Proceedings

ITISE 2015

International work-conference on Time Series

Granada

July, 1-3 2015

Proceedings ITISE 2015.

International work-conference on Time Series

Editors and Chairs:
Olga Valenzuela
Fernando Rojas
Hector Pomares
Ignacio Rojas

ISBN: 978-84-16292-20-2
Deposito Legal: Gr-891/2015

Edita e imprime: Copicentro Granada S.L

An ensemble strategy for forecasting the extra-virgin olive oil price in Spain

A. J. Rivera[1], M. D. Pérez-Godoy[1], F. Charte[2], F. J. Pulgar[1] and M. J. del Jesus[1]

[1] Dept. of Computer Science
University of Jaén. Spain
(arivera,lperez,fpulgar,mjjesus)@ujaen.es
[2] Dept. of Computer Science and Artificial Inteligence
University of Granada. Spain
francisco@fcharte.com

Abstract. Time series prediction is one of the key tasks in data mining, especially in areas such as science, engineering and business. It is possible to distinguish between fundamental analysis and technical analysis while dealing with time series in the business area. Fundamental analysis takes into account different exogenous variables such as expenses, assets or liabilities. Technical analysis summarizes information using technical indicators such as momentums, moving averages or oscillators. The most influential exogenous variables and technical indicators for the olive oil price have been already identified in previous studies. The objective of the present paper is to propose an ensemble strategy, based on dividing this set of exogenous variables and technical indicators into subsets of features for the base models. These base models use CO^{2}RBFN, a cooperative competitive algorithm for RBFNs, as learning algorithm. The obtained results show that the ensemble strategy outperforms both the base models and other classical soft computing methods.

Keywords: Ensemble, Radial Basis Function Networks, Forecasting, Time Series, Olive oil price

1 Introduction

Olive oil is an important business sector around the world. Spain is the first olive oil producing country, and Jaén is its most productive province. The agents involved in this sector are interested in the use of forecasting methods for the olive oil price. In this context, an accurate prediction of this price in the future could increase the global benefits.

In Spain there are different organizations implied in the monitoring, analysis and study of the olive oil sector, such as *Poolred* [3], an initiative of the Foundation for the Promotion and Development of the Olive and Olive Oil located in Jaén, or the Agency for the Olive Oil, promoted by the Ministry of Agriculture, Food

[3] http://www.oliva.net/poolred/

and Nature [4]. The data obtained from these institutions are time series, that is, sets of regular, time-ordered observations of a quantitative characteristic of an individual phenomenon. These observations are taken at successive periods or points in time. The problems in which the data is not independent, but also have a temporal relationship, are called time series forecasting problems.

Traditionally, statistics/econometrics methods [4] were the only ones used to deal with time series forecasting problems. However, in recent years soft computing methods [3] [21] have achieved accurate solutions. Ensemble technologies have been very common and have reached good results [22] [8] in the soft computing field.

In [15] we made a deep study about the influence of differences external variables and technical indicators over olive oil price forecasting. Classical soft computing methods forecast olive oil price time series using the most relevant characteristics, chosen by feature selection methods.

Taking into account the previous issues, in this paper we propose an ensemble strategy to forecast the extra-virgin olive oil price. Each individual model in the ensemble works over a subset composed by different external variables and technical indicators, those which have greater influence according to the carried out study. These base models use CO^{2}RBFN, a cooperative competitive algorithm for Radial Basis Function Networks (RBFNs), as learning algorithm. The obtained results show that ensemble strategies outperform both the base method and other typical soft computing methods.

This paper is organized as follows: Section 2 shows a brief description about time series and their forecast. The ensemble methodology is described in Section 3. CO^{2}RBFN, the learning algorithm used in any model in the ensemble, is outlined in Section 4. In Section 5 the ensemble proposed is explained. The experimental framework and the obtained results are provided in Section 6. Finally, the conclusions appear in Section 7.

2 Time Series

A time series is a set of regular, time-ordered observations of a quantitative characteristic related to an individual or collective phenomenon. The observations are taken at successive and, in most cases, equidistant periods of time. The goal of any basic forecasting method is to predict an outcome from a set of past values.

In general, the ultimate goal always is increasing our knowledge on a phenomenon or aspect of our environment data from past and present. Therefore, the main aim is to extract the regularities observed in the past behavior of the variable, i.e. obtain the mechanism that generates it, in order to have a better understanding of it in the future.

