Research on Time Series at the DiCITS Lab

J.M. Benitez, J.L. Aznarte, C. Bergmeir

Distributed Computational Intelligence and Time Series Labs
Department of Computer Science and Artificial Intelligence,
E.T.S. de Ingenierías Informática y de Telecomunicación,
CITIC-UGR, Universidad de Granada
e-mail: `J.M.Benitez@decsai.ugr.es`

Resumen The purpose of this communication is introducing the DiCITS Lab. One of the two cornerstones of this laboratory is the research on time series analysis and forecasting. DiCITS is one of the laboratories upon which the SCI2S research group at DECSAI-UGR is organized. We describe its current lines of research and some of its achievements. In particular, the software packages published with open-source licenses.

1. Introduction

DiCITS stands for *Distributed Computational Intelligence and Time Series* laboratory (`http://sci2s.ugr.es/dicits,dicits.ugr.es`). It is one of the laboratories upon which, the research group "Soft Computing and Intelligent Information Systems" (SCI2S) (`http://sci2s.ugr.es`) is organized. This research group is based at the Department of Computer Science and Artificial Intelligence (`http://decsai.ugr.es`) of the Universidad de Granada (`http://www.ugr.es`).

The research interest of the members of the laboratory are built on two wide scope topics: distributed computational intelligence, and time series analysis and forecasts.

The first topic focuses on techniques to develop high performance intelligent systems chiefly based on the Computational Intelligence (CI) paradigms. The current trends for high performance computer architecture go in the line of clusters of commodity components, towards distributed computing where independent processing units communicate through network connections. In addition to that, microprocessors architectures follow the trek of parallelization through the development of multi-core processors, leading to effective parallel processing. The overall goal for distributed computational intelligence is to increase the size of the problems already addressed or obtaining higher quality solutions by exploiting the vast computing power rendered by the current supercomputers of parallel-distributed nature.

The second topic is a well established one: the analysis and forecasting of time series, although with a new approach. Time series originate in the Statistics field and, have traditionally been studied with techniques developed within that field. However, researchers in the Computational Intelligence area soon realized that the power their techniques showed in the resolution of hard and non-linear problems could proved equally effective for time series. DiCITS researchers seek to develop new methods to analyze time series based on Computational Intelligence paradigms. Obviously, due to

Figura 1. DiCITS' logo.

the scope of the workshop for which this communication is intended, we focus on this second topic of DiCITS' activity.

This short paper purpose is to provide a current snapshot of the laboratory research lines and interest fields, and some of it most remarkable achievements.

The laboratory was founded in late 2008 within the bosom of a research group with a productive history in the Soft Computing —or Computational Intelligence— field, with specific interest in Machine Learning and Data Mining. Some of the group researchers with backgrounds in different and cooperative CI techniques developed a common interest in approaching time series analysis. The laboratory is still at its initial stages, but evolving at a steady pace. Its staff is growing with the inclusion of new students who engage in pursing Ph.D. degrees in the areas of interest of the laboratory.

The current staff of the laboratory is distributed as follows:

- Ph. D. members:
  - *Head*: José Manuel Benítez Sánchez
  - Antonio Araúzo Azofra
  - José Luis Aznarte Mellado
  - Daniel Molina Cabrera
  - Ana María Sánchez López
- Non-doctors members:
  - Christop Bergmeir
  - Manuel Martín Márquez
  - Manuel Parra Royón
  - Daniel Peralta Cámara
  - Ignacio Robles Paiz

This is reinforced by tight collaborations with researchers in international centers. For example:

- Marcelo C. Medeiros, Department of Economics. Pontifical Catholic University of Rio de Janeiro, Brazil.
- Gregorio I. Saniz, Rubén García at Computer Science Division, CARTIF, Valladolid.

2. Lines of research

In this section we describe the current lines of research at DiCITS in the time series field.

