ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

A Study on Privacy Preserving Data Mining

K.Sashirekha, B.A.Sabarish, Arockia Selvaraj

Department of Computer Science & Engineering, United Institute of Technology, Coimbatore, India.

Department of Information Technology, Amrita School of Engineering, Coimbatore, India.

Department of Computer Science & Engineering, Info Institute of Engineering, Coimbatore, India.

ABSTRACT:- Privacy Preserving and Data Mining addresses the problem of protecting the mobile users from the attackers. Privacy threat includes the process of predicting the movement pattern based on the statistical information collected. Intruder monitors the traffic models to predict the group movement and try to access the private information of mobile users. Privacy can be achieved by means of randomization, k-anonymization, and distributed privacy-preserving data mining. In order to provide better privacy multi-level frameworks are used. In this paper, an analysis is done on various methods of privacy preserving and multi-level trust policy, limitation while using large dimension data sets.

I. INTRODUCTION

Privacy-preserving uses the data mining techniques to protect the user's information from the intruders. The various methods involving randomization, k-anonymity model and l-diversity, Distributed privacy preservation, Downgrading Application Effectiveness. Privacy preserving can be used in many real time applications include the surveillance, identity check etc.

***Privacy-Preserving Data Publishing*:** These echniques tend to study different transformation methods associated with privacy. These techniques include methods such as randomization, k-anonymity ,and l-diversity. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining. Other related problems include that of determining privacy-preserving methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

Changing the results of Data Mining Applications to preserve privacy:

In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has spawned a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data.

A classic example of such techniques are association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

- Query Auditing: Such methods are akin to the previous case of modifying the results of data mining algorithms. Here, we are either modifying or restricting the results of queries. Methods for perturbing the output of queries are discussed in [8], whereas techniques for restricting queries are discussed in [9, 13].
- Cryptographic Methods for Distributed Privacy: In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function. In such cases, a

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information.

- Theoretical Challenges in high Dimensionality: Real data sets are usually extremely high dimensional, and this makes the process of privacy preservation extremely difficult both from a computational and effectiveness point of view.

II. RELATED WORK

A. The randomization method:

It is the process of adding extra noise and masks the attribute value of records. Level of noise should be chosen in an optimum way to mask the value of attributes in the records and records cannot be recovered. it is relatively simple, and does not require knowledge of the distribution of other records in the data the randomization method can be implemented at data collection time itself. Since it treats all records equally irrespective of their local density, outlier records are more susceptible to adversarial attacks. The quantity used to measure privacy should indicate how closely the original value of an attribute can be estimated. The work in uses a measure that defines privacy as follows: If the original value can be estimated with c% confidence to lie in the interval [α1, α2], then the interval width (α2 − α1) defines the amount of privacy at c% confidence level.

Attacks on Randomization*:*

The major technique used is a PCA analysis to choose the exact attributes to quantify and anonymize them. Using the correlation analysis if a pattern of data is identified it becomes easier to remove noise and get the actual information. The randomization approach can be used to preserve privacy in stream of data where there is no relation between the data.

Known Input-Output Attack: The attacker knows some linearly independent collection of records, and their corresponding perturbed version. In such cases, linear algebra techniques can be used to reverse- engineer the nature of the privacy preserving transformation.

Known Sample Attack**:** The attacker has a collection of independent data samples from the same distribution from which the original data was drawn. In such cases, principal component analysis techniques can be used in order to reconstruct the behavior of the original data.

B. k-anonymity model and l-diversity:

K-anonymity model is used to hide the information using generalization and suppression. It involves the process of identifying Quasi identifiers and reduces the level of granularity of data representation. Its named K-anonymity to show that the granularity level is reduced in such a way that each record maps into atleast K records in the data. The l-diversity model was designed to handle some weaknesses in the k-anonymity model since protecting identities to the level of k-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group.

In the method of generalization, the attribute values are generalized to a range in order to reduce the granularity of representation. For example, the address of an individual could be generalized to a range such as regions and district etc. to reduce the risk of identification. In the method of suppression, the value of the attribute is discarded to reduce the risk of identification with the help of public dataset to reduce the accuracy of retrieval.

