ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 246-252
doi: 10.18178/wcse.2019.06.037

An Improved Method for Electromagnetic Streaming Data Anomaly Detection

Degang Sun[1,2], Yulan Hu [1,2] , Zhixin Shi [1,+] and Guokun Xu [1]
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. The electromagnetic data is a kind of real-time, streaming data, the usage of wireless devices can alter the energy value at the corresponding frequency point. To detect the usage of wireless devices timely, an improved CUSUM method(H-CUSUM) is proposed, which combined a nonparametric CUSUM algorithm with an adaptive sliding window and tolerance factor. Four experimental data sets are collected to evaluate the proposed method, The result shows that the proposed method can detect the change in electromagnetic streaming data with high coverage and low latency.

Keywords: Change Point, H-CUSUM, Sliding Window, Tolerance Factor.

1. Introduction

The electromagnetic data[1] is gathered by the electromagnetic collector, through specific channels, the time-series data collected can be transmitted continuously to a repository as a data stream[2], it's necessary for us to detect the anomaly signals(e.g, wireless signals) during the transmission process. Currently, there are numerous anomaly detection frameworks and techniques[3][4][5][6][7] for time series anomaly detection, the time series anomaly detection techniques can be categorized into three main groups: statistical based, clustering based and nearest neighbor based approaches.

The CUSUM algorithm[8] is a statistical-based method. For the sake of low latency and low complexity considerations, it is suitable for streaming data anomaly detection. There have been several improvements for this method. In[9], an adaptive CUSUM method was proposed to meet the challenge of malicious network traffic detection, the threshold can be adaptively changed therefore reduced manual intervention. In[10], the author uses a vector to monitor multiple ports simultaneously, expand the original CUSUM method to multidimensional form. In[11], a markovian adaptive CUSUM(ACUSUM) scheme was proposed to analyze the future unknown shifts and evaluate the run length performance of the scheme. In[12], by using the defined time-slot structure, the author proposed a nonparametric CUSUM algorithm to surveillance the incoming network data. In[13], the author combined the sliding window and tolerance factor with the CUSUM algorithm, thus monitor the DDOS network data in an adaptive way.

The electromagnetic streaming data is essentially similar to network traffic data[14][15], once an attack occurs, the statistical attribute of the electromagnetic streaming data deviates simultaneously, so it is natural to transfer the above methods to electromagnetic streaming data anomaly detection. This paper put forward an improved multidimensional nonparametric adaptive CUSUM method based on adaptive sliding window and tolerance factor to detect the anomaly signals. For efficiency, the group area calculation is denoted as the detection feature. The contributions of this paper can be attributed to the following three aspects: (1). The group area calculation method was pull in to reduce the calculation complexity. (2). The nonparametric form

+ Corresponding author. Tel.: +8618600561912
E-mail address: shizhixin@iie.ac.cn.

of the method has been proposed to accommodate the unknown distribution of coming stream data. (3). The sliding window mechanism and the tolerance factor made the detection method implement in an adaptive way, which reduces human intervention.

The rest of the paper is organized as follows. In Section 2, the proposed method is presented in detail. In section 3, performance analysis is made in terms of detection probability and detection delays, besides, the transformation of main parameters during the detection process and their impact on experiment results are analyzed. And section 5 summarizes this paper.

2. Design of The Improved CUSUM Method

In this section, the proposed method is illustrated in detail, the description of the problem is given firstly, afterward, we set forth our method hierarchically.

2.1. Problem formulation

Fig.1 displays an electromagnetic streaming data. The X-axis represents the frequency points number, and the Y-axis represents the corresponding energy value. Each of the frequency points is likely to be attacked, thus it is needed to monitor all frequency points of one electromagnetic streaming data, the problem is converted into a problem of monitoring multidimensional independent variables in time series.

