Privacy Preserving Data Mining Techniques by using a Data Anonymization - based Framework

[1]**Arti K. Pandya**, [2]**Priyanka R. Raval**

[1]M.E. Student, [2]Assistant Professor,
Department of Computer Science
BHGCET, Rajkot, India.

***Abstract*— Data mining techniques are used to mine the knowledge from a large database. But sometime these knowledge can disclose sensitive information about the data holder or individuals. The goal of privacy preservation technique is to release aggregate information about the data, without leaking individual information about participants. Several techniques of privacy preserving data mining have been proposed in literature. Privacy in data mining can be obtained by various techniques like Perturbation, Anonymization and Cryptographic. This paper tries to reiterate various Privacy Preserving Data Mining (PPDM) techniques based on anonymization currently developed to meet the privacy issues in the process of data mining.**

***Index Terms*— data mining; privacy preserving; sensitive attributes; privacy; privacy preserving techniques**

I. INTRODUCTION

Mining a knowledge from a large amount of data is very use for business application, health care government agencies. But the process of mining a knowledge from a large database, the privacy of individual is became challenging task. As a result, a whole new set of approaches were introduced to allow mining of data, while at the same time prohibiting the leakage of any private and sensitive information. The majority of the existing approaches can be classified into two broad categories [1].

(i) methodologies that protect the sensitive data itself in the mining process, and

(ii) methodologies that protect the sensitive data mining results (i.e. extracted knowledge) that were produced by the application of the data mining.

The first category refers to the methodologies that apply perturbation, sampling, generalization / suppression, transformation, etc. techniques to the original datasets in order to generate their sanitized counterparts that can be safely disclosed to untrustworthy parties. The goal of this category of approaches is to enable the data miner to get accurate data mining results when it is not provided with the real data.

The second category deals with techniques that prohibits the disclosure sensitive knowledge patterns derived through the application of data mining algorithms as well as techniques for downgrading the effectiveness of classifiers in classification tasks, such that they do not reveal sensitive knowledge.

Now- a-days, Data Mining is used in many applications. There are certain areas where data mining if used without privacy may cause serious affects .These areas are the main research challenges and are mentioned below.

A. Cyber Terrorism, Insider Threats, and External Attacks

One of the major threats people face today is Cyber Crime [2]. Since most of our information is stored on electronic media and a lot of data is also available on internet or networks. Attacks on such areas might be dangerous and devastating for an individual. For example, consider the Banking system. If hackers attack a bank's information system and empty the accounts, the bank could lose millions of dollars. Therefore security of information is a critical issue. There are two types of threats –Outsider or Insider. An attack on Information System from someone outside the organization is called outsider threat, such as hackers, hacking Bank's computer systems and causing havocs. A more critical problem is the insider threat. Insider threat can be due to an intruder present in the organization. Members of an organization have studied their policies and business practices and know every bit of the information so it can affect the organization's information assets.

B. Credit Card Fraud and Identity Theft

Another area which requires attention is detecting frauds and thefts. Frauds may be credit card frauds [2]. These can be detected by identifying purchases made of enormous amounts. A similar and a more serious theft is identity theft. Here one pretends to be an identity of another person by obtaining that person's personal information and carrying out all

types of transactions under the other person's name. By the time, the owner finds out it is often far too late-the victims may already have lost millions of dollars due to identity theft.

In today's increasingly digital world, there is often a tension between safeguarding privacy and sharing information. On the one hand, sensitive data needs to be kept confidential; on the other hand, data owners are often motivated or forced to share sensitive information. Consider the following examples:

A. Aviation Safety

The Department of Homeland Security (DHS) checks whether any passengers on each flight from/to the United States must be denied boarding or disembarkation, based on several secret lists, including the Terror Watch List (TWL) [3]. Today, airlines surrender their passenger manifests to the DHS, along with a large amount of sensitive information, including credit card numbers [4]. Besides its obvious privacy implications, this modus operandi poses liability issues with regard to mostly innocent passengers' data and concerns about possible data loss. (See [5] for a litany of recent incidents where large amounts sensitive data were lost or mishandled by government agencies.) Ideally, the DHS would obtain information pertaining only to passengers on one of its watch lists, without disclosing any information to the airlines.

