Sliding CUSUM and its Solutions to Anomaly Detection problems in the field of Digital Currency

Thien Trang Nguyen Vu[1] ● Tung Nguyen Van Thanh[2] ● Lam Dang Thanh[3]

School of Information and Technology, Hanoi University of Science and Technology, Ha Noi, Viet Nam.

ABSTRACT

Detecting anomalies from times series data has become crucial in modern applications such as stock price irregular patterns detecting, oil drilling early warning systems, internet data stream, etc. Such real-time data applications experience significant influence of seasonality effect, concept drift phenomenon and unpredicted sudden change in attributes of data. This paper proposes a time series anomaly detection algorithm developed based on the combination of sliding window technique and CUSUM (Cumulative Sum). This algorithm is named SCUSUM (Sliding CUSUM). With the appearance of concept drift phenomenon, the application of the sliding window method is required to divide data into smaller segments, focus on volatility on the latest pattern. A modified two-sided cumulative sum algorithm is another factor to solve the problem of changing properties of data. SCUSUM first calculates the average of measurements in the latest window. Along with the previously defined values, the average and standard deviation of means of the windows are updated. After having required calculations, the cumulative sum is applied on the means of the windows. Each time a new data point is considered, SCUSUM takes into account the shift range between mean of the new window and average of all windows' mean. One special point is that the shift range needed to be detected and the thresholds are dynamic, which here are both chosen as the weighted standard deviation. This characteristic ensures that the parameters are consistent with the flow of data. With innovation and combination of simple algorithms, SCUSUM can detect abnormal trends of data despite facing up to concert drift. Experiments on Bitcoin price anomaly detection have shown great results that solve the challenge of an accurate but early enough algorithm. In the field of digital currency rate, not only the computation complexity of SCUSUM surpasses practical needs, but also it has detected early irregular patterns, opening up many potential possibilities.

[1] Thien Trang Nguyen Vu. Email: trang.nvt194459@sis.hust.edu.vn
[2] Tung Nguyen Van Thanh. Email: tung.nvt190090@sis.hust.edu.vn
[3] Lam Dang Thanh. Email: lam.dt194442@sis.hust.edu.vn

KEYWORDS
anomaly detection, sliding window, cumulative sum, concept drift, Bitcoin price

INTRODUCTION

It would not be an exaggeration to say that we are living in the era of digital marketplaces. Virtual currency is becoming more and more popular than ever before. More specifically, here we are talking about Bitcoin. It cannot be denied that the popularity and adoption of Bitcoin are increasing at a fast rate, despite its being such a young currency. The number of people interested in Bitcoin keep rising dramatically and no sign of stopping has been seen. Thus, this fever extends to the whole field of science. Under the perspective of statistics, many algorithms and new ideas have been introduced to monitor, analyze and predict Bitcoin rate trends.

When it comes to outstanding issues related to Bitcoin rate, it would be a big mistake not to mention the anomaly detection. Anomaly detection is defined as the process of looking for patterns in data that do not follow any model of normal behavior. Some might question what makes it big a deal of finding rare patterns. The matter here is the meaning behind the appearance of abnormalities. To the extent of science, such anomalies might lead to beautiful discoveries. In contrast, when the background is the business world, where we rarely want things out of order, even slightest irregularities can make a great impact. Early anomaly detection can solve many business problems. In the scope of this paper, where the topic being discussed is Bitcoin price, the sooner anomalies are detected, the better analysis and prediction can be proposed. This brings significant benefits for not only the companies, who are drastically coming up with strategies to compete with other competitors on the e-commerce wars, but also for the users of Bitcoin. It is obvious that anomaly detection plays an important role in every field, which leads to the fact that early detection is crucial.

Its importance requires a variety of algorithms to solve the anomaly-detecting problems in different goals and domains. The knowledge which has been applied to the development of the proposed methods are numerous, such as z-scores, sequential statistics, density-based algorithms, etc. Each method has its own strengths and meets different demands. In this paper, the need of anomaly detection in time series data with the presence of concept drift would be discussed. Note that here the concept drift is the phenomenon where unpredicted changes in data properties are experienced. This phenomenon is extremely annoying for analysts for the reason that it leads to wrong assumptions in the future. The problem now is not only about timely detection but also about the need for an accurate enough model that works great despite the appearance of concept drift. High accuracy and low accuracy, the two contrast features require new algorithms. Such a model was proposed and developed by our team based on two techniques, sliding window and CUSUM, which was named SCUSUM (Sliding CUSUM). In this paper, ideas and the works behind SCUSUM algorithm would be explored and applications of that on the real-world data of digital currency would be shown and analyzed.