As mentioned, the importance of time series analysis and forecasting has grown in science, engineering and business. Besides the statistics/econometrics

[4] http://aplicaciones.magrama.es/pwAgenciaAO/General.aao

methods [4], soft computing techniques [11] [3] have been developed in order to solve the time series forecasting problem. For example, [17] combines several approaches of soft computing, as Radial Basis Functions Networks (RBFNs), to research financial market efficiency, and [10] shows the application of soft computing techniques in prediction of an occupant's behavior in an inhabited intelligent environment. Usually, the results obtained by soft computing methods are similar, or even better, than those from traditional statistics models.

Concerning to the financial time series area [20], a fundamental and a technical analysis can be distinguished. Fundamental analysis involves delving into the financial statements by examining related economic and company-specific information. This involves looking at revenue, expenses, assets, liabilities, and all the other financial aspects of an organization. On the other hand, technical analysis takes a completely different approach, as it does not care about the intrinsic values of an organization. Technical analysis [13] [1] (sometimes called chartists) is only interested in the price movements of the market, identifying patterns and using them in order to predict future prices. Some example indicators used for technical analysis, known as technical indicators, are momentums, moving averages, oscillators, convergences-divergences, etc.

Nevertheless, a future value is usually influenced by some external (or exogenous) information (as harvest season when one tries to estimate the price of the product). In this case the external information can be incorporated into the model, usually as commonly-known exogenous variables. This external information can also be summarized by technical indicators [6].

In this paper some exogenous variables, such as stocks, exports or price indexes, and different technical indicators are used for forecasting the extra-virgin olive oil price.

3 Ensembles

An Ensemble [16] [23] is a set of learning (base/expert) methods whose outputs are combined in some way (typically by weighted or unweighted voting) to classify new examples. The main advantage of ensemble methods is that they achieve more accurate predictions than the individual base methods. The fundamentals of this success can be summarized as follows. The expert knowledge learned from the base methods can be combined. Complex problems can be decomposed into multiple subproblems, which are easier to understand and solve (divide-and-conquer approach). There is not a classical method that works for all problems (not free lunch theorem). According to [9] *"A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse"*. The authors consider two classifiers as diverse if they make different errors on new data points. There are several methods [5] for constructing ensembles:

– Subsampling the training examples: where base methods are trained on different data sets obtained by resampling a common training set (Bagging, Boosting).

- Manipulating the input features: base classifiers are trained on different representations, or different subsets of a common feature vector.
- Manipulating the output targets: the output labels of the classes are encoded into groups of different label-sets. A base classifier is trained to predict each one of these label-sets.
- Modifying the learning parameters of the classifier: different base classifiers are developed with disparate learning parameters, such as number of neighbors in a k Nearest Neighbor rule, number of neurons for a neural network, etc.

Aiming to obtain a unique output prediction from the ensemble of classifiers, there are different combination strategies that also depend on the constructing methodology. Some typical combination strategies are:

- Voting: each base method produces or votes a single class/prediction and the class with the majority vote on the ensemble wins.
- Averaging: each base method produces a confidence estimate, the winner is the class/prediction with the highest average value.
- Weighted averaging: the output of each base method is weighted according to its own performance over the training set.

4 Base Algorithm: CO^{2}RBFN

CO^{2}RBFN [14] is an evolutionary cooperative-competitive hybrid algorithm for designing RBFNs. In this algorithm each individual of the population represents, with a real representation, an RBF, and the entire population is responsible for the final solution. The individuals cooperate towards a definitive solution, but they must also compete for survival. In this environment, in which the solution depends on the behavior of many components, the fitness of each individual is known as its "credit assignment". In order to measure the credit assignment of an individual, three factors were proposed: the RBF contribution to the network output, the error in the basis function radius, and the degree of overlapping among RBFs.

There are four evolutionary operators that can be applied to an RBF: an operator that eliminates the RBF, two operators that mutate the RBF, and finally, an operator that maintains the RBF parameters in order to explore and exploit the search space and to preserve the best RBF, respectively.

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.

The main steps of CO^{2}RBFN are shown in the pseudocode 1.

