ANALYSIS AND REASONING OF DATA IN THE DATABASE USING FUZZY SYSTEM MODELLING

Dr.E.N.Ganesh
Dean, School of Engineering, VISTAS
Chennai - 600117

Abstract— **In this paper a new fuzzy system modeling algorithm is introduced as a data analysis and approximate reasoning tool. The performance of the proposed algorithm is tested in two different data sets and compared with some well-known algorithms from the literature. In the comparison two benchmark data sets from the literature, namely the automobile mpg (miles per gallon) prediction and Box and Jenkins gas-furnace data are used. The comparisons demonstrated that the proposed algorithm can be successfully applied in system modeling.**

Keywords— **Fuzzy system, Fuzzy clustering, Determination, Membership Functions and Miles Per Gallon.**

I. INTRODUCTION

Informally, one can define data analysis, as the search for structure in data. The data can be viewed as a collection of *n* objects, where each object is represented by means of *NV* attributes. However, unless the structure that is *hidden* in the system is identified, the data provides very little information if at all. Hence the objective of data analysis is to bring the hidden structure to the surface. In many real life situations the data represents an input-output relation. Once the hidden structure of the data is identified it is possible to infer conclusions for new data where only the values of the input attributes are known. This process is known as the reasoning process. In two-valued logic theory this process is carried out through the use of inference rules. One of the most notable inference rules is the modus ponens. Fuzzy system modeling emerged as an alternative approach to two valued data analysis and reasoning approaches. In fuzzy system modeling, the structure is represented by means of *fuzzy if-then* rules. In earlier approaches of fuzzy system modeling the structure hidden in the data, i.e., fuzzy if-then rules, were determined a priori subjectively from other sources such as experts' knowledge. However these rules varied among the experts, even for the same expert at different times. Later, more objective approaches were developed that identify the structure of the data from the historical data [4,8,9]. In fuzzy system modeling, inference is achieved by approximate reasoning. Approximate reasoning can be viewed as a process by which a possible imprecise conclusion is deduced from a collection of imprecise premises. As most of the classical two-valued concepts are "fuzzified" and introduced to the usage of fuzzy set and logic theory, modus ponens is also re-interpreted in fuzzy set theory. Zadeh [11] introduced the Generalized Modus Ponens (GMP) and provided a methodology known as Compositional Rule of Inference (CRI) that can be used to infer fuzzy consequents from given fuzzy premises. CRI may be defined as follows;

Let A be a fuzzy set in universe of discourse U and B be a fuzzy set in universe of discourse V. Let the fuzzy "rule" $A \Rightarrow B$ and the fuzzy "fact" A^* is given then

$$B^* = A^* o(A \Rightarrow B) \qquad (1)$$

$$\mu_{B^*}(y) = \vee_x (\mu_{A^*}(x) \wedge \mu_{A \Rightarrow B}(x,y)) \qquad (2)$$

Many authors proposed systematic approaches for fuzzy system modeling. Among those most notable one is the approached proposed by Sugeno-Yasukawa [9] which was further investigated by Nakanishi, Turksen and Sugeno [7]. Turksen-Bazoon [8] further made some improvements to this model and successfully applied in pharmacological data analysis. Later Emami *et. al.* [4] proposed a parametric approach for fuzzy system modeling and applied it to robotics[4]. In this paper we are going to provide a new fuzzy system modeling algorithm and demonstrate how it can be used as a data analysis and approximate reasoning tool.

II. FUZZY SYSTEM MODELING ALGORITHM

The most common problem with the existing fuzzy system modeling (FSM) approaches is their lack of a global structure in terms of system modeling. The concepts are imported from different domains such as pattern recognition, approximate reasoning, etc., and are adapted to new situations without particular care. The main difference in the proposed approach is that it is designed to adapt the system modeling to particular system behavior patterns.

The basic steps of the proposed approach are similar to the existing ones.

1. Fuzzy clustering of the output
2. Determination of the significance of the input

variables

3. Input membership assignment
4. Inference

The first three steps are the structure identification parts and the fourth step is the reasoning part.

A. Fuzzy Output Clustering

Clustering unlabeled data $X=\{X_1, X_2, \ldots X_n\} \subset R^p$ is the assignment of labels to the vectors in X and, hence, to the objects of X. In the case of hard clustering each point becomes a member of one cluster and has no membership in the rest of the clusters. However, in the case of fuzzy clustering each data point may be a member of more than one fuzzy cluster to a certain degree which is in [0,1].