This paper first explains related knowledge, then show detailly how SCUSUM works. Based on that, the algorithm would be applied to real Bitcoin rate data. Finally, the model would be evaluated based on the results and extended further to future potentials.

SCIENTIFIC BASIS KNOWLEDGE

SLIDING WINDOW TECHNIQUE

The main idea of window methods is to split an unbounded stream of data (events) into finite sets, which are called windows, based on specified criteria. Applications on windowing techniques are wide, most of which are time series classification, clustering, or anomaly detection. Based on the great advantages it brings, windowing is one of the most widely studied when it comes to processing methods for data streams.

Choosing time as the main criterion for truncating data, window methods are divided into three kinds for the data stream based on each's length: landmark, snapshot and sliding window. This paper focuses on the sliding window, which is the most frequently used and most practical. Given a continuous time series stream, the sliding window observes the most recent data points measurements and moves by a fixed step size along the time axis as new ones arrive. In other words, data points are grouped in a window that slides along the data stream according to a specified interval. The main difference compared to the other two types of window techniques is that sliding windows always remain the same size and same processing data range. The advantage of sliding windows is that the concept can be considered as a table which memorizes new added measurements and removes old ones.

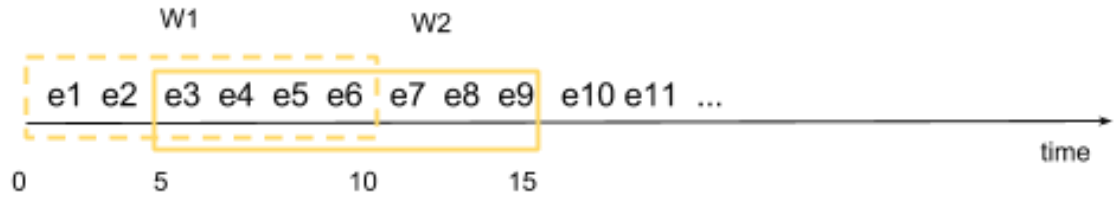

Figure 1. Sliding Window

As in this paper and most applications, sliding window technique is used to truncate data, analyze significance of the most recent time series segment and compare current window to preliminary analyzed data measurements.

CUSUM

The CUSUM (or cumulative sum control chart) is a statistical control chart based on hypothesis testing developed for i.i.d.[4] random variables. In statistical quality control, it is a sequential analysis technique which is typically used for monitoring change detection and irregular patterns, therefore can be used for step detection of a time series.

As its name implies, to detect changes in the distribution, CUSUM involves the calculation of cumulative sum, which represents the cumulative deviation between expected values and real

[4] Independent and identically distributed

observed value. These patterns are classified as anomalies when the upper or lower control limit exceeds a certain threshold. Combining with sliding window techniques, CUSUM chart is a method that is able to detect small shifts in statistical parameters, which in our paper is the process' mean, relative to the regular parameter.

The CUSUM chart requires four parameters:

- **μ** The expected mean of the process
- **σ** The expected standard deviation of the process
- **k** The size of the shift to be detected
- **H** The control limit

To be more specific, given a process characteristic X that follows a normal distribution N(μ, σ). Firstly, to simplify how the whole algorithm works, it will be instructive to explain the definition of the one-sided CUSUM chart. A commonly-used one-sided CUSUM chart is to track the individual cumulative sums of deviations from the mean, which are summed sequentially as follow:

$$So = 0$$

$$S_{n+1} = \max(0, S_n + X_n - \mu - k)$$

$$for\ n = 1, 2, \dots N$$

The CUSUM chart addresses out-of-control signals when the value of S exceeds the threshold H. Note that the above formula only detects changes in the positive direction. When it comes to finding both positive and negative changes, we need a dual CUSUM scheme. Here, CUSUM periodically computes two sums, which are the negative and positive deviations from the man. The high value and low value formulations are shown respectively below:

$$S_{n+1^+} = \max(0, S_n + X_n - \mu - k)$$

$$S_{n+1^-} = \min(0, S_n + X_n - \mu + k)$$

When the upper control limit or lower control limit exceeds the threshold H, deviations relating to this pattern are classified as anomalies.

