Constructing accurate and parsimonious fuzzy models with distinguishable fuzzy sets based on an entropy measure

Zhou, Shang-Ming

http://hdl.handle.net/10026.1/20378

Constructing Accurate and Parsimonious Fuzzy Models with Distinguishable Fuzzy Sets Based on an Entropy Measure

Shang-Ming Zhou and John Q. Gan

Department of Computer Science
University of Essex
Colchester CO4 3SQ
United Kingdom
szhoup@essex.ac.uk; jqgan@essex.ac.uk

Abstract. **Parsimony is very important in system modeling as it is closely related to model interpretability. In this paper, a scheme for constructing accurate and parsimonious fuzzy models by generating distinguishable fuzzy sets is proposed, in which the distinguishability of input space partitioning is measured by a so-called "local" entropy. By maximizing this entropy measure the optimal number of merged fuzzy sets with good distinguishability can be obtained, which leads to a parsimonious input space partitioning while preserving the information of the original fuzzy sets as much as possible. Different from the existing merging algorithms, the proposed scheme takes into account the information provided by input-output samples to optimise input space partitioning. Furthermore, this scheme possesses the ability to seek a balance between the global approximation ability and distinguishability of input space partitioning in constructing Takagi-Sugeno (TS) fuzzy models. Experimental results have shown that this scheme is able to produce accurate and parsimonious fuzzy models with distinguishable fuzzy sets.**

Keywords **- Interpretability; distinguishablity; fuzzy set merging; entropy; parsimonious fuzzy model.**

* **Corresponding author:**

Shang-Ming Zhou

Email: szhoup@essex.ac.uk; shmzhou@sohu.com

Tel: 0044-(0)1206-872136, Fax: 0044-(0)1206-872788

Constructing Accurate and Parsimonious Fuzzy Models with Distinguishable Fuzzy Sets Based on an Entropy Measure

Shang-Ming Zhou and John Q. Gan

Department of Computer Science
University of Essex
Colchester CO4 3SQ
United Kingdom
szhoup@essex.ac.uk; jqgan@essex.ac.uk

Abstract. **Parsimony is very important in system modeling as it is closely related to model interpretability. In this paper, a scheme for constructing accurate and parsimonious fuzzy models by generating distinguishable fuzzy sets is proposed, in which the distinguishability of input space partitioning is measured by a so-called "local" entropy. By maximizing this entropy measure the optimal number of merged fuzzy sets with good distinguishability can be obtained, which leads to a parsimonious input space partitioning while preserving the information of the original fuzzy sets as much as possible. Different from the existing merging algorithms, the proposed scheme takes into account the information provided by input-output samples to optimise input space partitioning. Furthermore, this scheme possesses the ability to seek a balance between the global approximation ability and distinguishability of input space partitioning in constructing Takagi-Sugeno (TS) fuzzy models. Experimental results have shown that this scheme is able to produce accurate and parsimonious fuzzy models with distinguishable fuzzy sets.**

Keywords **- Interpretability; distinguishablity; fuzzy set merging; entropy; parsimonious fuzzy model.**

I. INTRODUCTION

When adaptive learning algorithms are introduced into a fuzzy inference system, there might be a loss in the interpretability or transparency of the fuzzy system. Specifically speaking, undistinguishable fuzzy sets, as shown in Fig. 1, are often generated by most adaptive learning

1

algorithms due to their accuracy-oriented nature. As a result, it is difficult to assign distinct linguistic labels and semantic meaning to these fuzzy sets. This problem has drawn much attention recently [1]-[15].
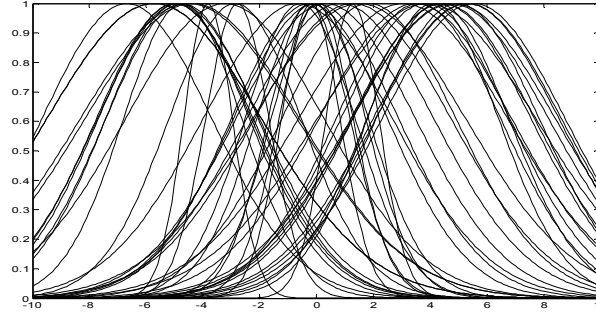

Fig. 1 Fuzzy sets without distinguishability

As one of the main objectives in system modeling, the development of reliable and transparent models is crucial for human beings to understand a real world system or natural phenomenon. Interpretable fuzzy system modeling is to produce a model that not only performs good global prediction, but also generates a parsimonious and understandable rule base, according to the Occham's razor (The simplest hypothesis or theory agreeing with the facts is the best one). In order to produce parsimonious fuzzy models, the ASMOD (Adaptive Spline Modeling of Observational Data) algorithm was developed [16] for constructing 0-order TS fuzzy models with ANOVA decomposition, and the MASMOD algorithm, a modified version for ASMOD, was proposed in [17] to construct first-order TS fuzzy models based on ANOVA decomposition. Furthermore, Gan and Harris suggested a hybrid learning scheme [18] combining the expectation-maximization (EM) algorithm, which is a technique for maximum likelihood or maximum *a posteriori* estimation [19], and the MASMOD algorithm for fuzzy local linearization modeling. The above-mentioned ANOVA decomposition based construction algorithms produce parsimonious models by adding submodels on lower-dimensional spaces into the model which starts from an empty model. This submodel addition process is controlled by some complexity penalty criteria such as structural risk minimization [16]. There are two possible problems in this type of algorithms. The first is that the model construction

process is slow due to the exhaustive search of effective submodels; whereas the second problem is that the model construction process may stop too early without achieving the expected model accuracy.

The motivation of this paper is to construct parsimonious fuzzy models by generating distinguishable fuzzy sets and preserving the global model accuracy at the same time. It is well known that in transparent fuzzy modeling, the interpretability of fuzzy models heavily depends on human's prior knowledge [5]. However, if there is no prior knowledge available, concise and distinguishable fuzzy sets which lead to a parsimonious partitioning are more interpretable and should be adopted in fuzzy modeling. As a matter of fact, the distinguishability condition is regarded as one of the most important aspects for the interpretability of fuzzy models [6]-[9]. In order to produce distinguishable fuzzy sets in fuzzy models, Setnes *et al* suggested a procedure to design triangular membership functions (MFs) by ordering the cores or modal values of fuzzy sets for clearly distinguishing their linguistic terms [10]. Espinosa and Vandewalle proposed an algorithm, named FuZion, to merge MFs whose cores are "too close" to each other [11]. In [12], an agglomerative approach based double-clustering technique was suggested to generate distinguishable fuzzy granules. Based on a multi-objective optimization method, Wang *et al* presented a fuzzy set agent based evolutionary approach to extract interpretable fuzzy rules, in which an interpretability based regulation strategy for merging similar fuzzy sets was used [13]. To obtain comprehensible initial fuzzy models, Roubos *et al* developed a covariance based model initialization method [14]. And for the effective initialization of interpretable fuzzy classifiers, a branch merging algorithm based on fuzzy value clustering of branches was applied in [15].

As a matter of fact, most existing algorithms for generating distinguishable fuzzy sets merge similar fuzzy sets based on similarity measures between fuzzy sets [20][21]. Although the similarity measure based merging procedure seems simple and is effective in many cases, extensive search is usually required, which leads to high computational load. More importantly, this kind of procedure partitions the input space based on input data only without using the information contained in the output data.

However, for system identification or modeling, the output data contains useful information for input space partitioning. Castellano *et al* suggested a possibility measure to replace the similarity measure of distinguishability, aiming to reduce computational load during the fuzzy set merging process [9], but the possibility measure based merging procedure still ignores the information provided by output data. Inspired by a deviation index suggested to measure the conciseness of input space partitioning [4], this paper proposes a "local" entropy to characterize the distinguishability of input space partitioning, based on which the distinguishability of input space partitioning can be optimized while the information in the original fuzzy sets is preserved as much as possible. Furthermore, this paper proposes a scheme to seek a balance between global approximation ability and distinguishability of input space partitioning based on this "local" entropy measure. The advantage of the proposed scheme is that it takes into account the information provided by input-output sample data pairs during the fuzzy set merging process. In such a way, accurate and parsimonious fuzzy models can be constructed from data using adaptive learning algorithms.

The organization of this paper is as follows. Section II proposes a compactness index for data space partitioning and defines a "local" entropy based on this index as a criterion to measure the distinguishability of input space partitioning. In section III, a fuzzy set merging algorithm is developed by using this entropy. Section IV presents a scheme for seeking a balance between global approximation ability and distinguishability of input space partitioning in constructing TS fuzzy models. Experimental results are given in section V, and section VI concludes this paper.

