Available online at www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 41 (2016) 68-75 EISSN 2392-2192

A Survey on Privacy Preserving Data Mining

S. Bharanya*, P. Amudha

Department of Computer Science and Engineering, Faculty of Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

*E-mail address: bharanya93@gmail.com

ABSTRACT

Privacy-preserving data mining has been considered widely because of the wide propagation of sensitive information over internet. A number of algorithmic techniques have been designed for privacy-preserving data mining that includes the state-of-the-art method. Privacy preserving data mining has become progressively popular because it allows sharing of confidential sensitive data for analysis purposes. It is important to maintain a ratio between privacy protection and knowledge discovery. To solve such problems many algorithms are proposed by various authors across the world. The main objective of this paper is to study various Privacy preserving data mining techniques and algorithms used for mining the item sets.

***Keywords*: Data mining, Privacy-preserving data mining, Perturbation, Frequent pattern mining**

1. INTRODUCTION

Data mining is the procedure of extracting the significant, available information from a large database. Frequent pattern mining determines the frequent relationships in large repositories of data, called patterns. On the other hand, the sensitive information about individuals being published may be disclosed, which create many privacy issues. Due to privacy issues many individuals are unwilling to share their data to the public that may leads to absence of data. Thus, privacy should be considered as an important in the field of Data Mining. Privacy Preserving Data Mining becomes a current research area to boom various privacy related issues.

(Received 01 February 2016; Accepted 10 February 2016; Date of Publication 11 February 2016)

This paper provides a wide study of various literatures and finally gives some conclusions based on the certain parameters. The rest of the paper is structured in such a way that Section II includes the classification of PPDM techniques. In section III various PPDM techniques and studies related to privacy issues, proposed by different authors are given. Finally, conclusion and future directions are given in Section IV.

2. PRIVACY PRESERVING DATA MINING (PPDM)

A. PPDM Classification

The PPDM techniques can be generally classified, into two scenarios as illustrated by Figure 1:

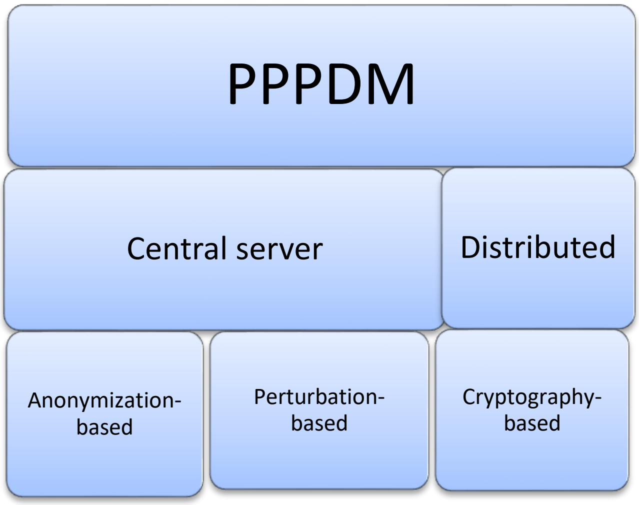

Figure 1. Classification of PPDM Hierarchy

i) Central Server ii) Distributed hierarchies are described here.

i) Central Server Scenario:

This Scenario is also known as Data Publishing Scenario because it tells how data is protected by the PPDM techniques before publishing it for data mining task. In this any number of the data owners/data miners are can be involved in the process of mining and handling privacy issues. The available information can be more adaptable. The best appropriate methods under this scenario are Anonymization and Perturbation.

ii) Distributed Scenario:

This scenario deals with steps involved in the mining process, for protecting against private databases that are carried out by the PPDM techniques. The data owners can get the aggregate results by combining their Databases. Perturbation based techniques are broadly used in this scenario. Few researches has focused on finding efficient protocols for specific problems of data mining such as 1) classification 2) clustering 3) pattern matching 4) association rule. Some of these studies are based on Central server scenario and few are based on Distributed scenario.

3. PPDM TECHNIQUES

In this section, to achieve privacy various PPDM algorithms and their techniques based on the above classification hierarchy are discussed in detail. The three major classical PPDM techniques: Anonymization, Perturbation and Cryptographic.