Among various fuzzy clustering algorithms the most common one is Fuzzy C-Means algorithm developed by Bezdek [1]. This algorithm is basically an objective function clustering algorithm based on FCM theorem [2]. FCM theorem states the necessary membership degrees of each individual to each cluster and the cluster centers is made in order to minimize the well known weighted within group sum of square error. However, this algorithm creates membership degree harmonics and semantically wrong boundary clusters [6]. The proposed output clustering algorithm solves both of these problems. Briefly, the proposal suggests the determination of the cluster centers (v_i) by applying any clustering algorithm in the literature, and to sort the cluster centers in ascending order such that $v_{[i]} < v_{[i+1]}$. The membership degree of the k^{th} data point (y_k) in the i^{th} cluster (B_i) and $i+1^{th}$ cluster (B_{i+1}) is determined as follows,

if $v_{[i]} \leq y_k \leq v_{[i+1]}$

$$\mu_{(Bi)}(y_k) = \|y_k - v_{[i+1]}\| / \| v_{[i]} - v_{[i+1]}\|$$

$$\mu_{(Bi+1)}(x_k) = \|y_k - v_{[i]}\| / \| v_{[i]} - v_{[i+1]}\|$$

otherwise,

$$\mu_{(Bi)}(y_k) = \mu_{(Bi+1)}(y_k) = 0$$

Furthermore the boundary sets are corrected by assigning full membership degrees for the points that are smaller (or larger) than the cluster center of the smallest (or largest) clusters in order to protect the semantic soundness.

In fuzzy clustering literature one of the major problems is the cluster validity problem, i.e. determination of the number of clusters c and selection of the m value (level of fuzziness). Many different functions are proposed in the literature [3,4,9,10] and the determination of the (m,c) pair is based on the optimization of these functions. However, the optimum (m,c) pair is not necessarily the best selection in terms of fuzzy system modeling approach. In the proposed approach the selection of c is based on minimization of the modeling error, which is a more suitable approach in terms of fuzzy system modeling performance. The proposed clustering method eliminates the level of fuzziness for Type 1 fuzzy system modeling.

B. Input Membership Assignment

Prior to input membership assignment one must determine the significance (as will be discussed in the next section) of the input variables. However, the proposed methodology is based on the modeling error. Therefore, first the proposed formation of the rule base will be introduced.

There are various ways of constructing the fuzzy rule base. In the literature the most common technique is to determine the input membership degrees by projecting the output fuzzy clusters onto input space [9,4]. The major problem with this approach is that the natural links between the input variables are ignored and they are assumed to be independent from each other. Furthermore the existing algorithms assume a single convex input fuzzy set for each fuzzy output set. It is demonstrated that both assumptions (independence of input variables, and convexity of input fuzzy set) produce invalid rule structures [6]. Hence a new approach is proposed where the output fuzzy clusters are projected onto n-dimensional input space. The second step is the classification of the intermediate values, which is achieved usually by fitting a line. This approach further approximates the representation of the hidden rules. In terms of providing a graphical rule base that explains the hidden rules, this approach is valuable. However, there is no need to make calculations with the approximated function. In the proposed algorithm the background calculations are done by using the actual data itself and a function is fit to the existing data only to provide a graphical representation of the hidden rules.

Suppose there are *ND* number of data vectors, and *NV* number of input variables. Let $X_1, X_2, \ldots X_{NV}$ be *NV* fuzzy linguistic variables in the universe of discourse $U_1, U_2, \ldots U_{NV}$ and Y be a fuzzy variable in the universe of discourse V. Let $X_k = [x_{k,1}, \ldots x_{k,NV}]$ denotes the input vector of the k^{th} data and y_k is the output. A data vector D_k is represented with (X_k, y_k) where X_k is a vector of *NV*-dimension. Let A_i be a fuzzy set in the universe of discourse ($U_1 \times U_2 \times \ldots U_{NV}$) and B_i be fuzzy set in universe of discourse V.

C. Algorithm Input Membership Assignment

1.Cluster the output variables, say c clusters, and determine $\mu_{Bi}(y_k)$, i.e., the membership degree of the k^{th} data in the i^{th} output cluster.

2.Use the reinterpretation of the extension principle and assign the same membership degrees for the n-dimensional input vector for each corresponding input variable and create a fuzzy input cluster A_i where the elements are n-dimensional vectors for each corresponding B_i.

$$\mu_{Ai}(X_k) = \mu_{Bi}(y_k)$$

Hence i^{th} rule, R_i is formed as;
R_i : If X isr A_i Then Y is B_i

3.Obtain a projected rule base for each input variable *independently* in order to provide a graphical representation of the rule base.

