Association Rule Hiding Approaches–A Survey

S. Sharmila, S.Vijayarani,
Department of Computer Science,
Bharathiar University, Coimbatore.

Abstract: Privacy preserving data mining is used to extract relevant knowledge from large amount of data and at the same time protect sensitive information from the data miners. People in business, hospitals, educational institutions, and banks need a secure and safe transaction of their data. To serve this need Privacy Preserving Data Mining (PPDM) was created. PPDM solves the problem related to designing accurate models about combined data without requiring the access to exact information in individual data record. PPDM is the most important research area for protecting the perceptive data or knowledge. Association rule hiding is one of the techniques of PPDM to protect the association rules generated by association rule mining. This study presents a survey of association rule hiding approach for preserving privacy of the user data. Association rule hiding methodology consists of five approaches namely Heuristic, Border, Exact, Cryptography and Reconstruction The study has briefly explained the approaches.

Keywords: Privacy preserving Data Mining, Association Rule Hiding approaches. Heuristic approach

I. INTRODUCTION

1.1 Privacy Preserving Data Mining

Privacy preserving data mining technique protects the privacy information. Privacy preserving techniques are classified on the basis of data distribution, data distortion, data mining algorithms, anonymization, data or rules hiding, and privacy protection. Intensive research findings over the decades have revealed that the existing privacy preserving data mining search approaches are still suffering from major inadequacies which include distributed clients' data to multi semi honest providers, overhead of computing global mining, and incremental data privacy issue in cloud computing, integrity of mining result, utility of data, scalability and overhead performance [1][2]. Thus, a robust, scalable model is essential to overcome these shortcomings. Furthermore, to protect the privacy of each client, proper anonymization of data is essential prior to publishing it. The connection between personal data and personal identification should be dispelled [11] [14].

The privacy preserving data mining involves two steps: First, sensitive raw data should be modified or trimmed out from the original database so that the recipient of the data does not compromise on privacy [4] [5]. Second, sensitive knowledge which can be mined from a database using data mining algorithms should also be excluded [3]. The main objective of privacy preserving data mining is to develop algorithms for modifying the original data so that the private data and knowledge remain private even after the mining process [6]. Privacy of the user data is preserved, and, at the same time, the mining models can be reconstructed from the modified data with reasonable accuracy [10].

In the recent times, differential privacy model has widely been explored to provide maximum security to the private statistical databases by minimizing the chances of record identification [7]. There are several trusted parties that hold datasets of sensitive information such as medical records, voter registration information, email usage, and tourism. The primary aim of this model is to provide global, statistical information about the data publicly available, and at the same time ensure that the privacy of the users, whose information is contained in the dataset, is protected [8].Generally, security protects the data against unauthorized access when transmitted across a network [9] [10]. However, upon reaching an authorized user, no additional constraints on revealing the personal information of an individual are imposed [11].

1.2 Techniques in Privacy Preserving Data Mining

Various approaches have been proposed in the existing literature for privacy-preserving data mining. They differ from each other with respect to their assumptions of data collection model and user privacy requirements [12]. There has been some research pertaining to how much information can be inferred, calculated or revealed from the data made available through data mining process, and how to minimize the leakage of information [13]. PPDM techniques are listed out below. Privacy Preserving Data Mining consists of two scenario central and distributed scenario. Figure 1 illustrates the techniques in Privacy Preserving Data Mining [19].

