Open access Journal International Journal of Emerging Trends in Science and Technology

A Review on Privacy Preserving Data Mining

Authors

Jyoti, Divya Rana

University school of information and communication technology, GGSIPU

Email: *jojo1392@gmail.com, divya2524@gmail.com*

ABSTRACT

Most data mining applications operate under the assumption that all the data is available at a single central repository, called a data warehouse. This poses a huge privacy problem because violating only a single repository's security exposes all the data. Although people might trust some entities with some of their data, they don't trust anyone with all their data. With the extensive amount of data stored in databases and other repositories it is very important to develop a powerful and effective mean for analysis and interpretation of such data for extracting the interesting and useful knowledge that could help in decision making. Data mining is such a technique which extracts the useful information from the large repositories. Knowledge discovery in database (KDD) is another name of data mining. Privacy preserving data mining techniques are introduced with the aim of extract the relevant knowledge from the large amount of data while protecting the sensible information at the same time. In this paper we review on the various privacy preserving data mining techniques like data modification and secure multiparty computation based on the different aspects. We also analyze the comparative study of all Techniques followed by the future research work

Keywords *Data Mining, KDD, Privacy Preservation.*

I. INTRODUCTION

In today's information age, data collection is ubiquitous, and every transaction is recorded somewhere. The resulting data sets can consist of terabytes of data, so efficiency and scalability is the primary consideration of most data mining algorithms. Naturally, ever-increasing data collection, along with the influx of analysis tools capable of handling huge volumes of information, has led to privacy concerns. Protecting private data is an important concern for society several laws now require explicit consent prior to analysis of an individual's data, for example but its importance is not limited to individuals: corporations might also need to protect their information's privacy, even though sharing it for analysis could benefit the company. Clearly, the trade-off between sharing information for analysis and keeping it secret to preserve corporate trade secrets and customer privacy is a growing challenge.

But is it even possible to perform large-scale data analysis without violating privacy? Given sufficient care, we believe the answer is yes. In this paper, we'll describe why data mining doesn't inherently threaten privacy; Many privacy preserving techniques are using some form of transformation to achieve privacy. Privacy preserving is mainly focused on data distortion, data reconstruction and data encryption technology. The implementation of PPDM techniques has become the demand of the moment. The goal of this paper is to present the review on privacy preserving techniques which is very helpful while mining process over large data sets with reasonable efficiency and preserve security.

According to [1] categorize data mining into five tasks:

- ***Exploratory data analysis (EDA)***. Typically interactive and visual, EDA techniques simply explore the data without any preconceived idea of what to look for.
- ***Descriptive modeling***. A descriptive model should completely describe the data examples include models for the data's overall probability distribution (density estimation), partitions of the *p* dimensional space into groups (cluster analysis and segmentation), and descriptions of the relationship between variables (dependency modeling).
- ***Predictive modeling: classification and regression***. The goal here is to build a model that can predict the value of a single variable based on the values of the other variables. In classification, the variable being predicted is categorical, whereas in regression, it's quantitative.
- ***Discovering patterns and rules***. Instead of building models, we can also look for patterns or rules. Association rules aim to find frequent associations among items or features, whereas outlier analysis or detection focuses on finding "outlying" records that differ significantly from the majority.
- ***Retrieval by content***. Given a pattern, we try to find similar patterns from the data set. The first three tasks output models that essentially summarize the data in various ways; the last two find specific patterns, but they're often generalized and don't reflect particular data items. Because these models generally don't contain individual data values, they don't present an Immediate threat to privacy. However, there's still the issue of inference. A "perfect" classifier, for example, would enable discovery of the target class, even if the individuals' target classes weren't directly disclosed. In practice, though, probabilistic inferences are more likely, giving the model's possessor a probabilistic estimate of private values. The more immediate privacy problem with data mining is based not on its results, but in the methods used to get those results.