The approach starts with the identification of quasi-identifier attributes and discretize them using the concept hierarchy. Quantitative attributes are grouped as range of intervals and categorical attributes into sets. For each group is identified as a single item. A tree is created from starting of a root node which is null. Each node is an attribute group and

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

an index is created for each node. Append nodes into the tree using the index and arrange based on the type, level of concept hierarchy involved. The number of nodes will increase exponentially in relation with number of dimensions in the dataset.

Personalized Privacy-Preservation

Level of privacy will vary from the user to user. For example, organization has the high level of privacy than individuals. In individuals also high profile people have the high level of privacy than normal human. Since the level of granularity varies from user to user the value of k in anonymization techniques will vary based on the record.

Utility Based Privacy Preservation

The process of privacy-preservation leads to loss of information for data mining purposes. This loss of information can also be considered a loss of utility for data mining purposes. Since some negative results on the curse of dimensionality suggest that a lot of attributes may need to be suppressed in order to preserve anonymity, it is extremely important to do this carefully in order to preserve utility.

C. Cryptographic Methods for Information Sharing and Privacy:

In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end . For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores. This requires secure and cryptographic protocols for sharing the information across the different parties. The data may be distributed in two ways across different sites:

Horizontal Partitioning:

In this case, the different sites may have different sets of records containing the same attributes.

Vertical Partitioning:

In this case, the different sites may have different attributes of the same sets of records.

D. Query Auditing and Inference Control:

Many private databases are open to querying. This can compromise the security of the results, when the adversary can use different kinds of queries in order to undermine the security of the data. For example, a combination of range queries can be used in order to narrow down the possibilities for that record. Therefore, the results over multiple queries can be combined in order to uniquely identify a record, or at least reduce the uncertainty in identifying it. There are two primary methods for preventing this kind of attack:

l -diversity Method

The k-anonymity is an attractive technique because of the simplicity of the definition and the numerous algorithms available to perform the anonymization. Nevertheless the technique is susceptible to many kinds of attacks especially when background knowledge is available to the attacker. Some kinds of such attacks are as follows:
Homogeneity Attack: In this attack, all the values for a sensitive attribute within a group of k records are the same. Therefore, even though the data is k-anonymized, the value of the sensitive attribute for that group of k records can be predicted exactly.

Background Knowledge Attack: In this attack, the adversary can use an association between one or more quasi-identifier attributes with the sensitive attribute in order to narrow down possible values of the sensitive field further. Clearly, while k-anonymity is effective in preventing identification of a record, it may not always be effective in preventing inference of the

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

sensitive values of the attributes of that record. Therefore, the technique of l-diversity was proposed which not only maintains the minimum group size of k.

t-closeness model

The t-closeness model is a further enhancement on the concept of l-diversity. One characteristic of the l-diversity model is that it treats all values of a given attribute in a similar way irrespective of its distribution in the data. This is rarely the case for real data sets, since the attribute values may be very skewed. This may make it more difficult to create feasible l-diverse representations. Often, an adversary may use background knowledge of the global distribution in order to make inferences about sensitive values in the data. Furthermore, not all values of an attribute are equally sensitive. For example, an attribute corresponding to a disease may be more sensitive when the value is positive, rather than when it is negative. In , t-closeness model was proposed which uses the property that the distance between the distribution of the sensitive attribute within an anonymized group should not be different from the global distribution by more than a threshold t. The Earth Mover distance metric is used in order to quantify the distance between the two distributions. Furthermore, the t-closeness approach tends to be more effective than many other privacy preserving
data mining methods for the case of numeric attributes.

E.Distributed Privacy-Preserving Data Mining

The key goal in most distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. Thus, the participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which have the same set of attributes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacy preserving data mining.

*Semi-honest Adversaries***:** Participants Alice and Bob are curious and attempt to learn from the information received by them during the protocol, but do not deviate from the protocol themselves. In many situations, this may be considered a realistic model of adversarial behavior.
*Malicious Adversaries***:** Alice and Bob may vary from the protocol, and may send sophisticated inputs to one another to learn from the information received from each other.

F. Privacy-Preservation of Application Results

In many cases, the output of applications can be used by an adversary in order to make significant inferences about the behavior of the underlying data.In this section, we will discuss a number of miscellaneous methods for privacy preserving data mining which tend to preserve the privacy of the end results of applications such as association rule mining and query processing. This problem is related to that of disclosure control in statistical databases, though advances in data mining methods provide increasingly sophisticated methods for adversaries to make inferences about the behavior of the underlying data. In cases, where the commercial data needs to be shared, the association rules may represent sensitive information for target-marketing purposes, which needs to be protected from inference.

