An Efficient Framework for Privacy Preserving in Correlated Data based on Tree Distribution Approach

Chaitrali S. Waghchaure[1] Prof. J. R. Yemul[2]
[1]Student [2]Assistant Professor
[1,2]Department of Information Technology
[1,2]Smt. Kashibai Navale College of Engineering Pune, India.

Abstract— Privacy preserving in Data Mining is study of achieving some data mining goals without scarifying the privacy of individuals. In this paper privacy has been provided to correlated dataset in which users fires a query to get the output. In general attributes in a dataset are sampled independently. However, in real-world attributes in a dataset are rarely independent. If one attribute is depend on another then it leads to privacy violation. The solution to this problem is provide privacy using correlated tree distribution approach in which query gets scrutinized to get proper attributes in the dataset. Different steps are processed like K-Means Clustering algorithm, Correlation Computation, Tree Creation. If-Then Rules are used to set ranges where the sensitive information has been resided. Access control policies are provided in order to protect information. Different mechanisms are used, like relaxed admissible mechanism Correlated Iteration Mechanism (CIM), Privacy preservation approach via data set complementation to maintain privacy. Many existed system have a drawback in which major drawback is all have huge time complexity. In tree distribution approach the drawback has been overcome in which it strictly focuses on finding correlation amongst data. To enhance the privacy in Non-IID (Independent and identically distributed) datasets proposed system uses tree based traversing technique for maintaining a strict access control over the attributes. The main aim of the system is to provide privacy when user searches globally.

Key words: K-Means Clustering, Correlation Computation, Correlation Tree, If-Then Rules

I. INTRODUCTION

In general records in a dataset are worked independently but if as per depth of particular dataset it seems that data in the dataset are depend on each other. In this case what happens, particular user fires a query on a dataset expecting the output as per query but if data is depend on any sensitive attribute data then it leads to leakage of particular information.

It is necessary to find out correlated attribute so that privacy can be applied to particular attributes. There are different kinds of queries user can design to get data. Therefore the careful examination of query in detail is necessary. In many existing system access control policies are not provided which again leads to privacy issue.

Data mining can be used to find patterns or relationships in their data. Knowledge discovery in databases processes in which data mining is one of the important steps which extract data from large datasets, also it includes the techniques/methods like artificial intelligence, machine learning and statistics. It has different emerging field in which data can be extracted from a dataset and transform it into an understandable form, is the main purpose of data mining process.

Privacy preserving in Data Mining is study of achieving some data mining goals without scarifying the privacy of individuals. Privacy preserving in distributed data has various applications. Each application has different conditions: What is meant by privacy, what are the desired results, how is the data correlated and distributed with each other, what is the highest distribution factor in dataset, etc. Data mining can extract knowledge from large data collections, sometimes these collections are divided among various parties. Privacy may prevent the parties from directly sharing the data and some types of information about the data. This paper presents some mechanism and shows how they can be used to solve several privacy-preserving data mining problems.

Gang Li, Tianqing Zhu, Ping Xiong, [1] proposed Correlated Iteration Mechanism in which while finding correlation between attributes iteration mechanism is used which takes more time to compute correlation factor. It also considers non-correlated attributes which is no longer necessary because privacy has been given to only correlated data. It will lead to space complexity problem. Also access control policies are poor. In some system there may also exist a mismatch between the estimated prior statistics and the true prior statistics, due to a small number of observable samples[4].Chaotic system generates highly unpredictable noise, it contains a couple of differential equations and needed to be solved for each point of the required number of solution points. The solution of these equations may be time consuming [6].The proposed system aims to find strict correlation amongst attributes in the datasets in which important data can be protected.

The remaining section of this paper is organized as follows: Section II includes related work. Section III includes Proposed Methodology. Section IV includes Results and discussions and Section V includes conclusion and future scope.

II. RELATED WORK

Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao[1], in this paper Each data owner inserts fictitious transactions to his private database, and encrypts items in the database with a substitution cipher. The fictitious transactions are used to mitigate frequency analysis attacks .Once the databases have been encrypted, they are outsourced to the cloud as part of the joint database maintained by the cloud. To allow the cloud to accurately mine the database data owners tag each transaction in their outsourced databases and joint database with an encrypted realness value (ERV) using customized homomorphic encryption scheme. A realness value (RV for short) is either 0 or 1.

