A NEW APPROACH FOR TIME SERIES FORECASTING BASED ON GENETIC ALGORITHM

Mahesh S. Khadka, Benjamin Popp, K. M. George, N. Park
Computer Science Department
Oklahoma State University
Stillwater, OK 74078

Abstract

In this paper, we propose a new fusion approach to predict time series based on Concordance and Genetic Algorithm. Different measures of concordances such as the Kendall's Tau, Gini's Mean Difference, Spearman's Rho, and a weak interpretation of the Weak Concordance are used to identify these generic trends. The concept is validated using Financial Time Series data *(S&P 500 Index)* as the sample data set.

1 Introduction

This paper proposes and implements a new fusion approach based on Concordance and Genetic Algorithm to predict time series in short term in the same or another time series. These generic trends in the time series are identified by the use measures of concordance such as the *Kendall's Tau, Gini's Mean Difference, Spearman's Rho*, and a weak interpretation of the *Weak Concordance*. This paper validates the concept using Financial Time Series data as the sample data set. We apply this method to forecast *S&P 500* Index.

Stock Market forecasting is considered as one of the most challenging tasks in present financial world. So a lot of attention has been made to analyze and forecast future values and behavior of financial time series. Many factors interact in the stock market including business cycles, interest rates, monitory policies, general economic conditions, traders' expectations, political events, etc. According to academic investigations, movements in market prices are not random. Rather they behave in a highly non-linear, dynamic manner [14]. The ability to predict the direction and correct value of the future stock market values is the most important factor in financial market to make money.

With the introduction of online trading, stock market has emerged as one of the investment venues where anyone can earn profits. So forecasting the correct value and behavior of stock market has become the area of interest. However, because of the high volatility of the underlying laws behind the financial time series, it is not any easy task to build such a forecasting model [15]. Numbers of forecasting techniques have been proposed so far with their own merits and limitations. Especially the conventional statistical techniques are constrained with the underlying seasonality, non-stationary and other factors *(Tambi, 2005)* [15].

The remainder of the paper is organized as follows: Section 2 describes the previous approaches to forecasting, Section 3 presents definitions of terms used in this paper, Section 4 presents the proposed fusion model, Section 5 presents the experimental results and analysis and finally conclusion and directions for further work are presented in Section 6.

2 Previous Approaches to Forecasting

There have been many ways in which the prediction of time series has been proposed, such as extrapolation, linear prediction etc. Generally there exist two classes of methods of prediction; Parametric Methods and Non-Parametric Methods [4].

2.1 Parametric Approach

The parametric approach assumes that we can predict the outcome of a time series data based on certain parameters on which the time series is dependent upon. The first stage of such approach typically involves the identification of the parameters on which the data depends. Then, a function or a set of functions on these parameters are constructed. The measures of the parameters are collected from the data and are then used in the set of functions to predict the value of the series.

The parametric approaches are classified into two types based on the types of functions that are used for prediction. They are linear parametric approach and non-linear parametric approach. Linear parametric approach emphasizes that the function or the set of

functions defined on the parameters be linear whereas the non-linear parametric approach emphasizes that these functions be non-linear.

Various other approaches are also taken for prediction of time series in economics such as Auto Regressive Moving Average, ARMA, Auto Regressive Integrated Moving Average ARIMA and the Seasonal ARIMA [4]. The ARMA method involves two parts, Auto Regression and the Moving Average, that is, it takes into consideration the regression models of data and also the moving average for analyzing the time series data. The ARIMA method is a generalization of the ARMA model and is obtained by integrating the ARMA model.

2.2 Non-Parametric Approach

In the Non-Parametric approach, we assume that the data is independent of any other parameters. Some of the Non-Parametric methods that are in use are Multivariate Local Polynomial Regression, Functional Coefficient Autoregressive Model, Adaptive Functional Coefficient Autoregressive Model and the Additive Autoregressive Model [5]. Since the behavior of the varieties decays exponentially with increase in the amount of past data, one of the proposed ways is to convert a multi-dimensional problem into one-dimensional problem by incorporating a single trajectory in the model [13].

Another non parametric approach is the use of perceptrons or neural networks [7]. There are many ways to implement such approach. The predictive perceptron model or neural network is created and the historical data is fed as input to the neural network for training. Once the neural network completes the training stage, it can then be used for prediction. Several methods include, conversion of input data into a symbolic representation with grammatical inference in recurrent neural networks to aid the extraction of knowledge from the network in the form of a deterministic finite state automaton [8], preprocessing of input data into Embedded Phase-Space Vectors using delay co-ordinates [9], using special types of networks called Dynamic System Imitator which have been proved to model dynamic complex data [10]. Another method of prediction involves choosing of the training dataset that closely resembles the time series in the "Correlation Dimension" [11]. In some cases, there are separate neural nets that are used to find undetected regularities in the input dataset [7]. Another way of prediction is to apply a neural network to fuzzy time series prediction using bivariate models to improve forecasting [12].