Suppose $X=\{x_n, n=1,2,...\}$ is a sequence of independent random variables observed sequentially, and $\{x_{n,m}\}$ denotes the m_{th} frequency point signal energy value at n_{th} the time. Under normal circumstances, for each frequency point the signal energy value obeys a certain distribution, generally, normal distribution, the mean of X (denoted by μ_X) is stable. At a certain moment (random and unknown), an anomalous event occurs and μ_X shifts. However, the parameterized CUSUM method can not well detect the shifts of the electromagnetic stream data since no prior distribution knowledge of upcoming data is known, thus nonparametric CUSUM method needs to be adopted. Besides, the original CUSUM method relies too much on the detection threshold, it can not adaptively change under different scenarios.

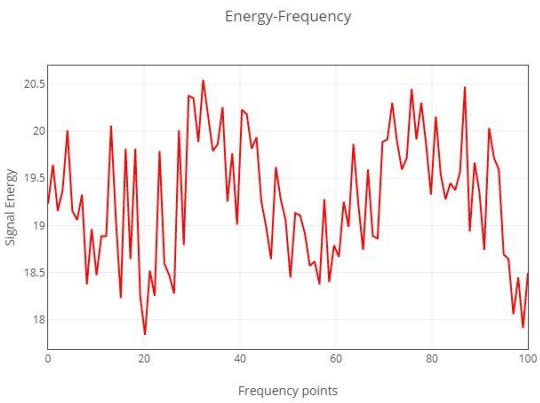


Fig.1: Energy-Frequency Diagram

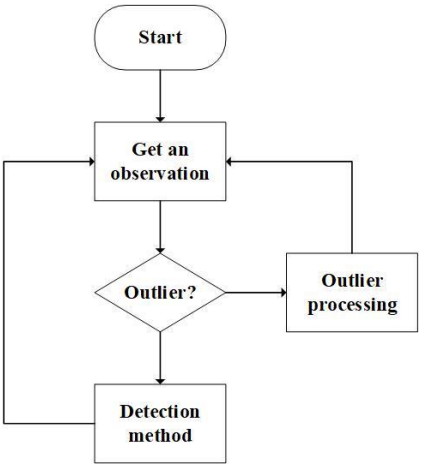


Fig. 2: Detection Process Diagram

Fig.2 shows a high-level process of the proposed detection method. It consists of two key procedures, the first is the outlier process procedure, and the second is where the improved CUSUM method performs. If the coming signal is judged an outlier, it won't be put into the second procedure.

2.2. Outlier process

The outliers are dealt with in this step. During the electromagnetic stream data transmission process, the structured data may be partially missing or not in accordance with common sense(for example, the signal energy value smaller than zero), for the sake of detection efficiency, these data are discarded directly. On the other hand, at each frequency point, a temporary threshold T_i and tolerance factor K_i are set to detect the intensive attacks, the number of successive outliers is cumulated by a counter C_i. Once the value of C_i increases to K_i, it is considered that the corresponding frequency is subjected to a strong attack, the alarm

information will be emitted immediately. Otherwise, the data will be put into the second procedure for further detection.

2.3. Detection method

This section elaborates on the proposed detection approach. Fig.3 displays the sketch map of the detection. The main steps of the second procedure are illustrated in Table 1.

