

Fuzzy Modeling in the Agro-Climatic Domain

Mercedes Valdés

Departamento de Ingeniería
de la Información
y las Comunicaciones,
Universidad de Murcia.
Campus de Espinardo.
E-30100. Murcia, Spain.
mvaldes@dif.um.es

Juan A. Botía

Departamento de Ingeniería
de la Información
y las Comunicaciones,
Universidad de Murcia.
Campus de Espinardo.
E-30100. Murcia, Spain.
juanbot@um.es

Antonio F. Gómez-Skarmeta

Departamento de Ingeniería
de la Información
y las Comunicaciones,
Universidad de Murcia.
Campus de Espinardo.
E-30100. Murcia, Spain.
skarmeta@dif.um.es

Abstract

Fuzzy Modeling is an effective approach for System Identification. In its turn, *Data Driven Fuzzy Modeling* (DDFM) extracts these models from a set of input-output observations about the system. One way to carry out a DDFM process is by means of a combination of techniques, each one solving one of the DDFM phases. In this paper, we apply hybridizations of clustering algorithms and neural networks (NN) in order to solve regression problems in the agro-climatic domain.

Keywords: data driven fuzzy modeling, hybridization, clustering, neuro-fuzzy networks, evapotranspiration, solar radiation.

1 Introduction

Data Driven Fuzzy Modeling (DDFM) is an approach to the extraction of models starting from input-output data of the target system. The generated models are represented as fuzzy inference systems (FIS) [6]. Three main stages compose a DDFM process: first, the number of rules must be detected (*rules number identification*). After that, a rough approximation to the set of rules is obtained (*rules generation*). These rules are based on fuzzy sets whose parameters are adjusted in the *parameter optimization* stage to better mimic the system behaviour. In the last years, many *hybrid DDFM techniques* have arisen. They are hybrid in the sense that they do the modeling combining several algorithms, each one facing one or more of the phases. One of the most successful hy-

brid approaches begins with clustering algorithms [1] to solve the first two phases. Every cluster detected in the input-output data is viewed as a potential fuzzy *IF – THEN* rule. These clusters can be projected into every dimension outlining the fuzzy sets involved in the rules antecedents and consequents. Afterwards, these rough set of rules can be adjusted. In order to do that, optimization methods as Artificial Neural Networks (ANN) [3] are usually used. In this work we apply this approach to the prediction of the water needs of a plant and to the interpolation of the solar radiation.

This paper is organized as follows: in section 2 the DDFM techniques applied in this paper are explained. Section 3 is devoted to the application of the described strategy to the prediction of the water needs of a crop. In section 4 the DDFM strategy is used to the interpolation of solar radiation. Finally, in section 5 the conclusions are summarized.

2 An Hybrid Approach to DDFM

In this paper we are concerned with Takagi-Sugeno (TS) FISs [5]. In such kind of FISs the consequents of the rules are linear functions of the input variables:

$$\begin{aligned} \text{IF } x_1 \text{ is } A_1 \text{ AND } \dots \text{ AND } x_p \text{ is } A_p \\ \text{THEN } y = a_0 + a_1x_1 + \dots + a_px_p \end{aligned} \quad (1)$$

With this type of consequents each rule describes a local behaviour of the system. A type of TS FIS where every consequent is a crisp value is considered a special case of TS FIS with zero-order poly-

nomials in the consequents, the so-called zero-order TS FIS.

Our hybrid DDFM strategy starts with an *auto-organizational clustering method* in order to detect the suitable number of rules and the initial centroids. These centroids are the starting point for a *clustering optimization method* resulting in a first approximation to the target FIS. Finally, a *neuro-fuzzy architecture* is initialized with the centroids detected. Afterwards, the adaptive capabilities of this network are used in order to adjust the FIS parameters.

The stage of rules number identification will be done by means of an auto-organizational clustering method. Auto-organizational clustering methods are those techniques that are able to detect on their own the number of clusters underlying the data. Therefore, these methods do not only detect the number of rules but also obtain clusters that can be used as initial values for a subsequent rule generation method. For example, the goodness of Fuzzy C-Means Clustering Method (FCM, see section 2) strongly depends on the initial data partition. So, its results can be improved if it is initially configured with the partition generated by the auto-organizational method.

