Forecasting Chaotic Stock Market Data using Time Series Data Mining

Mohammad Rafiuzzaman
Department of Computer and Information Engineering
Sakarya University, Sakarya, Turkey

ABSTRACT

An important financial subject that has attracted researchers' attention for many years is forecasting stock return. Many researchers have contributed in this area of chaotic forecast in their ways. Among them data mining techniques have been successfully shown to generate high forecasting accuracy of stock price movement. Nowadays, instead of a single aspects of stock market, traders need to use various aspects' forecasting to gain multiple signals and more information about the future of the markets. Aspects of Lyapunov, Entropy and Variance (ALEV) provide an approach for mining large stocks of time series data. This paper proposes a novel method for forecasting chaotic behavior of stock market's opening, high, low and closing price with time series data mining. The outcome of this study tries to help the investors in the stock market to decide the better timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks.

General Terms

Chaotic data mining, time series based data forecasting.

Keywords

Stock market, data mining, chaos data, data forecasting.

1. INTRODUCTION

Financial market is a chaotic, complex, non-stationary, noisy, nonlinear and dynamic system but it does not follow random walk process [3]. Investors have been trying to find a way to predict stock prices and to find the right stocks and right timing to buy or sell. To achieve those objectives, some research used the techniques of fundamental analysis [4] [5] [6], where trading rules are developed based on the information associated with macroeconomics, industry, and company. However, for short and medium-term speculations, fundamental analysis is generally not suitable.

In this paper, we analyzed the seemingly chaotic behavior of apple stock market. Using chaos analysis successful feature extraction can only be conducted if a system is shown to be deterministic and non-linear with some specific chaotic features which are discussed below:

1.1 Confirmation of nonlinearity and determinism of the system:

In order to confirm this, two conditions have to be fulfilled with an assumption.

First, linearity and stochastic property of the system data has to be disproved. This can be efficiently conducted by showing that system data is not in accordance with the null hypothesis. Which means that the time series of a system variable does not behave in a way that Gauss noise does. A Fourier transform of original time series has to be made in order to prove this. Its phases are then randomized in the frequency domain in interval $[0,2\pi]$. An inverse Fourier transform is made after that. Thus the shuffled values are returned to the time domain.

Second, if the values of the original time series correspond significantly to this, so called, surrogate time series, it will signify that the original time series is linear and stochastic. Then chaos analysis can't be applied there. However, if their values differ significantly (usually one order of magnitude), then the null hypothesis is disproved and the system is not both linear and stochastic [1]. Three options are left. But the system can still be:

- Deterministic & Non-linear
- Deterministic & Linear
- Stochastic & Non-linear

Next, an assumption is made for stock market time series data in general. It is assumed that the stock market data is not both linear and deterministic. Although this assumption is generally true, some stock rates may exhibit linear and deterministic behavior, usually those that have very low rate variability. Finally, to determine if the stock market rate exhibits deterministic or stochastic non-linear behavior, a method using attractor [2] reconstruction dimension d and correlation dimension D2 is used. If in some d dimensional description of the system attractor correlation dimension D2 comes into saturation, then the system can be considered deterministic. If too much noise exists in stock market data, then it is possible that the attractor is "masked" and so its correlation dimension never saturates, thus making stock market data system a stochastic system.

As we have discussed before, aspects of Lyapunov, Entropy and Variance (ALEV) provide an approach for mining large stocks of time series data. Three of these aspects are discussed below:

1.2 Lyapunov exponent:

The Lyapunov exponent is a measure for exponential convergence or divergence of a celestial body's trajectory with origins in Orbital Mechanics along a reference orbit. For calculating a Mean Local Lyapunov Exponent (MLLE or λn) of a given sequence in time [15]

$$\lambda = \frac{1}{\eta\tau}\sum_{i=1}^{n} \ln d_i$$

is given where n is the number of elements the sequence vector consists of. The distances d_i between equal located elements are logarithmic and summed up before a mean for the vector's behavior is built. Here τ is a delay factor to optional consider a passage of time.

