A Review on Various Privacy Preserving Techniques & Classifications Algorithms

Ankita Jain¹, Shubha Dubey², Anurag Jain³

Abstract—Privacy preserving data mining is one of the most demanding research areas within the data mining community. In many cases, multiple parties may wish to share aggregate private data without disclosing any private information at user side. Over the last few years this has naturally lead to a growing interest in security or privacy issues in data mining. More precisely, it became clear that discovering knowledge through a combination of different databases raises important security issues. New dimension of structure Trust (MLT) poses new challenges for perturbation-based PPDM. In distinction to the single-level trust situation wherever just one rattled copy is released, currently multiple otherwise rattled copies of the same knowledge are offered to knowledge miners at completely different sure levels. The a lot of sure an information manual labourer is, the less rattled copy it will access; it’s going to even have access to the rattled copies offered at lower trust levels. In this paper we are presenting some techniques to overcome problems related with privacy preservation and multi-level trust.

Keywords—Privacy Preservation Data Mining, Multi-Level Trust, PPDM, Perturbation.

I. INTRODUCTION

The multilevel trust (MLT) facing the new problems for privacy preservation that is based on perturbation PPDM. While talking about scenario of single level trust where only single perturbed copy is released, but now multiple differently perturbed copies of the similar data are available to data miners at different trusted levels. As far as most trusted data miner is concerned the less perturbed copy it can access; it may also have access to the perturbed copies available at lower trust levels. Likewise a data miner could access multiple perturbed copies through various other means [1].

Data Mining

Data mining is a recent emerging field with the supports of database, statistics and artificial intelligence. This helped to many organizations to gather huge amount of useful data or information. But generally gathering of useful data is quite difficult; hence knowledge extraction is also a big challenge in this respect.

Data mining also known as knowledge discover, it means useful information can be retrieved by existing stored data.

This can also introduced new concepts and algorithms such as association rule learning. It is also best suited for known machine-learning algorithms like inductive-rule learning to the setting where very large databases are involved [2].

Confidentiality is a common problem in data mining. The privacy is the great need of organization just because of sensible data storage as well as sharing. This sharing may lead mutual gain. A key utility of large databases today is research, whether it will be scientific or economic and market oriented. Increasing privacy and security consciousness has lead to increased research and development of methods that compute useful information in a secure manner. Companies could exchange information to boost productivity, but are prevented by fear of being exploited by competitors or antitrust concerns [3].

1.1 Privacy Preservation of Data Streams

A new topic within the space of privacy preserving data processing is that of information streams, within which knowledge grows speedily at a limitless rate. In such cases, the matter of privacy-preservation is kind of difficult since the info is being free incrementally [4].
1.2 Personalized Privacy Preservation

In many applications, different subjects have different requirements for privacy. For example, a brokerage customer with a very large account would likely have a much higher level of privacy-protection than a customer with a lower level of privacy protection. In such case, it’s necessary to individualize the privacy protection rule. In customized privacy preservation, we have a tendency to construct anonymization of the information specified totally different records have a unique level of privacy. The method uses condensation approach for personalized anonymization, while the method in [5] uses a more conventional generalization approach for anonymization.

1.3 Multi-Provider Outsourcing

An efficient DBMS system provides secure and reliable data storage and query execution without showing content of database. Presently distributed DBMS were used to maintain reliability and easy data access. Partitioning of data is performed in such a fashion as to ensure that the exposure of the contents of any one service provider does not result in a violation of privacy. The client executes queries by transmitting appropriate sub-queries to each service provider, and then piecing together the results at the client side. A better alternative is to deploy a distributed multi-party data sharing scheme among the service providers for them to share data while maintaining the data privacy at each site and answer the client’s queries directly [6].

1.4 Utility Based Privacy-Preserving Data Mining

Most privacy-preserving data mining methods apply a transformation which reduces the effectiveness of the underlying data when it is applied to data mining methods or algorithms. In fact, there is a natural tradeoff between privacy and accuracy, though this tradeoff is affected by the particular algorithm which is used for privacy-preservation. A key issue is to maintain maximum utility of the data without compromising the underlying privacy constraints. The issue of designing utility based algorithms to work effectively with certain kinds of data mining problems is addressed [7].

