Fuzzy Inference Based Autoregressors for Time Series Prediction Using Nonparametric Residual Variance Estimation

Federico Montesino Pouzols, Amaury Lendasse and Angel Barriga

***Abstract*—We apply fuzzy techniques for system identification and supervised learning in order to develop fuzzy inference based autoregressors for time series prediction. An automatic methodology framework that combines fuzzy techniques and statistical techniques for nonparametric residual variance estimation is proposed. Identification is performed through the learn from examples method introduced by Wang and Mendel, while the Marquard-Levenberg supervised learning algorithm is then applied for tuning. Delta test residual noise estimation is used in order to select the best subset of inputs as well as the number of linguistic labels for the inputs. Experimental results for three time series prediction benchmarks are compared against LS-SVM based autoregressors and show the advantages of the proposed methodology in terms of approximation accuracy, generalization capability and linguistic interpretability.**

I. Introduction

TIME series prediction and analysis in general is a recurrent problem in virtually all areas of natural and social sciences as well as in engineering. In the time series prediction field, prediction accuracy is not the only major goal. Understanding the behavior of time series and gaining insight into their underlying dynamics is a highly desired capability of time series prediction methods.

In the past, conventional statistical techniques such as AR and ARMA models have been extensively used for forecasting However, these techniques have limited capabilities for modeling time series data, and more advanced nonlinear methods including neural networks have been frequently applied. Fuzzy inference systems, despite its good performance in terms of accuracy and interpretability [1], have seen little application in the field of time series prediction as compared to other nonlinear modeling techniques such as neural networks and support vector machines.

In this paper, we propose a methodology framework to perform autoregressive time series prediction by means of fuzzy inference systems. We will call fuzzy autoregressors those autoregressors implemented as fuzzy inference systems. This is not to be confused with what is usually called fuzzy regression in the literature [2]. The methodology proposed here is intended to apply to crisp time series.

In practice, one founds two problems when building a fuzzy model for a time series: choosing the inputs to the inference system, and identifying the structure of the system The first problem is addressed by means of a priori feature selection techniques based on nonparametric residual variance estimation. The second problem is addressed by techniques for identification of fuzzy systems from numerical examples, such as the algorithm by Wang and Mendel (W&M) [1] and identification algorithms based on clustering techniques [3].

This paper also addresses a recent challenge in the field of time series prediction: long-term prediction (as a generalization to short-term prediction), for which lack of information and accumulated errors pose additional difficulties. Also, real world benchmarking time series, instead of synthetic series (chaotic but noise-free) are analyzed. Experimental results are compared against least-squares support vector machines (LS-SVM) [4], a well stablished method in the field of time series prediction.

The next section describes nonparametric residual variance estimation. In section III we propose a methodology framework and one concrete implementation based on well known algorithms. Section IV illustrates the methodology through a case study. Sections V and VI present and further discuss experimental results for a number of time series benchmarks.

II. Nonparametric Residual Variance Estimation: Delta Test

Nonparametric residual variance estimation (or nonparametric noise estimation, NNE) is a well-known technique in statistics and machine learning, finding many applications in nonlinear modeling [5].

Delta Test (DT) is a NNE method for estimating the lowest mean square error (MSE) that can be achieved by a model without overfitting the training set [5]. Given N multiple input-single output pairs, $(\bar{x}_i, y_i) \in R^M \times R$, the theory behind the DT method considers that the mapping between $\bar{x}_i$ and y_i is given by the following expression:

$$y_i = f(\bar{x}_i) + r_i,$$

where f is an unknown perfect fitting model and r_i is the noise. DT is based on hypothesis coming from the continuity of the regression function. When two inputs x and x' are close, the continuity of the regression function implies that outputs $f(x)$ and $f(x')$ will be close enough. When this

Federico Montesino Pouzols is with the Microelectronics Institute of Seville, CSIC, Scientific Research Council, Avda. Reina Mercedes s/n. Edif. CICA. E-41012 Seville, Spain (phone: +34 955 056 666; fax: +34-955-056-686; email: federico.montesino@imse.cnm.es).

