19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

Predicting SET50 Stock Prices Using CARIMA (Cross Correlation ARIMA)

Sornpon Wichaidit
Dept. of Computer Engineering
Faculty of Engineering
King Mongkut's Institute
of Technology Ladkrabang, Thailand
Email: s6601133@kmitl.ac.th

Surin Kittitornkun
Dept. of Computer Engineering
Faculty of Engineering
King Mongkut's Institute
of Technology Ladkrabang, Thailand
Email: kksurin@staff.kmitl.ac.th

***Abstract*—Investing in stocks is one of the most popular approaches for money investment. This paper aims to predict short-term stock prices of SET50 of Stock Exchange of Thailand (SET). The proposed method is called CARIMA (Cross Correlation Autoregressive Integrated Moving Average. The basic idea of CARIMA is to find the most highly correlated stock to predict the target one in addition to ARIMA predicted price. The results of CARIMA model yield better price trends (measured by 10-day correlation coefficient) while % MAEs (Mean Absolute Errors) are quite similar with those of ARIMA.**

Keywords - Stock, Prediction, ARIMA, Correlation, Time Series

I. Introduction

Today, Stock Trading is a very popular method of short-term investment. It is known that the time series of stock prices fluctuate greatly but not totally random. Therefore, the investigation of factors that the financial data has long attracted researchers from various different areas, such as: mathematicians, economists and more recently, computer scientists. With the advantages of information technique, a huge amount of stock trading data can be collected easily. The current focus is to design a method of extracting useful information from the collected data. Therefore, data mining has drawn considerable attentions from the young generations of investor in order to predict the changes or discover the patterns of the stock prices.

There are two analytical methods for stock trading: 1.Fundamental [3]; 2.Technical Analysis [4]. Fundamental analysis focuses on the central factors of a company. Such as financial statements, Earnings per Share (EPS), Return on Assets (ROA), Price-Earnings Ratio (P/E Ratio), Return on Equity (ROE), price and book value (P/BV). On the other hand, Technical analyses use the recent historical data of an individual stock, including opening, highest, lowest and closing prices and the volume of stock, to predict future stock price movements. There are many indicators that can be created from the data. Such as Moving Average Convergence Divergence (MACD), Average Directional Index (ADX), Exponential Moving Average (EMA), Double Exponential Moving Average (DEMA), and Relative Strength Index (RSI).

This paper proposes a simple yet effective method for predicting short-term movements of a target stock based on ARIMA Model with another stock with highest lead/lag correlation with the target one. The correlation coefficient can be effectively incorporated with the original ARIMA thus called Cross Correlation ARIMA (CARIMA). The performance of CARIMA is evaluated based on SET (Stock Exchange of Thailand) 50 dataset in terms of trend similarity and %Mean Absolute Error(MAE).

The rest of this paper is organized as follows. Section 2 presents the existing related work. Section 3 explains the theory of ARIMA and the proposed CARIMA method. Section 4 describes dataset and results. Eventually, Section 5 provides the conclusion including the future work.

II. Related Work

There is much work based on time series technique for analyzing and predicting trends of the stock market around the world. All of these researches have been using various factors that have impacts on the stock market as input to predict the trends.

C. Fonseka and L. Liyanage [5] developed an algorithm for predicting an individual stock from the Australian Stock Exchange(ASX) with correlation in 2008. Meanwhile, S. Chaigusin, et al [6], used a feedforward backpropagation neural network to predict the movement of SET index. The inputs of the models consist of seven nodes including the Dow Jones index, Nikkei index, Hang Seng index, Gold prices, Minimum Loan Rate (MLR), and the exchange rates of the Thai Baht and the US dollar. In 2012, A. Srisawat [7] applied an association rule mining technique for discovering relationships between individual stocks from SET. Recently, W. Weiqing and Y. Lav [8] used ARIMA Model to study volatile characteristics in the US dollar index itself in 2013. They used measurement statistical model to fit the ARIMA Model to predict US Dollar index movement for one month.

As mentioned above, all of them used various data mining techniques to predict the stock or index. Very few of them have used relationships between stocks to help in the stock price prediction process. As a result, we are going to take advantages of this gap to propose our method.