1.5 Cryptographic Methods for info Sharing and privacy

In several cases, multiple parties might need to share mixture personal knowledge, while not unseaworthy any sensitive info at their finish [8]. As an example, totally different superstores with sensitive sales knowledge might need to coordinate among themselves in knowing mixture trends while not unseaworthy the trends of their individual stores.

This needs secure and cryptanalytic protocols for sharing the data across the various parties. The info is also distributed in two ways that across totally different sites:

**Horizontal Partitioning:** during this case, completely different sites might have different sets of records containing identical attributes.

**Vertical Partitioning:** during this case, the various sites might have different attributes of similar sets of records. Clearly, the challenges for the horizontal and vertical partitioning case are quite different.

1.6 Privacy-Preserving Data Publishing

These techniques tend to study different transformation methods associated with privacy [9]. Other related problems include that of determining privacy-preserving methods to keep the underlying data useful and how they compare in terms of effectiveness in different scenarios.

Using additive perturbation based PPDM approach for multilevel trust is used for providing better flexibility and security. MLT-PPDM allows data owners to generate differently perturbed copies of its data for different trust levels. So, the data miners will have no diversity gain in their joint reconstruction of the original data. The scope of perturbation-based PPDM is extended to Multi-Level Trust (MLT-PPDM). Additionally MLT-PPDM considers only linear attacks but more powerful adversaries apply nonlinear techniques to derive original data and recover more information [10].

Privacy is turning into Associate in nursing progressively necessary issue in several data processing applications. A malicious knowledge mineworker might have access to otherwise flustered copies of constant knowledge through varied suggests that, and will mix these various copies to put together infer further data regarding the first knowledge that the information owner doesn't shall unleash. The rest of paper is organized as follows. In Section II explains about back ground of privacy preservation. In Section III explains about related work of privacy preservation data mining. Section IV describes conclusion.

II. BACKGROUND

Data mining (knowledge discovery from data) is defined as the non-trivial mining of implicit, earlier unknown, and potentially valuable information from large data sets or databases. Advances in hardware technology have increased the capability to store and record personal data about consumers and persons. Personal data may be used for a variety of intrusive or malicious purposes.
Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. Data integration and sharing have been a long-standing challenge for the database community. This need has become critical in numerous contexts, including integrating data on the Web and at enterprises, building ecommerce market places, sharing data for scientific research, data exchange at government agencies, monitoring health crises, and improving homeland security.

III. RELATED WORK

This section describes some related work to multi-level trust and privacy preservation data mining.

In 2012 Yaping Li et al. [1] presented Enabling Multilevel Trust in Privacy Preserving Data Mining. Privacy conserving data processing (PPDM) addresses the matter of developing correct models concerning mass knowledge while not access to specific data in individual knowledge record. A wide studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before knowledge area unit printed. The primary problem lies in preventing the information miners from combining copies at completely different trust levels to collectively reconstruct the initial data a lot of correct than what’s allowed by the information owner. All assumption and expand the scope of perturbation-based PPDM to construction Trust (MLT-PPDM) and also the additional trusty an information jack is, the less rattled copy of the info it will access. Preventing such diversity attacks is that the key challenge of providing MLT-PPDM services. Here address this challenge by properly correlating perturbation across copies at totally different trust levels and prove that this resolution is powerful against diversity attacks with regard to privacy goal. That is, for information miners World Health Organization have access to AN impulsive assortment of the rattled copies, this resolution stop them from conjointly reconstructing the first information additional accurately than the most effective effort exploitation a person copy within the assortment. This resolution permits a information owner to come up with rattled copies of its data for impulsive trust levels on demand. This feature offers information house owners most flexibility [1,2].

Hongwei Tian, Weining Zhang, Shouhuai Xu and Patrick Sharkey has proposed a new way of preserving data using knowledge based model for the sharing of data [3]. Here in this paper anonymized knowledge is used and on the basis of which a model is implemented.