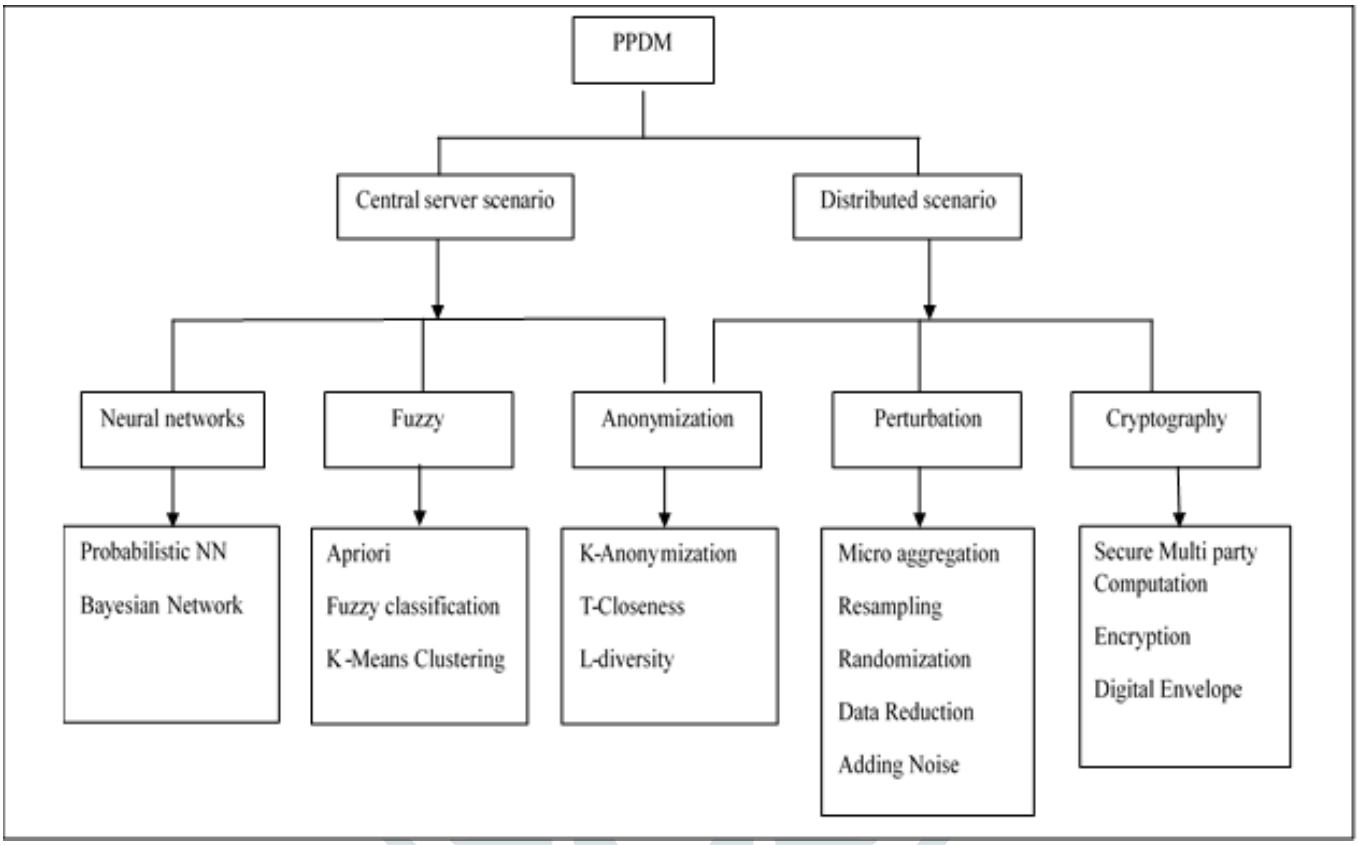

Fig 1: Techniques of PPDM

1.2.1 Central Server Scenario

In this scenario, a trusted third party Central Commodity Server plays an important role. Each of the contributing parties entrust the Central Commodity Server the task of preserving the privacy of individual contributing parties [15]. Before publishing the data, all the contributors transfer their data to the server [18].

1.2.2 1.1.2 Distributed Scenario

These scenarios have very specific goals and are based on heavy computation techniques like Secure Multiparty Computation (SMC) and Cryptographic techniques [16]. The contributing parties sanitize the data and privatize it. The mining can be performed by the data owners and their aggregate results are then used for finding the effective association rules [17] [27

1.2.3 Neural Network based

Neural network is a mathematical model or computational model based on biological neural networks [18]. Neural Network based PPDM is studied in literature to achieve privacy of individual contributing parties without compromising information loss [19] [20].

1.2.4 Fuzzy Logic Based

PPDM based on Fuzzy algorithms allow achieving anonymization without significant loss of information. The algorithms merge similar records into clusters [15]. Each cluster formed is distinct from other clusters and the records of each cluster are not distinguishable from those of other clusters [17].