978-1-4673-7825-3/15/$31.00 ©2015 IEEE

```
Augmented Dickey-Fuller Test

data:  stock[, "CLOSE"]
Dickey-Fuller = -2.2875, Lag order = 9, p-value = 0.4566
alternative hypothesis: stationary
```

Fig. 1. Unit Root Test of ADVANC

III. Our Proposed Method

The method used to develop ARIMA model for stock price prediction with combining correlation of the other stock (the stock that has lead/lag correlation to the focused stock) can be explained in subsections below

A. Unit Root Test

We make use of the Augmented DickeyFuller test(ADF) as the unit root test for the dataset whether it is stationary or not. So we test it on the closing price of an individual stock. The result of this test shown in Fig. 1 that the data is stationary.

B. ARIMA model

Autoregressive integrated moving average (ARIMA) model is generalisation of an autoregressive moving average (ARMA) model. ARIMA include differencing I to ARMA AR(autoregressive)+I(integrated)+MA(moving average). The equation of ARIMA Model is.

$$X_t = m + \sum_{t=1}^{p} \Theta_{i-p} \Delta_d X'_{t-i} - \sum_{t=1}^{q} \ominus_{i-q} \epsilon_{t-i} \qquad (1)$$

Where m is a constant, Θ and $\ominus$ are parameter of autoregressive and moving average parts, ϵ is error(white noise), i is lead/lag time variable (days), Δ is the difference, X' is a closing price of an individual stock, t is a day variable. p , d and q are orders of autoregressive, difference and moving average parts, respectively.

The ARIMA predicted stock price X_t is parameterized as

$$X_t = ARIMA(p, d, q) \qquad (2)$$

where AR: p is order of the autoregressive part, I: d is the degree of first differencing involved, and MA: q is order of the moving average part.

To select the best model of ARIMA(p,q,d), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be evaluated to estimate the quality as listed in Table 1. In this case, ARIMA(0,1,0) model is selected to predict ADVANC because of the smallest values of AIC and BIC. The model returns the smallest AIC of 4582.58 and relatively small BIC of 4587.26. in Fig. 2 show that there is no correlation between predicted ADVANC and the actual price movement of ADVANC.

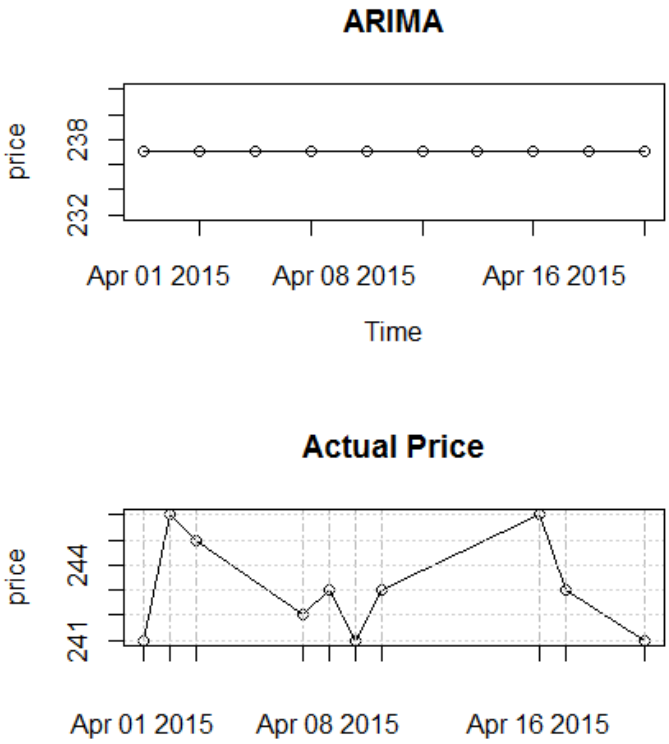

Fig. 2. ADVANC: ARIMA Predicted Price, X_t (Top) vs Actual Price, X'_t (Bottom)

TABLE I
ARIMA Model Selection of ADVANC based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

