# A new approach to dealing with missing values in data-driven fuzzy modeling

# A New Approach to Dealing With Missing Values in Data-driven Fuzzy Modeling

Rui J. Almeida, *Graduate Student Member, IEEE*, Uzay Kaymak, *Member, IEEE* and João M.C. Sousa, *Member, IEEE*

***Abstract*— Real word data sets often contain many missing elements. Most algorithms that automatically develop a rule-based model are not well suited to deal with incomplete data. The usual technique is to disregard the missing values or substitute them by a best guess estimate, which can bias the results. In this paper we propose a new method for estimating the parameters of a Takagi-Sugeno fuzzy model in the presence of incomplete data. We also propose an inference mechanism that can deal with the incomplete data. The presented method has the added advantage that it does not require imputation or iterative guess-estimate of the missing values. This methodology is applied to fuzzy modeling of a classification and regression problem. The performance of the obtained models are comparable with the results obtained when using a complete data set.**

## I. Introduction

When applying data analysis methods to real problems, the data sets may contain missing elements. A data set has partial missing data if some attribute values of a feature are not observed. Incomplete datasets present a big obstacle for many learning algorithms, that usually require a complete data set to build the model.

There are two forms of missing data [1]: missing completely at random (MCAR) and missing at random (MAR). Most approaches for dealing with partially missing datasets assume that missing values are missing completely at random (MCAR). Such missing values behave like a random sample and their probability does not depend on the observed data or the unobserved data [2], [3]. Thus, MCAR are interpreted as a random reduction of the dataset, which provide no further information for assigning incomplete feature vectors to clusters.

Depending on how the data were collected, the occurrence of a missing value can provide information about which class the incomplete feature vector might belong to, i.e. the probability for the occurrence of missing values is class-specific. In this case the values are said to be missing at random. MAR exists when missing values are not randomly distributed across all observations but are randomly distributed within one or more subsamples. Missing values of this kind provide additional information for the classification of the partially missing dataset and should be distinguished and also treated differently from feature values that are missing completely at random.

An example of values missing at random is intentionally unanswered questions in a survey conducted. A missing value in a questionnaire (e.g. for income) may indicate if the person was more or less willing to answer depending on their standing (class) within society. Similarly, on medical reports some attributes can be left blank, because they are inappropriate for some class of illnesses. Here the person providing the information feels that it is not appropriate to record the values of some attributes. The same kind of missing values can be found as unmarked sections on data sheets, when the options to choose do not apply to the example at hand (or more particular to its class).

If the proportion of incomplete data is small, it is possible to deal with missing values by deleting all incomplete data, and then execute the data analysis method on the remaining data. This is called the whole data strategy (WDS). The data analysis method is then executed on the remaining data only [4], [5], [2], [3].

However, if missing values are frequent, the data set size may be considerably reduced, yielding unreliable or distorted results. A way to minimize this extreme data reduction problem is presented in [6]. All features with a percentage of missing values higher than a threshold are removed and thereafter records containing missing values in this reduced feature space are removed. The result is a reduced but complete data set. This data set retains as many features as possible, to better capture all relevant relations in the data set, and also as many records as possible, to maintain a sufficient number of examples.

When data values are missing completely at random, they can be replaced or the data can be used with the missing values. Replacing missing values can potentially present disadvantages [7], [8], and is used if the missing values occur rarely or if they can be imputed with a high reliability. Widely used replacement methods use the variables mean, median or the most probable value as a replacement [1], [5]. Another possible approach to handle missing values is to extend the data analysis method so that the method can deal directly with incomplete data. In [9], a method for managing the incomplete input data in a Mamdani fuzzy system is proposed. This method does not use any sort imputation. Evolving models which iteratively use a best guest of the missing data were proposed for classification [10] and for identification of an ARX model [11]. In [4],

Rui J. Almeida is with Erasmus University Rotterdam, Erasmus School of Economics, Econometric Institute, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: rjalmeida@ese.eur.nl

Uzay Kaymak is with Erasmus School of Economics, P.O. Box 1738, 3000 DR Rotterdam and with Eindhoven University of Technology, School of Industrial Engineering, The Netherlands. E-mail: u.kaymak@ieee.org

João M.C. Sousa is with Technical University of Lisbon, Instituto Superior Técnico, Dept. of Mechanical Engineering, CIS/IDMEC – LAETA, Av. Rovisco Pais, 1049-001 Lisbon, Portugal. E-mail: jmsousa@ist.utl.pt

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

four strategies for using the Fuzzy c-Means algorithm (FCM) in incomplete data sets are proposed. In this work, two of the approaches iteratively find the best guess estimate of the missing data, and another approach uses the partial distance strategy (PDS). This fuzzy cluster analysis using the partial distance strategy is also discussed in [12].