1.2.5 Anonymization Based

It is used to modify the data in such a way that any information that could directly link data to individuals is removed .The data may not be encrypted and perturbed, but still some sort of precaution should be taken before releasing the data in terms of anonymization [18]. This is a kind of generalization of some attributes that protects against identity disclosure. Anonymization can be achieved by methods like generalization, suppression, data removal, permutation, swapping etc [19]. K-anonymity method is treated as the conventional anonymization method. Improved methods like l-diversity, t-closeness, km -anonymization, (α, k) anonymity, p-sensitive k-anonymity, anonymity, are described in [20],

1.2.6 Perturbation Based

Perturbation technique employs a mechanism to distort data prior to data mining. It consists of fake sensitive data which is based on statistical techniques. Once the local perturbed copy is generated the data miner can reconstruct the perturbed version to obtain the original data distribution [22] [23].

1.2.7 Cryptography Based

If the parties distributed across multiple sites are legally prohibited from sharing their datasets, a mining model to be built must be able to maintain the privacy of contributing parties. In Cryptography technique, Secure Multi-party Computation SMC challenge in securing multiparty computation is to allow the calculation of any function on a set of data distributed across multiple sites [15][16]. Each site has a portion of the data and calculation should be realized in such a way that any party can deduct in some way the data of other sites from its own data and calculation results [25].

II. ASSOCIATION RULE HIDING

Association Rule Hiding is a PPDM technique used in conjunction with Association Rule Mining method in transactional database [21]. Data Mining, with its approach using the association rule hiding method plays a major role in the protection of perceptive data or knowledge [50]. Association rule hiding is the process of modifying the original database in such a way that confident sensitive association rules disappear without leaving any effect on the data and the non-sensitive rules [16][17].

Association rule finds the frequently occurred patterns, associations and correlation between itemsets in operational databases, non-operational relational databases, and other information repositories [11] [19]. The task for the frequent itemset mining algorithm is then to find all common sets of items, defined as those itemsets that have at least a minimum support. Association rules are discovered by analyzing data for frequent if/then patterns and applying the support and confidence to describe the most significant correlation [10][22].

2.1 Association Rule hiding Strategy

A rule for illustration X => Y, can be realized in two ways: It can be done so by either decreasing the support of the item set X and Y below the threshold of minimum support or decreasing the confidence of the item set X and Y below the threshold of least confidence [14]. Decreasing the confidence of the rule X => Y can be done by either escalating the support of only X in transactions or decreasing the support of Y in transactions. Decreasing the support of the rule X => Y can be done by decreasing the support of the corresponding large item set XY [11][13].

2.2 Approaches of Association Rule Hiding Methodologies

- Heuristic-Based approach - Heuristic methods is one of the approaches of Association rule hiding, this process is used to prefer the appropriate data for sanitization to hide the sensitive information [16].
- Border-Based approach- It hides sensitive association rule by modifying the borders in the frame of the frequent and the infrequent item sets of the original database.
- Extract-based approach- proposed an approach to find optimal solution for rule hiding problem. This tries to minimize the distance between the original database and its sanitized version [14][16].
- Reconstruction-based techniques - This approach places the original data aside and start from knowledge base. To sanitize, it hides the sensitive rules by sanitizing item set lattice rather than sanitizing original dataset [23].
- Cryptography- based approaches - are used for multiparty computation, when database is distributed among several sites [19] [21]. Multiple parties may wish to share their private data, without leaking any sensitive information at their end. This approach is categorised as: vertically partitioned distributed data, horizontally partitioned distributed data [25].