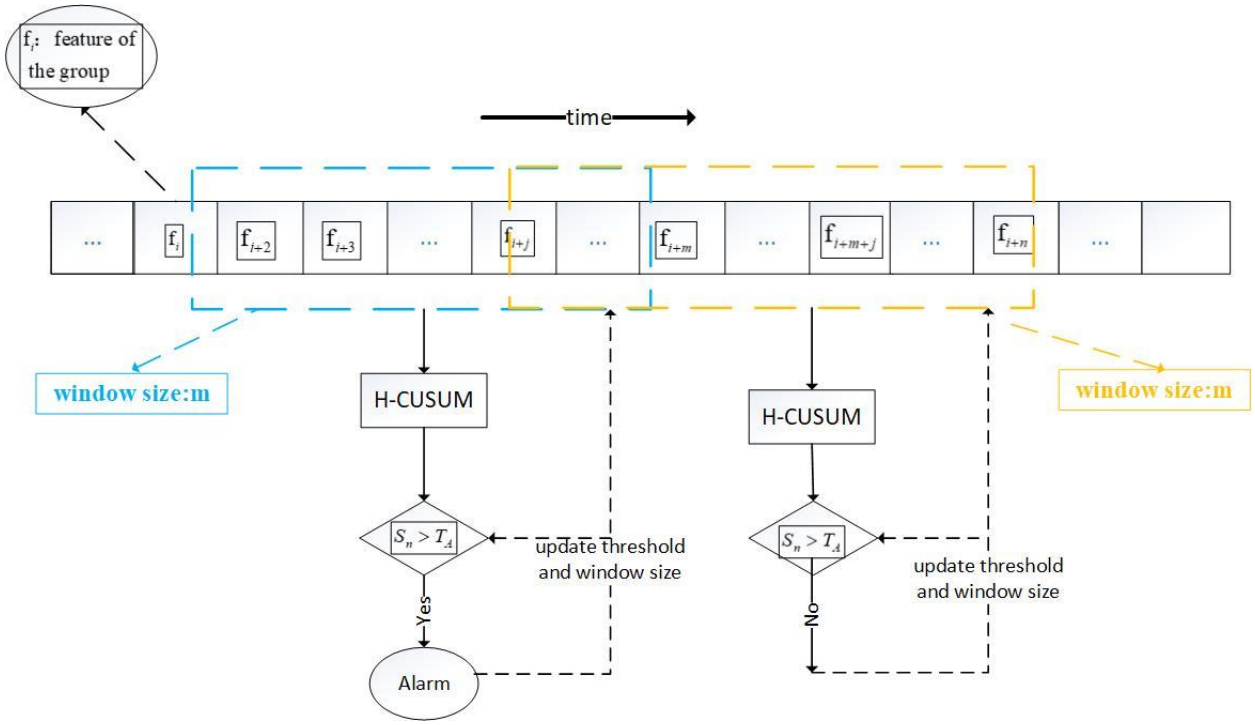


Fig.3: Second Step Flowchart

Table 1: The main steps of the second detection procedure

1) For each electromagnetic streaming data, traverse all frequency points, compose every t frequency points into one group, if there remain extra frequency points, compose the extra frequency points together as one group. Afterward, calculate the group area size of each group separately as the detection feature.
2) For each group, execute the improved CUSUM algorithm, calculate the cumulative sum.
3) Compare the cumulative sum with the respective thresholds, if there exists an attack, update the corresponding threshold and sliding window size.

During the monitoring process, it is required to observe multiple frequency points at the same time. In general, the number of frequency points in a frequency band can reach up to thousands. It will be a waste of computation if we calculate the cumulative sum at each frequency point one by one, thus it's natural to integrate the adjacent frequency points. While traversing an electromagnetic stream data, we converge every t frequency points together as a fixed-size group. Within each group, all the energy values of the frequency points are connected to form a curve, the area enclosed by the curve and X-axis is calculated, we use the area value as the detection feature. Suppose the energy values of t adjacent frequency points at time k are $\{x_{k,i}, x_{k,i+1}, ..., x_{k,i+t-1}\}$, the detection feature of each group can be calculated as $x_{k,i+1} + x_{k,i+2}, ... + x_{k,i+t-2} + \frac{x_{k,i} + x_{k,i+t-1}}{2}$.

For each upcoming streaming data, the same operation is implemented. Then, the cumulative sum of the time-series feature values at each group is calculated for further detection. In this step, by dividing the streaming data into multiple fixed-size groups and extracting the feature value of each group, the efficiency of the method is enhanced.

