Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 4 (2023)

Secure And Adaptive Data Stream Mining For New Generation Big Data

[1] A Ravi Kishore, [2] Dr Gururaj Murtugudde

[1]Research Scholar, Don Bosco Institute of Technology, Affiliated to Visvesvaraya Technological University
[2] Research Supervisor, Don Bosco Institute of Technology, Affiliated to Visvesvaraya Technological University

E-mail: [1] ravikishore4u.2010@gmail.com, [2] gururajmurtu@gmail.com

Abstract: As the era of new-generation big data applications unfolds, the need for secure and adaptive data stream mining has become increasingly paramount. Evolving databases, characterized by ever-changing data streams and dynamic data distributions, present unique challenges and opportunities. This paper addresses the crucial intersection of security and adaptability in the context of data stream mining for new-generation big data. First, we delve into the evolving landscape of big data, where real-time data streams from diverse sources drive decision-making processes. Ensuring the privacy and security of sensitive information within these data streams is a fundamental concern. We explore cryptographic techniques, anonymization methods, and access control mechanisms that safeguard data while allowing for meaningful analysis. We present novel adaptive algorithms and model update strategies that can continuously learn and adjust to changing data distributions. These approaches enable data stream mining to remain effective and accurate over time. This paper offers insights into the fusion of security and adaptability in data stream mining, providing a foundation for the development of robust and privacy-conscious solutions for the evolving landscape of new-generation big data applications.

Keywords: Big Data, Data Stream Mining, Security, Adaptive Algorithm.

1. Introduction

In the contemporary era of technology-driven innovation, the proliferation of data has catalyzed a paradigm shift in how organizations harness information for decision-making, insights, and value creation. This transformation is epitomized by the emergence of new-generation big data applications, characterized by the ingestion, analysis, and utilization of vast volumes of data in real-time or near real-time [1]. In this dynamic landscape, the ability to secure and adaptively mine data streams has emerged as a pivotal challenge and an imperative opportunity [2].

Traditional data analysis paradigms, which often rely on static datasets and batch processing, are ill-equipped to cope with the velocity, variety, and volume of data generated by new-generation applications [3]. The hallmark of these applications is the continuous influx of data streams from diverse sources such as sensors, social media, IoT devices, and more [4]. These data streams are not only vast but also ephemeral, with constantly evolving data distributions, making them inherently challenging to analyze and harness effectively [5].

This paper delves into the intersection of two critical facets of data stream mining for new-generation big data applications: security and adaptability. The security aspect addresses the vital concern of safeguarding sensitive information within data streams [6]. As organizations strive to leverage these streams for insights and decision-making, they must simultaneously protect the privacy and integrity of the data. The adaptability facet is equally indispensable, as static models and algorithms fall short in the face of dynamic data distributions and evolving patterns.

Security in Data Stream Mining: The security dimension of data stream mining encompasses the protection of sensitive information, compliance with data privacy regulations, and guarding against unauthorized access or data breaches [7]. The urgency of these concerns cannot be overstated, given the potential consequences of data leaks or misuse.

Security measures encompass various techniques, including cryptographic methods to ensure data confidentiality, anonymization approaches to protect privacy, and access control mechanisms to restrict data access based on roles and permissions [8]. These measures must be seamlessly integrated into data stream mining processes to enable secure analysis while upholding privacy principles.

Adaptability in Data Stream Mining: The adaptability aspect addresses the dynamic nature of data streams. In contrast to batch processing, where models are trained on static datasets, data stream mining requires continuous learning and adjustment to evolving data distributions [9]. Traditional machine learning models can degrade in performance over time as the underlying data evolves, necessitating the development of adaptive algorithms and model update strategies.

Adaptive data stream mining techniques enable models to autonomously recognize concept drift (changes in data patterns) and recalibrate themselves accordingly. These approaches ensure that data stream mining remains effective, accurate, and relevant in the face of shifting data landscapes [10].

