

Available online at www.sciencedirect.com**SciVerse ScienceDirect**

Procedia Computer Science 6 (2011) 201–206

Procedia
Computer Science

Complex Adaptive Systems, Volume 1
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2011- Chicago, IL

Stock Market Prediction with Multiple Regression, Fuzzy Type-2 Clustering and Neural Networks

David Enke^{a*}, Manfred Grauer^b, Nijat Mehdiyev^b

^a*Department of Finance and Operations Management, The University of Tulsa, Tulsa, OK 74104, USA*

^b*Department of Information Systems, University of Siegen, 57076, Siegen, Germany*

Abstract

Stock market forecasting research offers many challenges and opportunities, with the forecasting of individual stocks or indexes focusing on forecasting either the level (value) of future market prices, or the direction of market price movement. A three-stage stock market prediction system is introduced in this article. In the first phase, Multiple Regression Analysis is applied to define the economic and financial variables which have a strong relationship with the output. In the second phase, Differential Evolution-based type-2 Fuzzy Clustering is implemented to create a prediction model. For the third phase, a Fuzzy type-2 Neural Network is used to perform the reasoning for future stock price prediction. The results of the network simulation show that the suggested model outperforms traditional models for forecasting stock market prices.

© 2011 Published by Elsevier B.V. Open access under [CC BY-NC-ND license](#).

Keywords: Stock Market Prediction; Forecasting; Multiple Regression Analysis; Fuzzy type-2 Clustering; Fuzzy Neural Networks; Hybrid Model; Differential Evolution Optimization

1. Introduction

The goal of stock market prediction research is to seek a system which can predict stock price levels precisely. Taking into account the nonlinearities and discontinuities of the factors which are considered to impact stock markets, the selection process of a manageable amount of the financial and economic data is often viewed as a necessary initial stage of any stock market prediction model. Multiple Regression Analysis, Principal Component Analysis, Hurst Exponent Analysis, and Grey Relation Analysis can be used during this initial step [Ince and Trafalis, 2007; Hurst, 1951; Kung and Wen, 2007]. Given stock market model uncertainty, neuro-fuzzy techniques are viable candidates to capture stock market nonlinear relations, returning good forecasting results without prior knowledge of the input data statistical distribution [Atsalakis and Valavanis, 2009]. Atsalakis and Valavanis [2009] analyzed more than 100 related published articles that focus on artificial neural networks and neuro-fuzzy techniques derived and applied to forecasting stock markets. None of the reviewed authors of these articles used

* Corresponding author. Tel.: 918-631-2218; fax: 918-631-2037.

E-mail address: david-enke@utulsa.edu.

Differential Evolution (DE) based Fuzzy type-2 Clustering or Fuzzy type-2 Inference Neural Networks in their models. The following will present a hybrid model that combines Multiple Regression Analysis, Fuzzy type-2 Clustering, and a Fuzzy type-2 Neural Network (See Figure 1).

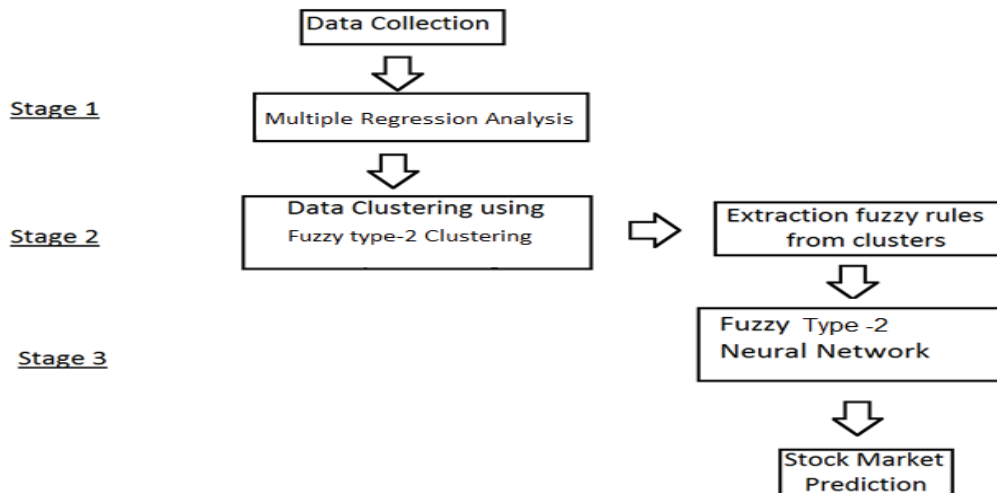


Figure 1: Proposed Hybrid Model

The remainder of the paper is organized as follows: Section 2 describes the application of Multiple Regression Analysis to reduce the dimensionality of the variable set. The structure of the Differential Evolution Optimization-based Fuzzy type-2 Clustering system that is implemented in this paper is then presented in Section 3. Section 4 discusses the architecture of the Fuzzy type-2 Neural Network used for predicting S&P 500 stock market prices. The experiments and empirical results of computer simulations are presented and described in Section 5. Finally, Section 6 provides a conclusion.

