# An Enhanced Privacy Preserving Techniques for Asynchronous Streaming Data Mining in Distributed Environment

S. Hariraman[1], Dr. S. Velmurugan[2]
[1]Assistant Professor, Department of Computer Science Engineering,
Thiruvalluvar College of Engineering and Technology, Vandavasi-604505, Tamil Nadu, India
[2]Professor & Head, Department of Electronics and Communication Engineering,
T.J.S. Engineering College, T.J.S. Nagar, Kavaraipettai, Chennai-601206, Tamil Nadu, India

***Abstract*:- This research deals with the study and analysis of privacy preserving techniques in collaborative data mining in order to improve the efficiency and effectiveness of asynchronous streaming data mining in distributed environment. Data mining is a process which uses different data analysis tools that discover patterns and relationships in data that can be used to make predictions. Security and Privacy protection has been a public policy firm for decades. However, rapid technological changes, the fast growing of the internet source and electronic digital source, and the development of more sophisticated methods of collecting, analyzing, and using personal data have made privacy a major public and government issues. The field of data mining is yield significance recognition to the availability of large amounts of data, easily collected and stored via computer systems.**

***Keywords: Data mining* in distributed environment *using CFS, CRM, FSP, PPDM***

## 1. INTRODUCTION

Currently, the large amount of data, accumulate from various channels, contains much personal data. When personal and sensitive data are published and/or analyzed, one important question to take into account is whether the analysis violates the privacy of individual user's whose data is referred to. The importance of information data that can be used to increase revenue cuts costs or both. Data mining software is one of a number of analytical tools for data analyze. It allows users to data analyze in privacy is growing constantly. For this reason, many research works have focused on privacy-preservation technology for asynchronous streaming data mining in a distributed environment, proposing novel techniques that allow extracting knowledge while trying to protect the privacy of users [1].

Data mining involves six common classes of tasks:

***Anomaly detection**:*

The identification of unusual data records, that might be interesting or data errors that require further investigation.

***Association rule learning**:*

Searches for relationships between variables, for example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

**Clustering:** It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

**Classification:** It is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

**Regression:** It attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.

**Summarization:** It providing a more compact representation of the data set, including visualization and report generation.

### *1.1 Privacy Preserving Techniques*

The privacy preserving techniques in data mining is the process that incorporates privacy fear when two parties are having secret databases and they wish to run a data mining algorithm on the unique of their databases without informative any redundant information. It is used to protect the privileged information and enable the use of research. As the input in data mining concentrates on massive data set, privacy preservation techniques is the best method which preserves the data in aggregated form. Due to the various types of privacy provided in data mining, the field has the highest advantage when concerned towards privacy factor.

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through data aggregation. Data aggregation involves combining data together (possibly from various sources) in a way that facilitates analysis (but that also might make identification of private, individual-level data deducible or otherwise apparent). This is not data mining peruse, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data were once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when the data were originally anonymous [2].

### *1.2 Security Issue in Data Mining*

In present years, privacy-preserving techniques in data mining have been studied broadly, because of the wide proliferation of sensitive electronic digital information on the internet. This has leading to increase interest about the privacy of the basic data. Currently, a number of techniques have been proposed for modifying or transforming the data in such a way to preserve privacy. Also privacy-preserving techniques in data mining find various applications in surveillance which is naturally supposed to be “privacy-violating” applications. The key is to design methods which continue to be effective, without compromising security. A number of algorithmic techniques have been designed for privacy-preserving in data mining such as bio-surveillance, facial de-identification, and identity theft using the following methods [3].

***i. Mining Methodology and User Interaction***

- Mining different kinds of knowledge in database
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data Mining query language and ad-hoc data mining
- Expression and visualization of data mining results
- Handling noise and incomplete data
- Pattern evaluation

***ii. Performance and Scalability***

- Efficiency and scalability of data mining algorithms
- Parallel, distributed and incremental mining methods

***iii. Issues Relating to the diversity of Data Type***

- Handling relational and complex types of data
- Mining information from heterogeneous databases and global information systems like web database.