5 Proposed ensemble strategy

The goal of the present paper is to build an ensemble in order to forecast the extra-virgin olive oil price time series. Each individual in the model deals with a

Algorithm 1 Main steps of CO^2RBFN

```
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
```

different set of input exogenous variables and technical indicators, those with the most influence over the time series. The whole sets of exogenous variables and technical indicators are chosen according to the paper [15], where a deep study about the influence of these variables and technical indicators over the olive oil price forecasting was made. CO^2RBFN is used as base learning algorithm for any individual model.

Following, the time series and the variables are presented and the ensemble is described.

5.1 Time series: exogenous variables and technical indicators

The time series contains the monthly extra-virgin olive oil price per ton at target from the 1st month of 2003 to the 12th moth of 2012 (see Figure 1). This time series has been obtained from the Ministry of Economy and Competitiveness http://www.mineco.gob.es, and the objective is to predict olive oil price per ton at target on a six-month horizon.

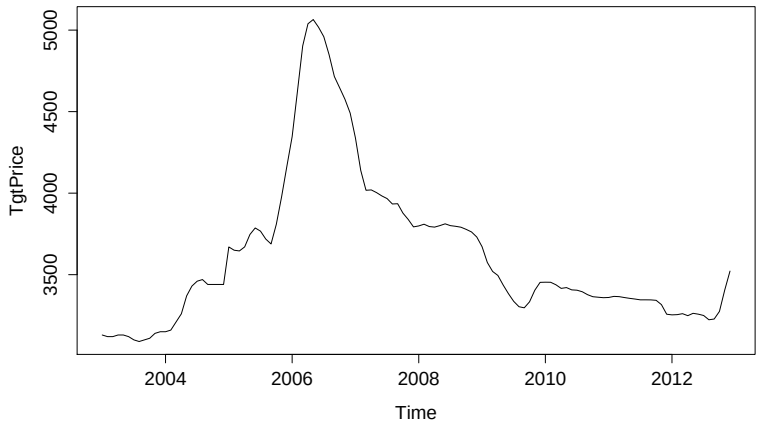

Fig. 1. Time series of the extra-virgin olive oil price

From the paper [15], the most influential variables and technical indicators in the extra-virgin olive oil price have been chosen to build the base ensemble models (Tables 1 and 2). In Table 2, i_t is the value of the index at time t, H_{t-k} and L_{t-k} are the highest and lowest values during a period of time k, respectively, and H_n and L_n are the highest and lowest values from the beginning of the time series, respectively.

Table 1. Exogenous variables used to forecast the olive oil price

Variable	Description	Source
Exports	Exports of olive oil	Agency for the Olive Oil
FoodCPI	Food Consumer Price Index	National Institute of Statistics
GenCPI	General Consumer Price Index	National Institute
Stock	Stocks of olive oil	Agency for the Olive Oil
TgtPrice	Target Price of the extra-virgin olive oil	Poolred

Table 2. Technical indicators and their formulas

Technical indicator	Description	Formula
Momentum 3	Measures the change of an index over a time span of three months	$i_t - i_{t-3}$
Momentum 6	Measures the change of an index over a time span of six months	$i_t - i_{t-6}$
Williams %R	Larry William's %R. A momentum indicator that measures overbought/oversold levels	$\frac{H_n - i_t}{H_n - L_n} \times 100$

5.2 Ensemble strategy

From the different alternatives for building ensembles, detailed in the Section 3, the one following the manipulation of input features methodology has been chosen. Specifically, we propose to build an ensemble composed by four models which deal with different subsets of input variables, exogenous variables and technical indicators, always using CO^2RBFN as base learning algorithm, since it is the most accurate algorithm in [15]. In Table 3 the input variables for the different models are shown. In all models the output variable is the six months ahead price of the extra-virgin olive oil. This division of the whole set of input

features has been heuristically built based on maintaining the presence of the whole set of exogenous variables in any model and promoting that each model was an expert on one technical indicator.

Table 3. Models of the Ensemble

Variable	Model 1	Model 2	Model 3	Model 4
Exports	X			
FoodCPI	X			
GenCPI	X			
Stock	X			
TgtPrice	X	X	X	X
Exports Momentum 3		X		
FoodCPI Momentum 3		X		
GenCPI Momentum 3		X		
Stock Momentum 3		X		
TgtPrice Momentum 3		X		
Exports Momentum 6			X	
FoodCPI Momentum 6			X	
GenCPI Momentum 6			X	
Stock Momentum 6			X	
TgtPrice Momentum 6			X	
Exports Williams				X
FoodCPI Williams				X
GenCPI Williams				X
Stock Williams				X
TgtPrice Williams				X

In order to combine the output of the base methods and taking into account that we address a prediction problem, the more typical strategies, which are weighted average and average, have been chosen.