An effective approach to the identification of complex nonlinear systems is to partition the data using product space fuzzy clustering into subsets and approximate each subset by a simple linear model [13], [14]. Fuzzy clustering divides the available data intro groups in which local linear relations exists between the inputs and the output. A rule-based model can be obtained from the available fuzzy partition matrix and from the cluster prototypes. This process, as well as the parametrization of the rules, membership functions, consequents and other parameters of the fuzzy model can be extracted in an automated way. This type of approach does not tolerate data with missing values.

In this work we present a novel method to estimate a Takagi-Sugeno model from data containing missing values, without using any kind of imputation or best guest estimation. For this purpose, we use the extension of the Fuzzy c-Means algorithm with the partial distance strategy, as studied in [4], [12] to cluster the incomplete data. From the fuzzy partition matrix $U$ we discuss two methods to identify the fuzzy sets in the antecedent of the rules and also, how to estimate the consequent parameters. Furthermore we study the performance of these models in a regression and classification setting and consider their added value in fuzzy modeling. The proposed method allows to identify Takagi-Sugeno fuzzy models from incomplete data by means of product-space clustering. This method does not require imputation or iterative guess-estimate of the missing values and the model performance is comparable to the results obtained when using a complete data set.

This paper is organized as follows. Section II briefly presents the Takagi-Sugeno models and how they can be identified using product space clustering of an incomplete data set. The experimental setup and results are presented, respectively, in Section III and Section IV. Finally the conclusions and future work are given in Section V.

## II. Fuzzy modeling with missing values

This section presents the basic outline of the proposed method for constructing fuzzy rule-based models by product-space clustering of data containing missing values. After some general considerations, we present the Takagi-Sugeno fuzzy models, and then proceed to explain how they can be derived from a data set $X$ containing missing values.

### A. General Considerations

A data set has partial missing data if some attribute values of a datum $\mathbf{x}_k$ are not observed. For example $\mathbf{x}_k = (x_{k,1}, ?, x_{k,3}, x_{k,4}, ?)$ has missing values in the second and fifth feature. Only the first, second and third feature are observed.

Given a numerical data set $X = [x_{1k}, \ldots, x_{nk}]^T, x_k \in \mathbb{R}^N$ the following subsets are defined:

$$\begin{aligned} X_W &= \{\mathbf{x}_k \in X | \mathbf{x}_k \text{ is a complete datum}\} \\ X_P &= \{x_{kj} \text{ for } 1 \leq j \leq n, 1 \leq k \leq N | \\ &\quad \text{the value for } x_{kj} \text{ is present in} X\} \\ X_M &= \{x_{kj} =? \text{ for } 1 \leq j \leq n, 1 \leq k \leq N | \\ &\quad \text{the value for } x_{kj} \text{ is missing from} X\} \end{aligned}$$

$X_W$ contains the data subset with the features that contain no missing values. $X_P$ is the subset of the data points which are present in the dataset and $X_M$ contains the data points missing in the dataset. For example, for a dataset $X$ with $n = 3$ features and $N = 4$ objects [4],

$$X = \begin{bmatrix} 2 & 3 & 2 & 1 \\ ? & 2 & 1 & 1 \\ ? & 4 & ? & 2 \end{bmatrix},$$

we obtain the subset $X_M = \{x_{12}, x_{13}, x_{33}\}$, $X_P = \{x_{11} = 2, x_{21} = 3, x_{22} = 2, x_{23} = 4, x_{31} = 2, x_{32} = 1, x_{41} = 1, x_{42} = 1, x_{43} = 2\}$ and

$$X_W = \begin{bmatrix} 3 & 1 \\ 2 & 1 \\ 4 & 2 \end{bmatrix}.$$

The missing values $x_{kj} \in X_M$ considered in this work are subject to the following constraints:

1) The values are missing completely at random.
2) Each original feature vector $\mathbf{x}_k$ retains at least one component.
3) Each feature has at least one value present.

The first constraint allows us to assume that the missing values do not provide any kind of information for assigning incomplete feature vectors to clusters. The second and third constraints simply state that the objects and the features without absolutely any information are disregarded, i.e., no object can have all relevant feature observations missing and that no feature can have all observations missing.

### B. Takagi-Sugeno Fuzzy Models

Takagi and Sugeno (TS) [15] introduced a fuzzy rule-based model where the rule consequents are crisp functions of the model input according to

$$R^k: \text{ If } \mathbf{x} \text{ is } A^k \text{ then } y^k = f^k(x), \ k = 1, 2, \ldots, K, \quad (1)$$

where $R^k$ denotes the $k$-th rule, $K$ is the number of rules, $\mathbf{x}$ is the antecedent variable, $y$ is the one dimensional consequent variable and $A^k$ is the (multidimensional) antecedent fuzzy set of the $k$-th rule. Each rule $k$ has a different function $f^k$ yielding a different value $y^k$ for the output.

The most simple consequent functions $f^k$ in (1) is the affine linear form, in which the rules look like:

$$R^k: \text{ If } \mathbf{x} \text{ is } A^k \text{ then } y^k = (\mathbf{a}^k)^T \mathbf{x} + b^k, \quad (2)$$

where $\mathbf{a}^k$ is a parameter vector and $b^k$ is a scalar offset. This model is called an affine TS model. The consequents of the

affine TS model are hyperplanes in the product space of the inputs and the output.

