As can be seen in Fig. 2, some data have a high correlation, near one, and they are arranged almost as a straight line. The GK fuzzy clustering algorithm, including the improvement proposed by Babuska and that we denote as GKM, is a good solution for the identification of the groups in this data set. Also, as a good way to establish the initial values for the GKM, the FCM is used and it provides the initial center of groups for the GKM. Both algorithms allow a good identification of the groups, as Fig. 2 shows, as well as the results given in the next section.

As the form of the classes is easily approximated by ellipses, the FCM algorithm was selected for the initial estimation of the class centers, and the GKM algorithm for the classifier; these algorithms provide good results as can be seen in Section III. Fig 2 shows the identified classes.

Moreover, due to great differences in consumption between different fixtures and activities, the characteristic space presents big regions without data. This leads us to select unbounded 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. In fact, some examples are considered in Section III and a threshold helps us to distinguish between data that is similar or equal to those of the AGUA data set and those from unusual or atypical data, which means, new data that are in regions where no examples exist and the classifier did not learn. So, a Gaussian function was selected to represent the fuzzy sets of the model, and it has the advantage that it only needs to calculate the center and dispersion for each membership function and it facilitates the analytical analysis of the model when necessary.

C. Fuzzy c-Means clustering algorithm

The FCM [5] is an algorithm where each data $z_k, k = 1, \ldots, n$, has a membership degree to every fuzzy set $A_i, i = 1, \ldots, c$. These membership degrees, noted $\mu_{A_i}(z_k)$ for the fuzzy set A_i and data z_k, can be simplified such that $\mu_{A_i}(z_k)$ can be written as μ_{ik}, the elements of the $U = [\mu_{ik}]$ matrix, with values in the interval [0, 1]. The FCM is based on the minimization of the objective function J_{FCM} given by (1).

$$J_{FCM}(Z; U, V) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ik}^{m}(z_k) \|z_k - v_i\|^2 \tag{1}$$

where $U = [\mu_{ik}] \epsilon M_{FCM}$, is the matrix that contains the membership degrees of the fuzzy partition of the Z data set; $V = [v_1, v_2, \ldots v_c]$ is the vector of prototypes or centers of the clusters; and $m \subset [1, \infty]$, is a factor that defines the degree of fuzziness of the partition.

Theorem FCM [5]: If $D_{ikA_i} = \|z_k - v_i\|_{A_i} > 0$ for all i and k; $m > 1$, and Z contains at least c distinct points, then $(U, V) \in M_{FCM} X \Re^{cn}$ may minimize J_{FCM} only if

$$\mu_{ik} = \left(\sum_{j=1}^{c} \left(\frac{D_{ikA_i}}{D_{jkA_i}} \right)^{2/(m-1)} \right)^{-1} \tag{2}$$
$$1 \leq i \leq c; \quad 1 \leq k \leq n$$

$$v_i = \frac{\sum_{k=1}^{n} \mu_{ik}^{m}(z_k)}{\sum_{k=1}^{n} \mu_{ik}^{m}}, \tag{3}$$
$$1 \leq i \leq c.$$

The most popular algorithm for the minimization of equation (1) is the Picard iteration, also known as *Alternating Optimization* (AO), through equations (2) and (3). This one consists of the estimation for each cycle of estimates for $V_{t-1} \Rightarrow U_t \Rightarrow V_{t-1}$, and it stops when the criterion $\|V_t - V_{t-1}\|_{err} \leq \varepsilon$ is satisfied.

D. Gustafson-Kessel clustering algorithm

The GK algorithm [7], [8] is an extension of the FCM algorithm. The GK algorithm employs an adaptive distance norm in order to detect clusters with different shapes. Contrary to the FCM, the GK algorithm gives clusters with different sizes in the different dimensions (clusters are ellipsoids not circles). In this case every cluster has its own norm matrix $\mathbf{A}_i$ for each fuzzy cluster A_i. The matrices $\mathbf{A}_i$ are an optimization variable for the distance calculus according to (4) and included in the criterion given by (5).

$$J_{GK}(Z; U, V, \mathbf{A}_i) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ik}^{m}(z_k) D_{ik\mathbf{A}_i}^{2} \tag{4}$$

$$D_{ik\mathbf{A}_i}^{2} = (z_k - v_i) \mathbf{A}_i (z_k - v_i)^T \tag{5}$$

Though the objective function (4) cannot be minimized directly because J_{GK} depends linearly on $\mathbf{A}_i$. It means that J_{GK} can be made as small as the matrices $\mathbf{A}_i$ which are defined positive. Therefore $\mathbf{A}_i$ is constrained to (6), where ρ_i represents the volume of the i-th cluster.