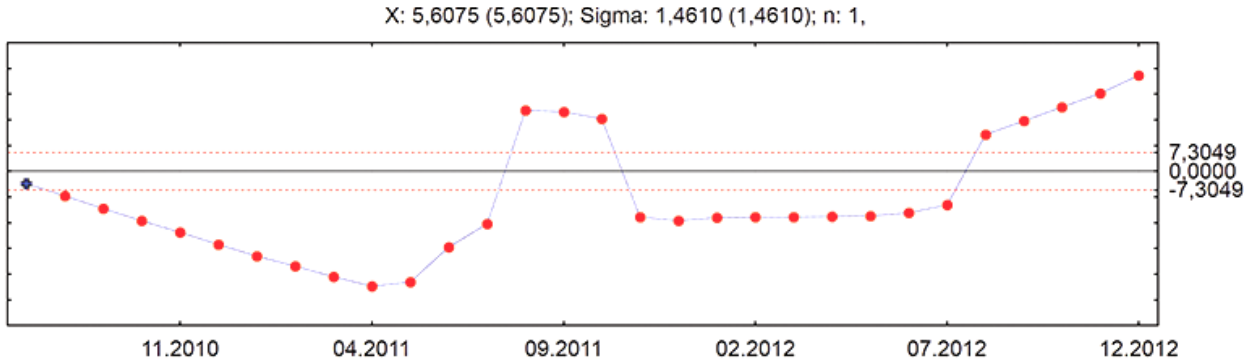

Figure 2. CUSUM BTC/USD 2010-2012.

An example of a CUSUM chart of Bitcoin exchange rate when converted to USD in 2010 to 2012 is shown above for clarifying the concept of this algorithm (Figure 2). Overall, the data fluctuated during the investigated time. At the beginning, the CUSUM rate gradually plunged until reaching the lowest peak in 2011 April, after which it started to rise. A high strong peak was hit in 2011 August with an approximately 2 month-time of stability up there. The cumulative sum again experienced a dramatic drop in the same year's October. After a while of remaining slight growth, the data took a strong leap again in 2012 July and kept rising up and hit the highest peak at the end.

SCUSUM

PROBLEMS

In time series, the phenomenon where statistical properties of the target variable change overtime in unforeseen ways is called concept drift. This kind of phenomenon changes patterns over time in underlying data foundations of a concept, which causes problems that the predictions become less accurate as time passes. The term concept commonly refers to the data value, the quantity to be predicted, or the target variable. In the financial domain, market behavior changes dramatically in contract prices, interest rates, inflation rates, budget announcements, and political and world events. In short, concept drifts are caused by changing real world conditions, it drifts a value away from the patterns and rules where it was perceived to be related to at historical time intervals.

Figure 3. Average Bitcoin price changing in 10 years (10 Aug 2010 – 10 Aug 2020)

Figure 4. Cumulative Bitcoin profit in 10 years if consistent invest on August – September every year (10 Aug 2010 – 10 Aug 2020)

Observations of Bitcoin rate for 10 years from 2010 to last year show (Figure 3), the data experience a decrease in the years' September on average. At this time of year, mutual fund managers, on average, typically sell losing positions before year end, investors change their portfolios at the end of summer to cash in, causing direct effect on other fields of the economy, changing inherent properties. Those strange patterns are presumed to be due to external factors, which in this case might be the consequence of the September Effect, and leads to the phenomenon concept drift.

The difficulty of concept drift is that the future distribution is unknown, which makes it more likely to misclassify the abnormal data. To solve such problems, sliding CUSUM for data stream anomaly detection was proposed with a combination of sliding window and control charts method.

MAIN IDEA

As mentioned before, the SCUSUM (Sliding CUSUM) was built based on two algorithms, sliding window and cumulative sum chart.

Firstly, the sliding window is adapted to enhance the trend analysis of the current data and previous data points, which helps to solve the problems caused by the phenomenon concept drift.

