An Experimental Study of Privacy Preserving Data Mining

Reetesh Kumar Rai[1], Arjun Singh Parihar[2]
[2] Professor
[1, 2] Shushila Devi Bansal College of Technology, Indore, M.P. – INDIA

Abstract- *Data mining is providing ease in analyzing data and making effective decisions using a set of process and large volume of information. Therefore in a number of business applications the data mining techniques are accepted now in these days. But in some of the scenarios the data has outsourced and combined with other data owners for finding long term impact of data. In this context the data security and privacy handling is essential in different applications thus the privacy preserving techniques with data mining is essential. The privacy preserving data mining (PPDM) is a technique by which the data owners can perform combined data analysis with data security and privacy. In this presented work we are investigating and implementing a PPDM model for association rule mining. The aim of this implementation is to demonstrate the essential components of the PPDM models. Additionally their functional aspects are discussed. The discussion involves the components such as client component, data gathering and preparation of combined dataset, processing of data and data discloser. The implementation of the simulation model is described using the JAVA technology and their performance is given in terms of time and space consumption. According to the findings we can say the increase in number of parties can increase the complexity of PPDM computation.*

Keywords- privacy preserving data mining, data analysis, association rule mining, cryptography, private and sensitive information, information security.

I. INTRODUCTION

The data mining is a classical research domain for classification, prediction, categorization and association rule building. The data mining algorithms are utilizing the bulk amount of data and identify the hidden patterns on the datasets. A number of applications are utilizing the concept of these techniques but some of them include the end client's personal and confidential data. Therefore discloser of such information can be damage the end client reputation and financial losses. In this context we need a secure and effective manner by which we can mine the data privately and reduce the data discloser risk. This domain of data mining is known as privacy preserving data mining (PPDM).

The PPDM is aimed at involving the multiple data owners who are agreed to mine their data in a common place without disclosing the data values and attributes for preserving the privacy of data owners and their clients. In this presented work are proposing a PPDM model for association rule mining. The methods and network based application has been designed to gather the data from different client. After data gathering of the data the system is preparing a new dataset in a server application. This combined dataset is will be used with the association rule mining algorithm for computing the association rules using the dataset. Additionally the rule distribution methodology is also presented by which the data owners can recover the obtained rules based on their contributed part of data privately.

II. PROPOSED WORK

The need of privacy preserving data mining is increasing day by day due to digital revolution and increasing potential of automated data analysis and decision making systems. In this context, the proposed work is provides a study of PPDM system design and their essential components. This chapter provides the details of such components which are utilized in different phases of PPDM based data analysis systems.

A. System Overview

In recent years there are number of applications are developed which are gathering our personal data. This data may be utilized for different business prospective such as a survey, our habits, internet usage patterns and many more. These applications are sometimes combining their data to others for utilizing the power of machine learning to discover the trends for their business growth and future work plan preparations. The data discloser of these applications needs to be secured because the sensitive or confidential information of the data owner may harm someone's personal and professional life. Therefore we need a privacy preserving technique to sanitize the sensitive information before disclosing it to any other trusted or un-trusted parties.

Therefore in this presented work we are studying and implementing the technique of PPDM as a study of required components and their basic modeling. The implemented techniques involve a network program for collecting data from the different agreed parties for contributing the data for mining. The second component involves the data sensitization process for hiding the attribute values. Thus the cryptographic methods are involved at the client end which encrypt and decrypt the data using a private key cryptographic technique. Next component includes a data mining algorithm which works on cryptographic data and provide the consequences of the combined data mining. Finally, the data decryption methodology is developed which demonstrating how the data is recovered privately for utilizing with the business domain. This section provides the overview of the proposed system and the next section discussing the details of the methodology used for simulation of PPDM model.

B. Methodology

The figure 2.1 demonstrates the different functional components of the studied model. The details and their working in PPDM modeling is discussed in this section.

