Predicting the price of Bitcoin using hybrid ARIMA and machine learning

Dinh-Thuan Nguyen, Huu-Vinh Le

University of Information Technology, VNU-HCM, Ho Chi Minh, Vietnam
thuannd@uit.edu.vn, vinhlh.10@grad.uit.edu.vn

Abstract. Bitcoin is one of the most popular cryptocurrencies in the world, has attracted broad interests from researchers in recent years. In this work, Auto-regressive Integrate Moving Average (ARIMA) model and machine learning algorithms will be implemented to predict the closing price of Bitcoin the next day. After that, we present hybrid methods between ARIMA and machine learning to improve prediction of Bitcoin price. Experiment results showed that hybrid methods have improved accuracy of predicting through RMSE and MAPE.

Keywords: Bitcoin prediction, ARIMA, machine learning, hybrid model.

1 Introduction

In October 2008, Bitcoin was firstly introduced by Satoshi Nakamoto in the report "Bitcoin: A Peer-to-Peer Electronic Cash System" [1]. In 2009, Nakamoto has released a software that created Bitcoin and had a large community using Bitcoin around the world so far. In recent years, Bitcoin trading has exploded into trading volume as well as the amount of money that spent on investment. They have boosted value of Bitcoin. It is estimated that the global value of Bitcoin comes to 72.1 trillion USD at the beginning of May 2019. Many investors are willing to spend large amounts of money on Bitcoin and hope to make a return. Therefore, Bitcoin price forecasting is one of the hot topics in recent years.

There are two main forecasting methods for predicting Bitcoin price. The first is that based on time series of Bitcoin price. The second is that found relationship between the price of Bitcoin and other indicators such as stock price, oil price, gold price,... In this study, we focus on method using time series. ARIMA model and machine learning algorithms such as Feedforward Neural Network (FFNN), Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), Support Vector Regression (SVR) will be implemented. Then, hybrid models are proposed to improve prediction. They are that combining ARIMA model and machine learning algorithms, combining models based on fluctuation interval.

2 Related Works

Bitcoin price prediction can bring returns to investors. Therefore, studies in Bitcoin price prediction are done increasingly. Devavrat Shah [2] presented the Bayes regression method to predict the price changes of Bitcoin in every 10 seconds. Based on this method, the author has come up with a simple strategy for trading Bitcoin. Then, the author has made 2872 transactions, profit achieved in 50 days about 89%.

João Almeida [3] applied an artificial neural network to predict the next day's Bitcoin's trend based on the price and volume of Bitcoin transactions in the previous days. Neural network models are installed and experimented with Theano library and MATLAB tool. The experiment showed that adding Bitcoin trade Volume as input data did not result in an increased performance.

Isaac Madan [4] collected Bitcoin dataset including 25 features and selected 16 features to predict the daily price change trending of Bitcoin. Prediction accuracy is up to 98.7%. At the same time, the time series of Bitcoin price collected in every 10 seconds. This dataset was trained for a random forest algorithm and a linear model to predict up-down Bitcoin price movement every 10 minutes. The prediction accuracy is about 50 - 55%.

Huisu Jang [5] proposed using the Bayesian neural network to analyze the volatility of Bitcoin price. The author also selected some characteristics Blockchain information that related to the supply and demand of Bitcoin to improve prediction. The author experimented with Bayesian neural network and some other linear and non-linear methods on Bitcoin price dataset. The result of experiment indicated that the Bayesian neural network is good for forecasting and can describes the big volatility of Bitcoin price.

Sean McNally [6] used LSTM model to forecast Bitcoin price. The experiment results showed that LSTM is better than traditional methods like ARIMA. The author also compared the performance of LSTM on CPU and GPU. It showed that GPU provided 67.7% higher performance than CPU.

Alex Greaves [7] used Bitcoin transaction graph that consists of blockchain information to forecast up-down Bitcoin price movement. In the study of the author, neural network is better classification than baseline model, linear regression, SVM.