1.3 Shannon entropy:

The Shannon-Entropy primary served as indicator for the problem of secure transmissions over insecure channels. With the definition Shannon entropy

$$S = -\sum_{i=1}^{n} p_i \,.\log_2(p_i)$$

an information content (E resp. S) is given in [16]. It uses the probability p_i that a vector element lies in a certain subspace i of the event space which has been divided into n subspaces. This indicator has already been in use in Linguistics with relation to SETI (Search for Extraterrestrial Intelligence), as a single description indicator for mining speech signals of aquatic bottlenose dolphins, arboreal squirrel monkeys and of humpback whales [17], [18].

1.4 Variance:

The Lyapunov exponent & Shannon entropy Combined together, offer the opportunity to identify patterns related to the notations stability and information content. As indicator for spreading the variance is used in addition.

For our research work, numerical stock quotes collected from yahoo/finance which are available in structured manner. Information about apple stock market is collected from this website [7] to predict the chaotic opening, high, low and closing price this stock market with time series data mining in this paper.

The rest of the paper is divided in this way: section 2 discusses some previous works done related to our research, section 3 discusses an overview of our system, section 4 show our experimental setup, section 5 shows the results that we got from our system and section 6 evaluates those results. Section 7 concludes the paper with some discussions about the possibility of our proposed system in this paper.

2. RELATED WORK

Many important changes have taken place in the environment of financial markets over the past two decades. The development of powerful communication and trading facilities has enlarged the scope of selection for investors. Forecasting stock return is an important financial subject that has attracted researchers' attention for many years. It involves an assumption that fundamental information publicly available in the past has some predictive relationships to the future stock returns [9]. In order to be able to extract such relationships from the available data, data mining techniques are new techniques that can be used to extract the knowledge from this data.

Data mining techniques have been introduced for prediction of movement sign of stock market index since the results of [10], where LDA, Logit and Probit and Neural network were proposed and compared with parametric models, GMM-Kalman filter. [11] Applied newly and powerful techniques of data mining, SVM and Neural network, to forecast the direction of stock index price based on economic indicators.

To obtain more profits from the stock market, more and more "best" forecasting techniques are used by different traders. Instead of a single method, the traders need to use various forecasting techniques to gain multiple signals and more information about the future of the markets. Authors in [12] collected five different approaches including SVM, Random forecast, neural network, Logit and LDA to predict Indian stock index movement based on economic variable indicators. From the comparison, the SVM outperformed the others in forecasting S&P CNX NIFTY index direction as the model does not require any priori assumptions on data property and its algorithm results global optimal solution which is unique. Huang et al in [13] also forecasted the movement direction of Japanese stock market (NIKKEI 225 index) by various techniques such as SVM, LDA, QDA, NN and the all-in-one combined approach. The SVM approach also gives better predictive capability than other models: LDA, QDA and NN, following the out-performance of the combined model. In the study, they defined the movement of the NIKKEI 225 index based on two main factors including American stock market, S&P 500 index, which is the most influence on the world stock markets including Japanese market, and the currency exchange rate between Japanese Yen and US dollar.

In this paper, we have analyzed the chaotic stock market data for apple stock from yahoo/finance to forecast its following incidences:

- Opening price,
- High price,
- Low price &
- Closing price.

3. SYSTEM OVERVIEW

Analyzing time series is the process of using statistical techniques to model and explain a time-dependent series of data points, while time series forecasting is the process of using a model to generate predictions (forecasts) for future events based on known past events. Examples of time series applications include: capacity planning, inventory replenishment, sales forecasting and future staffing levels. Time series data has a natural temporal ordering, which differs from typical data mining learning applications where each data point is an independent example of the concept to be learned, and the ordering of data points within a data set does not matter.