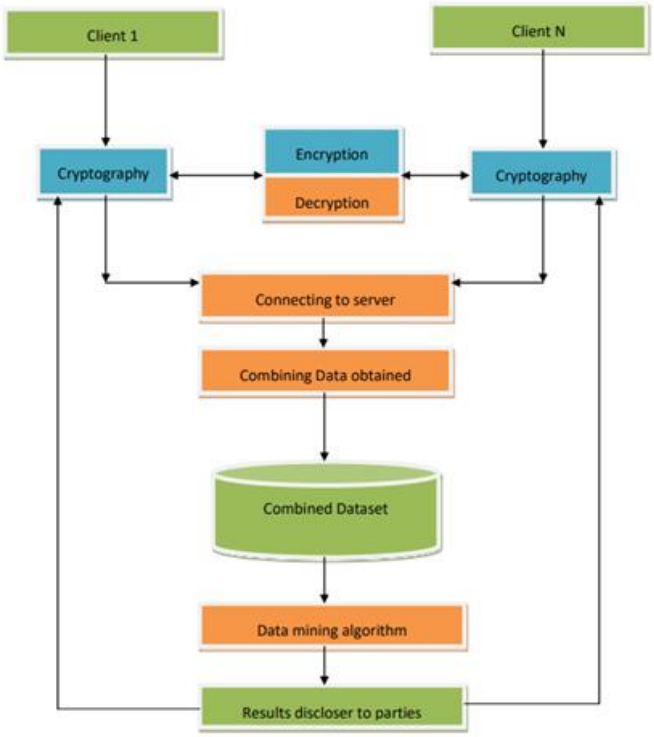

Figure 2.1 Proposed PPDM model for study

Clients: in this experiment the client is described as the data owners, data sources or any business institute, who is agree to combine their data on a common trusted place for mining the data. There are number of parties can be possible for contributing the data parts. But most of the clients are not believe on any one thus they want to send data privately. Therefore these clients are modifying their data by using various data sanitization approaches such as randomization, cryptography, and others. In this presented work the cryptographic approach is proposed for implementation. The implementation of cryptographic technique is easy and low cost. Additionally replaced with newer one easily, therefore it is one the popular approach of sanitization of data attributes.

Cryptography: here the cryptography is used for sanitizing the data in this context here we implemented AES algorithm. The AES algorithm is private key algorithm which accepts the personal key and data attributes for sanitization. The AES algorithm is chosen due to its efficiency and their ability to accept variable size key. Therefore different parties can add their own key according to their own privacy level requirement. In this experimental study we have just used a random string for generating the 128 bit key.

Encryption and Decryption: the AES algorithm usages the same key for encryption and decryption. Therefore the only who knows the secrete key can recover the encrypted data. Therefore all the data owners can use any hard key for providing security. In this experiment when the data goes outside then the data is encrypted and when it comes in then the data is decrypted using the same secure key. Therefore, when the data owner disclosing the data part then they can sensitize the data and when after processing the consequences of mining is received the data owner can recover the results only based on their own part of data.

Network connectivity: the entire process can be connected through personal or public network infrastructure to submit their data to the server. This server can be trusted, semi-trusted or un-trusted in nature. In this experimental study we proposed to use the private network for processing and submitting the data to the server. Therefore a network application is design which is running as two parts client and server. The client applications are available on the client's machine by which the client can encrypt or decrypt the data. Additionally by using the dedicated connectivity between both the parties and server the data is securely communicated to server.

Data combining approach: how the data is utilized and combined in the server to produce new dataset is a significant issue in the server end. Normally there are two different techniques are available for combining the data i.e. horizontal partitioned and vertical partitioned. If there are data sources has the common set of attributes then the model usages the horizontal partitioning method to combine the different sources of data. On the other hand if the data contributed has

the different set of attributes then the data combined in vertical manner.