The model is based on the generation of decision tree and pruning of the algorithms. The proposed methodology is beneficial for single and multiple source environments.

B.Swapna, R.VijayaPrakash has given privacy over data mining using operations without the quality data interruptions [5]. Here in this paper the technique implemented ensures that the privacy preservation is maintained while there should not be any lose of data. The technique also performs sanitization of the data while performing privacy.

V. Thavavel and S. Sivakumar has given privacy over distributed environment and creates a generalized framework for the unstructured data [5]. The technique is feasible for multi text documents and framework is generalized to make the system more effective.

Chris et al [6] presented potential research directions and challenges that need to be addressed in order to achieve privacy-preserving data integration. They also try to show some plausible solution ideas and also try to identify potential research directions and challenges that need to be addressed to perform privacy-preserving data integration. Increasing privacy and security consciousness has lead to increased research of methods that compute useful information in a secure manner. Privacy-preserving data mining deals with gaining knowledge after integration problems are solved.

Charu Aggarwal provided a first comprehensive treatment of the randomization approach in the presence of public information. They used this framework to illustrate a number of insights of the randomization method. They also showed the degrading effect of the dimensionality curse, and quantify the required perturbation level as a function of the dimensionality. Their analysis showed that the inclusion of public information in the analysis makes the goal of privacy preservation more elusive than previously thought for the randomization method [7].

Bayardo and Agrawal proposed a practical method for determining an optimal k-anonymization of a dataset. An optimal anonymization is one which perturbs the input dataset as little as is necessary to achieve k-anonymity, where “as little as is necessary” is typically quantified by a given cost metric. The ability to compute optimal anonymization lets us more definitively investigate the impacts of various coding techniques and problem variations on anonymization quality. It also allowed better quantifying the effectiveness of stochastic or other non-optimal methods. As per their result obtained offered method was feasible. They also demonstrated that despite the problems inherent hardness, provably optimal k-anonymizations can be obtained for real census data [8].
Li et al [9] described about outsourcing of data aggregation service. They also presented some decentralized peer-to-peer protocols for supporting data sharing across multiple private databases while minimizing the data disclosure among individual parties. They consider a number of important multi-party distributed outsourcing scenarios to support their protocols. They formalized the design goals and the notion of data confidentiality in terms of information revealed and offered a data privacy metric. A set of novel probabilistic computation protocols for important primitive operations were also described [9].

A new privacy preserving k-means clustering algorithm was proposed without compromising the security of individual. This algorithm uses random permutation due to which the communication overhead gets reduced [10].

Aggarwal et al [11] recommended a new method for anonymizing data records. To ensure privacy of the data records, they imposed the constraint that each cluster must contain no less than a pre-specified number of data records. They also provided constant-factor approximation algorithms to come up with such a clustering. The Multi-Level Trust in Privacy-Preserving Data Mining when integrated with partial information hiding methodologies help to find the right balance between maximum analysis results and keep the inferences that disclose private information about organizations or individuals at a minimum. Thus random rotation based data perturbation and K-anonymity are incorporated with MLT-PPDM to significantly enhance the data accuracy and to prevent the leakage of the sensitive data [11].

In 2008 Benjamin Fung et al [14] proposed an economical anonymization algorithmic rule to thwart the attacks within the model of continuous knowledge commercial enterprise. The majority thought of a single static unleash shield the information up to initial the primary unleash or first recipient. In sensible applications, knowledge is revealed endlessly as new knowledge arrives; a similar knowledge is also anonymized differently for a different purpose or a special recipient. In such situations, even once all releases square measure properly k anonymized, the obscurity of a private is also accidentally compromised if recipient cross-examines all the releases received or colludes with alternative recipients. Preventing such an attack are known as correspondence attacks, faces major challenges Associate in publishing formalized notion of attacks and presented a detection methodology and an anonymization algorithmic rule to forestall such attacks. Finally, they try to showed that each the detection and also the anonymization strategies square measure long to manage multiple releases and alternative privacy needs [14].