Takagi-Sugeno fuzzy models are suitable for identification of nonlinear systems and regression models. A TS model with affine linear consequents can be interpreted in terms of changes of the model parameters with respect to the antecedent variables as well as in terms of local linear models of the system [16], [13].

Several techniques can be used in fuzzy identification. One possibility is to use identification by product-space clustering [17]. This principle is to approximate a nonlinear problem by decomposing it into several subproblems. The information regarding the distribution of data can be captured by the fuzzy clusters, which can be used to identify relations between various variables regarding the modeled system. In general, the identification of fuzzy models is solved in two steps: structure identification and parameter estimation. Structure identification can be obtained by fuzzy clustering. Parameter identification consists of determining the parameters for the antecedent functions and the parameters of the rule consequents. These steps are briefly reviewed in the next sections.

### C. Product Space Clustering

In clustering, the aim is to partition the data set $X$ into $c$ clusters. A fuzzy partition $U$ of $X$ can be defined as a family of subsets $\{A_i|1 \leq i \leq c\} \subset P(X)$. The membership function $\mu_{ik}$ of data point $i = 1, \ldots, N$ to the cluster $c$, where $N$ is the number of data samples, are allowed to attain real values in $[0, 1]$.

There are several situations where performing the analysis in a complete, but reduced subset $X_W$ is not the desirable solution. Furthermore, when the number of missing values is large, WDS cannot be justified. In these cases, it is possible to calculate partial distances using all non-missing feature values, and scaling afterwards this quantity by the reciprocal of the proportion of components used [18]. An example of the $D_{ik}^2$ calculation of the Euclidean distance is [4]:

$$\begin{aligned} D_{ik}^2 &= \|z_k - v_i\|_2^2 \\ &= \|(?\ 2\ ?\ ?\ 5)^T - (6\ 7\ 8\ 9\ 10)^T\|_2^2 \\ &= \frac{5}{5-3}((2-7)^2 - (5-10)^2) \end{aligned} \tag{3}$$

The general formula for the partial distance calculation of $D_{ik}^2$ is given by

$$D_{ik}^2 = \frac{n}{I_k} \sum_{j=1}^{n} (x_{kj} - v_{ij})^2 I_{kj} \tag{4}$$

where

$$I_{kj} = \begin{cases} 0, \ if \ x_{kj} \in X_M \\ 1, \ if \ x_{kj} \in X_P \end{cases} \quad \text{for } 1 \leq j \leq n \text{ and } 1 \leq k \leq N \tag{5}$$

and $I_k = \sum_{j=1}^{n} I_{kj}$. A possible partial distance strategy (PDS) version of the FCM algorithm (FCM-PDS) is given by the following algorithm:

*Partial Distance Strategy Fuzzy c-Means Algorithm:*
*Step 0*: Given the data set $X$, choose the number of clusters $1 < c < N$, the weighting exponent $m > 1$, the termination tolerance $\epsilon > 0$. Initialize the partition matrix randomly $U^{(0)}$.
**repeat**
for $l = 1, 2, \ldots$

*Step 1*: Compute cluster prototypes (means):
$v_{ij}^{(l)} = \frac{\sum_{k=1}^{N} \left(\mu_{ik}^{(l)}\right)^m I_{kj} x_{kj}}{\sum_{k=1}^{N} \left(\mu_{ik}^{(l)}\right)^m I_{kj}}, \ 1 \leq i \leq c$

*Step 2*: Compute the squared inner-product distance norm:
$D_{ik}^2 = \frac{n}{I_k} \sum_{j=1}^{n} (x_{kj} - v_{ij})^2 I_{kj}, \ 1 \leq i \leq c, \ 1 \leq k \leq N$

*Step 3*: Update the partition matrix:
**if** $D_{ik}^2$ for $1 \leq i \leq c, \ 1 \leq k \leq N$ **then**
$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{ik}^2}{D_{jk}^2}\right)^{2/(m-1)}}$
**else**
$\mu_{ik}^{(l)} = 0$ if $D_{ik}^2 > 0$, and $\mu_{ik}^{(l)} \in [0, 1] with \sum_{i=1}^{c} \mu_{ik}^{(l)} = 1$
**end if**
**until** $\|U^{(l)} - U^{(l-1)}\| < \epsilon$

This algorithm is the regular FCM algorithm where the calculation of the distance $D_{ik}^2$ is substituted by (4) and also the cluster prototype centers $v_{ij}^{(l)}$, are calculated according to:

$$v_{ij}^{(l)} = \frac{\sum_{k=1}^{N} \left(\mu_{ik}^{(l)}\right)^m I_{kj} x_{kj}}{\sum_{k=1}^{N} \left(\mu_{ik}^{(l)}\right)^m I_{kj}} \tag{6}$$

This algorithm possesses all the standard convergence properties of FCM and the global convergence theory is applicable, as is the local alternating optimization convergence theory [4].

