An study on data mining methods for short-term forecasting of the extra virgin olive oil price in the Spanish market

P. Pérez[1] M. P. Frías[2] M. D. Pérez-Godoy[1] A. J. Rivera[1] M. J. del Jesus[1]
M. Parras[3] F. J. Torres[3]
[1]Department of Computer Science. University of Jaén. Spain
[2]Department of Statistics and Operations Research. University of Jaén. Spain
[3]Department of Marketing. University of Jaén. Spain

Abstract

This paper presents the adaptation of an evolutionary cooperative competitive RBFN learning algorithm, CO^{2}RBFN, for short-term forecasting of extra virgin olive oil price. The olive oil time series has been analyzed with a new evolutionary proposal for the design of RBFNs, CO^{2}RBFN. Results obtained has been compared with ARIMA models and other dada mining methods such as a fuzzy system developed with a GA-P algorithm, a multilayer perceptron trained with a conjugate gradient algorithm and a radial basis function network trained with a LMS algorithm. The experimentation shows the high efficacy reached for the applied methods, specially for data mining methods which have slightly outperformed ARIMA methodology.

1 Introduction

Olive oil has become an important business sector in a continuously expanding market. Spain is the first olive oil producing country in the world and Jaén its most productive province.

The official market for the negotiation of futures contracts on olive oil (http://www.mfao.es) is a society with the object of negotiating an adequate price for the olive oil at the moment that it is sold (and so paid) in a fixed time in the future. The agents involved in this sector are interested in the use of forecasting methods for the olive price. The managed data are the weekly extra-virgin olive oil price contributed by *Poolred*, http://www.oliva.net/poolred/, an initiative of the Foundation for the Promotion and Development of Olive and Olive Oil located in Jaén, Spain.

The managed data are a set of regular time-ordered observations of a quantitative characteristic of an individual phenomenon taken at successive periods or points of time, called time series. Time series forecast is an active research area and a typical paradigm to address it are ARIMA models introduced by Box and Jenkins [2] at the beginning of the seventies. Data mining, a research area to extract non trivial information contained in a database, has been also applied to time series forecasting. The prediction ability of data mining methods, as Neural Networks or Fuzzy Rule Based Systems [5, 11, 9, 4], is based on its universal functional approximation characteristics.

In this paper, CO^{2}RBFN, an evolutionary cooperative competitive learning method for Radial Basis Function neural Networks (RBFNs) is extended to address time series forecasting . The results obtained will be also compared with ARIMA methodology and others hybrid intelligent systems methods such as a Fuzzy System developed by a GA-P algorithm, a MultiLayer Perceptron Network trained with a Conjugate Gradient learning algorithm and a classical design method for Radial Basis Function Network learning.

To do so, this paper is organized as follow: in Section 2, the extension of CO^{2}RBFN for time series forecasting is presented. The study and results obtained for the forecast methods are detailed in Section 3. In Section 4, the conclusions and future work are outlined.

2 CO^{2}RBFN for time series forecasting

RBFNs [3] are an important Artificial Neural Network paradigm with interesting characteristics such as a simple topological structure or universal approximation ability [7]. They have been successfully used in time series prediction [12][11].

From a structural point of view, an RBFN is a feed-forward neural network with three layers: an input layer with n nodes, a hidden layer with m neurons or RBFs, and an output layer with one or several nodes. The m neurons of the hidden layer are activated by a radially-symmetric basis

```
1.  Initialize RBFN
2.  Train RBFN
3.  Evaluate RBFs
4.  Apply operators to RBFs
5.  Substitute the eliminated RBFs
6.  Select the best RBFs
7.  If the stop condition is not
    verified go to step 2
```

Figure 1. Main steps of Co^{2}RBFN.

function, $\phi_i : R^n \rightarrow R$, which can be defined in several ways, being the Gaussian function the most widely used.

The authors developed a hybrid cooperative competitive evolutionary proposal for RBFN design, CO^{2}RBFN, applied to the classification problem [8]. In this paper, a new version of CO^{2}RBFN is presented in order to deal with the time series forecasting problem. This new version also improves the efficiency and the balance between exploration-explotation in the evolutionary design process.

In this approach each individual of the population represents a basis function and the entire population is responsible for the final solution. The individuals cooperate towards a definitive solution but they must also compete for survival.