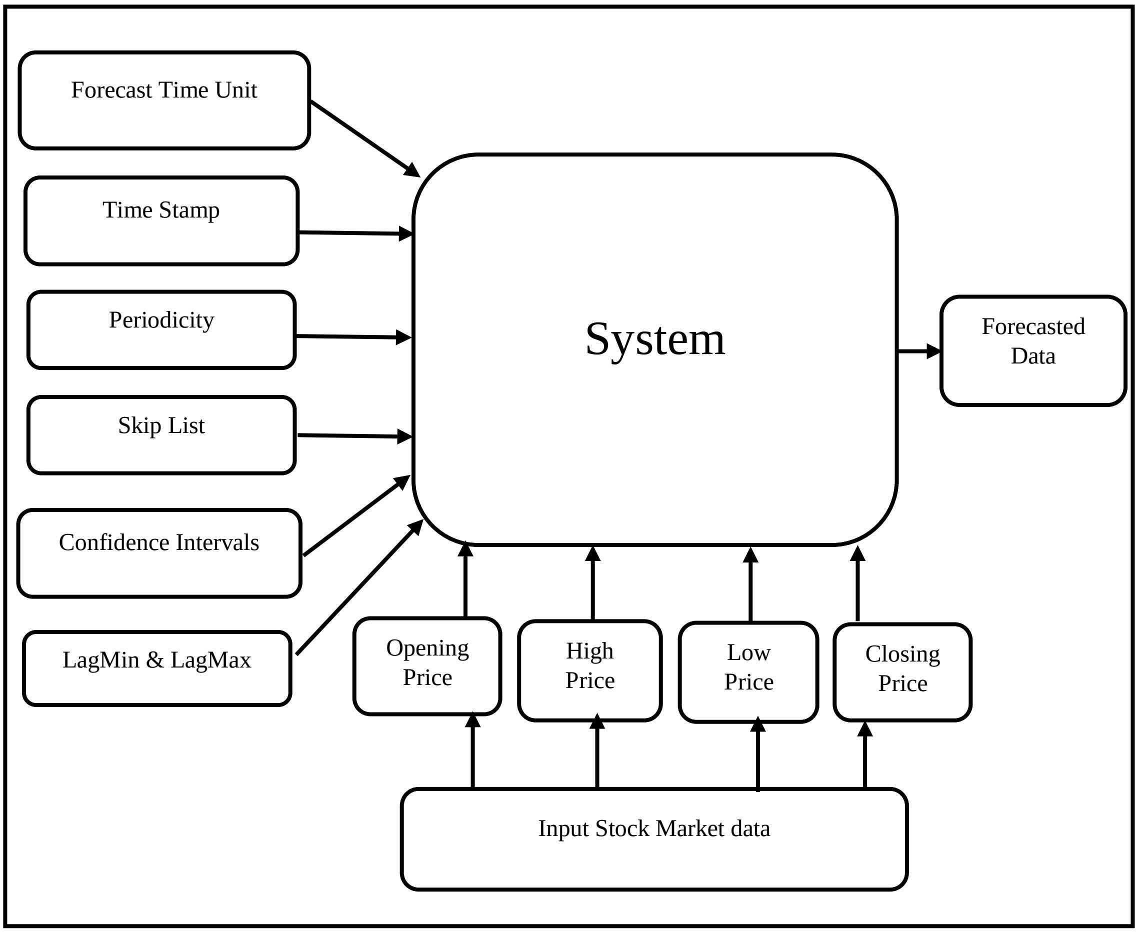


Figure 1: System framework.

For our time series based stock market forecasting we have used "appleStocks2011" data collected from Yahoo finance [7]. Figure 2 shows this dataset with which we have trained our system for future prediction. Due to shortage of space we have shown only some starting and ending part of the data. The general framework is depicted in Figure 1. It contains daily high, low, opening and closing price data, which we are going to forecast in our system for Apple computer stocks from January 3rd to August 10th 2011. Aside from theses parameters the data also includes a date time stamp.

3.1 Experimental setup:

For forecasting the apple stock market "appleStocks2011" following parameters have been used for the system:

- Number of time units to forecast,
- Time stamp,
- Periodicity,
- Skip list,
- Confidence intervals,
- Lag,

Evaluation metrics.