### D. Antecedent Membership Functions

Generally, the antecedent membership functions can be obtained by projecting the fuzzy partition onto the antecedent variables, or by computing the membership degrees directly in the product space of the antecedent variables.

*a) Antecedent Membership Functions by Projection:* The principle of generating antecedent membership functions by projection is to project the multidimensional fuzzy sets defined point wise in the rows of the partition matrix $U$ onto the individual antecedent variables of the rules. This method projects the fuzzy partition matrix onto the axes of the antecedent variables $x_j$, $1 \leq j \leq p$.

In order to obtain membership functions for the antecedent fuzzy sets $A_{ij}$, the multidimensional fuzzy set defined pointwise in the $i$th row of the partition matrix $U$ are projected onto the axes of the antecedent variables $x_j$, $1 \leq j \leq p$, by

$$\mu_{Aij}(x_{jk}) = \text{proj}_j(\mu_{ik}). \tag{7}$$

In order to obtain a model, the point-wise defined fuzzy sets $A_{ij}$ can be approximated by appropriate parametric functions.

Using FCM-PDS, the partition matrix $U$ is complete. If we assume that the antecedent variables $x_j$ are partially complete, the method of obtaining the antecedent fuzzy sets by projection can be used.

In general, it is considered that an advantage of this method over the multidimensional membership functions is that the projected membership functions can always be approximated in such a form that convex fuzzy sets are obtained.

When computing the degree of fulfillment $\beta_i(\mathbf{x})$ of the $i$th rule, the original cluster in the antecedent product space is reconstructed by applying the intersection operator in the cartesian product space of the antecedent variables:

$$\beta_i(\mathbf{x}) = \mu_{Ai1}(x_1) \wedge \mu_{Ai2}(x_2) \wedge \ldots \wedge \mu_{Aip}(x_p), \quad (8)$$

Where $\wedge$ denotes a t-norm. Suitable $t$-norms are the product or the minimum operator.

*b) Multidimensional Antecedent Membership Functions:* By computing the membership degrees directly in the space of the antecedent variables, each cluster represents one fuzzy rule, as in (2). The multidimensional membership functions $A^k$ are given analytically by computing the distance from the projection of the cluster center $v_i$ onto $X$, and then computing the membership degree in an inverse proportion to the distance, with the form:

$$\beta_i(\mathbf{x}) = \left[ \sum_{j=1}^{c} \left( D_{ik}^2(\mathbf{x}, v_i^x) / D_{jk}^2(\mathbf{x}, v_j^x) \right)^{2/(m-1)} \right]^{-1}. \quad (9)$$

where $D_{ik}^2$ is given by (4). The membership degree is computed directly for the entire input vector, without decomposition. The antecedents of the TS rules are simple propositions with multidimensional fuzzy sets given by (1), and $\beta_i(\mathbf{x}) = \mu_{Ai}(\mathbf{x})$.

### *E. Consequent Parameters*

The consequent parameters for each rule can be estimated by the least-squares method. A set of optimal parameters with respect to the model output can be estimated from the identification data set by ordinary least-squares methods. This approach is formulated as a minimization of the total prediction error of the model, or as a minimization of the prediction errors of the individual local models, solved as a weighted least-squares problems. In this work the global least squared method was used, as it aims at the minimization of the global prediction error and yields an optimal predictor, by taking into account the aggregation of the rules. The output of the system using the fuzzy-mean formula is:

$$y = \frac{\sum_{i=1}^{K} \beta_i(\mathbf{x}) [\mathbf{a}_i^T \mathbf{x} + b_i] I_{ij}}{\sum_{i=1}^{K} \beta_i(\mathbf{x})}, \quad (10)$$

where $I_{kj}$ is defined by (5). The degree of fulfilment $\beta_i(\mathbf{x})$ of the $i$-th rule can be computed using (8). The membership value $\mu_{Ai}(\mathbf{x})$ can be obtained from the parametrized membership functions obtained using the projection operator (7) or from the multidimensional membership function defined in (9).

In the case that the antecedent variable has at least one missing component at $x_{kj}$, the degree of fulfilment $\beta_i(\mathbf{x})$ will be complete for the case of the multidimensional antecedent membership functions and will have missing values for the case of projected antecedent membership functions. For the latter, we assume that the membership value $\mu_{Aikj}(x_{kj})$ will have membership value $\mu_{Aikj}(x_{kj}) = 1$. Since the value $x_{kj}$ is missing, no information is available, so it is possible to assume that any value of the domain is a suitable candidate. In fuzzy clustering any point belongs to a certain degree to every prototype cluster center. We represent the state as belonging to all member sets with certainty.

