OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

A Comparative Study on Privacy Preserving Datamining Techniques

M. Nithya[1], Dr. T. Sheela[2],

[1] Research scholar, Sathyabama University, Chennai & Assistant professor-II,
Sri sairam engineering college, chennai
[2] Professor, SriSairam engineering college, Chennai

Abstract: *Privacy protection is very important in the recent years for the reason of increasing in the ability to store data. In particular, recent advances in the data mining field have lead to increased concerns about privacy. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. The current practice in data publishing based on that what type of data can be released and use of that data. Recently, PPDM has received immersed attention in research communities, and many approaches have been proposed for different data publishing scenarios. In this comparative study we will systematically summarize and evaluate different approaches for PPDM, study the challenges ,differences and requirements that distinguish PPDM from other related problems, and propose future research directions.*

Keywords: *PPDM, Privacy-preserving; randomization; k-anonymity;*

I. Introduction

Data mining successfully extracts knowledge to support a variety of domains —marketing, weather forecasting, medical diagnosis, and national security —but it is still a challenge to mine certain kinds of data without violating the data owners' privacy.[1] For example, how to mine patients 'private data is an ongoing problem in health care applications .As data mining becomes more pervasive, such concerns are increasing. Online data collection systems are an example of new applications that threaten individual privacy. Already companies are sharing data mining models to obtain a richer set of data about mutual customers and their buying habits. A number of techniques such as classification, k-anonymity, association rule mining, clustering have been suggested in recent years in order to perform privacy preserving data mining. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. We analysis recent work on these topics, presenting general frameworks that we use to compare and contrast different approaches. We begin with the problem of focusing on different techniques of privacy preserving in section II,. In section III,we put rept attention to compare those methods and contrasted and finally we end up with conclusion and future work in subsequent sections.

II. Privacy Preserving Techniques

2.1 Anonymization Technique

When releasing micro data for research purpose,one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Samarati et al. [1]; Sweeney [2] introduced the *k*-anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least *k*-other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization. Unlike traditional privacy protection techniques such as data swapping and adding noise, information in a *k*-anonymous table through generalization and suppression remains truthful. In particular, a table is *k*- anonymous if the Ql values of each tuple are identical, to those of at least *k* other tuples. Table3 shows an example of 2-anonymous generalization for Table. With the help of table 1 and table 2 adversary can find the persons and their salary.in this case if we go for annoymiztion technique its somwhat difficult .If the adversary know the age and zipcode then easily he can find the salary of alice and carl with the help of tuple 1 and 3 in table3.

In general, *k* anonymity guarantees that an individual can be associated with his real tuple with a probability at most 1/*k*.

TABLE 1 MICRODATA

Sex	**Zip Code**	**Age**	**Salary**
F	40178	26	8000
F	40277	30	12000
M	40176	32	8000
F	40175	51	9000
F	40385	28	20000
M	40485	43	23000
M	40286	50	8000

TABLE2 POPULATION CENSUS

Name	**Sex**	**ZipCode**	**Age**
Alice	F	40178	26
Betty	F	40277	30
Carl	M	40276	32
Diana	F	40175	51
Ella	F	40385	28
Finch	M	40485	43
Gavin	M	40286	50

TABLE 3 A 2-AN0NYM0US TABLE

Sex	**ZipCode**	**Age**	**Salary**
*	4017-	26-32	8000
*	4027-	26-32	12000
*	4017-	26-32	8000
*	4017-	35-55	9000
*	4038-	26-32	20000
*	4048-	35-53	23000
*	4028-	35-53	8000

Even k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. There are two attacks: the homogeneity attack and the background knowledge attack.

2.2. Data perturbation approach

In this approach data will be modified so that it no longer represents the real world. Randomization and data swapping methods are two techniques which comes under this approach. Since this method does not reconstruct the original data values but only distributions, new algorithms need to be developed which use these reconstructed distribution in order to perform mining of the underlying data. This means that for each individual data problem such as classification, clustering, or association rule mining, a new distribution based data mining algorithm needs to be developed. For example, Agrawal [3] develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton and Rizvi and Haritsa[4] develop methods s for privacy-preserving association rule mining. While some clever approaches have been developed for distribution-based mining of data for particular problems such as association rules and classification, it is clear that using distributions instead of original records restricts the range of algorithmic techniques that can be used on the data [5].

In randomisation technique the noise will be added to the original data in randomly so that original record values cannot be guessed from the distorted data. Disadvantage in this method,it treats all records equally irrespective of their local density. Therefore, outlier records are more susceptible to adversarial attacks as compared to records in more dense regions in the data. For an example using this method ,randomly adding 50 with age attributes (instead of 26,26+50=72,80,82,etc) ,easily the adversary know that some of the noise added in that particular attribute. Second method in data perturbation method is data swapping ,in this method data values between attributes are randomly swapped .Using this method adversary can easily get the original records.

2.3 Cryptographic technique

Cryptographic technique is also used to provide privacy preserving data mining . This method became hugely popular [6] for two main reasons: Firstly, cryptography is a well-defined model for providing privacy, which includes methodologies for confirming and enumerating it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work [7] has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. So this method is not useful for provide the complete security for sensitive data in data mining.

