Privacy Preservation Data Mining: A Survey

Mausumidey
Dept. Of Computer Science & Engineering,
MaharanaPratap College of Technology,
Gwalior,India

Dr. AnamikaAhirwar
Dept. Of Computer Science & Engineering,
MaharanaPratap College of Technology,
Gwalior,India

ABSTRACT

Data mining is beneath the attack of privacy promoters due to confusion regarding what it really is and a accurate concern related how it's normally done. This paper presents how techniques from the community of security can modify data mining for its betterment, allowing all its advantages as it still maintaining its privacy.Large Volumes of precise personal data is regularly gathered and observed by various kinds of applications by the use of data mining, analyzing those data is profitable to the users of the application. It is a significant asset to users of the application such as governments for taking effective decisions or business organizations. But analyzing those data enables treats to the privacy if properly not done. This task targeted to disclose the information by preventing sensitive data. Different methods consisting k-anonymity, randomization and data hiding have been proposed for the same.

Keywords

Privacy,Data Mining, Anonymization, Perturbation, Generalization, Cryptographic, Privacy Preservation Data Mining [PPDM]

1. INTRODUCTION

In today's age of information, the collection of data is universal, and each transaction is stored some-where. The final data sets can include data in tera bytes or peta bytes, so scalability and efficient mining algorithms.

Normally, ever-growing is the main consideration of many data, collection of data, along with the arrival of analysis tools that are able for controlling large volumes of the information, has direct to the concerns of privacy. Preventing private data is most significant concern for society many laws are now needed to externally get approved before the analysis of every individual's data, for e.g. - Its importance is not restricted to only individuals: business organizations might also require preventing the privacy of their information, also even though sharing it for the purpose of analysis could help the company. Apparently, the trade-off between the sharing of information for analysis and preserving it as secret to preserve the secrets of corporate trade and the customer's privacy is an increasing challenge.

The most common definition of the privacy in cryptographic community restricts the information that is disclosed by distributed computation to information which can be generated from the assigned result of the calculation. Although there are many variations of the definition for privacy, for the intention of this consideration it uses the definition which compares the outcome of the exact computation to that an "ideal" calculation: Suppose first a party which is included in the real computation of the function (for example. a data mining algorithm). Suppose there is also an "ideal scenario", in which additionally to the actual parties there is also a "trusted party" that does not depart from the nature that it defined for him, and does not try to cheat. In ideal scenario all the parties that send their inputs to trusted party, also then calculates the function and again sends the exact results to other parties. Roughly speaking, a protocol is always secure only if anything to which an adversary can understand in the real world it can also understand in the ideal world, called from its own input and also from output that from the trusted party it receives[7].

Data mining is a currently a raising domain, that are connecting the three worlds of Artificial Intelligence, Databases, and Statistics. The age of information that has allowed several organizations to collect huge volumes of data. Though, the efficiency of this data is negligible if "meaningful information" or "knowledge" cannot be generated from it[8]. Data mining, also called as knowledge discovery, try to solve this requirement. In addition with to the standard statistical approach, data mining method searches for attractinginformation without requiring previous hypotheses. As a domain, it has discovered a new algorithms and concepts like association rule learning. Also it has implemented well known machine-learning algorithms like inductive-rule learning (for example, by the decision trees) to setting in which very huge databases are included. Data mining methods are used in research and business and are now becoming more popular with the time.

A main issue which arises in any of the masses group of data is of confidentiality. The requirement for privacy is because of law (for example, for the medical databases) or it can be encouraged by interests of business. Though, there are circumstances in which the data sharingcan direct to common gain. A basic utility of huge databases currently is research, if it is economic or scientific and market oriented. Hence, for e.g. the medical area has so much to receive for research by pooling data; as it can even attempt businesses with common interests. In spite of the potential gain, this is frequently not possible because of the confidentiality problem that arises.

2. LITRATURE REVIEW

Privacy-preservation of the sensitive information in the data mining approach is a significant topic in knowledge discovery systems and data communication. As an example, assume few hospitals needs to get important combined knowledge regarding a particular diagnosis from records of their patients whereas every hospital is not allowed this, because of the privacy rules, to expose private data of individuals.

Hence they are required to run a joint and safe protocol on their shared database to achieve the appropriate information. Several safe protocols have been suggested for data mining methods and machine learning approach like[1] for clustering, for decision tree classification, for Neural Networks, for association rule mining, and for Bayesian Networks. The basic concern of these type algorithms is to prevent the privacy of user's private data during they achieve useful information from the complete dataset.

Existing work in the privacy-preserving data mining has been expressed two problems. In first, the target is to prevent privacy of the customer by perturbing the values of data[2]. In this approach random noise data is discover to alter susceptible values, and the sharing of random data is used to create a new method for distribution of data that is nearly related to original data distribution without disclosing values of original data. The calculated original data distribution is used to again construct the data, and data mining approaches, like association and classifiers rules are implemented to the set of reconstructed data. Afterward clarification of this method has strengthened the computation of the original values depend on distorted data.