The Crucial Intersection: The confluence of security and adaptability in data stream mining presents a multifaceted challenge. Ensuring data security without compromising the adaptability and utility of data stream mining techniques is a delicate balance to strike. This paper explores the strategies, methodologies, and technologies that enable this balance, enabling organizations to harness the full potential of new-generation big data while preserving data security and analytical precision.

Furthermore, we discuss the practical implications of secure and adaptive data stream mining in various domains, including healthcare, finance, Internet of Things (IoT), and cybersecurity. These domains exemplify the diverse applications and pressing use cases where the fusion of security and adaptability in data stream mining is of paramount importance.

In the subsequent sections of this paper, we delve into the principles, techniques, real-world applications, and future directions of secure and adaptive data stream mining for new-generation big data. By addressing this critical intersection, we contribute to the advancement of data-driven decision-making in an era defined by dynamic data streams and evolving data landscapes.

2. Related Works

In this article [11], presents a thorough review of execution platforms for Data stream mining applications. In addition, a detailed taxonomic discussion of heterogeneous MDSM applications is presented. Moreover, the article presents detailed literature review of methods that are used to handle heterogeneity at application and platform levels. Finally, the gap analysis is articulated and future research directions are presented to develop next-generation Data stream mining applications.

The techniques are part of a reactive security strategy because they rely on the human factor, experience and the judgment of security experts, using supplementary technology to evaluate the risk impact and minimize the attack surface. This study [12] suggests an active security strategy that adopts a vigorous method including ingenuity, data analysis, processing and decision-making support to face various cyber hazards. Specifically, the paper introduces a novel intelligence driven cognitive computing security that is based exclusively on progressive fully automatic procedures.

The increase in volume, speed, and variety of data requires a more robust, accurate intrusion detection system capable of analyzing a huge amount of data. This work [13] proposes the creation of an intrusion detection system using stream classifiers and three classification layers - with and without a reduction in the number of features of the records and three classifiers in parallel with a voting system.

The popularizing of various types of network has diversified types, issues, and solutions for big data more than ever before. In this paper [14], provides recent research in data types, storage models, privacy, data security, analysis methods, and applications related to network big data. Finally, summarized the challenges and development of big data to predict current and future trends.

In order to extract valuable knowledge from data streams, one must be able to analyze the data as they arrive and make meaningful predictions. For this purpose, used fast incremental learners. There already exists a great community that is organizing various competitions on machine learning tasks for batch learners [15].

In this paper [16], a systematic review of big data streams analysis which employed a rigorous and methodical approach to look at the trends of big data stream tools and technologies as well as methods and

techniques employed in analysing big data streams. It provides a global view of big data stream tools and technologies and its comparisons.

The existing AI techniques that function in isolation exhibit clear limitations in developing a comprehensive platform due to the dynamicity of big data streams, high-frequency unlabeled data generation from the heterogeneous data sources, and volatility of traffic conditions. In this paper [17], proposed an expansive smart traffic management platform (STMP) based on the unsupervised online incremental machine learning, deep learning, and deep reinforcement learning to address these limitations.

Data mining technology can search for potentially valuable knowledge from a large amount of data, mainly divided into data preparation and data mining, and expression and analysis of results [18]. It is a mature information processing technology and applies database technology. The data in the database are processed and analyzed by studying the underlying theory and implementation methods of the structure, storage, design, management, and application of the database.

In this paper [19], used a systematic methodology to review the strengths and weaknesses of existing open-source technologies for big data and stream processing to establish their usage for Industry 4.0 use cases.

The massive streaming data generated and captured by smart building appliances and devices contain valuable information that needs to be mined to facilitate timely actions and better decision making. Machine learning and big data analytics will undoubtedly play a critical role to enable the delivery of such smart services. In this paper [20], gives the area of smart building with a special focus on the role of techniques from machine learning and big data analytics.

3. Proposed Model

Our proposed model encompasses key components and methodologies for addressing security and adaptability in dynamic data stream environments. Begin by providing a brief outline of the challenges posed by data stream mining in new-generation big data applications, emphasizing the need for both security and adaptability.

