International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.41 (2015)

Distinctive Multi-Party Secure Data Publishing for Vertically Segregated Data

S.Divya[1]/PG Scholar ,B.Santhosh Kumar[2]/AP(SG),Dr.S.Karthik[3]/Professor&Dean
Computer Science & Engineering
SNS College of Technology
Coimbatore, India
[1]subudivya91@gmail.com,[2] b.santhoshkumar@gmail.com, [3]profskarthik@gmail.com

***Abstract*- Differential privacy is a process of protecting the sensitive data from disclosure and holds strongest privacy. Differential privacy requires that computations that need to be insensitive to changes in any particular individual are record, thereby restricting data leaks through the results. The problem arises as where the details of an individual are shared by multiple parties. The existing method uses two party protocol and algorithm to release the private data in a secure way and it is applicable only for the two parties. So to overcome this problem a stochastic descent granite algorithm is framed for multi-parties to publish an anonymized view of the integrated data. A data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties. Experimental result of the proposed protocol can be more effectively secure the information than the two party protocols.**

***Keywords*- Data Mining, Privacy Preserving, Differential Privacy Techniques, Vertically Partitioning data, Stochastic Descent Granite Algorithm**

I INTRODUCTION

Data Mining is the practice of mechanically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data Mining involves the use of complicated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Data Mining is also known as Knowledge Discovery in Database (KDD). Data mining is solitary of the analysis steps in KDD. In a various application of data mining, privacy preserving techniques plays a significant role to prevent this move towards from intruders. Privacy preserving data mining techniques has emerged to address this issue. The concept of privacy preserving data mining has been proposed in response to these confidentiality concerns. Privacy preserving data mining aims at providing a transaction between sharing information for data mining analysis, on the one side, and defending information to preserve the privacy of the involved parties on the other side. The several approaches used by PPDM can be summarized as the data is altered before delivering it to the data miner. The data is distributed between two or more sites, which cooperate using a semi-honest protocol to learn global data mining results without revealing any information about the data at their individual sites. Privacy preserving data mining techniques clearly depend on the definition of privacy, which captures what information is sensitive in the original data and should be protected from either direct or indirect (via inference) disclosure. Privacy in the pasture of data management deals with the problem of concealing sensitive information about individual records. The main technique explored by the research literature follows the method of domain generalization. The ultimate goal is to conceal each individual tuple into an appropriately constructed group of data, in a way that an intruder cannot easily reason about the participation of individuals into the group.

In the privacy model, Differential Privacy is a technology that enables analysts to extract the useful answers from the database containing all the personal information and provides one of the strongest privacy guarantees. Its aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records. Differential privacy guarantees practical resolution to this dispute. Differential privacy is preserved the charity of any one individual to the answer of any question must be insignificant, in a precise mathematical sense. Differential Privacy is a relatively new privacy ensuring mechanism, but as the number and volume of databases with private data keep on to grow, this will continue to be a powerful and important tool.

In this paper, commonly used differential privacy mechanisms are interactive and non-interactive and these are involved in the exitisting methodology and proposed methodology.

II MOTIVATION

Differential privacy is a rigorous privacy grantees based upon two approaches: *interactive* and *non-interactive*.

A. Interactive Approach

In an interactive approach, a data miner can arise question through a private mechanisms and a database holder answers these queries in response. The multiple queries that to be posing in the data miner and database owner answers these queries in response. Comparative to fixed accuracy and privacy constraint, this mechanism can answer exponentially more queries than the previously best known interactive privacy mechanism. The interactive approach also referred in a Privacy Preserving Distributed Data Mining (PPDDM). In interactive approach of PPDDM, multiple data holder need to work out a function based on their inputs without sharing their data with anyone. In recent years, different protocols have been proposed for data mining tasks. However, none of the methods provide any privacy grantees on computed output. But the interactive algorithms posed to compute differentially count queries for both horizontally and vertically partitioned data.

The interactive approach is focus on the question database-answering, are not gladly applicable to PPDP, where the data publisher may not have complicated database management knowledge, which does not want to provide

an interface for database questions. A data publisher carries a hospital which has no purpose of being a database server, answering database queries is not part of its normal business.

B. Non-Interactive Approach

In a non- interactive approach, a database holder first anonymized the raw data and then released the anonymizes version for data analysis. Just the once the data are published. The data holder has no more control over the published data. The non- interactive means nothing but the data are sanitized and then release the data. The non-interactive approach also referred in a privacy preserving distributed data mining (PPDDM). In non- interactive approach in PPDDM, allows anonymizing data from different source of data release without revealing the sensitive information.