There are various areas for further research. Handling multiple parties is the first area. This algorithm is restricted to the two parties. Expanding it to the multiple parties is a non-trivial, particularly if it is taken into account the collusion between the parties. Another area for the future research is non-categorical attributes and quantitative association rule mining. This task is restricted to the Boolean association rule mining. The non-categorical attributes importantly raise the complication of the issue. If the party of data mining cannot achieve exact outcomes due to of privacy limitations imposed by the contributors of data, it can be ready to spend for extra data.

It is also suggest that computing the volume of private information in the terms of its financial value, as a pattern of creative property. The expense of every portion of data should be finding out in a "fair" manner, so that to influence the participation of this portion in the whole benefit. The paper carries the concept of fairness by the theory of coalitional games: the Shapley value and the core.

PPDM and Bridging game theory can create the theoretical base for the area of private data, in which all the candidates get appropriate rectification of contribution for their business. On the contrary anatomy method implement on sensitive tables decreases loss of information, due to its releases all quasi-identifier and sensitive values straightly in two different tables without implementing any suppression technique or even any generalization directs to maintaining data utility.In the paper[10] proposed a heuristic-dependent architecture for the preserving privacy in mining is presented for the frequent item-sets. Hiding the set of frequent-patterns is the main purpose of this architecture, storing highly private information. A set of algorithms have been proposed here whose work is to eliminate the transactional information only from the database, within the statistical disclosure monitor space they are also known as non-perturbative algorithms, dissimilar to those types of algorithms, which changes the information of previously declared algorithm by entering noise in the data, also termed as perturbative algorithms.

In paper[11] the performance is observed in regards of the encryption and the decryption operation's number that is needed by the particular algorithm. The end two observations, that is the cost of computation and average number of the operations, cannot gives an absolute observation, but to perform rapid comparison among the various algorithm they may be considered. Data collected from the various sources of the data and the various systems of operation is created by using ETL tools.From the Level 1 this clean and transformed data is saved in data warehouse. And that stored data in the data-warehouse is used in the mining. And in level 2, to find the patterns and to discover the knowledge from past data, algorithms of data-mining are used. As the data-mining privacy preservation technology is used to prevent the data from illegal access so the private data of each individual may be protected from being got misused. Another method which is commonly used is the Condensation method which can make constrained-clusters in data-set and after which they generate pseudo-data[12]. The concept of this approach is to condense or contract data into the various groups of the pre-defined size. And for every group, some statistics should be managed. This method is used in updating the data dynamically like stream issues.

3. PPDM TECHNIQUE

PPDM stands for privacy preservation data mining techniques which are used to extract the accurate data from huge database and at the same time provide the security on that data

3.1 Anonymization

As k-anonymity prevents against the disclosure of identity, it does not allowenough preservation against disclosure of attribute. There are mainly two types of attacks: the background knowledge attack and the homogeneity attack.

3.2 Perturbation Approach

This approach performs under the requirement that the service of data is not provided to recover or learn precise records. These limitations are automatically leads to few issues. As the approach does not again construct the values of original data but only the distributions, a new algorithm is needed to develop that uses these reconstructed distributions to do mining of the fundamental data.

3.3 Randomized Response Techniques

The randomization approach is a technique for the privacy-preserving data mining in that noise is merged to the data to mask the values of attribute of the records. The noise merged is large enough such that the values of individual record cannot be recovered[6].

3.4 Generalization Approach

Generalization includes substituting a value along with a less definite but semantically constant value. Generalization can be reached via local recoding or global recoding[3, 5].

3.5 Cryptographic Technique

In this ideal architecture the whole process is divided into the three main components the mediator, customer and a group of data service providers. Previously there is no interaction between the customer and the data provider.

And when the client sends a query, the mediator forwards the information to all data holders and via exchange ofthe acknowledgements, the mediator generates the connection with the data providers[4].

There are various techniques suggested in the area of the Privacy Preserving Data Mining but one exceed over the other on the basis of different criteria. Algorithms are categorized based on utility, performance, cost, complexity, etc. It has been presented a tabular comparison in a chronological order. Table 1 represents different Merits and Demerits of the Various Techniques of PPDM.

Here are different Merits and Demerits of the Various Techniques mentioned below in Table 1.

