Privacy Preserving Data Mining Techniques in a Distributed Environment

Mona Shah
Research Scholar, RK University
Rajkot, India-360020
JG College of Computer Applications, Ahmedabad,
India-380054.

Hiren D. Joshi, Ph.D
Associate Professor & Director (I/C)
School of Computer Science
Dr. BabaSaheb Ambedkar Open University
Ahmedabad, India-380060.

ABSTRACT

Data storing and retrieving has been important since decades in the world of information. It makes this process prolific, when the retrieved information becomes smartly meaningful. Data mining is this new flavor. In the recent years data mining is a wide spread and active area of research. Its meaningfulness has gained momentum due to its vast area of applications. One of the popular and potential sub-areas of data mining is preserving privacy while mining. Data mining tools bring a factor of threat to the data under study for subjects like medical history, banking/credit card details, judicial matters and a few more. In such sectors, the data can be sensitive and personal. Protecting such data becomes the key factor during the process of mining. Here is an attempt to study the techniques used to address the issue of privacy preserving data mining in a distributed database environment in the last decade.

General Terms

Privacy Preserving Data Mining

Keywords

Data mining, privacy preserving, distributed database, data security

1. INTRODUCTION

Data Mining, the process of extracting knowledge from huge sets of raw data, is a field of essence and interest for the researchers for more than two decades. It offers a strong approach to uncover the hidden patterns which can serve as a platform for predictions and forecasts. Evidently, this feature serves to give edge sharp competency in business and forming regulatory, still flourishing business strategies. This paper is organized into three main areas. 1. Introduction 2. Literature review. 3. An appraisal of these techniques

Sensitive and/or personal information carrier data protection is found to be equally important while mining. Such data can be individual information like bank account numbers, passport number, social security details (Adhar card in India), criminal history, medical health information, defense secrets, credit card numbers and so on to go. The privacy required is for the reasons viz. protecting customers, abiding by the law and terms, maintaining trust between business and customers, to prohibit misuse of personal information and to protect pooling of information by other competitive business units. To understand privacy preserving, one needs to realize how privacy can be violated and how it can be protected. Although all the features of mining might not be susceptible to privacy, a few of them might need robust protection. Also, the term privacy will vary for different applications. The degree of privacy may also be needed to be defined. A number of data mining techniques and mechanisms with privacy preserving have been proposed so far, and each of them has its own advantage.

In privacy preserving data mining, there are two major approaches. They are cryptographic techniques and other is data perturbation methods. Few other algorithms which are in this category are decision tree on randomized data and collaborative filtering.

One of the common approaches in case of distributed database is to collect all data in a central location and perform mining. The other approach is to build a model for mining at each location of the distributed database, run them at their place and collect the results in a common place. The latter has a gain of quickness with a compromise over accuracy while the former requires the tradeoff of time required in order to gain accuracy. So far, it has been persistently observed that the more is the amount of security preserved, the more one losses accuracy in results. Every data mining situation which requires privacy has a typical core of it which demands few additional features in general algorithms provided so far.

Distributed data mining fits sound into the portrait, where multiple parties are interested in mining identical data and who will benefit collectively with the results of mining. The sensitive information is protected from the multiple parties participating in mining. Such mining is seen as collaborative as well as cooperative environment.

2. LITERATURE REVIEW

The k-anonymity model by Sweeny [1] provides protection from k-1 parties. In other words, the results from this joint mining cannot be distinguished from at least k-1individuals. The most which can be concluded from this is that the facts belong to any one of the k parties, but which one is not known. In the case of a vertical partitioned data, an extension of K2 algorithm for Bayesian networks [2], Yang and Wright [3] proposes to mine the data at each local site. Encrypt the data of the site with k keys and send it to the next site and continue the process. Apriori algorithm by Vaidya and Clifton in [4] proposes a solution for vertically partitioned database. For each site participating in mining, the rule was applied at local site, and then encrypted with k keys. The result was sent to the next party and the process was repeated. The resulting intersecting set gives the result, which was never decrypted. An addition is one such approach is to add fake transactions, which would alter the mining result bare minimum. The role of miner and calculator was introduced in [5]. There are N parties. The role of a miner is to perform mining and that of a calculator is to do calculations. Neither the miner nor does the calculator have any part of the database. The role played by miner and calculator is same as the one in Apriori algorithm. This is applicable to both vertical and horizontal partitioned database. One stage of encryption is done at calculator level and it does not know which item is being checked. The miner sends the results to the participants. A variant of this method was done in [5], where the miner tells each participant to add pseudo random noise and send it to the calculator. For this, the miner sends the seed to the calculator, which is one of the N noises. For every created noise, the calculator subtracts it from each N database values. It then performs mining and sends the result to the miner. 1/N is the probability that the calculator will learn about the data with N parties coming together in the process.