Steve Y. Yang's research [8] focused on analyzing the relationship between the number of Bitcoin transactions and its price changes. The author has built a Bitcoin trading network with complex measurements that related to profitability and price volatility. In this study, author believed that this network is reliable for predicting Bitcoin price. Ferdiansyah [9] was Bitcoin-USD trading using SVM model to detect the current day's trend. Author concluded SVM and prediction can be used to forecast current day's trend of Bitcoin on the market. In the next study, neural networks will be used to improve accuracy of predicting.

The most of current studies concentrated on predicted trend of Bitcoin by implemented individual models. We will predict Bitcoin price in the next day with time series forecasting models and propose hybrid models to combine individual models to improve prediction.

3 Hybrid Methodology

3.1 Hybrid model based on ARIMA and machine learning

G. Peter Zhang [10] proposed to combine ARIMA and ANN for time series forecasting. The idea of this hybrid model is based on a combination of linear and non-linear components in time series. These two components are represented by the equation:

$$y_t = L_t + N_t$$

where y_t is time series, L_t is a linear component, N_t is a non-linear component.

Firstly, ARIMA model is used to capture the linear component, then the residuals from the linear model will contain only the nonlinear relationship. The residuals e_t at time t from the linear model is defined by:

$$e_t = y_t - \widehat{L_t}$$

where $\widehat{L_t}$ is the predicted value of the ARIMA model at time t.

The residuals can be modeled by using ANNs to discover nonlinear relationships. With n input nodes, the ANN model for the residuals will be:

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t$$

where f is a nonlinear function determined by the neural network and e_t is the random error. Finally the combined prediction will be:

$$\widehat{y_t} = \widehat{L_t} + \widehat{N_t}$$

where $\widehat{N_t}$ represents the prediction from ANN.

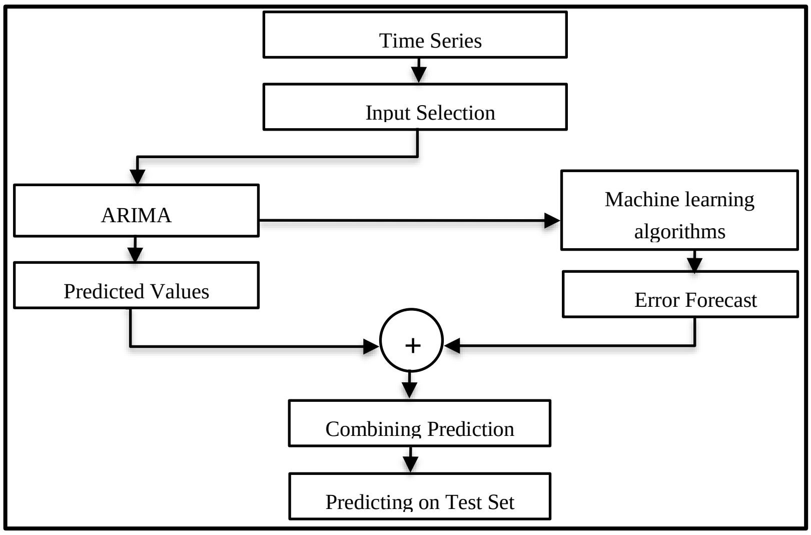

Figure 1. Hybrid ARIMA and machine learning

With Zhang's proposal, combining models are expanded with ARIMA and machine learning algorithms such as FFNN, CNN, LSTM and SVR. These combining methods promise a significant improvement in prediction instead of just combining ARIMA and traditional neural network.

3.2 Hybrid model based on fluctuation interval

One of the biggest challenges in predicting time series is its fluctuation. There has not yet been a really effective model to forecast on highly volatile time series. In this study, we propose a hybrid model that combines to forecast based on fluctuation interval. We considered that each model will be suitable to predict on specific fluctuations interval. For example, ARIMA model has good predictability on stationarity time series, that means it is less volatility. Meanwhile, machine learning algorithms and SVR will be able to capture nonlinear component, be large fluctuations. Therefore, this combination is very promising to improve prediction.

