Enhanced Equal Frequency Partition Method for the Identification of a Water Demand System

Antoni Escobet
Departament ESAII
Univ. Pol. Catalunya
Ed. MN2 - Campus Manresa
Av. Bases de Manresa 61-73
Manresa 08240, Spain
Phone: +34(93)877-7260
Fax: +34(93)877-7202
Toni@Bages.EUPM.UPC.ES

Rafael M. Huber
Inst. de Robtica i Inf. Ind.
Univ. Pol. Catalunya
Ed. NEXUS, Planta 2
Gran Capità, 2-4
Barcelona 08034, Spain
Phone: +34(93)401-5757
Fax: +34(93)401-5750
Huber@IRI.UPC.ES

Angela Nebot
Departament LSI
Univ. Pol. Catalunya
Mòdul C6 - Campus Nord
Jordi Girona Salgado, 1-3
Barcelona 08034, Spain
Phone: +34(93)401-5642
Fax: +34(93)401-7014
Angela@LSI.UPC.ES

François E. Cellier
Elect. & Comp. Engr. Dept.
University of Arizona
P.O.Box 210104
Tucson, AZ 85721-0104
U.S.A.
Phone: +1(520)621-6192
Fax: +1(520)621-8076
Cellier@ECE.Arizona.Edu

Keywords: Unsupervised partitioning, Fuzzy inductive reasoning, Water demand system.

Abstract

This paper deals with unsupervised partitioning. A first goal of this paper is to present an enhancement to the *Equal Frequency Partition (EFP)* method that allows to reduce, to some extent, the main drawback of this classical classification method, i.e. the data distribution dependency. A second goal of this work is to use the *Enhanced Equal Frequency Partition (EEFP)* method within the discretization process of the Fuzzy Inductive Reasoning (FIR) methodology for the identification of a model of a water demand system. It is shown that use of the EEFP method allows to obtain more accurate FIR models of the water demand system, reducing the prediction errors.

1 Introduction

The transformation of continuous variables into discrete variables is a common problem that arises in a large number of areas within the artificial intelligence field. The goal is to objectively partition the data into homogeneous groups in such a way that object similarity within a group and object dissimilarity between groups are maximized. Unsupervised partitioning assumes that the data is not labeled with class information. This is usually the case when dealing with dynamic features or variables. There exist a large number of unsupervised classification methods (Anderberg 1973; Bezdek *et al.* 1984; Li and Biswas 1999); one of the simplest being the equal frequency partition (EFP) technique. The EFP method has the advantage that it is extremely simple and, in a lot of cases, the data distribution obtained within the partitions or groups is quite reasonable. This method has been the one used most commonly in the discretization process of the Fuzzy Inductive Reasoning (FIR) methodology obtaining, usually, good results (Cellier *et al.* 1996; Nebot *et al.* 1996; Nebot *et al.* 1998). However, the EFP method is sensitive to data distribution, and good partitioning will only be obtained if the data distribution is more or less uniform in the sense that all possible behaviors of the system are represented with a comparable number of occurrences.

FIR, as all inductive modeling methodologies, is based on the data available from the system under study. Therefore it is necessary to have a rich amount of data representing all possible behaviors of the system in order to identify an accurate (optimal) model. If the data available from system observations represent all possible (physical) behaviors with a similar number of occurrences, then the use of the EFP method within the FIR methodology is indeed useful, and very good results are obtained by its use.

However, it can happen that although all possible behaviors are represented in the registered data, each has associated a different number of occurrences. For instance, it could be that a specific behavior of the system occurs frequently, and therefore, lots of data are registered of this situation. Some other behavioral pattern occurs rarely, and therefore, this situation is underrepresented in the data registered from the system.

The first goal of this paper is to present an enhancement to the EFP method to be used within the Fuzzy Inductive Reasoning methodology that allows to reduce, to some extent, the data distribution dependency. The second goal of this work is to use the Enhanced Equal Frequency Partition (EEFP) method within the discretization step of the FIR methodology for the identification of a model of a water demand system. The water distribution network carries water emanating from wells and rivers for human consumption in the city. It is required that the water arrives at the destination points with a certain pressure-flow. In the first part of the paper the EEFP method is described in detail, whereas in the second part, the water demand application is presented and the identification of FIR models is explained.

