An Insight in to Privacy Preserving Data Mining Methods

K. Srinivasa Rao* & B. Srinivasa Rao**

*Associate Professor, Department of Computer Science and Engineering, Swarna Bharathi College of Engineering, Khammam, Andhra Pradesh, INDIA. E-Mail: ksrao517@gmail.com
**Associate Professor, Department of Computer Science and Engineering, Medha Institute of Science & Technology for Women, Khammam, Andhra Pradesh, INDIA. E-Mail: srinuprec@gmail.com

***Abstract*—**Recent advances in information, communications, data mining, and security technologies have gave rise to a new era of research, known as Privacy Preserving Data Mining (PPDM). Several data mining algorithms, incorporating privacy preserving mechanisms, have been developed that allow one to extract relevant knowledge from large amount of data, while hide sensitive data or information from disclosure or inference. PPDM is a new attempt; thus, several research questions have often being asked. For instance: (1) How to measure the performance of these algorithms? (2) How effective of these algorithms in terms of privacy preserving? (3) Will they impact the accuracy of data mining results? and (4) Which one can better protect sensitive information? To help answer these questions, we conduct an extensive review on literature. We present a classification scheme, adopted from early studies, to guide the review process. Finally, we share directions for future research.

***Keywords*—**Data Mining; Data Perturbation; Knowledge Discovery; Privacy Preservation.

***Abbreviations*—**Privacy Preserving Data Mining (PPDM); Secure Multiparty Computation (SMC).

I. INTRODUCTION

INCREASING network complexity, affording greater access, sharing information and a growing emphasis on the Internet have made information security and privacy a major concern for individuals and organizations. Data mining is a well-known technology for automatically and intelligently extracting knowledge from large amount of data. Such a process, however, can also disclosure sensitive information about individuals compromising the individual's right to privacy. Moreover, data mining techniques can reveal critical information about business transactions, compromising the free competition in a business setting [Bertino et al., 2005]. Privacy preserving data mining (PPDM) is a new era of research in data mining. Its ultimate goal is to develop efficient algorithms that allow one to extract relevant knowledge from large amount of data, while prevent sensitive information from disclosure or inference.

PPDM research usually takes one of the three philosophical approaches: (1) data hiding, in which sensitive raw data like identifiers, name, addresses, etc. were altered, blocked, or trimmed out from the original database, in order for the users of the data not to be able to compromise another person's privacy; (2) rule hiding, in which sensitive knowledge extracted from the data mining process be excluded for use, because confidential information may be derived from the released knowledge. This problem is also commonly called the "database inference problem;" and (3) Secure Multiparty Computation (SMC), where distributed data are encrypted before released or shared for computations; thus, no party knows anything except its own inputs and the results.

PPDM is a fast growing research area. Given the number of different algorithms have been developed over the last years, there is an emerging need of synthesizing literature to understand the nature of problem, identify potential research issues, standardize new research area, and evaluate the relative performance of different approaches [Verykios et al., 2004; Bertino et al., 2005]. The main purpose of this study is to review the state-of-the-art in current PPDM research in order to better understand existing algorithms, answer research questions and move forward the field of research.

II. CLASSIFICATION FRAMEWORK FOR PPDM

In this paper, we propose to consolidate and simplify the taxonomy brought by Bertino et al., (2005). We propose to reduce the PPDM taxonomy into four levels: data distribution, purposes of hiding, data mining algorithms, and privacy preserving techniques (see figure 1).