The advantage of such a system over the parametric approach is that it is very robust, as it can adapt and respond to structural changes. The disadvantage of such an approach is that it can be very data intensive to get fully trained and cannot be used for any data set that is not huge [6].

3 Related Definitions

In this section, we outline our approach to forecasting the short term trends and define the terms that would be used in the rest of the paper.

3.1 Concordance

Concordance can be defined as the measure of agreement among raters.

Given the rating/ranking $X < x_1, x_2, ... >$ and $Y < y_1, y_2, ... >$ given by two judges say, then two pairs of rankings (x_i, y_i) and (x_j, y_j) are said to be concordant if $(x_i - x_j)(y_i - y_j) > 0$.

3.2 Kendall's Tau (τ)

The Kendall Tau Coefficient [1], developed by Maurice Kendall in 1938, is a statistic used to measure the degree of correspondence between two rankings and their significance. In other words, it measures the strength of association of the joint distribution of two or more variables.

Kendall's tau coefficient τ is defined as

$$\tau = \frac{2(n_c - n_d)}{n(n-1)}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and n is the number of elements in the dataset.

3.3 Spearman Rho (ρ)

The Spearman Rho correlation coefficient is a non-parametric measure of statistical dependence between two variables. It tells the magnitude and direction of the association between two variables that are on an interval or ratio scale. If X_i, Y_i be n data sets that are converted to ranks x_i, y_i, and the differences $d_i = x_i - y_i$ between the ranks of each observation on the two variables are calculated.

If there are no tied ranks, then ρ is given by

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2-1)}$$

3.4 Gini's Mean Difference (GMD)

This index is based on Gini's mean difference $D^{(m)}$ [1] computed on the totals ranks corresponding to each unit. It is more general measure of the agreement of m rankings. The index is used in a test for the independence of two criteria used to rank the units of a sample, against their concordance/discordance. For the present scenario, we consider *Gini's Mean Difference* with the case with two judges ($m = 2$). The value of $D^{(2)}$ ranges from *0* to *1* where *0* signifies very poor and *1* signifies high or perfect agreement between the judges. For n observations $< x_1, ..., x_n >$, relating to a quantitative variable X, *Gini's Mean Difference* (without repetition) can be defined as

$$GMD = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{n(n-1)}; i, j = 1, 2, ..., n; i \neq j$$

That is as the mean of the $n(n-1)$ absolute differences between every couple of different observations. If this measure of variability is applied to the total ranks $< T_1, ..., T_n >$, the following measure of agreement of m rankings is obtained.

3.5 Weak Concordance(w)

The Weak Concordance is a statistical measure of concordance for items that is calculated based on the varying successive values of a series. For all values in between, increasing value of w indicates increasing concordance. Weak Concordance (w) is defined as

$$w_x = \frac{\sum \delta_i}{n-1}; \text{ where}$$

$$\delta_i = \begin{cases} 1 & \text{if } (x_i - x_{i-1})(y_i - y_{i-1}) > 0 \\ -1 & \text{otherwise} \end{cases}$$

We use an adapted version of Weak Concordance to achieve Similarity Measure μ.

3.6 Pattern and Trend

A pattern $P < p_1, p_2, ..., p_n >$ is a distinctive combination or arrangement of elements in $\{0,1\}$ that is generated from the original time series. In this pattern, *0* represents decline of the next value in the data series where as *1* represents increment in the data series. A trend can be defined as sustained, relatively constant movement of a variable over a period of time.

4 Proposed Fusion Model

In this study, a fusion model of Concordance and Genetic Algorithm is developed to predict the future stock market behavior and values. In this method, we want to predict the future of stock using the history of other stocks. Since the past data is huge, we want to limit the past data so as to compare with the present using mathematical concordance. The weak *Tou, Gini,* and *Rho* Concordances of all the possible past segments are compared over a short period of time. This will come out with all the lengths and positions for high concordances. Higher the concordances and longer the matches, indicate better matches. A high concordance means that the trend is likely to continue, so we can use the past data to predict the future. To make the prediction as accurate as possible, we search the mathematical equation *g(x)* to map the past data to the future data to select which section of the past to use based on the concordances. The genetic program will then search for an equation such that

$$\forall k, g(p_k) \approx f_{k+n}$$

where k is a day in the past and n being the offset, in days. Specifically, we want to minimize $\sum(g(p_k) - f_{k+n})^2$ for all k by choosing the best possible function *g(x)*. The square makes larger differences matter much more than smaller differences. The function *g(x)* will get us close, but it will not be perfect. So we measure the error e_k for each term and subtract that error to get a better function. By extrapolating that error and using known values from the past, we can guess values that have not happened yet. This is done through genetic programming.