Now, the consequent parameter estimates can be obtained by solving a linear least-squares problem. Consider $(\theta^k)^T = [(\mathbf{a}^k)^T b^k]$, let $\Phi_e$ denote the matrix $[X, 1]$, and let $\Gamma^k$ denote a diagonal matrix in $\mathbb{R}^{N_d \times N_d}$ having the membership degree $\gamma_{ik} = \beta_{ik} / \sum_{j=1}^{c} \beta_{ik}$ as its $k$-th diagonal element. Denote $\Phi'$ the matrix in $\mathbb{R}^{N_d \times K(n+1)}$ composed from matrices $\Gamma^k$ and $\Phi_e$ as follows

$$\Phi' = [(\Gamma^1 \Phi_e), (\Gamma^2 \Phi_e), \ldots, (\Gamma^K \Phi_e)]. \quad (11)$$

Denote $\theta'$ the vector in $\mathbb{R}^{K(n+1)}$ given by

$$\theta' = [(\theta^1)^T, (\theta^2)^T, \ldots, (\theta^N)^T]^T. \quad (12)$$

The resulting least-squares problem, $\Upsilon = \Phi' \theta' + \epsilon$, has the solution

$$\theta' = [(\Phi)^T \Phi']^{-1} (\Phi)^T \Upsilon. \quad (13)$$

The optimal parameters $a^k$ and $b^k$ are given by

$$\begin{aligned} \mathbf{a}^k &= [\theta'_{s+1}, \theta'_{s+2}, \ldots, \theta'_{s+n}]^T, \\ b^k &= [\theta'_{s+n+1}], \text{ where } s = (k-1)(n+1). \end{aligned} \quad (14)$$

With the determination of the parameters $\mathbf{a}^k$ and $b^k$, the fuzzy model identification procedure is completed.

## III. Experimental Setup

Two data sets were used to test the proposed modeling approach in this paper. The classification problem uses the Altman data set for bankruptcy, which consists of five financial ratios as features and the feature class is bankruptcy or not bankruptcy [19]. Automobile miles per gallon (MPG) prediction is a nonlinear regression problem, in which several attributes of an automobile are used to predict the city-cycle fuel consumption in miles per gallon. This data set can be found in the UCI Repository Of Machine Learning Databases and Domain Theories [20]. It contains data collected from automobiles of various makes and models [21].

These examples were chosen because of their simplicity, which provides a suitable framework to study the effects of missing values in modeling. In both datasets $X = X_P = X_W$. Missing values are artificially included in both datasets, by randomly assigning a present value $x_{kj} \in X_P$ to the subset of missing values $X_M$. This random assignment

follows the constraints explained in Section II-A. Note that the output $y$ is kept complete without assigning missing values to it.

In all experiments the following parameters were fixed to the following values: fuzzy partition exponent $m = 2$, minimum amount of improvement $\epsilon = 1 \times 10^{-5}$ and maximum number of iterations $\rho = 100$, which were sufficient to achieve convergence for all trials.

For comparison purposes, we also build a fuzzy rule-based model that uses the complete set. Modeling the systems using the proposed approach entails the following experimental setup

1) Randomly assign missing values to the complete dataset $X$ set to obtain $X \bigcap X_M \neq \emptyset$.
2) Divide the data in training set $X^T$ and validation set $X^V$.
3) Cluster the training dataset $X^T$ using FCM-PDS, as explainded in Section II-C if the data contains missing values or FCM if the data is complete.
4) Obtain antecedent membership functions $A^k$ by projection onto the space of the input variables $x_j$ or by computing the membership degrees directly in the product space of the antecedents variables, as explained in Section II-D.
5) Compute the consequent parameters using global least squares estimation, as explained in Section II-E.
6) Validate the model using the validation set $X^V$.

For the validation step, we consider data with missing values as well as a complete dataset. This comparison allows us to check if the model is biased towards data with missing values. In all tests we used a simple holdout method for validation. This is the simplest kind of validation. The data set is separated into two sets, called the training set $X^T$ and the validation set $X^V$. A model is identified using the training set only. Then the model is asked to predict the output values for the out-of-sample data in the validation set, which it has not seen before. Only the results obtained with the validation set are reported. Since there is significant variation in the clustering results [4] we generated 50 trials and we compare the results across different methods, maintaining the data set constant for each trial.

The following nomenclature will be used to differentiate between models.

- The clustering method indicates if the model was built with data with missing values (FCM-PDS) or if the data was complete (FCM).
- The model built with incomplete data, was validated both data contaning missing values ($X^V \bigcap X_M^V \neq \emptyset$) and without missing values ($X^V \bigcap X_M^V = \emptyset$).
- The models derived with the antecedent membership functions obtained by projection onto the space of the input variables, are indicated by $A^k_{proj}$.
- The models derived with the antecedent membership functions obtained by computing the membership degrees directly in the product space of the antecedents variables are indicated by $A^k_{multi}$.

The simplest performance criterion for a regression problem is the root mean square error (RMSE) of the output error with respect to an independent set of out-of-sample data, with size $N_V$,

$$RMSE = \sqrt{\frac{1}{N_V}\sum_{k=1}^{N_V}(y_k - \hat{y_k})^2}, \tag{15}$$

where $\hat{y_k}$ is the output predicted by the model for the true output value $y_k$.