Secondly, as its name implies, the dual CUSUM algorithm is applied to integrate positive trend and negative trend of means of windows. Once either of the two control limits is exceeded, it would raise an alarm of irregular pattern.

Combined the two proposed techniques, a final model was developed as follow:

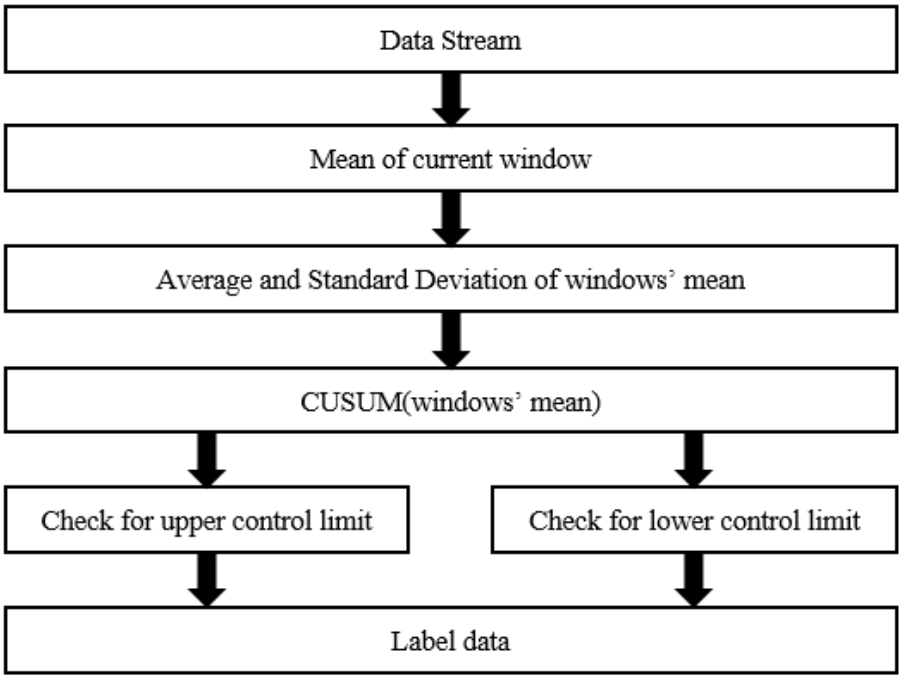

Figure 5. Model of DCUSUM-DS

NOTATIONS

L_W	Length of the windows

β	The output rate
M_L	The mean of the latest window
D_{mL}	The mean of windows' mean values
D_{SL}	The mean of the windows' standard deviation
S_{n+1}^+	The upper cumulative sum
S_{n+1}^-	The lower cumulative sum

FORMULAS AND PARAMETERS EXPLANATION

ACCEPTABLE VARIATION

$$S_{n+1^+} = \max(0, S_n + M_L - D_{mL} - \beta * D_{SL})$$

$$S_{n+1^-} = \min(0, S_n + M_L - D_{mL} + \beta * D_{SL})$$

As the original CUSUM formula stated above, the sum is added sequentially by a range of variation between the considered measurement and the mean value. To smooth the current significance of change and reduce the sensitivity of current data points, a range of acceptable variation is taken into account. In our proposal, we decided to compute the range by the weight standard deviation of windows' mean. With this modification, the differences which are between the allowed range would be ignored.

CONTROL LIMITS

One of the most important parameters of SCUSUM algorithm is the threshold. It is noticeable that the thresholds we chose, which decide the boundaries between normal trends and anomalies, is a nonstationary value. In order to be consistent with the change of data series over time, positive and negative standard deviations were selected as upper control limit and lower control limit, respectively. Despite changing over time, the standard deviation is a value relative to the previous data, hence its change would not be too significant. Based on the characteristics of this value, the standard deviation is a good threshold for detecting anomalous patterns.