II. A "LOCAL" ENTROPY TO MEASURE THE DISTINGUISHABLITY OF FUZZY SETS

For a given training sample set $\Xi := \left\{ (x^{(i)}, d^{(i)}) \right\}_{i=1}^{N}$, where $x^{(i)} = (x_1^{(i)}, \cdots, x_n^{(i)})^T \in X^n$ and $d^{(i)} \in Y$, a so-called sample fuzzy subset $\tilde{z}^{(i)}$ is defined on each sample as follows:

$$\widetilde{z}^{(i)}(z) = \exp\left(-\frac{\left\|z - z^{(i)}\right\|^2}{2\sigma^2}\right) \qquad (1)$$

where $z^{(i)} = (x^{(i)^T}, d^{(i)})^T \in \Xi \subset X^n \times Y$, and σ is a width parameter for the exponential function. Furthermore, a compactness index of this sample fuzzy subset is defined by

$$C(z^{(i)}) = \frac{\displaystyle\sum_{j \in \{j \neq i: \|z^{(i)} - z^{(j)}\| \leq \delta\}} \widetilde{z}^{(i)}(z^{(j)})}{\displaystyle\sum_{i=1}^{N}\left(\sum_{j \in \{j \neq i: \|z^{(i)} - z^{(j)}\| \leq \delta\}} \widetilde{z}^{(i)}(z^{(j)})\right)} \qquad (2)$$

where δ is a parameter indicating the size of the neighborhood of $z^{(i)}$, which is estimated as $\delta = 2\max_i\left(\min_{j(\neq i)}(\|z^{(i)} - z^{(j)}\|)\right)$. In terms of the definition by (2), large $C(z^{(i)})$ indicates that there are many fuzzy sets congregating around fuzzy set $\widetilde{z}^{(i)}$, whilst small $C(z^{(i)})$ reveals that there are few fuzzy sets close to $\widetilde{z}^{(i)}$. This property of the compactness index suggests that it bears an analogy with the data density around the center $z^{(i)}$ and contains useful information for fuzzy set merging. More importantly, because $\widetilde{z}^{(i)}$ are multidimensional fuzzy subsets defined on the input-output space, merging similar fuzzy sets on input space based on the compactness indices of sample fuzzy subsets actually takes into account the information provided by the output samples, which is very useful in fuzzy modeling. Obviously, this strategy is different from the most currently used fuzzy sets merging procedures in literature. However, it is difficult to assign meaningful semantic terms to multidimensional fuzzy sets [2], so most fuzzy modeling methods represent fuzzy rules in terms of fuzzy sets defined on each one-dimensional subdomain of input space. In this paper, the compactness index of sample fuzzy subset $\widetilde{x}_j^{(i)}$ on one-dimensional subdomain $X_{(j)}$ of X^n is calculated by projecting the compactness index of the multidimensional fuzzy subset onto subspace $X_{(j)}$ as follows:

$$C_j(x_j^{(i)}) = \max_{z \in \Xi} C(z \mid x_j = x_j^{(i)}) \qquad (3)$$

Now let us define the compactness index of a normal fuzzy set by associating it with the above defined sample fuzzy subset. For fuzzy sets $\{A_m^{(j)} \mid m = 1, \cdots, L_j\}$ on $X_{(j)}$, where L_j represents the number of fuzzy sets defined on $X_{(j)}$, the core center of $A_m^{(j)}$ is evaluated as

$$\beta_m^{(j)} = \frac{1}{2}\left(\beta_{m1}^{(j)} + \beta_{m2}^{(j)}\right) \tag{4}$$

where $\beta_{m1}^{(j)}, \beta_{m2}^{(j)}$ are the lower bound and upper bound of the core region of $norm(A_m^{(j)})$ that is defined as $norm(A_m^{(j)})(x_j) = A_m^{(j)}(x_j) / \sup_{x_j}(A_m^{(j)}(x_j))$. The compactness index of fuzzy set $A_m^{(j)}$ is defined by

$$MC_j(\beta_m^{(j)}) = \max_{x_j^{(i)} \in \Delta_m^{(j)}} C_j(x_j^{(i)}) \tag{5}$$

where $\Delta_m^{(j)} := \left\{ x_j^{(i)} \,\middle|\, |x_j^{(i)} - \beta_m^{(j)}| < \delta_j \right\}$, and δ_j is a parameter indicating the size of the neighborhood of $\beta_m^{(j)}$, which is estimated as $\delta_j = 2\max_i\left(\min_{m(\neq i)}(|x_j^{(i)} - x_j^{(m)}|)\right)$.

Assume the data space $\beta^{<j>} = \left\{\beta_m^{(j)} \mid m = 1, \cdots, L_j\right\}$ is partitioned into $\psi_k^{<j>}$, $k=1, 2, ..., K_j$, where $\psi_k^{<j>}$ is a subset of $\beta^{<j>}$ and $\beta^{<j>} = \bigcup_{k=1}^{K_j} \psi_k^{<j>}$. Based on the compactness index (5), an entropy is defined as follows to measure the distinguishability of this partition:

$$LE_{K_j}^j = LE\left(\left\{\psi_k^{<j>} \,\middle|\, k = 1, \cdots, K_j \right\}\right) = \sum_{k=1}^{K_j} E_k^j \tag{6}$$

where

$$E_k^j = -\sum_{\beta_m^{(j)} \in \psi_k^{<j>}} \left(\frac{MC_j(\beta_m^{(j)})}{S_k^j}\right) \log\left(\frac{MC_j(\beta_m^{(j)})}{S_k^j}\right) \tag{7}$$

and

$$S_k^j = \sum_{\beta_m^{(j)} \in \psi_k^{<j>}} MC_j(\beta_m^{(j)}) \quad (k = 1, \cdots, K_j) \tag{8}$$

6

For the sake of clarity in calculating the entropy, a partition matrix $U^{<j>}$ with the size of $K_j \times L_j$ is defined as follows:

$$U_{km}^{<j>} = \begin{cases} 1 & if \ \beta_m^{(j)} \in \psi_k^{<j>} \\ 0 & otherwise \end{cases} \tag{9}$$

and the entropy can be reformulated as

$$LE_{K_j}^j = -\sum_{k=1}^{K_j}\sum_{m=1}^{L_j} \widehat{U}_{km}^{<j>} \cdot MC_j(\beta_m^{(j)}) \log(\widehat{U}_{km}^{<j>} \cdot MC_j(\beta_m^{(j)})) \tag{10}$$

where

$$\widehat{U}_{km}^{<j>} = U_{km}^{<j>} \Big/ \left(\sum_{m=1}^{L_j} U_{km}^{<j>} \cdot MC_j(\beta_m^{(j)}) \right) \tag{11}$$

The above entropy can be regarded as an extension of the *a posteriori* entropy suggested by Kapur *et al* for image segmentation [22]. It is known that the entropy criterion in [22] is effective for finding an optimal grey level threshold for image segmentation. Therefore, it is expected that by maximizing the entropy (6) an optimal data space partition, denoted as $\beta^{<j>*} = \bigcup_{k=1}^{K_j^*} \psi_k^{<j>*}$ ($K_j^* < K_j$), can be obtained, in which $\psi_k^{<j>*}$ should be distinctively separated from each other. It should be noted that the proposed entropy measure (6) is different from the relative entropy measure in [4][5] for measuring the deviation of MFs from symmetry and the conciseness of fuzzy models. And the proposed entropy measure is also different from the fuzzy entropy proposed in [23] which was defined in terms of the membership degrees without considering the analogous density information around the core centers of fuzzy sets.

Because only neighboring samples are considered in (2), the entropy measure defined in (6) is called "local" entropy. The "local" concept in the compactness definition is important. When δ is large enough to cover all the samples, the "local" concept no longer exists. In the case of non-uniform samples, different values of δ will cause little difference. However, when the samples are uniform,

the entropy without using the "local" concept may not work as a measure of distinguishability. For example, in Fig. 2 for a non-uniform data set, the compactness indices with and without using the "local" concept are similar. However, as shown in Fig. 3, for a uniform data set, the compactness distributions based on the two indices are different. Fig. 3(bottom) indicates that the compactness indices around the samples are the same except for the marginal samples, which is what we expect for a uniform data set. We suggest using the compactness index defined by (2) to characterize the distinguishability of input space partitioning.
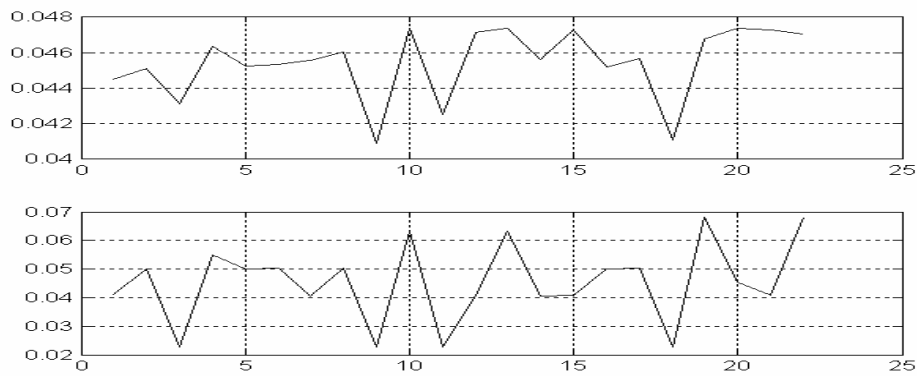
Fig. 2 Comparison of compactness indices on a non-uniform data set: compactness index based on all the samples (top), and compactness index based on neighboring samples only (bottom).
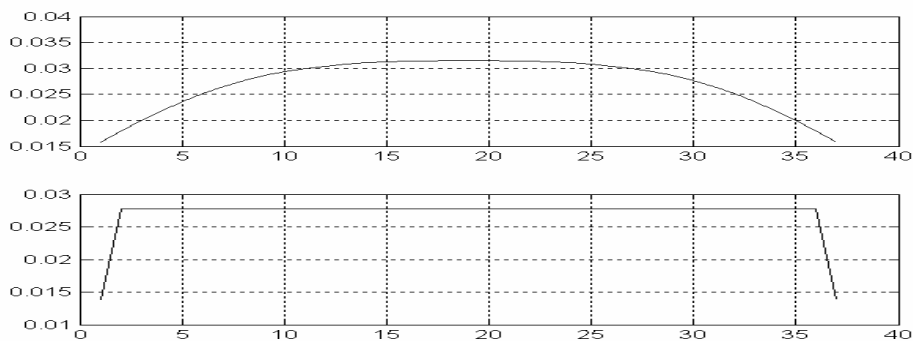

Fig. 3 Comparison of compactness indices on a uniform data set: compactness index based on all the samples (top), and compactness index based on neighboring samples only (bottom).