A possible way to solve a classification problem is to consider a regression system that consists of fuzzy if-then rules combined with a fuzzy inference mechanism. Assuming that the classes can be ordered, a post-processing step is applied to the output of the fuzzy inference system in order to determine the crisp class to which the feature vector $\mathbf{x}$ belongs.

The goal of a classification problem is to predict an unknown label $y$ based on an observed input $x$. For the testing data set, the classification obtained is compared with the actual label. If the two match, there is no error. If they do not match, then an error has occurred. The overall performance is measured by the accuracy. Accuracy is defined by the following ratio:

$$\text{Accuracy} = \frac{\sharp \text{CC}}{N}. \tag{16}$$

where $\sharp CC$ is the number of correct classifications and $N$ is the total number of examples.

## IV. Examples

This section reports the application results for the proposed approaches to the databases in study. The regression model is about the prediction of miles per gallon in automobile and the classification model is about the prediction of bankruptcy.

### A. Regression

Table I exhibits the average RMSE and standard deviation (in parenthesis), of the fuzzy models for different values of missing values. In this example we kept the order of the data at each trial, and randomized the placement of the missing values. The number of clusters used was $c = 3$. Using the fuzzy rule-based model constructed with the complete dataset, we obtained an average RMSE for the complete validation set $X^V \bigcap X_M^V = \emptyset$ of 2.868 and 3.151 for the models with the antecedent membership function $A^k_{proj}$ and $A^k_{mult}$, respectively.

As can be seen in Table I the obtained mean square-error of the models built with missing values increases with the number of missing values. Note that, although the obtained RMSE are higher than those for the models built with the complete dataset, the results can still be considered good, even for 70% of missing values. Furthermore, the results show that better results were obtained with models where the antecedent functions $A_{ik}$ are generated by projection.

Comparing the results obtained with the complete data $X^V \bigcap X_M^V = \emptyset$ and incomplete data $X^V \bigcap X_M^V \neq \emptyset$ it is possible to see that better results are obtained for the

TABLE I

RMSE FOR THE MPG DATABASE - MEAN (STD)

| Clust | FCM-PDS | | FCM-PDS | |
|---|---|---|---|---|
| Val.Data | $X^V \bigcap X_M^V \neq \emptyset$ | | $X^V \bigcap X_M^V = \emptyset$ | |
| MV | $A_{proj}^k$ | $A_{mult}^k$ | $A_{proj}^k$ | $A_{mult}^k$ |
| 5% | 4.419 (0.261) | 4.811 (0.860) | 3.878 (0.128) | 3.963 (0.124) |
| 10% | 4.475 (0.191) | 4.559 (0.169) | 4.052 (0.115) | 4.135 (0.091) |
| 20% | 4.568 (0.562) | 4.646 (0.150) | 4.186 (0.106) | 4.310 (0.098) |
| 30% | 4.665 (1.011) | 4.705 (0.160) | 4.306 (0.132) | 4.430 (0.130) |
| 40% | 4.736 (0.275) | 4.880 (0.241) | 4.418 (0.212) | 4.511 (0.125) |
| 50% | 4.930 (0.302) | 5.061 (0.222) | 4.599 (0.298) | 4.689 (0.266) |
| 60% | 5.057 (0.315) | 5.180 (0.270) | 4.941 (0.583) | 4.951 (0.402) |
| 70% | 5.405 (0.248) | 5.443 (0.235) | 6.353 (1.492) | 5.770 (0.948) |

complete data, if the percentage of missing values is below 60%. This indicates that the proposed methodology to build models with missing data is robust, and can compensate for the missing information, without distortion in the results.

An interesting aspect of the methodology proposed in this paper is the assignment of the membership value $\mu_{Aikj}(x_{kj})$ to missing value $x_{kj}$, for the case of projected antecedent membership functions. Table II exhibits the average RMSE and standard deviation (in parenthesis), of the fuzzy models for different values of $\mu_{Aikj}$, for the case of 30% of missing values.

TABLE II

RMSE DIFFERENT VALUES OF $\mu_{Aikj}(x_{kj})$ - 30% MV - MPG

| $\mu_{Aikj}(x_{kj})$ | $A_{proj}^k$ | $A_{mult}^k$ |
|---|---|---|
| 0.1 | 4.553 (0.207) | 4.705 (0.160) |
| 0.2 | 4.561 (0.251) | 4.693 (0.202) |
| 0.3 | 4.682 (1.096) | 4.686 (0.184) |
| 0.4 | 4.594 (0.479) | 4.675 (0.157) |
| 0.5 | 4.551 (0.190) | 4.700 (0.202) |
| 0.6 | 4.592 (0.249) | 4.711 (0.196) |
| 0.7 | 4.582 (0.221) | 4.727 (0.203) |
| 0.8 | 4.617 (0.222) | 4.760 (0.192) |
| 0.9 | 4.601 (0.214) | 4.708 (0.164) |
| 1 | 4.665 (1.011) | 4.705 (0.160) |

Table II shows that different values of $\mu_{Aikj}(x_{kj})$ do not influence the results. This stems from the fact that while estimating the model parameters, for each missing value $x_{kj}$ the same $\mu_{Aikj}(x_{kj})$ is assigned. This results in a model robust to the choice of the assigned value of $\mu_{Aikj}(x_{kj})$. Note that the slight difference between results can be explained by the randomization that occur in each trial.

