ISBN: 978-972-8924-40-9 © 2007 IADIS

SELF-ORGANIZING DATA MINING USING ENHANCED GROUP METHOD DATA HANDLING APPROACH

Godfrey C. Onwubolu[1], Petr Buryan[2] and Ajith Abraham[3]
[1]*School of Engineering and Physics, University of the South Pacific, Private Bag, Suva, Fiji. E-mail: onwubolu_g@usp.ac.fj*
[2]*Gerstner Laboratory, Department of Cybernetics, Czech Technical University, Technicka 2, 166 27 Prague, Czech Republic*
buryan@labe.felk.cvut.cz
[3]*Center of Excellence for Quantifiable Quality of Service (Q2S), Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, O.S. Bragstads plass 2E, N-7491 Trondheim, Norway*
ajith.abraham@ieee.org

ABSTRACT

Data Mining (DM) is a relatively recent technology that is employed in inferring useful knowledge that can be put to use from a vast amount of data. This paper presents the data mining processes applied to the seemingly chaotic behavior of stock markets which could be well represented using the enhance GMDH, and we compared its results with published results using neural network, TS fuzzy system and hierarchical TS fuzzy techniques. To demonstrate the capabilities of the different techniques, we considered Nasdaq-100 index of Nasdaq Stock Market^{SM} and the S&P CNX NIFTY stock index. We analyzed 7 year's Nasdaq 100 main index values and 4 year's NIFTY index values. This paper investigates the development of novel reliable and efficient techniques to model the seemingly chaotic behavior of stock markets. Experimental results reveal that all the models considered could represent the stock indices behavior very accurately and that the proposed e-GMDH approach is a useful for data mining technique for forecasting and modeling stock indices.

1. INTRODUCTION

Data Mining (DM) is an important component of the emerging field of knowledge discovery in databases (KDD). Large databases of digital information are ubiquitous. Current hardware and database technology allow efficient and inexpensive reliable data storage and access. With the exponential rate at which data is becoming available to user, one question that needs to be answered is, what else do we do with all the available data? This is where opportunities for KDD and consequently DM naturally arise. However, whether the context is business, medicine, science, engineering, or government, the datasets themselves in raw form are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. Along side with KDD, two technologies that have attention in terms of research and application are DM and Data Warehousing (DW), which helps set the stage for KDD through data cleaning and data access [1-6].

In this paper, we apply the Data Mining steps to prediction of stocks, which is generally believed to be a very difficult task. The process behaves more like a random walk process and time varying. The obvious complexity of the problem paves way for the importance of intelligent prediction paradigms. During the last decade, stocks and futures traders have come to rely upon various types of intelligent systems to make trading decisions [7-8]. Several intelligent systems have in recent years been developed for modeling expertise, decision support and complicated automation tasks etc [9][10]. In this paper, we analyzed the seemingly chaotic behavior of two well-known stock indices namely Nasdaq-100 index of Nasdaq^{SM} [11] and the S&P CNX NIFTY stock index [12]. Chen *et al*. [13] considered neural network, Takagi-Sugano-Fuzzy system (TS-FS) [14] and hierarchical fuzzy system [15]. Our research reported in this paper is to investigate the performance analysis of the enhanced Group Modeling of Data Modeling (e-GMDH) paradigms [16—19] as a self-organizing data mining technique for modeling the Nasdaq-100 and NIFTY stock market indices. We analyzed the Nasdaq-100 index value from 11 January 1995 to 11 January 2002 [11] and the NIFTY index from 01 January 1998 to 03 December 2001 [12]. For both the indices, we divided the entire

data into almost two equal parts. No special rules were used to select the training set other than ensuring a reasonable representation of the parameter space of the problem domain [8].