The credit apportionment of a given RBF is evaluated taking into account three factors: the RBF contribution to the network output, the error in the basis function radius, and the degree of overlapping among RBFs. The application of the operators is determined by a fuzzy rule-based system. The inputs of this system are the three parameters used for credit assignment and the outputs are the operators' application probability. To design the set of rules we take into account the fact that an RBF is worse if its contribution is low, its error is high and its overlapping is also high, otherwise an RBF is better. In this way the probability of eliminating an RBF is high when this RBF is worse and so on. The main steps of the algorithm are shown in figure1.

In this work, and for addressing short-term forecasting problems an extension of our previous version [8] has been developed with the following key differences: In the training phase only the weights of the net are learned instead of our previous version, where weights and RBF parameters were both trained. A new operator, *biased mutation*, has been introduced to complement the existent operators *remove, random mutation and null*. The goal of the biased mutation operator is to modify the parameters of the RBF by using local environment information. In the step 6, a new random method to introduce new RBFs has been added. For the forecasting task the number of outputs of the RBFN have set to one. The error to be minimized, defined as a parameter of the credit apportionment, is the mean absolute percentage error (MAPE) error.

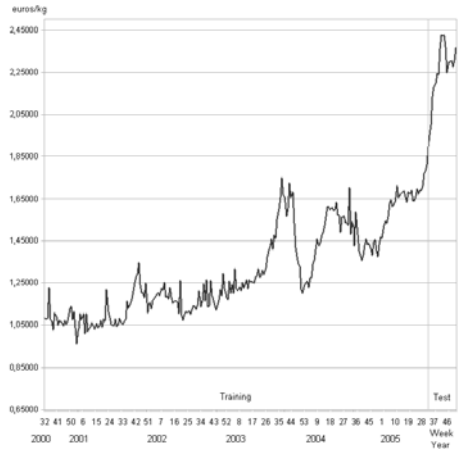

Figure 2. Weekly extra-virgin olive oil prices in Spain

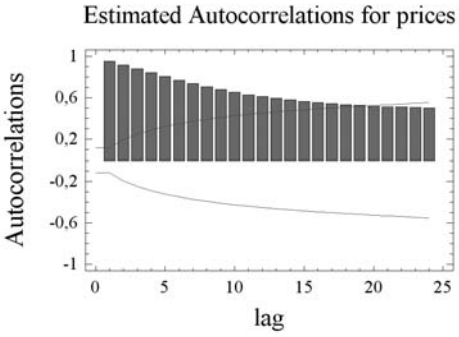

Figure 3. ACF for olive oil price series.

3 Experimentation and Results

The addressed problem is to carry out a short-term forecasting (next week) of the extra-virgin olive oil price. In this study, the data used are from the 32^{th} week (August) of 2000 year to the 52^{th} week (December) of 2005 year in Spain. The cases in the data set were divided into two subsets: one for training and the other one for testing. The data from the 32^{th} week of 2000 year to the 32^{th} week of 2005 year were used to train. Data from the 33^{th} week to the 52^{th} week of 2005 year are used to test the methods. Figure 2 shows the time series data and training and test datasets.

The error measures considered to evaluate the performance of the experiments are mean squared error (MSE $=\sum_i (d_i - y_i)^2/n$) and mean absolute percentage error (MAPE = $\sum_i (\mid (d_i - y_i)/d_i \mid)/n$), where d_i is the real value, y_i is the predicted value and n is the number of data.

