Privacy Preserving Data Mining and Data Exposure Technique and Performance Study

Ankita Sharma
M.Tech (IA & SE)
School of Computer Science and IT
Devi Ahilya Vishwavidyalaya, Indore, India

Pankaj Jagtap
Lecturer
School of Computer Science and IT
Devi Ahilya Vishwavidyalaya, Indore, India

ABSTRACT
The data mining is used in various applications in order to provide decisions, pattern matching and others. In order to mine data, it is required to initially learn from the data and then extract the patterns from data which is supplied in raw formats. Sometimes data is available in parts and need to conclude the outcomes from the data in secure manner. Therefore a technique is required that securely combine data from all the participating partitions of the data. Mine the combined data for extracting the meaningful pattern in a secured manner. Finally disclose the data in such manner which is utilized by the different other application, without disturbing the outcomes of the final application decisions. In order to design and demonstrate the three parties of data are considered for combining them, and mine them. And finally using the C4.5 classification algorithm is applied to find the final utilization on different application. To find that the mining decision is not varied from the initial values of mining the comparison is made between the disclosed data and initial dataset on the basis of accuracy of classification. The experimental results demonstrate a very fewer patterns are misclassified after in comparison of the initial dataset classification in terms of accuracy.

Keywords
Data Mining, Privacy Preserving Data Mining, Data Exposure, Classification, Vertical Partitioning Of Data

1. INTRODUCTION
Data mining and their techniques enable us to analyze the huge amount of data. The analysis is made in order to find the fruitful patterns for making the decisions in various applications such as prediction, classification, pattern recognition and others. Now in these days the computational domains are shifted towards the distributed computing. Therefore the data is partitioned either vertically or horizontally. The main motive of such data organization is to enhance the capability of storage and efficient data access. In this work the vertically partitioned data is considered for utilization during experimentation.

The vertical partitioned data is a kind of fragmentation of data which is performed in vertical manner. In order to understand this organization let a data set D contains four attributes such that $D = \{A_1, A_2, A_3, A_4\}$ and a class label C. the class label is associated with each instance of data for demonstrating the classes identified. When the data is partitioned vertically the set of attributes can divided in multiple parts, in this example we consider two parts of vertically partitioned data. Thus two different datasets are constructed using set of attributes given. Says $D_1 = \{A_1, A_2, C\}$ and $D_2 = \{A_3, A_4, C\}$.

The vertically partitioned data can be used in various application scenarios. For example a company's two department's A and B managing the records for a same clients but the utilization of the client's records are different in these departments. In addition of that for purpose for making decision for that client by using both the departmental data the data is aggregated and mined in a same place. In this example the departmental data is vertically partitioned and for privacy management the data can be mapped or encrypted using different formats. Let us consider a second example there are various different banks which have the account of client and they not wise to expose the client's privacy. But a central authority wants to make some hard decisions on the basis of the client's transactions. Then it is required to preserve the privacy in multiple phases. First when the data is communicated to central authority, second when the data is combined and mined. Finally, making decisions and disclosing the decisional values.

In the given two examples we are considering the second example where the end to end privacy during the data mining scope is required. This section provides the overview of the proposed work. The second section provides the review of available techniques and the next section involve a proposal of the securing the data mining in distributed computing environment. Finally the results and the conclusion of the work is presented.

2. LITERATURE SURVEY
This section provides the study about the recent research work performed on the targeted privacy preserving data mining.

Privacy preservation in data mining is to hold back the sensitive data from attackers. There are various existing methods available to preserve the data like perturbation, anonymization, randomization etc. each method has its own advantages and disadvantages. The trade-off between security and utility of data should be handled with standardizing methods for the PPDM. ***R. Hariharan et al [1]*** explained a method based on PPDM in data mining using cluster based greedy method. Application/Improvements: This method can be applied in sensitive data areas such as hospitals, Customer Management System, government survey, etc., where there is need for privacy preservation.

Recent advances in sensing and storing technologies have led to big data age where huge amounts of data are distributed across sites to be stored and analyzed. Indeed, cluster analysis is one of the data mining tasks that aim to discover patterns and knowledge through different algorithmic techniques such as k-means. Nevertheless, running k-means over distributed big data stores has given rise to serious privacy issues. Accordingly, many proposed works attempted to tackle this concern using cryptographic protocols. However, these cryptographic solutions introduced performance degradation issues in analysis tasks which do not meet big data properties. ***Zakaria Gheid et al [2]*** propose a novel privacy-preserving k-means algorithm based on a simple yet secure and efficient multiparty additive scheme that is cryptography-free. We designed our solution for horizontally partitioned data.

