Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

Survey Paper on Various Techniques of Privacy Preservation in Data Stream Mining

Nikita Bhegade [1], Sayali Kanase [2], Shreyas Kashetwar [3],Vaishnavi Chavan [4]
*Department of Computer Engineering,
Pimpri- Chinchwad College of Engineering, Pune, India

***Abstract* - With a tremendous increase in the need of mining live streaming data, it has resulted in the interest in the emerging field of data stream mining. The issue of privacy persists in it as sensitive data is involved in the streaming data. This paper describes the different approaches or techniques used for preserving the privacy of streaming data in real time data stream mining. Privacy preservation is one of the major challenges faced by mining of data streams. But time-to-time several techniques are proposed to address this challenge. In this survey paper we describe and compare the different approaches proposed for the same.**

Keywords- Cryptography, data anonymization, data stream mining, data security, privacy preservation, secure real-time stream processing

1. INTRODUCTION

a. Data Mining

Data mining is the extraction of knowledge from static data or finding the patterns from large sets of data. It is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is the analysis step of the "knowledge discovery in databases" process.

b. Data Stream Mining

Data stream is a continuous flow of data or Data stream is an ordered sequence of data elements that can be shown as S = {S1 , S2 , … , Sn(s)} where n(s) tends to infinity i.e. n(s)→∞.

Data stream mining is an extraction of knowledge from a continuous stream of data and processing data using incremental learning methods. It is the process of extracting knowledge structures from continuous, rapid data records .Data stream mining can be read only once or a small number of times using limited computing and storage capabilities.

c. *Inapplicability Of Data Stream Mining*

There are 3 main features in data stream mining which causes an inapplicability of standard data mining algorithms.

i. Large and infinite number of data lengths.
ii. Data arrival rate at the system is high.
iii. There is a potential change of different kinds in data distribution during data stream mining.

Due to the above mentioned features of streaming data, the various data mining techniques implemented on static data sets are inapplicable to be used for dynamic data sets.

2. BASIC METHODOLOGIES IN DATA STREAM MINING & CHALLENGES

As streaming data is characterized by large volume, variety and velocity, there is a need for efficient algorithms and techniques to address the mentioned issues for effective data processing to get the accurate desired results. Some of the techniques used are - Random sampling, Sliding Windows, Histograms, Multiresolutional Methods, Sketches and Randomized algorithms , Damped Window Model . The various proposed methodologies for information extraction from streaming data are- Clustering, Classification, and Association.

Data stream mining faces three principal challenges: volume, velocity, and volatility. Other challenges and research issues include concept drift, skewness of data, handling of incomplete or delayed information, infinite Data length, handling of the steady stream of information, effective result presentation from the mining data, scalability of the systems, as well as privacy preservation of the data in the process of extraction from the stream of data. Here our topic of study is about the privacy preservation challenge in data stream mining.

3. PRIVACY PRESERVATION IN DATA STREAM MINING

A. Introduction

Privacy preservation on data is a crucial process in data mining and data stream mining. Privacy preserving data mining methods are ever-demanding for secured and reliable information exchange over the internet. Various methods are developed for preserving the data privacy. The data needs to be secured from various attackers, hackers and unauthorized accesses that can lead to misuse of the data for malicious or other destructive activities that can lead to a higher level impact to the privacy of streaming data. The methodologies for preserving the privacy of sensitive data in mining continuous streams of data is very limited and an undergoing topic of research.

B. Methodologies For Privacy Preservation In Data Stream Mining

1. Introduction

The methodologies, techniques and algorithms proposed so far use the following approaches in general-

i. Data Anonymization techniques

Data anonymization means selecting and removing or encrypting personally identifiable data from data sets. There are following techniques for data anonymization:

a. K-anonymity

For data anonymization k-anonymity is the most efficient and secure technique used in data but traditional k-anonymization can not be used directly on streaming data. Therefore to achieve k-anonymity on streaming data there are two methods to achieve k-anonymity are suppression and generalization with a quasi identifier.

b. L-diversity

L-diversity is important when k-anonymity cannot provide security to sensitive data with lack of diversity in the data set and if the attacker has some knowledge of the data set. L-diversity is a method of diversifying the sensitive information on data streams in the block.In l-diversity values of sensitive information are well represented in the group.

c. SKY

SKY (Stream K-anonymity) is an algorithmic technique for continuously facilitating k-anonymity on data streams. The method tries to retain as much as possible information the output of the anonymized data stream.

d. CASTEL

CASTEL (Continuously Anonymizing Stream Data via Adaptive Clustering) is a clustering based technique which anonymizes data on the fly and simultaneously preserves the freshness of anonymized data by satisfying specified delay constraints.