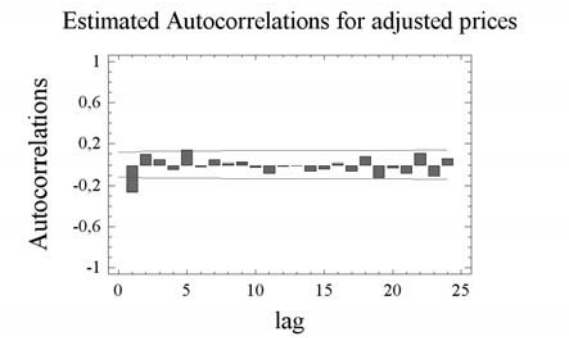

Figure 4. ACF for the differenced olive oil price series

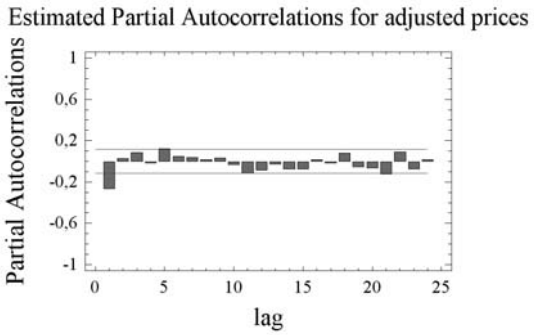

Figure 5. PACF for the differenced olive oil price series

In order to analyze the results obtained for CO^{2}RBFN, a comparison is performed with ARIMA methodology [2] and other data mining regression methods. To select the last ones a preliminary experimentation has been done with Keel tool [1]. From all the data mining regression methods included in this tool, the best results have been reached by Fuzzy GA-P [10], Multi-Layer Perceptron Conjugate Gradient [6] and Radial Basis Function Least Mean Square [3] methods.

A preliminary analysis of the weekly extra-virgin olive oil prices, portrayed in figure 2 shows a non-stationary series, because prices tend to increase over time. This inherent non-stationarity is also confirmed by the graph in figure 3, where the sample ACF (Autocorrelation Function) slowly attenuates for the given extra-virgin olive oil price series. Once the series is differenced, the non-stationarity is removed as it is shown by the sample ACF in figure 4. Based upon the identification information provided by the ACF and the PACF (Partial Autocorrelation Function), figures 4 and 5, of the differenced series, the weekly extra-virgin olive oil price series could be modeled by an ARIMA(0,1,1) or ARIMA(0,1,5).

For the selection of the features to include in the dataset,

Method	Test MAPE	Training MAPE	Total MAPE
ARIMA (0, 1, 5)	2,67	3,22	3,20
ARIMA (0, 1, 1)	3,32	3,24	3,24
Fuzzy GA-P	1,93	3,03	2,95
MLP-ConjGrad	2,36	2,56	2,54
RBFN-LMS	2,95	3,17	3,18
CO^{2}RBFN	1,94	2,66	2,67

Table 1. MAPE error for methods and datasets

Method	Test MSE	Training MSE	Total MSE
ARIMA (0, 1, 5)	0,00595	0,00254	0,00268
ARIMA (0, 1, 1)	0,01042	0,00265	0,00297
Fuzzy GAP	0,00308	0,00266	0,00269
MLP-ConjGrad	0,00408	0,00195	0,00210
RBFN-LMS	0,00752	0,00283	0,00317
CO^{2}RBFN	0,00393	0,00216	0,00229

Table 2. MSE error for methods and datasets

we have chosen the classical design of the patterns $(n+1, n, n-1, n-2, ...)$ where $(n+1)$ is the price to forecast and $n, n-1, n-2, ...$ and are past prices. This typical design has been used in different works [12, 4] and it allows to compare with the ARIMA methodology. Several experiments have been carry out with patterns composed by 3, 4, 5, 6, 7, 10 and 20 past prices and the best results for all the methods have been reached for the dataset with patterns with 5 past prices.

ARIMA methodology shows that the addressed time series data have a positive trend, and therefore it is convenient to differentiate such data in order to achieve a stationary series where training and test datasets values are in the same range.

In the experimentation, the number or neurons for CO^{2}RBFN (and so, the population size) is 9. For the methods used for the comparison, the parameters are the ones recommended by the authors.

3.1 Analysis of the results

Results for the all the methods considered are shown in tables 1 and 2. The graphical behavior in test dataset is shown in figure 6.

A first sight of the time series graph, figure 2, shows a non-trivial data series with a different final behavior to the initial phase. Despite this fact, ARIMA method and data mining methods have reached good results both in training and test datasets. Methods can be sorted, having into