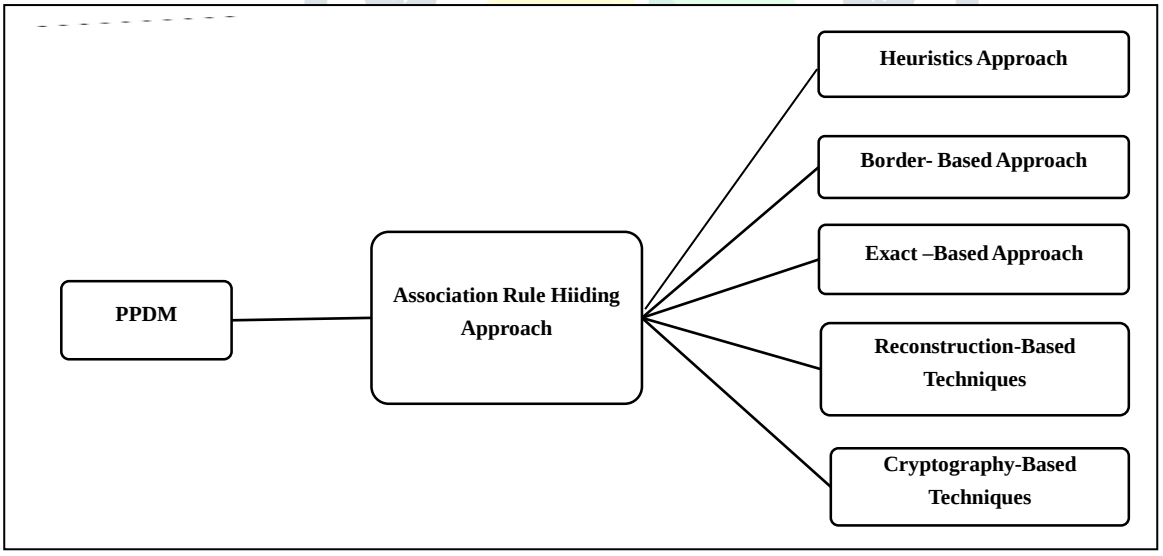

Fig.2: Approaches of Association rule hiding

Figure.2 illustrates the approaches of Association rule hiding. Association Rule Hiding approaches can be classified into five classes: heuristic based approaches, border based approaches, exact approaches, reconstruction based approaches, and cryptography based approaches.

III. Background Study of Association Rule Hiding

S.No	Author	Title/Year	Proposed Technique	Inference
1.	Guanling Lee∗ and Yi Chun Chen	Protecting sensitive knowledge in association patterns mining (2012)	Some open issues related to the Problem of association rules hiding were discussed.	First, the association rules hiding problem was a time consuming task, and developed an algorithm by applying parallel or cloud computation techniques to solve it efficiently. Second, in real-world applications, the transactions in the database might be deleted, inserted, or modified. Third, popularity of wireless network, global positioning system, and mobile devises, lots of spatiotemporal geo referenced data were produced and collected.
2.	S.Vijayarani A.Tamilarasi R.SeethaLakshmi	Tabu search based association rule hiding (2011)	Tabu search optimization technique	Proposed a protected sensitive association rules. The proposed technique was compared with the previous approaches. This approach has modified the sensitive rules accurately without affecting the non-sensitive rules and no false rules were generated.
3.	Mohammad azam chhipa Prof. Lalit gehlod	Survey on association rule hiding approaches (2011)	Association rule mining and techniques to hide sensitive rules	Discussed about data mining, association rule mining and techniques to hide sensitive rules which were mined through data mining. Author has studied about association rule and algorithms to mine association rule in data set, and also studied about studied different approaches to hide sensitive rules.
4.	S.Vijayarani S.Narmadha	Protecting Sensitive Association Rules in Privacy Preserving Data Mining using Genetic Algorithms (2011)	Genetic Algorithms	Author had investigated about how sensitive rules were protected from malicious data miner and also proposed genetic algorithm technique for hiding the sensitive rules. In this technique, all the sensitive were hidden; no false rules was generated and non-sensitive rules was not affected.
5.	Dr.S.Vijayarani R.Prasannalakshmi	Comparative analysis of association Rule generation algorithms in data Streams (2015)	Supervised Association Rule	This research work was Concentrated on how the traditional algorithms were used for generating association rules in data streams.. A number of rules generated by an algorithm and execution time were considered for performance Factors.
6.	Mahtab Hossein Afshari	Association rule hiding using cuckoo optimization algorithm(2016)	Cuckoo Optimization Algorithm	Hiding was performed using the distortion technique. Further in this study three fitness functions were defined which makes it possible to achieve a solution with the fewest side effects. Introducing an efficient immigration function in this approach has improved its ability to move from any local optimum.