TABLE I: Merits and Demerits of the Various Techniques

Techniques of PPDM	**Merits**	**Demerit**
ANONYMIZATION	This approach is used to prevent the identities of respondentduring releasing the reliable information. Whereas the k-anonymity prevents the disclosure of identity against, it does not allow enough protection against the disclosure of attribute.	There are two main attacks: the background knowledge attackand the homogeneity attack.Due to the restrictions of the k-anonymitymodel from these two perceptions. Firstly, itis very difficult for the database owner to find out which of the attributes are notpresent in external tables.The second restriction is that the k-anonymity model supposes a particular methodof attack, whereas in the real scenarios there is no such reason as why the attacker must not try otherapproaches.
PERTURBATION	Isolated treatment of the various attributes by perturbation method.	The approach does not again create the basicdata values, only the sharing and new types of algorithms must have been evolved that usesthese types of reconstructed distributions to work outmining of available data.
RANDOMIZED RESPONSE	The randomization approach is a technique that may be implementedeasily at the time of data collection. It has been presented to be very essential technique for covering eachdata individually in the privacy Preserving data mining. TheRandomization approach is very efficient.Though, it concluded in high loss of information.	Randomized Response approach is not applied for databases of multiple attribute.
GENERALIZATION	Generalized value is same in Global Recording. Information loss is less in Local Recording. It provides high level of Privacy in Global Recording.	And different in Local Recording.Information loss is more in Global Recording. Low level of privacy provided in Local Recording.
CRYPTOGRAPHIC	Cryptography provides a previously-explained scheme for privacy that consists of techniques for offering and measures it. A wide toolset is exists of cryptographic algorithmsand organize to apply algorithms of privacypreserving data mining.	This method is complicated to extent as when more than the little of the parties is involved. And also, it never expressed the questionof if the exposure of final result of data mining may gap the individual's records of privacy.

4. CONCLUSION

Currently, one of the major concerns is privacy to prevent the private data which they don't want to share. This paper targeted on the previous literatures in the domain of Privacy Preserving Data Mining. From this analysis, it has been found that there is no only one method which is constant in all fields. All techniques operate in a separate way based on the pattern of data also the kind of domain or application. Still from the analysis, it can be concluded that the Cryptography and Random Data Perturbation methods operate much better as compared with the other previous methods. For encryption of sensitive data, Cryptography is one of the best techniques. On the contrary, Data Perturbation will also provide help to protect data and therefore privacy is retained. Whereas all the introduced techniques are only comparative to this target of the privacy preservation, it is needed further to make perfect those methods or introduce few effective approaches.

REFERENCES

[1] Mohamed A.Ouda,Sameh A. Salem, Ihab A. Ali, and El-Sayed M.Saad "Privacy-Preserving Data Mining (PPDM) Method for Horizontally Partitioned Data", IJCSI International Journal of Computer Science Issues,ISSN (Online): 1694-0814, Vol. 9, Issue 5, No 1, September 2012.

[2] Li Liu , Kantarcioglu, M. , Thuraisingham, B. "Privacy Preserving Decision Tree Mining from Perturbed Data", System Sciences, HICSS '09. 42nd Hawaii International Conference, IEEE, ISBN:978-0-7695-3450-3, 2009.

[3] K. VenkataRamana and V.ValliKumari, "Graph Based Local Recoding For Data Anonymization", International Journal of Database Management Systems (IJDMS), ISSN: 0975-5705 (Online); 0975-5985(Print), Vol.5, No.4, August, 2013.

[4] Yashaswini, "A Survey on Privacy Preserving Data Mining", International Journal of Innovative Research in

Computer and Communication Engineering, ISSN(Online) : 2320-9801, Vol. 3, Special Issue 7, October 2015.

[5] ManolisTerrovitis _ Nikos Mamoulis _ PanosKalnis," Local and Global Recoding Methods for Anonymizing Set-valued Data", Research Center Mathematical and Computer Sciences and Engineering,KingAbdullah University of Science and Technology, ISSN:0949-877X,2011.

[6] Kiran P, S Sathish Kumar and DrKavya*"A Novel Framework using Elliptic Curve Cryptography for Extremely Secure Transmission in Distributed Privacy Preserving Data Mining",* An International Journal (ACIJ), ISSN 2229-726X, Vol.3, No.2,March 2012.

[7] Anand Sharma and vibhaojha "Privacy preserving Data Mining by Cryptography" in Springer-LNCS-CICS,1865 0929, "Recent Trends in Network Security and Applications". pp.576-581, Vol:89,2010.

[8] Malik, M.B, Ghazi, M.A. Ali, R., Computer and Communication Technology (ICCCT), "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", IEEE Third International Conference, ISBN**:** 978-1-4673-3149-4, 2012.

[9] M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in *proceedings of ICCCNTCoimbatore, India*, IEEE 2012.

[10] Kunwar Singh kushwah, AbhayPanwar, "A Privacy Preservation Technique Using Machine Learning Technique", International Journal of Engineering and Innovative Technology (IJEIT), ISSN: 2277-3754, Volume 4, Issue 8, February 2015.

[11] Mrs. S. Vasanthi, Ms. S. Nandhini, "Privacy Preserving using Association Rule in Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering, ISSN(Online): 2278-1323,Vol. 4, Issue 8, August 2015.

[12] Vishal RavindraRedekar, Dr. K.N.Honwadkar, "Privacy-Preserving Mining of Association Rules in Cloud", International Journal of Science and Research (IJSR), ISSN(Online): 2319-7064, Volume 3 Issue 11, November 2014.