A. Anonymization Based

Sometimes the original form of the data must be published publically. Though it is not perturbed or encrypted; some sort of safety measure should be implemented before releasing the data in case of anonymization. Anonymization can be achieved through methods such as generalization, data removal, suppression and swapping. K-anonymity method is considered as the classical anonymization method and the majority of the studies are based on k-anonymity. During the process of anonymization the quasi–identifier (QI), which is the mixture of person precise identifiers, is considered. Most common methods to achieve k –anonymity are to generalize. In this method the possibly identifying attributes are first mapped using aggregation for numeric data and swapping for nominal data. A different anonymization technique known, as Condensation is a statistical approach which constructs constrained clusters and then produces pseudo data from the figures of these clusters.

B. Perturbation Based

Disclosing a ‘perturbed’ version of a data is one of the data distortion method for privacy protection, before releasing it for data mining. One of the widely accepted perturbation technique adding noise from a known distribution. In both scenarios, central server as well as in distributed perturbation methods can be used. These methods use data distortion techniques such as adding noise, randomization etc. The classifier for the original dataset is built directly from the perturbed dataset. Perturbation techniques can be used for achieving privacy but there are certain limitations: i) it restricts the range of algorithmic techniques ii) loss of implicit information. A data distortion technique that masks the data by randomly modifying the data values is known as Randomization.

C. Cryptography Based

This approach addresses the problem of reconstructing missing values of building an accurate data mining model. Therefore they propose a cryptographic protocol based on decision-tree classification on horizontally partitioned databases. Cryptographic techniques are widely considered in distributed environment. In distributed data mining environments to maintain privacy a secure multiparty computation (SMC) technique is used. SMC depends on either one of these three techniques: 1) homomorphic encryption

2) circuit evaluation 3) secret sharing. Pseudonymization is an approach that breaks the link between personal and medical information. Pseudonyms can be constructed using the encryption technique and it can be performed either at the database-level or at the application-level.

4. PRIVACY PRESERVING DATAMINING

Srikant et al. [1] had introduced the new efficient algorithms like association rule learning and inductive-rule learning. The new model is introduced which is equal to that of multi- party computation which requires a secure protocol. Now days in business and research field's data mining techniques are becoming popular. In case of complex algorithms and large inputs the general solutions are not enough. For any private data mining the solutions is based on guiding principle and the number of bits along with independent computation done by the individual parties.

Ryang, H et al., [2] had presented an approach for Association rule mining technique, where the important rules are discovered and it is denoted as X → Y .Where X is a set of items and Y is an item. In real time applications each item can have different nature; therefore multiple minimum supports are considered by Association rule mining. A novel tree structure is constructed with single scan is proposed called MHU-Tree (Multiple item supports with High Utility Tree). High utility item sets with multiple minimum supports are mined using MHU-Growth (Multiple item supports with High Utility Growth), algorithm.

Fouad, M.R. et al., [3] proposed a technique to protect the privacy of individuals, organizations will try to disclose only the minimum amount of information. For hiding the identity they apply a set of transformations such as (1) data suppression (2) data generalization (3) data perturbation to the micro data. Data transformation has suffered by the two major problems, namely, scalability and privacy risk, to address this, a scalable algorithm is proposed that meets the differential privacy. Two contributions of this work are 1) a personalized anonymization technique is proposed and 2) The proposed aggregate formulation and specific sampling are combined to give an anonymization algorithm satisfying differential privacy.

Deokate Pallavi B and M.M. Waghmare [4] had focused on the major research issue as Data mining-as-a-service. An organization (data owner) can outsource its resources to a third party service provider (server). Data owner contains the private properties like items of the outsourced transaction database and association rules. To protect the privacy the data owner encrypts and sends data for mining queries and accepts the true patterns from the server. The author also describes the problems faced by the corporate privacy while outsourcing the transactional databases. An attack model is proposed based on previous knowledge and a privacy preserving outsourced data mining scheme is introduced. It ensures that the attacker's previous information is different from the data transformed.

