Qualitative Modelling of Time Series Using Self-Organizing Maps: Application to Animal Science

EMILIO SORIA♣, JOSE D. MARTIN♣, CARLOS FERNANDEZ♠,
RAFAEL MAGDALENA♣, ANTONIO J. SERRANO♣, ÁLVARO MORENO♣
♣Departamento de Ingeniería Electrónica, E.T.S.E, University of Valencia
C/ Dr. Moliner 50, 46100 Burjassot (Valencia).
♠Departamento Producción Animal, Universidad Cardenal Herrera CEU,
Edificio Seminario s/n. 46113 Moncada, Valencia, Spain

Abstract: - In this work, a new methodology for modeling qualitative temporal processes, is proposed. In this development, two stages are considered. First, models are obtained and adapted fitting their free parameters according to the existing time series. Second, Self-Organizing Maps (SOM) are used to establish a clustering of the data using the parameters obtained in the first stage. This methodology is successfully applied to a problem of milk yield prediction in goat herds.

Key-Words: - Time Series, Neural Networks, Mathematical Models.

1 Introduction

There is an extensive literature about time modeling of processes, including both linear and nonlinear models [1-3]. In all these models, there are a number of free parameters that should be fit. Once these parameters are fit, it should be analyzed the goodness of the model as well as its capacity of generalization.

There are many problems in which the processes to model do not present a unique behavior, that is, considering the same mathematical model base for all the processes, the value of the different parameters differs according to the process. As an example, we can think about a model that is a damped sine; the value of decay of the exponential and the frequency of the sine may vary for the different processes. In this situation, the development of local models for each process can be carried out after a classic analysis of clustering [4]. Therefore, this approach will need a first segmentation and characterization of the time series in order to carry out the modeling.

We propose a different approach in this communication. First, a basic model is taken into account and then, a clustering of processes is carried out using the parameters of the developed mathematical model.

The present communication shows the advantages of the approach proposed, complementing this description with an example of time series in Animal Sciences.

The remainder of the paper is organized as follows: in Section 2, the self-organizing maps are explained due to its essential role in this work. Section 3 presents the proposed procedure and, Section 4 the obtained results, ending up the paper with the conclusion of the work in Section 5.

2 Self-Organizing Maps

The Self-Organizing Map (SOM) is a neural model that is inspired on the fact that similar tasks take place in similar areas of the brain. In this model, neurons are arranged in two layers. The first layer is called the input or sensorial layer, and it consists of *n* neurons, being *n* the number of input variables. The processing is carried out in the second layer or competition layer, that forms the map of characteristics [3-4]:

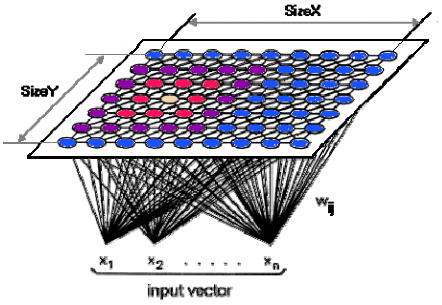

Fig. 1. SOM architecture.

In this structure, a vector of weights is assigned to each neuron of the output layer; this vector should have the same dimension as the input vectors, that is, for the neuron *i,j* (first index row, second column): $\vec{w}_{ij} = \left[w_{ij}^1 \ldots\ldots w_{ij}^m\right]$.

The processing of this network maps similar input patterns into close areas of the output layer of SOM [4]. To achieve this goal, the following learning algorithm for the synaptic coefficients is used [3-4]:

1) Weight initialization.

2) An input pattern is selected: $\vec{x}_k = \left[x_k^1 \ldots\ldots x_k^m\right]$.

3) Calculation of the similarity between the weights associated with all the neurons and the corresponding input vector. If we consider the Euclidean distance as a comparison measure, then:

$$d\left(\vec{w}_{ij}, \vec{x}_k\right) = \sum_{s=1}^{m} \left(w_{ij}^s - x_k^s\right)^2 \qquad (1)$$

4) Determination of the closest neuron to the input pattern, the so-called Best Matching Unit (BMU).