In this paper [2] Correlated Iteration mechanism is used to answer large number of queries, which proposed a

correlated data privacy for hiding information in NON-IID dataset. It successfully prevents the utility and enhances privacy guarantee while preserving privacy. It saves the privacy budget and decreases the noise for each query. Records in the datasets are often correlated with each other, and this may reveal extra information.

YilinShen and Hongxia Jin [4], in this paper relaxed admissible mechanism is used. With the help of this mechanism two contradictory goals they met i.e recommendation accuracy and privacy preservation. User's private data is required which put users at risk. User's perturbed data can be used in existing recommender system in order maintain privacy. Relaxed admissible mechanism is used to achieve feasible and useful perturbations. It benefits companies with high quality personalized services and strong privacy protection on perturbed data.

Anil Pratap Singh and Abhishek Mathur[7], proposes chaotic system based data perturbation technique with multilevel trust it is relatively very difficult to crack the privacy of the original data offered by the proposed method when compared with random perturbation based techniques especially when the number of published perturbed copies of same trust level increases. Chaotic system generates highly unpredictable noise; it contains a couple of differential equations and needed to be solved for each point of the required number of solution points. The solution of these equations may be time consuming.

Pui K. Fong and Jens H. Weber-Jahnke, [10] has proposed a approach called new privacy preserving approach via dataset complementation. It provides private information from the samples while keeping their utility. This approach converts the sample data sets into some unreal data sets so that any original data set is not reconstructed if a theft were to steal some of the contents. If training data sets are revealed, Privacy preservation via data set complementation will not work. To overcome this limitation a cryptographic privacy in preserving approach along with data set complementation can be used.

Haibat Jadhav, Prof. Pankaj Chandre[11] has used one of the significant and popular data mining process is association rule mining which is used to find frequent patterns in the given dataset in which the apriori algorithm are the most common for mining frequent item set. This paper explored DES algorithm in client side for generating the secret key to encrypt the items of the support table. Apriori algorithm is used at server side for getting transaction data from the client side by applying threshold or sigma value to filter out the item set whose frequency has to less than sigma value. The main focus of this paper is to achieve more security of the client side.

Julius Adebayo, Lalana Kagal[12] has presented a transformation procedure for large scale individual level data that produces output data in which no linear combinations of the resulting attributes can yield the original sensitive attributes from the transformed data. Iyer Chandrasekharan, P.K. Baruah Prashanti Nilayam[13] proposed a novel hybrid method to achieve k-support anonymity based on statistical observations on the datasets.

III. PROPOSED METHODOLOGY

A. Overview

The proposed method of privacy preserving using tree distribution approach is described in the Fig.1.The following steps are used to provide privacy.

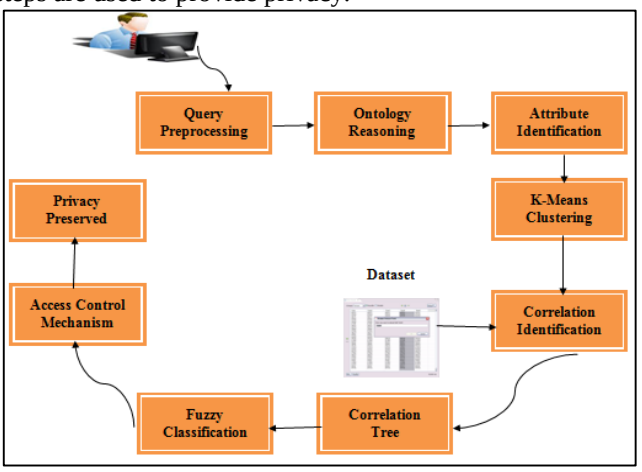

Fig. 1: Overview of Privacy Preserving using Tree Based Approach

B. Proposed System

Let 'S' be the system for privacy preserving for NON-IID dataset which contains input(I) as query from user.

Output (O) will be the resulted data with privacy.

