International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

A Study of Privacy Preserving Data Mining and Techniques

Varsha Patel[1], Namrata Tapaswi[2]

[1]ME in Computer Science and Engineering, Institute of Engineering and Science, IPS Academy, Indore, India
[2]Head of Department in Computer Science and Engineering, Institute of Engineering and Science, IPS Academy, Indore, India

---

Abstract - *Privacy preserving data mining is a technique of analysing bulk data for a shared data storage system. In this technique the main aim is to preserve the data sensitivity and data privacy with accurate analysis of combined data. In this paper the privacy preserving data mining is the key domain of investigation and a new system proposal. In this context a number of different recently developed PPDM based data mining techniques. In addition of that for designing an accurate and effective PPDM model a new data mining model is also suggested in this work. Finally the paper contains the future extension of the proposed work.*

Key Words: *(PPDM, centralized data mining, privacy and security, multi party data mining, decision mining*.

1. INTRODUCTION

The main aim of the data mining and its techniques is to analyses the data and obtain the application oriented patterns from data. Therefore in a number of applications it is accepted. In this context the main component of pattern recovery and data mining is "Data". The data mining algorithms are applied on the data and recover the patterns which can be used for decision making, prediction and other necessary task. But sometimes the complete information is not available on single place or single party. Therefore decision making capability is affected due to availability of fewer amounts of data attributes. Therefore it is required to combine multiple parties of data for finding effective decision making ability [1].

In this context, not all parties are agreed to combine their data openly due to security and privacy of the end data owners. Therefore in order to process and mine the data more effectively and with the needs of security and privacy preservation some new kinds of systems are required. In order to fulfill the requirement of security and privacy during processing the multiparty data the PPDM (privacy preserving data mining) is the suitable technique of information processing. Therefore in this presented work the main aim is to study and explore the techniques of PPDM and its applications in real world domain. In addition of that a new privacy preserving association rule mining technique is also proposed for design and implementation [2].

This section provides the basic overview of the proposed work involved in this survey paper and in next section some key terms and definitions are included for study and system understanding.

2. BACKGROUND

This section provides the understanding of the different key terms which are used for explaining the proposed concept of privacy preserving association rule mining.

PPDM: PPDM (privacy preserving data mining) is the sub-domain of data mining. In this environment the data mining techniques and algorithms are employed on data with the key aim of privacy and security of end user data. Therefore in order to implement such technique cryptographic techniques, noise based techniques, blocking based techniques are frequently used for preserving privacy and sensitivity of end data owner's privacy and security [3].

Vertically partitioned data: when the data is distributed among multiple parties with non-similar kind of attributes and similar class labels this kind of data partitions of database is termed as vertically partitioned data [4].

Horizontal partitioned data: in this environment the data is distributed among multiple parties with the similar number of attributes and class labels but the amount of data instances for all the parties are different such kind of data arrangement is known as the horizontal partitioned data [4].

Association rule mining: association rule mining technique is a data mining technique where the relationship among the data attributes are established on the basis of their frequencies and combination of different attribute values. Such kinds of techniques are used for prediction and decision rule building. There are apriori and FP-tree algorithms are available for association rule mining [5].

Multiparty data: in some complex cases the less attribute based information can affect the final decision making process for a business domain. Therefore multiple data owners are agreed to combine and process their data in a common place for mining the common decisions from entire collected data. This kind of data mining environment is termed as the multiparty data mining environment [6].

Sensitive data: different parties of data owners having the different set of attributes. These attributes may contain some confidential and private data of end user. The discloser of such kind of data can impact of someone's social or financial privacy is known as the sensitive part of data for example credit card information, date of birth, PAN card number and other private information [7].

Decision mining: the data or information is processed with the aim to generate decisions by evaluation of available attributes can be defined as the decision mining or rule mining. In this context the decision trees and association rule mining techniques are played essential role [8].

3. LITERATURE SURVEY

This section includes the different recently placed contributions and efforts for improving the existing technique of PPDM (privacy preserving data mining).