### *B. Classification*

Table III exhibits the obtained average accuracy and standard deviation (in parenthesis), of the fuzzy models for the different types of models. In this example we randomized the order of the data in each trial, as well as the locations of the missing values. The number of clusters used was $c = 2$. The average RMSE for the fuzzy rule-based model constructed with the complete dataset, was 0.900 and 0.950 for the model with the antecedent membership function $A_{proj}^k$ and $A_{mult}^k$, respectively. In this case we include a comparison with a general data imputation based on the expectation-maximization (EM) algorithm [22] as described in [23]. We use the EM algorithm to complete the data, obtaining $X_{EM}$ and then use FCM to derive a classification model. This strategy is referred as FCM-EM.

TABLE III

ACCURACY FOR THE ALTMAN DATABASE - MEAN (STD)

| Clust | FCM-PDS | | FCM-PDS | | FCM-EM | |
|---|---|---|---|---|---|---|
| Val.Data | $X^V \bigcap X_M^V \neq \emptyset$ | | $X^V \bigcap X_M^V = \emptyset$ | | $X_{EM}^V$ | |
| MV | $A_{proj}^k$ | $A_{mult}^k$ | $A_{proj}^k$ | $A_{mult}^k$ | $A_{proj}^k$ | $A_{mult}^k$ |
| 5% | 0.901 (0.064) | 0.942 (0.042) | 0.912 (0.058) | 0.956 (0.024) | 0.864 (0.058) | 0.937 (0.040) |
| 10% | 0.908 (0.070) | 0.933 (0.055) | 0.912 (0.061) | 0.962 (0.033) | 0.847 (0.072) | 0.928 (0.050) |
| 20% | 0.899 (0.062) | 0.908 (0.064) | 0.914 (0.069) | 0.968 (0.040) | 0.831 (0.084) | 0.908 (0.051) |
| 30% | 0.890 (0.078) | 0.907 (0.057) | 0.927 (0.065) | 0.957 (0.044) | 0.844 (0.082) | 0.892 (0.065) |
| 40% | 0.869 (0.068) | 0.881 (0.082) | 0.923 (0.053) | 0.959 (0.047) | 0.808 (0.089) | 0.903 (0.073) |
| 50% | 0.849 (0.079) | 0.853 (0.084) | 0.919 (0.065) | 0.951 (0.061) | 0.790 (0.098) | 0.868 (0.105) |
| 60% | 0.805 (0.087) | 0.779 (0.109) | 0.904 (0.086) | 0.934 (0.043) | 0.753 (0.096) | 0.824 (0.112) |
| 70% | 0.766 (0.126) | 0.760 (0.108) | 0.859 (0.138) | 0.897 (0.105) | 0.742 (0.112) | 0.718 (0.192) |

In general, the results obtained for the classification problem, are similar to those obtained in the regression problem. Table III shows that the obtained accuracy of the models built with missing values decreases as the number of missing values increases. Also, the obtained accuracy of the models built with missing values increases with the number of missing values. The obtained results using the proposed methodology (FCM-PDS) are in general better than the results obtained with the imputation methodology (FCM-EM), as the imputation of missing values bias the results. When comparing the accuracy obtained with the complete data $X^V \bigcap X_M^V = \emptyset$ and incomplete data $X^V \bigcap X_M^V \neq \emptyset$ it is possible to see that better results are obtained for the complete data. Note that if the percentage of missing values is below 50% the results are better than the average accuracy for the fuzzy rule-based model constructed with the complete dataset. During all our trials we noticed this

fact, and we conjecture that its due to the simplicity of this example combined with the differences both in the antecedent membership functions and the consequents of the models, since the difference is very small. Also for this case, the results obtained with the assignment of the membership value $\mu_{Aikj}(x_{kj})$ to missing value $x_{kj}$, for the case of projected antecedent membership functions, are very similar, so we do not report them.

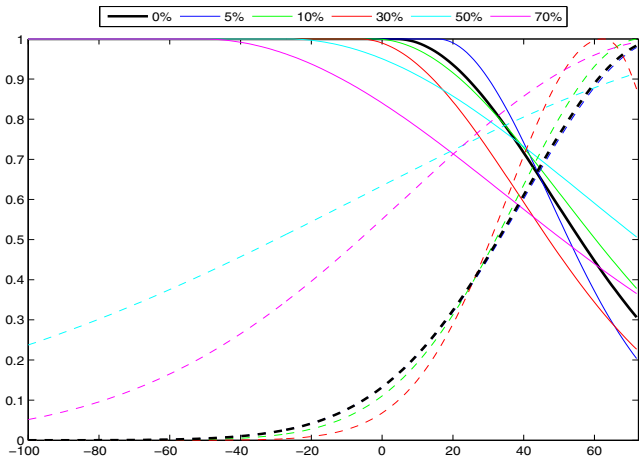


Fig. 1. Antecedent membership functions - % MV - Altman.