5) Update of the synaptic weights; if the quadratic function is used as similarity function, then:

$$\vec{w}_{ij} = \vec{w}_{ij} + \alpha \cdot h\left(NG, \vec{w}_{ij}\right) \cdot \left(\vec{x}_k - \vec{w}_{ij}\right) (2)$$

where α is the learning rate and *h(x)* is known as neighborhood function. This function is a function whose value depends on the distance between the BMU and the updated neuron. This function is thus responsible of maintaining the topological relationships among patterns when the input space is mapped into the two-dimensional space defined by the network architecture [4].

6) If the maximum number of iterations is reached, end; otherwise, go to step 2.

At the end of the learning stage, a two-dimensional map is obtained. This map provides qualitative, valuable information about how the input variables are related among themselves. With the SOM, a data clustering is obtained, determining similar groups within our data. It should be emphasized the role that SOM plays as a visual tool in data mining; it is possible to use a color-based representation of the components of the vectors corresponding to every neuron in the map, and therefore, it is straightforward to figure out the relationships amongst the input variables.

3 Proposed methodology

There are two main approaches to model a time series: either a parametric or a nonparametric model. There is still a third option, the so-called semiparametric modeling, but it is not necessary to study in depth this third option for the goals of this work [2]. Our methodology is based on the use of parametric models.

The parametric models assume a certain probability distribution, or a certain mathematical expression, as basis of the process that is wanted to describe. In these mathematical expressions, there are some terms that must be fit [2]. Therefore, the way to work with these models is first to define a basis model in order to adjust the parameters of this model and then, to check its performance.

However, one of the problems of modeling time processes is the existence of groups that, for the same basis model, have different values of the parameters. In these case, a classic approach to the problem is to carry out a clustering on the different processes depending on the time characteristics; after this clustering, the parameters can be fit to these groups. An extreme approach in this sense is often made in Animal Science [7-8]. In this approach, an average is carried out, and the parameters are adjusted according to this average. Nonetheless, the use of this approach involves losing much information about individual processes.

Our approach is based on fitting the parameters for each one of the time processes, and then, obtaining a SOM, which is characterized by these parameters. The main advantages of our approach are the following:

1. It is fully generic and can be applied to any parametric model (linear or non-linear).
2. The clustering is applied to parameters that depend on the modeling of the process, not on intermediate results The fact of using a SOM allows a qualitative analysis for each group.

Next section shows results achieved in a modeling example, using the proposed methodology in a real problem, which presents serious difficulties in its modeling and ulterior information extraction: the prediction of Milk Yield (MY) in dairy goats [5]. It is a strong non-linear problem, in which different goats have different time characteristics.

4 Results

Milk yield (MY) is a key factor in the management of goat livestock since it may become the main farmer income. MY is important to identify which goats are the best milk producers, also to analyze the current state of a certain livestock.
Therefore, MY prediction is very helpful as it enables the farmer to make decisions about livestock management before the end of lactation, when decisions are usually made. A homogenous group of dairy goats within a commercial farm (Excamur, S.L., Spain) was selected in this trial. This farm is a member of ACRIMUR (Murciano-Granadina Goat Breeder Association—Asociación Española de Criadores de la Cabra Murciano-Granadina). The selected farm is a representative sample of these kinds of farms in South-Eastern Spain. The data set used in this work was formed by 34 goats. A weekly follow-up was carried out for every goat, being the study period of 22 weeks.
The following figure shows the milk production for three different goats:

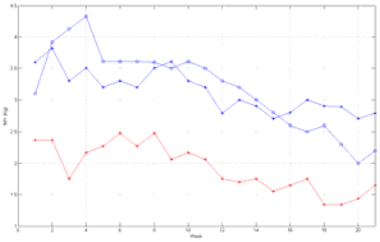

Fig. 2. Weekly milk yield for three goats.

In order to model MY of the different goats, a second order model was taken into account, although the followed methodology could be used for any type of model. The solution of a second order differential equation can give the following models:

Model_1:

$$f(t) = k_1 + k_2 \cdot e^{k_3 \cdot t} + k_4 \cdot e^{k_5 \cdot t} \quad (3)$$

Model_2:

$$f(t) = k_1 + k_2 \cdot e^{k_3 \cdot t} \cdot \cos(k_4 \cdot t) \quad (4)$$

Model_3:

$$f(t) = k_1 + k_2 \cdot e^{k_3 \cdot t} \cdot \sin(k_4 \cdot t) \quad (5)$$

The following step was to determine the parameters that appear in the equations (3)-(5) for each of the goats. A gradient procedure was followed using the least square function as cost function [2]. The third model was the most accurate one, and hence, this model was analyzed using our methodology. Figure 3 shows the achieved modeling.