4.1 Algorithm

The algorithm constitutes of two parts. First part is for the main Program and the second part is for the Genetic Program.

4.1.1 Algorithm for Main Program

1. Get stock data for all stocks we want to test.
2. Search for the pattern in the past that look very similar to the present pattern using *Kendall's Tau, Gini's Mean Difference and Spearman's Rho* as probabilistic distance measure.
3. Find the highest recorded *Tau Concordance* among of all matches.
4. Use *Genetic Program* to match the past match to present trend as close as possible. Use this program to estimate what will happen next "*now*" based on what happened next "*then*".

5. Repeat Steps 3 and 4 with Gini and Rho Concordances.

4.1.2 Algorithm for Genetic Program

1. Generate a population of random Polynomials.
2. Compute a "*fitness*" of each polynomial, defined by $\sum_{k=1}^{l}(g(p_k) - f_{k+n})^2$ where g is the genetic polynomial, p is the past data, f is the present data, and l is the length of the section found by the concordance measures.
3. Sort the polynomials according to their fitness. Then replace the lower half of the population through breeding and mutating the upper half, along with adding new random individuals.
4. Repeat Steps *2* and *3* until a sufficiently low fitness is attained.

5 Experimentation and Result

The algorithms are evaluated using *S&P* data for *200* days that can be selected based on the weak concordance. The data set is selected based on the best Tau, Rho and Gini. In most of the cases, the values obtained from these three parameters are same. So we can take any of them to select the predictor data set. But if there is significant difference in those values, then we have to select three different data sets based on them which will give us three different predictions. However, the predicted values do not differ much. In our experiments, we have predicted values for a week that comprises of 5 business days. However, we can predict for more days. The direction column on the tables below shows whether the direction of actual value is same or not to that of predicted values. This means if actual stock value increases or decreases compared to its previous day value and if the predicted value also increases or decreases accordingly, and then the direction is represented as "*Same*" otherwise *Different.*

5.1 Experiment with same *Tau, Gini* and *Rho*

Figure *1* shows *S&P* prediction from *5 April 2010* to *9 April 2010.* The predictor data set is from *28 February 2007* to *21 May 2007* with same T*au, Gini* and *Rho.*

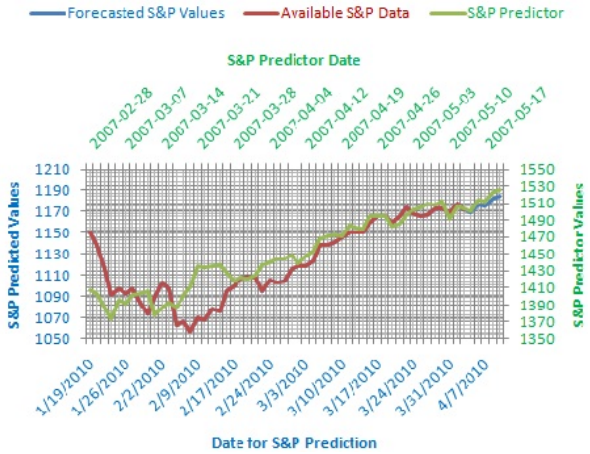

Figure 1 Stock Index Prediction for *S&P* from *04-05-2010* to *04-09-2010* with same *Tau, Gini* and *Rho.*

Table *1* shows the comparison between our predicted values and actual *S&P* values for the specified dates along with the percentage error.