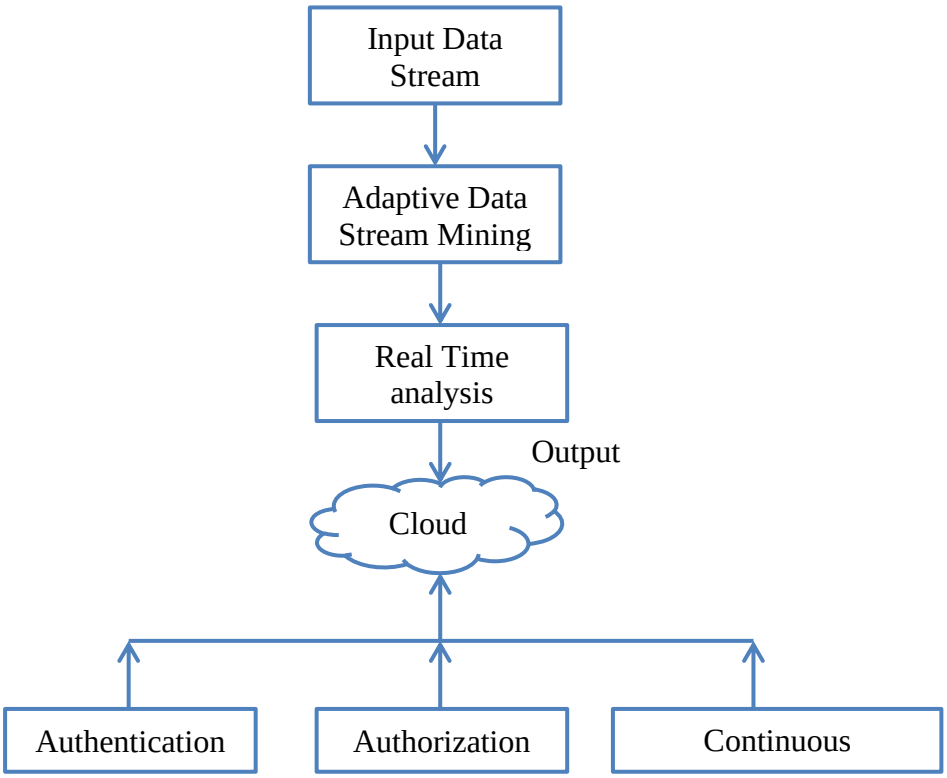

Fig 1: Architecture Diagram for Proposed Model

3.1 Input Data Stream

In the context of a Secure and Adaptive Data Stream Mining Model for New Generation Big Data, real-time events can serve as the input data stream. These events can be continuously generated and processed for various applications. The continuous data stream (Xt) consists of real-time events. Each event is represented as , where i is the index of the event within the stream at time t. Define a suitable format for representing each event, including its attributes (e.g., timestamp, source, event type, data fields).

3.2 Adaptive Data stream Mining

Adaptive Data Stream Mining refers to the process of continuously adjusting and optimizing data mining models and algorithms to cope with changing patterns, distributions, and characteristics of streaming data. In dynamic environments where data evolves over time, traditional static models may become outdated and less effective. Adaptive data stream mining aims to address these challenges by enabling models to learn and evolve alongside the data. This model incorporates adaptive learning mechanisms that can continuously analyze and adapt to changing data distributions.

Maintain model parameters (Θt) that adapt to the evolving characteristics of the real-time events.

Use the model to make real-time predictions or classifications based on the incoming events $(yt = f(xti, \Theta t))$.

Control the adaptation speed with a learning rate (αt). Update model parameters $(\Theta t + 1)$ based on detected concept drift and real-time events.

3.3 Real time analytics

Real-time analytics, also known as real-time data analytics or streaming analytics, refers to the process of analyzing and processing data as it is generated or ingested, enabling organizations to gain immediate insights and take rapid actions based on the latest information. This approach contrasts with traditional batch processing, where data is collected and processed in fixed time intervals. Real-time analytics is crucial in various domains and applications, including finance, healthcare, cybersecurity, IoT, and more.

Let Dt be a drift indicator at time t. $Dt = 1$ indicates a detected drift, and $Dt = 0$ otherwise.

