Abstract—This paper presents a novel, Information Theoretic Learning (ITL) method to model a fuzzy system for regression tasks that minimizes the Renyi's entropy of the error signal. An architecture based on a generalization of the well-known Adaptive-Network-Based Fuzzy Inference System (ANFIS) was used to perform such a modeling. The resulting method was tested on the prediction of future values for the Mackey-Glass chaotic time series. The results show that, when using the ITL cost function, the method returns better models in comparison with a Mean Squared Error (MSE)-guided cost function.

I. INTRODUCTION

Learning from examples is one of the most widely used techniques to model a process from its input-output data. The goal is to approximate the implicit global functional relation that maps each input pattern with its corresponding desired response. From the engineering perspective, this problem normally can be split in two steps: system’s structure identification and parameter optimization. The former is related with finding the appropriate structure for the system, while the latter comprises the task of adjusting its parameters on the context of the constraints imposed by the given structure. It is still a problem the modeling of such a system when the embedded relation is complex and nonlinear. Even more, in real problems data is normally affected by noise, especially in imprecise domains. In order to deal with imprecision and uncertainty, fuzzy systems have represented one the major successful solutions in these circumstances over the past years. Since the paper of Zadeh [1] on fuzzy sets, a lot of work has been successfully done on this field, and a number of architectures acting as mappers have been published. Fuzzy systems are based on fuzzy logic [2] and the concept of degree of membership, in which the truthfulness of an statement is no more bivalent but a real value $\mu \in [0,1]$. Fuzzy inference systems use fuzzy IF-THEN rules that lead their operation.

Let $X_1$, $X_2$ to be two input variables and $Z$ an output variable, where $A,B$ and $C$ are linguistic terms [3] characterized by appropriate membership functions [1], then a typical fuzzy rule is like:

\[
\text{if } X_1 \text{ is } A \text{ and } X_2 \text{ is } B \text{ then } Z \text{ is } C
\]

Such a rule can be implemented in several forms, attending to the concrete implementation of the logic connectives and the shape and parameters of the corresponding membership functions. Also the macrostructure of the general system is dictated by the way in which the rules are combined. Thus, fuzzy modeling represents a particular case of system modeling where one has to find out the correct structure and parameterization of the fuzzy relations.

From a machine learning perspective, we want a fuzzy system to have learning ability of updating and fine-tuning itself based on all the statistical structure provided by the input output pairs.

Adaptive Networks [4] are a superset of all kinds of feedforward neural networks in which each node performs a particular function (node function) based on the incoming signals and a set of parameters pertaining to this node. The type of node function may vary from node to node; and the choice of node function depends on the overall function that the network is designed to carry out. Jang and Sun [4-5] demonstrated the use of this type of networks to implement adaptive neuro-fuzzy systems.

As commonly happen in supervised training over neural-based systems, learning to find the optimal parameters for the mapping of the input-output data, is usually based on the minimization/maximization of a cost function that governs the system adaptation. The Mean Squared Error (MSE) has been the most popular criterion in the training to optimize the similarity between the output of the system and the desired response.

However, this measure relies heavily on the assumption of Gaussianity of the error signal, which is completely characterized by its first and second moment. If the error signal is not Gaussian distributed there is information that is not used to adapt the weights when the cost is minimized. Although such assumptions provide successful engineering solutions to most practical problems, it has become evident that when dealing with nonlinear systems, this approach needs to be refined [6]. A criterion that not only considers second-order statistics, but also takes into account higher order statistical behavior is much desired.

Information Theoretic Learning (ITL) has been recently developed extending the concept of mean squared error adaptation to include information criteria. Entropy, which
was introduced by Shannon [7], is a scalar quantity that provides a measure for the average information contained in a given probability function. As by definition, information is a function of the Probability Density Function (PDF), entropy as an optimally criterion, extends MSE since when entropy is minimized, all the moments of the error PDF (not only the second moments) are constrained.