II. Classification of Privacy Preserving Techniques

There are many approaches which have been adopted for privacy preserving data mining. We can classify them based on the following dimensions:

- Data distribution
- Data modification
- Data mining algorithm
- Data or rule hiding
- Privacy preservation

The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places. The second dimension refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include: • perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise), • blocking, which is the replacement of an existing attribute value with a "?", • aggregation or merging which is the combination of several values into a coarser category, • swapping that refers to interchanging values of individual records, and • sampling, which refers to releasing data for only a sample of a population. The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. We have included the problem of hiding data for a combination of data mining algorithms, into our future research agenda. For the time being, various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks. The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as “rule confusion”. The last dimension which is the most important refers to the privacy preservation technique used for the selective modification of the data. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized. The techniques that have been applied for this reason are:

- Heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values
- Cryptography-based techniques like secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the results, and
- Reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data. It is important to realize that data modification results in degradation of the database performance. In order to quantify the degradation of the data, we mainly use two metrics. The first one, measures the confidential data protection, while the second measures the loss of functionality.

III.REVIEW OF PRIVACY PRESERVING DATA MINING TECHNIQUES

In this section, we focus on the different PPDM techniques which are developed like data perturbation, blocking based, cryptographic techniques etc.

A. Data Perturbation Data

Perturbation [2][3] is a technique for modifying data using random process. This technique apparently distorts sensitive data values by changing them by adding, subtracting or any other mathematical formula. This technique can handle different data types: character type, Boolean type, classification type and integer. In discrete data [2], it is required to preprocess the original data set. The preprocessing of data is classified into attribute coding and obtaining sets coded data set. The method of average region to disperse the continuous data is used here. Discrete formula prescribed by Sativa Lohiya and Lata Ragha [8] is: A (max) - A (min)/n = length. A is continuous attribute, n is number of discrete, and length is the length of the discrete interval. The technique does not reconstruct the original data values, it only reconstructs the distribution. Data distortion or data noise are different names for data perturbation. It is very important and critical to secure the sensitive data and data perturbation plays an important role in preserving the sensitive data. Distortion is done by applying different methods such as adding noise, data transpose matrix, by adding unknown values etc[4]. In some perturbation approaches it is very difficult to preserve the original data. Some of these are distribution based techniques. In order to overcome this problem, new algorithm were developed which were able to reconstruct the distributions. This means that for every individual problem in classification, clustering, or association rule mining, a new distribution based data mining algorithm needs to be developed. In [5], develops a new distribution-based data mining algorithm for the classification problem, and Vaidya [6] and Rizvi [7] develop methods for privacy-preserving association rule mining. A new approach in data perturbation was introduced by Jahan, G.Narsimha and C.V Guru Rao [4].It was based on singular value decomposition (SVD) and scarified singular value distribution (SSVD) technique and having the feature of selection to reduce the feature space. In this method, different matrices have been introduced to compare or measure the difference between original dataset and distorted dataset. SSVD is efficient approach in keeping data utility, SVD also works better than other standard data distortion methods which add noise to the data to make it perturbed.

Each data dimension is reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations.

B. Blocking based technique

In blocking based technique [8][9], authors state that there is a sensitive classification rule which is used for hiding sensitive data from others. In this technique, there are two steps which are used for preserving privacy. First is to identify transactions of sensitive rule and second is to replace the known values to the unknown values (?). In this technique, there is scanning of original database and identifying the transactions supporting sensitive rule. And then for each transaction, algorithm replaces the sensitive data with unknown values. This technique is applicable to those applications in which one can save unknown values for some attributes. Authors in [8] want to hide the actual values, they replace ‘1’ by ’0’ or ‘0’ by ‘1’ or with any unknown (?) values in a specific transaction. The replacement of these values does not depend on any specific rule. The main aim of this technique is to preserve the sensitive data from unauthorized access. There may be different sensitive rules according to the requirements. For every sensitive rule, the scanning of original database is done. When the left side of the pair of rule is a subset of attribute values pair of the transaction and the right hand side of the rule should be same as the attribute class of the transaction then only transaction supports any rule. The algorithm replaces unknown values in the place of attribute for every transaction which supports that sensitive rule. These steps will continue till all the sensitive attributes are hidden by the unknown values.