2.1. Cooperation between Computational Intelligence and Statistics

Traditionally, time series analysis has been a field of Statistics. Statisticians have studied time series during the last couple of centuries, and have developed a deep insight of the characteristics of this particular problem. Their formal approach has given birth to a wealth of methods and techniques, as well as theorems and theoretical properties of them.

On the other hand, Computational Intelligence is a relatively new field which obtained great success in classification and regression problems amongst others. Researchers in this field have approached time series as a source of datasets to perform benchmarking and evaluation of new models or algorithms, and have established successful engineering applications which are comparable, in their results, to their statistical counterparts.

In our lab, we have developed a line of research which aims at gathering the advantages of both worlds, the statistical and the CI, in the modeling and forecasting of time series. Equivalence relations have been found between models coming from both disciplines, and this has allowed for a new look over the CI models under the statistical perspective.

Amongst other, some results of this theoretical research are theorems proving the asymptotic stationarity of fuzzy rule-based models (FRBM) [6, 8] and a hypothesis testing framework with applications like linearity tests against an FRBM [9] or an algorithm to determine the number of rules that an FRBM should have for a concrete time series. Also, diagnostic tests have been developed (heteroskedasticity or autocorrelation of the residuals, parameter constancy...) [10]. These results also help to promote the use of some CI paradigms in other fields where they have not received enough attention, for example in Econometrics [5].

An important step in time series modeling, as in may other modeling tasks, is preprocessing, with a leading role being played by *feature selection*. Having a rather deep implications in feature selection [4], we develop a rather effective method to combine different feature selection procedures to time series [14].

2.2. Empirical evaluation of predictors

Effective evaluation of predictors is quite an important issue in time series analyses. This is a rather complex problem for the number of factors involved: measure, forecast horizon, experimental design, or hypothesis tests, to name a few. Actually, no globally accepted consensus for a general methodology is available today. This is not the case in the Machine Learning field, where some general methodology is already in common use. Notwithstanding, there are some more stable proposals [13, 15].

At DiCITS we are concerned with this issue and try to provide some proposals as partial contributions towards a widely accepted empirical methodology. A first step was the in-depth analysis of the correctness and effectiveness of using cross-validation to evaluate the performance of Machine Learning-based predictors [11]. The study involved a set of typical machine learning methods used for one-step-ahead forecasting of a set of real and generated series. Different errors and error measures were used to assess their performance. The results indicate that for standard application scenarios no practical effects of the flaws in cross-validation theory are present, but on the other hand advantages of cross-validation to yield more stable/reliable results still prevail. A solution when using stationary time series might therefore be the use of a blocked version of cross validation, where the sets are not chosen randomly, but in data blocks.

2.3. Time series characterization

In order to accurately analyze and forecast future values for a time series it is very convenient to actually know well the defining properties of the time series. This way a much more effective selection of model and parameter estimation can be carried out. This is a basic well known fact. Some of the basic properties of the series should be identified straightforwardly: seasonality, stationarity, trends, ... This has lead to the tasks of time series segmentation and classification, which remain topics of current interest. A further advance in this line comes from the data complexity measures used in classification problems. A combined set of these values provides further insight of the underlying structure of the data helping to find one of the better classifiers for each specific case. One of the current interest at DiCITS is seeking for data complexity measures for time series which allows us to characterize a given time series and have much information for selection the appropriate predictor.

2.4. Applications

Our expertise in the time series field has been applied in a number of interesting real-world problems. We describe in this section some of them.

The forecasting of the airborne pollen concentration in the city of Granada is one of them . In collaboration with the aerobiology research group of the UGr, we used some CI models to predict the future values of this series [7]. This is of great interest given the increasing number of allergic-related problems.

Another area where time series have an important role is transportation. In particular, we were engaged in the project “Conservación de Infraestructuras Civiles basada en Inteligencia Computacional” (CIBIC, Infrastructure Conservation based on Computational Intelligence). This project was concerned with conservation tasks of different infrastructures. One of them was the railways for the Spanish High Speed Trains (AVE) network. We developed different predictors to forecast the monthly number of defects that appear in some lines of the AVE. The improved accuracy in the forecasts translates into a more effective maintenance policy and a reduction in costs.