$$\det |\mathbf{A}_i| = \rho_i \tag{6}$$

The following result is obtained:

$$\mathbf{A}_i = [\rho_i \det(F_i)]^{\frac{1}{n}} F_i^{-1} \tag{7}$$

$$F_i = \frac{\sum_{j}^{n} \mu_{ik}^{m}(z_k)(z_k - v_i)(z_k - v_i)^T}{\sum_{j=1}^{n} \mu_{ik}^{m}(z_k)} \tag{8}$$
$$1 \leq i \leq c$$

where n is the quantity of data, and F_i is the fuzzy covariance matrix of the i-th cluster. Babuska *et al* [8] have proposed an improvement for the calculus of the covariance matrix (8). This improvement avoids the proximity of the matrix to be singular when F_i is inverted, a common situation with small quantities of data. So, F_i is calculated according to (9), where a scaled identity matrix is added, and (10).

$$F_i = (1-\gamma)F_i + \gamma \det(F_0)^{\frac{1}{n}} I \quad (9)$$

Once the new value of F_i is calculated, it is necessary to determine its eigenvalues λ_{ij} and eigenvectors ϕ_{ij}, find $\lambda_{ijmax} = max_j \lambda_{ij}$, set $\lambda_{ij} = \frac{\lambda_{ijmax}}{\beta}$ for all j such that $\lambda_{ij} = \frac{\lambda_{ijmax}}{\lambda_{ij}} > \beta$, and finally reconstruct F_i by (10).

$$F_i = [\phi_{i,1}, \ldots, \phi_{i,n}] diag(\lambda_{i,1}, \ldots, \lambda_{i,n}) [\phi_{i,1}, \ldots, \phi_{i,n}]^{-1} \quad (10)$$
$$1 \leq i \leq c$$

An additional advantage of the GKM algorithm is its data scale independence, i.e. if data in any of their dimensions is multiplied by a constant, then the relative coordinates of the cluster centers and the matrix of membership degrees are identical. An inconvenience of this method is that it converges locally only. In the next section we show that this method is a good option for identification of the groups in the data set, except for clusters with important intersections where more flexibility on the shape of the clusters is desirable.

III. Test results with the AGUA data set

This section shows the results using the GKM algorithm after the unsupervised learning was carried out on the AGUA data set. The start point of this algorithm is the initial values provided by the FCM concerning the estimation of the class centers. Next, only the characteristics of a subset of the AGUA data set are used to achieve a better estimation of the clusters and their centers, as is for unsupervised learning. The AGUA data set has 10 classes with some of them very separated from the others, but some classes with an important intersection. Besides, the classifier has recognized data in an appropriate way even considering a threshold of 0.3.

In Fig. 2, 3 and 4 data were scaled to the interval [0, 1], and the prototypes were represented by solid circles, and level curves are in fact the membership degrees corresponding to each cluster. Fig. 2 contains even data and the clusters identified with the GKM algorithms; initially the odd AGUA data set is used for training. Fig. 3 contains odd data and the clusters identified with the GKM algorithm; here, the even AGUA data set is initially used for training.

Table I shows these test results, indicating data used for training and test of the classifier. Once the model was trained with the odd data from AGUA, it reached 91 % level of correct recognition. Once the classifier was trained with the even data AGUA, it reached a 92.8 % level of correct recognition. The biggest problem for good identification of data is the region that has the most important intersection between classes. For example, the $shower_2$ class with $shower_1$, $shower_3$ and WM_2 classes.

Table I shows a high recognition percentage of the classifier.

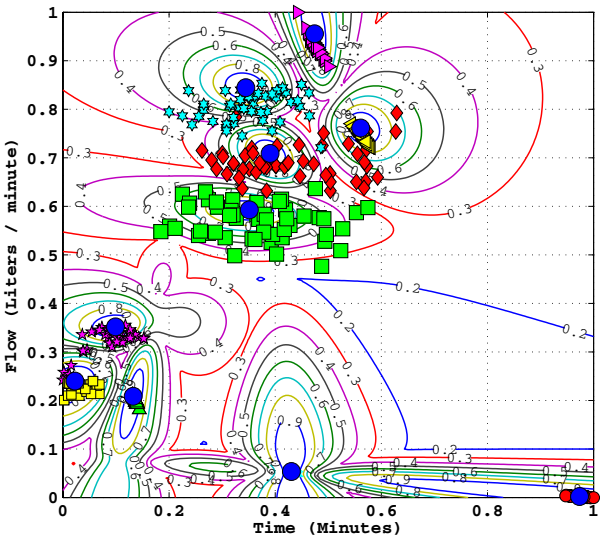

Fig. 2. Training with odd and test with even data.