Amaury Lendasse is with the Laboratory of Computer and Information Science of the Helsinky University of Technology. P.O. Box 5400, FIN-02015 HUT, Finland (phone: +358-9-451 3267; fax: +358-9-451 3277; email: lendasse@cis.hut.fi).

Angel Barriga Barros is with the Department of Electronics and Electromagnetism of the University of Seville, E-41012, Spain (phone: +34 955 056 666; fax: +34-955-056-686; email: barriga@us.es).

This work has been supported in part by project TEC2005-04359/MIC from the Spanish Ministry of Education and Science as well as project TIC2006-635 and grant IAC07-I-0205:33080 from the Andalusian regional Government.

978-1-4244-1819-0/08/$25.00©2008 IEEE

implication does not hold, it is due to the influence of the noise.

Let us denote the first nearest neighbor of the point $\bar{x}_i$ in the set $\{\bar{x}_1, \ldots, \bar{x}_N\}$ by $\bar{x}_{NN}$. Then the DT, δ, is defined as follows:

$$\delta = \frac{1}{2N} \sum_{i=1}^{N} \left| y_{NN(i)} - y_i \right|^2 ,$$

where $y_{NN(i)}$ is the output corresponding to $\bar{x}_{NN(i)}$. For a proof of convergence, refer to [6]. DT has been shown to be a robust method for estimating the lowest possible mean squared error (MSE) of a nonlinear model without overfitting. DT is useful for evaluating nonlinear correlations between random variables, namely, input and output pairs. This method will be used for a priori input selection.

III. Methodology Framework for Time Series Prediction with Fuzzy Inference Systems

Consider a discrete time series as a vector, $\bar{y} = y_1, y_2, \ldots, y_{t-1}, y_t$ that represents an ordered set of values, where t is the number of values in the series. The problem of predicting one future value, y_{t+1}, using an autoregressive model (autoregressor) with no exogenous inputs can be stated as follows:

$$\hat{y}_{t+1} = f_r(y_t, y_{t-1}, \ldots, y_{t-M+1})$$

Where $\hat{y}_{t+1}$ is the prediction of model f_r and M is the number of inputs to the regressor.

Predicting the first unknown value requires building a model, f_r, that maps regressor inputs (known values) into regressor outputs (predictions). When a prediction horizon higher than 1 is considered, the unknown values can be predicted following two main strategies: recursive and direct prediction.

The recursive strategy applies the same model recursively, using predictions as known data to predict the next unknown values. For instance, the third unknown value is predicted as follows:

$$\hat{y}_{t+3} = f_r(\hat{y}_{t+2}, \hat{y}_{t+1}, y_t, y_{t-1}, \ldots, y_{t-M+3})$$

It is the most simple and intuitive strategy and does not require any additional modeling after an autoregressor for 1 step ahead prediction is built. However, recursive prediction suffers from accumulation of errors. The longer the prediction term is, the more predictions are used as inputs. In particular, for prediction horizons greater than the regressor size, all inputs to the model are predictions.

Direct prediction requires that the process of building an autoregressor be applied for each unknown future value. Thus, for a maximum prediction horizon H, H direct models are built, one for each prediction horizon h:

$$\hat{y}_{t+h} = f_h(y_t, y_{t-1}, \ldots, y_{t-M+1}), \text{ with } 1 \leq h \leq H$$

While building a prediction system through direct prediction is more computationally intensive (as many times as values are to be predicted) it is also straightforward to parallelize. Direct prediction does not suffer from accummulation of prediction errors.