3.2 Setting “Number of Time Units to Forecast”:

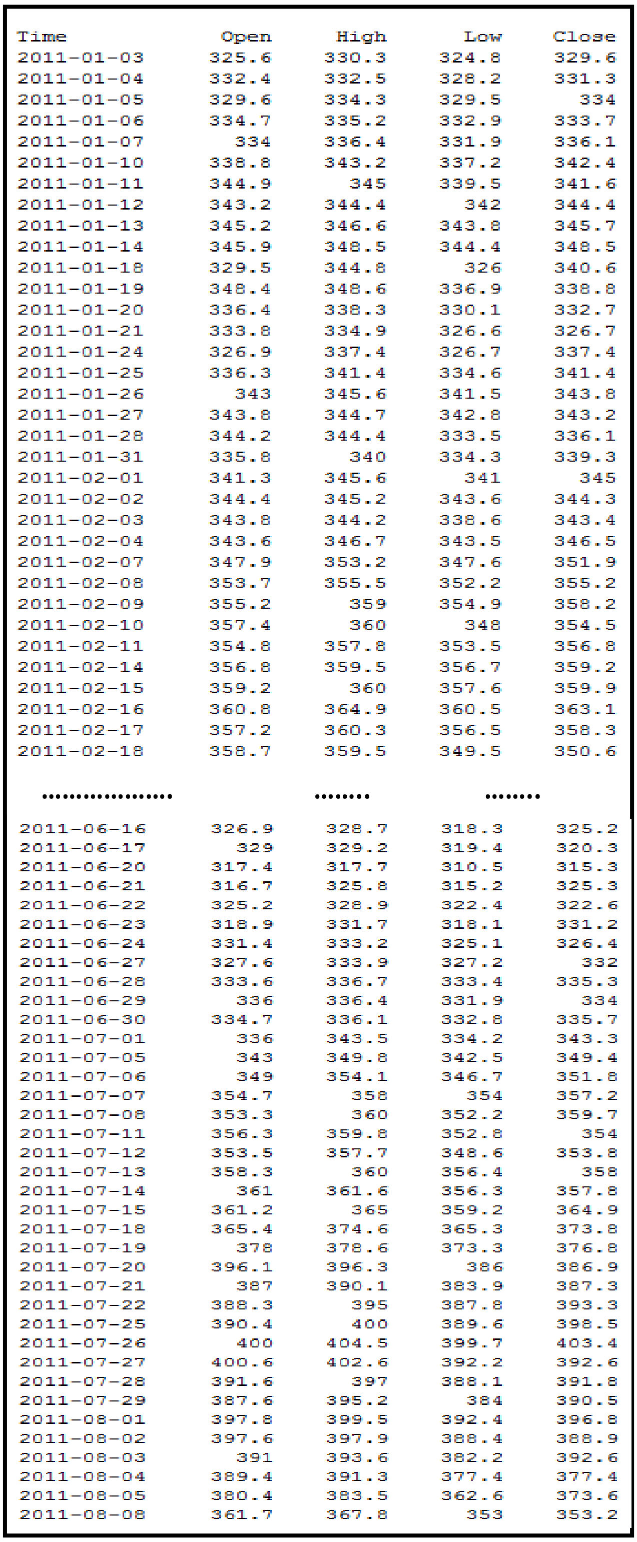
For forecasting "appleStocks2011" *15 days (half month)* has been set for the “*Number of time units to forecast”*. By defining 15 days for this parameter we are setting up our system to forecast 15 days of future stock market data from the last date of the input training dataset.

3.3 Setting “Time Stamp” data:

Date has been set for the *“Time Stamp”* parameter of our system. This allows us to select field in the data which holds the time stamp.

3.4 “Periodicity” setup:

We have selected “*Daily*” as the value for “*Periodicity*” in our system. Periodicity allows the user to specify the Periodicity of the data. So our system proposed in this paper is going to predict the stock price data of "appleStocks2011" on the daily basis.