Association rule mining and frequent itemset mining is two popular and widely studied data analysis techniques for a range of applications. In this paper, ***Lichun Li et al [9]*** focus on privacy preserving mining on vertically partitioned databases. In such a scenario, data owners wish to learn the association rules or frequent itemsets from a collective dataset, and disclose as little information about their (sensitive) raw data as possible to other data owners and third parties. To ensure data privacy, authors design an efficient homomorphic encryption scheme and a secure comparison scheme. Then propose a cloud-aided frequent itemset mining solution, which is used to build an association rule mining solution. The solutions are designed for outsourced databases that allow multiple data owners to efficiently share their data securely without compromising on data privacy. The solutions leak less information about the raw data than most existing solutions. In comparison to the only known solution achieving a similar privacy level as our proposed solutions, the performance of our proposed solutions is 3 to 5 orders of magnitude higher. Based on experiment findings using different parameters and datasets, we demonstrate that the run time in each of our solutions is only one order higher than that in the best non-privacy-preserving data mining algorithms. Since both data and computing work are outsourced to the cloud servers, the resource consumption at the data owner end is very low.

The collection and analysis of data are continuously growing due to the pervasiveness of computing devices. The analysis of such information is fostering businesses and contributing beneficially to the society in many different fields. However, this storage and flow of possibly sensitive data poses serious privacy concerns. Methods that allow the knowledge extraction from data, while preserving privacy, are known as privacy-preserving data mining (PPDM) techniques. ***Ricardo Mendes et al [10]*** surveys the most relevant PPDM techniques from the literature and the metrics used to evaluate such techniques and presents typical applications of PPDM methods in relevant fields. Furthermore, the current challenges and open issues in PPDM are discussed.

Paillier cryptosystem is extensively utilized as a homomorphic encryption scheme to ensure privacy requirements in many privacy-preserving data mining schemes. However, overall performance of the applications employing Paillier cryptosystem intrinsically degrades because of modular multiplications and exponentiation operations performed by the cryptosystem. In this study, ***Ismail San et al [11]*** investigate how to tackle with such performance degradation because of Paillier cryptosystem. They first exploit parallelism among the operations in the cryptosystem and interleaving among independent operations. Then, authors develop hardware realization of our scheme using field-programmable gate arrays. As a case study, they evaluate crypto processor for a well-known privacy-preserving set intersection protocol. Authors demonstrate how the proposed crypto processor responds promising performance for hard real-time privacy-preserving data mining applications.

Clinical Decision Support System (CDSS), with various data mining techniques being applied to assist physicians in diagnosing patient disease with similar symptoms, has received a great attention recently. The advantages of clinical decision support system include not only improving diagnosis accuracy but also reducing diagnosis time. In this paper, ***Ms. Meenal V. Deshmukh et al [12]*** have given the CDSS with some advance technologies like Support Vector Machine (SVM) classifier has offered many advantages over the traditional healthcare systems and opens a new way for clinicians to predict patient's diseases. As healthcare is the field in which Security of data related to patient diseases are needs to be more secure, for that in this paper, author has use RSA and Homomorphic encryption technique that properly meets the security Goals. Specifically, with large amounts of clinical data generated every day, support vector machine (SVM) classification can be utilized to excavate valuable information to improve clinical decision support system. Although clinical decision support system is quite promising, the flourish of the system still faces many challenges including information security and privacy concerns. In this they will use Homomorphic encryption technique to preserve the patient's privacy on the cloud. The patient's data can be compromised over the cloud. To overcome this scenario homomorphic encryption technique helps. The processing is done on the encrypted data; hence there is no chance of compromising privacy of patient's data.

With the popularity of smart handheld devices and the emergence of cloud computing, users and companies can save various data, which may contain private data, to the

cloud. Topics relating to data security have therefore received much attention. ***Chen-Yi Lin et al [13]*** study focuses on data stream environments and uses the concept of a sliding window to design a reversible privacy-preserving technology to process continuous data in real time, known as a continuous reversible privacy-preserving (CRP) algorithm. Data with CRP algorithm protection can be accurately recovered through a data recovery process. In addition, by using an embedded watermark, the integrity of the data can be verified. The results from the experiments show that, compared to existing algorithms, CRP is better at preserving knowledge and is more effective in terms of reducing information loss and privacy disclosure risk. In addition, it takes far less time for CRP to process continuous data than existing algorithms. As a result, CRP is confirmed as suitable for data stream environments and fulfils the requirements of being lightweight and energy-efficient for smart handheld devices.

4. PROPOSED WORK

The proposed work is aimed to design and develop a privacy preserving data model for association rule mining. In this context a data mining model is proposed. The different components of the proposed PPDM based association rule mining model is demonstrated in figure 1.