In this work, the Minimization of Error Entropy (MEE) is proposed as a more robust cost function applied for parameter optimization on the modeling of fuzzy systems, based on adaptive networks regression tasks. Effectively the fuzzy system becomes information theoretic, manipulating the error entropy. The objective is to demonstrate how this new measure can be used in order to guide the self-optimization of this kind of neuro-fuzzy systems and its validity for the tuning up of fuzzy membership functions. This work is a first step in a general framework of fuzzy modeling in terms of information theory. Other approaches concerning entropy manipulation applied to fuzzy sets can be found in [8].

II. MINIMUM ERROR ENTROPY CRITERION

Let $X$ be a random variable with probably distribution function (PDF) $f_r$, then the Renyi’s quadratic entropy [9] is defined as

$$H^2_r(X) = -\log \int f^2_r(x)dx.$$  

(1)

Given a set of data points $\{x_i\}_{i=1}^N$ drawn from $X$, the Parzen window estimate of the PDF [10] is

$$\hat{f}_{x,\sigma}(x) = \frac{1}{N} \sum_{i=1}^N \kappa_\sigma(x - x_i)$$

(2)

where $\kappa_\sigma(x - x_i)$ is the Gaussian kernel

$$\kappa_\sigma(x - x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - x_i)^2}{2\sigma^2}\right)$$

(3)

and $N$ is the number of data points and $\sigma$ the kernel size.

The kernel size or bandwidth is a free parameter that must be chosen by the user using concepts of density estimation, such as Silverman’s rule [11] or maximum likelihood. It has been experimentally verified that the kernel size affects much less the performance of ITL algorithms than density estimation [12], but a thorough treatment of this issue is beyond the scope of this paper.

Substituting (3) into (2) and after some mathematical manipulations [13], an estimator for (1) is obtained:

$$\hat{H}^2_r(X) = -\log IP(X).$$

(5)

$$IP(X) = \frac{1}{N^2} \sum_{j=1}^N \sum_{i=1}^N \kappa_{2\sigma}(x_j - x_i).$$

(6)

IP(X) stands for Information Potential (IP), and represents average interactions among the data samples. Now, consider the error between the desired and the actual outputs of the system $e = d - y$. It can be demonstrated that $H^2_r(e=0)$ is a global minimum and that the nonparametric estimator $\hat{H}^2_r$ preserves this property [13]. Thus minimization of Renyi’s quadratic entropy of the error as a criterion for parameter optimization is feasible, and it was demonstrated to be effective on the training of several types of systems, including traditional adaptive filters, neural networks, and various algorithms in machine learning [6, 12-15].

Using (1), (5) and (6), minimizing Renyi’s quadratic entropy of the error becomes finding $\xi$ such that

$$\xi = \min_{\omega} \hat{H}^2_r(E) = \min_{\omega} \left(-\log \int \hat{f}^2_r(e)de\right) = \max_{\omega} \frac{1}{N^2} \sum_{j=1}^N \sum_{i=1}^N \kappa_{2\sigma}(e_j - e_i)$$

(7)

since Renyi’s quadratic entropy is a monotonic negative function of the information potential. This results in the so-called Minimum Error Entropy (MEE) criterion.

III. A NEURO-FUZZY ARCHITECTURE FOR REGRESSION

Here we present an architecture to implement a fuzzy system based on adaptive networks. This architecture is based on the well-known structure ANFIS proposed by J.S. Jang [4] and later generalized by C.T. Sun [16].

FIGURE 1. Architecture for an adaptive-network-based fuzzy inference system

### Table I. Number of nodes and number of parameters per node for architecture of Figure 1.

<table>
<thead>
<tr>
<th>Layer</th>
<th>#Nodes</th>
<th>Parameters per Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$V$</td>
<td>1 (V+1)</td>
</tr>
<tr>
<td>2</td>
<td>$V$</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>$L_i$</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>$R$</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>$R$</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>$R$</td>
<td>1</td>
</tr>
</tbody>
</table>

$V$ is the number of input variables; $L_i$ is the number of linguistic terms for the variable $i$; and $R$ is the number of rules.

Figure 1 shows this architecture for the case of two input variables, $x_1$ and $x_2$, and one output variable $y$. Following Jang’s notation, in this kind of networks a circle denotes a node without parameters; otherwise a square is used. The proposed structure is organized into six functional layers following a feed-forward processing manner.