2 Enhanced equal frequency partition method

The equal frequency partition (EFP) method is undoubtedly one of the simplest classification methods available. It consist on distributing the system data into a predefined number of classes maintaining the same number of occurrences in each class. However, this method is sensitive to data distribution. In this section, a modification of the EFP method is proposed that exploits the advantages of the EFP technique while trying to reduce its drawbacks.

The idea behind the enhancement of the EFP method is simple. The EEFP method eliminates multiple observations of the same behavioral pattern determining if an observation is significantly different from another or not, then applies EFP to the remaining set of significantly different patterns to decide on a meaningful set of landmarks.

The EEFP method should take into account two relevant aspects. The first one is to decide which data values can be considered to be equal. In other words, it is required to define an interval, δ, that represents the set of observations that are similar and, therefore, that can be considered repetitions of the same occurrence. This is described graphically in figure 1.

The second aspect is to define the minimum number of similar observations required (samples that are inside the δ interval) in order to consider that this behavioral pattern is over-represented. This parameter, α, is also described in figure 1. If a number of similar observations greater than α is found in the data, redundant observations are eliminated. In contrast, if a set of similar observations with a number of elements lower than α is found in the data, all occurrences are kept.

As can be seen in the example of figure 1, all the values within the δ range are similar observations. α indicates the minimum number of occurrences necessary to assume that this behavioral pattern is over-represented. It is clear from the example that the number of similar values is greater than α, and therefore, redundant observations (shaded box) are eliminated from the data set.

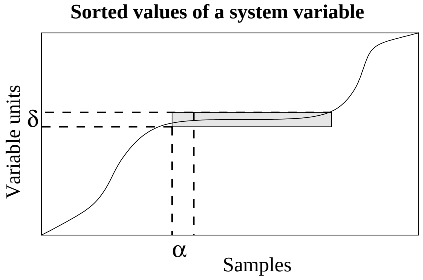

Figure 1. EEFP method parameters

It was decided to implement the δ and α values as input parameters to the algorithm as suitable values of these two parameters are quite dependent on the data. This solution is useful during the initial phase of algorithm development, because it allows to test different values of these parameters easily and to experiment with them in such a way that appropriate values can be found for the application at hand. Currently, we are working on the development of a FIR module that will perform a pre-study of the application data and propose meaningful default values for the δ and α parameters.

Once all the over-represented behavioral pattern are handled (processed), the classical EFP method i used to determine the landmarks from the resultin data set. The landmarks obtained are used to classify the original system data by means of the fuzzification function of the FIR methodology. The FII fuzzification process converts quantitative values int qualitative triples. The first element of the triple i the class value, the second element is the fuzzy membership value, and the third element is the side value The side value indicates whether the quantitative valu is to the left or to the right of the peak value of th associated membership function.

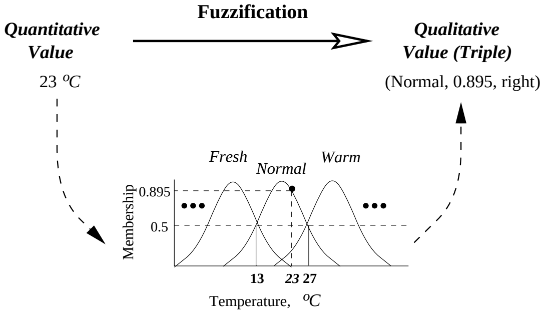

Figure 2. FIR fuzzification process

Figure 2 shows an example of fuzzification of the variable *Temperature*. For instance, a quantitative temperature value of $23°C$ is discretized into a qualitative class value of *'normal'* with a fuzzy membership function value of 0.895, and a side function value of *'right'* (since 23 is to the right of the maximum of the bell–shaped membership function that characterizes the class 'normal').