***iv. Issues Related to Applications and Social Impacts***

- Application of discovered knowledge, domain specific data mining tools, intelligent query answering, decision making.

### *1.3 Asynchronous Data*

Asynchronous data is data that is not synchronized when it is sent or received. In this type of transmission, signals are sent between the computers and external systems or vice versa in an asynchronous manner. This usually refers to data that is transmitted at intermittent intervals rather than in a steady stream, which means that the first parts of the complete file might not always be the first to be sent and arrive at the destination. Different parts of the complete data are sent in different intervals, sometimes simultaneously, but follow different paths toward the destination. The transfer of asynchronous data doesn’t require the coordination or timing of bits between the two endpoints [4].The transmission of asynchronous data is not prompted by a clock signal when sending the data to the receiver, unlike in synchronous methods, where sending data is measured against a time reference. Compared to synchronous transmission, asynchronous communication has a few advantages:

- ❖ It is more flexible and devices can exchange information at their own pace. Individual data characters can complete themselves so that even if one packet is corrupted, its predecessors and successors will not be affected.
- ❖ It does not require complex processes by the receiving device. This means that an inconsistency in the transmission of data does not result in a big crisis, since the device can keep up with the data stream. This also makes asynchronous transfers suitable for applications where character data is generated in an irregular manner.

## 2 LITERATURE SURVEY

Ricardo Mendes and Joao Vilela [5] proposed privacy-preserving techniques in data mining: methods, metrics, and applications. According to this paper there are three methods proposed to allow the extraction of knowledge from data while preserving the privacy of individuals. This paper surveys the most relevant Privacy Preserving Data Mining (PPDM) techniques from the literature and the metrics used to evaluate such techniques and presents typical applications of PPDM methods in relevant fields.
They presented a set of algorithms to handle the Association Rule hiding, Downgrading Classifier effectiveness and Query auditing and interference control. In Association Rule hiding the problem is NP-hard, requiring heuristic solutions also tend to non sensitive rules. In Query auditing and interference control has denying or blocking certain queries can also reveal information. So a sophisticated algorithm which can hide a set of rules with lesser CPU time is to be established.

M. E. Gursoy et al [6] proposed, privacy-preserving learning analytics: challenges and techniques. This paper aims to employ and evaluate such methods on learning analytics by approaching the problem from two perspectives: (1) the data is anonymized and then shared with a learning analytics expert, and (2) the learning analytics expert is given a privacy-preserving techniques interface that governs her access to the data. According to this paper Privacy protection mechanism for the learning analytics method does not destroy the utility (data publishing and mining methods). However, Promiscuity and carelessness in sharing personal information is a risk that cannot be addressed by a privacy mechanism enforced by an institution.

M. V. Ahluwalia et al [7] has proposed target-based, privacy preserving techniques, and incremental association rule mining. The main idea of this paper is to consider a special case in association rule mining where mining is conducted by a third party over data located at a central location that is updated from several source locations. The data at the central location is at rest while that flowing in through source locations is in motion. They impose some limitations on the source locations, so that the central target location tracks and privatizes changes and a third party mines the data incrementally.

The main advantage of this gradient based approach is that the frame work developed and able to mine the data in dynamic environment using quantitative association rule for secure big data mining.

J. Hua et al [8] has proposed an innovative solution for privacy-preserving techniques utility verification of the data published by non-interactive differentially private mechanisms. In this paper, they first proposed a privacy-preserving technique utility verification mechanism based upon cryptographic method for Diff Part-a differentially private scheme designed for set-valued data. This proposal can measure the data utility based upon the encrypted frequencies of the aggregated raw data instead of the plain values, which thus prevents privacy breach. Moreover, it is enabled to privately check the correctness of the encrypted frequencies provided by the publisher, which helps detect dishonest publishers. Also they extend this mechanism to Diff Gen-another differentially private publishing scheme designed for relational data. The demerit of this proposed solution always requires auxiliary dataset in cipher text which is impossible in a dynamic environment.