B. Law Enforcement

An investigative agency (e.g., the FBI) needs to obtain electronic information about a suspect from other agencies, e.g., the local police, the military, the DMV, the IRS, or the suspect's employer. In many cases, it is dangerous (or simply forbidden) for the FBI to disclose the subjects of its investigation. Whereas, the other party cannot disclose its entire dataset and trust the FBI to only extract desired information. Furthermore, FBI requests might need to be pre-authorized by some appropriate authority[6] (e.g., a federal judge). This way, the FBI can only obtain information related to authorized requests.

C. Healthcare

A health insurance company needs to retrieve information about its client from other entities, such as other insurance carriers or hospitals[6]. The latter cannot provide any information on other patients and the former cannot disclose the identity of the target client.

Other examples of sensitive information sharing include collaborative botnet detection [7] (i.e., service providers share their logs for the sole purpose of identifying common anomalies), interest sharing from smart phones [8], or preventing cheating in online gaming [9].

PRIVACY PRESERVING DATA MINING

Privacy-preserving data mining finds various applications in surveillance which is naturally expected to be "privacy-violating" applications. The key is to plan routines which continue to be viable, without compromising security. Various methods have been talked about for bio surveillance, facial de-identification, and data fraud [10]. Most systems for privacy calculations utilize some type of change on the data with a specific end goal to perform the privacy protection. Typically, such techniques diminish the granularity of representation to lessen the privacy. This lessening in granularity results in a few loss of viability of data administration or mining algorithms. This is the natural exchange off between information loss and privacy.

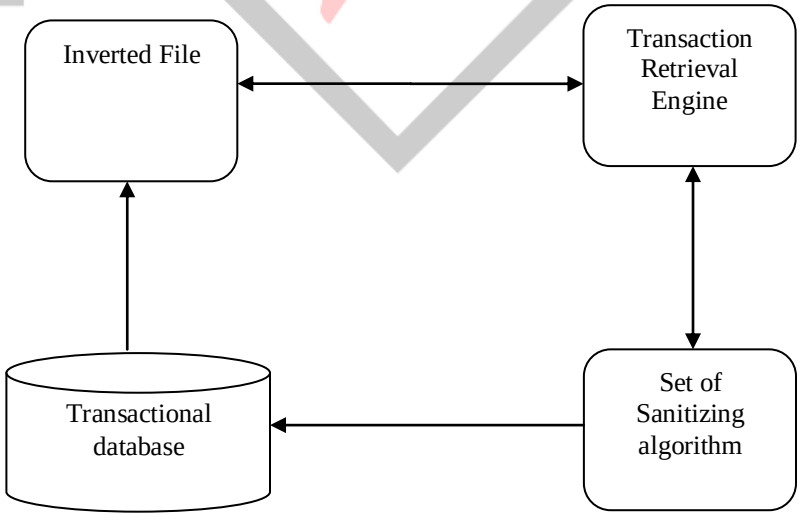

Figure 1: Framework of Privacy Preserving

Dimensions for Classification of PPDM Techniques

PPDM tends to transform the original data so that the results of data mining tasks should not defy the privacy constraints. Lists of dimensions on which PPDM techniques are based are [10]:

1. Data distribution:

This dimension is related to distribution of data. Data can be Centralized or Distributed. Distributed data can be of two types:
(i) Horizontal distribution: Refers to cases where different records reside in different places.
(ii) Vertical distribution: Refers to cases where all values of different attributes reside in different places.

2. Data Modification:

This dimension refers to modification of original values of data that are to be released for data mining task. Modification is carried out by using the following techniques:
(i) Perturbation: Perturbation of data is an easy and effective technique for protecting sensitive electronic data from unauthorized use.
(ii) Blocking: Blocking-based technique aims at hiding some sensitive information when data is shared for mining [11].
(iii) Aggregation: Data aggregation is a process in which information is gathered and expressed in a summary form for the purpose of statistical analysis [12].
(iv) Merging: Data merging refers to combination of several values of data.
(v) Swapping: Data swapping refers to interchanging values of various data records [13].
(vi) Sampling: Data sampling refers to releasing data for only a sample of population [7].

3. Data Mining Algorithms:

Data mining algorithms are applied on transformed data to get useful nuggets of information that where hidden previously.