III. A FUZZY SET MERGING (FSM) ALGORITHM

If there is no prior knowledge available, fuzzy sets generated by accuracy-oriented data-driven modeling methods such as neuro-fuzzy algorithms [24][25] are most likely undistinguishable. The FSM algorithm developed in this section, as delineated in Fig. 4, can be used to merge the undistinguishable fuzzy sets, producing new fuzzy sets with optimal distinguishability in terms of maximum "local" entropy.
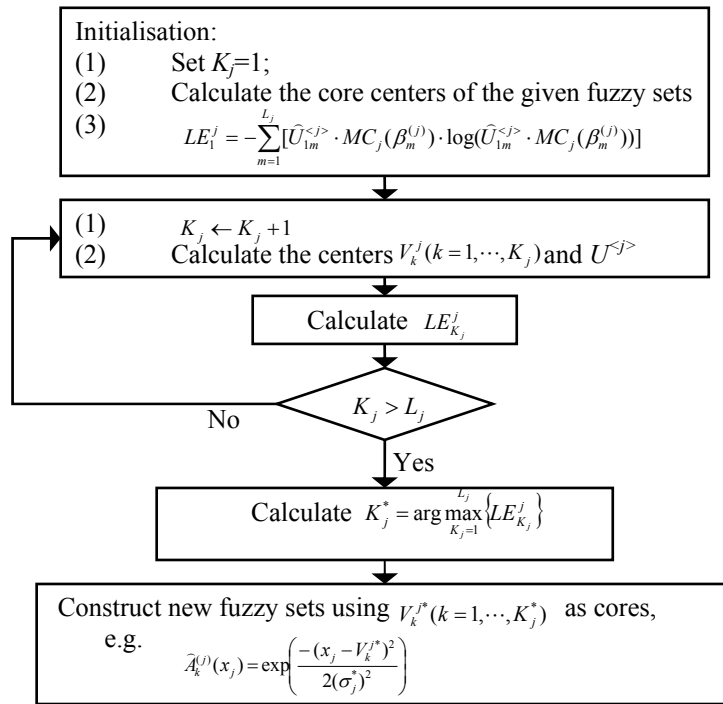

Initialisation:
(1)   Set $K_j=1$;
(2)   Calculate the core centers of the given fuzzy sets
(3) $LE_1^j = -\sum_{m=1}^{L_j}[\widehat{U}_{1m}^{<j>} \cdot MC_j(\beta_m^{(j)}) \cdot \log(\widehat{U}_{1m}^{<j>} \cdot MC_j(\beta_m^{(j)}))]$

(1) $K_j \leftarrow K_j + 1$
(2)   Calculate the centers $V_k^j(k=1,\cdots,K_j)$ and $U^{<j>}$

Calculate $LE_{K_j}^j$

$K_j > L_j$ No   Yes

Calculate $K_j^* = \arg\max_{K_j=1}^{L_j}\{LE_{K_j}^j\}$

Construct new fuzzy sets using $V_k^{j*}(k=1,\cdots,K_j^*)$ as cores,
e.g. $\widehat{A}_k^{(j)}(x_j) = \exp\left(\frac{-(x_j - V_k^{j*})^2}{2(\sigma_j^*)^2}\right)$

Fig. 4 Flowchart of the FSM algorithm

The FSM algorithm performs the following steps:

Step 1. Set $j=1$, and perform initialization on domain $X_{(j)}$

     1.1) Set $K_j=1$;

     1.2) Calculate the core centers of the fuzzy sets $\beta^{<j>} = \{\beta_m^{(j)} \mid m=1,\cdots,L_j\}$.

1.3) Set the partition matrix $U^{<j>}$ as follows: $U_{1m}^{<j>} = 1$, $m = 1, \cdots, L_j$, which means that every $\beta_m^{(j)}$ belongs to the current partition interval.

1.4) Calculate the compactness indices $MC_j(\beta_m^{(j)})$.

1.5) Calculate the initial "local" entropy:

$$LE_1^j = -\sum_{m=1}^{L_j} \widehat{U}_{1m}^{<j>} \cdot MC_j(\beta_m^{(j)}) \log(\widehat{U}_{1m}^{<j>} \cdot MC_j(\beta_m^{(j)}))$$

Step 2. Increase K_j by 1, and calculate the centers of all the partition intervals, V_k^j $(k = 1, \cdots, K_j)$, and the partition matrix $U^{<j>}$ by the unsupervised hard C-means (HCM) clustering algorithm. By now the optimal partition centers V_k^{j*} $(k = 1, \cdots, K^{j*})$ have been obtained. Without loss of generality, let $V_1^j < V_2^j < \cdots < V_{K_j}^j$.

Step 3. Calculate the "local" entropy $LE_{K_j}^j$ for this partition, and go to *Step 2* until $K_j = L_j$.

Step 4. Calculate $K_j^* = \arg \max_{1 \le K_j \le L_j} \left\{ LE_{K_j}^j \right\}$, that is, the optimal partitioning is to partition the data space $\beta^{<j>}$ into K_j^* regions. By now the optimal partition centers V_k^{j*} $(k = 1, \cdots, K^{j*})$ have been obtained.

Step 5. Construct new fuzzy sets with V_k^{j*} $(k = 1, \cdots, K_j^*)$ as cores. For example, the Gaussian MFs could be generated as follows:

$$\widehat{A}_k^{(j)}(x_j) = \exp\left(\frac{-(x_j - V_k^{j*})^2}{2(\sigma_k^{j*})^2} \right) \tag{12}$$

where the width parameter σ_k^{j*} can be determined as follows

$$\left(\sigma_k^{j*}\right)^2 = \min\left(-\frac{\left(V_k^{j*} - V_{k-1}^{j*}\right)^2}{2 \log \varepsilon_0}, -\frac{\left(V_{k+1}^{j*} - V_k^{j*}\right)^2}{2 \log \varepsilon_0} \right) \tag{13}$$

and $V_0^{j*} \triangleq \min\left(V_1^{j*}, \min_i x_j^{(i)}\right)$, $V_{K_j^*+1}^{j*} \triangleq \max\left(V_{K_j^*}^{j*}, \max_i x_j^{(i)}\right)$, ε_0 is a small positive number.

Step 6. Increase j by 1, and go to *Step 1.1* until $j=n$.

By using the above FSM algorithm, the original fuzzy sets could be merged optimally by maximizing the "local" entropy to achieve distinguishable fuzzy sets while preserving the original information as much as possible.

IV. CONSTRUCTION OF PARSIMONIOUS TS FUZZY MODELS BASED ON THE FSM ALGORITHM

Classical data-driven rule induction algorithms such as neuro-fuzzy algorithms [24][25] generate fuzzy sets with "too much" overlap due to their accuracy-oriented property. In this section, given a data set $\left\{(x^{(i)}, d^{(i)})\right\}_{i=1}^N$ (N is the number of training samples), an initial TS model [26] with grid partitioning of input space, named as TS_0, is obtained as follows:

$$R_l : \text{if } x_1 \text{ is } A_{l_1}^{(1)} \text{ and } \ldots \text{ and } x_n \text{ is } A_{l_n}^{(n)} \text{ then } y_l = a_{l0} + a_{l1}x_1 + \cdots + a_{ln}x_n \tag{14}$$

where $l = (l_1, \cdots, l_n)$, $1 \le l_1 \le L_1$, ..., $1 \le l_n \le L_n$, and L_j is the number of fuzzy sets on domain $X_{(j)}$. The overall system output is

$$y = \sum_{l_1=1}^{L_1} \cdots \sum_{l_n=1}^{L_n} w_l y_l \tag{15}$$

where $w_l = \dfrac{\tau_l}{\sum_{l_1=1}^{L_1} \cdots \sum_{l_n=1}^{L_n} \tau_l}$, and τ_l is the membership degree of $x = (x_1, \cdots, x_n)$ belonging to the fuzzy set A_l, *i.e.* ,

$$\tau_l = A_l(x) = \mathop{\text{T}}_{j=1}^n \left(A_{l_j}^{(j)}(x_j)\right) \tag{16}$$

11

where T is a triangular norm (*t-norm* for short) operator. A commonly used *t-norm* operator in the TS fuzzy model is the product, *i.e.*, $A_l(x) = \prod_{j=1}^{n} A_{l_j}^{(j)}(x_j)$.

Aiming to improve the interpretability of the TS model while preserving its global approximation performance defined by

$$E = \frac{1}{2}\sum_{i=1}^{N}\left\| d^{(i)} - y^{(i)} \right\|^2 \tag{17}$$

where $y^{(i)}$ is the model output, this paper proposes a scheme as shown in Fig. 5 to merge the fuzzy sets $\{A_{l_j}^{(j)}\}$ on every domain $X_{(j)}$ based on the FSM algorithm if a certain distinguishability of input space partitioning in TS_0 is considered to be lost in terms of the proposed entropy criterion.
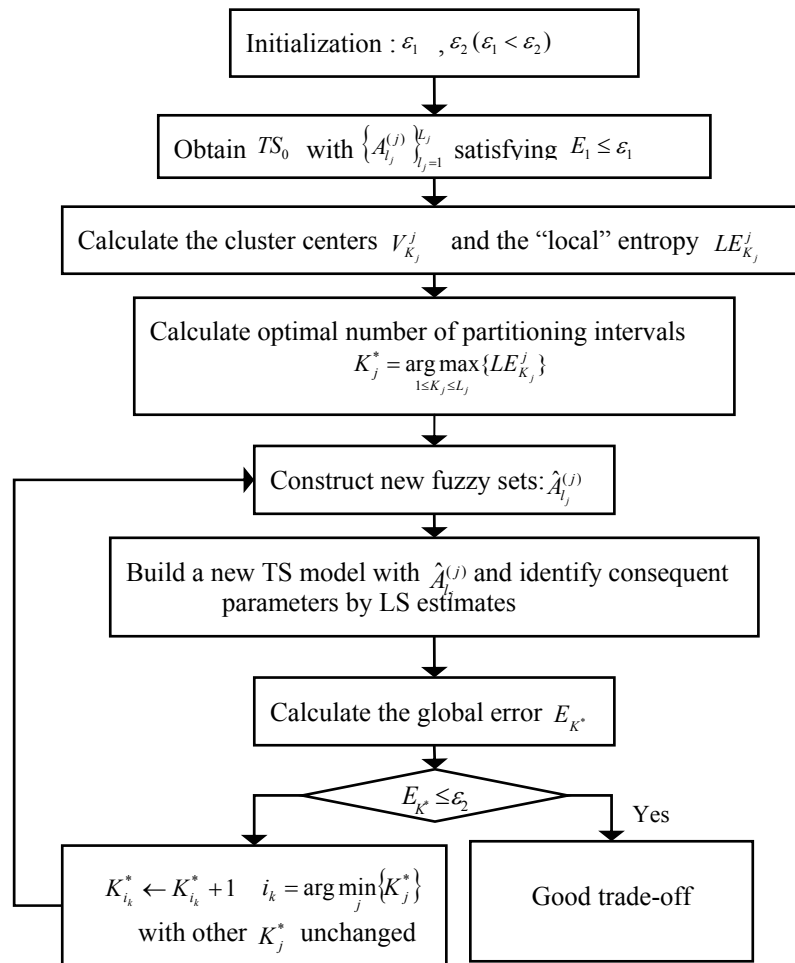
Fig. 5 Flowchart for seeking good trade-off between accuracy and distinguishability in the TS model

The steps in the proposed scheme are as follows:

Step 1. Initialization.