Table 2: Notions

1) *minWindow*: the minimum size of the sliding window.
2) *maxWindow*: the maximum size of the sliding window.
3) *toleranceFactor*: tolerance factor.
4) *attackCount*: the number of alarms.
5) T_A : adaptive threshold.
6) *offAlarm*: the number of intervals between two consecutive alarms to cancel one alarm.
7) *offFactor*: the number of intervals between two consecutive alarms.

Based on the first step, we elaborate on the main algorithm in the second step. The recursive version of the nonparametric CUSUM algorithm is present in[16], it is shown as follows.

$$\begin{cases} S_n = (S_{n-1} + Z_n)^+ \\ S_0 = 0 \end{cases}$$

where S_n denotes the test statistic and Z_n denotes the current feature value. If $S_n > T_A$, it means there exists an alarm. Fig.3 displays the sketch map of the second step, and Table 2 shows some important notions.

While calculating the cumulative sum of the time-series streaming data, the original CUSUM algorithm usually performs in two ways. One is cumulating the energy value of each element in sequence orderly, the other is to maintain a fixed-size window, and calculate the cumulative CUSUM sum of the window. However, each of the methods has its own shortcomings. For the first, it calculates the global cumulative CUSUM sum, successfully tackles the global attributes of the stream data, yet it losses the focus on recent stream data. The latter method can ease the problem of the first method to some extent, but it still has the problem of delaying the detection of a change point or even not detecting a change in the first place. The method aims to tackle this problem by maintaining a sliding window, its size can be auto-adjusted. Here the tolerance factor is introduced to filter the discrete alarms. Once the number of alarms(denoted by *attackCount*) cumulate up to *toleranceFactor*, an alarm will be produced**.** When the value of *attackCount* reaches *toleranceFactor*, the sliding window will reduce its size to a minimum value gradually. If the alarm continues, the sliding window will remain fixed size until a new change point is detected or the current change point is terminated. When the alarm stops, the factor *offAlarm* is introduced to describe the number of times the sliding window has been swiped after the current alarm ends, that is, for convenience, the method combine two consecutive alarms with an interval smaller than *offAlarm* to be one alarm, if the intervals of two alarms is smaller than *offAlarm*, the previous alarm won't be canceled. After the end of a complete alarm, the window size will gradually resume to the pre-change size.

The threshold parameter T_A is set to judge whether there is an alarm occurs, initially set to 0. During the detection process, the detection threshold is updated according to the value of T_A, if T_A is 0, the threshold will be set as $K_A * Z_n$, and remains constant, where K_A is a preset parameter which is consistent with the requirement on detection delay of the start of an anomaly. Once an alarm occurs, the average of three previous thresholds is employed to update the corresponding threshold, Fig.4 shows the improved idea.

Algorithm 1 Sliding window update

```
1: while data comes in do
2:     if S_n > T_A then attackCount = attackCount + 1
3:         if attackCount > toleranceFactor then
4:             1). reduce window size to minWindow gradually
5:             2). adjust threshold value
6:         end if
7:     end if
8:     if S_n <= T_A then offFactor = offFactor + 1
9:         if offFactor >= offAlarm then
10:            Alarm ends
11:        end if
12:    end if
13: end while
```

Fig.4: Sliding Window Update

3 Experimental Evaluation

In this section, the proposed improved algorithm is verified. The performance metrics include the detection probability (the percentage of attacks for which alarm was raised) and detection delay. For ease of analysis, the method only focuses on a fixed chosen group within an electromagnetic stream data. In addition, we seek to investigate the transformation of the sliding window size and the threshold value in time series. The initial parameters are set as follow, t =10, T_A = 0, *toleranceFactor*= 10, *minWindow*=3, *maxWindow*=10.

3.1 Electromagnetic data set

The use of wireless devices is continuous, there are four main types of wireless devices usage behaviors during monitoring.

1) Continuously uses with high intensity.
2) Continuously uses with low intensity.
3) Intermittent use with high intensity.
4) Alternating intermittent use with high intensity and low intensity.