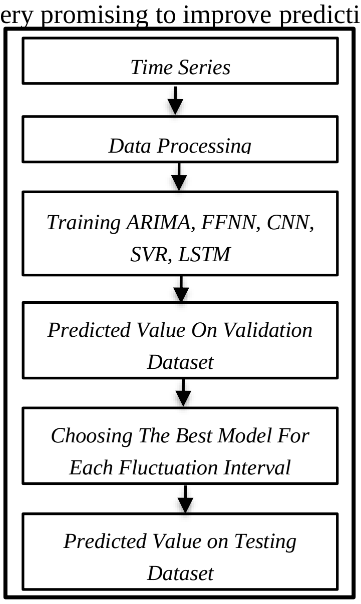
Time Series
Data Processing
Training ARIMA, FFNN, CNN, SVR, LSTM
Predicted Value On Validation Dataset
Choosing The Best Model For Each Fluctuation Interval
Predicted Value on Testing Dataset

Figure 2. Hybrid based on fluctuation interval

Firstly, we compute fluctuation of two consecutive time points as percentage of difference of 2 points by formula:

$$\text{fluctuation}(y_t, y_{t+1}) = \left|\frac{y_t - y_{t+1}}{y_t}\right| * 100$$

where y_t is the first time point, y_{t+1} is the second time point.

On the experiment dataset, determining the minimum and maximum fluctuation. And the frequency of volatility as a basis for dividing the volatility range. Of course, there are some outlier volatility with low frequency, we can ignore them and adjust the minimum and maximum volatility.

With a single series, the number of intervals can be be determined from [11]:

$$k = 1 + 3.3\log_{10} N$$

where N is the sample size.

We can apply above formula for dividing fluctuation interval with R is range of maximum and minimum fluctuation and N is the number of time points that is computed fluctuation.

Finally, we define the best model to forecast for each interval and make prediction. The procedure of the hybrid model concludes some steps below:

Step 1. Compute fluctuation between two consecutive time points.
Step 2. Find maximum and minimum of fluctuation in time series.
Step 3. Define fluctuation interval of range of fluctuation.
Step 4. Training ARIMA, FFNN, CNN, LSTM, SVR on training dataset with defining fluctuation(y_t) = fluctuation(y_{t-2},y_{t-1})
Step 5. Predicting values on validation dataset with trained models.
Step 6. Defining a rule of choosing models and using the best model for each interval to predict on test dataset.

4 Experiment

4.1 Closing price of Bitcoin dataset

Forecasting models will be implemented on the closing price dataset of Bitcoin, is collected from October 01, 2013 to June 08, 2019, including 2070 days on CoinDesk website. Based on the closing price chart of Bitcoin, it can be seen that the price of Bitcoin has big changes, especially in 2017 and continues to have fluctuations in 2018 and 2019.

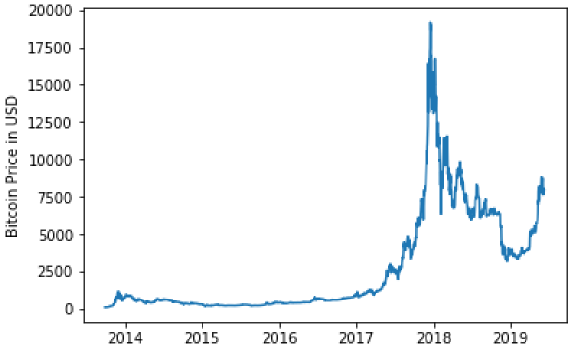

Figure 3. The closing price of Bitcoin chart

4.2 Software used

We used python 3.6 with Anaconda tools to manage libraries. A statistical library - Statsmodels used to build ARIMA model. Neural network models are installed with tensorflow library, Keras to build models quickly. Pandas library is used to process time series data, NumPy to calculate matrices/vectors and store training, validation and test data. Finally, the matplotlib library is used to draw illustrative charts.