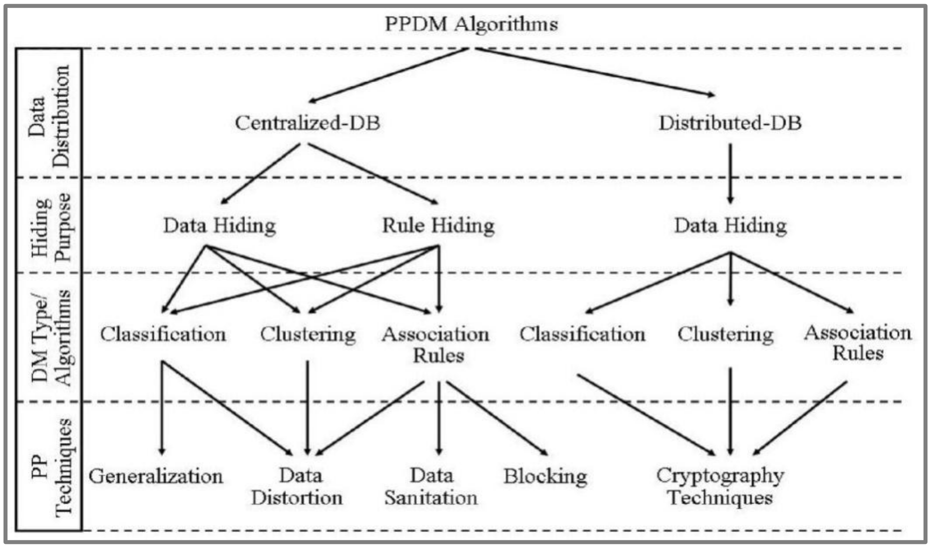

Figure 1: The Taxonomy of PPDM Algorithms

2.1. *Data Distribution*

The PPDM algorithms can be first divided into two major categories, centralized and distributed data, based on the distribution of data. In a centralized database (DB) environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases. Earlier research has been predominately focused on dealing with privacy preservation in a centralized DB [Du & Zhan 2003; Evfimievski et al., 2003, 2004; Islam & Brankovic, 2004; Natwichail et al., 2005; Oliveira et al., 2002, 2003, 2003A, 2003B, 2004, 2004A; Rizvi & Haritsa, 2002; Saygin et al., 2001, 2002; Verkios, et al., 2003; Wang et al., 2004; Xia et al., 2004]. The difficulties of applying PPDM algorithms to a distributed DB can be attributed to two reasons: first, the data owners have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data for data mining, the communication cost between the sites is too expensive. In today's global digital environment, most data are often stored in different sites, thus, more attention and research should be focused on distributed PPDM algorithms.

2.2. *Hiding Purposes*

The PPDM algorithms can be further classified into two types, data hiding and rule hiding, according to the purposes of hiding. Data hiding refers to the cases where the sensitive data from original database like identity, name, and address that can be linked, directly or indirectly, to an individual person are hided. In contrast, in rule hiding, we remove the sensitive knowledge derived from original database after applying data mining algorithms. Majority of the PPDM algorithms used data hiding techniques. This is especially true in a distributed database environment [Du & Zhan, 2002; Kantarcioglu & Clifton, 2002, 2003; Klusch et al., 2003; Lindell & Pinkas, 2000; Merugu & Ghosh, 2003; Vaidya & Clifton, 2002, 2003, 2005; Verkios et al., 2003; Yang et al., 2006; Zhan et al., 2005], as the techniques can be used to prevent individual information from being discovered by other parties in the joint computational process. Most PPDM algorithms hide sensitive patterns by modifying data [Du & Zhan, 2003; Oliveira et al., 2003B, 2004; Xia et al., 2004; Du & Zhan, 2002; Evfimievski et al., 2003, 2004; Kantarcioglu & Clifton 2002, 2003; Islam & Brankovic, 2004; Klusch et al., 2003; Lindell & Pinkas, 2000; Merugu & Ghosh, 2003; Rizvi & Haritsa, 2002; Vaidya & Clifton, 2002, 2003, 2005; Verkios, et al., 2003; Wang et al., 2004; Yang et al., 2006; Zhan et al., 2005]. Also, at present, the rule hiding techniques is only being adopted by association rule mining for centralized DB [Oliveira et al., 2002, 2003, 2003A, 2004A; Verkios et al., 2003; Saygin et al., 2001, 2002].