Date	Actual S&P Value[3]	Predicted S&P Value	Direction	Error (%)
04-19-2010	1197.52	1199.40	-	0.16
04-20-2010	1207.17	1204.12	Same	0.25
04-21-2010	1205.94	1201.74	Same	0.35
04-22-2010	1208.67	1207.41	Same	0.11
04-23-2010	1217.28	1217.28	Same	0.45

Table 1 Comparing Predicted and Actual Stock Values for *S&P* from *04-05-2010* to *04-09-2010* with same *Tau, Gini* and *Rho.*

5.2 Experiment with different *Tau, Gini* and *Rho*

Figure *2* shows *S&P* prediction from *19 April 2010* to *23 April 2010.* The predictor data set is from *04 February 1959* to *27 April 1959* with *Tau.*

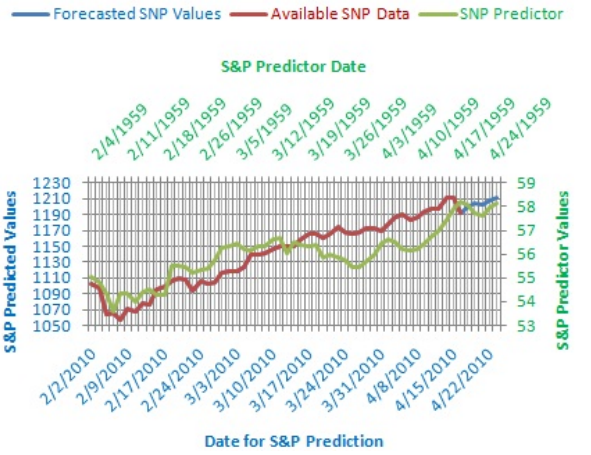

Figure 2 Stock Index Predictions for *S&P*

from *04-19-2010* to *04-23-2010* with *Tau*.

Table *2* shows the comparison between our predicted values and actual *S&P* values for the specified dates along with the prediction percentage error.

Date	Actual S&P Value[3]	Predicted S&P Value	Direction	Error (%)
04-19-2010	1197.52	1199.40	-	0.16
04-20-2010	1207.17	1204.12	Same	0.25
04-21-2010	1205.94	1201.74	Same	0.35
04-22-2010	1208.67	1207.41	Same	0.11
04-23-2010	1217.28	1217.28	Same	0.45

Table 2 Comparing Predicted and Actual Stock Values for *S&P* from *04-19-2010*to *04-23-2010* with *Tau*.

Figure *3* shows *S&P* prediction from *19 April 2010* to *23 April 2010*. The predictor data set is from *07 May 1987* to *28 July 1987* with *Gini*.

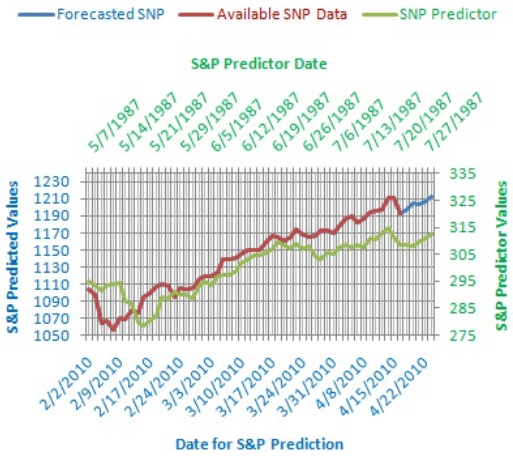

Figure 3 Stock Index Predictions for *S&P* from *04-19-2010* to *04-23-2010* with *Gini*.

Table *3* shows the comparison between our predicted values and actual *S&P* values for the specified dates along with the prediction percentage error.

Date	Actual S&P Value[3]	Predicted S&P Value	Direction	Error (%)
04-19-2010	1197.52	1198.06	-	0.05
04-20-2010	1207.17	1205.73	Same	0.12
04-21-2010	1205.94	1204.29	Same	0.14
04-22-2010	1208.67	1207.17	Same	0.13
04-23-2010	1217.28	1212.29	Same	0.41

Table 3 Comparing Predicted and Actual Stock Values for *S&P* from *04-19-2010* to *04-23-2010* with *Gini*.

Figure *4* shows *S&P* prediction from *19 April 2010* to *23 April 2010*. The predictor data set is from *23 March 1983* to *13 June 1983* with *Rho*.

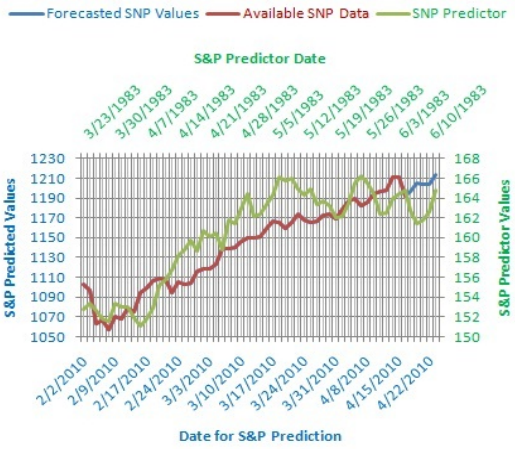

Figure 4 Stock Index Predictions for *S&P* from *04-19-2010* to *04-23-2010* with Rho.