In this paper, we follow the direct prediction strategy. In order to build each autoregressor, a fuzzy inference system is defined as a mapping between a vector of crisp inputs, and a crisp output. In principle, any combination of membership functions, operators and inference model can be employed, but the selection has a significant impact on practical results. As a concrete implementation, we use the minimum for conjunctions and implications, gaussian membership functions for inputs, singleton outputs and fuzzy mean as defuzzification method following the Mamdani defuzzification model. In this particular case a fuzzy autoregressor with M inputs for prediction horizon h is formulated as:

$$\mathcal{F}_h(\bar{y}) = \frac{\sum_{l=1}^{N_h} \min\left(\mu_{R_l^h}, \min_{1 \leq v \leq M} \mu_{L_l^{i,h}}(y_v)\right)}{\sum_{l=1}^{N_h} \min_{1 \leq v \leq M} \mu_{L_l^{i,h}}(y_v)}$$

Where N_h is the number of rules in the rulebase for horizon h, $\mu_{L_l^{i,h}}$ are gaussian membership functions for the input linguistic labels and $\mu_{R_l^h}$ are singleton membership functions.

The problem of building a regressor can be precisely stated as that of defining a proper number and configuration of membership functions and building a fuzzy rulebase from a data set of t sample data from a time series such that the fuzzy systems $\mathcal{F}_h(\bar{y})$ closely predict the $h-$th next values of the time series. The error metric to be minimized is the mean squared error (MSE).

We propose a methodology framework in which a fuzzy inference system is defined for each prediction horizon throughout the stages shown in figure 1. These stages are detailed in the following subsections.

A. Variable Selection

As first step in the methodology, DT estimates are employed so as to perform an a priori selection of the optimal subset of inputs from the initial set of M inputs, given a maximum regressor size M. Variable selection requires a selection criterion. We use the result of the DT applied to a particular variable selection as as a measure of the goodness of the selection. The input selection that minimizes the DT estimate is chosen for the next stages.

In addition, a selection procedure is required. For small (up to around 10-20) regressor sizes, an exhaustive evaluation of DT for all the possible selections (a total of $2^M - 1$) is feasible. We will call this procedure *exhaustive DT search*. Its main advantages is that the optimal selection is found.

For higher regressor sizes, forward-backward search of selections (FBS) [7] is employed. This procedure combines both forward and backward selection. Although this procedure does not guarantee optimality, a balance between performance and computational requirements is achieved.

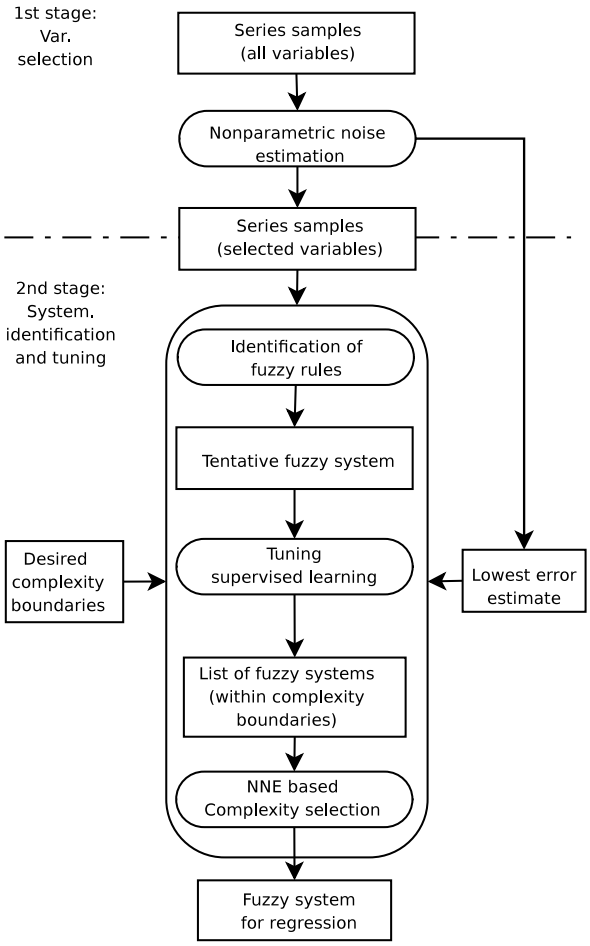

Fig. 1. Methodology Framework for Time Series Prediction.

B. System Identification and Tuning

This stage comprises three substages that are performed iteratively and in a coordinated manner. The whole process is driven by the third (complexity selection) substage, until a system that satisfies a training error condition derived from the DT estimate is constructed.