2.3. *Data Mining Tasks/Algorithms*

Currently, the PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups. A majority of the PPDM algorithms used association rule method for mining data [Evfimievski et al., 2003, 2004; Oliveira et al., 2002, 2003, 2003A, 2004A; Rizvi & Haritsa 2002; Saygin et al., 2001, 2002; Verkios et al., 2003; Xia et al., 2004; Kantarcioglu & Clifton 2002, 2003; Vaidya & Clifton 2005; Veloso et al., 2003], followed by classification [Du & Zhan, 2003; Islam & Brankovic, 2004; Natwichail et al., 2005; Wang et al., 2004; Du & Zhan, 2002; Kantarcioglu & Clifton, 2003; Lindell & Pinkas, 2000; Vaidya & Clifton, 2005; Yang et al., 2006], and then clustering [Oliveira et al., 2003B, 2004; Vaidya & Clifton, 2003; Klusch et al., 2003; Merugu & Ghosh, 2003].

2.4. Privacy Preservation Technique

Four techniques – sanitation, blocking, distort, and generalization – have been used to hide data items for a centralized data distribution. Data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined. The blocking approach replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Data distort protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but "global" properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value.

The privacy preservation technique used in a distributed database is mainly based on cryptography techniques. SMC algorithms deal with computing any function on any input, in a distributed network where each participant holds one of the inputs, while ensuring that no more information is revealed to a participant in the computation than can be inferred from that participant's input and output. Data distort is the most popular method used in hiding data [Du & Zhan, 2003; Evfimievski et al., 2003, 2004; Islam & Brankovic, 2004; Oliveira et al., 2003B, 2004; Rizvi & Haritsa, 2002; Xia et al., 2004], followed by data sanitation [Oliveira et al., 2002, 2003, 2003A, 2004A; Saygin et al., 2002; Verkios et al., 2003] and generalization [Natwichail et al., 2005; Wang et al., 2004]. If one wants to obtain data mining results from different data sources, then the only method can be used is a cryptography technique [Du & Zhan, 2002; Kantarcioglu & Clifton, 2002, 2003; Klusch et al., 2003; Lindell & Pinkas 2000; Merugu & Ghosh, 2003; Vaidya & Clifton, 2002, 2003, 2005; Veloso et al., 2003; Yang et al., 2006; Zhan et al., 2005]. Since the parties who use SMC operators cannot reveal anything from others except final results, it can have benefits of both accuracy of data mining results and the privacy of the database.

III. Suggestions for Future Work

There are many future research directions for privacy preserving data mining. First, present studies tend to use different terminology to describe similar or related practice. For instance, people used data modification, data perturbation, data sanitation, data hiding, and preprocessing as possible methods for preserving privacy; however, all are in fact related to the use of some types of technique to modify original data so that private data and knowledge remain private even after the mining process. Lacking a common language for discussions will cause misunderstanding and slow down the research breakthrough. Therefore, there is an emerging need of standardizing the terminology and PPDM practice.

Second, most prior PPDM algorithms were developed for use with data stored in a centralized database. However, in today's global digital environment, data is often stored in different sites. With recent advances in information and communication technologies, the distributed PPDM methodology may have a wider application, especially in medical, health care, banking, military and supply chain scenarios.

Third, data hiding techniques have been the dominated methods for protecting privacy of individual information. However, those algorithms do not pay full attention to data mining results, which may lead to sensitive rules leakages. While some algorithms are designed for preserving the rule such as with sensitive information, it may degrade the accuracy of other non-sensitive rules. Thus, further investigation, focusing on combining data and rule hiding, may be beneficial, specifically, when taking into account the interactive impact of sensitive and non-sensitive rules.

Fourth, although many machine learning methods have been used for classification, clustering, and other data mining tasks (e.g., diagnose, prediction, optimization), currently only the association rules method has been predominately used for classification. It would be interesting to see how to extend the current technique and practice into other problem domains or data mining tasks. Furthermore, it is important to find the privacy preserving technique that is independent of data mining task so that after applying privacy preserving technique a database can be released without being constrained to the original task.

Finally, identifying suitable evaluation criteria and developing benchmarks for algorithm selection are two important aspects in PPDM research. A framework for evaluating selected association rule hiding algorithms has been proposed by Bertino et al., (2005). Future research can consider testing the proposed evaluation framework for other privacy preservation algorithms, such as data distortion or cryptography methods.