ALGORITHM

ALGORITHM 1 SCUSUM

INITIALIZE L_W, β
COMPUTE M_L
UPDATE D_{mL}, D_{SL}
$S_{n+1^+} = \max(0, S_n + M_L - D_{mL} - \beta * D_{SL})$
$S_{n+1^-} = \min(0, S_n + M_L - D_{mL} + \beta * D_{SL})$
IF $S_{n+1^+} > D_{SL}$ **THEN**
    Label ← *Anomaly*
IF $S_{n+1^+} < -D_{SL}$ **THEN**
    Label ← *Anomaly*

To start with, the mean of the latest windows is computed. Based on this value combined with the previously calculated means, the next step is to update the average and standard deviation of the means of windows. After that, the high cumulative sum and low cumulative sum is calculated, which are defined by the acceptable variation between the average and the mean of the current considered window. Having all required values, comparisons with the limit controls are taken. Once the value exceeds one of the two thresholds, or standard deviation of windows' mean, the algorithm returns the label of the pattern.

RESULTS EVALUATION

The dataset we used in this paper to evaluate the algorithm is the bitcoin price history from the crypto currency exchange Binance platform, from 1/7/2021 to 31/7/2021. The time interval between two records is 1 minute. Each record includes Opening price, High price, Low price, and Closing price (OHLC format) data. The analysis parameter we use is the close price.

Like many other real-life data, this dataset comes without any label of which price is abnormal and which is not. Therefore, it's impossible to evaluate the algorithm with ROC or Precision-Recall based criteria, which are standard metrics of evaluation with labels. However, we can still take a look at the result and have an overview of how the algorithm performs.

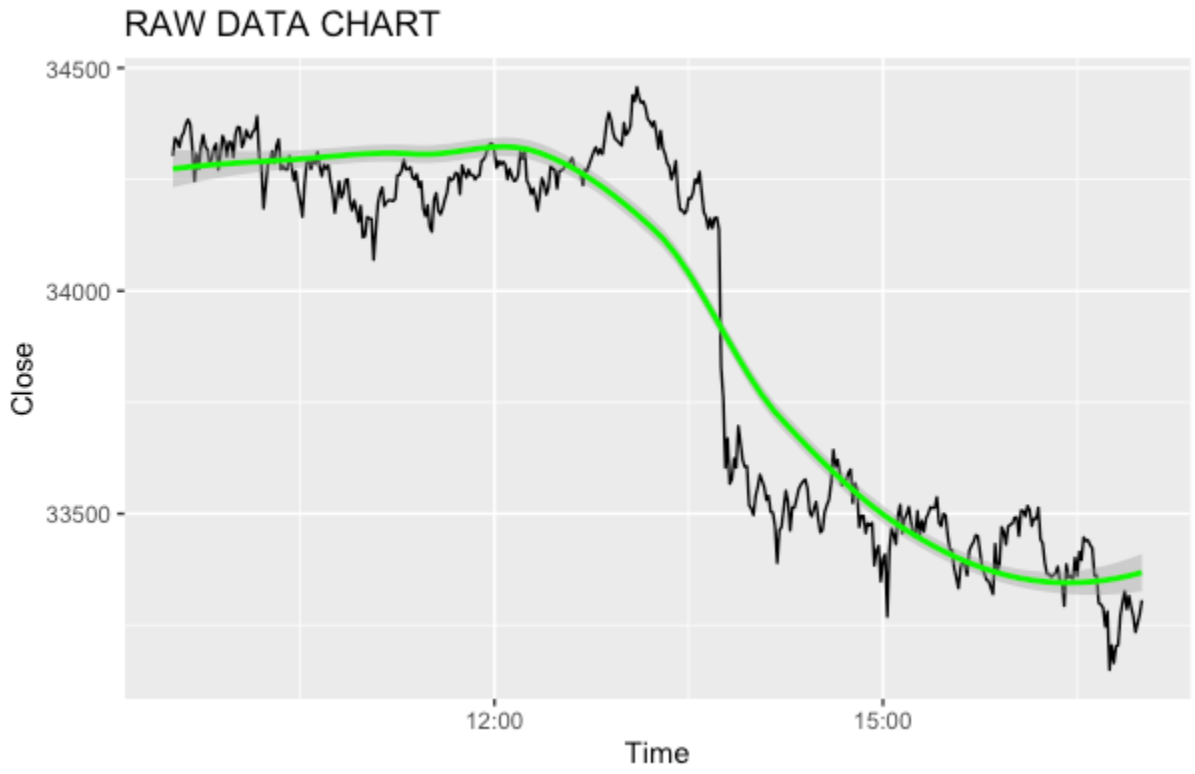

Figure 6. Raw data chart of Bitcoin price history dataset

A small part of our dataset is shown in figure 1, with the black line representing the price, and the green line is a smooth curve created by the Loess function. The anomaly we need to detect is the significant drop in the close price around 14:00. Other changes in the price are relatively small, so any signal caused by these changes can be considered as a false alarm.