4. Data hiding:

This dimension refers to whether the raw data or aggregated data should be hidden.

5. Privacy Preservation: T

This dimension refers to techniques that are used for protecting privacy.

ANONYMIZATION

Anonymization refers to a methodology where character or/and delicate data about record holders are to be covered up. It even accepts that delicate data should be retained for analysis [14].

There are four sort of quality of fundamental type of data [14]:

(i) Explicit Identifiers is a set of attributes containing information that identifies a record owner explicitly such as name, SS number etc.
(ii) Quasi Identifiers is a set of attributes that could potentially identify a record owner when combined with publicly available data.
(iii) Sensitive Attributes is a set of attributes that contains sensitive person specific information such as disease, salary etc.

Sometimes the data must be publically published in its original form. Even though it is not encrypted and perturbed, some sort of precaution should be implemented before releasing the data in terms of anonymization. This is a kind of generalization of some attributes which protects against identity disclosure. Anonymization can be obtained through methods such as generalization, suppression, data removal, permutation, swapping etc [15]. k-anonymity method is treated as the classical anonymization method and most of the studies are based on k anonymity. The others are based on its improved methods like l-diversity, t-closeness, k-anonymization, (α,k) anonymity, p-sensitive k-anonymity, (k,e) anonymity, which are described in [16] (t,k)-Hypergraph anonymization[17]

In 2013, researchers Abou-el-ela Abdou Hussien, Nermin Hamza, Hesham A. Hefny proposed a techniques which tried to solve the problem of common attacks on data mining by applying common attacks techniques for anonymization based Privacy Preserving Data Mining & Privacy Preserving Data Publishing. It uses k anonymity and l-diversity method of anonymization [18].

In 2014Atefeh Asayesh, Mohammad Ali Hadavi and Rasool Jalili model constraints as undirected hypergraphs and formally cluster attribute relations as hyperedge with the t-means-clustering algorithm. In addition, anonymization is carried out

with a k-anonymity method in every cluster for which the parameter k can vary in each cluster, to attain more flexibility and less information loss with respect to utility.[17]

In 2016 ahemad ali Mubarak,emad and hatem Proposed an approach to categorical data privation based on domain based semantic rule to overcome the similarity attacks.[21]

Attack on Anonymization

a) Homogeneity Attack:

Alice and Bob are hostile neighbours. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice sets out to find what ailment Bob is suffering from. Alice finds the 4 unacknowledged table of current inpatient records published by the hospital, thus she knows that one of the records in this table contains Bob's data. Since Alice is Bob's neighbour, she knows that Bob is a 31-year-old American male who lives in the postal division 13053. In this way, Alice knows that Bob's record number one among 9, 10, 11, or 12. Presently, all of those patients have the same medical condition (disease), along these lines Alice concludes that Bob has cancer.[19]

b) Background Knowledge Attack:

Alice has a pen companion named Umeko who is admitted to the same hospital as Bob, and who has some patient records. Alice knows that Umeko is a 21 year old Japanese female who currently lives in postal district 13068. In light of this information, Alice learns that Umeko's information is present in record number 1, 2, 3, or 4. Without additional information, Alice is not certain whether Umeko contracted an infection or has heart disease. On the other hand, it is also well-known that the Japanese have a too low incidence of heart disease. Along these lines Alice concludes with close certainty that Umeko has a viral infection.[19]

c) Skweness Attack:

At the point when the overall dispersion is skewed, satisfying l-diversity does not counteract characteristic disclosure.[20]

d) Similarity Attack:

At the point when the touchy attributes values in an equivalence class are distinct however semantically similar, an adversary can learn essential information.[17]

EVALUATION FRAMEWORK

The evaluation framework recommended for assessing and evaluating data modification-based techniques, is in accordance with the following eight criteria[22]:

Privacy Loss: is defined as difficulty level in estimating the original values from the perturbed data values.

Information Loss: is defined based on the amount of important data information, which needs to be saved after perturbation for data mining purposes.

Data Mining Task: is defined based on a data mining task, which contains the possibility to mine it, after applying the privacy preserving techniques.

Modifying the Data Mining Algorithms based on the needs, notifies the change for the existed data mining algorithms, in order to mine over the modified dataset. Preserved Property: that is, data information, which was already saved after applying the privacy preservation techniques.