Four data sets are collected corresponding with the above situations, recorded as DS-1, DS-2, DS-3, and DS-4. Table 3 summarises the characteristics of each data set.

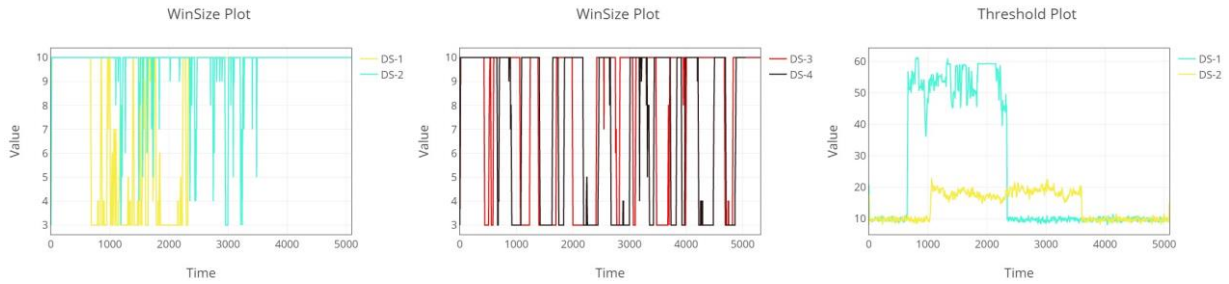
Table 3. Characteristics of Different Data Sets

Dataset	Data Set Size	Attack Intensity	Attack Duration	Number of Attacks
DS-1	5034	high	600s	1
DS-2	5088	low	600s	1
DS-3	5313	variable	417s	9
DS-4	5053	variable	610s	11

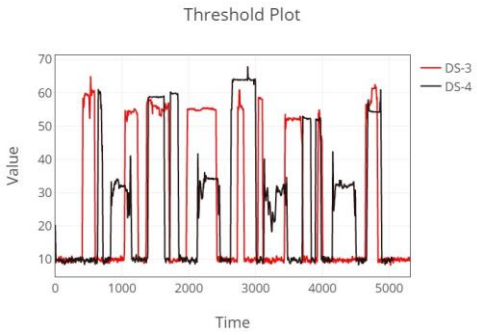
3.2 Result

The improved method is implemented on the above four data sets. Fig.5(a) shows the detection probability on four data sets. The attack behavior of DS-1 and DS-2 is continuous, so the detection probability is relatively higher compared with Dataset DS-3 and DS-4, all the detection probability is greater than 90%, it means that the proposed method can detect almost all the anomalies with high coverage. In data set DS-4, the attack behaviors are variable, among all the attack behaviors, the cumulative sum of some low-intensity attack behaviors could not surpass the corresponding threshold due to lack of accumulation time. Fig.5(b) shows the average detection delay of four data sets, the detection delay of DS-1 and DS-2 were directly calculated, while the detection delay of DS-3 and DS-4 were calculated by the average value of multiple attack behaviors, the detailed detection delays of DS-3 and DS-4 are shown in Fig.5(c).

(a) Detection Probability (b) Detection Delay (c) Detection Delay Details



(d)WinSize Transformation (e)WinSize Transformation (f) Threshold Transformation



(g) Threshold Transformation

Fig. 5: Experiment Results

In addition, to visually reflect the adaptive characteristics of our method, the transformation of the sliding window size and the threshold value of each data set during the detection process are recorded. For the sake of distinction, the window size transformation of DS-1 and DS-2 are put together, DS-3 and DS-4 together, as shown in Fig.5(d) and Fig.5(e) respectively. As can be seen from the figures, the window size adaptively reduced when an attack occurs, and restored to the pre-change size when the attack ends. Fig.5(f) shows the threshold size transformation of DS-1 and DS-2, Fig.5(g) shows the threshold size transformation of DS-3 and DS-4, it can be seen that under different attack behaviors, the detection threshold is also adaptively hanged. Through the experiments, it can be seen that the proposed method avoids computing the mathematical distribution of electromagnetic stream data, by introducing the sliding window mechanism and tolerance factor, the method has good flexibility and efficiency. The proposed method has been used in actual detection.