D. Significant Input Selection

In the proposed algorithm the selection of the significant input variable concept is fuzzified. In the literature various methodologies are proposed, such as Sugeno-Yasukawa's [9] RCI method that adds an input variable that increases the modeling performance most with a nearest neighbor strategy until the objective does not improve any more, or Emami *et. al.* [4] method where the significant inputs are determined by comparing the core of the trapezoidal input fuzzy clusters. However both of these algorithms determine if the input variable is significant or insignificant, dichotomously. In this work a new approach is provided which determines the degree of significance of the input variables. In real life, to classify an input, as significant or not, doesn't reflect reality since some variables are *more* significant then others. This approach may be interpreted as determining the similarity of the data with a weighted Euclidean distance measure. A hill-climbing algorithm similar to simulated annealing is proposed in order to determine the weight vector associated with the input variables.

E. Algorithm Significance Determination

```
1.Initialize the significance of each input, Sig(j)=1/NV
2.While the termination criteria not satisfied do
    For i=1 to number of input variables (NV) do

    1.Increase the significance of j'th input by $\varepsilon$, Sig(j):=Sig(j)+ $\varepsilon$
    2.Decrease the significance of the remaining inputs by $\varepsilon$ /(NV-1) temporary error:= 0
    3.For k=1 to number of training data do
    4.Form a rule base by using the (training data–$k^{th}$ data )
    5.Predict the error for $k^{th}$ data
    6.Temporary error:=temporary error + error
    7.Average error:=temporary error/number of training
      data
7.Select the minimum two average errors obtained for each increment of significance of input variables and select the best one randomly Save the minimum error found until this stage and the significance combination that is used to reach this minimum error
8.The significance combination is the one that produces the minimum error
```

Recall that the number of clusters is selected with respect to the model error. Hence, the above algorithm is iterated for each possible cluster size. The random selection in *step 2.2* is used in order to avoid cycles while searching the space of alternatives. Also, one must be careful at *step 2.1.2* to avoid obtaining negative significance degrees. This is achieved by not allowing negative significance and redistributing and normalizing the significance values at the end of each iteration. In this way, the summation of the significance values for the inputs will be one after each iteration. There are possible termination criteria that may be used such as a fixed number of iterations or termination if some number of consecutive iterations does not modify the minimum error by a certain amount. In this study we used the former approach and set the iteration size to 100. Finally the best cluster size and significance combination in terms of the average training error is used in the final model.

III. FUZZY INFERENCE

The first three steps of the proposed approach, described above, are the parts where the fuzzy if-then rules are identified, i.e. structure hidden in the data. By using the rule base so developed, one can infer the output for a new input data. The following general schema achieves the fuzzy inference,

1- Determine degree of firing for each rule $\tau_i(X^*)$
2- Infer by using an implication operator
3- Aggregate the output of each rule
4- Defuzzify the output

Some modifications are necessary for the modeling approach that is proposed. First of all degree of firing of each rule cannot be calculated as a conjunction of each separate input variable because the in the proposed approach the input variables are not treated as separate independent features but kept as an *n*-dimensional data (object) vector. For this purpose a *k-nearest neighborhood algorithm* is proposed in order to determine the membership degree of the given input in the *n*-dimensional antecedent input cluster. Basically the degree of matching is determined by first determining the closest *k-NN* based on the weighted Euclidean distance measure where the weights associated with each dimension is the significance degree of the input variables. Based on the nearest neighbors, one can determine the degree of firing of each rule. Recall that based on the proposed schema each training data has total membership degree equal to 1 in each rule and is a member of only two clusters. Hence the test data is also restricted to this constraint. Further details of this approach can be found in [6].

The proposed inference schema is similar to the position gradient methodology proposed by Sugeno [9],

$$y^{*} = \Sigma(\tau_i(X^{*})\times v_{[i]}) \qquad (3)$$

where $v_{[i]}$ is the cluster center of the i^{th} rule and $\tau_i(X^{*})$ is the degree of match (or firing) of the test data, i.e., estimated membership degree of X^{*} to the n-dimensional cluster of A_i, antecedent of the i^{th} rule.

IV. EXPERIMENTAL RESULTS

The proposed algorithm is applied to two benchmark data sets available in the literature, namely the automobile miles per gallon (mpg) prediction data and Box and Jenkins gas-furnace data.