2. THE ENHANCED GROUP MODELING OF DATA MODELING

The basics steps involved in the original Group Method of Data Handling (GMDH) modeling approach [16] are as follows:

Preamble: collect regression-type data of *n*-observations and divide the data into training and testing sets: $x_{ij};\ y_i \qquad i=1,2,...,n;\ j=1,2,...,m$

Step 1: Construct ${}^{m}C_2$ new variables $Z_1, Z_2, Z_3, ..., Z_{\binom{m}{2}}$, in the *training dataset* for all independent variables (columns of X), two at a time $\left(x_{i,k-1}, x_{i,k};\ i\in[1,m], k\in\left[2,\binom{m}{2}\right]\right)$ and construct the regression polynomial:

$$Z_1 = A + Bx_1 + Cx_2 + Dx_1^2 + Ex_2^2 + Fx_1x_2 \quad \text{at points } (x_{11}, x_{12}) \tag{1}$$

$$Z_k = A + Bx_{k-1} + Cx_k + Dx_{k-1}^2 + Ex_k^2 + Fx_{k-1}x_k \quad \text{at points } (x_{i,k-1}, x_{i,k}) \tag{2}$$

Step 2: For each of these regression surfaces, evaluate the polynomial at all *n* data points (i.e. using *A*, *B*, *C*, *D*, *E*, and *F* obtained from $x_{i,k-1}, x_{i,k};\ y_i$ for training)

Step 3: Eliminate the least effective variables: replace the columns of X (old variables) by those columns of Z (new variables) that best estimate the dependent variable *y* in the testing dataset such that

$$d_k^2 = \sum_{i=n_t+1}^{n}\left(y_i - z_{i,k}\right)^2, \qquad k\in\left[1,2,...,\binom{m}{2}\right] \tag{3}$$

Order Z according to the least square error $d_k \mid \|d_j\| < R$ where R is some prescribed number chosen *a priori*. Replace columns of X with the best Z's $(Z_{<R})$; in other words $X_{<R} \leftarrow Z_{<R}$

Step 4: Test for convergence. Let $DMIN = d_k$. If $DMIN_k = DMIN_{k-1}$ go to Step 1, else stop the process.

Since the introduction of GMDH, there have been variants devised from different perspectives to realize more competitive networks. Considering the diversity problem, three new pruning techniques, which are embedded in the enhanced e-GMDH structure [20] along with the classical "best-of" selection, all of them having been inspired by the evolutionary algorithms, are (1) classical (MIA)-GMDH "best-of" approach (2) roulette wheel selection (3) semi-randomized selection (random factor selection) and (4) totally randomized selection.

3. STOCK INDEX PREDICTION USING GMDH PARADIGM

3.1 The Data Set

We considered 7-year stock data for Nasdaq-100 Index and 4-year data for NIFTY index. Our target is to develop efficient forecast models that could predict the index value of the following trade day based on the opening, closing and maximum values of the same on a given day. Test data was presented to the trained soft computing models and the output from the network was compared with the actual index values in the time series. The assessment of the prediction performance of the different soft computing paradigms was done by quantifying the prediction obtained on an independent data set. The Root Mean Squared Error (RMSE), Maximum Absolute Percentage Error (MAP) and Mean Absolute Percentage Error (MPE) and Correlation Coefficient (CC) were used to study the performance of the trained forecasting model for the test data. *MAP* is defined as follows:

$$MSE = \sqrt{\frac{\sum (y(t) - \hat{y}(t))^2}{N}} = \sqrt{\frac{\sum e^2(t)}{N}}, \quad (4)$$

$$MAP = \max(\frac{|y(t) - \hat{y}(t)|}{y(t)} \times 100\%) \quad (5)$$

where $y(t)$ is the actual index value on a particular day and is the forecast value of the index on that day. Similarly *MPE* is given as

$$MPE = \frac{\sum \frac{|y(t) - \hat{y}(t)|}{y(t)}}{N} \times 100\% = \frac{\sum \frac{|e(t)|}{y(t)}}{N} \times 100\% \quad (6)$$

where N represents the total number of days.

3.2 Experimental Results for NIFTY index

For simulation, the five-day data sets were prepared for the NIFTY index from 01 January 1998 to 03 December 2001. The experimental results for MAP, MPE, and RMSE are presented. The output of the RMSE criterion for the Nifty index is shown in Figure 1.

