International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013
ISSN 2229-5518

AN EFFICIENT CRYPTOGRAPHIC APPROACH FOR PRESERVING PRIVACY IN DATA MINING

T.Sujitha[1], V.Saravanakumar[2], C.Saravanabhavan[3]

1. M.E. Student, Sujiraj.me@gmail.com
2. Assistant Professor, visaranams@yahoo.co.in
3. Assistant Professor,profbhavan@gmail.com

Department of Computer Science and Engineering, Kongunadu College of Engineering and Technology, Trichy

ABSTRACT

Data mining techniques have been developed in many applications. The goal of data mining process is to extract the information from a large data set and transform it into an understandable format. Data mining is used by many companies with a customer focus on the financial, communication and marketing organization. The privacy preserving data mining (PPDM) is used to extract the relevant knowledge from large amount of data while protecting at the same time sensitive information. In the existing system considered the private property of the corporate data, the data owner transforms the data and ships it into the server for privacy. But there is enough privacy is not found and lacking of computational resources can outsource its mining needs to third party service provider. In this paper, we propose the efficient cryptography techniques are used to protect the privacy to the data owner. By using these techniques, the third parties cannot able to attack the real transaction database. The clustering techniques are used to group the sensitive data from the dataset.

Key words- Clustering, Cryptography, Data Mining, Privacy, Privacy-Preserving Data Mining, Sensitive Data.

I. INTRODUCTION

Data mining techniques have been developed in many applications and researches. But, it also brings the problem of privacy. Privacy is one of the most important properties that an information system. For this reason, several works have been committed to incorporating privacy preserving techniques with data mining algorithms in order to prevent the sensitive data during the knowledge discovery process. When a client sends its database to the server, some sensitive patterns are hidden from its database according to the some specific privacy policies. In recent years, more researchers highlight the seriousness of the problem about privacy.

Privacy preserving is most important problem in data mining techniques. The main challenge of existing data mining algorithm is extracting the data while maintaining the privacy of datasets. Due to this concern, the privacy preserving data mining (PPDM) techniques has been introduced. The main concern in privacy preserving data mining is the sensitive pattern mining. Privacy Preserving Data Mining techniques are used to modify the database through the insertion of false information in order to hide the sensitive information. The problem of existing data mining techniques are improve the privacy but it can be done with increasing cost, computation and overhead will be occur [2,11].

The privacy preserving techniques includes data sanitization, cryptography techniques and access control methods. The data sanitization techniques is the process, which hiding the sensitive information in the data sets. In this method uses the many techniques to provide the privacy to the dataset. The access control methods are used to limit and control the access to host system and applications via a communication links. The cryptography techniques are used to provide a strong security and privacy and also provide a correct and efficient implementation. The cryptography protocols would enable secure communications by addressing the authentication [2]. The Privacy Preserving Data Mining introduce privacy preserving data publishing model and multi-

analyze some related technologies, k-anonymity, relational k-anonymity, 1-diversity and perturbation of privacy preserving.

II. RELATED WORKS

The privacy preserving data mining is the most important research technique in data mining process. The shariq J.Rizvi and Jayant R.Haritsa [11] introduce a MASK scheme (Mining Associations with Secrecy Konstraints), this technique will provide a high degree of privacy and accuracy to the user. It is one of the probabilistic approaches, used to estimate the support of frequent item sets by using association rule mining. In this approach used to extract the hidden knowledge from the large datasets. The Rahat Ali Shah and Sohail Asghar [12] proposed privacy preserving genetic algorithm (PPGA). This algorithm is a modification technique. This technique modifies the database until the support or confidences are drop below the threshold. These techniques are applicable for N binary dataset, small dataset as well as large dataset. The genetic algorithm is used to provide the optimal solution to the complex problem. The problem of outsourcing data mining task within a corporate privacy framework offers a formal protection against information disclosure. It shows that how the data owner can recover the correct data mining results efficiently. It is fully based on the background knowledge; if attacker identifies the background knowledge of the process it will be harmful.

Chunhua Su and Kouichi Sakurai [13] introduce the association rule mining using FP-tree; it will provide the frequent item set, by focus on the privacy. Elisa Bertino, Dan Lin, and Wei Jiang [14] the heuristic based techniques are mainly used for centralized datasets. The cryptography techniques are used for protecting privacy in a distributed database by using encryption techniques and this paper used to evaluate the effectiveness of the PPDM algorithms. In this paper used to identify the privacy level, hiding failure, data quality and complexity [14].

Another related issue is secure multiparty mining over distributed datasets. Data on which mining is to be performed is partitioned, horizontally or vertically, and distributed among several parties. The partitioned data cannot be shared and must remain private but the results of mining on the union of the data are shared among the participants, by means of multiparty secure protocols [11] [1]. They do not consider third parties. This approach partially implements corporate privacy, as local databases are kept private, but it is too weak for our outsourcing problem, as the resulting patterns are disclosed to multiple parties.

