機械学習アルゴリズムによるビットコイン価格変動の予測

Prediction of Bitcoin Price Movements with Machine Learning Algorithms

新立 拓也 [1]　ピフル　ルカーシュ [1]　海蔵寺　大成 [1]

Takuya Shintate[1], Lukáš Pichl[1], and Taisei Kaizoji[1]

[1] 国際基督教大学
[1]International Christian University

Abstract: We study the limits of prediction accuracy of Bitcoin price data in CNY currency using tick data from the OKCoin Bitcoin exchange (source: Kaiko data). The tick data contain the price, volume, and trade direction, and are transformed to the OHLCV format using standard methods. In this report, we deploy the Support Vector Machine algorithm by Vapnik to estimate the sign of the hour-to-hour transaction return using a sampling moving window of varying size on the past data. Several kernel functions are validated. Our first results for all months of the year 2015 show that the hit ratio accuracy level (the fraction of correctly predicted upward or downward events) does not exceed 60%. It remains to be established whether this low result corresponds to the causal extraction limit inherent in the data, or whether it can be improved by deploying other methods, such as LSTM networks in deep learning.

Introduction

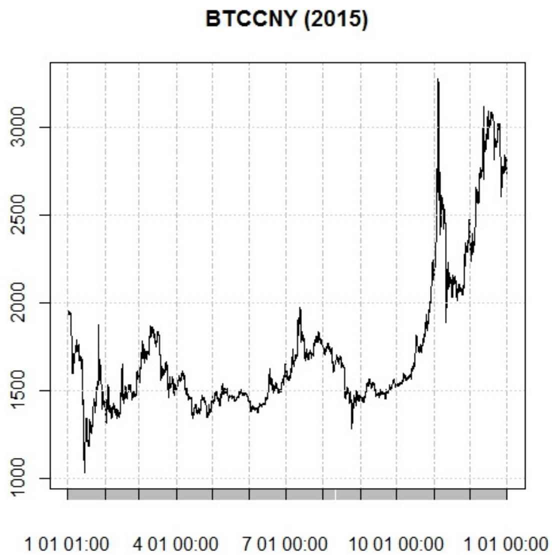
Bitcoin price data time series denominated in standard currencies are subject to many extreme events, since bubbles and crashes in the cryptocurrency markets are quite common. For instance, at the beginning of 2017, the Bitcoin price was about 1000 USD, and since then it has seen the rise to almost 5000 USD on September 1st of the same year, followed by a sharp drop in just a two-week period to about 3000 USD-level on September 14th, when China (accounting for major source of Bitcoin demand) announced suspension of Bitcoin trades at domestic exchanges. Therefore the BTC (XBT) price series are a serious benchmark both for econometric and machine learning algorithms.

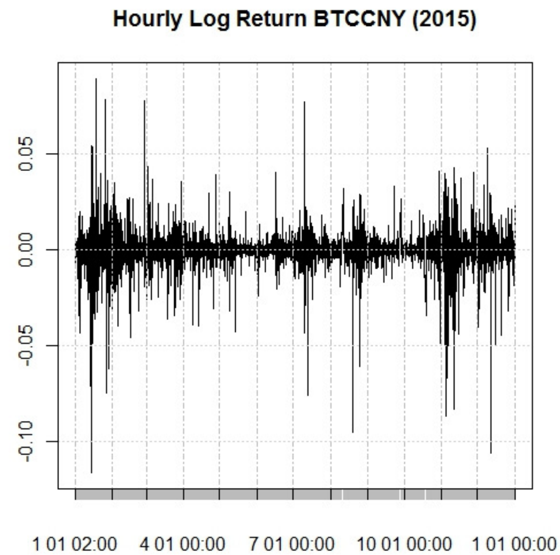
The available literature related to Bitcoin time series prediction is relatively limited, with the exception of the burst of mostly economic publications this year [1-3]. The insight from the field of computational intelligence still remains rather limited. We have analyzed the BTCEUR price times series using recurrent neural network in Elman topology in a previous work [4], finding out that that on the daily scale the MSE of the predicted data (5-day ahead prediction) scales as the Realized Volatility computed from intraday trading data.

Support Vector Machine

When predicting the logarithmic returns on equidistant time grid, the MSE is typically used as the accuracy criterion for training neural networks. It is, nevertheless, more intuitive and an easier subject to interpretation to focus just on the hit ratios, i.e. the trend data (class 1 for logarithmic return $R>0$, class -1 for $R<0$; no-tick data excluded). Then we have a binary classifier problem, for which, according to the standard findings in the literature [5], the Support Vector Machine method is considered to be superior. As quoted in [5], “The underlying motivation for using SVMs is the ability of this methodology to accurately forecast time series data when the underlying system processes are typically nonlinear, non-stationary and not defined a-priori.”

The present dataset of hourly data for BTCCNY prices is broken to 12 monthly segments, and for each of them, 2/3 of the data is used for SVM training, and the remaining part for hit ratio validation. We adopt the R-package kernlab [6] implementing the original method of Vapnik [7]. The relative constant for penalizing mis-predictions is set as $C=10$. Figure 1 depicts the BTCCNY time series in 2015 and shows the clustering behavior of the logarithmic return on the 1 hour scale.





Results and Discussions

We have experimented with various polynomial kernels of orders 1, 2, 3, and 4, tuning the scale and offset parameters, and the Radial Basis function kernel, with manually tried and automatically determined Gaussian function width. The highest validation scores were obtained for the automatically set RBF kernel. These results are summarized in Table 1 for the input space dimension (size of the predictor sampling window) N=5.

Table 1 SVM-prediction hit ratio for BTCCNY 2015

JAN	FEB	MAR	APR	MAY	JUN
52.08	52.53	50.85	51.80	53.11	50.86
JUL	AUG	SEP	OCT	NOV	DEC
51.45	50.22	56.87	56.88	53.04	58.75

As the above table indicates, the SVM-predicted results are rather in accord with the efficient market hypothesis, suggesting a low level of causal content in the time series data. We have been currently implementing the LSTM deep-learning networks in an attempt to confirm whether the data in Table 1 are an upper bound or not. At the conference, we will also briefly review our previous work on BTC time series analysis [4].

References

[1] Bariviera, A. F., Basgall, M. J., Hasperué, W., and Naiouf, M.: Some stylized facts of the Bitcoin market, *Physica A: Statistical Mechanics and its Applications* Vol. 484, pp. 82-90, (2017)

[2] Blau, B. M.: Price dynamics and speculative trading in bitcoin. *Research in International Business and Finance* Vol. 41, pp. 493-499, (2017)

[3] Katsiampa, P.: Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters,* Vol. 158, pp. 3-6, (2017)

[4] Kodama, O., Pichl., L., Kaizoji, T.: Regime change and trend prediction for Bitcoin times series data, in Proceedings of CBU International Conference on Innovations in Science and Engineering, Prague, 2017, DOI: http://dx.doi.org/10.12955/cbup.v5.954

[5] Sapankevych, N, I., Sankar, R.: Time Series Prediction Using Support Vector Machines: A Survey, IEEE Computational Intelligence Magazine, Vol. 4, No. 2, (2009)

[6] Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., kernlab -- An {S4} Package for Kernel Methods in {R}, Journal of Statistical Software, Vol. 11, No. 9, pp. 1-20, (2004).

[7] Cortes, C., Vapnik, V.: Support Vector Networks, Machine Learning Vol. 20, pp. 273-297, (1995)