C. Cryptographic Technique

Cryptography is a technique through which sensitive data can be encrypted. It is a good technique to preserve the data. In [10] , authors introduced cryptographic technique which is very popular because it provides security and safety of sensitive attributes . There are different algorithms of cryptography available .But this method has many disadvantages. It fails to protect the output of computation. It prevents privacy leakage of computation. This algorithm does not give fruitful results when it talks about more parties. It is very difficult to apply this algorithm for huge databases. Final-data mining result may break the privacy of individual’s record.

D. Condensation Approach

Another approach used is Condensation approach .It was introduced by Charu C. Aggarwal and Philip [11] which builds constrained clusters in the data set and after that produces pseudo-data. The basic concept of the method is to contract or condense the data into multiple groups of predefined size. For each group, certain statistics are maintained. This approach is used in dynamic data update such as stream problems. Each group has a size of at least ‘k’, which is referred to as the level of that privacy-preserving approach. The higher the level, the high is the amount of privacy. They use the statistics from each group in order to generate the corresponding pseudo-data. This is a simple privacy preservation approach but it is not efficient because it leads to loss of the information.

E. Hybrid technique

Privacy preservation is a very huge field. Many algorithms have been proposed in order to secure the data. Hybrid technique is a new technique through which one can combine two or more techniques to preserve the data. Sativa Lohiya and Lata Ragha [8] proposed a hybrid technique in which they used randomization and generalization. In this approach first they randomize the data and then generalized the modified or randomized data. This technique protects private data with better accuracy; also it can reconstruct original data and provide data with no information loss. Many other techniques can also be combined to make a hybrid technique such as Data perturbation, Blocking based method, Cryptographic technique, Condensation approach etc.

IV.COMPARISON BETWEEN TECHNIQUES

Data Mining Algorithms are classified on the basis of performance, utility, cost, complexity, tolerance against data mining algorithms etc. We have shown a tabular comparison in table1 of the work done by different authors in a chronological order.

S.No	Authors	Year of publication	Technique used for PPDM	Approach	Result and Accuracy
1.	Y.Lindell, B.Pinkas [10]	2000	Cryptographic Technique	A technique through which sensitive data can be encrypted. There is also a proper toolset for algorithms of cryptography	This approach is especially difficult to scale when more than a few parties are involved. Also it does not hold good for large databases
2.	J. Vaidya and C. Clifton[6]	2002	Association Rule	Distribution of data vertically into segments.	Distribution Based Association Rule Data Mining provides privacy.
3.	Hillol Kargupta, Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar[3]	2003	Data Perturbation	They tried to preserve data privacy by adding random noise, while making sure that the random noise still preserves the “signal” from the data so that the patterns can still be accurately estimated.	Randomization-based Techniques are used to generate random matrice
4.	CharuC.Aggarwa, Philip S. Yu[11]	2004	Condensation Approach	This approach works with pseudo-data rather than with modifications of original	The use of pseudo-data no longer necessitates the redesign of data mining