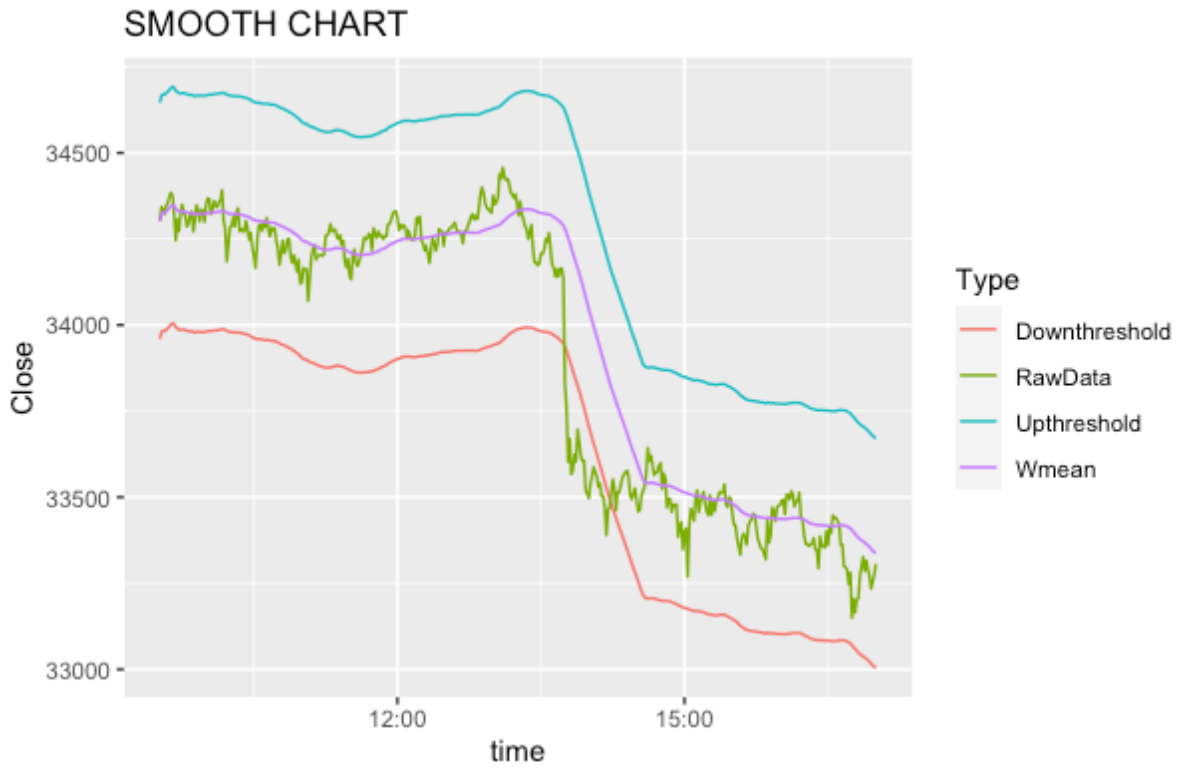

Figure 7. Data analysis diagram

As shown in Figure 7, the data in the normal range are enveloped in the upper and lower threshold line. When there is an abnormal rise or abnormal drop, the feature value will exceed the threshold line. The anomaly detection algorithm deals with the current point and several nearby data. After the cumulative number of deviation data exceeds the set demarcation rate, the SCUSUM reaches the classification criteria.

Figure 8 maps the raw data with the characteristic of SCUSUM. The top graph shows the dual CUSUM chart of the data, integrating positive trend and negative trend of means of windows, while the bottom graph is drawn from the original data in the same interval of time. The chosen length for the sliding window is 50. The upper and lower threshold change over time based on the standard deviation of the data. The anomaly is detected near 14:00, when the CUSUM negative variable exceeds the lower threshold. The CUSUM positive variable had a slight rise at about 13:20, but its value was still under the upper threshold so no alarm was raised. This result is totally consistent with the original intention of the algorithm design. One thing worth mentioning is the time interval between the drop in the data and the alarm signal is relatively small, only a few minutes. As discussed before, a quick detection rate can bring great benefits for both finance companies and individual users when applied to the cryptocurrency market.