1) Stage 2.1: System identification: In this substage, the structure of the inference system (linguistic labels and rule base) is defined. For the concrete implementation analyzed in this paper, identification is performed using the W&M algorithm driven by the DT estimate. Though many modifications to the original algorithm have been proposed throughout the years, for the sake of simplicity we adhere to the original algorithm specification in [1] as implemented in version 3.2 of the Xfuzzy design environment [8].

For identification, one or more parameters are usually required that specify the potential complexity of the inference system. Thus, the desired boundaries of complexity for the systems being built are additional inputs to the process. In the case of the W&M algorithm, the number of labels per input must be specified a priori. Our approach is to explore systems in an increasing order of complexity, from the lowest possible number of labels up to a maximum specified as complexity boundary. The same number of labels is used for each input.

This iterative identification process for increasing grid partitions of the universe of discourse stops when a system is built such that the training error is lower than the DT estimate or a threshold based on the DT estimate. The selection is made by comparing the error after the next (tuning) stage.

2) Stage 2.2: System Tuning: We consider an additional tuning step in the methodology as a substage separated from the identification substage. Note that in some cases (as in the H&G [9] algorithm), these two substages can be integrated into a standalone algorithm. The tuning process is driven by one or more error metrics.

As concrete implementation for this paper we apply the Levenberg-Marquardt supervised learning algorithm driven by the normalized MSE (NMSE). All the parameters of the membership functions of every input and output are adjusted using the algorithm implementation in the Xfuzzy development environment [10].

3) Stage 2.3: Complexity Selection: As last step, the complexity of the fuzzy autoregressors (measured as the number of linguistic labels per input in our concrete implementation) is selected depending on the DT estimate. The first (simplest) system that falls within the error range defined by the DT-NNE is selected.

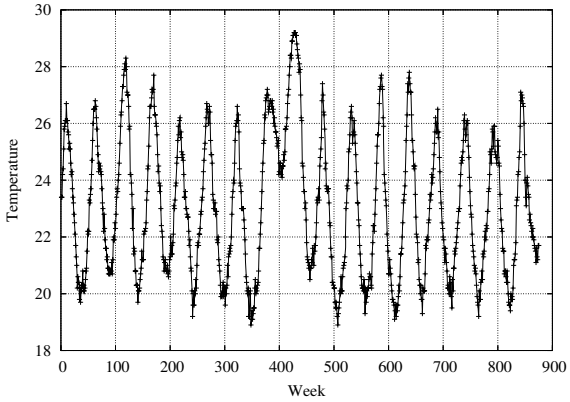

Fig. 2. ESTSP´07 competition data set (875 samples).

IV. Case Study and Validation: ESTSP 2007 Competition Dataset

For the purposes of validating and illustrating the proposed methodology framework and concrete algorithms and criteria, we analyze the data set from the ESTSP 2007 time series prediction competition [11] (ESTSP´07). This data set (see figure 2) consists of 875 samples of temperatures of the El Niño-Southern Oscillation phenomenon.

The original ESTSP´07 series is splitted into two subsets: a training set (first 475 samples) and a second set (last 400 samples) that will be used for validation. We will call

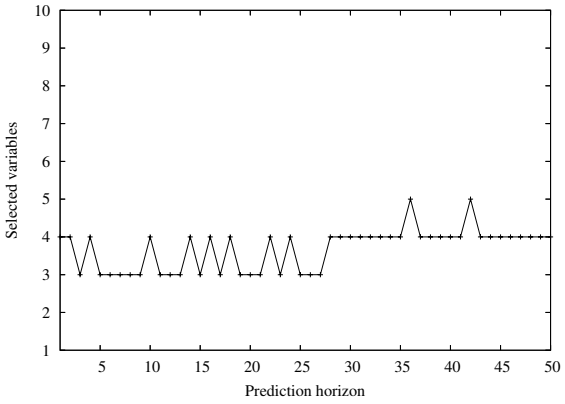

Fig. 3. ESTSP´07-T: Number of selected variables for horizon up to 50. DT based selection with exhaustive search. Maximum regressor size 10.