Association Rule Hiding -Recent years have seen tremendous advances in the ability to perform association rule mining effectively. Such rules often encode important target marketing information about a business. Some of the earliest work on the challenges. Two broad approaches are used for association rule hiding:
Distortion: The entry for a given transaction is modified to a different value. Since, we are typically dealing with binary transactional data sets, the entry value is flipped.

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Special Issue 3, July 2014

Blocking: Entry is not modified, but is left incomplete. Thus, unknown entry values are used to prevent discovery of association rules.

G. Limitations of Privacy: The Curse of Dimensionality

Many privacy-preserving data-mining methods are inherently limited by the curse of dimensionality in the presence of public information. For example, the technique in analyzes the k-anonymity method in the presence of increasing dimensionality. The curse of dimensionality becomes especially important. when adversaries may have considerable background information, as a result of which the boundary between pseudo-identifiers and sensitive attributes may become blurred. This is generally true, since adversaries may be familiar with the subject of interest and may have greater information about them than what is publicly available. This is also the motivation for techniques such as diversity in which background knowledge can be used to make further privacy attacks. Thus, the data loses its utility for the purpose of data mining algorithms. The broad intuition behind the result in is that when attributes are generalized into wide ranges, the combination of a large number of generalized attributes is so sparsely populated, that even two anonymity becomes increasingly unlikely.

Applications of Privacy-Preserving Data Mining

The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. Some of these applications such as those involving bio-terrorism and medical database mining may intersect in scope.

III. SUMMARY

In this paper, a survey of the broad areas of privacy-preserving data mining and the underlying algorithms is done. The broad areas of classification includes Privacy-preserving data publishing, Privacy-Preserving Applications, Utility Issues, Distributed Privacy, cryptography and adversarial collaboration are analyzed.

A variety of data modification techniques such as randomization and k-anonymity based techniques has been studied and analyzed based on their activities. A complete study is done on for distributed privacy-preserving mining, and the methods for handling horizontally and vertically partitioned data. issue of downgrading the effectiveness of data mining and data management applications such as association rule mining, classification, and query processing. The limitation of privacy preserving as the increase in the dimension also analyzed and application which can employ the privacy algorithm is also studied.

REFERENCES

[1] Adam N., Wortmann J. C.: "Security-Control Methods for Statistical Databases: A Comparison Study. ACM Computing Surveys", 21(4), 1989.
[2] Agrawal R., Srikant R. "Privacy-Preserving Data Mining". Proceedings of the ACM SIGMOD Conference, 2000.
[3] Jagannathan G.,Wright R., "Privacy-Preserving Distributed k-means clustering over arbitrarily partitioned data". ACM KDD Conference, 2005.
[4] Jagannathan G., Pillaipakkamnatt K.,Wright R., "A New Privacy-Preserving Distributed k-Clustering Algorithm". SIAM Conference on Data Mining, 2006.
[5] Kantarcioglu M., Clifton C., "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data". IEEE TKDE Journal, 16(9), 2004.
[6] Charu C. Aggarwal, Philip S. Yu "An Introduction to Privacy-Preserving Data Mining "
[7] Josep Domingo-Ferrer, , "ASurvey of Inference Contro lMethods for Privacy-PreservingDataMining"
[8] Agrawal R., Srikant R., ThomasD. "Privacy-PreservingOLAP". Proceedings of the ACM SIGMOD Conference, 2005.
[9] Agrawal D. Aggarwal C. C. "On the Design and Quantification of Privacy-Preserving Data Mining Algorithms". ACM PODS Conference, 2002.
[10] AggarwalC., Pei J.,ZhangB. "AFramework for Privacy Preservation against Adversarial Data Mining." ACM KDD Conference, 2006.
[11] Jiang W., Clifton C, "Privacy-preserving distributed k-Anonymity" . Proceedings of the IFIP 11.3 Working Conferences on Data and Applications Security, 2005.
[12] JohnsonW., Lindenstrauss J."Extensions of LipshitzMapping intoHilbert Space" , Contemporary Math. vol. 26, pp. 189-206, 1984.