IV. Conclusions

PPDM has recently emerged as a new field of study. As a new comer, PPDM may offer a wide application prospect but at the same time it also brings us many issues / problems to be answered. In this study, we conduct a comprehensive survey on 29 prior studies to find out the current status of PPDM development. We propose a generic PPDM framework and a simplified taxonomy to help understand the problem and explore possible research issues. We also examine the strengths and weaknesses of different privacy preserving techniques and summarize general principles from early research to guide the selection of PPDM algorithms. As part of future work, we plan to apply the proposed evaluation framework to formally test a complete spectrum of PPDM algorithms.

REFERENCES

[1] Y. Lindell & B. Pinkas (2000), "Privacy Preserving Data Mining", *Advances in Cryptology – CRYPTO 2000, Springer-Verlag*, Pp. 36–54.

[2] Y. Saygin, V. Verykios & C. Clifton (2001), "Using Unknowns to Prevent Discovery of Association Rules", *ACM SIGMOD Record*, Vol. 30, No. 4.

[3] Y. Saygin, V. Verykios & A. Elmagarmid (2002), "Privacy Preserving Association Rule Mining", *Proceedings of 12th International Workshop on Research Issues in Data Engineering (RIDE).*

[4] J. Vaidya & C. Clifton (2002), "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Pp. 639–644.

[5] M. Kantarcioglu & C. Clifton (2002), "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD).*

[6] W. Du & Z. Zhan (2002), "Building Decision Tree Classifier on Private Data", *Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (PSDM'02)*, Maebashi City, Japan, Pp.1–8.

[7] S.R.M. Oliveira & O.R. Zaïane (2002), "Privacy Preserving Frequent Itemset Mining", *Proceedings of the IEEE International Conference on Privacy, Security and Data Mining (PSDM'02)*, Maebashi City, Japan, Pp. 43–54.

[8] J. Rizvi & R. Haritsa (2002), "Maintaining Data Privacy in Association Rule Mining", *Proceedings of the 28th Very Large Data Base Conference (VLDB'02)*, Hong Kong, China, Pp. 682–693.

[9] S.R.M. Oliveira & O.R. Zaïane (2003), "Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining", *Proceedings of the 7th International Database Engineering and Applications Symposium (IDEAS'03)*, Hong Kong, China, Pp. 54–65.

[10] S.R.M. Oliveira & O.R. Zaïane (2003A), "Protecting Sensitive Knowledge by Data Sanitization", *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03)*, Melbourne, Florida, USA, Pp. 613–616.

[11] S.R.M. Oliveira & O.R. Zaïane (2003B), "Privacy Preserving Clustering by Data Transformation", *Proceedings of the 18th Brazilian Symposium on Databases*, Manaus, Amazonas, Brazil, Pp. 304–318.

[12] J. Vaidya & C. Clifton (2003), "Privacy-Preserving K-Means Clustering over Vertically Partitioned Data", *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery in Data (KDD'03)*, Washington D.C., USA, Pp. 206–215.

[13] M. Kantarcıoglu & C. Clifton (2003), "Assuring Privacy When Big Brother is Watching", *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Privacy & Security*, Pp. 88–93.

[14] M. Klusch, S. Lodi & G. Moro (2003), "Distributed Clustering based on Sampling Local Density Estimates", *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, Acapulco, Mexico, Pp. 485–490.

[15] W. Du & Z. Zhan (2003), "Using Randomized Response Techniques for Privacy-Preserving Data Mining", *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, D.C.

[16] A. Evfimievski, J. Gehrke & R. Srikant (2003), "Limiting Privacy Breaches in Privacy Preserving Data Mining", *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, San Diego, California, Pp. 211–222.

[17] S. Merugu & J. Ghosh (2003), "Privacy-Preserving Distributed Clustering using Generative Models", *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03)*, Melbourne, FL, USA, Pp. 211–219.

[18] A. Veloso, Jr. W. Meira, S. Parthasarathy & M. de Carvalho (2003), "Efficient, Accurate and Privacy-Preserving Data Mining for Frequent Itemsets in Distributed Databases", *Proceedings of the 18th Brazilian Symposium on Databases*, Manaus, Amazonas, Brazil, Pp. 281–292.