ARIMA	AIC	SC
p=1,d=0,q=0	4594.27	4608.31
p=0,d=1,q=0	4582.58	4587.26
p=0,d=0,q=1	6846.23	6860.27
p=1,d=1,q=0	4584.58	4593.94
p=1,d=0,q=1	4596.27	4614.99
p=0,d=1,q=1	4584.58	4593.94
p=1,d=1,q=1	4586.58	4600.62

C. Cross Correlation with Lead/Lag i days

We assume the closing price is the one that shows the realest price movement behaviour rather than High Low Open price. Cross Correlation analysis of closing prices is a measure of the interrelationship between two stocks prices with a function of lead/lag time of one company relative to another. The correlation coefficient can be calculated with respect to lead/lag time i days as defined in Eq.(3).

$$\rho_{x_t y_t}(i) = \sum_{t=1}^{796} [(X_t - u_X)(Y_{t-i} - u_Y)]/\delta_X \delta_Y \qquad (3)$$

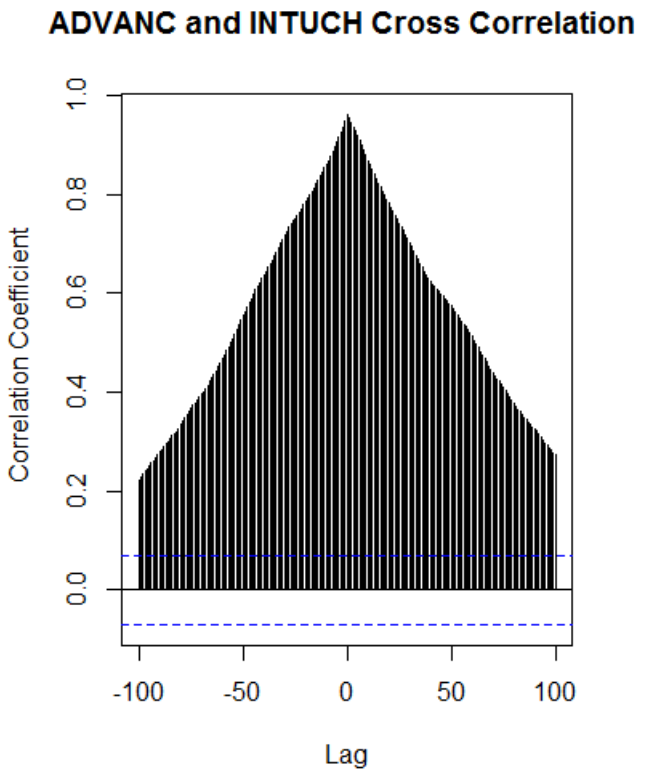

Fig. 3. Cross Correlation Coefficient $\rho_{X_t Y_t}(i)$ with Lead/Lag i between BAY and BH

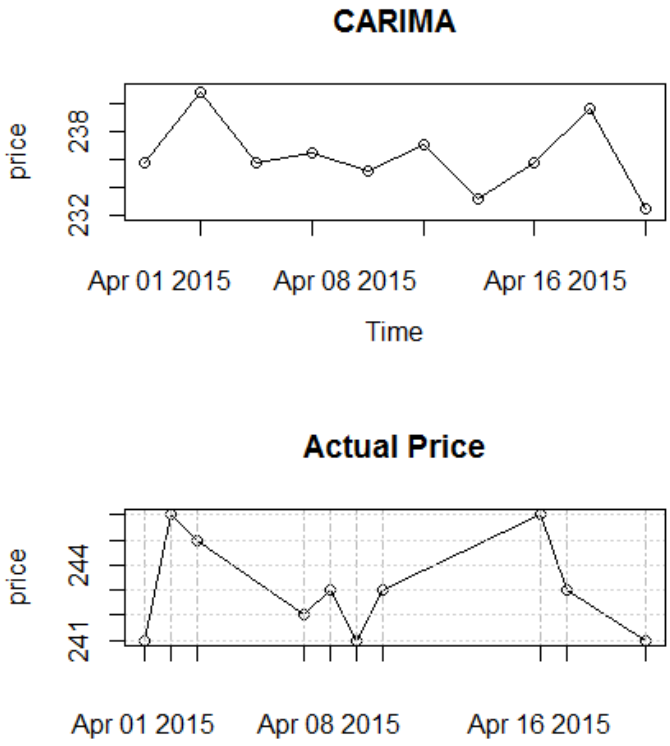

Fig. 4. ADVANC: CARIMA Predicted Price $\hat{X}_t$ with Lead, i = 10 from INTUCH (Top) vs Actual Price, X'_t (Bottom)

where t is the day variable that starts from January 1st, 2012 to March 31st, 2015, i is the lead/lag time variable in days, δ_X is a standard deviation of stock X, δ_Y is a standard deviation of stock Y, u_X is a mean of stock X and u_Y is a mean of stock Y.