Set up an accuracy threshold ε_1 for the initial TS model and a tolerant accuracy threshold ε_2 for the TS model with merged fuzzy sets ($\varepsilon_1 < \varepsilon_2$).

Step 2. Obtain an initial TS model satisfying ε_1 with a set of initial fuzzy sets $\left\{ A_{l_j}^{(j)} \right\}_{l_j=1}^{L_j}$ on domain $X_{(j)}$ ($j = 1, \cdots, n$). Calculate the core centers of fuzzy sets $\left\{ A_{l_j}^{(j)} \right\}_{l_j=1}^{L_j}$.

Step 3. Calculate the cluster centers and the "local" entropies of L_j possible partitions of the data space $\beta^{<j>}$. At the K_j-th partitioning, there are K_j intervals. The K_j cluster centers are represented as $V_{K_j}^{(j,1)}, \cdots, V_{K_j}^{(j,K_j)}$ ($K_j = 1, \cdots, L_j$). The "local" entropy for the K_j-th partitioning is denoted as $LE_{K_j}^j$.

Step 4. Determine the trade-off between global accuracy and distinguishability of fuzzy sets.

4.1) Calculate the optimal number of partitioning intervals in terms of maximum "local" entropy:

$$K_j^* = \underset{1 \leq K_j \leq L_j}{\arg \max} \{ LE_{K_j}^j \} \tag{18}$$

4.2) Use *step 6* of the FSM algorithm to construct K_j^* new fuzzy sets $\hat{A}_{l_j}^{(j)}$ ($l_j = 1, \cdots, K_j^*$) whose cores are $V_{K_j^*}^{(j,1)}, \cdots, V_{K_j^*}^{(j,K_j^*)}$, ($j = 1, \cdots, n$).

4.3) Use the new merged fuzzy sets $\hat{A}_{l_j}^{(j)}$ to build a TS model, named TS_{K^*}, in which $\hat{A}_{l_j}^{(j)}$ is the premise part of rule $\hat{R}_l$ and the consequent parameters are obtained by least-square (LS) estimates, where $K^* = (K_1^*, \cdots, K_n^*)$.

4.4) Calculate the global error E_{K^*} for model TS_{K^*}.

4.5) If $E_{K^*} > \varepsilon_2$, set $K^*_{i_k} : K^*_{i_k} + 1$, where $i_k = \arg\min_j \{K^*_j\}$, keep other K^*_j unchanged, and go to

4.2) until $E_{K^*} \le \varepsilon_2$,

The above scheme aims to find a best trade-off between required model accuracy and fuzzy set distiguishability. It is possible that the fuzzy sets with the optimal distinguishability in terms of the maximum "local" entropy cannot meet the required model accuracy. When this happens, further partitioning is performed on the domain where there is the smallest number of fuzzy sets, as indicated in step 4.5, to make the least sacrifice of the fuzzy set distinguishability. Actually, due to $\varepsilon_1 < \varepsilon_2$ there always exists a solution in this scheme. At least TS_0, which is actually TS_L, will satisfy $E_{K^*} \le \varepsilon_2$, where $L = \prod_j L_j$.

It should be noted that in the TS model, in addition to the distinguishability of input space partitioning, there is a further challenging problem, *i.e.* , the TS local models often exhibit eccentric behaviors that are hard to be interpreted, which imposes limits to the model application [27]. A desirable interpretability about the TS local models is that these local models should match the global model well in the corresponding local regions, that is to say, a good interpretable local model can dominate the system behavior in the corresponding local region. A requirement directly related to this interpretability is that MFs should have large core regions and be less overlapped among adjacent fuzzy sets [28]. Actually, this requirement for MFs design is in line with the heuristic criterion suggested in [23] and [29] for fuzzy cluster interpretation. This heuristic criterion states that "good" clusters are actually not very fuzzy [23]. Although fuzzy algorithms are used in data clustering, the aim of the clustering is to generate a "harder" partitioning of the data set [29], by which a better interpretation of input space partitioning can be achieved. In order to obtain MFs with large bounded core and less overlapping, Hoppner and Klawonn [30] proposed a novel clustering algorithm by using the distance to the Voronoi cell of a cluster rather than to the cluster prototype, and furthermore they

assigned a "reward" to membership degrees that are near to 0 and 1. However, to improve the TS local model interpretability will generally degrade the global model accuracy performance, so a trade-off between global model accuracy and local model interpretability should be obtained. One of the benefits from the above merging process, which results in a more compact rule base in the TS model, is that it makes the local models tend to dominate the behaviors of the TS model in most local areas, which would provide good local model interpretability, at the same time, the global system performance is taken into account.

V. EXPERIMENTAL RESULTS

In this section, two examples are given to evaluate the proposed method for improving the interpretability of input space partitioning while preserving the global accuracy.

The first example is a benchmark problem of modeling a time series of the intensity of a variable white dwarf star. This dataset is available on http://www-psych.stanford.edu/~andreas/Time-Series/. The data samples, which are noisy, discontinuous and nonlinear in nature, represent the optical oscillations of a physical system. In our experiment 350 samples were used, each with two inputs and one output, *i.e.*, $v_t = f(v_{t-1}, v_{t-2})$, where v_t is the intensity of the star at time t. These samples were divided into a training set with 300 samples and a test set with the remaining 50 samples. The thresholds ε_1 and ε_2 were set as 0.015 and 0.0315 respectively. An initial TS model, denoted as TS_0, was produced by the ANFIS method [25]. The global prediction results by TS_0 on the training data and test data are depicted in Fig. 6, in which the approximation errors are 0.0143 on the training data and 0.0346 on the test data. Although the global approximation accuracy of TS_0 is very good, the distinguishability of its MFs is poor, as shown in Fig. 7.
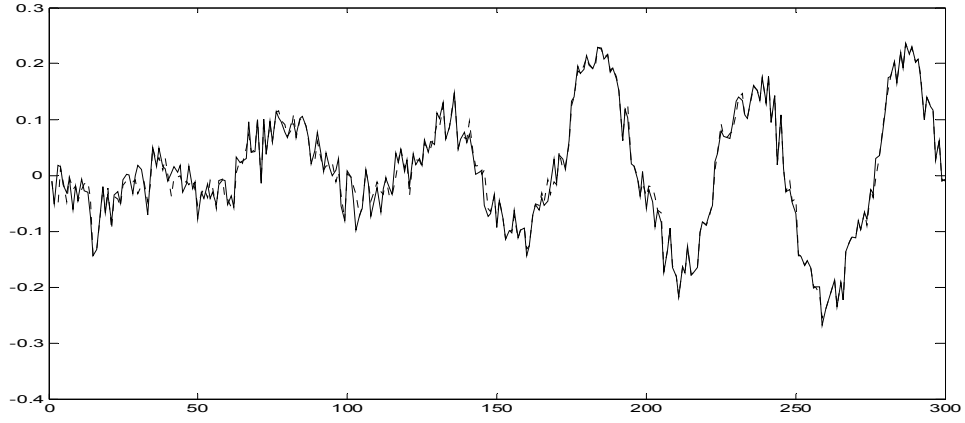
(a)
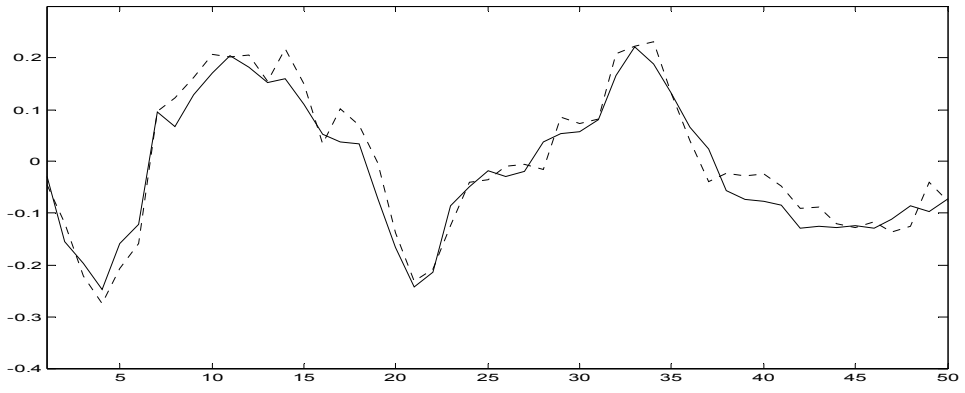
(b)

Fig. 6 Prediction results by TS_0 on training samples (a) and test samples (b): solid line (SL) represents desired
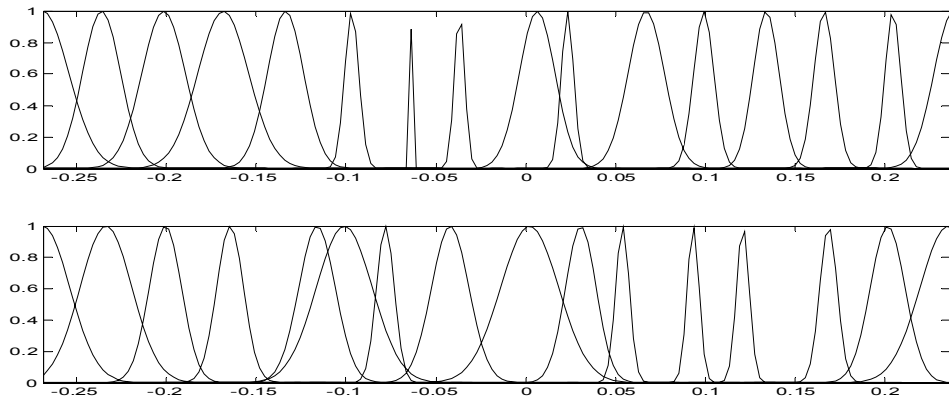
output and dotted line (DOL) represents model output.