3.3 Experimental Results for NASDAQ Index

For simulation, the three-day data sets were prepared for the NASDAQ index from 11 January 1995 to 11 January 2002. The experimental results for MAP, MPE, and RMSE are presented. In this section we show all the graphical outputs (Figures 2-6) for the Nasdaq index since visualization is an important aspect of Data Mining. The output of the RMSE criterion for the Nasdaq index is shown in Figure 4. Our proposed e-GMDH approach realized a model for the Nasdaq index represented as:

$$\hat{y}(t) = -0.1676 + 0.5905x(t-1) + 0.3929x(t-2) + 0.0306x(t-3)$$
$$+ 0.0004x(t-1)^2 + 0.0166x(t-2)^2 + 0036x(t-3)^2 - 0.0133x(t-1) \times 0.0155x(t-2)$$
$$+ 0.0133x(t-1) \times 0.0230x(t-3) - 0.0155x(t-2) \times 0.0230x(t-3)$$

For comparison purpose, the training and test performances of hybrid neural network-particle swarm optimization (*NN-PSO*), *fuzzy-TS*, *hybrid TS-FS* and *e-GMDH* proposed in this paper for modeling stock index are shown in Table 1. The statistical analysis performances of the four learning methods (test data) are shown in Table 2.

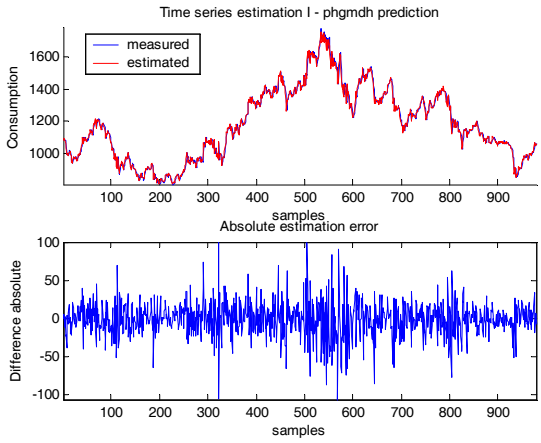

Figure 1. The actual and predicted Nifty index using RMSE criterion

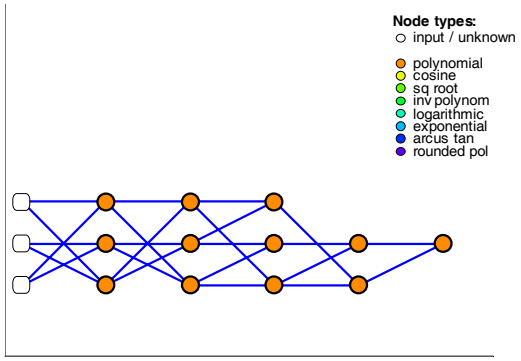

Figure 2. The *e-GMDH* network connections after pruning for the Nasdaq index problem

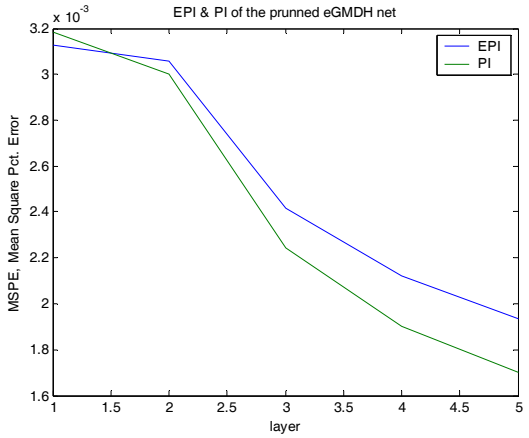

Figure 3. Nasdaq index training (PI) and testing (EPI) performances for different layers