4.1 Automobile Miles Per Gallon (MPG) Prediction

Automobile MPG prediction is a six input single output regression problem. The gasoline consumption of the cars are to be predicted based on some of their attributes. These attributes are number of cylinders, displacement, horsepower, weight, acceleration and model year. The original data is available in ftp://ftp.ics.uci.edu/pub/machine-learning-databases/auto-mpg/auto-mpg.names . After removing the data vectors that have some missing attribute values 392 data vectors are left. Two thirds of the data is randomly selected as the training set and the remaining data is used as the test set. For this data set a comparison is made with the Turksen-Bazoon [8] algorithm, which is a slightly modified version of the well known Sugeno-Yasukawa method [9] where the *m* (level of fuzziness) is selected as 2. The RMSE for the prediction of the test data results for Turksen-Bazoon (T-B) and the proposed approach (PA) is presented in Table 1. In Fig. 1, the actual *vs.* the prediction of the proposed algorithm of miles per gallon consumption for 145 test data is presented.

Table 1. The comparison of the predictive performances in terms of RMSE of the prediction.

	T-B	PA
RMSE	3.29	2.61

Jang [5] also analyses the same data with the Adaptive Network-based Fuzzy Inference System (ANFIS). It is not possible to compare the proposed algorithm with the results presented in [5] because the same train and test data is not used. However in order to provide some insight Jang presents a test error of 2.98 and a training error of 2.61.The significance degrees of the input variables obtained by the proposed schema, Turksen-Bazoon model and ANFIS is presented in Table 2. Note that the selection for the ANFIS is obtained from [5].

Table 2. The significance degrees of the input variables. S represents that the input variable is considered as significant and I represents that it is insignificant for T-B and ANFIS.

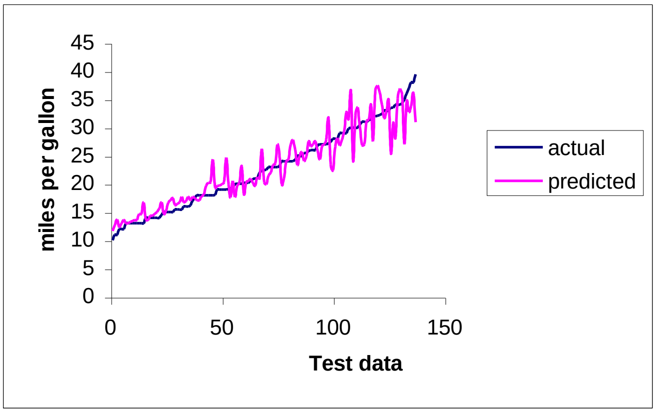

Figure 1. The actual *vs.* predicted graph of gasoline consumption of the proposed approach for the test data.

4.2 Box and Jenkins Gas-Furnace Data

Box and Jenkins gas furnace data is a single input single output time series data for a gas furnace process with gas flow rate *u(t)* as the input and *y(t)*, the CO_2 concentration as the output. Sugeno -Yasukawa [9] considered 10 input variables which are *y(t-1), y(t-2), y(t-3), y(t-4), u(t-1), u(t-2), u(t-3), u(t-4), u(t-5)* and *u(t-6)* as candidates to effect the output *y(t).* The original data set contains 296 data pairs and with these settings only 290 of them can be used. The proposed approach is applied and the results are compared with the Turksen and Bazoon [8] model and the Adaptive Network-based Fuzzy Inference System (ANFIS) proposed by Jang [5]**.** The first 145 data is used as the training and the next 145 data is used as the test data as was suggested in [5]. The RMSE of the prediction of the test data for each algorithm is presented in Table 3. The exact RMSE of the test prediction of ANFIS is not tabulated explicitly in [5], but a figure is presented where you can interpolate that the test data prediction RMSE for ANFIS is greater than 0.52. In Table 4, the significant variables (or the degree of significances) are presented. For the proposed approach the significance determination algorithm proposed in [6] and for ANFIS the significant inputs presented in [5] is used**.** Note that in the tables Turksen-Bazoon [8] approach is denoted as T-B and the proposed approach is denoted as PA.

Table 3. The comparison of the results in terms of predictive performances of the test data.

	T-B	ANFIS	PA
RMSE	1.03	0.52	0.43

Table 4. The significance degrees of the input variables. S represents that the input variable is considered as significant and I represents that it is insignificant for T-B and ANFIS.

From Table 3, the proposed schema performs better than the other two algorithms in terms of the test data prediction RMSE. The proposed approach has an RMSE 0.42, where as Turksen-Bazoon model has the largest RMSE of 1.03 and ANFIS has a prediction RMSE at least 0.52. From the structure identification process *y(t-1)*, *u(t-4)* and *u(t-6)* are specified to be the most significant inputs and some significance is assigned to *y(t-2)* and *u(t-3)***.**The actual *vs.* prediction of the CO_2 level of the 145 test data is presented in Fig. 2.