Fig. 7 MFs for variable v_{t-1} (top) and v_{t-2} (bottom) in TS_0.

The proposed method was then adopted to improve the interpretability of input space partitioning while trying to preserve the approximation ability of the TS model. The "local" entropies of the possible partitions for each input variable were obtained using the compactness index (5) by taking into account the information of input-output samples. As shown in Fig. 8, the partition with 6 clusters for v_{t-1} and 5 clusters for v_{t-2} is optimal in terms of maximum "local" entropy. Based on this observation, the results of fuzzy set merging are depicted in Fig. 9. However, the corresponding TS model, denoted as $TS_{(6,5)}$, produces an approximation error of 0.0317 on the training data, which is higher than the preset error threshold ε_2. By using the step 4 in the proposed scheme in section IV, it has been found that the TS model $TS_{(6,6)}$ built by using 6 merged fuzzy sets for v_{t-1} and 6 merged fuzzy sets for v_{t-2}, as shown in Fig. 10, can satisfy the tolerant approximation error threshold. Fig. 11 shows the prediction results by $TS_{(6,6)}$ with prediction errors of 0.0314 on the training data and 0.0496 on the test data respectively.
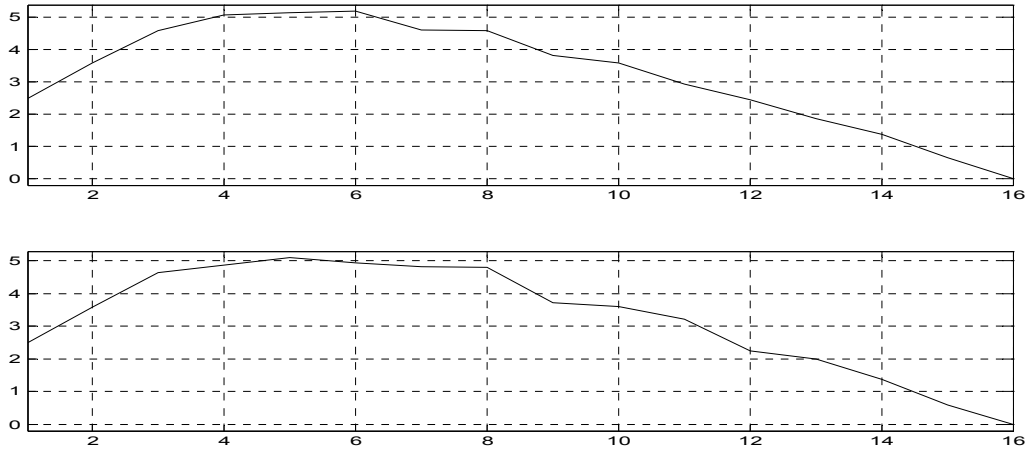
Fig. 8 The "local" entropies of the possible partitions for variables v_{t-1} (top) and v_{t-2} (bottom)
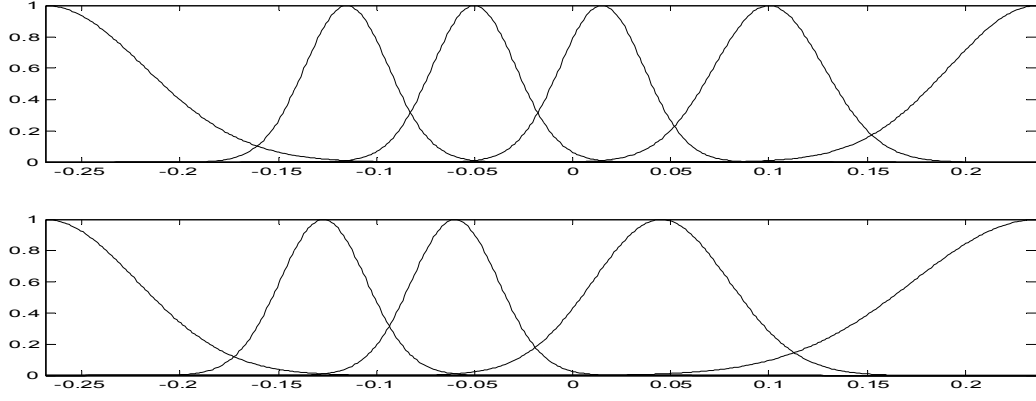
Fig. 9 MFs for v_{t-1} (left) and v_{t-2} (right) with the optimal partition in terms of maximum "local" entropy
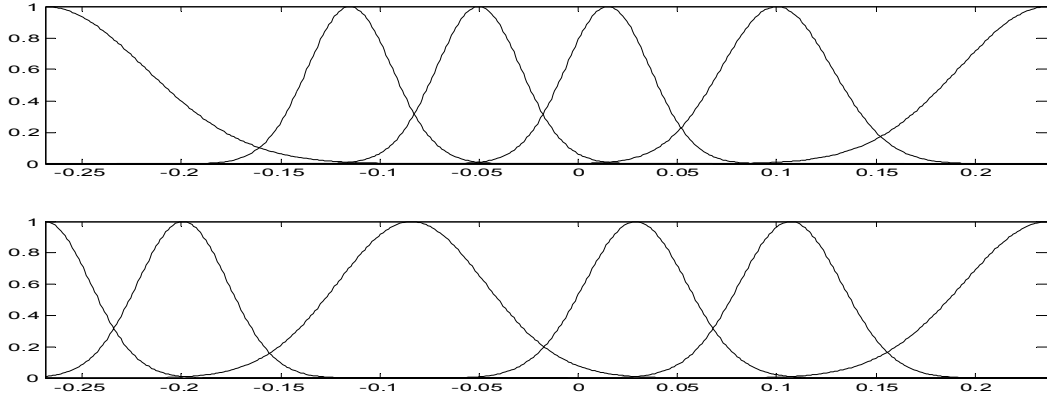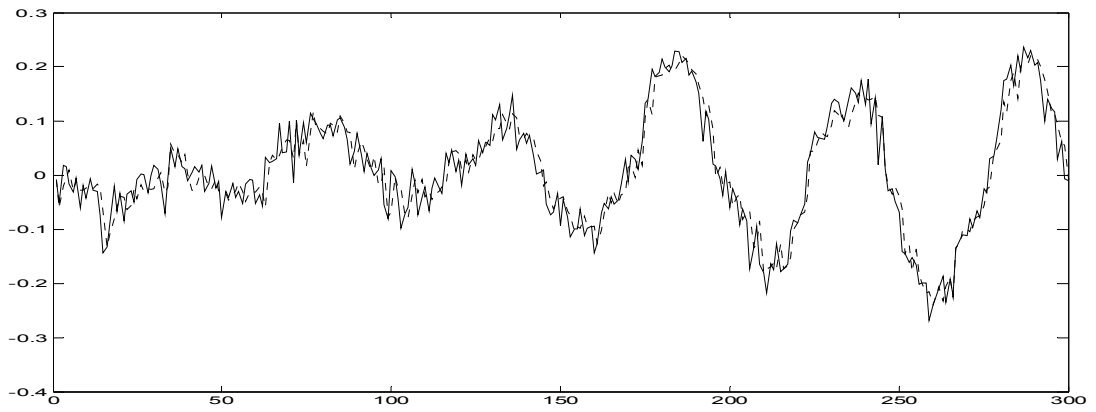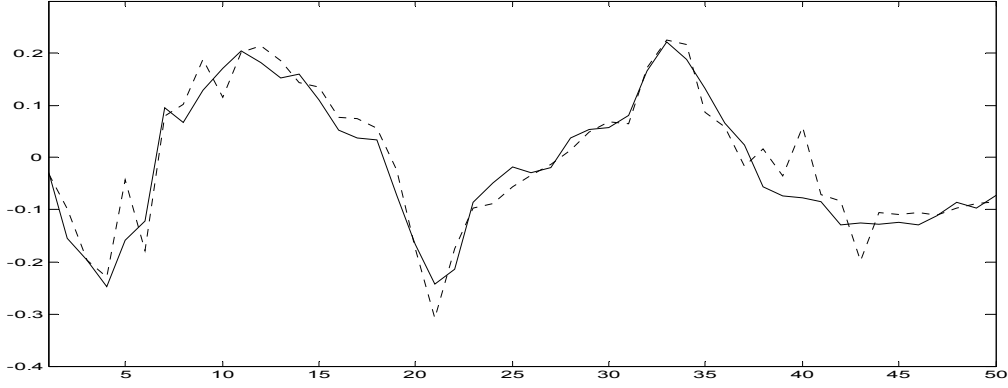

Fig. 10 MFs v_{t-1} (left) and v_{t-2} (right) used in $TS_{(6,6)}$

(a)

(b)

Fig. 11 Prediction results by $TS_{(6,6)}$ on training samples (a) and test samples (b): SL represents desired output

and DOL represents model output.