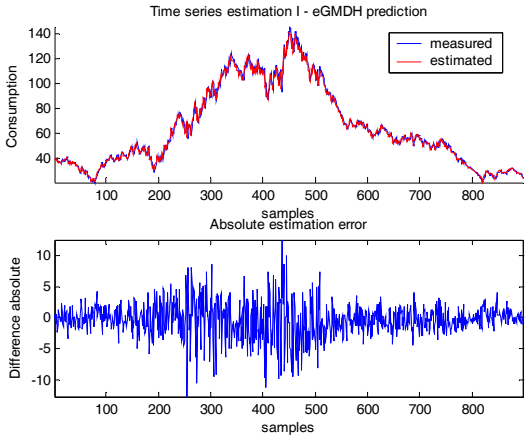

Figure 4. The actual and predicted Nasdaq index using RMSE criterion

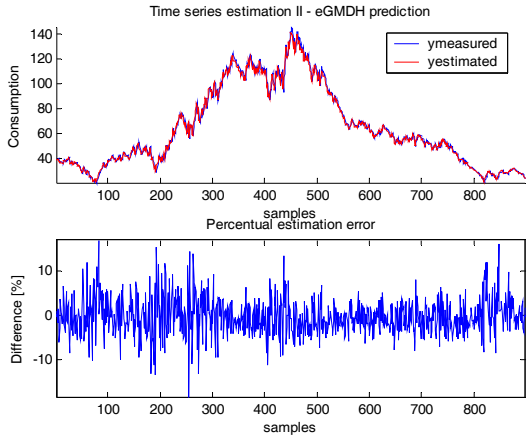

Figure 5. The e-GMDH percentage error for the Nasdaq index problem

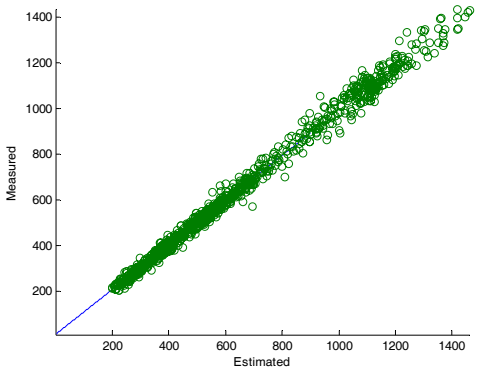

Figure 6. The GMDH prediction and measured values for the Nasdaq index problem

Table 1. Empirical comparison of RMSE results for four learning methods

	NN-PSO	Fuzzy-TS	H-TS-FS	e-GMDH
Training results (RMSE)				
Nasdaq-100	0.02573	0.02634	**0.02498**	0.04123
NIFTY	0.01729	0.01895	**0.01702**	0.01850
Testing results (RMSE)				
Nasdaq-100	0.01864	0.01924	**0.01782**	0.04398
NIFTY	**0.01326**	0.01468	0.01328	0.01998

Table 2. Statistical analysis of four learning methods (test data)

	NN-PSO	Fuzzy-TS	H-TS-FS	e-GMDH
Nasdaq-100				
Correlation coefficient	**0.997704**	0.997538	0.997698	0.9963
MAP	141.363	156.464	138.736	**16.103**
MAPE	6.528	6.543	6.205	**3.1350**
NIFTY				
Correlation coefficient	0.997079	0.997581	0.997685	**0.994**
MAP	27.257	30.432	27.087	**5.418**
MAPE	3.092	3.328	3.046	**1.346**

4. CONCLUSIONS

In this paper, we have demonstrated how the chaotic behavior of stock indices could be well represented by different hybrid learning paradigms. Empirical results on the two data sets using four different learning models clearly reveal the efficiency of the techniques. In terms of RMSE values, for Nasdaq-100 index, H-TS-FS performed marginally better than other models and for NIFTY index, NN approach gave the lowest generalization RMSE values. For both data sets, H-TS-FS has the lowest training error. For Nasdaq-100 index (test data), Fuzzy TS has the highest correlation coefficient but the lowest value of MAPE and MAP value was achieved by using the H-TS-FS model. Highest correlation coefficient, and the best MAPE/MAP values for NIFTY index were achieved using the H-TS-FS trained using GP-like evolutionary algorithm and the PSO model. The number of fuzzy rules obtained by direct fuzzy method is 27 for Nasdaq-100 data and 243 for NIFTY data. The number of fuzzy rules for obtained by H-TS-FS is 18 for Nasqad-100 data and 99 for NIFTY data. A low MAP value is a crucial indicator for evaluating the stability of a market under unforeseen fluctuations. In terms of the MAP and MAPE criteria, the proposed e-GMDH performs significantly better than NN-PSO, Fuzzy TS and H-TS-FS for both the Nifty and Nasdaq-100 indices.