3 Water demand application

The system to be modeled is the water distribution network of the city of Sintra in Portugal. The goal of the water distribution network is to carry water emanating from wells and rivers for human consumption in the city. It is required that the water arrives at the destination points with a certain pressure-flow. To this end, the network has water reservoirs, valves that regulate the amount of water, and pump stations. Figure 3 represents a simplified diagram of the Sintra water distribution system.

As it is shown in figure 3, the simplified diagram of the water distribution network is composed of 7 reservoirs that must provide the requested water of each demand. However, there is data available for 6 of these reservoirs only, namely: Mabrao, Pimenta, Cotao, Ranholas, Rinchoa and Merces.

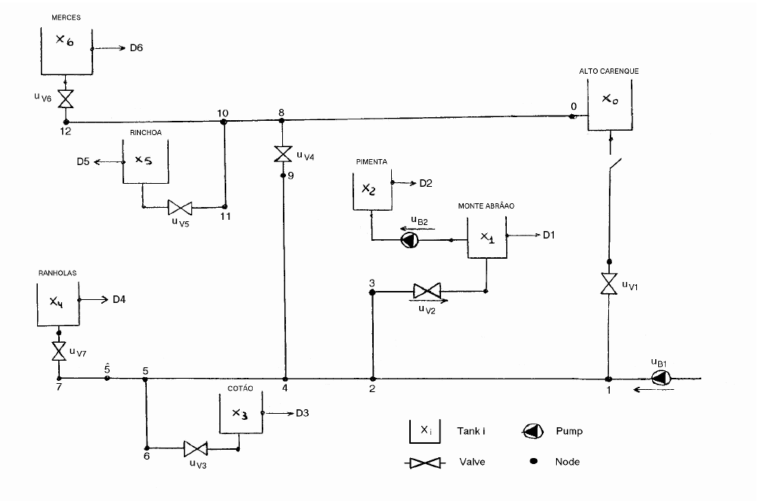

Figure 3. Simplified diagram of the water demand system

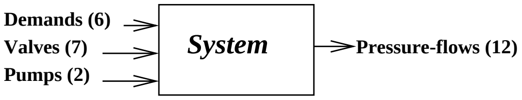

Figure 4. System inputs and outputs

The water demand network can be viewed as a system where the inputs are the water demands, the valves opening and the state of the pumps, whereas the outputs are the pressures in each node. The inputs and outputs of the system are summarized in figure 4.

The water demands for each reservoir are measured data stemming from the water network. The values of the other input variables are obtained from the simulation of a control model of the water demand system. From the control point of view, it is necessary to regulate the pumps and the valves, and if the reservoirs are placed at a high altitude, it may also be necessary to control the turbines because they take advantage of the kinetic energy. The state of the system is represented by the flow, the pressures and the reservoir levels.

Discretization of the system variables

The first step to obtain the pressure-flow models is to discretize the input and output variables by means of the fuzzification process of the FIR methodology. To this end, both the EFP and the EEFP methods have been used to compute the landmarks of all system variables. The first variables to be discretized are the water demands. The upper plot of figure 5 shows the D_1 water demand signal that corresponds to the Mabrao reservoir.

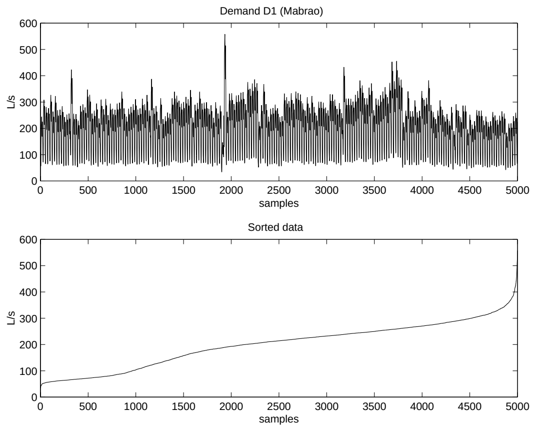

Figure 5. Data distribution of the D_1 *demand* (Mabrao reservoir)

The signal ranges from a value of 34 l/s (liters per second) to a value of 400 l/s, except for a few specific hours when the demand is higher than the upper limit. The lower plot of figure 5 shows the sorted data. This plot can be interpreted as the distribution function of a histogram. For example, there are 2000 samples with a water demand of less than 200 l/s. The resulting signal presents itself as fairly linear, except during the rightmost interval that contains the outliers.