Data Type: it points out the types of data, which could be numerical, binary, or categorical.

Indistinguishability Level: it is in accordance with level of indistinguishability of different records of the original dataset.
Data Dimension: it is defined based on the purpose of PPDM technique for preserve Dimensional information, which could be single-Dimensional or Multi-Dimensional.

CONCLUSION

This paper gives the challenges, applications and framework of privacy preservation and the dimension on which is the privacy preservation techniques is classified then different anonymization based privacy preservation and attacks in it is describes and the evaluation framework is describes. Finally we conclude there does not exists a single privacy preserving data mining algorithm that outperforms all other algorithms on all possible criteria like performance, utility, cost, complexity, tolerance against data mining algorithms etc.

REFERENCES

[1] Aris Gkoulalas-Divanis and Vassilios S. Verikios, "An Overview of Privacy Preserving Data Mining", Published by The ACM Student Magazine, 2010.
[2] M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in proceedings of ICCCNT Coimbatore, India, IEEE 2012.
[3] Federal Bureau of Investigation: Terrorist Screening Center. http:// www.fbi.gov/terrorinfo/counterrorism/tsc.htm
[4] Sherri Davidoff: What Does DHS Know About You? http://philosecurity.org/ 2009/09/07/what-does-dhs-know about-you
[5] Caslon Analytics: Consumer Data Losses.http://www.caslon.com.au/datalossnote.htm
[6] Emiliano De Cristofaro, Yanbin Lu , Gene Tsudik,"Efficient Techniques for Privacy-Preserving Sharing of Sensitive Information",2012
[7] Nagaraja, S., Mittal, P., Hong, C., Caesar, M., Borisov, N.: BotGrep:" Finding Bots with Structured Graph Analysis." In: Usenix Security (2000)
[8] De Cristofaro, E., Durussel, A., Aad, "I.: Reclaiming Privacy for Smartphone Applications." In: PerCom (2011)
[9] Bursztein, E., Lagarenne, J., Hamburg, M., Boneh, D.:" OpenConflict: Preventing Real Time Map Hacks in Online Games." In: S&P (2011)
[10] Mynavathi, R., N. Sowmiya, and D. Vanitha. "Survey of Various Techniques to Provide Multilevel Trust in Privacy Preserving Data Mining."
[11] Stanley R. M. Oliveira and Osmar R. Zaiane, "Privacy-Preserving Data Mining on the Web:Foundations and Techniques".
[12] http://searchsqlserver.techtarget.com/definition/data-aggregation.
[13] Lai Xu, Katalin Tarney and Sandor Imre, "Research and development in E-business through Service Oriented Solutions", 2013, chapter 4, pp 74.
[14] Benjamin C M Fung, Ke Wang, Rui Chen, Philip S Yu, "Privacy Preserving Data Publishing: A Survey of recent developments", ACM Computing Surveys, Vol. 42, No. 4, Article 14, June 2010.
[15] Asmaa H.Rashid and Prof.dr. Abd-Fatth Hegazy, "Protect Privacy of Medical Informatics using K-Anonymization Model", IEEE Explore.
[16] Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo, "A Survey on Privacy Preserving Approaches in Data Publishing", First International Workshop on Database Technology and Applications, 2009.
[17] Atefeh Asayesh, Mohammad Ali Hadavi and Rasool Jalili," (t,k)-Hypergraph anonymization: an approach for secure data publishing" 25 September 2014 in Wiley Online Library
[18] Hamza, Nermin, and Hesham A. Hefny. "Attacks on anonymization-based privacy-preserving: a survey for data mining and data publishing." Journal of Information Security 4: 101, 2013.
[19] Ninghui Li Tiancheng Li,Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-diversity, ICDE 2007, pp. 106–115
[20] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam. "ℓ-Diversity: Privacy Beyond k-Anonymity". Department of Computer Science, Cornell University.
[21] ahemad ali Mubarak,emad and hatem," semantic anonymization in publishing Categorical sensitive attributes",IEEE 2016.
[22] Freny Presswala, Amit Thakkar , Nirav Bhatt," Survey on Anonymization in Privacy Preserving Data Mining" International Journal of Innovative and Emerging Research in Engineering Volume 2, Issue 2, 2015.