The bloom filter mechanism by Siu Man Lu and Ling Qiu in [6] was applied on centralized database but works equally well for distributed databases. It maps every item to a binary vector with a bitwise OR operator. A secret key is inserted before converting the data into bloom filters. The data converted to such bloom filters is irreversible. One cannot get back the original data. So, the original data cannot be inferred. The presence of a particular pattern is checked by a bitwise AND operation. If the result is true then the pattern p is not contained in transaction T has a very low probability (false positive).On the other hand, if the result is false then the pattern p is certainly not contained in transaction T.

“Combine without owner” by Vladimir and Ahmed [7] is a technique that involves multiple parties participation with each encrypting the data. It starts with the first contributor A, who has a public encrypt key k for encryption. The encrypted data from party A goes to the next party B, who then mixes his rows with A, removing duplicates, shuffles it and then sends it to party C. With party C, the process is repeated. After the last participant, the data comes back to the original sender party A. Party A decrypts all transactions of which its own transactions are identifiable. Then it publishes the union of all data to the parties.

The method in [8] uses sample selection and matrix decomposition method both. It is called as SS-SVD method. (Sample Selection-Singular Value Decomposition).A sample selection is performed using WCNN (Weighted Condensed Nearest Neighbor) method and preserves only selected samples. Only selected samples are preserved and they are replaced till it equal to the actual selected size. The singular value decomposition (SVD) is used to disturb the original data. Naïve Bayesian classifier is used to generate transition probability matrix in [9]. The entire data is divided into i lines and j columns, where is the number of lines/records and j is the number of attributes. With the help of a transition probability j number matrices, each of size i*i are generated. In each of these matrices, the positions of highest values are ordered and the corresponding matrix of original value is arranged. The accuracy rate of this is calculated using conditional probability.

SDQ (Secure Group Differential Private Query) mentioned in [10] is the work which is a combination of differential privacy and secure multiparty computation (SMC). This work tries to minimize the leak happening during the computations of mining and making intermediate results available to everyone. The results have shown that intermediate results can also reveal the actual data or hint towards it.

In [11], the work of Shamir [12] has been used and extended. An algorithm for privacy preserving distributed decision tree has been given which is based on Shamir's secret and ID3 algorithm. They have given the solution to implement ID3 algorithm for constructing decision trees over an arbitrary number of distributed resources in a privacy preserving manner. Lindell and pinkas [13] have proposed ID3 algorithm for two parties for privacy preserving data mining. This idea is extended to multiples parties, although one of such idea has already been given by [14]. Shamir's secret sharing method is used. A dealer D distributes a secret value v among n peers. To reconstruct the secret, any k (k<<n) peers is required. Here, they have shown, how Shamir's work can be used to privately compute the sum of secret values without letting others know about it.

They have used three phase for summing up the data viz. Distribution, intermediate and final computation phase. In the distribution phase, n parties decide on the degree k of a polynomial. During the intermediate computation phase, each party adds up all the shares it received from other parties and then sends this intermediate result to all other parties. In the final computation phase, each party P_i can compute the sum of secret values using the intermediate results it received during the previous phase.

In [15], a data mining privacy algorithm by decomposition (DMPD), which has been taken up classification accuracy and k-anonymity constraint as background.

While in [16], Xun and Zhang propose, Naïve Bayes privacy preserving classifier for horizontally partitioned distributed data. Distributed data mining can be classed into two categories [17]. The first is server-to-server where data are distributed across several servers. The second is client to-server where data reside on each client while a server or a data miner performs mining tasks on the aggregate data from the clients. This multi-party protocol is built on the semi-trusted mixer model, in which each data site sends messages to two semi-trusted mixers, respectively, which run the proposed two-party protocol and then broadcast the classification result.