Table *4* shows the comparison between our predicted values and actual *S&P* values for the specified dates along with the prediction percentage error.

Date	Actual S&P Value[3]	Predicted S&P Value	Direction	Error (%)
04-19-2010	1197.52	1197.03	-	0.05
04-20-2010	1207.17	1204.73	Same	0.21
04-21-2010	1205.94	1203.32	Same	0.22
04-22-2010	1208.67	1204.04	Same	0.38
04-23-2010	1217.28	1213.56	Same	0.31

Table 4 Comparing Predicted and Actual Stock Values for *S&P* from *04-19-2010* to *04-23-2010* with *Rho*.

From the above experiments, we can see that the direction of prediction is accurate in all experiments whereas the predicted values are closer to the actual values regardless of what we choose *Tau, Gini* or *Rho*.

6 Conclusion and Further Works

In this paper we proposed a new time series forecasting tool that combines concordance and Genetic Algorithm. From the experimental results, the performance of this forecasting tool is much more accurate than some of the existing tools not only in direction but also the values. The error percentage is merely not more than 2There is a lot of scope for future work in this proposed model tough. Currently we are working on developing an updated theory than can predict the future values even more precise and accurate than this model. We are trying to add up some more influence parameters on this model so that the forecasting would be more accurate.

References

[1] Kendall, M., A New Measure of Rank Correlation," *Biometrika*", Vol. 30, pp. 81-89, 1938.

[2] Borroni C.G., Zenga M., A test of concordance based on Gini's Mean Difference, "*Statistical Methods and Applications*, Vol. 16, pp. 289-308, 2006.

[3] Yahoo Finance Website (http://www.finance.yahoo.com).

[4] Gemai Chen, Bovas Abraham, Greg W. Bennett, Parametric and non-parametric modelling of time series: An empirical study, "*EnvironMetrics*", Vol. 8, pp. 63-74, 1997.

[5] Jianqing Fan, Qiwei Yao, "*Non-Linear Time Series*", Springer, 2003.

[6] Chai Quek, Michel Pasquier, and Neha Kumar, Novel Recurrent Neural Network Based Prediction System for Trading, "*International Joint Conference on Neural Networks*", IEEE, pp. 2090-2097, 2006.

[7] H. White, Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns, "*Proceedings of the 2nd Annual IEEE Conference on Neural Networks*", II: pp. 451-458, 1988.

[8] C. Lee Giles, Steve Lawrence, A. C. Tsoi, Noisy Time Series Prediction using a Recurrent Neural Network and Grammatical Inference, "*Machine Learning*", Vol. 44, Number 1/2, July/August, pp. 161-183, 2001.

[9] Zhang, J. Shu-Hung Chung, H. Wai-Lun Lo, Chaotic Time Series Prediction Using a Neuro-Fuzzy System with Time-Delay Coordinates, "*IEEE Transactions on Knowledge and Data Engineering*", Vol. 20, Issue: 7, 2008.

[10] Mohammad Bodruzzaman, Magdi A. Essawy, Iterative Prediction of Chaotic Time Series Using a Recurrent Neural Network, "*Artificial Neural Network in Engineering (ANNIG) Conference*", 1996.

[11] Chen Feng, Guangrong Ji, Wencang Zhao, Rui Nian, The prediction of the financial time series based on correlation dimension, "*Lecture Notes in computer Science*", Vol. 3612, 2005.

[12] Yu and Huarng, T.K. Yu and K. Huarng, A bivariate fuzzy time series model to forecast the TAIEX, "*Expert Systems with Applications*" Vol. 34 (4), pp. 2945-2952, 2008.

[13] Aneiros-Prez Germn, Vieu Philippe, Nonparametric time series prediction: A semi-functional partial linear modeling, "*Journal of Multivariate Analysis*", Vol. 99, Issue 5, pp. 834-857, 2008.

[14] Rohit Choudhary, Kumkum Garg, A Hybrid Machine Learning System for Stock Market Forecasting, "*World Academy of Science, Engineering and Technology*" , Issue 39, 2008.

[15] Md. Rafiul Hassan, Baikunth Nath, Michael Kirley, A fusion model of HMM, ANN and GA for stock market forecasting, "*Expert Systems with Applications*", Issue 33, pp. 171-180, 2007.