TABLE I
PERCENTAGE OF DATA CORRECTLY RECOGNIZED.

Class number	Class	Percentage of recognition Training (odd-even)	Testing (even-odd)
1	$Toilet_1$	100 %	100 %
2	$Toilet_2$	100 %	100 %
3	$Toilet_3$	100 %	100 %
4	Sink	78 %	86 %
5	Washbowl	100 %	100 %
6	$Shower_1$	94 %	88 %
7	$Shower_2$	62 %	62 %
8	$Shower_3$	76 %	92 %
9	WM_1	100 %	100 %
10	WM_2	100 %	100 %
	Total:	91 %	92.8%

However, in order to probe with data located in regions where there are no data of AGUA, a set of 10 examples was artificially generated; the results concerning these examples are in Table II where 6 examples were classified as atypical. As we are using unbounded membership functions, every input data is classified in one of the 10 classes, no matter how small their membership degrees are, though this is hard to justify especially for data that are far away from the classes.

Fig. 4 shows these examples. A more logical option for these cases is to take into account a threshold such that, if the maximum membership degree of an example to the 10 classes is smaller than the threshold, the example can be considered as atypical. For the results of Table II we take 0.3 as the threshold, but taking 0.2, for example, produces only one atypical example. From the results of Table II this last value is a better option. However, looking at Fig. 4, 0.3 allows maintaining better classes with regard to the AGUA data set.

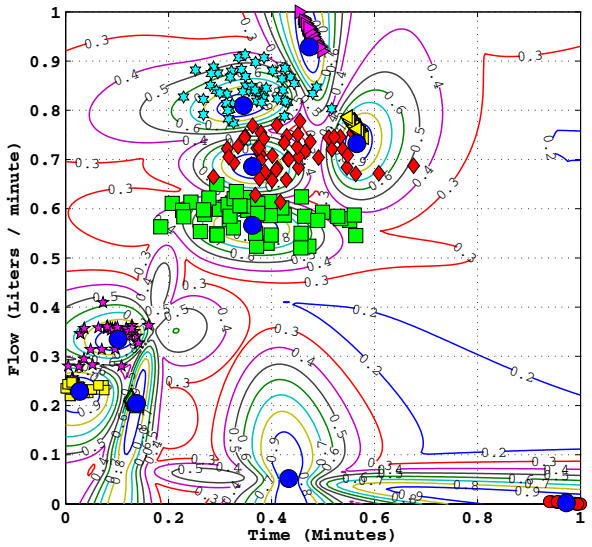

Fig. 3. Training with even and test with odd data.

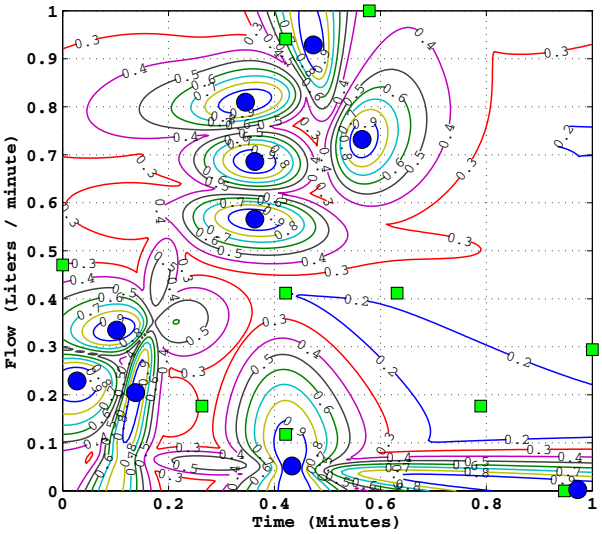

Fig. 4. Results of test with new data.

TABLE II
RESULTS OF TEST WITH 10 NEW EXAMPLES.

Examples	Inputs values Time (min)	Flow (L/min)	Class
1	10.0	9.5	9
2	13.0	10.0	Atypical
3	07.0	3.0	Atypical
4	21.0	4.0	Atypical
5	02.0	5.5	4
6	10.0	5.0	Atypical
7	10.0	2.5	3
8	17.0	3.0	Atypical
9	14.0	5.0	Atypical
10	20.0	1.5	2

So, we kept this value. Fig. 4 shows the relative position of these examples to the classes generated by the 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. The results of this section show that the classifier, using unsupervised learning, achieves a high recognition percentage of 91 %. The only disadvantage is that it finds ellipsoidal clusters which do not have the best decision edges in regions where important intersections between classes exist, as for the shower and washing machine subclasses.