Let T_n be the data set, Q_n be the set of query keywords on the basis of users query.

The correlation factor is being calculated in the datasets in which numbers of records are considered as $x_1, x_2 \ldots x_n$ which belongs to T_n.

Privacy can be preserved on the basis of what kind of query user is firing. So Q_n is considered set of query keywords which are related to sensitive information present in the dataset.

Different processes are used to find the strict correlation factor.

1) Preprocessing

Preprocessing is done on query given as input in which special symbols (.,;') are removed. Stop words (is, are, was, were, it, but) are removed. Stemming is used to get proper word without any 'ing', 'ly', 'ed'.

Output is a set of preprocessed data attributes.

2) Extracted Feature

Input to this step is preprocessed query.

Output is a extracted features.

Features gets extracted in which occurrence of attributes gets calculated. Also pronoun from query gets extracted.

3) Ontology Reasoning

Extracted attributes are scrutinized in such a way that it gives meaning so it will be easy to understand. To identify hierarchy of relation OWL file is scanned. So that attribute gets identified.

4) Cluster Computation

Input for this step is labeled dataset that is dataset is labeled for computing the Euclidean distance. Output is in the form of clusters. Euclidean distance is computed using following equation (1).

$$d(p,q) = d(q,p) = \sqrt{((q_1 - p_1)^2) + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2)} \quad (1)$$

Where
p= Euclidean vectors, q= Euclidean vectors

Algorithm for cluster computation

- Step 1: Initialize the center of the clusters
- Step 2: Attribute the closest cluster to each data point
- Step 3: Set the position of each cluster to the mean of all data points belonging to that cluster
- Step 4: Repeat steps 2-3 until convergence

5) Correlation Computation

Input to this step is match vector in which particular attribute relation identified in the clusters. Output is correlation among the attributes and can be calculated using following equation (2)

$$C = \frac{I_a}{T_a} \qquad (2)$$

Where,
I_a= Total number of attribute relation identification
T_a = Total number of excited attributes

IV. RESULTS AND DISCUSSION

Experimental evaluation is carried on system of privacy preservation on non IID dataset which is developed with the approach of tree pattern analysis. Proposed system is developed on java based windows machines which uses Netbeans as IDE. System is put under hammer for crucial testing for its authenticity as mentioned in below tests.

The Privacy Preserving using Tree Distribution Approach system uses gives privacy to dataset as how sensitive information a user ask to system. The main aim of the PPTDA system is to find the correlated information. On providing access control mechanism user will get output. The Privacy Preserving using Tree Distribution Approach system is compared with Association Rule Mining Approach. In PPTDA system 'Adult' dataset is used for computation.

Evaluation measures like precision, recall, F-measure can easily be calculated using following formulae.

A. Precision

Precision and recall are two widely used metrics for evaluating performance. Precision is used to measure exactness. Precision is "how useful the search results are". Precision is the number of relevant attributes divided on the probabilistic irrelevant attributes. This is shown in the following formula (3).

$$\text{Precision: } \frac{Relevant}{Relevant + Probablistic\ Irrelevant} \qquad (3)$$

B. Recall

Recall is "how complete the results are". Recall is a measure of completeness. Recall is the number of relevant attributes divided on the total relevant not identified. This is shown in the following formula (4).

$$\text{Recall: } \frac{Relevant}{Relevant + Relevant\ Not\ Identified} \qquad (4)$$

C. F-Measure

F-Measure is calculated as the harmonic mean of precision and recall (5). This gives a score that is a balance between precision and recall. F-Measure combines them into one score for easier usage.

$$F = \frac{2 \times precision \times recall}{precision + recall} \qquad (5)$$

The ARMA system [1] implements privacy preserving using association rule mining and frequent item set mining gives the precision 33%, recall 100% and F-measure 49.62%. The privacy preserving using tree distribution approach system gives better results than the association rule mining approach such as precision 42%, recall 100%, F-measure 59.15% as shown in Fig.2.