It was decided to discretize the 6 demand variables into 3 classes each. Three classes seem to be enough for capturing the dynamic behavior of these signals. Once the number of classes is defined, both the EFP and the EEFP methods can be applied to obtain the landmarks. In order to compute the landmarks when the EFP method is used, it is necessary to divide the ordered signal into three classes, each one containing the same number of occurrences. Therefore, the lower landmark of class 1 is the smallest value of the sorted signal, the upper landmark of the same class is the value that corresponds to one third of the total number of samples, and so on. The landmarks of the 3 classes when using the EFP method for the D_1 demand signal (figure 5) are shown in table 1. The third column shows the number of occurrences within each class.

Class	**Landmarks**	**NofO**
1	34.0-172.3	1666
2	172.3-244.3	1666
3	244.3-557.5	1668

Table 1. Landmarks of the D_1 *demand* when using the EFP method

In order to compute the landmarks when the EEFP method is used, it is necessary to determine the values of the δ and α parameters (see figure 1). The criterion that have been adopted in the application at hand, is to consider that two observations are similar if they differ less than 1% of the amplitude range of all observations. Therefore, the δ value in that case is of 1%. On the other hand, it has been considered that an α value of 10% is acceptable taking into account the total number of samples available.

The EEFP algorithm is applied with the predetermined parameter values obtaining as a result the sorted original signal without the data associated to over-represented behavioral patterns. The landmarks are then computed from the new signal by using the EFP method, as has been already explained.

As can be seen from the lower plot of figure 5 there are few similar observations. Therefore, the landmarks obtained when applying the EFP and the EEFP methods are exactly the same in this case. The same process has been used to obtain the landmarks for the other 5 water demand signals. As it happened in the case of the Mabrao reservoir, the water demand data for the Pimenta, Cotao, Ranholas, Rinchoa, and Merces reservoirs don't exhibit over-represented behaviors and, therefore, the use of the EFP method produces a reasonable classification for these variables.

The next input variables that should be discretized are the 7 valves that can be regulated from 0% to 100% of opening. The observations registered from the second valve are presented in figure 6.

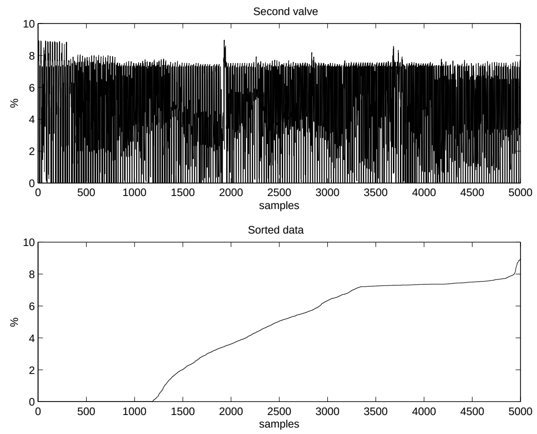

Figure 6. Data distribution of the *second valve*

In the upper plot of this figure, the observed trajectory of the valve is presented. As can be seen, the valve operates with varying degrees of opening ranging between 0% and 10%. The lower plot of the same figure shows the ordered data. There is a high number of observations (more than 1000) with an opening of 0%. Therefore, when the EFP method is used, the computation of the landmarks become distorted due to the over-represented behavioral pattern. As in the case of the water demand variables, it is decided to discretize all 7 valve signals into three classes each. Table 2 shows the landmarks of the second valve obtained with the EFP method.