Kun Liu et al [15] suggested Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. Specifically, they explore the possibility of using multiplicative random projection matrices for constructing a new representation of the data. The transformed data is released to the data miner. It can be proved that the inner product and Euclidean distance are preserved in the new data. They assumed that the private data is from the same continuous real domain and all the parties are semi honest that means there is no collusion between parties and all the parties follow the protocol properly. Without loss of generality, they demonstrate this technique in a two-party-input scenario where Alice and Bob, each owning a private database, want a third party to analyze their data without seeing the raw information. This technique can be easily modified and applied to other input cases [15].

Guan Wang et al [16] proposed Inference Analysis in Privacy-Preserving Data Re-publishing. The goal of inference analysis in PPDR is to find out potential privacy breaches from all the published data sets. From statistical perspective, inference analysis basically tries to identify if the inference channels among the published data sets can reduce the uncertainty of certain individuals’ SA values to a level that can lead to privacy breaches. This method allows data to be arbitrarily updated. More importantly, our method allows data publishers to use any arbitrary combination of the two popular data disguises methods, bucketization and generalization; moreover, data publishers do not need to follow any specific pattern when they conduct data disguising. They described a generic method to conduct inference analysis across multiple published datasets in PPDR. They also formulate the problem as a well-studied maximum entropy estimation problem, and use standard non-linear programming tool to solve it. Experimental results demonstrate the effectiveness of this approach [16].

Optimal random perturbation at multiple privacy levels was proposed by Xiao et al [7]. They presented the algorithms of multi-level uniform perturbation that are tentatively robust even under the collusion of recipients. Specifically, this algorithm achieves two crucial properties. Firstly they can archive highest retention probability of data sharing. Secondly same way methods were used to compute ordinary permutation. This is more effective algorithm for uniform perturbation to perform analytical tasks like frequent itemset mining, counting and decision tree classification. This technique is incremental, because it does not assume that the privacy levels of future releases are known in advance.
Instead, a new request can arrive at any time, at any privacy level, no matter how many requests have been handled previously [17].

Chen, Keke, and Ling Liu proposed Privacy preserving data classification with rotation perturbation. This approach exploits the task-specific information about the datasets to be classified, aiming at producing a robust data perturbation that exhibits a better balance between loss of privacy and loss of information, without performance penalty. Concretely, they observed that the multi-dimensional geometric properties of datasets are the critical “task-specific information” for many classification algorithms. One intuitive way to preserve the multi-dimensional geometric properties is to perturb the original dataset through rotation transformation. They have identified and proved that kernel methods, SVM classifiers with the three popular kernels, and the hyperplane-based classifiers, are the three categories of classifiers that are rotation-invariant [18].

IV. CONCLUSION & FUTURE WORK

Privacy preserving data mining is methods over the chance to build models and extract patterns without disclosing private data. Privacy preservation methods for knowledge discovery can be categorized into two groups, data perturbation methods and cryptographic methods. Methods of the first group use data distortion, such as adding uniform noise, with the purpose of hiding private data, or, more formally, of guaranteeing k-anonymity. Methods of the second group are used for collaborative model learning: P >=2 parties contribute their data for the learning of a shared model according to protocols that prevent the disclosure of the contributed data. In this paper we are presenting some common techniques in the fields of multi-level trust and privacy preservation data mining.

As in the work proposed in [1] is only perturbation-based PPDM, the work is limited in the sense that it considers only linear attacks. More powerful adversaries may apply nonlinear techniques to derive original data and recover more information. Hence further enhancement will be prevention from such attacks and made the algorithm more private and secure.

REFERENCES


AUTHOR’S PROFILE

- Ankita Jain, Computer Science dept, Radharaman Institute of Technology, RGPV University, India. E-mail: (jankita27@gmail.com)
- Shubha Chaturvedi, Computer Science dept, Radharaman Institute of Technology, India. E-mail: (chaturvedishubha07@gmail.com)
- Anurag Jain, HOD Computer Science dept, Radharaman Institute of Technology, India. E-mail: (anurag.akjain@gmail.com)