ISSN:0975-9646

Privacy Preserving Data Mining

Smita D Patel, Sanjay Tiwari

Computer Science and Engineering,
Rajasthan Technical University

Arya Institute of Engineering and Technology,
Jaipur, India

Abstract— **Through data mining collect large amount of data in many organizations. A key value of huge databases today is technical or financial research. In a huge collection of data there arises a key issue that is privacy. Due to personal interests, medical databases or business interests privacy is needed. Due to privacy infringement while performing the data mining operations this is often not possible to utilize large databases for scientific or financial research. To address this problem, several privacy-preserving data mining techniques are used. The aim of privacy preserving data mining (PPDM) is to extract relevant knowledge from large amounts of data while protecting at the same time sensitive information.**

Keywords—**Data Mining, Cryptography, Secret Sharing, Secure Multi Party Computation, Yao's protocol, Homomorphic Secret Sharing.**

I. INTRODUCTION

Information system must persuade one of the most important properties as Privacy. For this basis, several efforts have been dedicated to incorporating privacy preserving techniques with data mining algorithms in order to prevent the revelation of sensitive information during the knowledge finding. Existing privacy preserving data mining techniques can be classified according to the following five different Dimensions (i) the modification applied to the data (perturbation, substitution, generalization, encryption and so on) in order to sanitize (ii)data distribution (centralized or distributed) (iii) the data type (single data items or complex data correlations) that needs to be protected from disclosure (iv)the data mining algorithm which the privacy preservation technique is designed for(v)heuristic or cryptography-based approaches. cryptography-based algorithms are designed for protecting privacy in a distributed scenario by using encryption techniques while heuristic based techniques are mainly conceived for centralized datasets,. Heuristic-based algorithms just projected aim at defeat sensitive raw data by applying perturbation techniques based on probability distributions. Furthermore, several heuristic-based approaches for hiding both raw and aggregated data through a hiding techniques (k-anonymization, adding noises, data swapping, generalization and sampling) have been developed, first, in the context of association rule mining and classification and, more recently, for clustering techniques.

II. DATA DISTRIBUTION

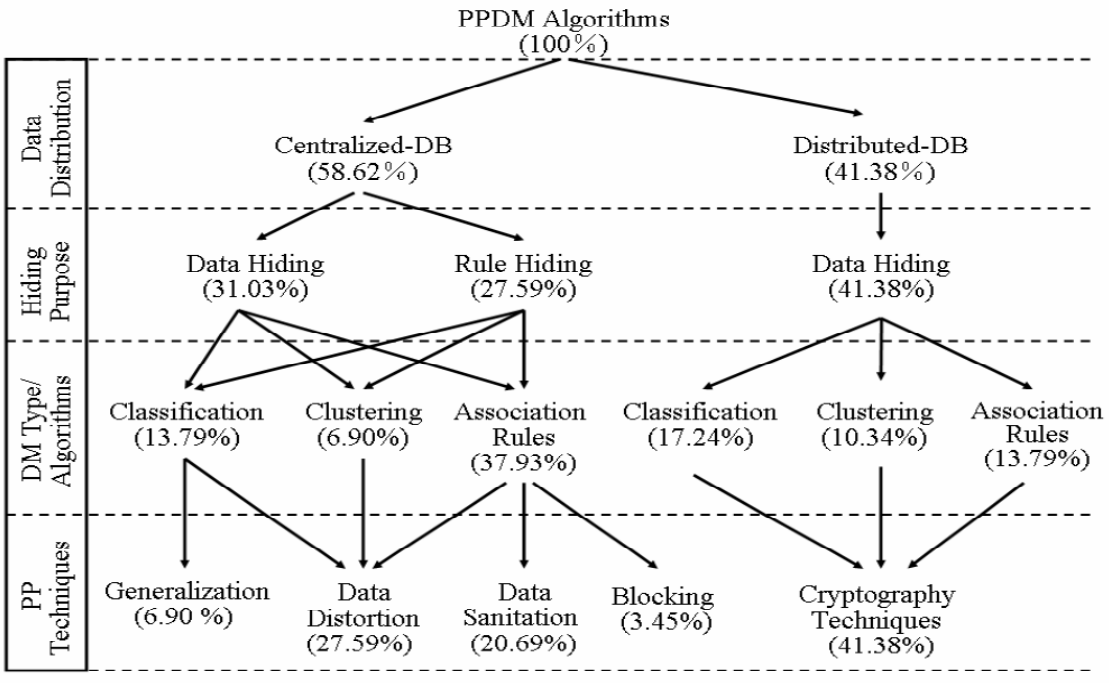

Fig. 1 Taxonomy of PPDM algorithms. [8]

III. THE GENERIC PRIVACY PRESERVING PROBLEM

The problem of learning something without revealing ones own data is not new. It was proposed way back in 1982 by Yao [5]. It has become very important as data has started to grow a million times faster and along with it the demands to keep it for oneself only. When the problem was proposed the web had just come out of its infancy, today it is mature and large and spread to the remotest corners of the world. The authors marked back then that explosive progress in networking, storage, and processor technologies has led to the creation of ultra large databases that record unprecedented amount of transactional information [1].

