International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 11, November-2014

A Survey on Privacy Preserving Data Mining Techniques using Differential Privacy

S. Divya[1] /PG Scholar
Computer Science & Engineering
SNS College of Technology
Coimbatore, India

B. Santhosh Kumar[2]/AP(SG)
Computer Science & Engineering
SNS College of Technology
Coimbatore, India

Dr. S. Karthik[3]/Professor & Dean
Computer Science & Engineering
SNS College of Technology
Coimbatore, India

Abstract :- **Data Mining is one of the processes of extracting the interesting or useful information from the large database. Here, we review about the data mining and privacy preserving concepts. Privacy preserving in data mining techniques plays an important role in recent years over the internet and the social networks. Among the privacy preserving model, Differential Privacy is a technology that enables analysts to extract the useful answers from the databasecontaining all the personal information and provides one of the strongest privacy guarantees. In this survey paper, it mainly described about the differential privacy techniques which offers the strong individual protection, interactive and non-interactive approach and their applications to prevent the data from the unauthorized user or customer.**

Keywords- ***Data Mining, Privacy Preserving, Differential Privacy Techniques***

I INTRODUCTION

Data mining has attracted a great deal of attention in the information industry and in society as whole in recent years. In a Wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data Mining is also known as Knowledge Discovery in Database(KDD). Data mining is one of the analysis steps in KDD.Inreal-life application of data mining, privacy preserving techniques plays an important role to prevent this approach from intruders. First, the problem might be the member should join on the online social networks, to create a profile, using the different application offered by the service, it possibility to easily share the information with selected contacts or public. Second, the problem might be the sophiscated data mining algorithm. Third, the problem might be increase in large data storage. Privacy preserving data mining techniques has emerged to address this issue.The main goal of privacy preserving data mining(PPDM) is how to protect the sensitive information or private knowledge leaking in the data mining process and also obtain an accurate result of data mining.The privacy preserving techniques in data mining a several approach is summarized into 3 levels:

First level of PPDM is focus with protecting the sensitive information and sensitive data or profile such as id, name, address.Second level of PPDM is focus with data mining algorithm in order to protect the extraction of sensitive information during the knowledge finding.Third level of PPDM is focus with protecting the sensitive knowledge which is showed by data mining.

In the privacy model, differential privacy is one of the techniques for releasing statistical information about a database without revealing information about its individual entities and its prove its one of the strongest privacy guarantees. Itsaims 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 contributions of any one individual to the answer of any question must be insignificant, in a precise mathematical sense. Differential privacycreates the tantalizing possibility of privacy-preserving data analysis that is both useful and provably secure.

In this survey paper, commonly used differential privacy mechanisms are interactive and non-interactive and these are involved in the application of differential privacy to the health data is presented.

II DIFFERENTIAL PRIVACY MECHANISM

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 pretense question through a private mechanisms and a database holder answers these queries in response. The multiple queries to be pose 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 in PPDDM,multiple data holder need to work out a function based on their inputs without sharing their data with others.

In recent years, different protocols have been proposed for data mining tasks. However, none 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 question database-answering, are not gladly applicable to PPDP, wherethe data publisher may not have complicated database management knowledge, not want to provide an interface for database question. A data publisher, such as ahospital, has no purpose of being a database server; answering database queries isnot part of its normal business.

B. Non-Interactive Approach

In a**non- interactive approach,**a database holder first anonymized the raw data and then release theanonymizes 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 data holders without disclosing data from one party to another. The non- interactivequery model is a statisticaldisclosure control, in which the data recipient canscan and submit one query to the system. This kind of non-interactive query model maynot fully deal with the information needs of data recipients because, in some other cases, itis very complicated for a data recipient to exactly construct a query for a data miningtask in one shot.

C. Comparison Interactive Vs. Non – Interactive

When Compared to an interactive approach Versus a Non- interactive approach , non- interactive approach gives greater flexibility sincedata holder can perform their required analysis and data investigation, such as mining patterns in a particulargroup of records, envisage the transactions containing theexact pattern or trying different modeling methods andparameters.

III APPLICATION OF DIFFERENTIAL PRIVACY TO PREVENT DATA

Application of Health Data is one of the areawhich givethrust for differential privacy.The disclosure of health data has a number of uniqueness that need to be considered in any practical mechanism used to preserve privacy. Differential privacyconcernare technological, and some are more social, these concernsmust all be considered before a practical health data differentialprivacy system is developed and it is used to actually protect thegeneral public's private data.

A. Data types

Health data contains definite data such as diagnosis codes, procedurecodes, drugs dispensed, laboratory test ordered, and geographical data about the patient and obtained. There is also numericdata such as age, length of stay in hospital, and time as last visit.Both types of variables need to be addressed by any practical solution.

B. User receipt

As it is not a technical concern, the paper points out some significantconcerns about user acceptance of new differential privacyapproach. Health data is often disclosed by professionals that havea set method and established code. Because it will be challenging toconvince users to discard their established code, in a petite term, a non-interactive mechanism would be most suited forthis community.

C. Social Networks

One of the applicationsplaysa important role in social networks using differential privacy. The main Challenge of Preserving privacyin a social network because of the properties of individual node protected from each other. Many online social-network services such as Facebook allow users to configure their individual privacy policy with a high level of granularity.While the differential privacy algorithm givesus assurances about the effect of the removal of a certain node fromthe database, there is the matter of influence over other nodes. Forexample, if a tuple contains information about a user Alice, deletingAlice's tuple would not remove the influence of this tuple on thedata. This influence causes there to be evidence of Alice's tupleeven after Alice's tuple has been removed from the data.

D.Convincing the Public

Convincing the public that stewardship of their data is being conducted in a responsible way is becoming a necessary objective. For example, patients and providers have expressed concerns about the disclosure and use of health information, and there is proof that patients adopt privacy protective behaviors when they concern about how their own information is being used or disclosed, especially among vulnerable patient groups.A prosperity of data about individuals is constantly accumulate in various databases in the form of medical report, social network graphs, movability traces in cellular network, explore logs, and picture ratings.

There are many expensive which are used by the users of such datasets, but it is very hard to realize these used while preventing privacy. Even when data collectors try to protect the privacy of their customers by releasing anonymized. This data often reveals much more information than intended. To consistently prevent such privacy violations, we want to restore the current ad-hoc solutions with a principled data release mechanism that offers strong, provable privacy guarantees. Differential privacy allows us to reason formally about what an adversary could learn from released data, the breakdown of which have been the cause of privacy violation in the past.

IV CONCLUSION

This survey paper has presented a novel approach of privacy preserving concepts thatis based on the data mining techniques which can be used over the internet and other social networks. 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. The main goal of using the differential privacy techniques is to provide one of the strongest privacy grantees in the problem of private data publishing. As theusage of data mining for potential intrusive purposes using personally identifiable information increases, privately using these results will become more important. The above PPDM using differential privacy techniques show that it's possible to ensure privacy guarantees and It plays an important role in PPDP using the interactive and non-interactive approach and discussed about them which one is best and better to use in the networks. The highlightedapplication of differential privacy is useful to prevent the data during the data releasing and potentially lead to more adoption differential privacy in health care.

V 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'1 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'1 Conf. Data Engg., 2010.
[5] K. Chaudhuri, "Near Optimal Differentially Private Principal Components", Proc. Conf. Neural Information Processing Systems, 2012.
[6] A. McGregor, "The Limits of Two-Party Differential Privacy", Proc. IEEE Symp.Foundations of Computer Science (FOCS '10), 2010.
[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'1 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,July2011.