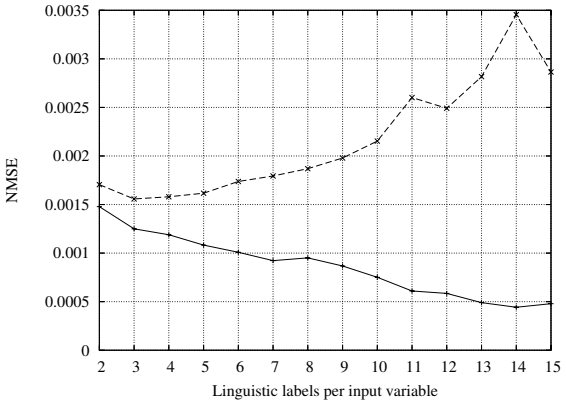

Fig. 4. ESTSP´07-T: Errors for horizon 1, exhaustive DT based selection of inputs. Continous line: training error. Dashed line: validation error.

this series ESTSP´07-T. Though one of the major goals of the proposed methodology is to avoid the requirement of validation and test series, we define two subsets in order to validate the methodology with the residual noise estimator and algorithms being used.

A maximum regressor size of 10 and a prediction horizon of 50 are considered. As first stage within our methodology, DT is performed on the training set for all the possible variable selections ($2^{10} - 1$) and the one with lowest DT estimate is chosen. This process is performed independently for each prediction horizon. The number of selected variables is shown in figure 3.

As second stage, the W&M algorithm is applied to the training set in order to identify fuzzy inference systems. These models are then tuned through supervised learning using the Levenberg-Marquardt algorithm over the training set. The process is repeated for increasing numbers of linguistic labelsper input, starting from 2. Within this iterative process, the DT estimate is used to check whether the best possible approximation has been achieved, i.e., the right compromise between model complexity and training error has been found.

After the tuning substage, there is a considerable performance increase as for accuracy (the MSE decreases around 1 order of magnitude). In particular, tuned systems with a low number of rules perform better than untuned systems with a much greater complexity. Thus, the supervised learning substage also contributes to reducing model complexity.

The DT estimate threshold for horizon 1 is $1.26 \cdot 10^{-3}$ and, as shown in figure 4, the fuzzy system with 3 linguistic labels per input is chosen as autoregressor for horizon 1. Figure 5 shows the training and validation errors of the fuzzy autoregressors. Training and test errors of LS-SVM models are also shown. LS-SVM models were built with the same autoregressor size, input selection and training subset, using RBF kernels, gridsearch as optimization routine and crossvalidation as cost function. From figure 5, two main conclusions can be drawn from the comparison:

- As for generalization capability, the overall superiority of fuzzy regressors is specially evident for long-term prediction (beyond horizon 25).
- Training and generalization errors are much closer for fuzzy models than for LS-SVM models.

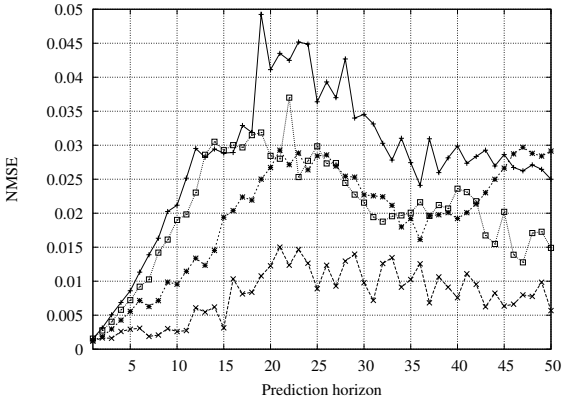

Fig. 5. ESTSP´07-T: comparison of our methodology against LS-SVM. Generalization errors of LS-SVM models ($+$). Generalization errors of fuzzy models ($\square$). Training errors of fuzzy models ($*$). Training errors of LS-SVM models ($\times$).

Figure 6 shows the predictions for the first 50 values after the training set together with a fragment of the actual time series.

