Parallel Outlier Detection for Streamed Data Using Non-Parameterized Approach

Harshad Dattatray Markad, Zeal College of Engineering and Research, Pune, India
S. M. Sangve, Zeal College of Engineering and Research, Pune, India

ABSTRACT

Outlier detection is used in various applications like detection of fraud, network analysis, monitoring traffic over networks, manufacturing and environmental software. The data streams which are generated are continuous and changing over time. This is the reason why it becomes nearly difficult to detect the outliers in the existing data which is huge and continuous in nature. The streamed data is real time and changes over time and hence it is impractical to store data in the data space and analyze it for abnormal behavior. The limitations in data space has led to the problem of real time analysis of data and processing it in FCFS basis. The results regarding the abnormal behavior have to be done very quickly and in a limited time frame and on an infinite set of data streams coming over the networks. To address the problem of detecting outliers on a real-time basis is a challenging task and hence has to be monitored with the help of the processing power used to design the graphics of any processing unit. The algorithm used in this paper uses a kernel function to accomplish the task. It produces timely outcome on high speed multi-dimensional data. This method increases the speed of outlier detection by 20 times and the speed goes on increasing with the increase with the number of data attributes and input data rate.

KEYWORDS

Anomaly Intrusion Detection, Compute Unified Device Architecture (CUDA), Gaussian Detection Scheme, Graphics Processing Unit (GPU), Outlier Detection, Parallel Execution

INTRODUCTION

Outlier detection mechanism is the detecting of a different pattern or an unusual pattern that is different from the rest of the normal data set. Outlier detection is usually done to indicate and identify the defective data or a behavior. Sometimes outlier detection is done to analyses the data for security and scientific interest. Instead of discarding the data, researchers sometimes compose the data pattern in the form of data mining technique so that same pattern can be detected very easily in the near future.

The data incoming to the system is in the form of chunks or streams. The data in the form of chunks are usually datagram. The datagram is similar in size and is possible to detect and identify the outliers in it very easily. When the data is the form of streams, it becomes difficult to analyze the data on the regular basis as the stream is continuous and enormous data comes into the system. As the streamed data is continuous in nature, it becomes nearly impractical to store it in a memory and analyze it thoroughly for abnormal behavior. This gave rise to the use of stream outlier detection mechanism which works on one pass basis. The streamed data is put into the tunnel of data outlier
detection mechanism such that the outlier is detected in a single pass which overcome the need to store the data in the memory and analyze it fragment by fragment.

Although the density-based outlier detection approaches are proven to be accurate, they are also known to be computationally demanding. Therefore, when using kernel density estimation to detect outliers in a high volume, high-speed data stream, we need to speed up the computation to keep up the rate of input that is stream data. The purpose of designing the Graphics Processing Units (GPUs) is to handle high parallel workloads also to execute thousands of threads which are concurrent. With the introduction of CUDA (Compute Unified Device Architecture) (Nair et al., 2011; Hudlicka, 2011; Zhang et al., 2017), it is also called as architecture for general purpose parallel computing and programming model, GPU computing became more and more popular in general-purpose data mining applications. To comply with the real-time processing requirements of streaming data, we use parallel processing powers of GPUs to accelerate kernel density estimation and generate timely manner results. As compare to multi-core implementation our experimental results show that method achieved 20 times higher speed on real-world datasets.

In the mechanism of outlier detection, sometimes the nature of data changes for prolong period of time. Hence it is very difficult to mark the behavior as outlier as soon as the behavior is detected. This gave rise to the mechanism of dividing the data into batches also called as windows. The portion of the data stream called window is analyzed for outlier behavior.

Once the detection algorithm finds the outlier, before marking it as an outlier behavior, it checks the series of windows before and after the outlier-window so that it makes easy to decide the outlier occurrence. It is also said that the entire data set over a period of time has to be considered before coming over to the final decision of abnormal behavior.

As shown in the Figure 1, there are 2 windows having cluster named as A, A’, B, B’ and H. As researchers see that in the first window, there are 2 clusters named as A and A’. As these clusters are separated from each other and has sufficient number of members, researchers cannot consider either of them as an outlier.

The second window also has the same case. The second window has B and B’. It would be very early to consider either of them as an outlier as the behavior of the network or the data generated over the network has a variation. This gave rise to the most dynamic solution of cumulative resulting where the behavior of both the results are considered together to identify the outlier or abnormal behavior.

Once the cumulative result is obtained, the cluster named as H is neither collaborated with other clusters nor has sufficient members to treat it as a cluster. It leads to the final conclusion that the cluster H is an abnormal behavior indentified over the streamed data and can be considered as an outlier.

To address the problem of cumulating, this paper has considered the global view of the data set so as to ensure that no usual data pattern should be marked as outlier. The function used for outlier detection in this paper uses kernel density to find out the statistics of the incoming data packets and store the pattern results to the storage instead of storing the entire data. These patterns are used to analyze the pattern of the current window.

The outlier function which is based on the density of the kernel has to be computationally fit and fast so that to keep up with the fast and continuous incoming of streamed data packets. The introduction of CUDA (Compute unified device architecture) has led to the execution of parallel outlier detection. This increased the speed of our density outlier function. The results are also obtained on timely manner.

**RELATED WORK**

Various methods have been proposed for outlier detection. One such method is K- nearest neighbor method. The method used by Nadungodage et al. (2016) is widely used methods for outlier detection. K-nearest neighbor method uses various training sets where the incoming traffic is tested against the training set. This algorithm can only be used in concrete and stable applications. The data input is unique and never changing. This approach does not make any assumption about the ever-changing...
Related Content

Stiffness Modeling and Analysis of Passive Four-Bar Parallelogram in Fully Compliant Parallel Positioning Stage
[www.igi-global.com/chapter/stiffness-modeling-analysis-passive-four/76439?camid=4v1a](www.igi-global.com/chapter/stiffness-modeling-analysis-passive-four/76439?camid=4v1a)

Survey of Recent Cyber Security Attacks on Robotic Systems and Their Mitigation Approaches
[www.igi-global.com/chapter/survey-of-recent-cyber-security-attacks-on-robotic-systems-and-their-mitigation-approaches/180078?camid=4v1a](www.igi-global.com/chapter/survey-of-recent-cyber-security-attacks-on-robotic-systems-and-their-mitigation-approaches/180078?camid=4v1a)
Development of a Novel Robotic Catheter Manipulating System with Fuzzy PID Control
[www.igi-global.com/article/development-novel-robotic-catheter-manipulating/68864?camid=4v1a](www.igi-global.com/article/development-novel-robotic-catheter-manipulating/68864?camid=4v1a)

Long-Term and Short-Term Traffic Forecasting Using Holt-Winters Method: A Comparability Approach with Comparable Data in Multiple Seasons