6 Experimental framework

In order to conduct the forecasting task, the data partitions have been done using the rolling-window technique [19].

In the rolling-window method, forecasts for a fixed horizon are performed by sequentially moving values from the test set to the training set, and changing the forecast origin accordingly. The amount of data used for training is kept constant, so that as new data is available, old data from the beginning of the series are discarded. This method mitigates the influence of data from the distant past.

The four partitions obtained can be seen in Table 4. In the first partition the data from January 2003 to December 2008 were used for training and the data from January 2009 to December 2009 were used for testing. In the second

partition the year 2003 was discarded and the training partition comprise from January 2004 to December 2009, whereas the test partition is set from January 2010 to December 2010, and so on.

Table 4. Data sets

Data set	Training years	Test years
Test2009	2003 2004 2005 2006 2007 2008	2009
Test2010	2004 2005 2006 2007 2008 2009	2010
Test2011	2005 2006 2007 2008 2009 2010	2011
Test2012	2006 2007 2008 2009 2010 2011	2012

6.1 Results and analysis

To estimate the forecast accuracy of the methods, the Mean Absolute Percentage Error (MAPE) has been considered (1):

$$MAPE = \sum_{i}^{z}(\mid 100(f(x_i) - y(x_i)/y(x_i) \mid)/z, \tag{1}$$

where $f(x_i)$ is the predicted output of the model, $y(x_i)$ is the desired output and z is the size of the test set.

CO^{2}RBFN has been run with its default configuration: 10 neurons and the number of generations is set to 200. It has been executed 5 times for each partition.

Firstly, in Table 5 the MAPE mean test results per partition of the two ensemble strategies and the four base models are shown. As can be observed, ensemble strategies not only achieve best results for individual partitions, but also in average where the two ensemble strategies outperforms any base model.

Table 5. Mape test: Ensemble vs Individual Models

Year	Model 1	Model 2	Model 3	Model 4	Ensemble weighted average	Ensemble average
2009	4.464	4.925	4.960	4.003	3.957	3.932
2010	2.518	2.831	3.884	2.451	2.101	2.102
2011	2.230	2.904	3.181	2.496	2.338	2.338
2012	6.634	6.515	6.851	6.417	6.379	6.379
Mean	3.961	4.294	4.719	3.842	3.694	3.688

After that, in Table 6 the results of the ensemble strategies are compared with classical soft computing methods using the whole set of input variables: the base method, CO^{2}RBFN, and other typical soft computing methods, such

as a Fuzzy System developed with a GA-P algorithm (GFS-GAP) [18], a MultiLayer Perceptron Network trained using a Conjugate Gradient learning algorithm (MLP-CG) [12], and NU-SVM [7], a Super Vector Machine based method. The implementation of the other methods has been obtained from KEEL [2]. The main parameters used are set to the values indicated by the authors. Firstly, from the obtained results can be observed that CO^{2}RBFN outperforms the other soft computing methods both in average and for almost all the individual partitions.

Table 6. Mape test: Ensemble versus Model with all variables

Year	CO^{2}RBFN	GFS-GAP	MLP-CG	NU-SVM	Ensemble weighted average	Ensemble average
2009	5.013	5.134	14.406	8.673	3.957	3.932
2010	2.402	4.595	8.972	3.065	2.101	2.102
2011	2.650	2.537	3.609	3.802	2.338	2.338
2012	7.276	6.721	8.446	7.649	6.379	6.379
Mean	4.335	4.747	8.858	5.797	3.694	3.688

Furthermore, it must be highlighted that two ensemble methodologies, composed with different sets of input variables, outperform CO^{2}RBFN, the base method, with the whole set of input variables, both for any partition and for the total average. This fact confirms some grounds of the ensemble methodologies, such as the successful combination of the expert knowledge or the divide and conquer principle, explained in section 3.

7 Conclusions

In the business time series area is typical to apply procedures as fundamental analysis and technical analysis. While fundamental analysis promotes the use of economical exogenous variables in order to forecast the output variable, technical analysis proposes a set of technical indicators for the same goal.