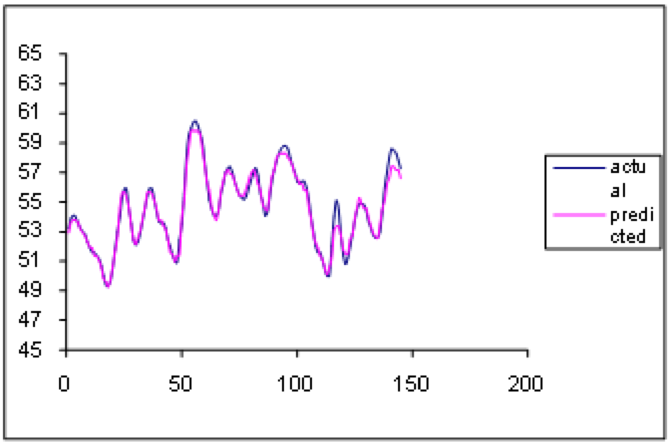

Figure 2. The actual *vs.* predicted graph of the test data for CO_2 level at time *t* for the proposed approach

V. CONCLUSION

In this paper, a new approach for structure identification problem is presented. The proposed algorithm preserves the natural links between the input variables and treats them as an *n*-dimensional input vector. Also a new approach of significance degrees for input variables is introduced, and a probabilistic hill-climbing algorithm is proposed. It is demonstrated that the proposed fuzzy system modeling algorithm can be used effectively for data analysis and approximate reasoning.

VI. REFERENCES

[1] Bezdek J. C., 1981 "Pattern recognition with fuzzy objective function algorithms", *Plenum Press*, New York and London.
[2] Bezdek J. C., 1973, "Fuzzy mathematics in pattern classification", *Ph. D. thesis*, Cornell University, Ithaca, NY.
[3] Bezdek J. C.; "Cluster validity with fuzzy sets", *Journal of Cybernetics*, 3, 58-72, 1974.
[4] M.R. Emami, I.B. Turksen, A.A.Goldenberg, 1999 "A unified parametrized formulation of reasoning in fuzzy modeling and control" *Fuzzy Sets and Systems*, 108, 59-81.
[5] Jang J.R.,1996 "Input Selection for ANFIS Learning", *Proceedings of IEEE*, 1493-1499.
[6] Kilic K, Sproule B.A., Turksen I.B, Naranjo C.A., 2002 "Fuzzy System Modeling in Pharmacology: A new Algorithm", *Fuzzy Sets and Systems*, 130 (2), 253-264.
[7] Nakanishi H., Turksen I.B, Sugeno M., 1993"A review and comparison of six reasoning methods", *Fuzzy Sets and Systems*, 57, 257-294.
[8] Sproule B.A., Bazoon M., Shulman K.I., Turksen I.B., Naranjo C.A.,1997 "Fuzzy logic pharmacokinetic modeling: Application to lithium concentration prediction", *Clinical Pharmacology Therapy*, 62, 29-40.
[9] Sugeno M., Yasukawa T.A.,1993 "A Fuzzy Logic Based Approach to Qualitative Modeling", *IEEE Transactions on Fuzzy Systems*, 31, 7-31.
[10] Xie X.L. and Beni G.A., 1991 "Validity measure for fuzzy clustering", *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 3, 841-846.

[11] Zadeh L.A.,1973 "Outline of a new approach to the analysis of complex systems and decision processes", *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3, 28-44.

Table 2. The significance degrees of the input variables. S represents that the input variable is considered as significant and I represents that it is insignificant for T-B and ANFIS.

1)	*#Cylinders*	Displacement	Horsepower	Weight	*Acceleration*	Model
P.A	0.08	0.15	0.23	0.30	0.12	0.12
T-B	I	I	S	S	I	S
ANFIS	I	I	I	S	I	S

Table 4. The significance degrees of the input variables. S represents that the input variable is considered as significant and I represents that it is insignificant for T-B and ANFIS.

1) Algorithms	*y(t-1)*	*y(t-2)*	*y(t-3)*	*y(t-4)*	*u(t-1)*	*u(t-2)*	*u(t-3)*	*u(t-4)*	*u(t-5)*	*u(t-6)*
T-B	S	I	I	I	I	I	I	I	I	S
ANFIS	S	I	I	I	I	I	S	I	I	I
PA	0.59	0.02	0.00	0.00	0.00	0.00	0.04	0.16	0.00	0.18