Class	Landmarks	NofO
1	0.01-2.685	1666
2	2.685-7.19	1665
3	7.19-8.97	1669

Table 2. Landmarks of the *second valve* when using the EFP method

In this case, the first class represents almost exclusively the values of 0% of opening. This situation is not desirable because clearly it is an over-representation of that system behavior. The landmarks obtained using the EEFP method for the second valve signal are shown in table 3. The application of the EEFP method allows to obtain a more representative distribution of the data within the classes.

Class	Landmarks	NofO
1	0.01-4.73	2383
2	4.73-7.28	1199
3	7.28-8.97	1418

Table 3. Landmarks of the *second valve* when using the EEFP method

The last input variables to be discretized is the state of the pumps. In the water network studied, only two pumps (UB1 and UB2) can be controlled. The UB1 pump provides water to node 1, whereas the UB2 pump provides water to the Pimenta reservoir corresponding to the D_2 demand. Each pump is composed of two motors, that can either be both stopped, both pumping, or one stopped and one pumping. This is the reason why we propose to not use an equal frequency partition method for these variables, but instead lump the individual binary states of both motors into a single ternary variable, where each ternary state represents one of the three possible situations as shown in table 4.

Once all input variables have been discretized, it is the turn of the 12 output variables. It was decided to discretize the pressure-flows into three classes, as it has been done for all input variables. The pressure-flows are measured in meters of water column. Figure 7 shows the distribution data of the pressure-flow at node 4.

Classes	State
1	Zero motors working
2	One motor working
3	Two motors working

Table 4. Classification of the *pump* variables

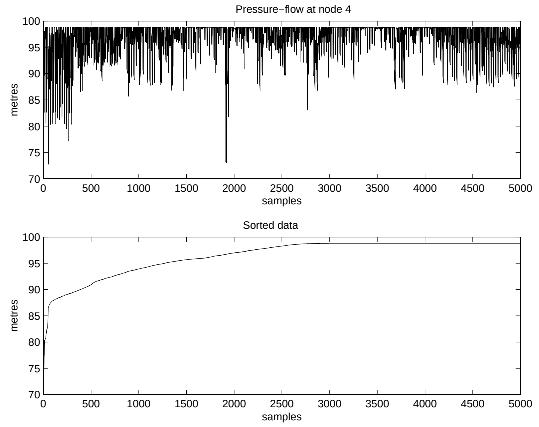

Figure 7. Data distribution of the *pressure-flow* at node 4

In this node, the pressure-flow takes values within the range of 70 to 100 meters of water column. If we analyze the ordered data (lower plot of figure 7), it can be observed that more than one/third of the total number of samples have a value of 98.8 meters. Therefore if we use the EFP method to compute the landmarks, it happens that values of 98.8 can be found in two different classes. This situation is obviously undesirable and it is not allowed in the fuzzification process of FIR methodology. This is the reason why the upper landmark of class 2 and the lower landmark of class 3 (that are the same value) are modified in such a way that all the 98.8 observations are included in class 3. The landmarks obtained are presented in table 5.

Class	Landmarks	NofO
1	72.74-95.95	1665
2	95.95-98.7	1057
3	98.7-98.8	2278

Table 5. Landmarks of the *pressure-flow* at node 4 when using the EFP method

Also in this case, the EEFP method is used to compute the landmarks. Due to the high number of repeated occurrences found in the data (figure 7) it is to be presumed that the EEFP algorithm will give a better distribution of the data within the three classes. Table 6 contains the landmarks obtained when using the EEFP method.

Class	Landmarks	NofO
1	72.74-93.84	979
2	93.84-96.8	978
3	96.8-98.8	3043

Table 6. Landmarks of the *pressure-flow* at node 4 when using the EEFP method

At this point the landmarks of all input and output variables have been obtained by means of the EFP and EEFP methods. Now the fuzzification process of the FIR methodology can be applied to each variable in order to obtain qualitative representations of the given signals. As explained before, the FIR fuzzification function converts each quantitative value into a qualitative triple that contains the class, the membership and the side values (see figure2). With the qualitative data available, the identification of qualitative pressure-flow models can take place.