In the present example, the predictability assures the fact that the decrease in trade is only a temporary cyclic variation that is perfectly under control. Our research was to predict the share price for the following trade day based on the opening, closing and maximum values of the same on a given day. Our experiment results indicate that the most prominent parameters that affect share prices are their immediate opening and closing values. The fluctuations in the share market are chaotic in the sense that they heavily depend on the values of their immediate forerunning fluctuations. Long-term trends exist, but are slow variations and this information is useful for long-term investment strategies. Our study focuses on short term, on floor trades, in which the risk is higher. However, the results of our study show that even in the seemingly random fluctuations, there is an underlying deterministic feature that is directly enciphered in the opening, closing and maximum values of the index of any day making predictability possible. Empirical results also show that there are various advantages and disadvantages for the different techniques considered. There is little reason to expect that one can find a uniformly best learning algorithm for optimization of the performance for different stock indices. This is in accordance with the no free lunch theorem, which explains that for any algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class [21].

REFERENCES

[1] Fayyad, U. M.; Djorgovski, S. G.; and Weir, N. 1996. From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey. *AI Magazine* 17(2): 51–66.

[2] Fayyad, U. M.; Haussler, D.; and Stolorz, Z. 1996. KDD for Science Data Analysis: Issues and Examples. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 50–56. Menlo Park, Calif.: American

[3] Association for Artificial Intelligence.

[4] Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1–30. Menlo Park, Calif.: AAAI Press.

[5] Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. 1996. *Advances in Knowledge Discovery and Data Mining.* Menlo Park, Calif.: AAAI Press.

[6] Brachman, R., and Anand, T. 1996. The Process of Knowledge Discovery in Databases: Human-Centered Approach. In *Advances in Knowledge Discovery and Data Mining,* 37–58, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.: AAAI Press.

[7] Elder, J., and Pregibon, D. 1996. A Statistical Perspective on KDD. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 83–116. Menlo Park, Calif.: AAAI Press.

[8] Abraham A., Nath B. and Mahanti P.K.: Hybrid Intelligent Systems for Stock Market Analysis, Computational Science, Springer-Verlag Germany, Vassil N Alexandrov et al (Editors), USA, (2001)337-345.

[9] Abraham A., Philip N.S., and Saratchandran P.: Modeling Chaotic Behavior of Stock Indices Using Intelligent Paradigms, International Journal of Neural, Parallel and Scientific Computations, USA, Volume 11, Issue (1,2), (2003)143-160.

Leigh W., Modani N., Purvis R. and Roberts T.: Stock market trading rule discovery using technical charting heuristics, Expert Systems with Applications 23(2), (2002)155-159.

[10] Leigh W., Purvis R. and Ragusa J.M.: Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support, Decision Support Systems 32(4),(2002)361-377.

[11] Nasdaq Stock Market*SM*: http://www.nasdaq.com

[12] National Stock Exchange of India Limited: http://www.nse-india.com

[13] Yuehui Chen, Ajith Abraham, Ju Yang and Bo Yang, Hybrid Methods for Stock Index Modeling, 2005 International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'05), China, Lecture Notes in Computer Science, Volume 3614, Springer Verlag, pp. 1067- 1070, 2005.

[14] Takagi, T. and Sugeno, M.: Fuzzy identification of systems and its application to modeling and control. IEEE Trans. Syst. Man, Cybern., 15, (1985) 116-132

[15] Wang, L.-X. : Analysis and design of hierarchical fuzzy systems. IEEE Trans. Fuzzy Systems, 7, (1999) 617-624

[16] A. G. Ivakhnenko, Polynomial Theory of Complex Systems, IEEE Transactions on Systems, Man, and Cybernetics, 1971, pp. 364-378.

[17] J.-A. Mueller, F. Lemke, A. G. Ivakhnenko, GMDH algorithm for complex systems modeling, Mathematical Modeling of Systems, 1997.

[18] H. R. Madala, A. G. Ivakhnenko, Inductive Learning Algorithms for Complex Systems Modeling, CRC Press, Boca Raton, 1994.

[19] Farlow S. J. (Ed.): Self-organizing methods in Modeling GMDH-type Algorithms, Marcel Decker, N.Y.,1984.

[20] Buryan, P. and Onwubolu, G. C., 2007, Design of Enhanced MIA-GMDH Learning Networks, (review in process).

[21] Macready W.G. and Wolpert D.H.: The No Free Lunch theorems, IEEE Trans. On Evolutionary Computing, 1(1), (1997)67-82.