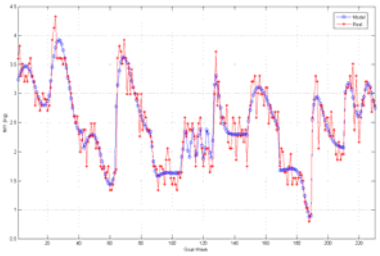

Fig. 3. Weekly milk production and modeling using the third model. The different goats are arranged in the *x*-axis.

As it is shown in Fig. 3, the modeling offered by the third model was accurate enough. The following step was to obtain a SOM which was trained by using the values of the parameters that appear in (5): *k1*, *k2*, *k3*, *k4*. An architecture with hexagonal neighbors was used, being the size of the map equal to 65, according to the guidelines proposed in [4]. A batch algorithm was used to train the neural network. Finally, the map of components for the experimental data set was obtained, as it is shown in Fig. 4.

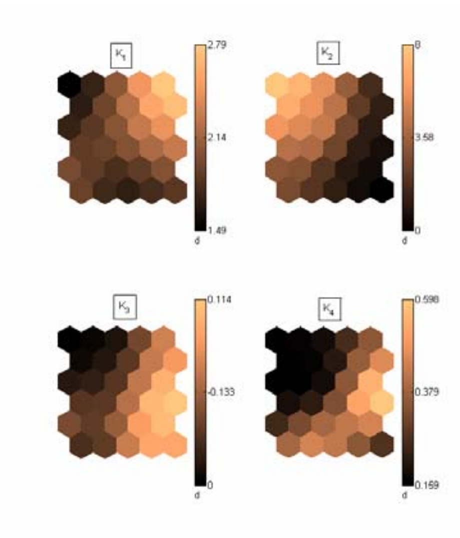

Fig. 4. Map of components.

Once the map was obtained and using (5), it was possible to analyze the different time behaviors that appeared in the data set, and it was also possible to compare the variation of the different parameters altogether. As an example, the upper-left corner of the map shows the lowest values of the parameter *k1* which correspond with the constant level throughout the weeks; it also corresponds with the highest values of the parameter *k2* (factor that multiplies the exponential). Moreover, *k3* has a value close to zero, hence, this exponential becomes as a multiplying factor of 1. Finally, the factor with lowest variation (or frequency) corresponds to parameter *k4*. It means that these goats, whose production is more or less constant throughout the time, are clustered in the same group.

A second example of information extraction can be obtained from the lower-right area of parameter *k2* in Fig. 4; as it is shown, the tonality is rather dark, i.e., this parameter is very close to zero. Therefore, the group of goats allocated in this area, present a more or less constant MY throughout the lactation and this production depends on the constant value of parameter *k1*. In this area, there are 6 BMUs, which stands for a total of 17 goats (50% of the studied goats). Neuron 28 is the one that includes the highest number of goats (6 goats).

It should be emphasized the relevance of the qualitative analysis that can be carried out on any area of the map and, for any parametric model.

5 Conclusion

In this communication, a new methodology for the qualitative analysis of parametric models has been proposed. It is completely general and can provide a descriptive knowledge of the processes involved in the modeling. Its effectiveness has been demonstrated on a nonlinear process, obtaining relevant qualitative information.

References

[1] G. Box, G.M Jenkins, G. Reinsel, *Time Series Analysis: Forecasting & Control*. Prentice 1994.
[2] O. Nelles, *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Springer, 2000.
[3] S. Haykin, *Neural Networks: A Comprehensive Foundation,* Prentice Hall, 1999.
[4] T. Kohonen, *Self Organizing Maps*, Springer, 2000.
[5] C. Fernández, E. Soria, J.D. Martín, A.J. Serrano, *Neural Networks for Animal Science: Two case studies*. Expert Systems with Applications, vol 31, pp 444-450, 2006.