Moreover, we demonstrate that our scheme resists against adversaries passive model.

The privacy preserving data mining is playing crucial role act as rising technology to perform various data mining operations on private data and to pass on data in a secured way to protect sensitive data. Many types of technique such as randomization, secured sum algorithms and k-anonymity have been suggested in order to execute privacy preserving data mining. ***Rajesh N et al [3]***, on current researches made on privacy preserving data mining technique with fuzzy logic, neural network learning, secured sum and various encryption algorithm is presented. This will enable to grasp the various challenges faced in privacy preserving data mining and also help us to find best suitable technique for various data environment.

Association rule mining and frequent item-set mining is two popular and widely studied data analysis techniques for a range of applications. ***Lichun Li et al [4]*** focus on privacy preserving mining on vertically partitioned databases. In such a scenario, data owners wish to learn the association rules or frequent item-sets 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, we design an efficient homo-morphic encryption scheme and a secure comparison scheme. Then propose a cloud-aided frequent item-set 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. 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 proposed solutions is 3 to 5 orders of magnitude higher. Based on our experiment findings using different parameters and datasets, we demonstrate that the run time in each of 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.

Privacy preservation is that the most targeted issue in information publication, because the sensitive data shouldn't be leaked. For this sake, several privacy preservation data mining algorithms are proposed. ***V. Shyamala Susan and T. Christopher [5]*** provide, feature selection using evolutionary algorithm and data masking coupled with slicing is treated as a multiple objective optimization to preserve privacy. To start with, Genetic Algorithm (GA) is carried out over the datasets to perceive the sensitive attributes and prioritize the attributes for treatment as per their determined sensitive level. In the next phase, to distort the data, noise is added to the higher level sensitive value using Hybrid Data Transformation (HDT) method. In the following phase slicing algorithm groups the correlated attributes organized and by this means reduces the dimensionality by retaining the Advanced Clustering Algorithm (ACA). With the aim of getting the optimal dimensions of buckets, tuple segregating is accomplished by Meta-heuristic Firefly Algorithm (MFA). The investigational consequences imply that the anticipated technique can reserve confidentiality and therefore the information utility is additionally high. Slicing algorithm allows the protection of association and usefulness in which effects in decreasing the information dimensionality and information loss. Performance analysis is created over OCC 7 and OCC 15 and our optimization method proves its effectiveness over two totally different datasets by showing 92.98% and 96.92% respectively.

3. PROPOSED SOLUTION

This section provides the working of the proposed methodology. Therefore first the initial overview of the system is demonstrated and then the system is formulated using the algorithm steps.

3.1 System Overview

Data mining is an application which is used to analyze data in order to find the valuable patterns from the data. To recover these patterns from the data, the data mining algorithms are applied. These algorithms are process the information by which the significant patterns can be learned. Using this learning of algorithms the different applications are functioning such as pattern recognition, prediction and others. The learning of the algorithms is either supervised or unsupervised in nature. The supervised learning techniques require the predefined patterns for learning purpose. On the other hand the unsupervised learning technique not requires the pre-defined patterns. In this presented work the supervised learning technique is used for design and investigation.

In various data mining scenarios the data is available in parts and the data owners are worried about the confidentiality of data. Therefore the security and privacy management in data mining techniques are major concern in today's applications. In this context the data leakage and other kind of privacy issues can also a significant concern in the domain of privacy preserving techniques. In this proposed work a privacy preserving data mining technique is proposed that works in three major parts. That is simple to implement and efficient during the computation. Additionally not much increases the computational overhead during the data exposer. In these algorithms first phase the data is processed using a session key which is randomly distributed using the server side. In next part the data is aggregated and in final steps the data mining algorithm is applied on data for learning or utilizing with the applications.

This section provides the overview of the proposed work and in the next section the detailed methodology of the system is described.

3.2 Methodology

The proposed work is demonstrated in a client server environment. It is assumed that the client is a kind of data owner and having the part of data. Additionally the server is a trusted authority who collects the data from a different number of clients. In order to combine the data and process the data using server first user need to make connection request. As the server accepts the connection request the server generate a session key for the particular user. The client accepts the session key and process the part of data using the following manner. In order to demonstrate the functional aspects of the system here only numerical data is considered.