3.4 Secure Data Handling

Monitor and compute a security metric St to ensure data privacy and integrity throughout the process.

Encrypt sensitive data within the stream using a strong encryption algorithm $E(x, k)$.

Apply anonymization techniques $A(x)$ to protect individual privacy.

Control adaptation speed with a learning rate αt.

Update model parameters using $\Theta t + 1 = \Theta t + \alpha t \cdot \Delta\Theta t$ based on detected concept drift.

Implement access control mechanisms $AC(xti))$ to restrict data access within individual events based on roles and permissions.

Ensure data integrity through hash functions $H(xti))$ applied to event data.

Quantify the trade-offs between data security (privacy preservation) and model adaptation (accuracy) for real-time events to make informed decisions.

3.5 Pseudocode for Proposed Model

```
Initialize model parameters, security mechanisms, and adaptation settings
while True do:
   # Continuously ingest real-time data stream
   data_point = ingest_data_stream()
   # Secure data handling
   encrypted_data = encrypt_data(data_point)
   anonymized_data = anonymize_data(encrypted_data)
      # Real-time prediction/classification
   prediction = predict(anonymized_data, model_parameters)
      # Concept drift detection
   drift_detected = detect_concept_drift(anonymized_data)
```

```
if drift_detected:
    # Model adaptation
    update_model_parameters(anonymized_data, model_parameters)
    # Security monitoring
security_metric = monitor_security(anonymized_data)
# Evaluate and store results
accuracy = evaluate_model(prediction, true_label)
log_results(accuracy, security_metric)
    # Check for real-time alerts
if security_metric >= security_threshold:
    raise_alert("Security breach detected!")
# Implement real-time actions based on predictions and alerts
take_real_time_actions(prediction, security_metric)
```

Continuously monitoring all incoming real-time events for immediate decision-making and insights. By incorporating real-time events as the input data stream, the model can adapt to changing patterns and distributions, ensuring both data security and effective data analysis in a continuously evolving data landscape.

4. Results and Discussions

By analyzing large volumes of data, patterns can be identified and used to develop models that can help to optimize real time data streaming and improve its performance as given in table 1.

Table 1: Methodology used for processing with total time taken

Methodology	No. of Processed records/sec	Total Time Taken
Kaggle datasets	143	5.5 mins
Hive (20 tasks)	263	4.1 mins
Postgre SQL	354	6.7
Hadoop	172	6.2
Our Proposed Method	2850	1.3