In [18], k-TTP (trusted third party) attempt has been made, where k-privacy is defined as the privacy attained when no participant learns statistics of a group of no less than k participants. This method does not require all-to-all communications and hence is suitable for real world application involving large number of systems in a distributed environment.

Keivan and Negar [19] make a different approach to address the solution to PPDM (Privacy Preserving Data Mining) in a distributed environment. They consider vertically partitioned data. They apply rankings to the attributes on each local site using support vector machine (SVM) algorithm and transferring the local rankings to a central site. At central site, they are merged while managing privacy. Jaideep and Murat in [20] have given Naïve bayes classifier for vertically and horizontally partitioned data with different aspects like communication and computation cost, for nominal and numerical attributes, effect of collusion and added that security always comes with a cost.

3. APPRAISAL OF THE AVAILABLE TECHNIQUES

Different approaches for privacy preserving data mining under different parameters come with their own advantages. The more the level of security, more is the computation cost. Different solutions when evaluated on different factors like amount of privacy attained, complexity, scalability and overall performance, the user might be provided with the aid to choose among the options based on the application area and the number of users participating in the process.

A concise expression of all the work is considered in a tabular format in Table I.

Sr. No.	Paper Title	Technique	Advantage	Findings
1.	L. Sweeney, "K-anonymity: a model for protecting privacy"	K-Anonymity protection	It can prevent the attacker from linking the information to map to the actual data holder. It confuses the attacker by the resulting linkages of a single data to ambiguous multiple data holders.	Focuses on person specific data and the property to be protected is the identity of the subjects.
2.	Zhiqiang Yang and Rebecca N. Wright, Member IEEE, "Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data"	Secure distributed computation with Bayesian network	Parties do not learn about individual data.	Applicable to only two parties and not suggested for large databases. Participating agency learn about the resulting network along with the parameters. They learn about the intermediate results also.
3.	J.Vaidya, C.Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data"	Secure computation of scalar product	Serves as efficient secure solution in distributed environment. Preventing disclosure of values at every site or the possibility to find a specific value.	A two-party algorithm for efficiently discovering frequent item-sets with minimum support levels, without either site revealing individual transaction values. Applicable to only two parties. Limited to Boolean association rule mining. The relevant parties know that which item-set is currently computed[5]
4.	Alex Gurevich, Ehud Gudes, "Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation"	Concept of two new entities: Miner and Calculator with each having different role and none of them have data.	Available for vertical and horizontal partition both.	The results are known to each party. There is a very littlie prob. of exposure if the local support happens to be same as support threshold.
5.	Siu Man Lu, Ling Qiu, "Individual Privacy and Organizational Privacy in Business Analytics"	Concept of bloom filter, where input data is converted into a stream of binary bits.	Computationally efficient and irreversible coding scheme. Also, gives false positive and false negative rates.	It has a tradeoff of precision versus storage requirements to some extent.
6.	Vladimir Estivill-Castro Ahmed HajYasien, "Fast Private Association Rule Mining by A Protocol for Securely Sharing Distributed Data"	Combine without owner technique. Introduces the concept of one party playing the role of protocol driver.	Works on horizontally partitioned data for 3 or more parties. A third outside party is not required.	The party that can be trusted with the role of shuffling the data and publishing to all other parties is matter to ponder over.
7.	Guang Li and Yadong Wang, "Privacy-Preserving Data Mining Based on Sample Selection and Singular Value Decomposition"	Matrix decomposition based data perturbation method involving sample selection for PPDM called as sample selection singular value decomposition (SS-SVD)	Above attribute extraction in matrix decomposition method, when sample selection is applied, it gives more accurate results. Better results than Non–negative matrix factorization (NMF) and singular value decomposition method (SVD) based on similar lines of matrix decomposition.	Deletes the repeated samples and does not use incomplete samples.

