UNIVERSITI TEKNOLOGI MARA

AUTONOMOUS ANOMALY DETECTION USING DENSITY-BASED FEATURES IN STREAMING DATA

MUHAMMMAD YUNUS BIN IQBAL BASHEER

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Computer Science)

College of Computing, Informatics and Mathematics

July 2023

ABSTRACT

The rise of Industrial Revolution 4.0 (IR4.0) technology, such as the Internet of Things (IoT), leads to the existence of massive volumes of data. The phenomenon produces a vast volume and variety of data and increases production speed. Consequently, to handle these data, computer algorithms must adapt to their characteristics. Due to its massive volume, variety, and velocity, it contains a lot of insightful patterns. These patterns may include both normal and anomalies data. Anomalies are important to be detected as its existence may require immediate attention and actions. The anomaly data deviate far from normal and may feed wrong information that might lead to wrong decisions and predictions. Hence, it is critical for an anomaly detection algorithm to detect data anomalies patterns. Nonetheless, the process of detecting anomalies in streaming data is laborious. The available algorithms will face difficulties due to the abundance of data produced over time. Furthermore, it needs to operate fast. This research focuses on anomaly detection in streaming data. We built three algorithms to detect anomalies in the streaming data autonomously. These algorithms are data-driven and do not require thresholds or predefined assumptions. They are nonparametric and have no assumptions on the distribution of data. Autonomous anomaly detection (AAD) is enhanced to receive streaming data. It is called multithreaded autonomous anomaly detection for streaming data (MAAD) which works asynchronously while using recursive updates to calculate required mechanisms such as mean and average scalar products. After that, autonomous anomaly detection for streaming data (AADS) is proposed to detect anomalies in any amount of data. The AADS algorithm uses evolving methods which are evolving autonomous data partitioning (eADP) and non-weighted frequency equations. Finally, the AADS is enhanced to operate parallelly, called parallel autonomous anomaly detection for streaming data (PAADS). It is because the parallel mechanism is able to handle high-speed streaming data. The proposed algorithms were evaluated to test their speed in handling streaming data. The performance tests are also conducted to assess whether each algorithm can detect most of the true anomalies. The data is supplied using IoT devices, and benchmark datasets are also presented to test the algorithm's performance. As a result, based on ionosphere benchmark dataset, the proposed PAADS achieved 100% precision and recall rate with no false alarm rate. Meanwhile, MAAD achieves 96% precision and 72% recall rate with 0.5 false alarm rate. In addition, the PAADS algorithm also performs better in another benchmark dataset. Although the proposed algorithms worked well in streaming data, there are several limitations that need to be addressed. The proposed algorithms are computationally efficient but need to enhance so that it is memory efficiency. It is because the algorithms accumulate the incoming data, and there is no forgetting factor used. Finally, this research contributes toward the new foundation of anomaly detection algorithms that operate much better in terms of speed and detecting true anomalies than the previous AAD and streaming TEDA which is a nonparametric and autonomous algorithm.

ACKNOWLEDGEMENT

Alhamdulillah and praise to Allah, the Almighty because of His blessing, I was able to finish this project within the time given provided. A special thanks to my supervisor Dr Azliza Mohd Ali and my co supervisor, madam Nurzeatul Hamimah Abdul Hamid for guiding me in the whole process of completing this project. Thanks to my hero, my father who provide many supports while conducting this project. Not to forget, Allahyarhamah , my mother who act like motivator whenever I become hopeless. Last but not least, I would like to thank all my friends for providing me moral support in order to complete this project.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Research Background

Detecting anomalies in the streaming environment is important so that timely decisions can be made and maintain normal data patterns. However, streaming data needs to be processed as soon as possible because it cannot be paused. Streaming data, such as those produced by the Internet of Things (IoT) devices enter the data cloud in real time (Bomatpalli & Vemulkar, 2016). The harvested IoT data then needs to be analysed for insights. However, as the speed of the arriving data increases, the algorithm needs to adapt. Consequently, critical data could be missed (Kolajo et al., 2019).

Streaming data also produces a variety of data. Hence, the algorithm not only needs to adapt to data velocity, but it also requires agility to handle the variety of data produced. The algorithm must learn from time to time (Angelov, 2014b). Streaming data is one of the data that has speed and dynamically changing in the pipeline (Rettig et al., 2015). Therefore, due to the characteristics of streaming data, the algorithm for handling them must be free from any predefined parameters. The training stage is also not required as we cannot pre-assumed data distribution (Costa et al., 2014). The memory also needs to be used efficiently because memory is limited, and streaming data always accumulates along the time (Bomatpalli & Vemulkar, 2016; Tellis & D'Souza, 2018).

Streaming data carries valuable information, which can be focused on mining useful information from data stream. Unfortunately, the information extracted from data may not always be correct and meaningful. Predictive models built upon imprecise information will inevitably produce poor decisions under certain conditions. The imprecise information is because the presence of anomalies. An anomaly data cannot fit with other normal data (Tellis & D'Souza, 2018). It is also difficult to detect (Foorthuis, 2020) and unpredictable such as when did it exist and how does it look like. Sometime, these anomalies will try to camouflage in order to prevent its identity from being revealed.