Figure 1 shows the obtained antecedent membership functions $A_{ik}$ of models built with different percentages of missing values. It is interesting to note, that the reduction of information caused by the missing values, results in less separation between clusters in the product space. This is reflected in an increase of the support of the fuzzy set $A_{ik}$ with the increase in percentage of missing values. Note that support of a fuzzy set $A$ is defined as $\text{supp}(A) = \{\mathbf{x} \in X | \mu_A(\mathbf{x}) > 0\}$.

## V. Conclusions

Takagi-Sugeno fuzzy models are suitable for identification of nonlinear systems and regression models and can be identified by product-space clustering, but this methodology is not directly applicable if missing values are present. This paper proposes a new approach to data-driven fuzzy modeling of data with missing values. We propose a methodology to identify Takagi-Sugeno fuzzy models by means of product-space clustering of incomplete data. This approach does not require imputation or iterative guess-estimate of the missing values. The methodology is applied to a classification and regression problem. The performance of the obtained models are comparable with the results obtained when using a complete data set, with the added advantage that our method does not require a guess estimation of the true value. Future research will concentrate on applying this methodology to other problems and on how to use the fuzzy models of incomplete data to estimate the missing values.

## References


[1] H. Timm, C. Döring, and R. Kruse, "Differentiated treatment of missing values in fuzzy clustering," in *Proceedings of the 10th International Fuzzy Systems Association World Congress (IFSA 2003), Lecture Notes in Artifical Intelligence*, T. Bilgiç, B. D. Baets, and O. Kaynak, Eds. Istanbul, Turkey: Springer-Verlag, June 2003, vol. 2715, pp. 354–361.

[2] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data, Second Edition*. New York: Wiley and Sons, 2002.

[3] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, 1997.

[4] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 31, no. 5, pp. 735–744, October 2001.

[5] J. Han and M. Kamber, *Data Mining Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2000.

[6] M. Setnes and U. Kaymak, "Fuzzy modeling of a client preference from large data sets: an application to target selection in direct marketing," *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 1, pp. 153–163, 2001.

[7] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recogn.*, vol. 41, no. 12, pp. 3692–3705, 2008.

[8] H. Timm, C. Döring, and R. Kruse, "Fuzzy cluster analysis on partially missing datasets," in *Proc. of the European symposium on Intelligent Techniques, hybrid Systems and their implemetation on smart adaptive systems (EUNITE 2002)*, Albufeira, Portugal, Setember 2002, pp. 421–426.

[9] S. Pospiech-Kurkowska, "Processing of missing data in a fuzzy system," in *Information Technologies in Biomedicine*, 2008, pp. 453–460.

[10] M. R. Berthold and K.-P. Huber, "Missing values and learning of fuzzy rules," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 2, pp. 171–178, 1998.

[11] A. Isaksson, "Identification of arx-models subject to missing data," *Automatic Control, IEEE Transactions on*, vol. 38, no. 5, pp. 813–819, 1993.

[12] H. Timm and R. Kruse, "Fuzzy cluster analysis with missing values," in *Fuzzy Information Processing Society - NAFIPS, 1998 Conference of the North American*, Aug 1998, pp. 242–246.

[13] R. Babuška, *Fuzzy Modeling for Control*. Boston, Dordrecht, London: Kluwer Academic Publishers, 1998.

[14] J. Sousa and U. Kaymak, *Fuzzy Decision Making in Modeling and Control*. Singapore: World Scientific Pub. Co., 2002.

[15] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its aplication to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15, no. 1, pp. 116–132, 1985.

[16] U. Kaymak and R. Babuška, "Compatible cluster merging for fuzzy modelling," in *Proceedings of 1995 IEEE International Conference on Fuzzy Systems, International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium, FUZZ-IEEE/IFES'95*, vol. 2, no. 2, Yokohama, Japan, March 1995, pp. 897–904.

[17] R. Babuška and H. B. Verbruggen, "Constructing fuzzy models by product space clustering," pp. 53–90, 1997.

[18] J. Dixon, "Pattern recognition with partly missing data," *IEEE Transactions on Systems, Man and Cybernetics,*, vol. 9, no. 1, pp. 617–621, 1979.

[19] E. I. Altman, "Financial ratios, discriminate analysis and the prediction of corporate bankruptcy," *Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.

[20] P. M. Murphy, "UCI repository of machine learning databases and domain theories," 1985, http://www.ics.uci.edu/-mlearn/MLRepository.html; ftp.ics.uci.edu: pub/machine-learning-databases.

[21] K. Driessens and S. Džeroski, "Combining model-based and instance-based learning for first order regression," in *ICML '05: Proceedings of the 22nd international conference on Machine learning*. New York, NY, USA: ACM, 2005, pp. 193–200.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[23] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, pp. 853–871, October 2001.