A. S. Syed Navaz et al., [5] proposed a technique in order to protect data sets. The inference rules need to be prevented by removing/adding certain item sets in the transactions. The purpose of hiding the Inference rules is to, so that the user may not be able to discover any valuable information and also any organization can release their data without the fear of 'Knowledge Discovery In Databases'. The major threat to the database security is uncovering the hidden patterns. Two fundamental approaches to protect sensitive rules are that, by hiding the frequent sets of data items the rules are prevented from generating and by locating their self-confidence below a user-specified onset the importance of the rules get reduced.

N. V. Muthu Lakshmi and Sandhya Rani, proposed a technique to improve the performance of the business or organizations the users makes use of strategic decisions with help of association rules found by researchers using many algorithms. To get mutual benefits data or information is shared to many users therefore in association rule mining threat occurs. Best solutions with minimum side effects are provided by exact hiding approaches among the existing techniques in this paper. In order to hide delicate rules by expressing constraint satisfaction problem, a modified inline algorithm is proposed along with the concepts of positive and negative border sets. By accepting divide and conquers technique, the efficiency of the proposed algorithm is increased.

Li, Y., et al. [6] presents a way to preserve privacy of data; a random perturbation process to individual values is introduced by perturbation-based PPDM approach. In this work, the scope of PPDM based on perturbation to Multi-Level Trust (MLT-PPDM) is extended. If the data miner is more trusted, then they can access the data with less perturbed copy. The key challenge of providing MLT-PPDM services is to prevent such diversity attacks. This solution prevents the data miners from accessing the perturbed copies by reconstructing the original data using any individual copy in the collection.

Tseng, V. S et al., [7] proposed an UP-Growth (Utility Pattern Growth) algorithm, for pruning candidate item sets they used set of techniques for mining high utility item sets. High utility item sets info's are upheld in a data structure named UP-Tree (Utility Pattern Tree, so that the candidate item sets generations are done with only two database scans. As a result the performance of UP-Growth is compared with the state-of-the-art algorithms. The trial results show that UP-Growth not only reduces the number of candidates effectively but also perform better than other algorithms considerably in terms of execution time.

Jiawei Ha et al., [8] had used a mining top-k frequent closed patterns whose length should not be minimum`, where k represents number of frequent patterns and minsignifies the minimal length of each pattern. The performance shows that in most cases, TFP performs better than two frequent closed pattern mining algorithm, CLOSET and CHARM with the best tuned min support. Thus he conclude that mining top-k frequent closed patterns without min support is more desirable for frequent pattern mining, than the traditional min support-based mining.

Mohammed J. Zaki and Ching-Jui Hsiao had proposed an efficient algorithm for Closed Association Rule Mining where only the set of closed frequent item sets is mined rather than mining all frequent items. Here it is assumed that any rule between item sets is equivalent to the some rule between closed item sets. Another algorithm known as CHARM is proposed for mining all closed frequent item sets. A wide experimental assessment on a number of actual and artificial databases shows that CHARM is better than a previous method via an order of magnitude.

Jieh-Shan Yeh and Po-Chiang Hsu [9] had made a study that mainly focus on privacy preserving utility mining (PPUM) and also presents two new algorithms, HHUIF and MSICF, in order to hide sensitive item sets, so that the oppositions cannot mine them from the altered database. This also reduces the impact for hiding sensitive item sets from the disinfected database. The results show that HHUIF achieves reduced miss costs than MSICF and also has a lower difference ratio than HHUIF between original and disinfected databases.

Unil Yun et al., [10] uses an algorithm named MU-Growth (Maximum Utility Growth) that is used for mining process by the candidates along with two techniques. Moreover, a MIQ-

Tree (Maximum Item Quantity Tree) tree structure is generated in order to capture database information with a single-pass. The proposed data structure is reorganized for reducing overrated utilities. Performance evaluation of MU-Growth shows that, it not only decreases the number of candidates but also overtakes state-of-the-art tree-based algorithms with overrated methods in terms of their runtime with alike memory usage.