e. UBDSA

UBDSA (Utility Based Approach for Data Stream Anonymization) has two optimization objectives: 1) Minimizing Data aging by average delay minimization. 2) Minimizing data loss to improve data quality. In UBDSA data utility is a function for data quality and data aging and therefore UBDSA tries to maximize data utility

.

ii. Cryptographic techniques

Cryptography is a technique of encrypting and decrypting data to preserve the privacy of the data. Cryptographic technique is generally used when there is data sharing over the internet between different parties.There are different cryptographic algorithms for Preserving the privacy of data stream are as follow:

a. Public - key cryptography algorithm

In public - key cryptography data can be encrypted with public key and can only be decrypted with private key. It is an asymmetric encryption technique.

b. RSA (Rivest - Shamir - Adleman) algorithm

RSA is a public-key cryptosystem. In RSA a user generates and publishes a key based on two large prime numbers with auxiliary value.

c. AES (Advanced Encryption Standard) algorithm

AES is a symmetric block cipher technique. There are different modes of AES for data stream cryptography CTR mode is used. In CTR (Counter Mode) Encryption and Decryption is performed using many threads at a time.

d. RC4

RC4 is a stream cipher symmetric key algorithm. In RC4 the same key used for encryption as well as decryption and data stream is XORed with the generated key by RC4 algorithms.

iii. Differential Privacy techniques

iv. Perturbation Technique

Perturbation is used to preserve the privacy of the statistical data . Perturbation is a technique in which original data value is changed with synthetic data value so that there should be no difference between statistical information computed from perturbed data and original data.

C. Literature Survey Of Different Methodologies proposed so far

As discussed in the paper titled “Secure Stream Processing for Medical Data” [1] the following are mentioned-

The objective of the paper is to provide an efficient method for privacy preservation of real time streaming medical patients records data. For which the authors have claimed to have contributed a proof of concept for a privacy-preserving streaming platform for patients medical records data, and a study on the overhead introduced by privacy-preserving processing techniques. The methodologies they used were - A server side equipped with INTEL SGX, and a client package

With a sensor and a gateway included, Multiple sensors monitor cardiac data which is processed on the cloud as input, SGX-SPARK that is a modified version of Apache spark used at server side, Critical parts of processing (RR-intervals) are processed inside enclaves.

The conclusions derived were (i) Proposing of a streaming platform that grants executions on remote, untrusted, servers or clouds with data and code confidentiality and integrity. (ii)Health sensitive data from an untrusted cloud provider is protected.

The need for privacy preserving of patients sensitive medical report data is explained and thus a need to come for some solutions is stated. The focus here is on the granting of access control of the medical data. ‘Apache Spark’ is used for stream processing. Specific algorithmic approach is not used; rather an architectural structure is proposed which provides enclaves for data (similar like an encapsulation) and thus protects its privacy.

The authors Alireza Jolfaei, Krishna Kant and Hassan Shafei in the paper “Secure Data Streaming to Untrusted Road Side Units in Intelligent Transportation System ” [2] have proposed a light weight mechanism for privacy preservation of IOT sensor streaming data and have used “permutation only encryption ” algorithm. Their objective being- To preserve privacy of streaming data generated by vehicles sent to monitoring traffic sensors. The methodologies discussed throughout the paper are – formation of vehicles form groups and choose a leader that collects data from other vehicles and communicates with RSUs.

For access control, division of streams into chunks and enforce time/event based access to them. The encryption is done by a leader vehicle which allows vehicle ID obfuscation, easier key management, and allows operation on encrypted data.

The contribution of the paper being a lightweight permutation mechanism for preserving the confidentiality of sensory data is proposed. The conclusion was derived that the privacy of streaming vehicle data collected by road traffic sensors is preserved by applying the proposed algorithm.

The paper titled as “Secure Processing of Stream Cipher Encrypted Data Issued from IOT Application to a connected Knee Prosthesis” [3] has discussed the following points –

The objective being to propose a secure protocol that allows processing encrypted data emitted by a medical IOT device. The authors in this paper also throw light on the need of privacy preservation in data stream mining. The input they have used to address the condition is Medical IOT streaming data which is a major source of continuous streaming data. A lightweight stream cipher CLGG (Combined Linear Congruential Generator) is used to encrypt data at the IMD (Implantable Medical Device) side and of a new cryptostream conversion protocol, Stands on a cryptosystem conversion and a data packing strategy.

The components involved in the proposed approach are - Connected Implantable medical device framework, (the framework being a part of the Followknee project being developed for a connected knee prosthesis), a Cryptosystem conversion (CrC) which converts a CLGG encrypted data into a D-J encrypted data, Processing of Stream Cipher Encrypted Data and Data packing and communication complexity reduction. It further allows jointly filtering-thresholding data. Experimental results show that our solution is practical in real application contrary to the state of the art based on fully homomorphic cryptosystems. The authors claim that any IOT having enough capacity to implement a CLCG cryptosystem can benefit from the proposed solution.