Each node at layer 1 implements a weight of importance $w_i$ associated with the respective input variable $i$. This allows...
the implementation of feature selection during the learning adaptation, according to the evolution of the weight assigned to the corresponding input variable. As closer the weight is to 1, the more the influence of the variable on the posterior input fire strength (layer 2). On the contrary as long as the variable is less important on the construction of the output, this weight will tend to approximate to 0.

Every node at layer 2 is a square node associated with a parameterized membership function \( \mu_A(x_i) \) where \( x_i \) is one of the input variables and \( A \) is the linguistic term associated with this node function. Each node connected with an input variable represents a different fuzzy set for the corresponding variable. Thus the resulting degree of membership, taking into account the corresponding weight \( w_i \) of each rule's consequence. Therefore, is equivalent to the firing strength of each fuzzy rule is calculated as the average of each rule's consequence. The final output of the system. Note that in this architecture the consequence of each rule is a linear combination of the inputs, and the final output is obtained as the weighted average of each rule's consequence. Therefore, is equivalent to a Takagi-Sugeno's type fuzzy inference system of order one [18]. However, other types of fuzzy reasoning methods can also be implemented with this model [4].

IV. Fuzzy modeling

As previously mentioned on the introduction, system modeling mainly consists of two parts: structure identification and parameter optimization. On fuzzy modeling, and taking the proposed architecture as reference, the former implies finding the appropriate fuzzy partitions on the input space, as well as determining the number of rules, and their antecedent and consequent parts. Several methodologies and algorithms were proposed through the literature to handle the problem of structure identification on fuzzy inference systems [16, 19-20]. Here a clustering algorithm is used to solve the structure identification problem. The algorithm proposed by L. Chiu [21] is a fuzzy extension of the Mountain Method [22] for estimating the number and initial locations of cluster centers from a set of \( n \)-dimensional points. The algorithm, also called **Subtractive Clustering**, considers each data point as a potential cluster center and defines a measure of potential for each data point \( x_i \)

\[
P_i = \sum_{j=1}^{n} e^{-\frac{r_a^2}{\sum_{j=1}^{n} r_b^2}}
\]

with \( r_a \) a positive constant defining a radius of neighborhood. Once the potential of every point has been computed, the point with highest potential is selected as the first cluster center. Subsequently, in function of a decay parameter \( r_b \), and in proportion to the distance to the first cluster, the potential is reduced (subtracted) from the rest of data points. The point with the remaining highest potential is then selected as the second cluster center. The algorithm continues until the maximum potential reaches a certain threshold.

After the cluster centers have been determined, each cluster center can be thought as a prototypical data point that exemplifies a characteristic behavior of the system. Hence, each cluster center can be used as the basis of a rule that describes the system behavior [21].

Once determined the structure of the FIS one can start finding the optimal values for the parameters. Taking into account Table I and the fact that, in the proposed method, the input is partitioned according to Gaussian membership functions of mean \( \mu \) and standard deviation \( \sigma \) (thus, 2 parameters), one can calculate the total number of parameters to optimize. Of course, the number of clusters returned by the clustering algorithm, and therefore, the number of parameters to optimize, depends on the complexity of the data set as well as the values for the parameters \( r_a, r_b \) and thresholds for the stopping criteria of the clustering algorithm.

 Independently on the searching strategy used to search across the space of possible values, usually, parameter optimization algorithms are based on the minimization of a cost function. MSE is still the most extensively used form of
cost function in machine learning. Here we propose MEE criterion described in Section III as a new criterion for adapting the parameters of the FIS.

Parameter’s optimization minimizing a cost function comprises searching across the range of possible values for the parameters. At each iteration, the cost function is pretended to return a lower value. Typical algorithms to search in the weights space are based on the direction marked by the gradient on the error surface. However, these locally guided techniques often suffer from the problem of local minima. Here we employed a Genetic Algorithm (GA) which is a kind of global search strategy, in order to find out the parameters that minimize Renyi’s quadratic entropy. Inspired in biological nature selection, the GA repeatedly modifies a population of individual solutions and performs evolution through operations of selection, crossover and mutation [23].