Pressure-flow model identification

The qualitative model identification process of the FIR methodology is responsible for finding causal spatial and temporal relations between system variables and therefore to obtain the best model (called *mask* in the FIR nomenclature) that represents the system. The identification function evaluates all possible masks and concludes by means of an entropy reduction measure, which of them has the highest quality.

Once the best model has been identified, it can be applied to the qualitative data matrices resulting in a fuzzy rule base that, in FIR terminology, is called the *behavior matrix*. Once the behavior matrix and the mask are available, predictions of future states of the system can be made using the FIR fuzzy inference engine. This process is called fuzzy forecasting. Th FIR inference engine is a specialization of the k-nearest neighbor rule, commonly used in the pattern recognition field. For a deeper inside to the FIR methodology, the reader is referred to (Nebot *et al.* 1998).

In this section, the FIR qualitative identification function is used to obtain two models for each one of the 12 pressure-flow variables. The first model is identified from the qualitative data obtained when the EFP method is used to compute the landmarks, whereas the second model is identified from the qualitative data obtained from discretization when the EEFP method is used. Once the best models are identified for each variable, the fuzzy forecast function of the FIR methodology is used to predict a subset of the data not used in the identification process. The prediction errors obtained are computed by means of the formula presented in equation 1.

$$MSE = \frac{E[(y(t) - \hat{y}(t))^2]}{y_{\mathrm{var}}} \cdot 100\% \qquad (1)$$

The FIR model obtained for the pressure-flow at node 4 when the EFP method is used to compute the landmarks is described in equation 2.

$$P4(t) = \tilde{\mathrm{f}}(V4(t), V6(t), P4(t-1), P4(t-24)), \qquad (2)$$

In this formula, the mask (best model) is represented in equation format for simplification. This formula suggests that the current value of the pressure-flow at node 4 depends somehow on the value of the fourth valve at the present time, the value of the sixth valve also at the present time, and on the values of the pressure-flow at node 4 one hour and one day in the past. In equation 2, $\tilde{\mathrm{f}}$ denotes a *qualitative relationship*. It does not stand for any (known or unknown) explicit formula, but only represents a generic causal relationship. The quality associated with that model is 0.7492.

The model presented in equation 2 is then used to forecast the pressure-flow at node 4 during one day (24 samples). It does not make sense in the application at hand to predict for more than one day into the future, because one day suffices for the purpose of controlling the input variables in an optimal manner. The upper plot of figure 8 shows the real *vs.* the predicted signals of the pressure-flow at node 4 when the model described in equation 2 is used. The solid line represents the measured signal, whereas the dashed line represents the forecast. The MSE error in percentage (see equation 1) obtained is 13.3112%.

As can be seen from the plot, the predicted signal follows the real curve up to a certain degree. It is evident that the prediction obtained for the first 9 hours is quite poor.

The FIR model obtained for the pressure-flow at node 4 when the EEFP method is used to compute the landmarks is described in equation 3.

$$P4(t) = \tilde{\mathrm{f}}(V4(t), V4(t-15), P4(t-1), P4(t-24)) \qquad (3)$$

The model described in equation 3 differs in one component from the model obtained when the classical EFP method is used. Notice that now, the output variable at present time depends on the value of the fourth valve fifteen hours in the past and not on the value of the sixth valve at the present time. The associated

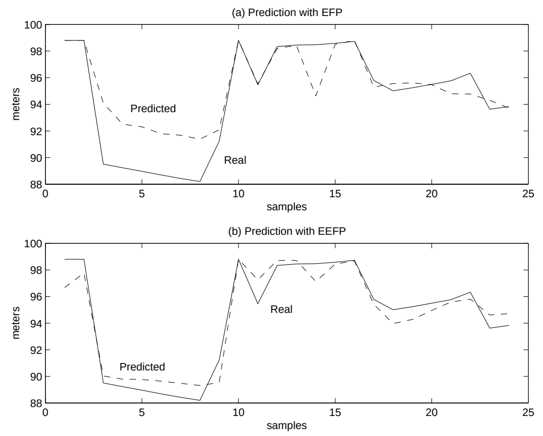

Figure 8. Prediction of the *pressure-flow* at node 4 with EFP and EEFP method

quality of the new model is 0.7765, i.e., slightly higher than the quality obtained for the previous model.