One of the most used auto-organizational techniques is the **Subtractive Clustering Method (SCM)** [2]. This algorithm considers each data point as a potential cluster center. In order to decide which ones become centroids, a measure of potential is calculated for every candidate. A point with many nearby points has greater potential than one with few close data. Therefore, a radius r_a inside of which points contribute to the potential calculus must be fixed. After the selection of the current cluster, the potential for every candidate is reduced. This is done by the subtraction of a quantity that is inversely proportional to the distance to the last detected cluster. This reduction can be controlled fixing the value r_b . This parameter determines the radius defining the neighborhood which will have measurable reductions in potential.

This process of center selection and potential reduction is repeated until the stopping criterium is fulfilled (see [2]).

The clustering method just described, selects a set of cluster centers among an existing set of candidates. Let it be C its cardinal. However, perhaps the “real” centroids are not in that set. So, methods for clusters optimization are necessary.

The **Fuzzy C-Means Clustering Method (FCM)** [1] is one of the most used in order to optimize an initial partition of the data. Given a set of examples, the algorithm must be provided with the previously obtained set of C centroids and an initial fuzzy C -partition of the data. Then, this algorithm optimizes the C -partition in such a way that certain cost function J is minimized. This function is the weighted within groups sum of squared-errors.

At each step, three operations are successively carried out until the centroids are stable with respect to a given tolerance (ϵ): (1) computation of the centroids (assuming that the membership degrees composing the partition are constant numbers); (2) calculus of the distances from each data to every centroid; (3) update the membership degree of each data point to every cluster (assuming that the centroids are constant numbers).

We must remark that this algorithm allows the control of the clusters fuzziness setting certain m parameter, in such a way that the greater m the fuzzier the clusters. On the other hand, the closer to its lower bound (that is 1) the crisper the clusters. The more widespread value is 2. From now on we call HCM (**Hard C-means Method**) to the FCM algorithm with $m = 1$ (crisp partition).

Through the sequential application of an auto-organizational clustering method as SCM and a cluster optimization algorithm as FCM or HCM a first approximation to a FIS is obtained. Centroids obtained by SCM are the initial centroids to FCM. Then the clusters generated by FCM are a good first approximation to the structure of the target FIS. These clusters are the initial parameters of the fuzzy sets of a potential TS FIS. The next step is the tuning of those parameters. This stage is described in the next section.

In order to optimize the parameters involved in the fuzzy rules (that is, parameters of the fuzzy sets and coefficients of the consequents), in this paper we use a neuro-fuzzy approach. Neuro-

fuzzy mechanisms merge the adaptive capabilities of artificial neural networks (ANN) with the human-readable information representation provided by the FISs. Concretely we use the **Adaptive Neuro-fuzzy Inference System (ANFIS)** [3] that is functionally equivalent to a TS FIS. It is compound of five layers. The nodes in layer 1 and layer 4 correspond to the antecedent and consequent parameters of the TS FIS respectively. We have supposed fuzzy sets with bell-shaped membership functions (MF):

$$bell(x; a, b, c) = \frac{1}{1 + \left| \frac{x-c}{a} \right|^{2b}} \quad (2)$$

Therefore, the parameters c of the fuzzy sets in the antecedents layer are initialized with the centroids obtained in the previous phase of DDFM and then the training is carried out during several *epoches* (iterations). Every epoch of the training is compound of a forward and a backward pass. In the forward pass the network is evaluated for every input data and the rule consequent parameters are identified by means of the least-squares estimator. Afterwards, errors for every training data are calculated and, in the backward pass, error signals are propagated and the rule antecedents are modified through backpropagation.

In the next sections the applications of these techniques to problems from the agro-climatic domain are explained.

3 DDFM Applied to the Prediction of the Hydric Needs of a Crop

The estimation of irrigation water needs of a crop is an important problem in zones where water is scarce as in the Region of Murcia, in southeastern of Spain. The base to calculate these needs is the estimation of certain magnitude called *Reference Evapotranspiration* (ET_0). ET_0 is the quantity of water that a plant loose because of the transpiration. Its value depends on the climatic conditions and the type of crop of the plant. Once the ET_0 is obtained, the crop evapotranspiration ET_c is calculated through the equation $ET_c = K_c \times ET_0$, where K_c is a constant to adapt the evapotranspiration to a particular growth phase, type of plant etc. Finally, the quantity of water (N_t) needed is calculated as $N_t = ET_c - P_e$, where P_e is the rain.