Sr. No.	Paper Title	Technique	Advantage	Findings
8.	Xing Yang, Yubao Liu, Zhan Li, and Jiajie Mo, "Privacy Preserving Naïve Bayesian Classifier Based on Transition Probability Matrix"	Proposes a method called Naïve Bayesian Classifier based on Transition probability matrix (NBCTPM)	Can be used for processing non-char data. Has efficiency parallel to that of a Naïve Bayesian classifier (NBC). Holds accuracy along with privacy as compared to NBC.	Applicable for classification problems and can be considered for increasing the precision.
9.	Ning Zhang, Ming Li, Wenjing Lou, "Distributed Data Mining with Differential Privacy",	Uses Secure Group Differential Private Query (SDQ), a combination of secure multiparty computation and differential privacy.	Attempts to limit information leak which happens during release of all intermediate results. The SDQ method can achieve stronger efficiency than available secure multi party computation (SMC) approaches and better accuracy over existing differential privacy solutions.	Gives more effective results when the large number of parties are participating and the data is sparsely located
10.	F. Emekci* , O.D. Sahin, D. Agrawal, A. El Abbadi, "Privacy preserving decision tree learning over multiple parties"	Using ID3 algorithm over multiple parties- a scalable secured distributed ID3 for building a decision tree.	The contributing parties cannot learn the secret value of another party even if they exchange their shares with each other. Suitable for large number of participations	Assumes that all participants do computation honestly in distribution and intermediate phase. Might be computationally intensive.
11.	Yehuda Lindell, Benny Pinkas, "Privacy preserving data mining"	Constructing a tree applying ID3 algorithm recursively for two parties.	Require few rounds of computation and lesser bandwidth as compared to generic solutions in the same area.	Assumes that each attribute is categorical. Depends on how best predicting attribute is chosen.
12.	Benny Pinkas, "Cryptographic techniques for privacy-preserving data mining"	Demonstrates the generic construction for a two party and multi party secure multiparty computation (SMC) method.	Comparison on trust, efficiency and communication criteria available	Easier to implement on two party than on multi party.
13.	Nissim Matatov, Lior Rokach , Oded Maimon, "Privacy-preserving data mining: A feature set partitioning approach"	K-anonymity is implemented using a different method referred as Data Mining Privacy by decomposition (DMPD). The dataset is divided into such a way that they follow k-anonymity. Also, after rejoining those tables, it still follows k-anonymity. For each projection, a classifier is trained. An unlabelled instance is classified by combining the classification of all classifiers.	Better performance in relation to existing k-anonymity based solutions.	A tradeoff between the computation cost and the number of increasing partitions in larger data sets is observed.

Sr. No.	Paper Title	Technique	Advantage	Findings
14.	X. Yi, Y. Zhang, "Privacy-preserving naive Bayes classification on distributed data via semi-trusted mixers"	Privacy preserving Naïve Bayes classifier for horizontally partitioned data for two party and multi-party. Among all the parties in the mining process, to parties are semi trusted and they work as mixer, to whom all data is sent. They run the two party protocol and broadcast the classification result.	More efficient in communication mode and secure in terms of privacy for multi-party protocol as compared to the one given in [21].	Assumption is that the two parties working as trusted mixers do not conclude.
15.	N. Zhang, S. Wang, W. Zhao," A new scheme on privacy-preserving data classification"	Algebraic technique based approach. Introducing a component called perturbation guidance (PG). The data miner sends the data perturbation guidance to the data provider, which in turn uses it and provides perturbed data to the miner.	Can build classifiers more accurately with less information leakage as compared to the existing randomization techniques. It can be integrated as middleware with the existing systems. Also builds an upper bound of the error introduced to the predictive accuracy of the classifier built, which makes the prediction more accurate.	Assumes there are two class label attributes with the values 0 and 1 and the data set is categorical.
16.	Bobi Gilburd, Assaf Schuster, and Ran Wolff, "k-TTP: A New Privacy Model for Large-Scale Distributed Environments"	K-TTP privacy, a generalization of the trusted third party (TTP) model.	More flexible than TTP. Therefore it is suitable for large distributed systems.	Assumes that the database under consideration is updated over time and no transactions are deleted.
17.	Keivan Kianmehr ,Negar Koochakzadeh, "Privacy-Preserving Ranking over Vertically Partitioned Data"	Provides method for ranking problem based on SVM method for a situation where different sites contain different attributes for the same set of entities.	Each participating site learns only ranking model of entities without knowing the attributes in other sites and finally the global ranking model will be built.	It can be extended to data from different platform domains.
18.	Jaideep Vaidya, Murat Kantarcıoglu, Chris Clifton, "Privacy-preserving Naïve Bayes classification"	Extension of Naïve Bayes Classifier	Presents Naïve Bayes Classifier for horizontally and vertically partitioned data.	Provides mining models equivalent to that of a situation where the data would have been integrated.
19.	M. Kantarcioglu, J. Vaidya, "Privacy preserving naive Bayes classifier for horizontally partitioned data"	Using Naïve Bayes classifier for distributed data using secure sum and logarithm.	Algorithm available for stringent privacy requirement.	It allows partial visibility of data to few parties, but no one can have knowledge of all data. This feature gives exact result. A tradeoff is observed between security achieved and the computation cost.