S. Liu et al [9] has proposed a practical schema for privacy-preserved data sharing over distributed data streams. According to this paper distributed data sharing with privacy-preserving techniques requirements given a data demander requesting data from multiple distributed data providers. The main objective is to enable the data demander to access the distributed data without knowing the privacy of any individual provider. The problem is challenged by two questions: how to transmit the data safely and accurately; and how to efficiently handle data streams? They first proposed a practical method, Shadow Coding, to preserve the privacy in data transmission and ensure the recovery in data collection, which achieves privacy preserving computation in a data-recoverable, efficient, and scalable way. They also provide practical techniques to make Shadow Coding efficient and safe in data streams.

The main merit of this proposed method is better for distributed data sharing along with privacy preservation with ensuring of recovery.

The main demerit of this proposed solution required to study attack model and address the distributed data sharing problem in a synchronous distributed environment is to be carried out.

## 3 PROBLEM IDENTIFICATION AND PROPOSED SOLUTION

The process of preserving privacy techniques between the different users is provided. When different users are implicated in data mining all users require sending their data to trusted common centre in order to perform mining. The centralized collaborative data mining (CDM) setting requires adequate software support. In the privacy concern all users need to send their data to trusted general centre to perform the mining. It is very difficult for a user to trust the other users then this situation is called Privacy Preserving Collaborative Data Mining (PPCDM). Collaborative data mining brings light to several problems, i.e., related to model evaluation.

The number of human and knowledge based solutions to these problems are time consuming or domain dependent. Subsequently, in situations with concerns related to privacy it is highly insignificant for a user to trust the other users and in such a situation, the process is referred as Privacy Preserving Data Mining (PPDM) technique. Multiple parties have a private data set and that desire to collaboratively conduct association rule mining without release their private data to each other or any other parties.

Dileep Kumar Singh [15] proposed to realize that many of the techniques that were developed for the past two decades or soon the privacy problem can now be used to handle privacy. One of the challenges to securing databases is the privacy problem. Privacy is the process of users posing queries and deducing unauthorized information from the legitimate responses that they receive. This problem has been discussed quite a lot over the past two decades. However, data mining makes this problem worse. Users now have sophisticated tools that they can use to get data and deduce patterns that could be sensitive. Without these data mining tools, users would have to be fairly sophisticated in their reasoning to be able to deduce information from posing queries to the databases. That is, data mining tools make the privacy problem quite dangerous. While the privacy problem mainly deals with secrecy and confidentiality which are beginning to see many parallels between the privacy problems.

As with other aspects of data mining, while technological capabilities are important, there are other implementation and oversight issues that can influence the success of data distribution.

### *3.1 Problem Identification:*

- One issue is data quality, which refers to the accuracy and completeness of the data being analyzed.
- A second issue is the interoperability of the data mining software and databases being used by different agencies.
- A third issue is mission creep, or the use of data for purposes other than for which the data were originally collected.

***To overcome these problems, we propose the following solution***

The proposed method consists of participants, trusted third party, rule generation for private-sharable data, multi-sessions and instance generation. The trusted third party center is a common trust centre which assigns with the session authentication to all participants. Formation and generation of private data rule created the participants to preserve their privacy information. The instance invention along with rule generation ensures the internal malicious participant when it invokes data in unauthorized sessions.

First, proposed scheme for privacy-preservation technique for collaborative file sharing presented with CFS mechanism using Anchor method by preserving the privacy of data quality in a reliable manner.