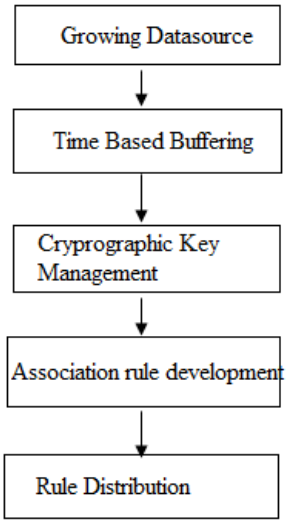

Fig.- 1:-Proposed solution methodology

In the given solution first a growing transactional data source is required which collect the data or data transactions from the different possible parties. In order to synchronize the data source and the appearance of the data in the target data source the time based buffer scheme is developed. That collects the transaction samples from time to time. In the third phase the key management or token management technique is applied to keep in track the security and privacy of data aggregated into a single database. The key management of the participating parties are managed in such way by which the re-collaboration of parties are not affected and continuous secure data or rule outcomes can be available. Finally a central authority processes the data using the association rule mining algorithm and distributes the generated rules securely.

5. CONCLUSION

The privacy preserving data mining is relatively new technique then classical data mining and it's techniques. However as the traditional data mining technique the PPDM techniques utilizing the similar algorithms and mining techniques but both the approaches are much similar in their application and data processing abilities. In this presented paper the different techniques and applications of privacy preserving data mining techniques are investigated. Additionally a new PPDM based association rule mining technique is also described for design and implementation. The proposed technique is based on the existing data mining techniques and their privacy preserving scenarios. Therefore it is a promising and effective technique that combine goodness o different PPDM model. This section provides conclusion of the presented work in this paper and in next section the future extension of the proposed work is presented.

6. FUTURE WORK

This paper provides a study about the privacy preserving data mining technique. That can be used in a number of applications where the different parties are combining their data for finding common decisions. Therefore a new privacy preserving association rule mining technique is proposed in this paper. In near future the proposed technique is implemented using the suitable technology and their performance is provided with comparative study with the similar technique.

REFERENCES

[1] Tao Xie and Suresh Thummalapenta, David Lo, Chao Liu, "DATA MINING FOR SOFTWARE ENGINEERING", Published by the IEEE Computer Society AUGUST 2009.

[2] Vinoth kumar J, Santhi V, "A Brief Survey on Privacy Preserving Techniques in Data Mining", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 18, Issue 4, Ver. V (Jul.-Aug. 2016), PP 47-51.

[3] Rajesh N, Sujatha K., A. Arul Lawrence, "Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 133 – No.7, January 2016.

[4] Saurabh Karsoliya, "Privacy Preserving Classification of heterogeneous Partition Data through ID3 Technique", International Journal of Emerging Trends

& Technology in Computer Science (IJETTCS), Volume 1, Issue 4, November – December 2012.

[5] Suzan Wedyan, "Review and Comparison of Associative Classification Data Mining Approaches", World Academy of Science, Engineering and Technology International Journal of Industrial and Manufacturing Engineering Vol:8, No:1, 2014.

[6] S. B. Javheri, U. V. Kulkarni, "A Survey on Privacy Preserving Machine Learning Techniques for Distributed Data Mining", International Journal of Computer Sciences and Engineering, Vol.-6, Issue-6, June 2018.

[7] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, Sara Kiesler, "My Data Just Goes Everywhere: User Mental Models of the Internet and Implications for Privacy and Security", Symposium on Usable Privacy and Security (SOUPS) 2015, July 22-24, 2015, Ottawa, Canada.

[8] Sjors Otten, Marco Spruit and Remko Helms, "Towards decision analytics in product portfolio management", Decision Analytics (2015) 2:4, DOI 10.1186/s40165-015-0013-7.

[9] Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao, "Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases", Transactions on Information Forensics and Security, 1556-6013 (c) 2016 IEEE.

[10] Ricardo Mendes And Joao P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications", 2169-3536, 2017 IEEE, VOLUME 5, 2017.

[11] Ismail San, Nuray At, Ibrahim Yakut and Huseyin Polat, "Efficient paillier cryptoprocessor for privacy-preserving data mining", Security Comm. Networks 2016; 9:1535–1546.

[12] Ms. Meenal V. Deshmukh, Prof. Pritish A. Tijare, Prof. Swapnil N. Sawalkar, "A Survey on Privacy Preserving Data Mining Techniques for Clinical Decision Support System", INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET), VOLUME: 03 ISSUE: 05 | MAY-2016.

[13] Chen-Yi Lin, Yuan-Hung Kao, Wei-Bin Lee and Rong-Chang Chen, "An efficient reversible privacy-preserving data mining technology over data streams", SpringerPlus (2016) 5:1407, DOI 10.1186/s40064-016-3095-3.