4. CONCLUSION

Here, a number of solutions provided for privacy preserving in a distributed environment have been considered. More and more contributions work towards achieving superlative efficiency and accuracy in the existing solutions with least disclosures and premium scalability. New solutions in a similar backdrop work from two party to multi-party implementations. Privacy preserving data mining is an attention seeking area among the research community and still has more to explore and offer for practical solutions. The spectrum of its applications has a fair chance to broaden with the global scenario.

5. REFERENCES

[1] Sweeney L. 2002 K-anonymity: A model for protecting Journal on Uncertainty, fuzziness and Knowledge based systems.

[2] Cooper g. and Herskovits E. 1992 A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, vol. 9, no. 4, pp. 309-347

[3] Yang Z. and Wright R. N. Member IEEE 2006 Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 9

[4] Vaidya J. and Clifton C. 2002 Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In Proceedings of SIGKDD 2002, Edmonton, Alberta, Canada.

[5] Gurevich A. and Gudes E. 2006 Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation. 10th International Database Engineering and Applications Symposium (IDEAS'06), IEEE.

[6] Siu Man Lu and Qiu L. 2007 Individual Privacy and Organizational Privacy in Business Analytics Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07), IEEE

[7] Vladimir Estivill-Castro Ahmed HajYasien 2007 Fast Private Association Rule Mining by A Protocol for Securely Sharing Distributed Data" IEEE.

[8] Guang Li and Yadong W. 2011 Privacy-Preserving Data Mining Based on Sample Selection and Singular Value Decomposition, International Conference on Internet Computing and Information Services.

[9] Yang X., Liu Y., Zhan L, and Jiajie M. 2011 Privacy Preserving Naïve Bayesian Classifier Based on Transition Probability Matrix. Seventh International Conference on Computational Intelligence and Security.

[10] Zhang N., Ming L. and Wenjing L. 2011 Distributed Data Mining with Differential Privacy. IEEE.

[11] Emekci F. ,Sahin O. D., Agrawal, A. and El Abbadi, 2007 Privacy preserving decision tree learning over multiple parties

[12] Shamir A. 1979 How to share a secret. Communications of ACM.

[13] Lindell Y. and Pinkas B. 2002 Privacy preserving data mining. Journal of Cryptology 15 (3) (2002) 177–206.

[14] Pinkas B. 2003 Cryptographic techniques for privacy-preserving data mining. SIGKDD Explorations.

[15] Matatov N., Rokach L. and Maimon O. 2010 Privacy-preserving data mining: A feature set partitioning approach.

[16] Yi X. and Zhang Y. 2008 Privacy-preserving naive Bayes classification on distributed data via semi-trusted mixers. Information Systems

[17] Zhang N, Wang S. and Zhao W. 2005 A new scheme on privacy-preserving data classification. In Proceedings of KDD'05.

[18] Gilburd B., Schuster A. and Wolff R. 2004 k-TTP: A New Privacy Model for Large-Scale Distributed Environments. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), Seattle, WA, USA.

[19] Kianmehr K. and Koochakzadeh N. 2012 Privacy-Preserving Ranking over Vertically Partitioned Data. PAIS 2012, Berlin, Germany.

[20] Vaidya J., Kantarcıoglu M. and Clifton C. 2008 Privacy-preserving Naïve Bayes classification. The VLDB Journal.

Kantarcioglu M. and Vaidya, J. 2003 Privacy preserving naive Bayes classifier for horizontally partitioned data. in: Proceedings of IEEE Workshop on Privacy Preserving Data Mining.