Time	Open	High	Low	Close
2011-01-03	325.6	330.3	324.8	329.6
2011-01-04	332.4	332.5	328.2	331.3
2011-01-05	329.6	334.3	329.5	334
2011-01-06	334.7	335.2	332.9	333.7
2011-01-07	334	336.4	331.9	336.1
2011-01-10	338.8	343.2	337.2	342.4
2011-01-11	344.9	345	339.5	341.6
2011-01-12	343.2	344.4	342	344.4
2011-01-13	345.2	346.6	343.8	345.7
2011-01-14	345.9	348.5	344.4	348.5
2011-01-18	329.5	344.8	326	340.6
2011-01-19	348.4	348.6	336.9	338.8
2011-01-20	336.4	338.3	330.1	332.7
2011-01-21	333.8	334.9	326.6	326.7
2011-01-24	326.9	337.4	326.7	337.4
2011-01-25	336.3	341.4	334.6	341.4
2011-01-26	343	345.6	341.5	343.8
2011-01-27	343.8	344.7	342.8	343.2
2011-01-28	344.2	344.4	333.5	336.1
2011-01-31	335.8	340	334.3	339.3
2011-02-01	341.3	345.6	341	345
2011-02-02	344.4	345.2	343.6	344.3
2011-02-03	343.8	344.2	338.6	343.4
2011-02-04	343.6	346.7	343.5	346.5
2011-02-07	347.9	353.2	347.6	351.9
2011-02-08	353.7	355.5	352.2	355.2
2011-02-09	355.2	359	354.9	358.2
2011-02-10	357.4	360	348	354.5
2011-02-11	354.8	357.8	353.5	356.8
2011-02-14	356.8	359.5	356.7	359.2
2011-02-15	359.2	360	357.6	359.9
2011-02-16	360.8	364.9	360.5	363.1
2011-02-17	357.2	360.3	356.5	358.3
2011-02-18	358.7	359.5	349.5	350.6
....		....	....	
2011-06-16	326.9	328.7	318.3	325.2
2011-06-17	329	329.2	319.4	320.3
2011-06-20	317.4	317.7	310.5	315.3
2011-06-21	316.7	325.8	315.2	325.3
2011-06-22	325.2	328.9	322.4	322.6
2011-06-23	318.9	331.7	318.1	331.2
2011-06-24	331.4	333.2	325.1	326.4
2011-06-27	327.6	333.9	327.2	332
2011-06-28	333.6	336.7	333.4	335.3
2011-06-29	336	336.4	331.9	334
2011-06-30	334.7	336.1	332.8	335.7
2011-07-01	336	343.5	334.2	343.3
2011-07-05	343	349.8	342.5	349.4
2011-07-06	349	354.1	346.7	351.8
2011-07-07	354.7	358	354	357.2
2011-07-08	353.3	360	352.2	359.7
2011-07-11	356.3	359.8	352.8	354
2011-07-12	353.5	357.7	348.6	353.8
2011-07-13	358.3	360	356.4	358
2011-07-14	361	361.6	356.3	357.8
2011-07-15	361.2	365	359.2	364.9
2011-07-18	365.4	374.6	365.3	373.8
2011-07-19	378	378.6	373.3	376.8
2011-07-20	396.1	396.3	386	386.9
2011-07-21	387	390.1	383.9	387.3
2011-07-22	388.3	395	387.8	393.3
2011-07-25	390.4	400	389.6	398.5
2011-07-26	400	404.5	399.7	403.4
2011-07-27	400.6	402.6	392.2	392.6
2011-07-28	391.6	397	388.1	391.8
2011-07-29	387.6	395.2	384	390.5
2011-08-01	397.8	399.5	392.4	396.8
2011-08-02	397.6	397.9	388.4	388.9
2011-08-03	391	393.6	382.2	392.6
2011-08-04	389.4	391.3	377.4	377.4
2011-08-05	380.4	383.5	362.6	373.6
2011-08-08	361.7	367.8	353	353.2

Figure 2: Training dataset

3.5 Defining "Skip list":

Sometimes it may happen that for a daily trading data of a given stock, the market is closed for trading over the weekend and on public holidays, so these time periods do not count as an increment and the difference. Skip list is used to supply this types of time periods that are not to be considered as increments. For our system we have set our skip list as:

weekend, 2011-01-17@yyyy-MM-dd, 2011-02-21, 2011-04-22, 2011-05-30, 2011-07-04.

Stock price data of these given dates will not be forecasted as those dates are to be considered as holidays for our system.

3.6 Setting up "Confidence intervals":

For the *"Confidence Intervals"* we have set 95% which is also the default value for this parameter. So 95% of the true target values of our data fell within the interval as this parameter is used for the system to compute confidence bounds on the predictions that it makes. The system uses predictions made for the known target values in the training data to set the confidence bounds. The confidence intervals are computed for each step-ahead level independently.

3.7 Lag creation:

Lagged variables are the main mechanism by which the relationship between past and current values of a series can be captured by propositional learning algorithms. It creates a "window" over a time period. The number of lagged variables created determines the size of that window. The basic configuration panel uses the Periodicity setting to set reasonable default values for the number of lagged variables (and hence the window size) created.