Traditionally, the ET_0 value has been estimated starting from climatic data and applying analytically defined mathematical models. One of the more widespread model is the *Class A Pan Evapotranspiration Model* (CAPEM) [4]. It is based on the measure of levels of water in a pan and certain climatic variables influencing the water consumption: wind speed and humidity. The formula relating the ET_0 with these variables is $ET_0 = K_p \times E_{pan}$, being E_{pan} the evaporation in the pan given in *millimetres/day* and being K_p a coefficient calculated as:

$$K_p = a_0 + a_1U + a_2H_r + a_3d + a_4H_r^2 + a_5d^2 + a_6UH_r^2 + a_7dH_r^2 \quad (3)$$

where U is the wind speed, H_r is the air relative humidity, d is the distance from the plantation to the pan and a_i with $0 \leq i \leq 7$ are coefficients obtained in an ideal situation. However, those ideal conditions can be reproduced in a laboratory but not outside. For this reason, experts need to adapt these coefficients a_i to the particular conditions of the region. Nevertheless, this adjustment is not an easy task and it is often based on trial-error.

Let us consider the function $f_{ET_0} : U \times H_r \times E_0 \times d \rightarrow ET_0$, and let us apply our hybrid DDFM strategy in order to obtain a FIS to estimate it. For this aim, we have at our disposal a data base with measures from 64 climatic stations. Although hourly and daily measures are available, they are very noisy and hence they are not suitable for learning. So we use a training set with weekly averaged data corresponding to a three years period (2580 tuples). The tuples have the format $\{U, H_r, E_0, d, ET_0\}$.

We measure the accuracy of the model obtained by the combination SCM+HCM+ANFIS by means of the modeling root mean square error (RMSE). The equation for RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i^t - \hat{y}_i^t)^2}{n}} \quad (4)$$

being y_i^t and \hat{y}_i^t the inferred and desired outputs for the i -th data and being n the number of data. The traditional CAPEM model obtains an error

of 1.69 while the FIS obtained by the combination SCM+HCM+ANFIS error is 0.28. This FIS is also compared with the result of training a multilayer perceptron (MLP) with 10 hidden nodes achieving an error of 0.34.

4 Interpolating the Solar Radiation

The climatic network of southeastern of Spain is compound of 64 stations with different kind of climatic sensors. However, not all the stations have the same type of sensors and hence, not all climatic variables can be measured in every station. On the other side, the variables needed in the CAPEM calculus can be measured by means of low cost sensors that are present in every station. This is the reason for the widespread use the CAPEM formula instead of other existing analytical models more reliable proposed by the Food and Agriculture Organization of the United Nations (FAO) ¹.

Nevertheless there exist many more reliable methods based on other climatic variables. One of them is the so-called *FAO- Penman-Monteith* model. According to the experts, this is one of the most suitable models in southeastern of Spain due to the particular conditions of this region. However, many of the stations in the south-east climatic network lack technology to measure solar radiation. Therefore, it is interesting to apply DDFM to the problem of interpolating the radiation in a point, starting from the measures gathered in the stations possessing the needed technology. In the two next subsections we describe two approaches for solving this problem.

4.1 One Reference Station Approach

We have at our disposal a Digital Elevation Map (*DEM*) of the Region of Murcia. Therefore we have the coordinates (x, y, z) in the space $X \times Y \times Z$ for all the climatic stations, where X represents the longitude, Y is the latitude and Z corresponds to the height over the sea level.

The set of stations able to measure the radiation is referred as E_R (that is to say, E_R is the set of their *DEM* coordinates). We try to obtain a

FIS in order to infer the radiation in a point m taking as input: the week of the year (from 1 to 52), the radiation measured in another point p of the *DEM* map and the relative position of m with regard to p . Therefore, our problem is to approximate the function:

$$f : [1, 52] \times RAD \times X \times Y \times Z \rightarrow RAD$$

Every data in the training set is a tuple $(w, r_i, x_{io}, y_{io}, z_{io}, r_o)$ where $w \in \{1, 52\}$ represents the week of the year, r_i and r_o are the measures of radiation in stations e_i and e_o respectively with $e_i, e_o \in E_R$ and $e_i \neq e_o$. The station e_o is the location whose radiation we want to know and e_i is the reference station (that is, the station whose radiation is taken as input). Therefore, r_o is the ideal output while r_i is the input to the model.