2.4 Secure Multiparty Computation

This method reveals nothing except the results between two parties who does not want to revel their data sources using this we can prevent our sensitive attribute. Anyhow this approach contains miniature drawbacks such as Trusted Third Party Model, Semi-honest Model. In Third party model data will be shared through the third party, so the third party comes to know the data sources. In Semi-honest Model**,** Consider a secure sum functionality which simply outputs the sum of the local input of the participants. With two parties, the output reveals the input of the other party.

III. Several challenges with PPDM

Iyengar [8] demonstrated that data can be transformed in such a way as to protect individual identity. He suggests that random data can replace any individually identifiable information. The author's argument is that there is a tradeoff between privacy and information loss with this method. Thuraisingham [9] first suggested that privacy issues occur in data mining and that this is a generalization of the inference problem. The inference problem refers to an issue when a user can infer new knowledge by executing successive queries against a database. He also noted that this may cause ethical issues based on how the information is going to be used. Du and Zhan [10] proposed a randomized response technique to perturb data so that users cannot tell whether the data contains truthful information or false information. They used a decision-tree classifier along with randomization methods to perturb the data so that aggregate results still show some degree of accuracy, while at the same time maintain individual privacy. One drawback with this approach is that it only focused on Boolean data types to test their technique. Du and Zhan also neglected to define exactly what tolerances are acceptable during data mining with privacy preservation. Narayanan and Shmatikov [11] demonstrated that data can be encrypted in such a way that users can still use the information contained within it. Their study used provably secure techniques while permitting certain types of queries to be generated. A limitation to their study was that they only examined its use on small databases, not for larger databases. In order for their approach to work, they developed a new query language. Their approach may also be impractical if a user wanted to use widely available databases such as Microsoft SQL Server or Oracle. Generalization for k-anonymity losses considerable amount of information, especially for high-dimensional data due to the curse of dimensionality. In order for generalization to be effective, records in the same bucket must be close to each other so that generalizing the records would not lose too much information. Bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not. Also bucketization requires a clear separation between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs. By separating the sensitive attribute from the QI attributes, bucketization breaks the attribute correlations between the QIs and the SAs. To improve the attribute correlation slicing[12] technique has been introduced. Using slicing techniques can improve attribute correlation but it does not achieve 100 percent but it is better than k anonymity and l-diversity approaches.

IV. Merits and Demerits of PPDM techniques

PPDM Techniques	**Merits**	**Demerits**
Anonymization technique	This technique is used to protect user identities while releasing information. While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against sensitive attribute.100% accuracy can achieve.	There are two attacks: The homogeneity attack and the background knowledge attack.
Data perturbation approach	Independently the noise can add to the attributes	This method does not reconstruct the original data values ,to reconstruct the original data distribution, new algorithms have been developed . RANDOMIZED RESPONSE : Randomly the noise will be added and swapping technique have been used.

Cryptographic	Cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. There exists a vast toolset of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms.	This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.
Slicing	More efficient and better data utility compare to anonymity and l diversity method	Randomly generate the associations between column values of a Bucket. This may lose data utility. Random data transmission have been used.

V. Conclusion

With the development of data analysis and processing technique, the privacy disclosure problem about individual or company is inevitably exposed when releasing or sharing data to mine useful decision information and knowledge, then give the birth to the research field on privacy preserving data mining. In this paper, we presented different issues and reiterate privacy preserving methods to distribute ones and the methods for handling horizontally and vertically partitioned data. While all the purposed methods are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods. To address these issues, following problem should be widely studied.

1. In distributed privacy preserving data mining areas, efficiency is an important issue. We should try to develop more efficient algorithms and attain a balance between disclosure cost, computation cost
2. Privacy and accuracy is a pair of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched.
3. Side-effects are inevitable in data cleansing process. How to reduce their negative impact on privacy preserving needs to be considered carefully. We also need to define some metrics for measuring the side-effects resulted from data processing.

REFERENCES

[1] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2001.

[2] P. Samarati,(2001). Protecting respondent's privacy in micro data release. In IEEE Transaction on knowledge and Data Engineering,pp.010-027.

[3] L. Sweeney, (2002)."k-anonymity: a model for protecting privacy ", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570.

[4] Evfimievski, A.Srikant, R.Agrawal, and Gehrke J(2002),"Privacy preserving mining of association rules". In Proc.KDD02, pp. 217-228.

[5] Hong, J.I. and J.A. Landay,(2004).Architecture for Privacy Sensitive Ubiquitous Computing", In Mobisys04, pp. 177- 189.

[6] Laur, H. Lipmaa, and T. Mieli' ainen,(2006)."Cryptographically private support vector machines". In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 618-624.

[7] Ke Wang, Benjamin C. M. Fung1 and Philip S. Yu, (2005) "Template based privacy preservation in classification problems", InICDM, pp. 466-473.

[8] Iyengar, V. (2002). Transforming data to satisfy privacy constraints. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 279-288.

[9] Verykios, V., Bertino, E., Fovino, I., Provenza, L., Saygin, Y., & Theodoridis, Y. (2004). State of- the-art in privacy preserving data mining. ACM SIGMOD Record, 33(1), 50-57.

[10] Du, W., & Zhan, Z. (2003). Using randomized response techniques for privacy-preserving data mining. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 505-510.

[11] Narayanan, A., & Shmatikov, V. (2005). Obfuscated databases and group privacy. Paper presented at the Proceedings of the 12th ACM Conference on Computer and Communications Security, Alexandria, VA.

[12] “Slicing: A New Approach for Privacy Preserving Data Publishing” - Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and Ian Molloy IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012