For our system we have used *Minimum lag* = 1, which means that a lagged variable will be created that holds target values at time – 1. *Maximum Lag* = 10, which means that a lagged variable will be created that holds target values at time - 10. All time periods between the minimum and maximum lag will be turned into lagged variables.

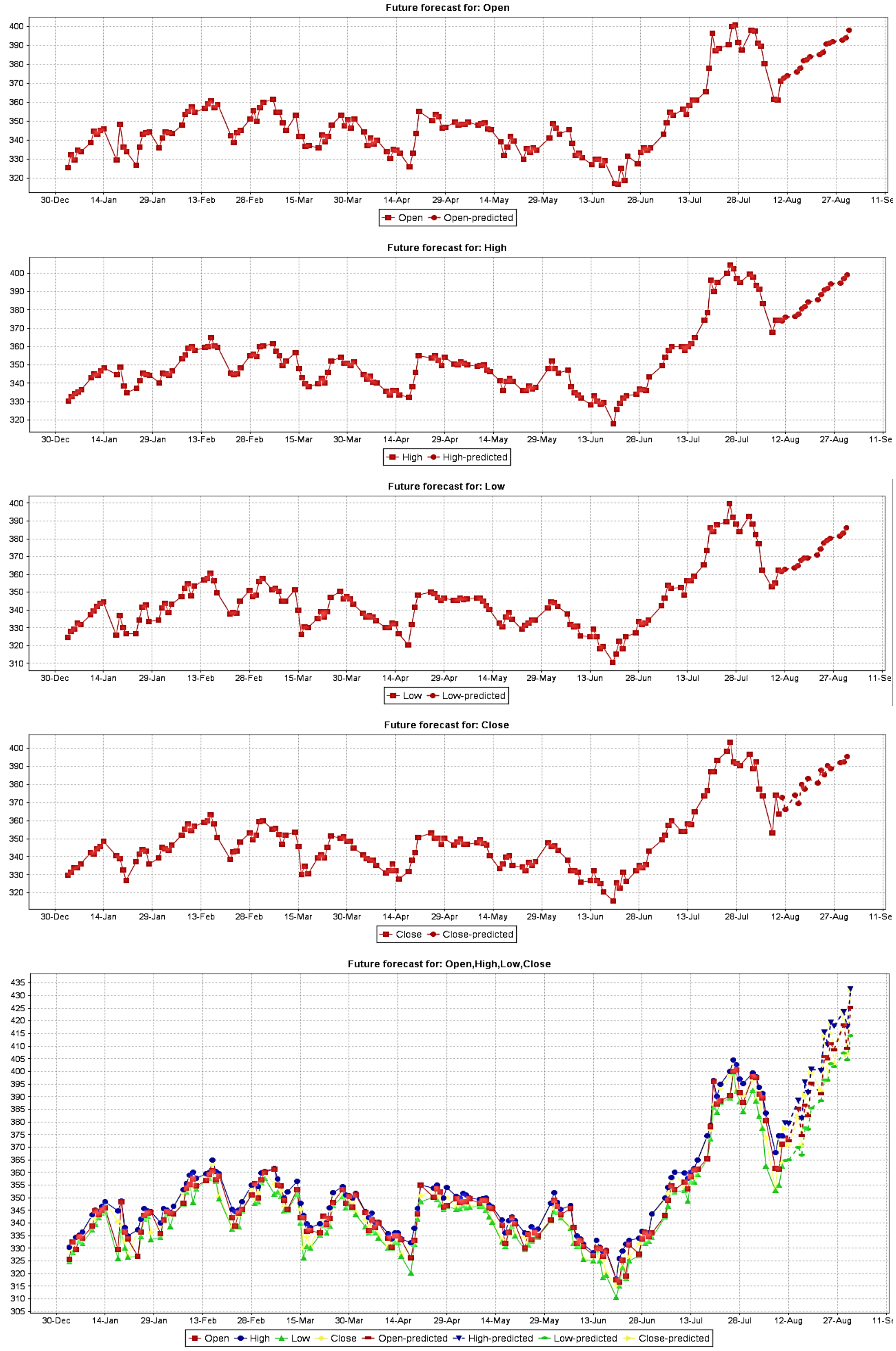


Figure 3: Forecasted data for opening, high, low and closing price of apple stock market. The first four sections show separate forecast and the last sections shows combined forecast for each indices.