Phuong-Thanh Laet al., [11] explored the lattice concept, which is a well-organized structure between theories. For real-world applications discovering an effective strategy for dynamically appraising the lattice is an important issue, where the new transactions are regularly inserted into databases. This study suggests a method for building the initial frequent closed item sets lattice from the original database in order to build an effective storage structure for mining association rules. The lattice is updated when new transactions are introduced and also the number of database rescans is reduced in the maintenance process over the complete database. This algorithm is related with building a lattice in batch mode so that they can demonstrate their effectiveness.

Ting Wang and Ling Liu [12] presents an organized study on protecting output privacy problems in data mining, and mainly, stream mining: (i) it highlight the importance by viewing the problem as output privacy does not guarantee the protection of input privacy; (ii) Also present a general disclosure model and inferencing that feats the intra window and inter window privacy holes in stream mining output; (iii) it also propose a light-weighted counter measure that effectively removes these holes without explicitly detecting them, while reducing the loss of output accuracy; (iv) while maintaining hard privacy guarantee, further improve the basic scheme by enchanting account of two sorts of semantic constraints, directing at excellently preserving utility-related semantics; (v) lastly, a wide experimental evaluation over both real data and synthetic, so that efficiency of this approach gets validated.

Gangin Leeet al. [13] proposed a novel algorithm, based on sliding window model of weighted best frequent pattern mining over data streams (WMFP-SW) is done, to obtain the weighted maximal frequent patterns replicating recent information done on data streams. He concluded by saying that the report of Performance experiments that MWFP-SW overtakes the previous algorithms in terms of runtime, usage, scalability and memory.

Dongwon Lee and Sung-Hyuk Park [14] presents a utility-based association-rule mining method that valuates association rules by calculating their explicit business benefits. Based on previous studies, three key elements such as opportunity, effectiveness, and probability are identified. To apply the utility-based mechanism to the processing of large transaction databases, constructed a functional algorithms, with heightened attention paid to their strategies, and based on real-world databases they are evaluated. Experimental results show that the proposed approach can offer users with better business welfares than the high-utility item set, suggesting numerous important strategic suggestions for both research and practice.

Grigorios Loukides and Aris Gkoulalas-Divanis [15] had proposed a novel approach for anonymizing data that satisfies the data publishers' utility needs and includes low information loss. To achieve this, an accurate information loss measure and an operational anonymization algorithm that explores a large part of the problem space got introduced. A general untried study, by medical statistics, proves that this approach permits many times more accurate query answering than the state-of-the-art methods, in terms of their efficiency.

C. Saravanabhavan and R. M. S. Parvathi [16] incorporated the privacy preserving concept into the before developed weighted utility mining approach. In this, by considering the sensitive item sets they have presented an effective algorithm for mining of privacy preserving high utility item sets. The algorithm contains three major stages, 1) Data sanitization, 2) Building of sensitive utility FP-tree and, 3) Mining of complex utility item sets. The testing had been carried out using real and also synthetic dataset. The presentation of the proposed algorithm is assessed with the help of the evaluation metrics such as Database difference ratio and Miss cost.

Stanley R.M. Oliveira and Osmar R. Zaiane [17] addresses the privacy problem against unauthorized secondary use of information. They introduce a family of geometric data transformation methods (GDTMs) which ensure that the mining procedure will not disturb privacy up to a certain degree of security. It focuses mainly on privacy preserving data clustering, particularly on partition-based and hierarchical methods. While preserving general features for clustering analysis, the proposed methods alter only confidential numerical attributes to meet privacy requirements. It demonstrates that our methods are real and provide acceptable values for balancing privacy and accuracy.

Elena Dasseni et al., [18] observes the privacy issues of a broad group of rules, which are known as association rules. If the discovered hazard of these rules is above a certain privacy threshold, those rules must be considered as sensitive. Occasionally, delicate rules should not be revealed to the public since, it may be used for impeding sensitive data, or they may provide advantage for business competitors.

5. CONCLUSION

Data mining is a large and sensitive domain where the Privacy preservation plays a vital role in providing data privacy. This paper gives a study of various research papers that have used different privacy preserving data mining technique to provide accurate results. Since, no such technique exists which overcomes all privacy issues, research in this direction can make major contributions.The future work can be carried out using any one of the existing techniques or using a combination of these or by developing entirely a new technique.