The new model is then used to predict the same data as before, obtaining the results shown in the lower plot of figure 8. As can bee seen from the figure, the prediction obtained is more accurate, resulting in an MSE error of only 3.2376%. It is evident that, at least in this case, the use of the EEFP method helped to obtain more reasonable distributions of the original data into classes, leading to better fuzzifications and a more accurate model.

	EFP	**EEFP**
node 1	3.0602%	1.1269%
node 2	2.5627%	1.5212%
node 3	2.2279%	2.5324%
node 4	13.3112%	3.2376%
node 5	21.0761%	3.3052%
node 6	3.4005%	1.2636%
node 7	0.9704%	0.9838%
node 8	1.2997%	0.4703%
node 9	13.1315%	2.0776%
node 10	0.4109%	0.1219%
node 11	0.4109%	0.2103%
node 12	0.3999%	0.2429%

Table 7. MSE of the pressure-flow models at nodes 1-12

The prediction errors computed for the pressure-flow models at all 12 nodes are shown in table 7. The first column of the table contains the MSE prediction errors obtained when the EFP method is used to compute the landmarks of all system variables. Taking into account that the errors are in percentages, the results obtained are quite acceptable, except at nodes 4, 5, and 9 for which higher forecasting errors are found. The results obtained when the EEFP method is used are presented in the second column of the same table. As can be seen, the errors were reduced considerably at nodes 4, 5, and 9. However, the errors of most of the other models were also reduced.

4 Conclusions

In this paper, an enhancement to the classical equal frequency partition method is proposed. The EEFP method allows to obtain a better distribution of the data into classes, while maintaining the simplicity of the EFP method. The new algorithm is specially useful in those situations where the different system behaviors are not represented within the data with similar numbers of occurrences. The FIR methodology is chosen in this work to model a real system, the water distribution network of a city of Portugal. The classical EFP an the new EEFP methods are used in the fuzzification process of the FIR methodology, and are compared from the point of view of the prediction accuracy of the models identified from the classified data. In this research it is shown that the use of the EEFP method allows the FIR methodology to synthesize models that represent better the system behavior. The prediction errors obtained when the EEFP method was used in the fuzzification process are usually lower than the ones obtained when the classical EFP method was used. More importantly, none of the models exhibits a poor forecasting quality any longer.

References

Anderberg, M. 1973. Cluster Analysis for Applications. John Wiley & Sons, Inc., London.

Bezdek, J.C., R. Ehrlich and W. Full. 1984. "*FCM*: The fuzzy *c*-means clustering algorithm." Computers & Geosciences 10, no. 2-3: 191-203.

Li, C. and G. Biswas. 1999. "Finding Behavior Patterns from Temporal Data using Hidden Markov Model based Unsupervised Classification." In Proceedings of the 1999 CIMA:Computational Intelligence Methods and Applications (Rochester, NY, June 22-25), 266-272.

Cellier, F.E., A. Nebot, F. Mugica and A. de Albornoz. 1996. "Combined Qualitative/Quantitative Simulation Models of Continuous–Time Processes Using Fuzzy Inductive Reasoning Techniques." International Journal of General Systems 24, no. 1-2: 95-116.

Nebot, A., F.E. Cellier and D.A. Linkens. 1996. "Synthesis of an Anaesthetic Agent Administration System Using Fuzzy Inductive Reasoning." Artificial Intelligence in Medicine 8, no. 3: 147-166.

Nebot, A., F.E. Cellier and M. Vallverdú. 1998. "Mixed Quantitative/Qualitative Modeling and Simulation of the Cardiovascular System." Computer Methods and Programs in Biomedicine 55: 127-155.