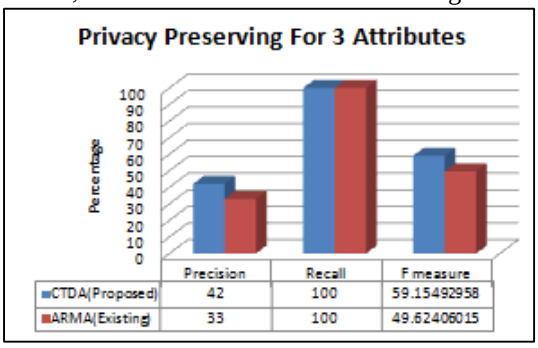

Fig. 2: Comparative Results using Evaluation Parameter Precision, Recall and F-measure (For 3 Attributes)

Likewise different identified attributes are considered i.e. as shown in the fig 3 the privacy preserving for four attributes is shown.

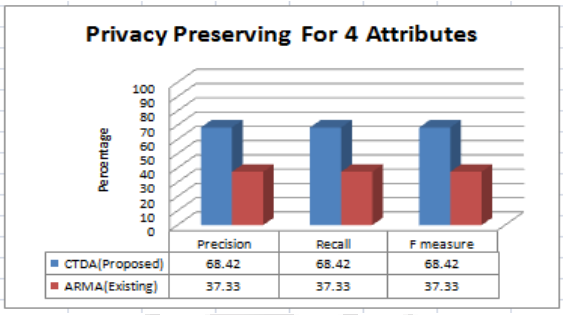

Fig. 3: Comparative Results using Evaluation Parameter Precision, Recall and F-measure (For 4 Attributes)

The privacy preserving for five attributes is shown in the following fig. 4.

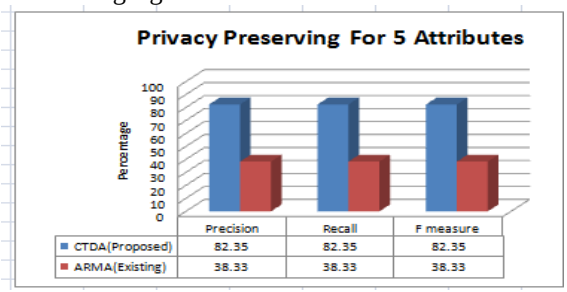

Fig. 4: Comparative Results using Evaluation Parameter Precision, Recall and F-measure (For 5 Attributes)

The privacy preserving for five attributes is shown in the following fig. 5.

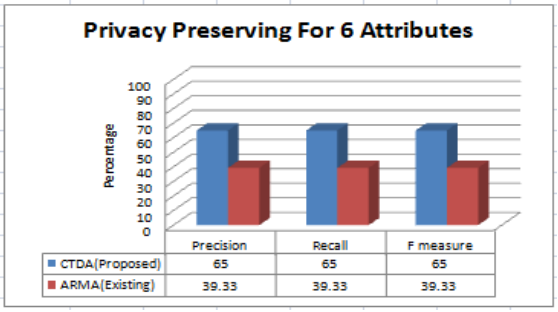

Fig. 5: Comparative Results using Evaluation Parameter Precision, Recall and F-measure (For 6 Attributes)

The privacy preserving for five attributes is shown in the following fig. 6.

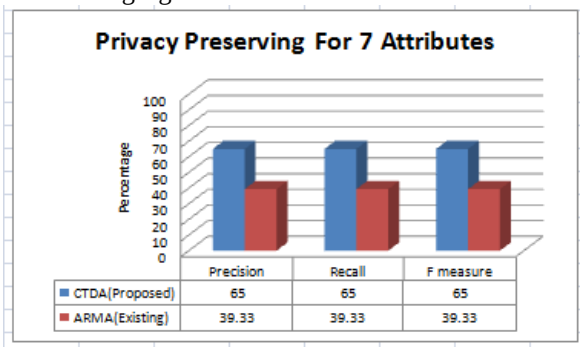

Fig. 6: Comparative Results using Evaluation Parameter Precision, Recall and F-measure (For 7 Attributes)

V. CONCLUSION AND FUTURE SCOPE

This system gives comprehensive and general approach named correlated tree based approach for discovering informative knowledge in complex dataset for any kind of data mining applications. In proposed method, Access control policies are provided which is not provided in the existing system. . In PPTDA system privacy has been given to identify attributes is based on the correlation amongst them but in the ARMA system [1] the attributes are scrutinized using frequent itemset mining in which only identified attributes gets privacy. The main aim of PPTDA (proposed system) is to find the Correlated information and provide privacy. Also different mechanisms as well as algorithms have been proposed that how the sensitive information can be protected. The system evaluated with the help of evaluation parameter that is Precision, Recall and F-Measure.