Secondly, the data miners, the tasks are colliding with these concerns using analytic customer relationship management (CRM); it often analyzes customer data with the specific intent of understanding individual behavior of data distribution by different agencies. Associated rule mining has been used to join the data sets of each user to form one sequential file efficiently. To provide the enhancement design and implement the secured file sharing through online by assigning a secured file block ID and a participant ID using binary tree. Secured file transferring scheme among the networks in the system by way of assigning a secured file block ID to each file maintained by the user. A participant in the network has one or more files which are ready to be shared with the other users in the network. When sharing data with binary tree representation using File Security Packet (FSP), disclosure is limited by familiar mechanisms consisting of strong identities, capabilities and end-to-end encryption. Finally, mission creep or data elected and public officials should be informed of the costs and consequences to consumers, businesses, and the economy of legislative or regulatory proposals to an enhanced privacy preserving techniques for asynchronous streaming data mining in distributed environment.

## REFERENCE:

[1] Think Before You Dig: Privacy Implications of Data Mining & Aggregation Archived 2008-12-17 at the Way back, NASCIO Research Brief, and September 2004.

[2] Darwin Bond-Graham, Iron Cage book - The Logical End of Facebook's Patents, Counterpunch.org, 2013.12.03

[3] Verykios V. S., Bertino E., Fovino I. N., Provenza L. P., Saygin Y., Theodoridis Y.: State-of-the-art in privacy preserving data mining. ACM SIGMOD Record, v.33 n.1, 2004.

[4] https://www.techopedia.com/definition/26893/asynchronous-data

[5] R. Mendes and J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications," in *IEEE Access*, vol. 5, pp.10562-10582,2017. DOI:10.1109/ACCESS.2017.2706947

[6] M. E. Gursoy, A. Inan, M. E. Nergiz and Y. Saygin, "Privacy-Preserving Learning Analytics: Challenges and Techniques," in *IEEE Transactions on Learning Technologies*, vol. 10, no. 1, pp. 68-81, 1 Jan.-March 2017.

[7] M. V. Ahluwalia, A. Gangopadhyay, Z. Chen and Y. Yesha, "Target-Based, Privacy Preserving, and Incremental Association Rule Mining," in *IEEE Transactions on Services Computing*, vol. 10, no. 4, pp. 633-645, 1 July-Aug. 2017.

[8] J. Hua, A. Tang, Y. Fang, Z. Shen and S. Zhong, "Privacy-Preserving Utility Verification of the Data Published by Non-Interactive Differentially Private Mechanisms," in *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2298-2311, Oct. 2016.

[9] S. Liu, Q. Qu, L. Chen and L. M. Ni, "SMC: A Practical Schema for Privacy-Preserved Data Sharing over Distributed Data Streams," in *IEEE Transactions on Big Data*, vol. 1, no. 2, pp. 68-81, 1 June 2015.

[10] H. Rong, H. Wang, J. Liu and M. Xian, "Privacy-Preserving k-Nearest Neighbor Computation in Multiple Cloud Environments," in *IEEE Access*, vol. 4, pp. 9589-9603, 2016.

[11] P. S. Wang, F. Lai, H. Hsiao and J. Wu, "Insider Collusion Attack on Privacy-Preserving Kernel-Based Data Mining Systems," in *IEEE Access*, vol. 4, pp. 2244-2255, 2016.

[12] J. Wang, C. Deng and X. Li, "Two Privacy-Preserving Approaches for Publishing Transactional Data Streams," in *IEEE Access*, vol. 6, pp. 23648-23658, 2018.

[13] L. Xu, C. Jiang, Y. Chen, Y. Ren and K. J. R. Liu, "Privacy or Utility in Data Collection? A Contract Theoretic Approach," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1256-1269, Oct. 2015.

[14] K. Nissim, S. Vadhan, D. Xiao, "Redrawing the boundaries on purchasing data from privacy-sensitive individuals", *Proc. 5th Conf. Innovat. Theoret. Comput. Sci.*, pp. 411-422, 2014.

[15] Dileep Kumar Singh, Vishnu Swaroop, “Data Security and Privacy in Data Mining: Research Issues & Preparation”, in International Journal of Computer Trends and Technology- volume 4 Issue 2- 2013