In comparison with the proposed scheme, the following commonly used similarity measure [20][21] was also adopted to merge similar fuzzy sets to improve the distinguishability of input space partitioning for TS_0:

$$sim(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$ (19)

where A and B are two fuzzy sets, $|\cdot|$ and $\cap$ represent the cardinality of a set and the intersection respectively. In the similarity based merging procedure with Gaussian MFs, the core center and width parameter of a merged fuzzy set are usually calculated as the mean values of the core centers and width parameters of the similar fuzzy sets separately. By setting the similarity threshold as 0.2 (this was chosen for a fair comparison, *i.e.*, comparable fuzzy set distinguishability and global model accuracy), the similar fuzzy sets for each input variable were merged as shown in Fig. 12, and a TS model, denoted as TS_{sim}, was built up based on these merged fuzzy sets. Fig. 13 illustrates the prediction results by TS_{sim}, which produces approximation errors of 0.021 on the training data and 0.082 on the test data respectively. It can be seen that although the distinguishability of input space

19

partitioning for TS_0 has been improved to some extent by the similarity based merging method, the proposed scheme has achieved both better distinguishability of input space partitioning and better generalization performance than the similarity based merging method on this benchmark problem.
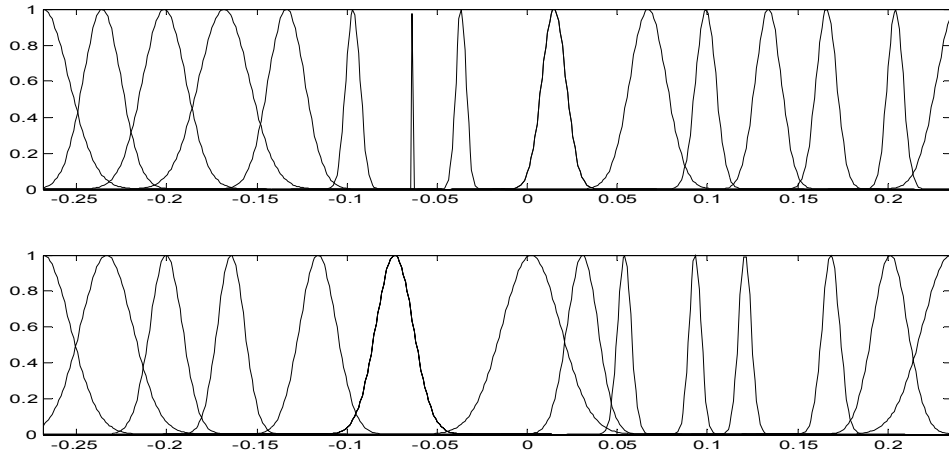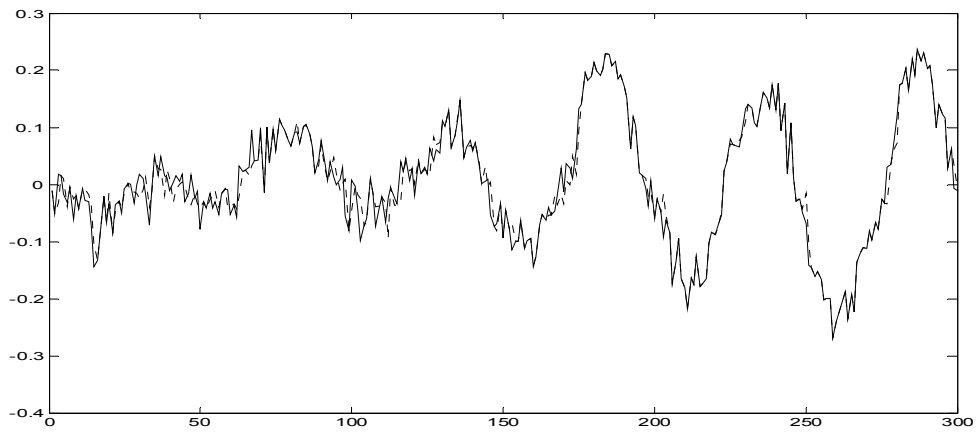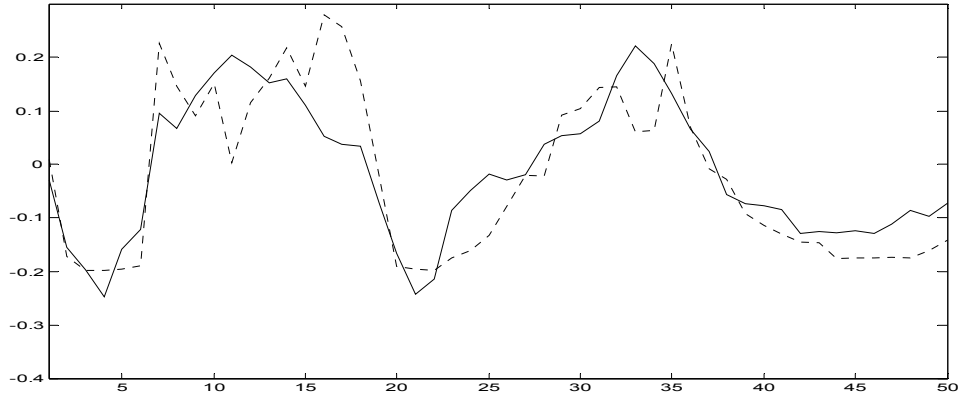

Fig. 12 MFs merged in terms of the similarity.

(a)

(b)

Fig. 13 Prediction results by TS_{sim} on training samples (a) and test samples (b): SL represents desired output

and DOL represents model output.

Furthermore, an input space partitioning with uniformly distributed fuzzy sets as shown in Fig. 14 was used to build a TS model TS_{uni}, which produced an approximation error of 0.021 on the training data. However, as shown in Fig. 15, TS_{uni} has poor generalization ability on the test data, although it was also optimised with a least square procedure to obtain the consequent parameters.
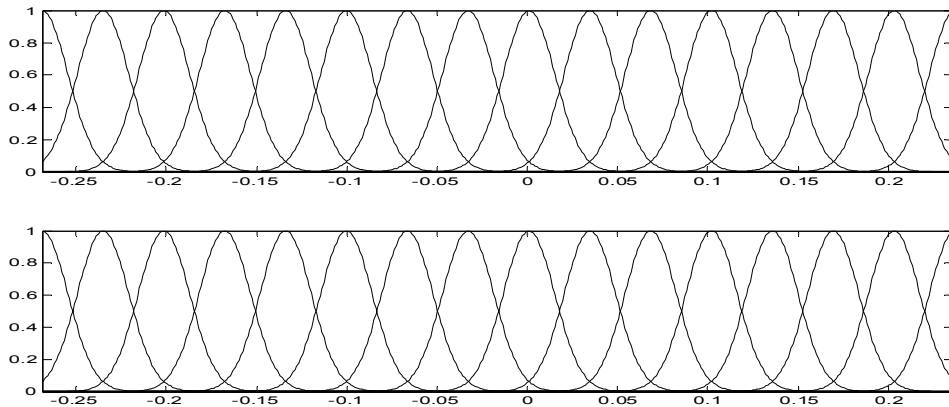
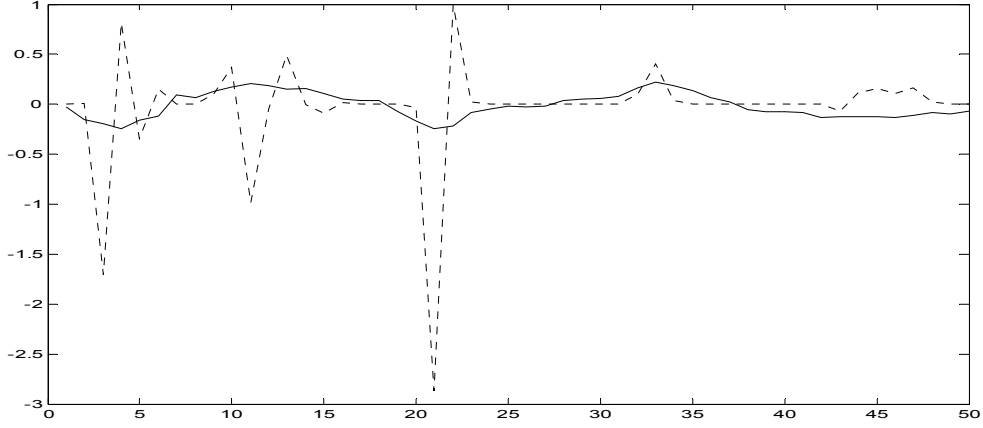Fig.14 An input space partitioning with uniformly distributed fuzzy sets

Fig. 15 Prediction results of TS_{uni} on test samples: SL represents desired output and DOL represents model output.

The second experiment is to recover an original signal from data highly contaminated by noise. In this experiment, the noisy signal is generated by

$$v = \widetilde{v} + \upsilon \tag{19}$$

$$\widetilde{v} = 3\sin(\pi x)/(1 + x^2) \tag{20}$$

where $\widetilde{v}$ is the original signal, υ is a Gaussian noise with zero mean and variance 0.1. 350 samples $\left\{(x^{(i)}, d^{(i)})\right\}_{i=1}^{N}$ (N=350) were generated with random $x^{(i)} \in [-4.5, -1]$ and $d^{(i)}$ obtained by (19). The uniformly distributed fuzzy sets as shown in Fig. 16 were firstly used to build a TS model TS_0 with ε_1=0.3. The signal recovering results by TS_0 and its local models are depicted in Fig. 17, from which we can see that the local model interpretability is very poor.