V. Experimental Results

In this section, the proposed concrete implementation of the methodology framework described is applied to two time series prediction problems, namely the Poland electricity time series prediction benchmark and one of the series of the NN3 forecasting competition.

A. Poland Electricity Benchmark

This time series (PolElec henceforward) represents the normalized average daily electricity demand in Poland in the 1990´s. The benchmark consists of a training set of 1400

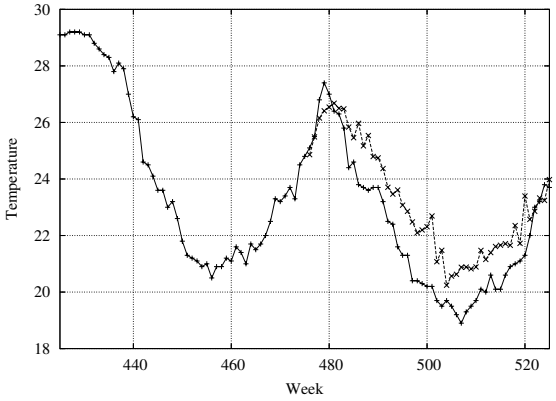

Fig. 6. ESTSP´07-T: Prediction of 50 values after the training set. Continous line: actual time series. Dashed line: predictions.

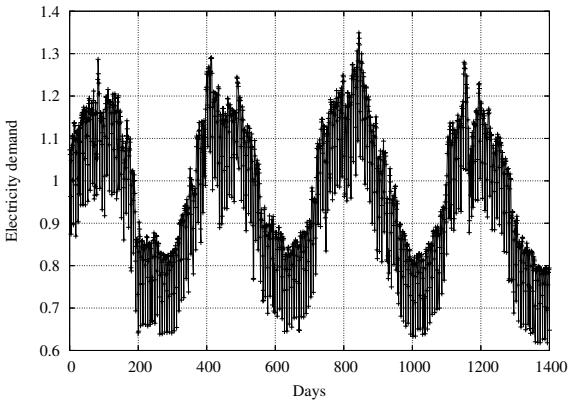

Fig. 7. PolElec: training series (1400 samples).

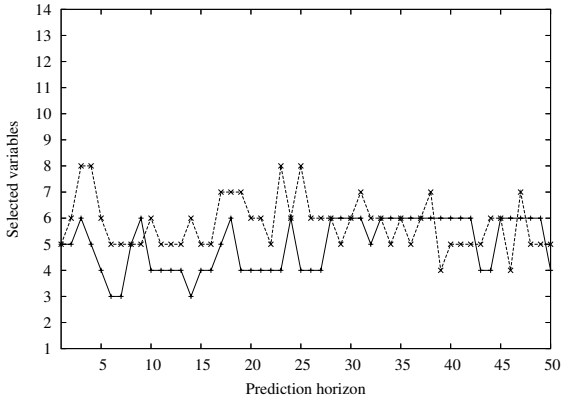

Fig. 8. PolElec: Number of selected variables (exhaustive DT based selection). Regressor sizes 7 (continuous line) and 14 (dashed line).

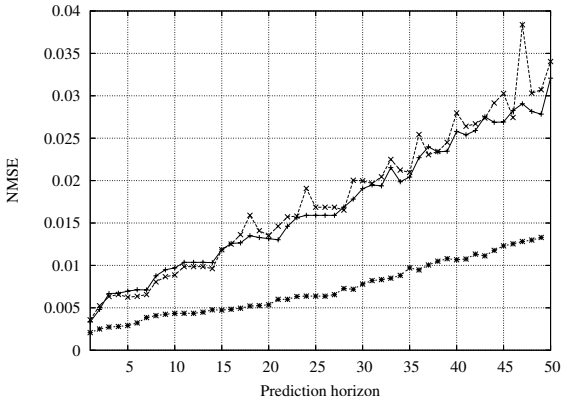

Fig. 9. PolElec: NDT estimates ($*$), training (+) and test (x) errors of fuzzy autoregressors. Maximum regressor size 7. Exhaustive DT based selection of inputs.

samples, shown in figure 7, and a test set of 201 samples. It has been shown that the dynamics of this time series is nearly linear [12]. Besides the yearly periodicity, a clear weekly periodicity can be seen on smaller time scales.