2. Multiple Regression Analysis

Recent studies in stock market prediction suggest that there are many factors which are considered to be correlated with future stock market prices. Nonetheless, using too many financial and economical factors can overload the prediction system [Thawornwong and Enke, 2003; Hadavandi et al., 2010; Chang and Liu, 2008; Esfahanipour and Aghamiri, 2010]. As a result, one of the initial and most challenging steps of stock market prediction is determining the manageable amount of the input variables which have the strongest forecasting ability and can be used as inputs to a prediction system (the Fuzzy type-2 Neural Network in the proposed model). In order to understand which of the input variables are significant and driving the output, Multiple Regression Analysis is implemented in our hybrid model. The Multiple Regression Analysis is performed on 25 finance and economic variables to both reduce the dimensionality of the variable set and identify which variables have a strong relationship with the market price of the S&P 500 Index for the subsequent testing period. Example variables include past prices, volume, technical indicators, T-bill rates, certificate of deposit rates, credit ratings, producer/consumer price indexes, industrial production levels, and money supply levels.

The input variables are collected on a monthly basis from January 1, 1980 to January 1, 2009, for a total of 361 months. The variables with inappropriate t-statistics and p-values are excluded from the list of inputs. According to the experiment results, the Multiple Regression Analysis method identified the 3-month T-bill (T-Bill3) rate, 3-month Certificate of Deposit (CDR3) rate, past S&P 500 (SP500) Index price level, past Money Supply (M1) level, recent Industrial Production (IP) reading, and the recent Producer Price Index (PPI) as significant and relevant variables in the regression model (with p-values less than 0.005). Analysis of the t-statistics and significance value of each variable suggested that these variables contain relevant information about the future stock prices. Given that the absolute t-statistics values of each variable was greater than 1 and the p-values (or significance values) are less than 0.005, the selected variables are believed to have strong forecasting ability. The regression analysis resulted in

the following equation for predicting the S&P 500 Index price level for the following month given variation in the inputs:

$$SP500_{t+1} = -33.690 + (11.999 * T\text{-}Bill3_t) + (1.365 * IP_t) + (.059 * M1_t) - (.781 * PPI_t) - (9.124 * CDR3_t) + (.955 * SP500_t) \quad (1)$$

The relationship generated from the regression (Equation 1) indicates that the $SP500_{t+1}$ is a function of the intercept (-33.690) plus various coefficients multiplied times the relevant variables. For this formula, positive changes in $T\text{-}Bill3_t$, $SP500_t$, IP_{t-1} , and $M1_{t-1}$ have positive effects on the prediction of the stock market level for the next month ($SP500_{t+1}$), while the positive changes of $CD3_t$ and PP_{t-1} have negative effects. The R-squared value of the model is 0.994, implying that the equation explains 99.4 percent of the variation of the future stock market price.

Table 1: Model Summary

R	R Squared	Adjusted R Squared	Std. Error of Estimate
0.997	0.994	0.994	35.604

3. Differential Evolution Optimization-based Fuzzy type-2 Clustering

Different types of clustering analysis have been applied by various researchers in numerous fields of science. The agglomerative hierarchical clustering and the nonhierarchical clustering are the two main types of clustering techniques. The agglomerative hierarchical clustering methods, which are used as an explanatory statistical technique to determine the number of clusters of data sets, include Single, Complete, and Average linkage methods. These methods are appropriate for both qualitative and quantitative variables, whereas the Centroid and Wards methods are applied only for quantitative variables [Johnson and Wichern, 2002]. K-means, Fuzzy c-means, SOM neural networks and Differential Evolution-based Fuzzy Clustering are the main types of nonhierarchical clustering analysis [Aliev et al., 2011, Gardashova et al. 2010].