We choose the highly correlated coefficient whose $|\rho_{X_t Y_t}(i)| \geq 0.8$ at time lag $t - i$. It means that stock Y_t is leading stock X_t by i days. As a result, stock Y_{t-i} can improve ARIMA Model to predict stock X_t price movement. The example of cross correlation between ADVANC and INTUCH shown in Fig. 3.

D. CARIMA: Cross Correlation ARIMA

After we get the results of ARIMA price X_t predicted up to 10 days from April 1-18, 2015. The highly correlated stock Y whose $|\rho_{X_t Y_t}(i)| \geq 0.8$ that leads stock X for i days is picked. By shifting back stock Y to day $t - i$, the Rate of Change (ROC) from Y_{t-i} is applied with respect to X_t. Finally, the CARIMA predicted $\hat{X}_t$ is a product of $\rho_{X_t Y_t}(i)$, X_t, and ROC of Y_{t-i} as shown in Eq.(4).

$$\hat{X}_t = \rho_{X_t Y_t}(i) \times X_t \times (1 + \frac{Y_{t-i+1} - Y_{t-i}}{Y_{t-i}}) \qquad (4)$$

where $\rho_{X_t Y_t}(i)$ is the correlation coefficient between stock x and y, X is the ARIMA predicted price, Y is the most correlated stock, t is day variable and i is the lead time variable in days.

IV. Experiment and Results

In this section, there are 2 main activities. Firstly, we explain the details of SET50 dataset. Secondly, we present and discuss the results.

A. SET50 Dataset

The SET50 dataset is collected from Stock Exchange Thailand (SET). It consists of the 50 most valuable companies (stocks) in Thailand. We use their closing prices to find the correlation coefficient of each individual stock to the others with the lead and lag time. The dataset starts from January 1st, 2012 ($t = 1$) to March 31, 2015 ($t = 796$).

B. Results and Discussion

The stock names, 10-day correlation coefficients, and %MAEs between CARIMA $\hat{X}_t$ and actual prices X'_t vs. ARIMA X_t and actual prices X'_t are listed in Table 2. NA in $\rho_{X_t X'_t}$ column means Not Available because the ARIMA result is a straight line resulting divided by zero. The price trends of CARIMA are closer to the actual prices than those of ARIMA alone (measured by 10-day correlation coefficients), although the %MAEs (Mean Absolute Errors) are quite similar. %MAEs of CARIMA is not better than %MAEs of ARIMA.

TABLE II
10-DAY CORRELATION COEFFICIENTS AND %MAE OF CARIMA, $\hat{X}_t$ AND ACTUAL STOCK PRICES,X'_t VS. ARIMA,X_t AND ACTUAL STOCK PRICES,X'_t, (NA: NOT AVAILABLE)

Stock	$\rho_{\hat{X}_t X'_t}$	$\rho_{X_t X'_t}$	$\%\text{MAE}_{\hat{X}_t X'_t}$	$\%\text{MAE}_{X_t X'_t}$
JAS	0.58	NA	2.29	2.13
ADVANC	0.40	NA	2.85	2.51
RATCH	0.35	NA	0.84	0.51
GLOBAL	0.33	-0.39	1.53	1.32
IVL	0.31	NA	2.07	2.01
THCOM	0.30	NA	3.96	2.23
BANPU	0.27	NA	3.65	3.20
TCAP	0.15	NA	1.86	1.46
BCP	0.14	NA	4.61	4.78
EGCO	0.10	-0.87	1.53	1.51
KKP	0.08	-0.15	0.89	0.74
PTTEP	0.07	-0.56	6.82	6.36
CENTEL	0.04	NA	3.94	4.10
TTW	0.03	-0.88	1.38	1.41
BAY	-0.09	0.32	4.62	5.17
BH	-0.11	-0.75	3.67	1.74
GLOW	-0.17	-0.74	2.64	2.46
TRUE	-0.20	0.80	3.84	4.07
BDMS	-0.22	-0.39	2.62	2.42
CPN	-0.33	NA	3.20	3.61
SCB	-0.38	-0.51	1.77	1.33
PS	-0.48	-0.71	2.62	2.31