References

[1] Srikant R., Agrawal R., Privacy-preserving data mining. SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data May 2000 Pages 439–450. https://doi.org/10.1145/342009.335438

[2] Ryang, H., Yun, U., & Ryu, K. Discovering high utility item sets with multiple minimum support. *Intelligent Data Analysis*, vol. 18, no. 6, pp. 1027-1047, 2014

[3] Fouad, M. R., Elbassioni, K. M., &Bertino, E. (2014). A supermodularity-based differential privacy preserving algorithm for data anonymization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7), pp. 1591-1601.

[4] Deokate Pallavi B.,& M.M. Waghmare. Privacy-Preserving in Outsourced Transaction Databases from Association Rules Mining, 2014. Corpus ID: 13552173

[5] A. S. Syed Navaz, M. Ravi, & T. Prabhu. Preventing Disclosure of Sensitive Knowledge by Hiding Inference. *International Journal of Computer Applications* 63(1) (2013) 32-38. DOI: 10.5120/10431-5104

[6] Li, Y., Chen, M., Li, Q., & Zhang, W. (2012). Enabling multilevel trust in privacy preserving data mining. *IEEE Transactions on Knowledge and Data Engineering,* 24(9), pp. 1598-1612.

[7] Tseng, V. S., Shie, B. E., Wu, C. W., & Yu, P. S. (2013). Efficient algorithms for mining high utility itemsets from transactional databases. *IEEE Transactions on Knowledge and Data Engineering,* 25(8), pp. 1772-1786.

[8] Jiawei Han, Jianyong Wang, Ying Lu and P. Tzvetkov, Mining top-k frequent closed patterns without minimum support. *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* 2002, pp. 211-218, doi: 10.1109/ICDM.2002.1183905.

[9] Jieh-Shan Yeh, & Po-Chiang Hsu. HHUIF and MSICF: Novel algorithms for privacy preserving utility mining. *Expert Systems with Applications,* 37(7), pp. 4779-4786, 2010

[10] Yun, U., Ryang, H., & Ryu, K. (2014). High utility itemset mining with techniques for. reducing overestimated utilities and pruning candidates. *Expert Systems with Applications,* 41(8), pp. 3861-3878

[11] Phuong-Thanh La, Bac Le, Bay Vo, Incrementally building frequent closed itemset lattice. *Expert Systems with Applications,* Volume 41, Issue 6, 2014, Pages 2703-2712, https://doi.org/10.1016/j.eswa.2013.11.002

[12] Wang, T., & Liu, L. (2011). Output privacy in data mining. *ACM Transactions on Database Systems,* 36(1), pp. 1-34. article 1.

[13] Lee, G., Yun, U., & Ryu, K. (2014). Sliding window based weighted maximal frequent pattern mining over data streams. *Expert Systems with Applications*, 41(2), pp. 694-708.

[14] Dongwon Lee & Sung-Hyuk Park, Utility-based association rule mining: A marketing solution for cross-selling. *Expert Systems with Applications,* 40(7), pp. 2715-2725, 2013.

[15] Grigorios Loukides & Aris Gkoulalas-Divanis. Utility-preserving transaction data anonymization with low information loss. *Expert Systems with Applications*, 39(10), pp. 9764-9777, 2012.

[16] C. Saravanabhavan & R. M. S. Parvathi. Privacy preserving sensitive utility pattern mining. *Journal of Theoretical and Applied Information Technology* 49(2) (2013) 496-506

[17] Stanley R.M. Oliveira & Osmar R. Zaiane. Privacy Preserving Clustering by Data Transformation. *Inaugural Issue of Journal of Information and Data Management* Vol. 1 No. 1 (2010) 2010, 37-51

[18] Dasseni E., Verykios V.S., Elmagarmid A.K., Bertino E. (2001) Hiding Association Rules by Using Confidence and Support. In: Moskowitz I.S. (eds) Information Hiding. IH 2001. Lecture Notes in Computer Science, vol 2137. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45496-9_27