In a previous work, the authors of the present paper have identified the most influential exogenous variables and technical indicators for the extra-virgin olive oil price forecasting by means of feature selection algorithms. Moreover, classical soft computing methods were trained to achieve an accurate prediction of the olive oil price.

Aiming to improve those predictions, this paper proposes an ensemble strategy based on dividing the whole set of input features into subsets of features based on technical indicators.

The obtained results show that the ensemble strategy outperforms both the individual models as well as the classical soft computing methods, including the base learning method, CO^{2}RBFN, with the whole set of input features.

Acknowledgments: F. Charte is supported by the Spanish Ministry of Education under the F.P.U. National Program (Ref. AP2010-0068). This paper is

partially supported by the Spanish Ministry of Science and Technology under the Project TIN 2012-33856, FEDER founds.

References

1. S. Achelis. *Technical Analysis from A to Z, 2nd ed.* McGraw-Hill, 2000.
2. J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. of Mult.-Valued Logic & Soft Computing*, 17:255–287, 2011.
3. G.S. Atsalakis and K.P. Valavanis. Surveying stock market forecasting techniques - part ii: Soft computing methods. *Expert Systems with Applications*, 36(3, Part 2):5932–5941, 2009.
4. G. Box, G. Jenkins, and G. Reinsel. *Time series analysis: forecasting and control,4th Edition.* Wiley, 2008.
5. T. G. Dietterich. Ensemble methods in machine learning. *LNCS*, 1857:1–15, 2000.
6. H. Dourra and P. Siy. Investment using technical analysis and fuzzy logic. *Fuzzy Sets Syst.*, 127(2):221–240, 2002.
7. R. E. Fan, P. H. Chen, and C. J. Lin. Working set selection using the second order information for training svm. *Journal of Machine Learning Research*, 6:1889–1918, 2005.
8. A. Grigorievskiy, Y. Miche, A. M. Ventelä, E. Séverin, and A. Lendasse. Long-term time series prediction using OP-ELM. *Neural Networks*, 51(0):50–56, 2014.
9. L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.
10. S. Mahmoud, A. Lotfi, and C. Langensiepen. Behavioural pattern identification and prediction in intelligent environments. *Applied Soft Computing*, 13(4):1813–1822, 2013.
11. A. Mochón, D. Quintana, Y. Sáez, and P. Isasi. Soft computing techniques applied to finance. *Applied Intelligence*, 29(2):111–115, 2008.
12. F. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6:525–533, 1990.
13. J.J. Murphy. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications.* New York Institute of Finance, 1999.
14. M. D. Pérez-Godoy, A. J. Rivera, M. J. del Jesus, and F. J. Berlanga. CO^2RBFN: An evolutionary cooperative-competitive RBFN design algorithm for classification problems. *Soft Computing*, 14(9):953–971, 2010.
15. A. J. Rivera, P. Pérez-Recuerda, M. D. Pérez-Godoy, M. J. del Jesus, M. P. Frías, and M. Parras. A study on the medium-term forecasting using exogenous variable selection of the extra-virgin olive oil with soft computing methods. *Applied Intelligent*, 34(3):331–346, 2011.
16. L. Rokach. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics and Data Analysis*, 53(12):4046–4072, 2009.
17. V. Sakalauskas and D. Kriksciuniene. Tracing of stock market long term trend by information efficiency measures. *Neurocomputing*, 109:105–113, 2013.
18. L. Sánchez and I. Couso. Fuzzy random variables-based modeling with ga-p algorithms. *In: B. Bouchon, R.R. Yager, L. Zadeh (Eds.) Information, Uncertainty and Fusion*, pages 245–256, 2000.

19. L. J. Thasman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.
20. R. S. Tsay. *Analysis of Financial Time Series, 3ed*. 2010.
21. B. Vanstone and T. Hahn. *Designing Stock Market Trading Systems: with and without Soft Computing*. Harriman House, 2010.
22. W. Yan. Toward automatic time-series forecasting using neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1028–1039, 2012.
23. J. Yang, X. Zeng, S. Zhong, and S. Wu. Effective neural network ensemble approach for improving generalization performance. *IEEE Transactions on Neural Networks and Learning Systems*, 24(6):878–887, 2013.