We will show the results obtained for two different maximum regressor sizes: 7 and 14. In both cases, input selection was performed by exhaustive search of the lowest DT estimate. The number of selected variables is shown in figure 8

For 7 steps ahead prediction, considering the notation for discrete time series introduced in section III, three input variables are selected to predict y_{t+7}: y_t, y_{t-1} and y_{t-5}. As an example of the interpretability of the models developed, let us suppose that the last 7 daily electricity demand measurements that are available correspond to the demand for a week from monday through sunday. Then, the fuzzy autoregressor predicts the demand for next Sunday based on the last known daily demand (Sunday), the demand of last Saturday and the demand of last Tuesday. A sample rule from the fuzzy inference based autoregressor would read as follows:

IF Tuesday was High AND Saturday was Low AND
Sunday was Low THEN NextSunday ← "0.92"

Where "0.92" is used as linguistic label for a singleton output centered approximately at 0.92.

TABLE I
TRAINING AND TEST ERRORS OF LS-SVM AND FUZZY MODELS AVERAGED FOR HORIZONS 1 THROUGH 50. ERRORS GIVEN AS NMSE. MAXIMUM REGRESSOR SIZE SPECIFIED BETWEEN PARENTHESIS.

	LS-SVM		Fuzzy inference	
Series	Training	Test	Training	Test
ESTSP´07-T (10)	$7.93{\cdot}10^{-3}$	$2.79{\cdot}10^{-2}$	$1.94{\cdot}10^{-2}$	$2.04{\cdot}10^{-2}$
PolElec (7)	$1.16{\cdot}10^{-2}$	$3.57{\cdot}10^{-2}$	$1.70{\cdot}10^{-2}$	$1.78{\cdot}10^{-2}$
PolElec (14)	$1.04{\cdot}10^{-2}$	$3.24{\cdot}10^{-2}$	$1.58{\cdot}10^{-2}$	$1.82{\cdot}10^{-2}$

Figure 9 show the DT estimates as well as training and test errors for the regressor with maximum size 7. The average training and test error of LS-SVM models are shown together with the errors of fuzzy models in table I. The accuracy

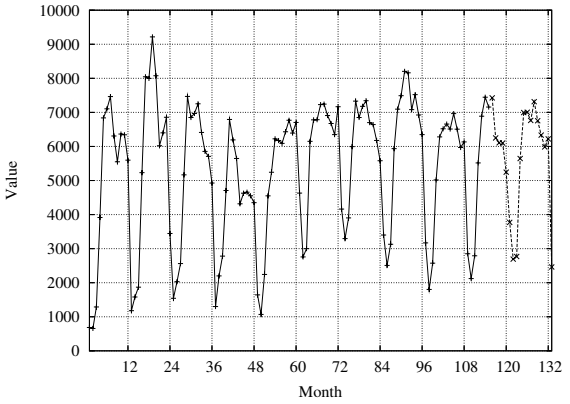

Fig. 10. NN3_104 series. 115 known values (continuous line) and 18 predictions (dashed line).

of fuzy autoregressors is better with no exception for any prediction horizon.

B. NN3 Competition

The NN3 forecasting competition [13] comprises a set of 111 series with monthly measures of financial variables. The next unknown 18 values have to be predicted. We analyze the time series number 104. The known values and predictions are shown in figure 10. These predictions were obtained using a maximum regressor size of 18. Variable selection was performed through exhaustive search up to size 12 extended with forward-backward search up to size 18.

From the plot, it can be concluded that the cyclic behavior of the series is correctly identified and the predictions are within reasonable boundaries. This result shows that the methodology employed can perform well when the training series is small.

VI. Discussion

A fundamental advantage of autoregressive time series prediction with fuzzy inference systems is that the rule based models can be linguistically interpreted by humans. For some time series, the most accurate rulebases have a low number of rules (below 15 or 10 rules). When the most accurate system has a high number of rules, there is still the possibility to build simpler, approximate models with a degree of accuracy of the same order of the most accurate model.