Table 1 Processing Data

Input: dataset D = $\{A_1, A_2 \ldots, A_n\}$, Session Key S Output: processed data P_d
Process: 1. $for(i = 1; i \leq n; i+=)$ a. $A_{max} = FindMax(A_i)$ b. $A_{min} = FindMin(A_i)$ c. $for(j = 1; j \leq A_i.Length; j++)$ i. $NewV = \frac{Current\ Value - A_{min}}{A_{max} - A_{min}} XS$ ii. Update new value d. End for 2. End for

After processing of the data, Data is communicated to the server. Server combines the data obtained from different parties. During this process the data is again processed using the session keys according to the end user. The table 1 process is repeated but for computing the new values the session key is divided from the values. Additionally according to the class labels the entire data set is again generated. Finally the data is produced to C4.5 algorithm to process and learning with the input data. Finally the data is again converted using the session key for discloser of data values. This process is implemented using the JAVA technology and their performance is computed in various performance parameters. The next section includes the performance of the algorithm with the real data and the processed data.

4. RESULTS ANALYSIS

This section provides the discussion about the results evaluation of the proposed privacy preserving technique for secured data mining. Therefore the computed parameters are discussed in this section.

4.1 Accuracy

The accuracy of any data mining technique demonstrates the accurately classified patterns over the total amount of samples present for classification. The accuracy of system or algorithm can be evaluated using the following formula:

$$Accuracy = \frac{total\ correctly\ classified\ data}{total\ data\ to\ classify} X100$$

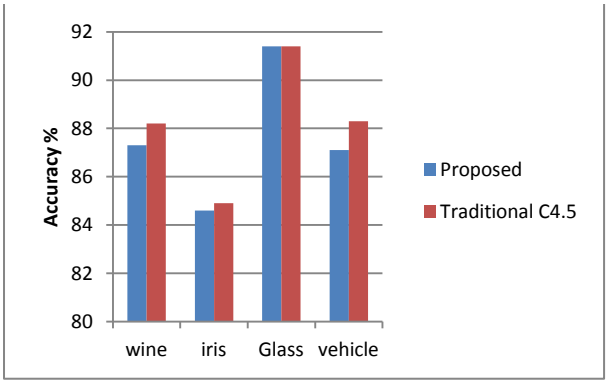

Figure 1 Accuracy

The performance of proposed data processing technique and with real data processing is given using the figure 1. In this diagram X axis contains the datasets on which experiments are performed and in Y axis the accuracy of algorithms in terms of percentage. According to the obtained results the data and their performance is not varied from actual learning patterns. Therefore the proposed technique is acceptable for privacy preserving data mining applications and their exposure.

4.2 Error Rate

The error rate is measures of the incorrectly classified data from the input data to be classify. The error rate of the data mining algorithm is computed using the following formula:

$$error\ rate = 100 - accuracy$$

Or

$$error\ rate = \frac{incorrectly\ classified\ data}{total\ data\ to\ classify} X100$$

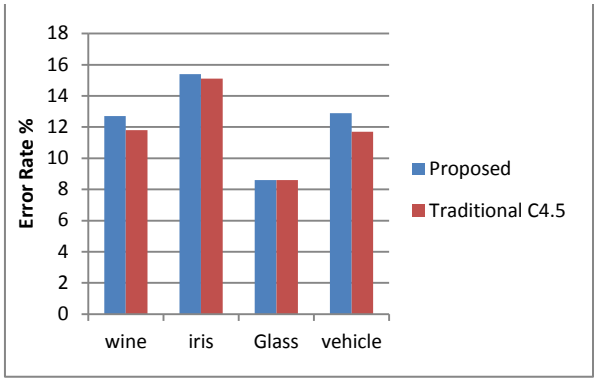

Figure 2 Error Rate

The figure 2 shows the comparative performance of the implemented algorithms. To represent the performance of algorithms the X axis contains the datasets used for experiments and the Y axis shows the error rate percentage. According to the obtained results the error is not fluctuating in both the algorithms due to change of data thus the proposed technique is acceptable for privacy preserving data mining.