The widespread installation of renewable energies in the electrical system has raised the need for prediction of the renewable production, as this is a critical issue for the transmission system operators as well as for the productors, who bid in a futures market. Our group is collaborating with European research centers in the development of CI models tailored to the short-term forecast of the wind and solar production.

As well, determining the ampacity of the lines as a function of the weather is an interesting area in which we are developing operational solutions. The aim is to facilitate a better use of the lines through dynamic rating, as its forecast allows for much efficient use of renewable energies in the control room.

3. Software for time series analysis

A necessary step to design and evaluate new methods for time series analysis is software development. The efforts at DiCITS with regard to this issue have focused on three classes of methods: artificial neural networks, evolutionary algorithms, and threshold autoregressive models. Three are the programming languages that we used: R [17], C++ and Java. A leading guide for our programming efforts is generality and reusability, so that the software is easier to use in later projects and for other researchers and/or practitioners. When there are a set of modules or libraries that we consider of general interest, we package them together and make them available under some kind of open-source license. This way our code is ready for use by other researchers to advance upon our progresses. In this regard, three are the main contributions of DiCITS, which are described in the sequel:

3.1. Artificial neural networks

Neural networks are a frequent tool used for time series forecasting. The best overall neural network simulator is SNNS, "Stuttgart Neural Network Simulator (SNNS)," [19], and the most used programming language by researchers in time series is R. So we decided to bring the functionality of SNNS to the R community. This way the RSNNS package was born [12].

SNNS includes many different learning mechanisms for feed-forward neural networks, such as Backpropagation in different versions, Quickprop, Rprop, and Counterpropagation. For time series research, the implemented time-delay architecture is relevant, as well as the recurrent network types that are present, like Jordan and Elman networks, or recurrent cascade correlation (RCC). Not directly related to time series are the association and clustering algorithms like self-organizing maps (SOM), Hopfield networks, associative memories, adaptive resonance theory networks (ART), and others, that are also part of the software.

We ported the SNNS kernel and some other relevant algorithmic parts of the software to C++, and encapsulated it in a single class to allow for parallel instances, i.e. for parallel use of multiple neural networks. Taking into account the powerful capabilities of R for scripting, parallelization, and visualization, using RSNNS considerably facilitates the inclusion of SNNS-algorithms in state-of-the-art experiment designs and data analysis procedures. So, RSNNS can be considered a comprehensive neural network standard package.

3.2. Threshold autoregressive models

Threshold autoregressive (TAR) models [18] are a Statistics proposal to time series analysis, where the nonlinear model is composed of several linear models. A threshold variable and a corresponding threshold determine, which linear model will be used for prediction in the current situation. If the threshold variable is a past value of the time series, the model is called self-exiting TAR (SETAR). Instead of just switching the regimes in a winner-takes-all manner, smoother transitions can be desirable, which result in the definition of smooth TAR (STAR) models. For modeling the smooth transition, the logistic function, the exponential function, or the Gaussian function are common, defining the respective models LSTAR, ESTAR, and NSTAR. Another extension of the methods is the use of a neural network to compute the threshold, leading to the neuro-coefficient STAR (NCSTAR) [16], which uses the logistic function, or the NCGSTAR, that has Gaussian activation. Orthogonal enhancements are made for the determination of the number of regimes. An iterative procedure based on statistical tests for nonlinearity (see section 2.1) is used to successively add regimes.

The tsDyn package [3] implements a large set of nonlinear models for univariate and multivariate time series. Our group is working on regime switching models for univariate time series, i.e. the implementation of the theory presented in section 2.1. Within the methods implemented in tsDyn are: SETAR and LSTAR models with fixed numbers of regimes (maximal three), and LSTAR, NSTAR, NCSTAR, and NCGSTAR models using the iterative building procedure.