Because the main purpose of ARIMA is to predict price movement as close to the actual price as possible. But CARIMA incorporates the cross correlation coefficient with ARIMA in order to improve the correlation of the price movements. For example, CARIMA predicted prices of ADVANC predicted by INTUCH can be plotted in Fig. 4 and compared with those of ARIMA in Fig. 2. We can see that those of CARIMA are more correlated while the predicted price from ARIMA is a straight line thus not correlated to the actual price. Although %MAE are slightly higher than ARIMA. CARIMA can yield better price trends as $\rho_{\hat{X}_t X'_t}$ are better than $\rho_{X_t X'_t}$.

Since the dataset is not big enough, only 50 individual stocks, it is hard to find the leading/lagging stocks with high correlation coefficients to make the better prediction. Moreover, most of the stocks have the highest/ least correlation coefficients in lag i = 0 that can be predicable that stock always move up and down together in the same day. Its not reasonable to use that to help ARIMA to predict individual stock price movement.

V. CONCLUSION

This paper aims to incorporate cross correlation coefficient with ARIMA to predict short-term (daily) SET50 stock price movement. The proposed method is called CARIMA to predict stock A using the most highly correlated stock B within 10-day lead. The empirical results obtained and compared CARIMA predicted prices $\hat{X}_t$ and ARIMA predicted X_t against the actual prices X'_t. In terms of performance evaluation, CARIMA prices are more correlated to the actual prices and their %MAEs are similar to those of ARIMA. For future work, we intend to apply two or more stocks as predictors to CARIMA for each individual SET50 stock. It is expected that they can improve the prediction performance even more.

REFERENCES

[1] P. S.P. Cowpertwait and A. W. Metcalfe, *Introductory Time Series with R*, Springer, 2009, pp.137-155

[2] C. M.Conover and D. R. Peterson, *The Lead-Lag Relationship between the Option and Stock Markets prior to Substantial Earnings Surprises and the Effect of Securities Regulation*. Journal of Financial and Strategic Decisions, Spring 1999, Vol. 12 No. 1

[3] B. Graham , J. Zweig and W. Premchaiswadi, *The Intelligent Investor: The Definitive Book on Value Investing. A Book of Practical Counsel*, Collins Business Essentials, 2006

[4] S. Nison, *Beyond Candlesticks: New Japanese Charting Techniques Revealed*, Wiley Finance, 2009

[5] C. Fonseka, and L. Liyanage, *A Data mining algorithm to analyse stock market data using lagged correlation*, 4th International Conference on Information and Automation for Sustainability, 2008, pp.163 - 166

[6] S. Chaigusin, C. Chirathamjaree, and J. Clayden, *A Data mining algorithm to analyse stock market data using lagged correlation*, Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation, 2008, pp. 670 - 673

[7] A. Srisawat, *Discovery Stock Trading Patterns: A Case Study of Thai Stock Market*, International Journal of Intelligent Information Processing, 2012, Vol. 3 Issue 1, pp.1-9

[8] W. Weiqing and Y. Lav, *A Study of the USDX Based on ARIMA Model A Correlation analysis between the USDX and the Shanghai index*. Consumer Electronics, Communications and Networks (CECNet), 3rd, 2013

[9] J. A. Ryan, J. M. Ulrich, and W. Thielen, *quantmod: Quantitative Financial Modelling Framework*, Version: 0.4-5, https://cran.r-project.org/web/packages/quantmod/index.html, 2015

[10] R.J. Hyndman, *forecast: Forecasting Functions for Time Series and Linear Models*, Version: 6.1, https://cran.r-project.org/web/packages/forecast/index.html, 2015