Some researchers have conducted studies to compare the effectiveness of different clustering techniques. The proposed model utilizes a Differential Evolution-based Fuzzy type-2 Clustering technique. Fuzzy clustering is a well-established paradigm used to generate the initial type-2 fuzzy “If-Then” model [Hwang and Chung-Hoon Rhee, 2007]. For the proposed model, Fuzzy type-2 Differential Evolution-based Clustering is used since it has been proven to produce results that better suit the application of type-2 If-Then rules [Aliev et al., 2011]. This removes the uncertainty in choosing the “ m parameter” existing in Fuzzy c-means by suggesting a solution for a range of its values covering $\{1.4, 2.6\}$, a meaningful range for “ m .” Adequate choice of m is very important as it plays a visible role in forming the shape of resulting fuzzy clusters [Aliev et al., 2011].

As the experiments have shown, using a type-2 Fuzzy Clustering method provides better location of the cluster centers, and subsequently results in a better fuzzy rule model. This in turn allows capturing more uncertainty, while delivering higher robustness against the imprecision of the data. The objective function is as follows (with n data vectors, $P = \{p_1, p_2, \dots, p_n\}$ inputs; prototype v_j of the j^{th} cluster generated by the fuzzy clustering; membership degree u_{ij} of the i^{th} data belonging to the j^{th} cluster represented by the prototype v_j):

$$J_{m1} = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^{m_1} \|p_i - v_j^{(1)}\| \rightarrow \min, \quad J_{m2} = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^{m_2} \|p_i - v_j^{(2)}\| \rightarrow \min \quad (2)$$

subject to constraints:

$$0 < \sum_{i=1}^n u_{ij} < n \quad (j = 1, 2, \dots, c) \quad \text{and} \quad \sum_{j=1}^c u_{ij} = 1 \quad (i = 1, 2, \dots, n)$$

The vector $\tilde{\mathbf{v}}_i$ is formed as:

$$\tilde{\mathbf{v}}_i = [\min(\mathbf{v}_i^{(1)}, \mathbf{v}_{Ind_i}^{(2)}), \max(\mathbf{v}_i^{(1)}, \mathbf{v}_{Ind_i}^{(2)})] \text{ where } Ind_i = \arg \min_j \|\mathbf{v}_i^{(1)}, \mathbf{v}_j^{(2)}\| \quad (3)$$

In the proposed approach, Differential Evolution [Price et al., 2005] is used for minimization of the objective function (Equation 1) since it acts as a global search algorithm and is expected to be more advantageous than standard Fuzzy c-means for the case of a large number of highly-dimensional data vectors. From this fuzzy model, the linguistic hedges approach [Aliev et al., 2011] can be used to derive the corresponding interpretable linguistic model. Fragments of the Fuzzy type-2 IF-THEN model(rules) discovered by fuzzy clustering are shown below:

1) IF S&P500 is about 1222 AND T-Bill3 is about 1.69 AND CDR3 is 2.96 AND PPI is about 136.93 AND M1 is about 498.56 AND IP is about 12.37
THEN S&P500 is about 1044

.....

7) IF S&P500 is about 878.79 AND T-Bill3 is about 5.19 AND CDR3 is 5.90 AND PPI is about 97.40 AND M1 is about 679.06 AND IP is about 50.49
THEN S&P500 is about 899.20

4. The Structure of Type-2 Fuzzy Inference Neural Network

The structure of the proposed type-2 Fuzzy Inference Neural Network (T2FINN) [Aliev et al., 2011] is shown in Figure 2. Layer 1 maps inputs to type-2 fuzzy terms used in the rules. Layer 2 defines nodes representing the chosen rules. Each rule node performs the Min operation on the outputs (interval valued membership degrees) of the incoming links from the previous layer. Layer 3 consists of output term membership functions of type-1. Layer 4 computes the fuzzy output signal for the output variables. Layer 5 realizes the defuzzification using the Center-of-Gravity (COG) defuzzification method.