IV. CONCLUSIONS

The purpose of this work was to show the results of a classifier, based on unsupervised learning, such that it is able to identify for each event, from duration and flow, the point of consumption where water was consumed in a house. The FCM and GKM algorithms work appropriately except for the region with important intersection of the subclasses, where all the classification mistakes occurred. Besides, due to the unsupervised learning, it is possible to apply this solution automatically to houses even with different conditions.

According to the results, the classifier reaches a 91% and 92.8 % level of correct recognition, which can be qualified as a good result, particularly if this could be used as an important element for a supervisory system acting as an aid for users to achieve better water conservation. However, the goal remains to get better results for a classifier based on an unsupervised learning as results are much better using supervised learning [16]. This can be done by combined classifiers or making better use of typicality, although it is also necessary to increase the data set to other houses and other family types with different behaviors. These results will be showed in forthcoming works. Finally, this research enables us to measure the water in just one place, as measuring in many locations, for many reasons, would not be feasible.

V. ACKNOWLEDGMENTS

This research has been partially supported by University of Guadalajara under projects 60789 and 76817.

REFERENCES

[1] Corona-Nakamura M. A; Ruelas R; Ojeda-Magaña B. Problemática relativa al uso y consumo de agua. Technical report, SEP-No.Reg.-03-2007-120515105500-01, 2007.

[2] *Kendall, A., Gaines III, B.: Computer-implemented method and system for estimating facility water consumption, United States Patent 20050015210, January, 2005.*

[3] Sahely H; Kennedy C; Adams B. Developing sustainability criteria for urban infrastructure systems. *Canadian Journal of Civil Engineering*, Vol. 32, No 1:72–85, 2005.

[4] *Kennedy S; Reducing Domestic Water Consumption -A Rainwater-Greywater Hybrid Recycling System in Liverpool, England, The Management School, The University of Sheffield, 2007.*
[5] Bezdek J. C. *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
[6] Balasko B; Abonyi J; Feil B. *Fuzzy Clustering and Data Analysis Toolbox For Use with Matlab*. PhD thesis, Department of Process Engineering University of Veszprem, Hungary, 2005.
[7] Gustafson E. E and Kessel W. C. Fuzzy clustering with a fuzzy matriz de covarianza. *Proceedings of the IEEE CDS, San Diego CA*, pages 761–766, 1979.
[8] Babuska R; van der Veen P.J; and Kaymak U. Improved covariance estimation for gustafson-kessel clustering. *International Conference on Fuzzy Systems*, pages 1081–1085, 2002.
[9] Kaufman L and Rousseeuw P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, 1990.
[10] MacQueen J. B. Some methods for classification and analyisis of multivariate observation. In *Proceedings of the Fifth Berkley Symposium on Methematical Statistic and Probability*, Unversity of California Press, Berkley, Vol 1, pages 281-297, 1967.
[11] Krishnapuram R. and Keller J. The possibilistic c-means algorithm: Insights and recommendations. *International Conference on Fuzzy Systems*, Vol 4, No 3:385–393, 1996.
[12] Gath I and Geva A. B. Unsupervised optimal fuzzy clustering. *IEEE Trans. Pattern Anal. Machine Intell*, pages 773–781, 1989.
[13] Höppener F; Klawonn F; Kruse R; and Runkler T. *Fuzzy Cluster Analysis, Methods for classification, data analysis and image recognition*. Chistester, United Kingdom, 2000.
[14] Corona-Nakamura M. A. *Desarrollo de un Modelo Neuro-difuso y su Aplicación en el Diagnóstico de Consumo de Agua de una Zona Urbana*. PhD thesis, Universidad de Guadalajara, December 2008.
[15] Jang J-S R; Sun C-T; Mizutani E. *Neuro-Fuzzy and Soft Computing: a computational approach to learning and machine intelligence*. 1997.
[16] Corona-Nakamura M. A; Ruelas R; Ojeda-Magaña B; Andina D. Classification of domestic water consummation using anfis model. In *World Automation Congress (WAC), Aloha, Hawii, USA*, 2008.
[17] Shah J. Z and Salim N. Fuzzy clustering algorithms and their applications to chemical datasets. In *Postgraduate Annual Research Seminar*, 2005.