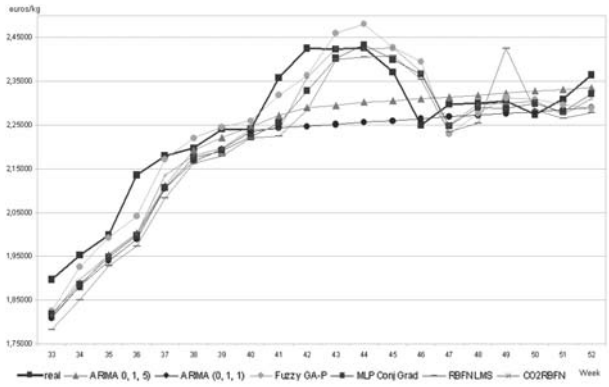

Figure 6. Olive oil price and predicted values

account their test error results, as follow: Fuzzy GA-P, CO^2RBFN, MLP-ConjGrad, ARIMA(0,1,5), RBFN-LMS, ARIMA(0,1,1). The best results have been obtained for data mining methods, concretely CO^2RBFN and Fuzzy GA-P algorithms obtain a similar MAPE error, but Fuzzy-GAP outperforms CO^2RBFN in MSE error. However CO^2RBFN outperforms Fuzzy-GAP in training and total error. From the behavior shown in figure 6 we can conclude that data mining methods follow the shape of the original data. ARIMA models are not able to fit this shape, mainly in the "hill" that test dataset shows in the middle of the graph. This fact is due to ARIMA models only have MA (moving average) component but do no have AR (autoregressive component), which implies a low efficiency for these kind of shapes.

4 Concluding remarks

The goal of our research has been to carry out a short-term forecasting of the weekly extra virgin olive oil price. To do so, an evolutionary cooperative competitive algorithm for RBFN learning, CO^2RBFN, has been adapted to address forecasting tasks. Firstly, olive oil price time series has been analyzed by means of ARIMA methodology. This analysis has demonstrated the non-stationarity of time series and the convenience of differentiate it. CO^2RBFN results has been compared with the forecasting models obtained by ARIMA and with others data mining methods. Both, data mining and ARIMA methods have obtained good results, specially data mining methods as Fuzzy GA-P or CO^2RBFN algorithms. As future work, exogenous features such as meteorology or econometrics facts can be taken into account in order to increase the performance of the forecast. Long-term predictions of the price olive oil will be also addressed.

Acknowledgments: *Supported by the Spanish Ministry of Education and Science under project TIN-2005-08386-C05-03 and the Andalusian Research Plan under project P05-TIC-00531*

References

[1] J. Alcalá, L. Sánchez, S. García, M.J. Del Jesus, S. Ventura, J.M. Garrell, J.Otero, C. Romero, J. Bacardit, V.R. Rivas, J.C. Fernández, F. Herrera. KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems. Soft Computing. In press, 2008.

[2] G. Box ,G. Jenkins. Time series analysis: forecasting and control. San Francisco: Holden Day, 1976.

[3] D. Broomhead, D. Lowe. Multivariable functional interpolation and adaptive networks. Complex System. 2:321-355, 1998.

[4] H.C. Co, R. Boosarawongse. Forecasting Thailand's rice export: Statistical techniques vs. artificial neural networks. Computers and Industrial Engineering,53(4):610-627, 2007.

[5] M. Khashei, S. Reza, M. Bijari. A new hybrid artificial neural networks and fuzzy regression model for time series forecasting. Fuzzy Sets and Systems 159 (7): 769-786, 2008.

[6] F. Moller. A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw. 6:525-533, 1990.

[7] J. Park, I. Sandberg. Universal approximation using radial-basis function networks. Neural Comput., 3:246-257,1991.

[8] M.D. Pérez-Godoy, A.J. Rivera, M.J. Jesús,I. Rojas. CoEvRBFN: an approach to solving the classification problem with a hybrid cooperative-coevolutive algorithm. LNCS 4507:324-332, 2007.

[9] R. Pino, J. Parreno, A. Gomez, P. Priore. Forecasting next-day price of electricity in the Spanish energy market using artificial neural networks. Engineering Applic. of Artificial Intelligence, 21(1):53-62, 2008.

[10] L. Sánchez, I. Couso. Fuzzy Random Variables-Based Modeling with GA-P Algorithms. Information, Uncertainty and Fusion, 245-256, 2000.

[11] M. Ture, I. Kurt. Comparison of four different time series methods to forecast hepatitis A virus infection. Expert Systems with Applications, 31(1):41-46, 2006.

[12] B. Whitehead, T. Choate. Cooperative-competitive genetic evolution of radial basis function centers and widths for time series prediction. IEEE Trans. on Neural Networks, 7(4): 869-880, 1996.