IV. CONCLUSION

Privacy preserving in data mining has a significant value in business operations. When data is shared through a network, Data is visible to data users. Solving this problem is a formidable challenge. Data mining techniques provide better results for the safe transaction of the data. This study presents a survey of Association Rule Hiding approaches to solve privacy problems. The main objective of PPDM is to incorporate the traditional data mining techniques to transform the data with the view to mask sensitive information, and a major challenge in this process is to efficiently transform the data and recover its mining outcome from the transformed one. Thus, the study examines the overhead for preserving privacy of growing data, and the integrity of mining result. The study discusses Heuristic-based approach in detail with the aim to solve the major problems associated with privacy preservation.

REFERENCES

[1]. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2006.

[2]. A comprehensive review on privacy preserving data mining Yousra Abdul Alsahib S. Aldeen1,2*, Mazleena Salleh1 and Mohammad Abdur Razzaque1. SpringerPlus (2015) 4:694 DOI 10.1186/s40064-015-1481-x.

[3]. Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey Alpa Shah, Ravi Gulati , Research Gate March 2016 DOI: 10.5120/ijca2016909006

[4]. S. Vijayarani, A. Tamilarasi and R. SeethaLakshmi, "Privacy Preserving Data Mining Based on Association Rule-A Survey". In Proc. of the International Conference on Communication and Computational Intelligence-2010, pp. 99-103.

[5]. N. R. Radadiya, N. B. Prajapati, and K. H. Shah, ``Privacy preserving in association rule mining," Int. Journal of Innovative Res., vol. 2, no. 4, pp. 203_213, 2013.

[6]. Jain Y.K. (2011), 'An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining', International Journal of Computer Science and Engineering, pp.96-104

[7]. Gayatri Nayak and Swagatika Devi (2011), 'A Survey On Privacy Preserving Data Mining: Approaches And Techniques', International Journal of Engineering Science and Technology, pp.2127-2133

[8]. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD'93), pages 207–216, 1993.

[9]. Maryam Fouladfar A heuristic algorithm for quick hiding of association rules , ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 1, No.13 , January 2015 ISSN : 2322-5157

[10]. Juggapong Natwichai Xue Li, "A Reconstruction-based Algorithm for Classification Rules Hiding".

[11]. Gayathiri P and Dr. B Poorna, "Association Rule Hiding Techniques for Privacy Preserving Data Mining: A Study"

[12]. E. Pontikakis, Y. Theodoridis, A. Tsitsonis, L. Chang, and V. S. Verykios. A quantitative and qualitative analysis of blocking in association rule hiding. In Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society (WPES 2004), pages 29–30, 2004

[13]. Khyati B. Jadav, "A Survey on Association Rule Hiding Methods".

[14]. K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In Proceeding of the Fifth IEEE International Conference on Data Mining (ICDM 2005), pages 466–473, 2005.

[15]. S. Jha, L. Kruger, and P. McDaniel. Privacy preserving clustering. In Proceedings of the 10th European Symposium on Research in Computer Security (ESORICS 2005), pages 397–417, 2005.

[16]. V. S. Verykios, ``Association rule hiding methods," Wiley Interdiscipl. Rev., Data Mining Knowledge. Discovery, vol. 3, no. 1, pp. 28_36, 2013.

[17]. Vassilios S. Verykios, "A Survey of Association Rule Hiding Methods for Privacy".

[18]. C. Modi, U.P. Rao and D.R.Patel, "A Survey on Preserving Privacy for Sensitive Association Rules in Databases" Springer-Verlag Berlin Heidelberg 2010, pp. 538-544.

[19]. J. Vaidya, and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," In proc. Int"l Conf. Knowledge Discovery and Data Mining, pp. 639–644, July 2002.

[20]. M.Mahendran "An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach".

[21]. X. Sun, and P. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In: Proc. Fifth IEEE Int"l. Conf. Data Mining (ICDM 2005), pp. 426–433, 2005.

[22]. A. Gkoulalas-Divanis, V. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding," In: Proc. ACM Conf. Information and Knowledge Management (CIKM 2006), pp. 748–757 2006.

[23]. Y. Guo, "Reconstruction-Based Association Rule Hiding," In Proc. Of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 2007.

[24]. M. Kantarcioglu, and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp. 1026-1037, Sept. 2004.

[25]. V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. ACM SIGMOD Record, 33(1):50–57, 2004.