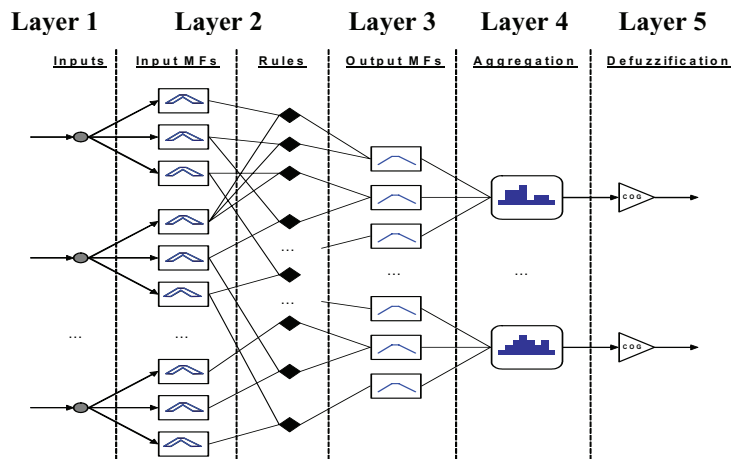


Figure 2: Structure of the type-2 Fuzzy Inference Neural Network

5. Computer Simulation

The goal of the hybrid model is to determine the next stock price value by using the observed numerical data. The dataset used for testing contains 360 records. As determined during regression analysis, the input included the 3-month T-bill (T-Bill3) rate, 3-month Certificate of Deposit (CDR3) rate, past S&P 500 (SP500) Index level, past Money Supply (M1) level, recent Industrial Production (IP) reading, and the recent Producer Price Index (PPI). The output was the next S&P 500 Index value. For the simulation, the Differential Evolution-based Fuzzy type-2

clustering model included seven clusters, max iteration =1000, exponent =2, and population size=200. Parameters of the type-2 neural network (that was initiated during the clustering procedure) are further adjusted by the Differential Evolution algorithm on the training series (80% of all data). Figure 3 provides examples of the produced type-2 membership functions.

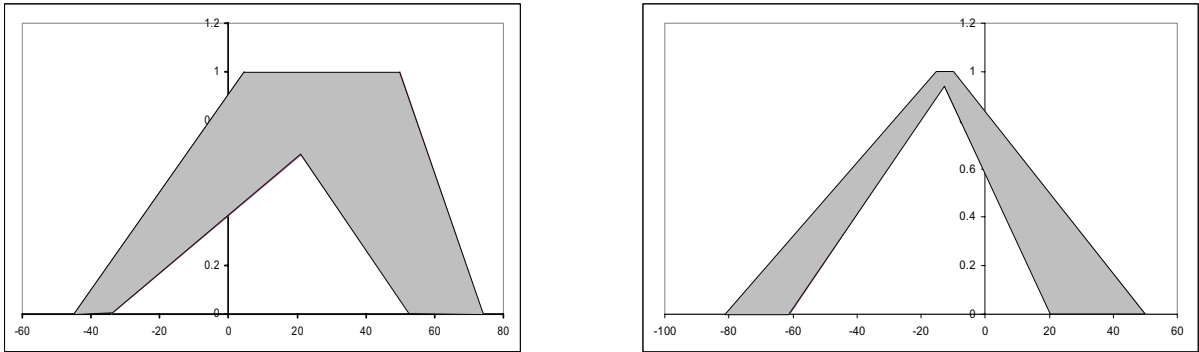


Figure 3: Examples of Parameters of the Fuzzy type-2 Network after Training of the Data Series

Performance of the type-2 neural network on both the training and testing data series is shown in Figure 4.