Combined dataset: in this presented work we have used the vertical partitioned technique for combining the different set of data. Therefore when a party sends their data then we aggregate the data according to their columns and keep the same class label for identifying the indexes of the data. After combining the data we are prepared a CSV file on which the attributes of combined data has written. This CSV file will be used in further process of data mining.

Data mining algorithm: the different types of data mining algorithms are available for processing the data. But in our study we found that most of PPDM models are utilizing the rule based data mining techniques. the advantage of the utilizing the rule based model is that, these models are producing rules which may be combination of different attributes and a rage of values by which the purpose of mining is completed as well as the data is not required to be disclose more precisely for maintaining the utility of the data. Therefore in this experiment we are utilizing the apriori algorithm for generating the association rules.

Results discloser: normally there are two types of techniques available by which the consequences of the data mining are disclosed i.e. privately and publically. That depends on the purpose of data is going to be use. For example when the consequences are going to be used for research purpose then it is disclosed on public platforms for and challenge or academic use. On the other hand when the data is privately disclosed the business intelligence utilizes these consequences for developing future plans or other proactive use. In this presented work the provision is made to disclose the data privately between participants.

III. RESULT ANALYSIS

The proposed work is motivated to study and implement an efficient and lightweight PPDM model development. Therefore the implemented model is evaluated in this chapter for finding the efficiency of the model in terms of time and space complexity.

A. Time Consumption

The PPDM models include the security and privacy management additionally employs the data mining algorithms for recovering the data patterns. Therefore these algorithms may have long running time and computationally expensive systems. Thus the time consumption is an essential parameter for system evaluation in PPDM scenarios. The time consumed for processing the data using a PPDM model is calculated using the following equation:

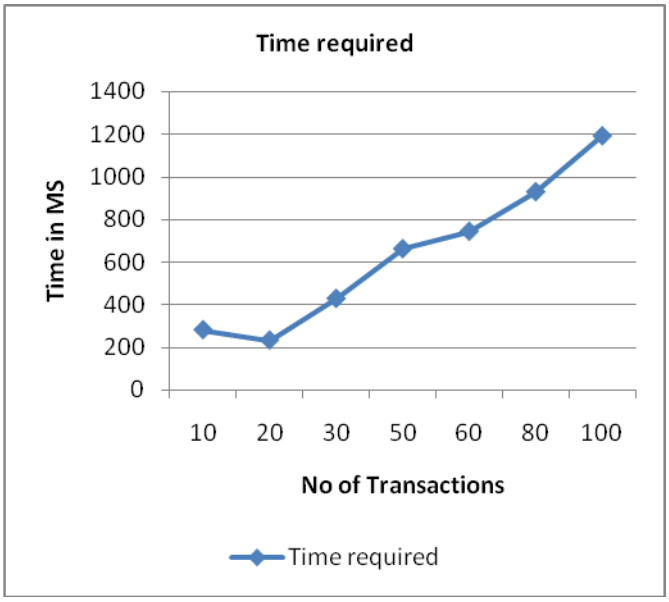

Figure 3.1 time consumption

$$time\ used = end\ time - start\ time$$

The total time required for processing of the proposed PPDM model for different number of transactions are demonstrated in table 3.1 and figure 3.1. The figure visualizes the time requirements of the model with increasing size of transactions. The X axis of this diagram includes the amount of transaction sets used for processing and Y axis provides the time utilized in terms of milliseconds (MS). According to the obtained results the amount of time requirement is increases as the amount of transactions sets are increases.