The proposed fuzzy set merging scheme was then used to improve the distinguishability of partitioning of input space and thus the interpretability of local models. The error threshold ε_2 was set as 0.31. The "local" entropies of the possible partitions, as shown in Fig. 18, were obtained in terms of the compactness index (5) by taking into account the information of input-output samples. The partition with 5 clusters is optimal in terms of maximum "local" entropy. The merged fuzzy sets are

illustrated in Fig. 19. However, the TS model $TS_{(5)}$, constructed using these 5 merged fuzzy sets, cannot satisfy the tolerant approximation error threshold. By inserting an extra fuzzy set using the step 4 in the proposed scheme, a TS model with 6 merged fuzzy sets, $TS_{(6)}$, achieves an approximation error of 0.3094. These 6 merged fuzzy sets are shown in Fig. 20, and the signal recovering result by $TS_{(6)}$ and its local models are illustrated in Fig. 21. It can be seen from Fig. 20 and Fig. 21 that not only the distinguishability of input space partitioning in $TS_{(6)}$ is much better than in TS_0, but also its local models exhibit much better interpretability than in TS_0. It is interesting to notice that there is overfitting problem in TS_0 due to the high level of noise in the training data and too many local models (thus too many free parameters). By merging the fuzzy sets in TS_0, the overfitting problem is combated to some extent and the obtained model $TS_{(6)}$ behaves smoothly, which is a desired property in noise cancellation.
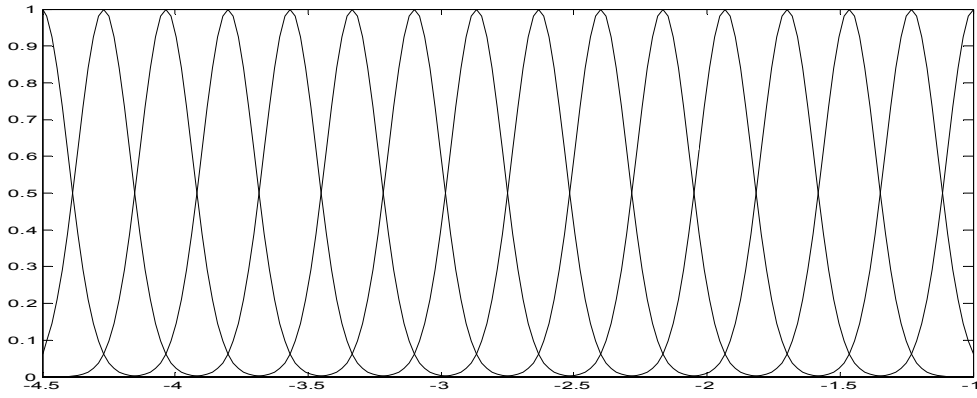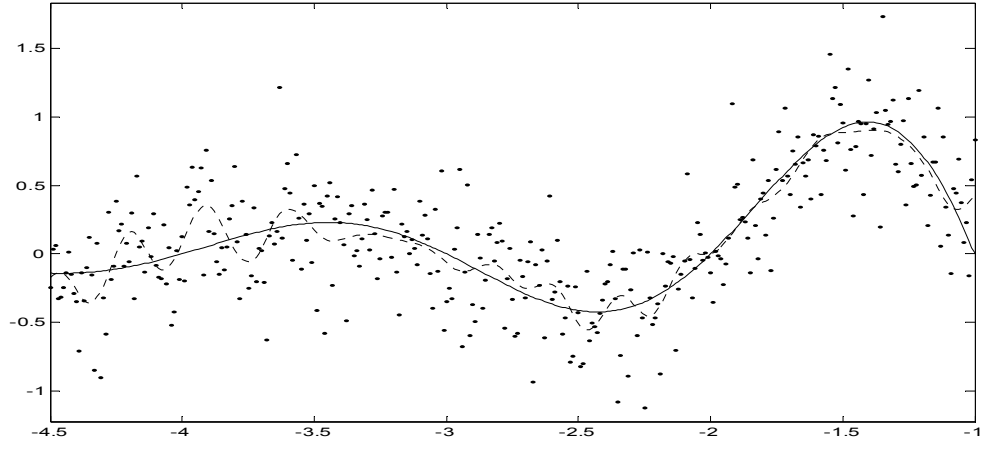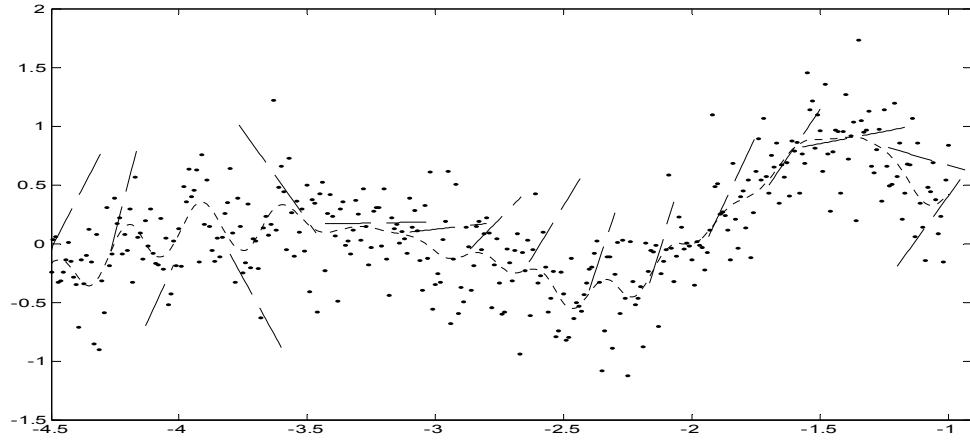

Fig.16 Uniformly distributed fuzzy sets for TS_0

(a)

(b)

Fig. 17 Signal recovering by TS_0 (a) and its local models (b): in (a) SL represents the original signal $\widetilde{v}$ and

DOL represents the recovered signal; in (b) DOL represents the recovered signal and dashed lines (DSLs)
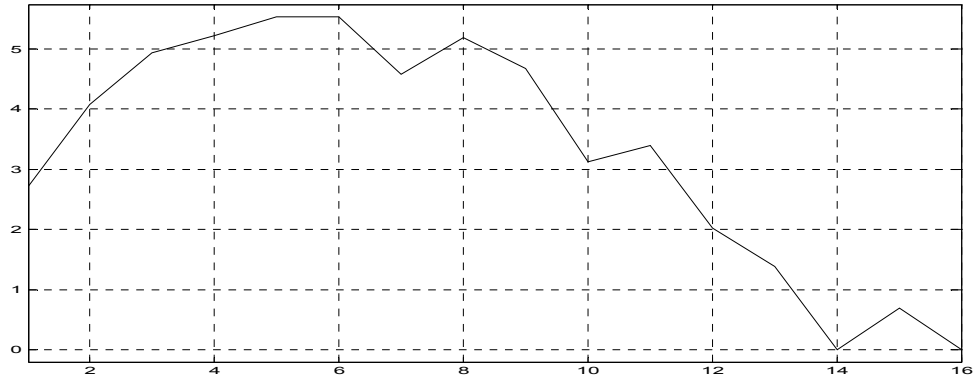
represent the outputs of local models

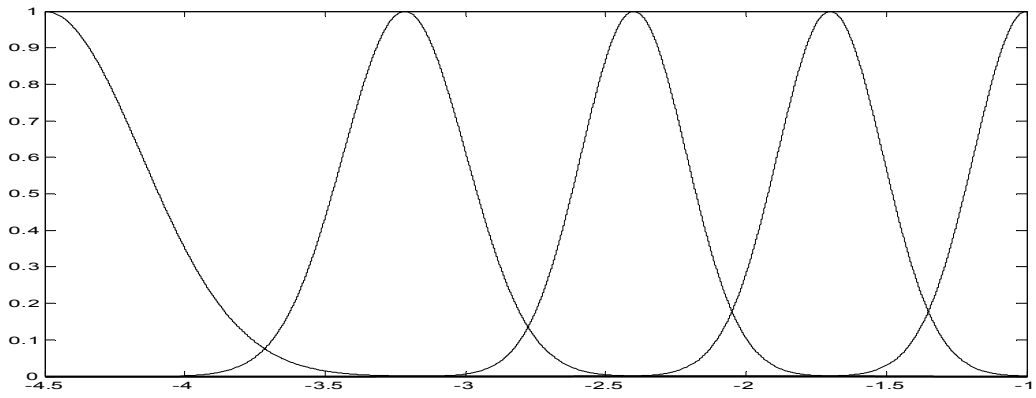Fig.18 The "local" entropies of the partitions of input space

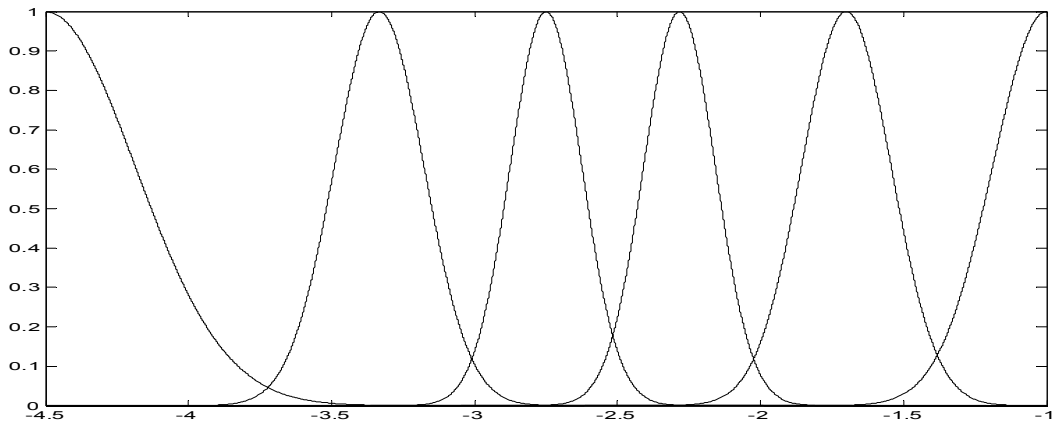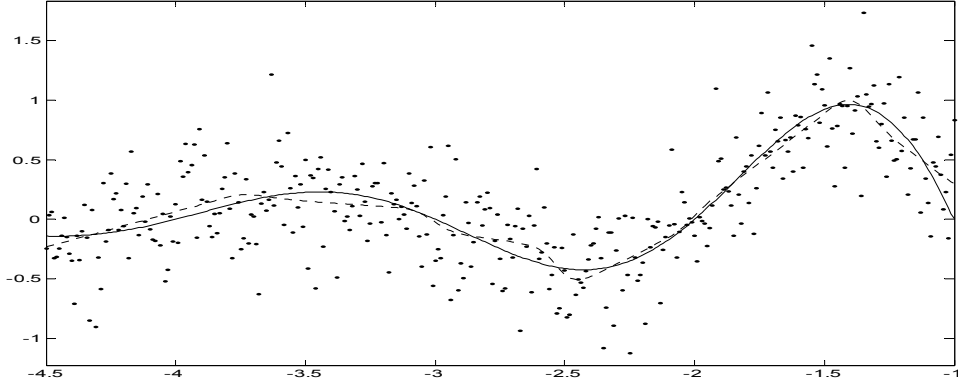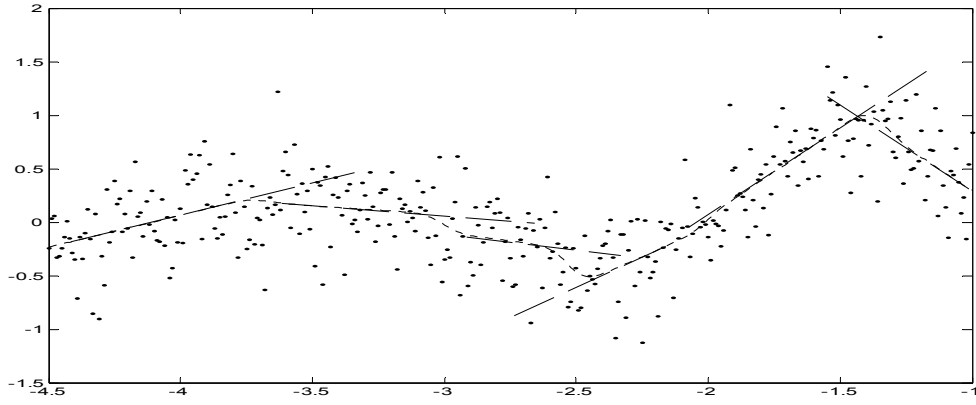Fig. 19 MFs with optimal partition in terms of maximum "local" entropy