Solar radiation is measured in *watts/metres*² and its domain is $[0, 500]$ *watts/metres*². Finally, $(x_{io}, y_{io}, z_{io}) = (|x_i - x_o|, |y_i - y_o|, |z_i - z_o|)$, is the coordinates vector of station e_o with regard to e_i . The training set is compound of 12578 weekly data. After the SCM+HCM+ANFIS strategy a FIS with 37 rules is obtained with a training RMSE of 15.53.

Once we have a FIS able to interpolate the radiation, a reference station must be chosen from the climatic network. The data measured in the chosen reference station will act as inputs to our FIS when it is in use. In order to decide this, we validate our model regarding to several data sets. In such a way that, in each set, a different reference station is considered. For example, the set D_n is the one with $e_n \in E_R$ acting as reference station. Therefore, it is compound of examples whose r_i components come from e_n and whose r_o components came from any other station $e_m \in E_R$ with $e_n \neq e_m$.

In table 1 the validation errors are shown. The first column represents the codes of every reference station. The second one is the number of data. The third one is the averaged RMSE. As we can see, the best errors are produced when the reference station is either CR12 or MO31, while the worst one is obtained when AL62 is used. This is logical taking into account their respective geographic locations. In fact CR12 and MO31 are

¹<http://www.fao.org>

Table 1: Validation Errors

REFERENCE STATION	e_n	#data D_n	RMSE
AL62		6979	28.71
CI52		6389	25.41
CA91		6455	25.14
CR12		6023	23.03
JU12		6820	26.82
JU51		6860	25.63
JU61		7015	27.08
LO61		6365	25.65
MO31		6559	23.78
MO41		4880	24.48
MU62		6452	25.07
TP11		4360	27.46
TP42		6859	24.96

much more centrally situated than the others.

In figure 1 the radiation values for a period of one year inferred by our model for AL62 using as reference station CR12 (line with crosses) are compared with the real measures (solid line).

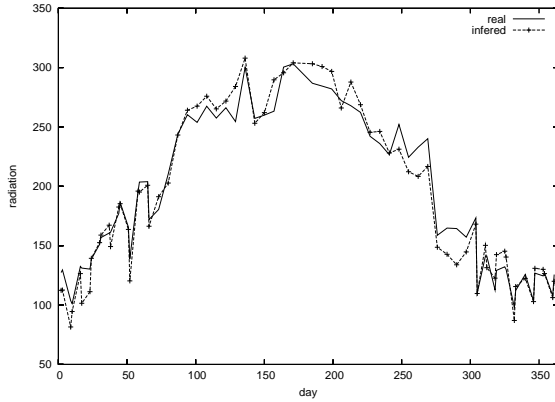


Figure 1: Weekly real radiation in station AL62 (solid line) and the radiation inferred when CR12 is the reference station (line with circles).

4.2 Several Reference Stations Approach

A different approach is to use several reference stations to provide the input values needed to interpolate the output radiation. Let us consider again the set E_R of all the climatic stations which are able to measure the radiation, and let c be its cardinality. As it has been said in the previous section, for every station $e_i \in E_R$ with $i = 1, \dots, c$, its coordinates in the space $DEM = X \times Y \times Z$ are available. We can assume that every point $e_i \in E_R$ is the centroid of a fuzzy cluster S_i with $i = 1, \dots, c$, dividing the space DEM . Let us consider the definition of the membership degree given in FCM [1]. We com-

pute the membership of point p to cluster S_i by the equation:

$$\mu_{S_i}(p) = \left(\sum_{j=1}^c \left(\frac{\|p - e_i\|}{\|p - e_j\|} \right)^{\frac{2}{m-1}} \right)^{-1} \quad (5)$$

with $e_i, e_j \in E_R$. We can also assume, that every cluster S_i defines the antecedent of a zero-order TSK rule R_i in this way:

$$R_i : \text{If } p \text{ is } S_i \text{ then } r = rad_i$$

being r the output variable representing the radiation and being rad_i the crisp value corresponding to the current radiation measured in the i -th station. Taking into account the time factor:

$$R_i(t) : \text{If } p \text{ is } S_i \text{ then } r = rad_i(t)$$

In such a way that $R_1(t), R_2(t), \dots, R_c(t)$ compose a zero-order TSK model whose output for point p is:

$$r(p) = \frac{\sum_{i=1}^c \mu_{S_i}(p) rad_i(t)}{\sum_{i=1}^c \mu_{S_i}} \quad (6)$$