We are more concerned in privacy preservation with context to data mining algorithms. This is one point where the privacy can be trapped. Suggestion from paper that data mining and data warehousing go hand-in- hand: Most tools operate by gathering all data into a central site, then running an algorithm against that data [6]. However, privacy concerns can prevent building a centralized warehouse—data may be distributed among several custodians, none of which are allowed to transfer their data to another site. It should be noted that what data mining algorithms produce is knowledge, and that data mining results rarely violate privacy, as they generally reveal high-level knowledge rather than disclosing instances of data. However, the concern among privacy advocates is well founded, as bringing data together to support data mining makes misuse easier. The problem is not data mining, but the way data mining is done [7].

1. The Solutions to the problem

PPDM is a new era of research in data mining, where data mining algorithms are analysed for possible infringement in privacy. PPDM research usually takes one of the three philosophical approaches: (1) data hiding, in which sensitive raw data like identifiers, name, addresses, etc. were transformed, jammed, 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) secure multiparty computation, where distributed data are encrypted before released or shared for computations; and (3) rule hiding, in which sensitive knowledge extracted from the data mining process be excluded for use, because private information may be derived from the released knowledge; thus, no party knows no matter which except its own inputs and the results. The crucial goal of PPDM is to develop efficient algorithms that allow one to extract relevant knowledge from a large amount of data, while prevent sensitive data and information from leak or deduction [10].

The third approach can be broadly called cryptographic approach to solve PPDM problems. A very nicely classifies all the proposed solutions into various categories depending on what methods are used in [8]. Various ways to handle PPDM problems including the cryptographic approaches available in [4][3].

2. The Quantifiers of efficient solution

The most general parameters for analysing efficient solution are overall performance in all the areas. A framework is but required to get exact comparative measures. Attempts have been made in past to generalize a framework. One is proposed by Bertino et al. The framework they identified was based on the following evaluation dimensions [9]:

-- Efficiency. The ability of a privacy preserving algorithm to execute with good performance in terms of all the resources implied by the algorithm;

– Scalability. This factor evaluates the efficiency trend of a PPDM algorithm for increasing sizes of the data from which relevant information is mined while ensuring privacy;

– Data quality after the application of a privacy preserving technique. Considered both as the quality of data themselves and the quality of the data mining results after the hiding strategy is applied;

– Hiding failure. The portion of sensitive information that is not hidden by the application of a privacy preservation technique;

– Privacy level offered by a privacy preserving technique. It estimates the degree of uncertainty, according to which sensitive information can still be predicted even if it has been hidden.

Such framework allows one to assess the different features of a privacy preserving algorithm according to a variety of evaluation criteria.

3 Privacy Preserving Techniques

The most used technique yet has been secure multiparty computation. As mentioned earlier, the basic privacy preservation problem is a classical multiparty problem. Cryptography-based SMC has the highest accuracy in data mining and good privacy preservation capability as well; however, it has strict usages as it is only applicable to a distributed data environment [6].

Elements	Computational Cost	Privacy Preserving	Accuracy of Mining	Scalability
Hiding Purpose:				
Data Hiding	Low	Contingent	Contingent	High
Rule Hiding	High	Contingent	Contingent	Low
Data Mining Tasks:				
Classification	Low	N/A	Contingent	High
Clustering	High	N/A	Contingent	Low
Association Rule	Low	N/A	Contingent	High
Privacy Preserving Technique:				
Sanitation	Medium	Medium	Medium	Low
Distortion	Low	High	Low	High
Blocking	Medium	Low	Medium	Low
Generalization	Low	High	Medium	High
Cryptography	High	High	High	Low

Fig. 2 Relative Performance of PPDM components. [8]

According to privacy preservation techniques PPDM algorithm can be further divided. Techniques of PPDM is – 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 repeatedly used item sets such that receptive patterns cannot be mined. The blocking approach replaces certain attributes of the data with a question mark. In this view, the minimum maintain and confidence level will be changed into a minimum period. As long as the support and/or the assurance of a sensitive rule lie below the middle in these two ranges, the secrecy of data is likely to be protected.

People have used data modification, data perturbation, data sanitation, data hiding, and pre-processing 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.