4.3 Split time series into training, validation, tesing dataset

In the experiment with individual models, the closing price of Bitoin dataset will be divided into 2 parts: 90% of the time points used for training, 10% of the remaining time points will be used for model testing. The experiment of hybrid ARIMA with machine learning and regression, the dataset is divided into 3 parts: 80% of the time points used for training, 10% used for validation, the remaining 10% used for testing.

In the experiment of hybrid based fluctuation interval, the dataset is divided into sections to match nested cross validation. The dataset is divided into 10 equal parts. During the first training session, 50% was used for training, the next 10% for validation, the next 10% for model testing. Just like that, until the last training session, 80% of the training is used, 10% for validation, the remaining 10% for testing.

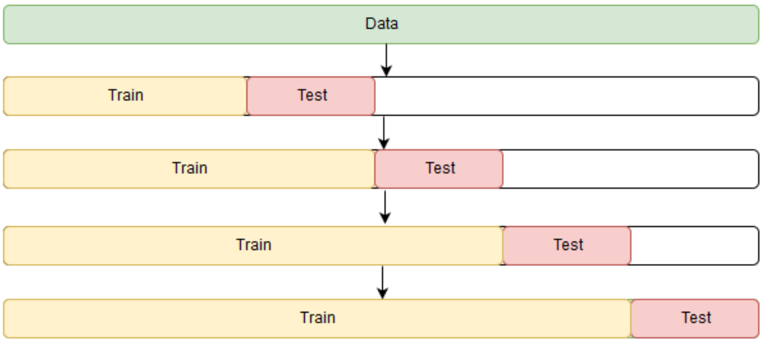

Figure 4. Nested Cross Validation

4.4 Evaluation time series forecasting models

To evaluate the forecasting quality of the models, this paper uses RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) measurements:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| * 100$$

where n is the number of time points in test dataset, y_t is expected values, $\hat{y}_t$ is predicted values.

4.5 Predicting the price of Bitcoin

Experiment of individual models: ARIMA (6, 1, 5) is selected for predicting the Bitcoin price. Neural networks are trained automatically with lags observation from 1 to 14 and epochs includes 100, 200, 500. After experiment, FFNN model with 5 lags, 9 hidden nodes and 100 epochs are used. CNN model with 5 lags, 6 hidden nodes and 200 epochs have the smallest error. LSTM model with lags 5, 100 hidden units and 100 epochs are used. SVR with kernel “rbf” and 4 lags have the smallest error.

Table 1. Error of individual models

Model	RMSE	MAPE (%)
ARIMA	**214.6536**	**2.5626**
FFNN	252.1017	3.6087
CNN	280.0976	4.1937
SVR	232.5739	2.8558
LSTM	219.3169	2.5852

With the closing price of Bitcoin, ARIMA and LSTM models have a good result with RMSE và MAPE. Meanwhile, CNN is not good for time series as expected.

Experiment hybrid model based on ARIMA with machine learning and regression: Hybrid ARIMA_CNN has the best forecasting with RMSE and MAPE.

Table 2. Error of hybrid model based on ARIMA with machine learning and regression

Hybrid model	RMSE	MAPE (%)
ARIMA_FFNN	213.3135	2.5849
ARIMA_CNN	**213.0082**	**2.5563**
ARIMA_LSTM	214.1901	2.5347
ARIMA_SVR	213.8141	2.6144

Experiment hybrid model based on fluctuation interval: The minimum fluctuation is 0.000679% and the maximum fluctuation is 35.849315%. Based on histogram of fluctuation, it showed that fluctuation is very little over 25% and they are outlier. Therefore, the range of fluctuation is R = [0-25]. Applying the formula of Herbert A. Sturges [15], there are 12 fluctuation intervals, that are [0-2), [2-4), [4-6), [6-8), [8-10), [10-12), [12-14), [14-16), [16-18), [18-20), [20-22), [22-24). The fluctuation is beyond the range of R in the data set, they belong to [22-24) interval.