The non-interactive algorithm is to securely integrate horizontally partitioned data from multiple or various data holders without disclosing data from one party to another. The non- interactive query model is a statistical disclosure control, in which the data recipient can scan and submit one query to the system. This kind of non-interactive query model may not fully deal with the information which describe the needs of data recipients because, in some other cases, it is very complicated for a data recipient to exactly construct a query for a data mining assignment in one shot.

C. Comparison Interactive Vs. Non – Interactive

When Compared to an interactive approach Vs a Non-interactive approach , non- interactive approach gives greater flexibility since data holder can perform their required analysis and data investigation, such as mining patterns in a particular group of records, visualize the transactions containing the exact pattern or trying different modelling methods and parameters.

D. Two party Vs. multiparty

Data are owned by different set of attributes of same set of individual are held by two parties. In a multi party scenario, the data owners want to achieve the same tasks as multi parties on their integrated data without sharing their data with others

III EXISTING METHODOLOGY

The private data publishing is used as the existing system where different attributes for the same set of individuals are held by two parties. Differential privacy is a rigorous privacy model that makes no assumption about an adversary's background knowledge. A differentially private mechanism ensures that the probability of any output (released data) is equally likely from all nearly identical input data sets and thus guarantees that all outputs are insensitive to any individual's data. In other words, an individual's privacy is not at risk because of the participation in the data set. In particular, it presents an algorithm for differentially private data release for vertically partitioned data between two parties in the semi-honest adversary model. To achieve this, first present a two-party protocol for the exponential mechanism. It achieves the two-party algorithm that releases differentially private data in a secure way according to the definition of secure multiparty computation. This protocol can be used as a sub-protocol by any other algorithm that requires the exponential mechanism in a distributed setting. It solves the distributed and non-interactive scenario.

A. Dataset upload and viewing process

Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as age and salary of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. In this implementation adult data set is used for the mining process.

B. Data Pre-process

Data populate process is performed to transfer textual data into relational database. Meta data provides the information about the database transactions. Data cleaning process is initiated to correct noisy transactions. Missing values are updated using aggregation based data substitution mechanism.

C. Data Generalization

The data generalization process, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '$20 < \text{Age} \leq 30$' , etc. The data has 2-anonymity with respect to the attributes 'Age', 'Gender' and 'State of domicile' since for any combination of these attributes found in any row of the table there are always at least 2 rows with those exact attributes. The attributes available to an adversary are called "quasi-identifiers". Each "quasi-identifier" tuple occurs in at least k records for a dataset with k anonymity.

A generalization replaces some values with a parent value in the taxonomy of an attribute. The reverse operation of generalization is called specialization. A suppression replaces some values with a special value, indicating that the replaced values are not disclosed. The reverse operation of suppression is called disclosure.

D. Two-Party Protocol for Exponential Mechanism

Exponential mechanism chooses a candidate that is close to optimum with respect to a utility function while preserving differential privacy. In the distributed setting, the same candidates are owned by two parties while records are horizontally-partitioned among them. Consequently, need a private mechanism to compute the same output while ensuring that no extra information is leaked to any party. The two-party protocol for exponential mechanism in a distributed setting is presented.

E. Sensitive score updating

User will request the needed data this requested data will be sent to the admin having the data set. Administrator can decide which data can be provided to the user. Based on the user requests all the details of the user will be viewed by the admin such as user name, password. Admin will decide which data can be provided to the third party based on their particular details. For ex. Party A table details or Party B table details.

IV PROPOSED METHODOLOGY

A. Stochastic Gradient Descent (SGD) Algorithm

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning. SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Given that the data is sparse, the classifiers in this module easily scale to problems with more than 10^5 training examples and more than 10^5 features. The advantages of Stochastic

Gradient Descent are Efficiency and Ease of Implementation (lots of opportunities for code tuning). The disadvantages of Stochastic Gradient Descent include SGD requires a number of hyper parameters such as the regularization parameter and the number of iterations. SGD is sensitive to feature scaling.

B. Anonymous privacy based distribution

Input: Data set (Adult data set)

Output: Distributed Anonymous data

- Initially preprocess the raw data set AS = $\{a_1,..,a_n\}$, with a_i the attribute at position i.
- Generalize the data set with general attributes.
  - Remove the null value attribute data
  - Select attributes for generalize the original data set AS.
  - Explicit identifier selection to do the prediction of sensitive data SS: $c_j=(w_i)2$
  - Sensitive attributes privacy mechanism
- Anonymizes sensitive attributes which are predicted as sensitive.