[19] S. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin and Y. Theodoridis (2004), "State-of-the-art in Privacy Preserving Data Mining", *ACM SIGMOD Record*, Vol. 33, No. 1, Pp. 50–57.

[20] K. Wang, S. Yu & S. Chakraborty (2004), "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection", *Proceedings the 4th IEEE International Conference on Data Mining (ICDM'04)*, Brighton, United Kingdom, Pp. 249–256.

[21] Y. Xia, Y. Yang, Y. Chi & R.R. Muntz (2004), "Mining Association Rules with Non-uniform Privacy Concerns", *Technical Report CSD-TR No. 040015*, Univ. of California.

[22] S.R.M. Oliveira & O.R. Zaïane (2004), "Achieving Privacy Preservation When Sharing Data for Clustering", *Proceedings of the International Workshop on Secure Data Management in a Connected World (SDM'04), In Conjunction with the 30th Very Large Data Base Conference (VLDB'04)*, Toronto, Canada, Pp. 76–82.

[23] S.R.M. Oliveira, O.R. Zaïane & Y. Saygin (2004A), "Secure Association Rule Sharing", *PAKDD*, Pp. 74–85

[24] A. Evfimievski, R. Srikant, R. Agarwal & J. Gehrke (2004), "Privacy Preserving Mining of Association Rules", *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining (KDD'02)*, Edmonton, Alberta, Canada, Pp. 217–228.

[25] M.Z. Islam & L. Brankovic (2004), "A Framework for Privacy Preserving Classification in Data Mining", *Proceedings of the 2nd workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalization (AISW'04, AWDM&WI'04, AWSI'04)*, Dunedin, New Zealand, Pp. 163–168.

[26] E. Bertino, I. Fovino & L. Provenza (2005), "A Framework for Evaluating Privacy Preserving Data Mining Algorithms", *Data Mining and Knowledge Discovery*, Vol. 11, No. 2, Pp. 121–154.

[27] J.Z. Zhan, S. Matwin & L. Chang (2005), "Privacy-Preserving Collaborative Association Rule Mining", *DBSec*, Pp. 153–165.

[28] J. Vaidya & C. Clifton (2005), "Privacy-Preserving Decision Trees over Vertically Partitioned Data", *Proceeding of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security (DAS'05)*, Storrs, CT, USA, Pp. 139–152.

[29] J. Natwichai1, X. Li & M. Orlowska (2005), "Hiding Classification Rules for Data Sharing with Privacy Preservation", *Proceedings of 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'05)*, Copenhagen, Denmark, Pp. 468–477.

[30] Xiaodan Wu, Yunfeng Wang, Chao-Hsien Chu, Fengli Liu, Ping Chen & Dianmin Yue (2006), "A Close Look at Privacy Preserving Data Mining Methods", *PACIS 2006 Proceedings*, Paper 32.

[31] Z. Yang, S. Zhong & R.N. Wright (2006), "GrC. Privacy-Preserving Model Selection", *Proceedings of the IEEE International Conference on Granular Computing.*

K. Srinivasa Rao received M.Tech in Computer Science and Engineering from University College of Engineering, JNTU, Kakinada. He received B.Tech in Computer Science & Engineering from JNTU, Hyderabad. And now presently working as Head of the Department, Computer Science & Engineering, Swarna Bharathi College of Engineering, Khammam. His research interests include Data Warehousing and Data Mining, Mobile Computing and Image Processing, Computer Networks.

B. Srinivasa Rao received his M.E. Degree in Computer Science and Engineering from Anna University, Tamilnadu, India and B.Tech degree in Computer Science and InformationTechnology from Dr.Paul Raj Engineering College, Jawaharlal Technological University, A.P, India. At present he is working as Associate Professor in the Department of Computer Science and Engineering, Medha institute of Science and Technology for Women, Khammam. His research interests include data warehousing and mining, information security, networking techniques, formal languages and automatatheory, cloud computing.