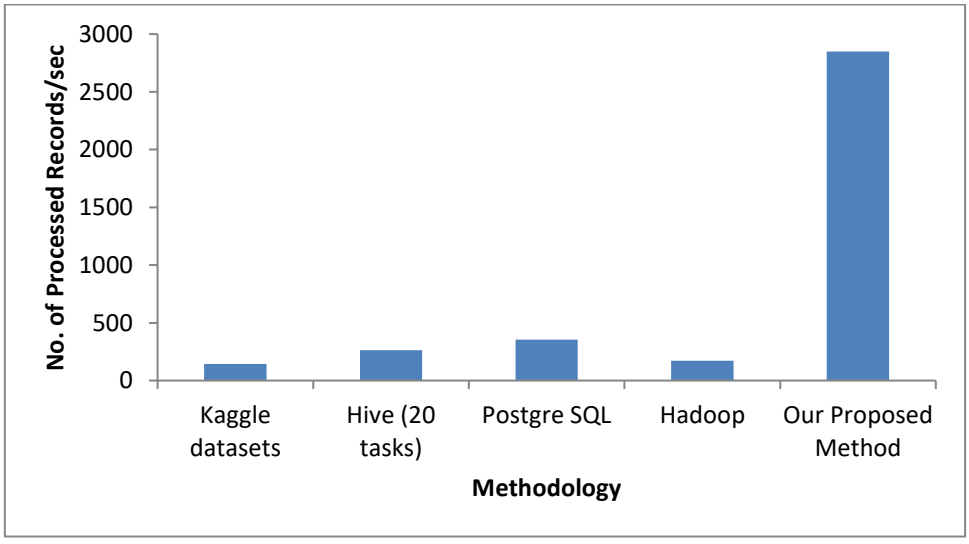

Fig 2: Number of Processed Records / second

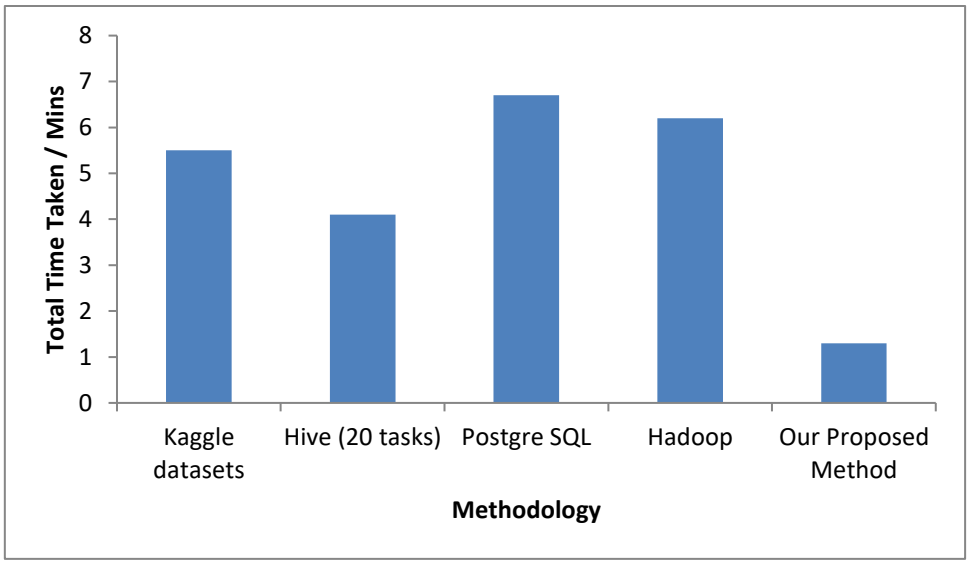

Fig 3: Total Time Taken for different methodologies

From the above results shown in fig 2 and 3, it clearly shows that our proposed model gives better results than other methologies.

5. Conclusions

This proposed model provides a structured framework for tackling the complexities of secure and adaptive data stream mining in the context of new-generation big data applications. Tailor the model and equations to your specific use case and research requirements while keeping a balance between data security and model adaptability. Summarize the key findings, emphasizing the significance of the model in addressing the challenges of secure and adaptive data stream mining with real-time events in dynamic big data environments. By incorporating real-time events as the input data stream, the model can adapt to changing patterns and distributions, ensuring both data security and effective data analysis in a continuously evolving data landscape. Adaptive Data Stream Mining is essential for various applications, including fraud detection, network monitoring, financial forecasting, and anomaly detection in IoT. It enables organizations to extract valuable insights and make informed decisions in dynamic, high-velocity data environments.

References

[1] Habeeb, R. A. A., Nasaruddin, F., Gani, A., Hashem, I. A. T., Ahmed, E., & Imran, M. (2019). Real-time big data processing for anomaly detection: A survey. *International Journal of Information Management*, *45*, 289-307.

[2] Mehmood, E., & Anees, T. (2020). Challenges and solutions for processing real-time big data stream: a systematic literature review. *IEEE Access*, *8*, 119123-119143.

[3] Li, W., Koo, C., Hong, T., Oh, J., Cha, S. H., & Wang, S. (2020). A novel operation approach for the energy efficiency improvement of the HVAC system in office spaces through real-time big data analytics. *Renewable and Sustainable Energy Reviews*, *127*, 109885.

[4] Cakir, A., Akın, Ö., Deniz, H. F., & Yılmaz, A. (2022). Enabling real time big data solutions for manufacturing at scale. *Journal of Big Data*, *9*(1), 1-24.

[5] Sanla, A., & Numnonda, T. (2019, July). A comparative performance of real-time big data analytic architectures. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (pp. 1-5). IEEE.