In order to test this model we divide the set E_R into two subsets: the one whose stations are used to define the fuzzy rules of the model (let us call it I_R) and the one whose stations will be used for validation (T_R). In order to build a model based on the most representative stations, we define I_R as the result of applying the algorithm SCM to the set E_R . In this way, the selection of centroids where a station able to measure the radiation exists is ensured (since this algorithm selects centroids among the elements of the input set, E_R in this case) On the other side, we define T_R as $E_R - I_R$. The cardinality of E_R (stations with radiation sensors) is $c = 29$. Applying SCM to E_R we obtain the clusters set I_R with 13 centroids. In this way our FIS is compound of 13 rules, one per centroid.

Finally, the cardinality of T_R is 16.

When the radiation on stations in T_R is inferred from the fuzzy model generated in base of I_R , the averaged RMSE is 16.42.

In table 2 the RMSE for every station in T_R are summarized. It can be observed that the worst

error values is obtained in JU12. This is coherent with the fact that in JU12 the radiation values are almost always much greater than in the other stations caused by a fault in its sensors. On the other hand, let us take into account that, although a set E_R with 29 stations with radiation sensors is available, we are just using 13 for defining the fuzzy model and 16 in order to validate it. Therefore the results in table 2 are worse than if we were using the whole set E_R , that is to say, all the available radiation measurements.

Table 2: Validation RMSE using test data from every station in T_R

TEST STATION	# DATA	FIS RMSE
AL31	366	12.30
AL51	366	15.42
CA21	366	13.24
CA52	353	14.22
CA72	366	16.78
CA91	363	15.33
CI32	366	15.18
JU12	364	32.78
LO61	365	18.05
MO12	360	16.93
MO41	360	11.03
MO51	366	15.00
MO61	366	10.88
MU31	364	16.79
MU62	271	13.23
TP81	279	11.55

Finally, in figure 2 the real and inferred radiation in station MO61 is shown for two months.

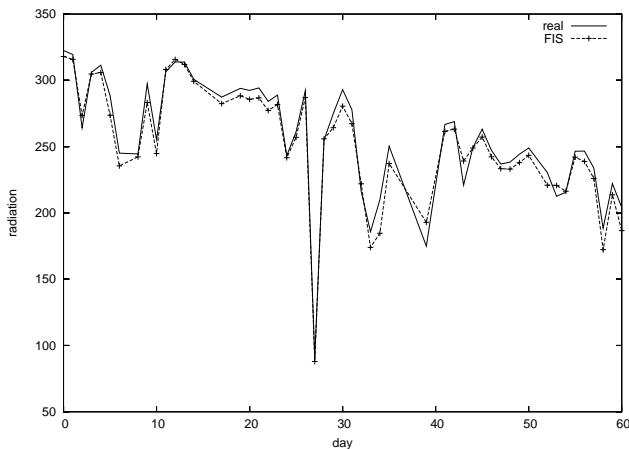


Figure 2: Real radiation in station MO61 compared to inferred radiation.

5 Conclusions

Data Driven Fuzzy Modeling (DDFM) is an effective approach for system identification which represents knowledge based on fuzzy *IF – THEN* rules. One of the most successful approaches to DDFM is the use of combination of different techniques each one solving a phase of the process. In this work we have applied two different approaches of Fuzzy Logic to the calculus of the water needs of a crop. These experiments remark the fact that, the DDFM in general and the hybrid DDFM in particular are successful approaches for the identification of complex systems.

Acknowledgements

This work has been supported by the Spanish Science and Technology Ministry (MCYT) by means of the Fuzzy KIM (TIC2002-04021-C02-02) project.

References

- [1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, USA, 1981.
- [2] S.L. Chiu. Fuzzy model identification based on cluster estimation. *Journal of the Intelligent and Fuzzy Systems*, 2(3):267–278, 1994.
- [3] J.-S. R. Jang, C.-T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing*. Matlab Curriculum. Prentice Hall, 1997.
- [4] Smith M., Allen R. G., and Pereira L. Revised fao methodology for crop water requeriment. In *Proceedings of the International Conference on Evapotranspiration and Irrigation Scheduling*, San Antonio., Texas, 1996.
- [5] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetic*, 15:116–132, 1985.
- [6] L. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions of Systems, Man and Cybernetics*, SMC-3:28–44, 1973.