4. EXPERIMENTAL RESULTS

Experiment carried out for designing the chaotic stock market forecasting system proposed in this paper is obtained using Weka 3.7.11 [14] and one personal computer with the configurations of: Intel Corei7 2.40 GH Processor, 8 GB of RAM, 1 TB hard disk, 64-bit Windows 8 OS, 1-2 Mbps bandwidth

The final experimental resultant dataset is shown in figure 3. It depicts the forecasted data for the desired indices of apple stock market. As we have defined the *Number of time units to forecast* as 15, so our designed system in this paper forecasts of only for 15 days stock market prices which have been shown on the graph. Table 1 shows the forecasted dataset for 15 days that we have got from our system.

Table 1: Forecasted dataset for 15 days

Time	Open	High	Low	Close
2011-08-11	373.9931	379.6548	364.7039	377.508
2011-08-12	372.5838	379.4417	365.0356	370.7665
2011-08-15	385.3409	388.6493	369.8565	382.1451
2011-08-16	374.6635	381.2983	366.9881	370.6978
2011-08-17	386.6532	395.7506	377.6815	390.5927
2011-08-18	382.3465	391.4903	377.0648	382.1776
2011-08-19	395.2087	400.8912	385.5063	399.8229
2011-08-22	391.14	400.266	388.3961	392.55
2011-08-23	405.8832	415.634	396.6132	413.9773
2011-08-24	405.0284	410.3438	396.4675	396.4561
2011-08-25	411.0001	419.3275	403.0245	418.7212
2011-08-26	408.6367	417.9138	401.9233	402.939
2011-08-29	418.5518	423.6215	407.2569	423.4176
2011-08-30	409.0094	417.564	404.5736	406.8019
2011-08-31	425.2091	432.6686	414.1057	432.3061

5. RESULT EVALUATION

Once the forecaster has been trained on the data, it is then applied to make a forecast at each time point (in order) by stepping through the data. These predictions are collected and summarized, using various metrics, for each future time step forecasted. Evaluated data have been shown in Table 2. For our system we have computed the following two metrics for evaluation:

- Mean absolute error (MAE): The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes [8]. For our system the mean absolute error is given by:

$$\frac{sum\big(abs(predicted - actual)\big)}{N}$$

- Root mean squared error (RMSE): The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions. Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors [8]. For our system RMSE is given by:

$$sqrt\left(\frac{sum\big((Predict - actual)^2\big)}{N}\right)$$

Table 2: Evaluation on training data

Target	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead	6-steps-ahead	7-steps-ahead	8-steps-ahead	9-steps-ahead	10-steps-ahead	11-steps-ahead	12-steps-ahead	13-steps-ahead	14-steps-ahead	15-steps-ahead
Open															
N	143	142	141	140	139	138	137	136	135	134	133	132	131	130	129
Mean absolute error	2.2886	4.7186	5.8591	7.0707	7.7987	8.2577	8.8408	9.3026	9.3732	9.3778	9.5531	9.3214	9.3987	9.4555	9.6452
Root mean squared error	3.4427	5.8448	7.6596	8.9986	9.9735	10.6565	11.2542	11.5705	11.889	11.9602	12.0959	12.0016	12.1308	12.2786	12.4093
High															
N	143	142	141	140	139	138	137	136	135	134	133	132	131	130	129
Mean absolute error	2.4801	4.2786	5.8635	6.794	7.5369	8.1824	8.6616	8.9252	9.112	9.2934	9.2609	9.1385	9.1489	9.2105	9.3681
Root mean squared error	3.2813	5.5279	7.3863	8.658	9.6746	10.3831	11.0173	11.3118	11.6307	11.7358	11.7655	11.7121	11.7808	11.8705	12.0063
Low															
N	143	142	141	140	139	138	137	136	135	134	133	132	131	130	129
Mean absolute error	2.9709	5.1524	6.4803	7.5107	8.1454	8.6724	9.2207	9.4746	9.5484	9.6192	9.658	9.7287	9.605	9.5885	9.8425
Root mean squared error	3.9588	6.5385	8.0728	9.2805	10.1946	10.9637	11.5907	12.0136	12.243	12.289	12.3658	12.3033	12.3801	12.4956	12.7176
Close															
N	143	142	141	140	139	138	137	136	135	134	133	132	131	130	129
Mean absolute error	3.7691	5.1063	6.6556	7.4207	7.9725	8.5111	8.9916	9.1525	9.3078	9.367	9.4317	9.3318	9.4828	9.4221	9.6745
Root mean squared error	4.7011	6.5293	8.2354	9.2627	10.151	10.7331	11.4113	11.7558	11.9799	11.9658	11.9699	11.9006	12.1092	12.1443	12.4374