4 Conclusion

In this paper, a multidimensional nonparametric CUSUM algorithm (H-CUSUM) based on the sliding window and tolerance factor, which divide an electromagnetic stream data into multiple fixed-size groups and calculate the area features in each group as detection feature. The method did these first by eliminating the outliers in each electromagnetic data. During the monitoring process, the cumulative sum of the feature values in the sliding window is calculated, the window size and the threshold value are adaptively adjusted. A tolerance factor is introduced to control the sensitivity of the alarm which can filter invalid alarms. The experiments on four different kinds of data sets are conducted, the performance analysis of the method are carried out through two indicators, namely detection probability, and detection delay. In addition, the transformation of the sliding window size and the threshold value are given. The algorithm has been put into use in the actual monitoring, in the future, we will continue to pay attention to related works in this area.

5 References

[1] Rijo, Luiz. "Modeling of electric and electromagnetic data." (1977).

[2] Babcock, Brian, et al. "Models and issues in data stream systems." Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2002.

[3] Yamanishi K , Takeuchi J I . A unifying framework for detecting outliers and change points from non-stationary time series data[C]// Eighth Acm Sigkdd International Conference. ACM, 2002.

[4] Aggarwal, Charu C., et al. "A framework for clustering evolving data streams."Proceedings of the 29th international conference on Very large data bases-Volume 29. VLDB Endowment, 2003.

[5] Assent, Ira, et al. "Anyout: Anytime outlier detection on streaming data." International Conference on Database Systems for Advanced Applications. Springer, Berlin, Heidelberg, 2012.

[6] Cao, Lei, et al. "Scalable distance-based outlier detection over high-volume data treams." Data Engineering (ICDE), 2014 IEEE 30th International Conference on IEEE, 2014.

[7] Salehi, Mahsa, et al. "Fast memory efficient local outlier detection in data streams." IEEE Transactions on Knowledge and Data Engineering 28.12 (2016): 3246-3260.

[8] Granjon, Pierre. "The CuSum algorithm-a small review." (2013).

[9] Yu, Ming. "A nonparametric adaptive CUSUM method and its application in network anomaly detection." International J. Advancements in Computing Technology 4.1 (2012): 280-288.

[10] Sun, Zhi-Xin, Yi-Wei Tang, and Yuan Cheng. "Router anomaly traffic detection based on modified-CUSUM algorithms." Ruan Jian Xue Bao(Journal of Software) 16.12 (2005): 2117-2123.

[11] Shu, Lianjie, Wei Jiang, and Zhang Wu. "Adaptive CUSUM procedures with Markovian mean estimation." Computational Statistics & Data Analysis 52.9 (2008): 4395-4409.

[12] De Oca, Veronica Montes, et al. "A cusum change-point detection algorithm for non-stationary sequences with application to data network surveillance." Journal of Systems and Software 83.7 (2010): 1288-1297.

[13] Sun, Degang, et al. "An Improved NPCUSUM Method with Adaptive Sliding Window to Detect DDoS Attacks." International Conference on Information and Communications Security. Springer, Cham, 2015.

[14] Benson, Theophilus, Aditya Akella, and David A. Maltz. "Network traffic characteristics of data centers in the wild." Proceedings of the 10th ACM SIGCOM conference on Internet measurement. ACM, 2010.

[15] Miao, Yuantian, et al. "Comprehensive analysis of network traffic data." Concurrency and Computation: Practice and Experience 30.5 (2018): e4181.

[16] Basseville, Mich èle, and Igor V. Nikiforov. Detection of abrupt changes: theory and application. Vol. 104. Englewood Cliffs: Prentice Hall, 1993.