IV. CLASSICAL TECHNIQUES

1. Secure Multiparty Computation

The concept of Secure Multiparty Computation was introduced in [11]. The basic idea of Secure Multiparty Computation is that a computation is secure if at the end of the computation, no party knows anything except its own input and the results. One way to view this is to imagine a trusted third party – everyone gives their input to the trusted party, who performs the computation and sends the results to the participants. It also defines Secure Sum, Secure Set Intersection and Secure set Union techniques [3].

2. Secret Sharing

Secret sharing was introduced by Shamir back in 1979 [2]. The idea is that one party has a secret which it distributes among n other parties in a way that none of the n parties alone can recover the secret. As a matter of fact the secret is shared in a way that the information of at least t of the n parties is needed to recover the secret, where t is a predefined threshold. Any attempt by less than t parties to recover the secret will fail and they will not learn anything about the secret.

3. *Homomorphic Secret Sharing*

Informally speaking, (m, t)-secret sharing is a method to share a secret among m parties in such a way that t−1 or less colluding parties cannot compute any information about the secret; but t arbitrary parties can recover the secret. A data holder that wishes to share his secret s will create m secret-shares $s1,...,s_m$ and send one share to each party.

4. *Homomorphic Encryption*

Let Epk(.) denote the encryption function with public key pk and Dsk(.)denote the ecryption function with private key sk. A public key cryptosystem is called additive homomorphic if it satisfies the following requirements: (1) given the encryption of plaintexts m1 and m2, Epk(m1)and Epk(m2), there exists an efficient algorithm to compute the public key encryption of m1 +m2, such that Epk(m1 +m2):= Epk(m1)+h Epk(m2). (2) given a constant k and the encryption ofm1, Epk(m1), there exists an efficient algorithm to compute the public key encryption of k·m1, such that Epk(k·m1):= k ×h Epk(m1).

5. *Yao's protocol*

Yao first proposed two-party comparison problem and developed provably secure solution. It was extended to multiparty computation by Goldreich [12]

Yao's Millionaire Problem

Essentially the problem is Alice and Bob are two millionaires who want to find out who is richer without revealing the precise amount of their wealth. Multi-party computation has been considered by the theoretical cryptography community for a long time, starting with the pioneering work of Yao [11] in 1986. Yao's garbled circuit construction is relatively simple, and runs in a constant number of rounds. Yao's construction still remains the most attractive choice for generic secure two-party computation.

V. CONCLUSIONS

In this paper present a clear view of current scenario in privacy preserving data mining area from the angle of cryptograpgy The cryptography techniques initiate in their original years have support due to development and their smart use made by researchers has brought new wave of solutions.

Cryptography-based SMC has the highest accuracy in data mining and good privacy preservation capability. The Solutions using conventional cryptography methods have failed to defeat the scalability parameter in their performance evaluation.

REFERENCES

[1] Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. ACM SIGMOD Record, 29(2), 439-450.

[2] Lindell, Y., & Pinkas, B. (2002). Privacy Preserving Data Mining. Journal of Cryptology, 15(3), 177-206.

[3] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y. (2002). Tools for privacy preserving distributed data mining. ACM SIGKDD Explorations Newsletter, 4(2), 28-34.

[4] Pinkas, B. (2002). Cryptographic techniques for privacy-preserving data mining. ACM SIGKDD Explorations Newsletter, 4(2), 12-19.

[5] Yao, A. C. (1982). Protocols for Secure Computations. IEEE.

[6] Kantarcioglu, M., & Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 16(9), 1026-1037.

[7] Vaidya, J., Lafayette, W., & Clifton, C. (2003). Privacy-Preserving K -Means Clustering over Vertically Partitioned Data. ACM SIGKDD '03.

[8] Wu, X., Chu, C.-hsien, Wang, Y., Liu, F., & Yue, D. (2007). Privacy Preserving Data Mining Research : Current Status and Key Issues. ICCS, LNCS 4489(2), 762-772

[9] Bertino, E., Fovino, I. N., & Provenza, L. P. (2005). A Framework for Evaluating Privacy Preserving Data Mining Algorithms. Data Mining and Knowledge Discovery, 11(2), 121-154.

[10] Elisa Bertino,Dan Lin and Wei Jiang.A Survey of Quantification of Privacy Preserving Data Mining Algorithms.ACT.

[11] Yao, A. C.-chih. (1986). How to Generate and Exchange Secrets. Exchange Organizational Behavior Teaching Journal, (1), 162-167

[12] Goldreich, O., Micali, S., & Wigderson, A. (1987). How To Play Any Mental Game or A Completeness Theorem for Protocols with Honest Majority. ACM.