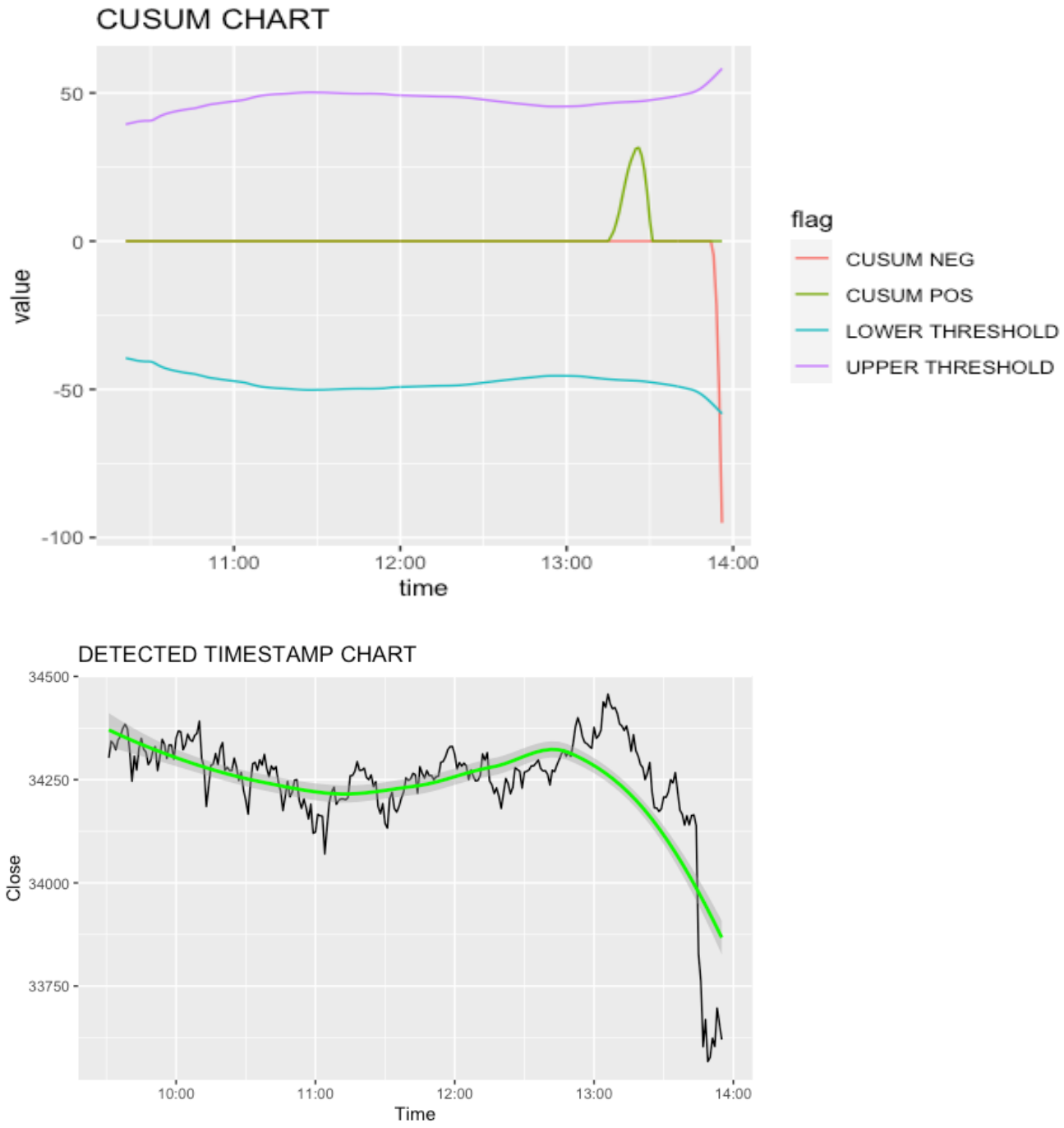

Figure 8. Feature mapping of SCUSUM

Finally, we need to consider the complexity of the algorithm. The complexity of each step is only linear time, directly proportional to the length of the sliding window. In this particular problem, the sampling rate is low so the computational cost is not really of concern, but many other anomaly detection problems need to deal with industrial data streams, which generally have the sampling frequency 1Hz or 0.2Hz. In that case, having a low time complexity and computational cost certainly is a big advantage.