3.3. Evolutionary algorithms

KEEL [1, 2] is a Java software framework for evolutionary computation that is developed within the research group SCI2S. We are working towards its use for time series forecasting, and its interaction with the R programming language. In particular, we are combining Statistics procedures with metaheuristics for model identification and parameter estimation.

In addition to this we are extending the KEEL Dataset repository with time series datasets (`http://sci2s.ugr.es/keel/datasets.php`). This is an on-going project where datasets freely available from the Internet are made available in a convenient point. We also include synthetic time series used in our empirical studies. In addition to original data, extra content is provided: partitioned data in several formats, data description, results for some predictors, . . .

4. Conclusions

The DiCITS lab is composed by a group of researchers interested in time series analysis and forecasts. In this communication, its main lines of research are described. Furthermore a brief overview of the software packages developed and released is offered.

Acknowledgments

The elaboration of this communication has been partially supported by the Spanish Ministerio de Ciencia e Innovación under grant TIN2009-14575.

References

1. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing 17(2-3), 255–287 (2011)

2. Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernández, J., Herrera, F.: Keel: A software tool to assess evolutionary algorithms for data mining problems. Soft Computing 13(3), 307–318 (2009)
3. Antonio, Narzo, F.D., Aznarte, J.L., Stigler, M.: tsDyn: Time series analysis based on dynamical systems theory (2009), r package version 0.7
4. Arauzo, A.: Un sistema inteligente para selección de características en clasificación. Ph.D. thesis, Universidad de Granada (2006)
5. Aznarte, J., Alcalá-Fdez, J., Arauzo, A., Benítez, J.: Fuzzy autorgressive rules: towards linguistic time series modelling. Econometric Reviews 30, 609–631 (2011)
6. Aznarte, J., Benítez, J., Castro, J.: Smooth transition autoregressive models and fuzzy rule-based systems: Functional equivalence and consequences. Fuzzy Sets and Systems 158(24), 2734–2745 (2007)
7. Aznarte, J., Benítez, J., Nieto-Lugilde, D., de Linares Fernández, C., de la Guardia, C., Sánchez, F.: Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. Expert Systems with Applications 32(4), 1218–1225 (2007)
8. Aznarte, J., Benítez, J.: Equivalences between neural-autoregressive time series models and fuzzy systems. IEEE Trans. Neural Networks 21(9), 1434–1445 (2010)
9. Aznarte, J., Medeiros, M., Benítez, J.: Linearity testing against a fuzzy rule-based model. Fuzzy Sets & Systems (2010), (en prensa)
10. Aznarte, J., Molina, D., Sánchez, A., Benítez, J.: A test for homoscedasticity of the residuals in fuzzy rule-based forecasters. Applied Intelligence (In press) (2011)
11. Bergmeir, C., Benítez, J.: On the use of cross-validation for time series predictor evaluation. Information Sciences (submitted) (2010)
12. Bergmeir, C., Benítez, J.M.: Neural Networks in R using the Stuttgart Neural Network Simulator: RSNNS (2010), R package version 0.3-1
13. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
14. García, R., Sainz, G.I., Benítez, J.M.: Frasel: Aggregation of feature ranking methods for time series modelling. Information Sciences (Sometido) (2011)
15. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180, 2044–2064 (2010)
16. Medeiros, M., Veiga, A.: A flexible coefficient smooth transition time series model. IEEE Transactions on Neural Networks 16(1), 97–113 (2005)
17. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009), ISBN 3-900051-07-0
18. Tong, H.: Non-linear time series: a dynamical system approach. Oxford University Press, Oxford, UK (1990)
19. Zell, A. et al.: SNNS Stuttgart Neural Network Simulator User Manual, Version 4.2. IPVR, University of Stuttgart and WSI, University of Tübingen (1998)