The methodology developed does not require a validation stage and thus the whole available data set can be used as training data to build autoregressive models. Several procedures have been shown to play a key role in achieving good approximation accuracy while keeping low complexity: variable selection, application of a supervised learning algorithm for tuning, and using DT-NNE for selecting the number of linguistic labels per input. The use of DT estimates has been shown to be advantageous in two main aspects:

- It does not only improve accuracy but also increases interpretability by decreasing the number of inputs.
- It is a robust solution to the problem of selecting the proper system complexity.

While LS-SVM are usually praised for their good generalization performance, we have shown that fuzzy autoregressors clearly outperform LS-SVM based autoregressors in terms of generalization capability. As far a s computational requirements is concerned, the methodology proposed has a very low cost compared against the LS-SVM method. A Java based implementation of the methodology presented is consistently between 1 and 3 orders of magnitude faster than the optimized C implementation of LS-SVM.

VII. Conclusion

We have developed an automatic methodology framework for long-term time series prediction by means of fuzzy inference systems. Experimental results for a concrete implementation of the methodology confirm good approximation accuracy and generalization capability.

Linguistic interpretability for both short-term and long-term prediction as well as low computational cost are two remarkable advantages over common time series prediction methods. Also, the proposed methodology has been shown to outperform LS-SVM based predictions in terms of approximation accuracy.

References

[1] L. Wang and J. M. Mendel, "Generating Fuzzy Rules by Learning from Examples," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 4, pp. 1414–1427, Dec. 1992.

[2] Yun-Hsi O. Chang and Bilal M. Ayyub, "Fuzzy regression methods - a comparative assessment," *Fuzzy Sets and Systems*, vol. 119, no. 2, pp. 187–203, Apr. 2001.

[3] S. L. Chiu, "A Cluster Estimation Method with Extension to Fuzzy Model Identification," in *IEEE Conference on Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence.*, vol. 2, no. 3, Orlando, FL, 1994, pp. 1240–1245.

[4] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002, ISBN: 981-238-151-1.

[5] A. J. Jones, "New Tools in Non-linear Modelling and Prediction," *Computational Management Science*, pp. 109–149, Sep. 2004.

[6] E. Liitiäinen, A. Lendasse, and F. Corona, "Non-parametric Residual Variance Estimation in Supervised Learning," in *WANN 2007, International Work-Conference on Artificial Neural Networks*, San Sebastián, Spain, Jun. 2007, pp. 63–71.

[7] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji, and A. Lendasse, "Methodology for Long-Term Prediction of Time Series," *Neurocomputing*, In Press, Corrected Proof Available online 21 May 2007.

[8] F. J. Moreno-Velo, I. Baturone, S. Sánchez-Solano, and A. Barriga, "Rapid Design of Fuzzy Systems With Xfuzzy," in *12th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'03)*, St. Louis, MO, USA, May 2003, pp. 342–347.

[9] C. M. Higgins and R. M. Goodman, "Fuzzy Rule-Based Networks for Control," *IEEE Transactions on Fuzzy Systems*, vol. 2, no. 1, pp. 82–88, Feb. 1994.

[10] F. J. Moreno-Velo, I. Baturone, A. Barriga, and S. Sánchez-Solano, "Automatic Tuning of Complex Fuzzy Systems with Xfuzzy," *Fuzzy Sets and Systems*, vol. 158, no. 18, pp. 2026–2038, Sep. 2007.

[11] "ESTSP´07 European Symposium on Time Series Prediction: Prediction Competition," http://www.estsp.org, Aug. 2007.

[12] A. Lendasse, J. Lee, V. Wertz, and M. Verleyssen, "Forecasting Electricity Consumption using Nonlinear Projection and Self-Organizing Maps," *Neurocomputing*, vol. 48, no. 1, pp. 299–311, Oct. 2002.

[13] "NN3 Artificial Neural Network & Computational Intelligence Forecasting Competition," http://www.neural-forecasting-competition.com, Aug. 2007.