In future, more work is needed on further improving in privacy concern. It can be applied for new applications. Although the techniques and algorithms used for preserving privacy are advancing fast, however, a lot of problems in this field of study remain unsolved. System can enhanced to work on heterogeneous dataset.

REFERENCES

[1] Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao , "Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases" IEEE transactions on Information Forensics And Security, Volume:11, May 2016

[2] Tianqing Zhu, Ping Xiong, Gang Li, "Correlated Differential Privacy: Hiding Information in Non-IID Dataset" Information Forensics and Security, IEEE Transactions on,Volume:10, Feb. 2015

[3] C. Gokulnath, M.K. Priyan, E. Vishnu Balan, Prof. K.P. Rama Prabha and Prof. R. Jeyanthi, "Preservation of Privacy in Data Mining by using PCA Based Perturbation Technique" 2015 International Conference on Smart Technologies and Management 978 -1-4799-9855-5/15/$31.00 ©2015 IEEE

[4] YilinShen and Hongxia Jin, "Privacy-Preserving Personalized Recommendation: An Instance-based Approach via Differential Privacy", IEEE International Conference on Data Mining, 2014.

[5] Ali Makhdoumi, Nadia Fawaz, "Privacy-Utility Tradeoff under Statistical Uncertainty", Fifty-first Annual Allerton Conference 2013.

[6] Jun Zhang, Yang Xiang, Yu Wang, WanleiZhou, Yong Xiang, and Yong Guan, "Network Traffic Classification Using Correlation Information", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 1, JANUARY 2013.

[7] Anil Pratap Singh and AbhishekMathur , "A Chaotic Based approach for Privacy Preserving Data Mining Applications with Multilevel Trust", 2013 InternationalConference on Green Computing, Communication and Conservation of Energy (ICGCE), 978-1-4673-6126-2/13/$31.00 c 2013 IEEE.

[8] RahenaAkhter ,RownakJahanChowdhury, Keita Emura, Tamzida Islam, Mohammad ShahriarRahman, NusratRubaiyat, "Privacy-Preserving Two-Party k-Means Clustering in Malicious Model", IEEE 37th Annual Computer Software and Applications Conference Workshop, 2013..

[9] M Balamurugan, J Bhuvana and S ChenthurPandian, "Shared and Secured Data Partitioning For Privacy Preserving of Collaborative File Transfer in Multi Path Computational Mining ",978-1-4673-1989-8/12/$31.00 ©2012 IEEE.

[10] Pui K. Fong and Jens H. Weber-Jahnke, "Privacy Preserving Decision Tree Learning using Unrealized Data Sets", IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 2, FEBRUARY 2012

[11] Xiaoyan Zhu, Momeng Liu, and Min Xie, "Privacy-Preserving Affinity Propagation Clustering over Vertically Partitioned Data", Fourth International Conference on Intelligent Networking and Collaborative Systems, 2012.

[12] Haibat Jadhav, Prof. Pankaj Chandre, "Association Rule Mining Methods for Applying Encryption Techniques in Transaction Dataset", 2016 International Conference on Computer Communication and Informatics (*ICCCI* - 2016), Jan. 07 – 09, 2016

[13] Julius Adebayo, Lalana Kagal, "A Privacy Protection Procedure for Large Scale Individual Level Data", 2015 IEEE

[14] Iyer Chandrasekharan, P.K. Baruah Prashanti Nilayam," Privacy-Preserving Frequent Itemset Mining in Outsourced Transaction Databases" , International Conference on Advances in Computing, Communications and Informatics (ICACCl), 2015

[15] https://archive.ics.uci.edu/ml/datasets/Adlt