CONCLUSION

In this paper, our team proposes the SCUSUM algorithm to detect the anomaly in time-series data. The algorithm uses the sliding window technique and CUSUM control chart to integrate positive and negative trends in the data, while reducing the effect of concept drift. From the experiment, it is shown that the algorithm can early detect abnormal trends in data, which makes great application in the finance market, such as digital currency. The computation complexity of the algorithm even surpasses the practical needs of the field, opening the potential for application in other fields that require low computational cost like industrial data stream anomaly detection.

One further step we might take is experimenting to see the influence of the parameter in the algorithm, such as the length of the window, on the final result. The algorithm should be tested with the labeled datasets for better estimation of its accuracy. The comparison with other existing anomaly detection algorithms is also essential. Finally, the applicability of the algorithm in the real world needs to be verified.

REFERENCES

1. Lesti, G.; Spiegel, S. 2017. A Sliding Window Filter for Time Series Streams.
2. Guang Li; Jie Wang; Jing Liang; Caitong Yue. 2018. The Application of a Double CUSUM Algorithm in Industrial Data Stream Anomaly Detection. DOI: 10.3390/sym10070264.
3. Sliding Windows, Understanding Sliding and Tumbling Windows, Developing Apache Storm Applications. https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.5/developing-storm-applications/content/understanding_sliding_and_tumbling_windows.html
4. Implementing Windowing Computations on Data Streams. Developing Apache Storm Applications. https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.5/developing-storm-applications/content/implementing_windowing_computations_on_data_streams.html
5. CUSUM, Wikipedia. https://en.wikipedia.org/wiki/CUSUM
6. Kinga, F. 2016. CUSUM Anomaly Detection.
7. Yi Zhao; Fugee Tsung; Zhaojun Wang. 2004. Dual CUSUM control schemes for detecting a range of mean shifts. DOI: 10.1080/07408170500232321.
8. Olufowobi, H.; Ezeobi, U.; Muhati, E.; Robinson, G.; Young, C. 2019. Anomaly Detection Approach Using Adaptive Cumulative Sum Algorithm for Controller Area Network. DOI: 10.1145/3309171.3309178.
9. Tim. “What are Concept Drifts in Time Series Data?”. https://www.iunera.com/kraken/fabric/concept-drift/
10. Harries, M..; Kim, H. 1996. Detecting Concept Drift in Financial Time Series Prediction using Symbolic Machine Learning.
11. Concept drift, Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Concept_drift

12. Siera. Why Anomaly Detection is Important?. https://avianaglobal.com/why-anomaly-detection-is-important/
13. Aryng. Why Anomaly Detection is important for businesses?. https://aryng.com/blog/anomaly-detection/
14. Anomaly detection, Wikipedia. https://en.wikipedia.org/wiki/Anomaly_detection
15. Beginners Guide: What is Bitcoin?.https://coincentral.com/beginners-guide-what-is-bitcoin/
16. Szettela, B. 2016. The Use of Control Charts in the Study of Bitcoin's Price Variability. DOI: 10.5772/66360.
17. Seasonax. https://app.seasonax.com/assets/btc-usd-cc?h=eJyrVkpUsjI2MtBRKipWslIyMjA0MLAwNFAC8lPBfCMYv1LJKjpWR6kApAwoZAYUKgApAWowB7JTlKyAhiSXgalcJSvDWgCk1hRd
18. Speck, D. 2020. FOR BITCOIN, THE SEASONALLY WEAKEST PERIOD OF THE YEAR IS BEGINNING NOW. https://www.seasonax.com/research/bitcoin-seasonally-weakest-period
19. Gallant, C. 2020. Why People Say September Is the Worst Month for Investing. https://www.investopedia.com/ask/answers/06/septworstmonth.asp
20. Binance [Dataset]: https://data.binance.vision/?prefix=data/spot/monthly/klines/BTCUSDT/1m/