Fig. 20 MFs for $TS_{(6)}$

(a)

(b)

Fig. 21 Signal recovering by $TS_{(6)}$ (a) and its local models (b): in (a) SL represents the original signal $\widetilde{v}$ and DL represents the recovered signal; in (b) DOL represents the recovered signal and DSLs represent the outputs of local models

It is also interesting to notice that in the second experiment after merging the initial fuzzy sets based on the proposed scheme only one local model tends to dominate the behavior of the TS model in most local areas, which leads to good local model interpretability. However, one point which may be raised regarding the proposed scheme is with regard to the choice of parameters ε_1 and ε_2. Clearly this will be data specific. ε_1 controls the global model accuracy, which has been adopted in most accuracy-oriented data-driven fuzzy modeling methods. ε_2 controls the trade-off between global model

accuracy and the distinguishability of input space partitioning (thus local model interpretability). With specific data, the trial-and-error procedure is appropriate in determining how much the global model accuracy can be degraded in order to get an interpretable fuzzy model with distinguishable fuzzy sets.

VI. CONCLUSION

In this paper, an entropy measure is introduced to characterize the distinguishability of fuzzy sets, based on which an effective FSM algorithm is proposed to choose the optimal number of merged fuzzy sets with good distinguishability by maximizing the entropy. Different from the existing merging algorithm, the proposed FSM algorithm takes into account the information provided by input-output sample pairs to improve the distinguishability of input space partitioning. Furthermore, a scheme to seek a good trade-off between global model accuracy and distinguishability of input space partitioning in constructing TS fuzzy models has been proposed. Experimental results have shown the effectiveness of the proposed method in constructing accurate and parsimonious TS fuzzy models. Further improvement can be made by combining the proposed entropy measure and global performance measure in a learning process to optimally update the model parameters.

ACKNOWLEDGMENT

REFERENCES

[1]   C.J. Harris, X. Hong, and Q. Gan, *Adaptive Modeling, Estimation and Fusion from Data: A Neurofuzzy Approach*. Springer, 2002.

[2] S. Guillaume, "Designing fuzzy inference systems from data: an interpretability-oriented review," *IEEE Trans. on Fuzzy Systems*, Vol. 9, No. 3, pp.426-443, 2001.

[3] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena (ed), *Interpretability Issues in Fuzzy Modeling,* number 128 in Studies in Fuzziness and Soft Computing, Springer–Verlag, Heidelberg, 2003

[4] T. Suzuki and T. Furuhashi, "Conciseness of fuzzy models," in J. Casillas, O. Cordón, F. Herrera, L. Magdalena, editors, *Interpretability Issues in Fuzzy Modeling,* number 128 in Studies in Fuzziness and Soft Computing, Springer–Verlag, Heidelberg, 2003.

[5] T. Furuhashi and T. Suzuki, "On interpretability of fuzzy models based on conciseness measure," *Proc. of the 10th IEEE Int. Conf. on Fuzzy Sets* (FUZZ-IEEE'01), Melbourne Australia, December 2001.

[6] J.V. de Oliveira, "Semantic constraints for membership function optimization," *IEEE Trans on Systems, Man and Cybernetics-Part A*, Vol.29, No.1, pp.128–138, 1999.

[7] Y. Jin, B. Sendhoff, "Extracting interpretable fuzzy rules from RBF networks", *Neural Processing Letters*, Vol.17, No.2, pp.149-164, 2003.

[8] M. Setnes, R. Babuska, U. Kaymak, and H.R. van Nauta Lemke, "Similarity measures in fuzzy rule base simplification," *IEEE Trans. on Systems, Man and Cybernetics-Part B*, Vol.28, No.3, pp. 376-386, 1998.

[9] G. Castellano, A.M. Fanelli, and C. Mencar, "Similarity vs. possibility in measuring fuzzy sets distinguishability," *Proc. of 5th International Conference on Recent Advances in Soft Computing (RASC)*, Nottingham, United Kingdom, December 16-18, 2004.

[10] M. Setnes, R. Babuska, and H.B. Verbruggen, "Rule-based modeling: precision and transparency," *IEEE Trans. on Systems, Man and Cybernetics-Part C*, Vol.28, No.1, pp.165-169, 1998.

[11] J. Espinosa and J. Vandewalle, "Constructing fuzzy models with linguistic integrity from numerical data-AFRELI algorithm," *IEEE Trans. on Fuzzy Systems*, Vol.8, No.5, pp. 591-600, 2000.

[12] G. Castellano, A.M. Fanelli, and C. Mencar, "Generation of interpretable fuzzy granules by a double-clustering technique," *Archives of Control Science, Special Issue on Granular Computing*, Vol.12, No.4, pp. 397-410, 2002.

[13] H. Wang, S. Kwong, Y. Jin, W. Wei and K. Man, "Agent-based evolutionary approach to interpretable rule-based knowledge extraction," *IEEE Trans. on Systems, Man, and Cybernetics-Part C*, In press, 2005

[14] J.A. Roubos, M. Setnes, and J. Abonyi, "Learning fuzzy classification rules from labeled data," *Int. J. of Information Sciences*, Vol.150, No.1-2, pp.77-93, 2003.

[15] X.Z. Wang, B. Chen, G. Qian, and F. Ye, "On the optimization of fuzzy decision trees," Fuzzy *Sets and Systems*, Vol. 112, No.1, 117–125, 2000.

[16] T. Kavli, "ASMOD-An algorithm for adaptive spline modeling of observation data," *Int. J. Control*, Vol.58, No.4, pp.947-967, 1993.

[17] Q. Gan and C.J. Harris, "Fuzzy local linearization and local basis function expansion in nonlinear system modeling," *IEEE Trans. on Systems, Man and Cybernetics-Part B*, Vol.29, No.4, pp.559-565, 1999.

[18] Q. Gan and C.J. Harris, "A hybrid learning scheme combining EM and MASMOD algorithms for fuzzy local linearization modeling," *IEEE Trans. on Neural Networks*, Vol.12, No.1, pp.43-53, 2001.

[19] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussions)," *J. Roy. Statist. Soc. B*, vol. 39, pp. 1–39, 1977.

[20] R. Zwick, E. Carlstein, and D.V. Budescu, "Measures of similarity among fuzzy concepts: a comparative analysis", *Int. J. of Approximation Reasoning*, Vol.1, No.2, pp.221-242, 1987.

[21] D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Applications*. New York: Academic, 1980.

[22] J.N. Kapur, P.K. Sahoo and A.K.C. Wong, "A new method for grey-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics and Image Processing*, Vol.29, No.3, pp.273-285, 1985.

[23] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[24] L-X. Wang, *Adaptive Fuzzy Systems and Control*, Englewood Cliffs, NJ: Prentice-Hall, 1994.

[25] J.-S.R. Jang, "ANFIS: Adaptive-network-based fuzzy inference system," *IEEE Trans. on Systems, Man and Cybernetics,* Vol.23, No.3, pp.665-685, 1993.

[26] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. on Systems, Man and Cybernetics*, Vol.15, No.1, pp.116-132, 1985.

[27] J. Yen, L. Wang, and C.W. Gillespie, "Improving interpretability of TSK fuzzy models by combining global learning and local learning," *IEEE Trans. on Fuzzy Systems*, Vol. 6, No. 4, pp. 530-537, 1998.

[28] S.-M. Zhou and J.Q. Gan, "Improving the interpretability of Takagi-Sugeno fuzzy model by using linguistic modifiers and a multiple objective learning scheme," *Proc. of International Joint Conference on Neural Networks (IJCNN'04)*, Budapest, Hungary, 2004, pp.2385-2390.

[29] I. Gath and A.B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.11, No.7, pp.773-781, 1989.

[30] F. Hoppner, and F. Klawonn, "A new approach to fuzzy partitioning," *Proc. of the Joint 9th IFSA Congress and 20th NAFIPS Int. Conf.*, Vancouver, Canada, 2001, pp1419-1424.