[6] Awan, M. J., Farooq, U., Babar, H. M. A., Yasin, A., Nobanee, H., Hussain, M., ... & Zain, A. M. (2021). Real-time DDoS attack detection system using big data approach. *Sustainability*, *13*(19), 10743.

[7] Zaki, N. D., Hashim, N. Y., Mohialden, Y. M., Mohammed, M. A., Sutikno, T., & Ali, A. H. (2020). A real-time big data sentiment analysis for iraqi tweets using spark streaming. *Bulletin of Electrical Engineering and Informatics*, *9*(4), 1411-1419.

[8] Arulkumar, V., Sridhar, S., Kalpana, G., & Guruprakash, K. S. (2022). Real-Time Big Data Analytics for improving sales in the Retail Industry via the use of Internet of Things Beacons. In *Expert Clouds and Applications: Proceedings of ICOECA 2022* (pp. 111-126). Singapore: Springer Nature Singapore.

[9] Tu, L., Liu, S., Wang, Y., Zhang, C., & Li, P. (2020). An optimized cluster storage method for real-time big data in Internet of Things. *The Journal of Supercomputing*, *76*, 5175-5191.

[10] Watkins, D. (2021). Real-time big data analytics, smart industrial value creation, and robotic wireless sensor networks in Internet of things-based decision support systems. *Economics, Management, and Financial Markets*, *16*(1), 31-41.

[11] ur Rehman, M. H., Liew, C. S., Wah, T. Y., & Khan, M. K. (2017). Towards next-generation heterogeneous mobile data stream mining applications: Opportunities, challenges, and future research directions. *Journal of Network and Computer Applications*, *79*, 1-24.

[12] Demertzis, K., Tziritas, N., Kikiras, P., Sanchez, S. L., & Iliadis, L. (2019). The next generation cognitive security operations center: adaptive analytic lambda architecture for efficient defense against adversarial attacks. *Big Data and Cognitive Computing*, *3*(1), 6.

[13] Schuartz, F. C., Fonseca, M., & Munaretto, A. (2022, October). A Distributed Platform for Intrusion Detection System Using Data Stream Mining in a Big Data Environment. In *2022 6th Cyber Security in Networking Conference (CSNet)* (pp. 1-7). IEEE.

[14] Lv, Z., Song, H., Basanta-Val, P., Steed, A., & Jo, M. (2017). Next-generation big data analytics: State of the art, challenges, and future research topics. *IEEE Transactions on Industrial Informatics*, *13*(4), 1891-1899.

[15] Boulegane, D., Radulovic, N., Bifet, A., Fievet, G., Sohn, J., Nam, Y., ... & Choi, D. W. (2019, December). Real-time machine learning competition on data streams at the IEEE big data 2019. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 3493-3497). IEEE.

[16] Kolajo, T., Daramola, O., & Adebiyi, A. (2019). Big data stream analysis: a systematic literature review. *Journal of Big Data*, *6*(1), 47.

[17] Nallaperuma, D., Nawaratne, R., Bandaragoda, T., Adikari, A., Nguyen, S., Kempitiya, T., ... & Pothuhera, D. (2019). Online incremental machine learning platform for big data-driven smart traffic management. *IEEE Transactions on Intelligent Transportation Systems*, *20*(12), 4679-4690.

[18] Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., ... & Lyu, J. (2020). Brief introduction of medical database and data mining technology in big data era. *Journal of Evidence-Based Medicine*, *13*(1), 57-69.

[19] Sahal, R., Breslin, J. G., & Ali, M. I. (2020). Big data and stream processing platforms for Industry 4.0 requirements mapping for a predictive maintenance use case. *Journal of manufacturing systems*, *54*, 138-151.

[20] Qolomany, B., Al-Fuqaha, A., Gupta, A., Benhaddou, D., Alwajidi, S., Qadir, J., & Fong, A. C. (2019). Leveraging machine learning and big data for smart buildings: A comprehensive survey. *IEEE Access*, *7*, 90316-90356.