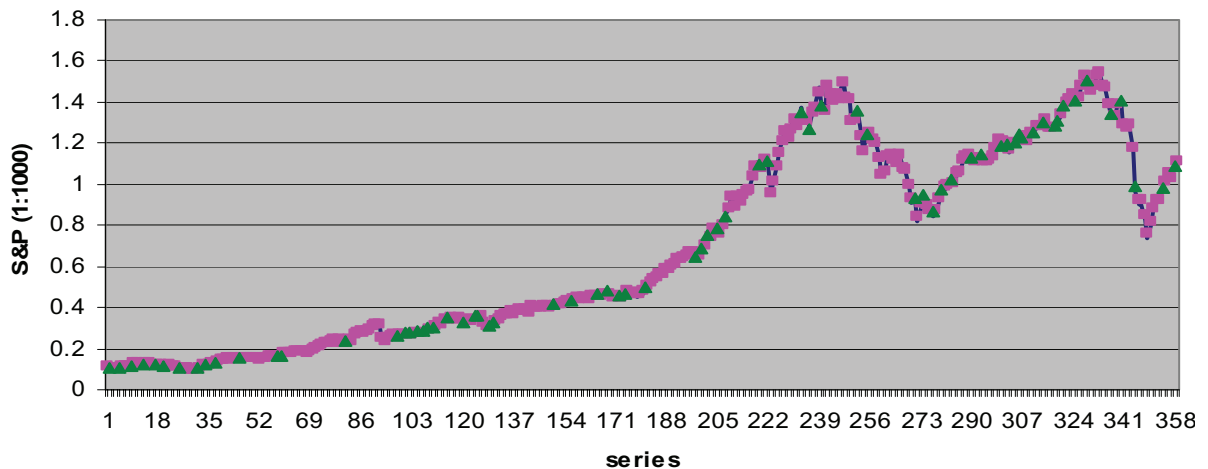


Figure 4: Fuzzy type-2 Neural Network Performance

The RMSE error comparison with a type-1 model is shown in Table 2.

Table 2: Comparison of Performance (type-1 compared to type-2)

	RMSE
Type-1 approach	0.948
Type-2 (the suggested approach)	0.909

6. Conclusion

Taking into account that gradient-based optimization methods (such as Fuzzy c-means) may not give a solution that reaches the global minimum (since it may get stuck in a local minimum), stock price prediction has been studied

using a Fuzzy type-2 Neural Network with Fuzzy type-2 clustering. Furthermore, since the standard iterative scheme may not be directly applicable to the considered problem when various distance functions are used, Differential Evolution-based algorithms were used for solving the clustering optimization problem. Further refinements were made by supplying the Differential Evolution optimization to the Fuzzy Neural Network inference system. When applied to the problem of stock market forecasting, simulations resulted in a lower prediction error when a Fuzzy type-2 approach is used as compared to a Fuzzy type-1 approach.

References

- Aliev R. A., W. Pedrycz, B. Guirimov, R. R. Aliyev, U. Ilhan, M. Babagil, and S. Mammadli, "Type-2 Fuzzy Neural Networks with Fuzzy Clustering and Differential Evolution Optimization," *Information Sciences* (2011): pp. 1591-1608.
- Atsalakis G. S., and K. P. Valavanis, "Surveying stock market forecasting techniques – Part II: Soft computing methods," *Expert Systems with Applications*, Vol. 36, Issue 3, Part 2 (2009): pp. 5932-5941.
- Chang, P.C., and H.C. Liu, "A TSK type fuzzy rule based system for stock price prediction," *Expert Systems with Applications*, Vol. 34 (2008): pp. 135-144.
- Esfahanipour, A., and W. Aghamiri, "Adapted neuro-fuzzy inference system on indirect approach TSK fuzzy rule base for stock market analysis," *Expert Systems with Applications*, Vol. 37 (2010): pp. 4742–4748.
- Gardashova, L. A., N. Mehdiyev, J. I. Ramazanov, and R. R. Aliyev, "Extraction Rules From Data By Using Differential Evolution Based Fuzzy Clustering Method," *Sixth World Conference on Intelligent Systems for Industrial Automation*, Tashkent, Uzbekistan (2010): pp. 322-329.
- Hadavandi, E., H. Shavandi, and A. Ghanbari, "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting," *Knowledge-Based Systems*, Vol. 23, Issue 8 (2010): pp. 800-808.
- Hurst, H.E., "Long-term storage of reservoirs: an experimental study," *Transactions of the American Society of Civil Engineers*, Vol. 116 (1951): pp. 770-799.
- Hwang C., and F. Chung-Hoon Rhee, "Uncertain fuzzy clustering: Interval Type-2 Fuzzy Approach to C-Means," *IEEE Transactions on Fuzzy Systems*, Vol. 15, No. 1 (2007): pp. 107-120.
- Ince, H., and B. T. Trafalis, "Kernel principal component analysis and support vector machines for stock price prediction," *IIE Transactions*, Vol. 39, No. 6, (2007): pp. 629-637.
- Johnson, R.A., and D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, 2002.
- Kung, C.Y., and K. L. Wen, "Applying Grey Relational Analysis and Grey Decision-Making to evaluate the relationship between company attributes and its financial performance - A case study of venture capital enterprises in Taiwan," *Decision Support Systems*, Vol. 43, Issue 3 (2007): pp. 842-852.
- Price, K., R. Storn, and J. Lampinen, Differential evolution – a practical approach to global optimization, Springer, 2005.
- Thawornwong, S., and D. Enke, "The Adaptive Selection of Financial and Economic Variables for Use with Artificial Neural Networks," *Neurocomputing*, Vol. 56 (2003): pp. 205-232.