4.3 Memory usage

The algorithms require the significant amount of main memory for computing the algorithm task. This amount of memory requirements are termed as the memory usage or the space complexity of algorithm. The space complexity of the algorithms in both scenarios is demonstrated using figure 3. In this diagram X axis shows the data sets used for experimentation and Y axis shows the memory usages in terms of KB (kilobytes). According to the obtained results the proposed methodology requires some additional memory as compared to the traditional C4.5 algorithm. But the memory utilization is acceptable for the improving secure outcomes.

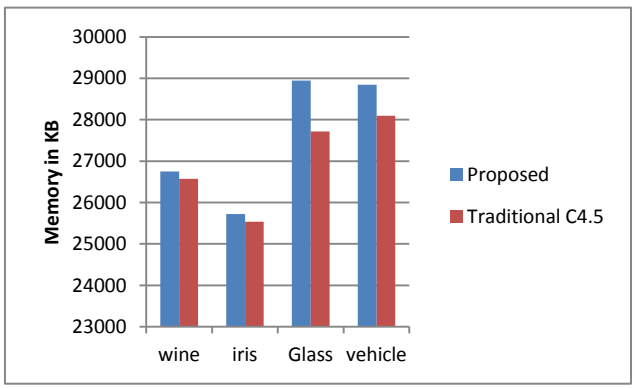

Figure 3 Memory Usage

4.4 D. Time consumption

Any algorithm needs an amount of time for computing the learning model this time is termed as the time consumption or the time complexity of algorithm. The time complexity of algorithm can be computed using the following formula:

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

The figure 4 contains the performance of both the algorithm namely traditional C4.5 data model and proposed secure technique. The X axis of diagram contains the data sets on which the experiments performed and the Y axis shows the time consumption in terms of milliseconds. According to the obtained results the time consumption of the proposed technique is little bit higher as compared to traditional technique but acceptable for security requirements.

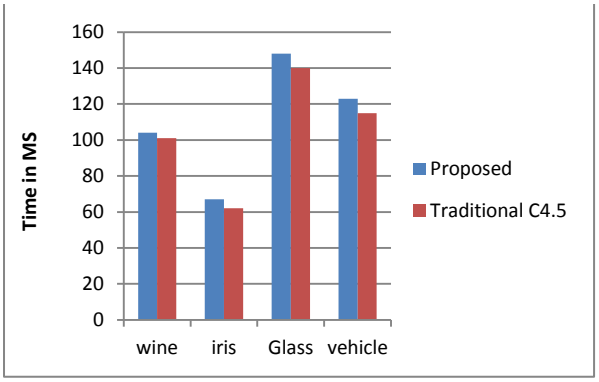

Figure 4 Time Consumption

5. CONCLUSION

This paper is a study of privacy preserving data mining technique. Therefore the key requirement of the privacy in data mining environment is described first. In further the various available contributions and research in the domain of privacy preserving data mining is studied. Finally a new method which is lightweight and efficient in nature is proposed for implementation. The proposed method of the privacy preserving data mining is implemented and their performance is computed. According to the obtained outcomes the proposed technique is efficient and provides accurate results as the real or original dataset provide. Thus the main aim of the proposed methodology is achieved successfully.

6. ACKNOWLEDGEMENT

I am extremely thankful to Mr. Sanjay Tanwani, Head of Department, SCSIT, DAVV, Indore for giving me a golden opportunity to my education.

I am thankful to Mr. Pankaj Jagtap, Lecturer, SCSIT, DAVV, Indore for his perfect guidance by giving timely suggestions throughout the tenure of my research and also for his continuous supervision and valuable guidance for improvements and completion of my research successfully.

7. REFERENCES

[1] R. Hariharan, C. Mahesh, P. Prasenna and R. Vinoth Kumar, "Enhancing Privacy Preservation in Data Mining using Cluster based Greedy Method in Hierarchical", Indian Journal of Science and Technology, Vol 9(3), January 2016

[2] Zakaria Gheid, Yacine Challal, "Efficient and Privacy-Preserving k-means clustering For Big Data Mining", IEEE TristCom, Aug 2016, Tianjin, China pp.791 - 798, 2016, TrustCom

[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] 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

[5] V. Shyamala Susan and T. Christopher, "Privacy Preserving Data Mining Using Multiple Objective Optimization", ICTACT Journal On Soft Computing, October 2016, Volumn: 07, ISSUE: 01.