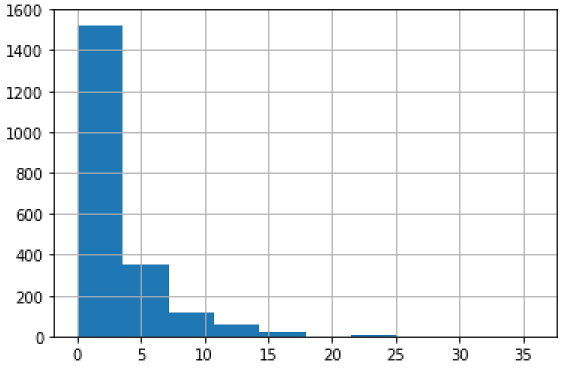

Figure 5. Histogram of fluctuation in the closing price of Bitcoin

Using nested cross validation, we did experiment and choosed models corresponding to interval below: interval [0, 2): ARIMA, interval [2,4): ARIMA, interval [4,6): ARIMA, interval [6,8): SVR, interval [8,10): LSTM, interval [10, 12): CNN, interval [12, 14): ARIMA, interval [14,16): CNN, interval [16,18): CNN, interval [18,20): ARIMA, interval [20,22): LSTM, interval [22, 24): ARIMA. Prediction on test dataset, error of hybrid model is RMSE= 214.2755, MAPE= 2.5483%.

5 Conclusion

The time series forecasting is a difficult problem, but is very important in economics and finance. In the recently, predicting the price of Bitcoin has attracted broad interests from researchers. This work has presented hybrid methods to improve prediction of Bitcoin in the next day. The experiment showed that, the most of hybrid models are possible to improve prediction through measurements of errors such as RMSE, MAPE. The Bitcoin price often depends on investor psychology and investors' interest. In the future, Google trending or investor sentiment on social networks will be analyzed and combine with time series models to improve prediction.

References

1. Satoshi Nakamoto: Bitcoin: A Peer-to-Peer Electronic Cash System (2008).
2. Devavrat Shah, Kang Zhang: Bayesian regression and Bitcoin. arXiv preprint arXiv:1410.1231 (2014).
3. João Almeida, Shravan Tata, Andreas Moser, Vikko Smit: Bitcoin prediction using ANN. Neural Networks (2015).
4. Isaac Madan, Shaurya Saluja, Aojia Zhao: Automated Bitcoin trading via machine learning algorithms. Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep. (2015).
5. Huisu Jang, Jaewook Lee: An empirical study on modeling and prediction of Bitcoin prices with bayesian neural networks based on blockchain information. In: IEEE Access, vol. 6, pp. 5427–5437 (2018).
6. Sean McNally: Predicting the price of Bitcoin using machine learning. Ph.D. dissertion, School Comput., Nat. College Ireland, Dublin, Ireland (2016).
7. Alex Greaves, Benjamin Au: Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin. Tech. Notes in Standford Univ. (2015).
8. Steve Y Yang, Jinhyoung Kim: Bitcoin Market Return and Volatility Forecasting Using Transaction Network Flow Properties. In: Computational Intelligence, IEEE Symposium Series on. IEEE, pp. 1778–1785 (2015).
9. Ferdiansyah, Edi Surya Negara, Yeni Widyanti: BITCOIN-USD Trading Using SVM to Detect The Current day’s Trend in The Market. Journal of Information Systems and Informatics (2019).
10. G. Peter Zhang: Times series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, vol. 50, pp. 159–75 (2003).
11. T.T. Soong: Fundamentals of probability and statistics for engineers. John Wiley & Sons, USA (2015).