6. CONCLUSION

Determining the Stock market forecasts has always been challenging work for business analysts. In this paper, we attempted to make use of these huge chaotic in nature data to predict the stock market indices. If we combine both these chaotic data and numeric time series analysis the accuracy in predictions can be achieved. Investors can use this prediction model to take trading decision by observing market behavior. Enhancements of this system is focused to help in improving more accurate predictability in stock market regardless how chaotic the stock market data can be.

7. REFERENCES

[1] Kaplan, D., Glass, L., "Understanding Nonlinear Dynamics", Springer-Verlag, 1995.

[2] Schuster, H. G., "Deterministic Chaos: An Introduction",Physik-Verlag GmbH, 1984.

[3] Lo, A.W., & MacKinlay, A.C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test, Review of Financial Studies 1, 41-66.

[4] Wu, M.C., Lin, S.Y., and Lin, C.H., (2006) “An effective application of decision tree to stock trading”, Expert Systems with Applications, 31, pp. 270-274.

[5] Al-Debie, M., Walker, M. (1999). “Fundamental information analysis: An extension and UK evidence”, Journal of Accounting Research, 31(3), pp. 261–280.

[6] Lev, B., Thiagarajan, R. (1993). “Fundamental information analysis”, Journal of Accounting Research, 31(2), 190–215.

[7] http://finance.yahoo.com/q/hp?s=AAPL&a=00&b=3&c=2011&d=07&e=10&f=2011&g=d

[8] T. Chai and R. R. Draxler. “Root mean square error (RMSE) or mean absolute error (MAE)?” Geosci. Model Dev. Discuss., 7, 1525–1534, 2014 www.geosci-model-dev discuss.net/7/1525/2014/ doi: 10.5194/gmdd-7-1525-2014.

[9] Enke, D., Thawornwong, S. (2005) “The use of data mining and neural networks for forecasting stock market returns”, Expert Systems with Applications, 29, pp. 927-940.

[10] Chen, An-Sing, Daouk, Hazem & Leung, Mark T. Application of Neural Networksto an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index. [Online] Available: http://ssrn.com/abstract=237038 or DOI: 10.2139/ssrn.237038 (July 2001).

[11] Kim, K.J. (2003). Financial time series forecasting using support vector machines, Neuralcomputing, 55, 307-319.

[12] Kumar, Manish and Thenmozhi, M. Forecasting Stock IndexMovement: A Comparisonof Support Vector Machines and Random Forest, IndianInstitute of Capital Markets 9th Capital Markets Conference Paper. [Online] Available: http://ssrn.com/abstract=876544 (February 06, 2006).

[13] Huang, W., Nakamori, Y. & Wang, S.Y. (2005). Forecasting stock market movement direction with support vector machine, Journal of Computers & Operational Research, pp. 2513-2522.

[14] Mark Hall, Lan written, Eibe Frank. Data mining: Practical machine learning tools and techniques. ISBN: 978-0-12-374856-0. January 2011.

[15] H. Troger, A. Steindl, Introduction into Bifurcation Theory - Lecture Notes Marie Curie, Vienna University of Technology, Release May 5th, Austria, 2007.

[16] H.G. Schuster, Deterministic Chaos, An Introduction, Verlag Chemie, Weinheim, Germany, 1988.

[17] B. McCowan, L.R. Doyle, S.F. Hanser, “Using Information Theory to Assess the Diversity, Complexity and Development of Communicative Repertoires”, Journal of Comparative Psychology, Vol. 116, No. 2, pp. 16-172, American Psychological Association (APA), Washington D.C., USA, 2002.

[18] L.R. Doyle, “Talking With Your Mouth Full: The Feeding Calls of the Humpback Whale”, Space.com, Imaginova Corp., New York, USA, Jan. 26th 2007.