V. RESULTS

Fuzzy modeling using the architecture presented on Section III was tested for prediction of future values in the Mackey-Glass (MG) chaotic time series [24] using both MSE and MEE minimization procedures. The time series is defined by the following differential equation:

$$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1 + x^10(t-\tau)} - bx(t)$$  \hspace{1cm} (11)

With \(a=0.2\), \(b=0.1\) and \(\tau = 17\), the time series is chaotic, exhibiting a cyclic but not periodic behavior. The fourth-order Runge-Kutta method was used to obtain the time series values. Initial condition \(x(0) = 1.2\) was used and \(x(t)\) was calculated for \(0 \leq t \leq 2000\) assuming \(x(t) = 0\) for \(t < 0\).

The goal is to predict the value of the series at some point \(x = t + P\), knowing values of the series up to the current \(x = t\). The standard method consists in creating a mapping from \(D\) points spaced \(\Delta\) apart, i.e., \((x(t-D-1)\Delta), \ldots, x(t-\Delta), x(t))\). Values of \(D = 4\) and \(\Delta = P = 6\) were used; therefore prediction for 6 steps on the future is intended, which is one well-reported problem in the literature [4, 21, 25]. According with this configuration and the estimated dimension of the attractor to be \(\approx 2.1\) [25], learning patterns were thus obtained as:

\([x(t-24), x(t-18), x(t-12), x(t-6), x(t)]\) \hspace{1cm} (12)

For comparison purposes among the two cost functions, a 10-fold cross-validation was used with the training set and the accuracy of the prediction was measured over the test sets using the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE}$$ \hspace{1cm} (13)

As previously mentioned, the application of the subtracting clustering algorithm for structure identification returns a fuzzy partitioned input space based on Gaussian membership functions. In order to keep the standard deviation parameter \(\sigma\) positive, linear constraints are placed on the GA to bound the search space to bring positive values for \(\sigma\). Similarly, we would also like to maintain the possible values for weights on layer 1 in the range [0,1].

Initial conditions were established the same for both MSE and MEE in order to make both procedures comparables. GA evolved for 40 generations using different number of individuals in the population. When using IP, and after training, the bias weights of the output processing elements were adjusted to yield zero error mean over the training set. Kernel size for (6) was adapted each iteration using the Silverman’s rule [11].

Table II shows the results for the simulations over the test sets. RMSE achieved on the 10-fold is indicated as mean±standard deviation. Results clearly show that information potential criterion achieve the best results.

VI. DISCUSSION AND CONCLUSIONS

Fuzzy systems allow us to come up within domains in which either data is imprecise or the inference process is affected by uncertainty. Modeling such systems is not an easy task; one has to tune up a lot of parameters, e.g. the type, number and values of the appropriate membership functions. The precise number of parameters heavily depends on the problem size and the concrete fuzzy architecture. Also when no expert knowledge is available, this problem aggravates, and the use of automatic procedures to tune up system’s parameters becomes necessary. In this work we have reported a methodology to perform fuzzy modeling in regression problems. We took a generalization of the ANFIS structure as reference to perform this task. Modeling is considered as two sub-problems: (i) structure identification and (ii) parameter optimization. While the former is solved applying a well-kwon clustering algorithm, the main contribution resides in the second sub-problem. Here we contribute proposing a new criterion to adjust the parameters of the fuzzy system based in information theory. ILT offer us a framework for machine learning in which we can employ cost functions that uses more information from the error PDF than those based on the typical L2-norms, as in the case of the MSE. It is interesting that while the majority of the effort in machine learning is usually done on the development of new learning techniques, it seems that no much attention is directed on the role that cost functions play to guide the learning algorithm. Minimum Error Entropy
criterion was already used with success in several types of systems, including traditional adaptive filters, neural networks, and various algorithms in machine learning [12-15]. Here we demonstrated its validity in modeling of fuzzy systems. Future work in this field will include to experiment with other measures born from ITL principles and investigate their capacity on some other fuzzy applications such as the classification task. Also in a more practical application domain, the authors are interested in incorporating these fuzzy modeling techniques in constructing a system with capabilities to reasoning, dealing with imprecision and uncertainty, in the field of sleep diseases diagnosis [26-27].

REFERENCES