Some papers have suggested an architecture and have used software for stream processing like “Spark Streaming” and “Apache Storm” for the implementation of some algorithms. The paper title “Spark Framework for Real-Time Analytic of Multiple Heterogeneous Data Streams” [4]

The authors in the paper have proposed a single Spark Streaming Application to process large amounts of heterogeneous IOT data which is an important issue regarding processing of mixed and varying data. Using Spark Structured Streaming, this paper introduces a Spark Streaming framework for multiple heterogeneous data streams which allows the deployment of multiple heterogeneous data stream processing. The proposed framework integrates multiple data streams into one single Spark application, which is not possible in the usual Spark application as it can handle only a single data stream at a time. The proposed framework also utilizes Spark’s FAIR Scheduling scheme to resolve job scheduling problems.

With this proposed framework, the authors claim to reduce the amount of Spark-Submit required; also minimizing the coding redundancy in Spark application, thereby reducing the monitoring difficulty caused by running a large amount of Spark application, and finally solving the scheduling problem of job queuing by using Spark’s FAIR scheduling mode to assign each stream an equal share of resources. Such Framework if implemented successfully will be able to process real time heterogeneous streams and privacy preserving algorithms and techniques can be very well integrated in them.

The authors Bunrong Leang , Rock-Won Kim , Kwan-Hee Yoo - “Real-Time Transmission of Secured PLCs Sensing Data” [5] gives the cryptographic technique for preserving privacy of PLCs data using Apache Kafka and Apache spark.

Secure transmission of PLCs data takes place through 3 layers of the proposed system. Sensing layer , Kafka server layer , data consumption and processing layer are three main layers. Kafka producer encrypts the data using the public key. provided by the kafka consumer and then it sends the data to the kafka server.

When a kafka server receives a message from a kafka producer , immediately the consumer has to be notified to subscribe to the new message. lasly , the consumer describes the message and uses the private key to decrypt the message.

In the consumer section Apache Spark is used for a real time data processing and analysis on the decrypted message. After that process and analytical result is stored in the database or visualized to the client.

4. CONCLUSION

Need of mining of data streams has raised a lot of challenges in the area of research. In this paper, we have addressed the issue of privacy preservation in data stream mining, its essential need for secure stream processing in various applications, the various techniques and algorithms being proposed and implemented in real-time stream processing applications so far have also been discussed by the literature survey prepared.

5. REFERENCES

[1] Maxime Pistono, Reda Bellafqira and Gouenou Coatrieux –“Secure Processing of Stream Cipher Encrypted Data Issued from IOT Application to a connected Knee Prothesis” , 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.

[2] Carlos Segarra, Enric Muntan´e, Mathieu Lemay, Valerio Schiavoni, and Ricard Delgado-Gonzalo – “Secure Stream Processing for Medical Data” , 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.

[3] Alireza Jolfaei, Krishna Kant, Hassan Shafei -“Secure Data Streaming to Untrusted Road Side Units in Intelligent Transportation System ” , 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 13th IEEE International Conference on Big Data Science And Engineering, 2019.

[4] Tanwa Sirisakdiwan, Natawut Nupairoj- "Spark Framework for Real-Time Analytic of Multiple Heterogeneous Data Streams", 2019 2nd International Conference on Communication Engineering ,and Technology, 2019.

[5] Bunrong Leang , Rock-Won Kim , Kwan-Hee Yoo - "Real-Time Transmission of Secured PLCs Sensing Data" , 2018 IEEE Confs on Internet of Things , Green Computing and Communication Cyber, Physical and Social Computing , Smart Data , Blockchain , Computer and Information Technology , Congress on Cybermatics , 2018.

[6] Yifan Tian , Jiawei Yuan , Yantian Hou -"PDF-DS: Privacy-Preserving Data Filtering for Distributed Data Streams in Cloud" , 2019 International Conference on internet of Things (iThings) and IEEE Green Computing and Communication (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) , 2019.

[7] Syed Navqvi , Sean Endervy , Ian Williams , Waqar Asif , Muttukrishnan Rajarajan , Cristi Potlog , Monica Florea - "Privacy-Preserving Social Media Forensic Analysis for Preventing Policing of Online Activities" , 2019 IEEE

[8] Adrian Pérez-Resa , Miguel Gracia-Bosque , Carlos Sánchez-Azqueta , Santiago Celma - "A new Method for Format Preserving Encryption in High-Data Rate Communication" , 2020 IEEE.

[9] Jinyan wang , Chaoji Deng , XianXian Li - "Two Privacy-Preserving Approaches for Publishing Transactional Data Stream " 2018 IEEE Translations and content mining 2018.