Table 3.1 time consumption

Transaction size	Time required
10	281
20	232
30	428
50	662
60	744
80	930
100	1193

B. Memory Usage

The memory usage is also known as the space complexity of an algorithm. The memory usages is measured using the process based method. In other terms for executing a process in java the amount of main memory utilized is termed

as memory usage of the process. This will be computed using the following formula:

$$memory\ used = total\ assigned - free\ space$$

Table 3.2 memory usages

Transaction size	Memory usage
10	1033
20	1328
30	1431
50	1492
60	1588
80	1629
100	1772

The memory usage of the proposed model is demonstrated using figure 3.2 and table 3.2. In this diagram the X axis shows the amount of transactions involved for processing and the Y axis demonstrate the required size of memory in terms of kilobytes (KB). According to the obtained performance the proposed model utilizes the acceptable amount of memory and not much increases with the different process involved with the PPDM model for privacy and security management. Thus the proposed model is acceptable and efficient for processing the association rules.

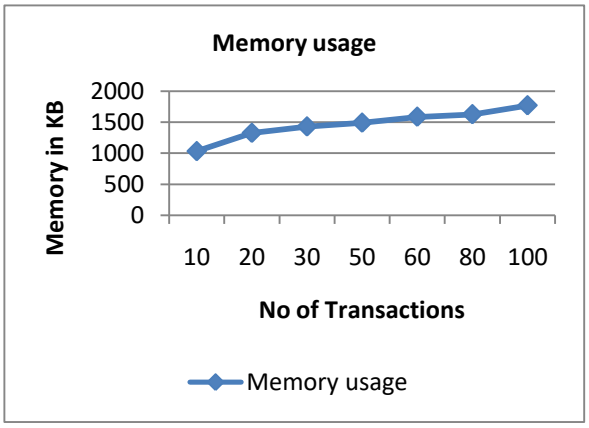

Figure 3.2 memory usage

IV. CONCLUSION & FUTURE WORK

In this chapter we are providing the summary of the proposed study and conducted experiments. In addition, the future extension of the proposed study is also described.

A. Conclusion

The data mining techniques are widely used now in these days. The data mining is set of methods and techniques which are used for data exploratory study. This study can also be used for classification, prediction, and association rules. Therefore these techniques will be used in a number of business domains such as banking, finance and business intelligence. But in some of the applications requires the sanitization of the private data or sensitive data attributes because the sensitive and private data can harm the privacy of the end client and also suffers to damage of financial and social reputation. In this context, we to employ the data privacy and security techniques when we implementing the multiparty data mining techniques. The multi-party data mining has combining the data from different data owners and mine the data processing.

Therefore, in this presented work we are investigating the different privacy preserving data mining techniques. These techniques are utilizing the security and privacy preserving methods for processing the combined data of different data owners and results the combined consequences to each party with disclosing any sensitive and private information. Further a PPDM model for association rule mining is developed to explore the required components of the PPDM system. In this model we have implemented AES cryptographic concept for encrypting the data at the client end. After data encryption the data is communicated to a common server using a network application developed using JAVA. The server application gathers data from different clients and prepares a common dataset. The combined data is processed using the data mining algorithm specifically using the apriori algorithm and produces the encrypted rules which are distributed to all the parties. The client uses their private key and decrypts their rules based on their own part of data. The described method is validated using the time and memory complexities for finding their efficiency. The summary of their performance evaluation is described in table 4.1.

Table 4.1 performance summary

S. No.	Parameters	Observations
1	Time consumption	The time is increasing with the increase of amount of transactions increase.
2	Memory usage	The memory consumption is also increasing but not as much which is acceptable

According to the obtained results we found the proposed model is efficient in terms of time and memory. This model will be extendable for future work and the real world use.

B. Future Work

In this presented work we have studied and implemented a PPDM model for association rule mining technique. That model is efficient and able to deal with the basic privacy requirement of the PPDM models. For extension of the given system the following future extension will be proposed.