				data, this helps in better preservation of privacy than techniques which simply use modifications of the original data	algorithms, since they have the same format as the original data.
5.	A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam [12]	2006	L-Diversity Algorithm	If there are ‘l’ ‘well represented’ values for sensitive attribute then that class is said to have L-Diversity	It is better than KAnonymity in preserving Data mining.
6.	Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira[13]	**2010**	Anonymization	Anonymization is a technique for hiding individual’s sensitive data from owner’s record. K-anonymity is used for generalization and suppression for data hiding	Background Knowledge and Homogeneity attacks of K-Anonymity Algorithm do not preserve sensitivity of an individual.
7.	P.Deivanai, J. Jesu Vedha Nayahi and V.Kavitha[14]	**2011**	Hybrid Approach	Hybrid Approach is a combination of different techniques which combine to give an integrated result.	It uses Anonymization and suppression to preserve data.
8.	George Mathew, Zoran Obradovic[16]	**2011**	Decision Tree	An approach which is technical, methodological and should give judgmental knowledge.	A graph-based framework for preserving patient’s sensitive information.
9.	Jinfei Liu, Jun Luo and Joshua Zhexue Huang[8]	**2011**	Rating Based Privacy Preservation	A novel algorithm which overcomes the curse of dimensionality and provides privacy.	It is better than KAnonymity and LDiversity
10.	Khaled Alotaibi, V. J. Rayward-Smith, Wenjia Wang and Beatriz de la Iglesia[9]	2012	Multi-Dimensional Scaling	A non linear dimensionality reduction technique used to project data on lower dimensional space.	The application of nonmetric MDS transformation works efficiently and hence produces better results.
11.	Savita Lohiya and Lata Ragha[8]	**2012**	Hybrid Approach	A combination of K-Anonymity and Randomization.	It has a better accuracy and original data can b reconstructed.

V. CONCLUSION

In today's world, privacy is the major concern to protect the sensitive data. People are very much concerned about their sensitive information which they don't want to share. Our survey in this paper focuses on the existing literature present in the field of Privacy Preserving Data Mining. From our analysis, we have found that there is no single technique that is consistent in all domains. All methods perform in a different way depending on the type of data as well as the type of application or domain. But still from our analysis, we can conclude that Cryptography and Random Data Perturbation methods perform better than the other existing methods. Cryptography is best technique for encryption of sensitive data. On the other hand Data Perturbation will help to preserve data and hence sensitivity is maintained .In future, we want to propose a hybrid approach of these techniques

REFERENCES

1. David Hand,Heikki Mannila,Padhraic Smyth,"Principal of data Mining",A Bradford book,The MIT Press ,Cambridge Massachusetts,London ,England
2. J. Liu, J. Luo and J. Z. Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements", in proceedings of 11th IEEE International Conference on Data Mining Workshops, IEEE 2011
3. H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", in proceedings of the Third IEEE International Conference on Data Mining, IEEE 2003.
4. T. Jahan, G.Narsimha and C.V Guru Rao, "Data Perturbation and Features Selection in Preserving Privacy" in proceedings of 978- 1-4673-1989-8/12, IEEE 2012.
5. R. Agrawal and A. Srikant, " Privacy-preserving data mining", in proceedings of SIGMOD00, pp. 439-450
6. J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data", in The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA, July 2002, IEEE 2002.
7. Evfimievski, A.Srikant, R.Agrawal, and Gehrke , "Privacy preserving mining of association rules", in proceedings of KDD02, pp. 217-228
8. S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.
9. A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database" , in proceedings of International Symposium on Computer Science and Society, IEEE 2011
10. Y. Lindell, B.Pinkas, "Privacy preserving data mining", in proceedings of Journal of Cryptology, 5(3), 2000.
11. C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in proceedings of International Conference on Extending Database Technology (EDBT), pp. 183–199, 2004. 746
12. A. Machanavajjhala, J.Gehrke, D. Kifer and M. Venkitasubramaniam, "I-Diversity: Privacy Beyond kAnonymity", Proc. Int'l Con! Data Eng. (ICDE), p. 24, 2006.
13. Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, "Efficient Multi-Dimensional Suppression for K-Anonymity", in proceedings of IEEE Transactions on Knowledge and Data

Engineering, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE 2010.

14. P.Deivanai, J. Jesu Vedha Nayahi and V.Kavitha," A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data" in proceedings of International Conference on Recent Trends in Information Technology, IEEE 2011
15. K. Alotaibi, V. J. Rayward-Smith, W. Wang and Beatriz de la Iglesia, "Non-linear Dimensionality Reduction for PrivacyPreserving Data Classification" in proceedings of 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security,Risk and Trust, IEEE 2012
16. G. Mathew, Z. Obradovic," A Privacy-Preserving Framework for Distributed Clinical Decision Support", in proceedings of 978-1- 61284-852-5/11/$26.00 ©2011 IEEE.