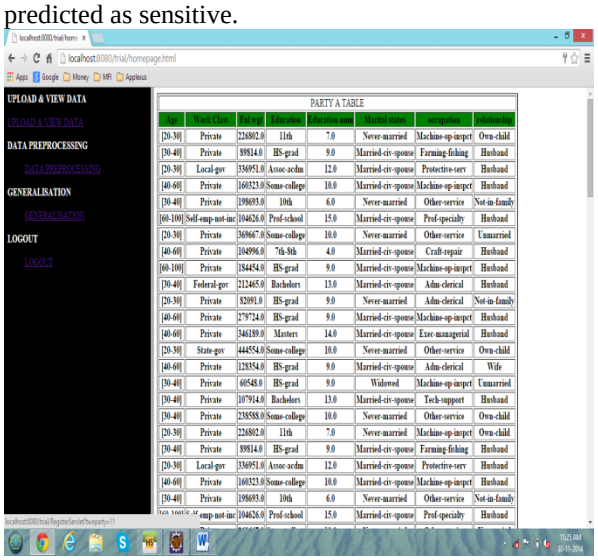

Figure Original Table Party A

Figure Original Table Party B

V CONCLUSION

This paper has presented a novel approach of privacy preserving concepts that is based on the data mining techniques which can be used over the internet and other social networks. The main goal of using this differential privacy technique is to provide one of the strongest privacy grantees in the problem of private data publishing among various techniques. The PPDM using differential privacy techniques show that it's possible to ensure privacy guarantees. In the existing methodology, the private data publishing which is used where the different attributes can be held by two parties. Here two-party protocol is used for anonymized view of integrated data and provide the privacy for individual records.

By using the proposed methodology with the existing methodology, the proposed methodology technique tells about the Stochastic Gradient Descent (SGD) Algorithm which is mainly used for partitioning the data that is more secure the data or information. The proposed algorithm is more efficient in securing the data of each individual's data.

REFERENCES

[1] N. Mohammed, "Secure Two-party Differentially Private Data Release for Vertically Data", Vol. 11, No. 1, 2014.

[2] N. Mohammed, R. Chen, "Differentially Private Data Release for Data Mining", ACM Int'l Conf. Knowledge Discovery and Data Mining, 2011.

[3] B. C. M. Fung, "Privacy-Preserving Data Publishing: A Survey of Recent Developments", ACM Computing Surveys, Vol. 42, No. 4, pp. 1-53, June 2010.

[4] X. Xaio, "Differential Privacy Via Wavelet Transforms", Proc. IEEE In'l Conf. Data Engg., 2010.

[5] A. McGregor,"The Limits of Two-Party Differential Privacy", Proc. IEEE Symp. Foundations of Computer Science (FOCS '10), 2010.

[6] Bunn.P and Ostrovsky.R, (2007) , "Secure Two-Party K-Means Clustering," Proc. ACM Conf. Computer and Comm. Security (CCS '07).

[7] C. Dwork, "A Firm Foundation for Private Data Analysis", Comm. ACM, Vol. 54, No. 1, pp. 86-95, 2011.

[8] Fida K. Dankar, "Practicing Differential Privacy in Health Care: A Review", CHEO Research Institute, November 2005.[9] A. Friedman, "Data Mining with Differential Privacy", Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '10), 2010.

[10] K. Chaudhuri, "Differentially Private Empirical Risk Minimization", J. Machine Learning Research, Vol. 12, pp.1069-1109, July 2011.

[11] N. Li, T. Li and Venkatasubramanian, "t-closeness: privacy behind k-anonymity and l- diversity", Proc. IEEE Int'l Conf. Data Engg. 2007.

[12] Clifton.C, Kantarcioglu.M, "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 28-34.

[13] Fung B.C.M, Wang. K, Chen.R, and Yu. P.S, (2010), "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, pp. 1-53.

[14] Xu, Yang, et al. (2014)"A Survey of Privacy Preserving Data Publishing using Generalization and Suppression." Appl. Math 8.3,pp. 1103-1116.

[15] Jiang.W and Clifton.C,(2006), "A Secure Distributed Framework for Achieving k-Anonymity", Very Large Data Bases J., vol. 15, no. 4, pp. 316-333.