1. The multiple parties are increasing the dimension of the data which increases the complexity of the data processing. In near future it is required to implement some privacy preserving dimensionality reduction algorithm
2. The PPDM models are suffers from the quality checking and validation of resultants thus we need to explore the methods for validation of privacy and data utility

REFERENCES

[1] L. Li, R. Lu, K. K. R. Choo, A. Datta, J. Shao, "Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases", 1556-6013 (c) 2016 IEEE

[2] D. Shah, H. Isah, F. Zulkernine, "Stock Market Analysis: A Review and Taxonomy of Prediction Techniques", Int. J. Financial Stud. 2019, 7, 26; doi:10.3390/ijfs7020026

[3] K. K. Nivethithaa, Dr. S. Vijayalakshmi, "Survey on Data Mining Techniques, Process and Algorithms", Journal of Physics: Conference Series 1947 (2021) 012052, IOP Publishing, doi:10.1088/1742-6596/1947/1/012052

[4] J. M. David, K. Balakrishnan, "Significance of Classification Techniques in Prediction of Learning Disabilities", https://arxiv.org/ftp/arxiv/papers/1011/1011.0628.pdf

[5] F. Alam, S. Pachauri, "Comparative Study of J48, Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA", Advances in Computational Sciences and Technology, ISSN 0973-6107 Volume 10, Number 6 (2017) pp. 1731-1743, © Research India Publications

[6] V. Sathya, and Dr. V. Gayathiri, "Encryption-Based Techniques For Privacy Preserving Data Mining", International Journal of Scientific & Engineering Research, Volume 8, Issue 4, April-2017

[7] V. Baby, N. Subhash Chandra, "Privacy-Preserving Distributed Data Mining Techniques: A Survey", International Journal of Computer Applications (0975 – 8887), Volume 143 – No.10, June 2016

[8] S. Dhanalakshmi, S. Mekala, T. Swathi, "A Review on Privacy Preservation in Distributed Data Mining", International Journal for Scientific Research & Development| Vol. 6, Issue 02, 2018 | ISSN (online): 2321-0613

[9] Chirag Mewada and Rustom Morena, "Trie Based Improved Apriori Algorithm to Generate Association Rules", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 10, October 2016

[10] Giannotti, Fosca, et al. "Privacy-preserving mining of association rules from outsourced transaction databases." IEEE Systems Journal 7.3 (2013): pp. 385-395.

[11] Richhariya Vineet, and Prateek Chourey, "A Robust Technique for Privacy Preservation of Outsourced Transaction Database", ISSN (E) (2014): 2321-8843.

[12] V. Richhariya, and P. Chourey, "A Robust Technique for Privacy Preservation of Outsourced Transaction Database", ISSN (E) (2014): 2321-8843.

[13] R. Lu, H. Zhu, X. Liu, J. K. Liu, J. Shao, "Toward Efficient and Privacy-Preserving Computing in Big Data Era", 0890-8044/14/$25.00 © 2014 IEEE, IEEE Network • July/August 2014

[14] L. Li, R. Lu, K. K. R. Choo, A. Datta, J. Shao, "Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases", DOI 10.1109/TIFS.2016.2561241, IEEE, Transactions on Information Forensics and Security

[15] X. Yi, F. Y. Rao, E. Bertino, A. Bouguettaya, "Privacy-Preserving Association Rule Mining in Cloud Computing", ASIA CCS'15, April 14–17, 2015, Singapore Copyright c 2015 ACM 978-1-4503-3245-3/15/04

[16] I. San, N. At, I. Yakut, H. Polat, "Efficient paillier cryptoprocessor for privacy-preserving data mining", SECURITY AND COMMUNICATION NETWORKS Security Comm. Networks 2016; 9:1535–1546

[17] J. Li, Z. Liu, X. Chen, F. Xhafa, X. Tan, D. S. Wong, "L-EncDB: A lightweight framework for privacy-preserving data queries in cloud computing", Knowledge-Based Systems xxx (2014) xxx–xxx, 2014 Elsevier B.V. All rights reserved.

[18] S. Dubey, A. Sen, "Data Mining Based on Association Rule Privacy Preserving", Binary Journal of Data Mining & Networking 5 (2015) 16-21