W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis introduce the securable process during transactions between third party (outsourcing). It is based on the a one-to-n item substitution cipher techniques for encrypting and decrypting process. The results show that our technique is highly secure with a low data transformation cost. It ran a generic algorithm that has background knowledge about the frequencies of the item sets. It can be easily broken with the help of background knowledge. [4]. I. Molloy, N. Li, and T. Li propose an approach for outsourcing association rule mining. It maps a set of real items into a set of false items, and then maps each transaction non-deterministically. It analyzes both the security and costs associated with outsourcing association rule mining. It indicates that outsourcing association rule mining not be practical, if the data owner is concerned with data confidentiality.

III. PRIVACY MODEL

Let D denote the original TDB that the owner has. To protect the identification of individual items, the owner applies an encryption function to D and transforms it to $D^{'}$, the encrypted database. We refer to items in D as plain items and items in $D^{'}$ as cipher items. We use I to denote the set of plain items and E to refer to the set of cipher items.

IV. PRIVACY PRESERVING DATA MINING METHODS

In this section describes the brief introduction of privacy preserving data mining methods.

A. Privacy preserving in data mining methods

1. The k-anonymity method

An important method for privacy de-identification is a method of k-anonymity [1]. The motivating factor behind the k-anonymity technique is that many attributes in the data can be considered pseudo-identifiers, which can be used in conjunction with public records in order to uniquely identify the records. For example, if the identifications from the records are removed, attributes such as the birth date and zip code can be used in order to uniquely identify the identities of the underlying records. The idea in k-anonymity is to reduce the granularity of representation of the data in such a way that a given

record cannot be distinguished from at least (k - 1) other records.

2. The Randomization Method

The randomization technique uses data distortion methods in order to create private representations of the records. In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be used for data mining purposes. Two kinds of perturbation are possible with the randomization method:

Additive Perturbation:

In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re designed to work with these data distributions.

Multiplicative Perturbation:

In this case, the random projection or random rotation techniques are used in order to perturb the records.

3. Cryptographic methods for Information Sharing and Privacy

In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores. This requires secure and cryptographic protocols for sharing the information across the different parties. The data may be distributed in two ways across different sites:

Horizontal Partitioning:

In this case, the different sites may have different sets of records containing the same attributes.

Vertical Partitioning:

In this case, the different sites may have different attributes of the same sets of records.

B. The encryption method for privacy preserving data mining:

The growth of Internet has triggered tremendous opportunities for distributed data mining, where people jointly conducting mining tasks based on the private inputs they supplies. These mining tasks could occur between mutual un-trusted parties, or even between competitors, therefore, protecting privacy becomes a primary concern in distributed data mining setting. Distributed privacy preserving data mining algorithms require collaboration between parties to compute the results or share no-sensitive mining results, while provably leading to the disclosure of any sensitive information.

In general, distributed data mining involves two forms: horizontally partitioned data and vertically partitioned data. Horizontally partitioned data means that each site has complete information on a distinct set of entities, and an integrated dataset consists of the union of these datasets. In contrast, vertically partitioned data has different types of information at each site; each has partial information on the same set of entities. Most privacy preserving distributed data mining algorithms are developed to reveal nothing other than the final result. Kantarcioglu and Clifton [2] studied the privacy-preserving association rule mining problem over horizontally partitioned data. Their methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task. Lindell et al. researched how to privately generate ID3 decision trees on horizontally partitioned data.

V. SYSTEM ARCHITECTURE

In this paper, our goal is to devise an encryption scheme which enables formal privacy guarantees to be proved, and to validate this model over large-scale real-life transaction databases (TDB). The architecture behind our model is illustrated in Fig. 1.

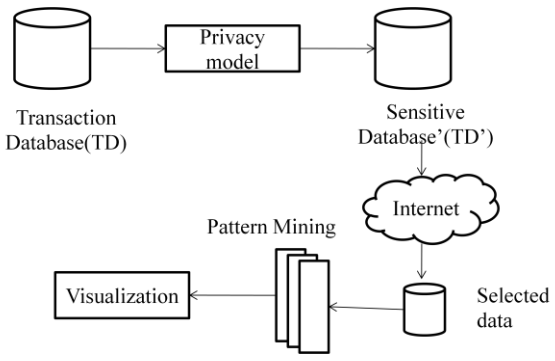

Fig. 1.System Architecture

We make the following contributions. First, to identify the user-defined sensitive items and grouped into the sensitive database (D'). This sensitive

database D' further processed by data mining technique to evaluate the patterns.

Second, we develop an cryptography techniques are used for key generation by using this key the plain text will be converted to cipher text.

Third, to allow the Encryption/Decryption technique, it is the process of transforming transaction database (TDB) TD into its sensitive database TD'.

Finally, pattern discovery module, there is recover the original patterns from the extracted pattern received from the data owner.

VI. ENCRYPTION/DECRYPTION SCEHEME

The encryption/decryption scheme is the process of transforming transaction database TD into its sensitive database TD'. It has been encrypted by using cryptography techniques for each plain item. The privacy preserving module, there is hash functions are used. It is used to efficient storage and fast retrieval of items. The main use of hash function is maps n keys to n integers.

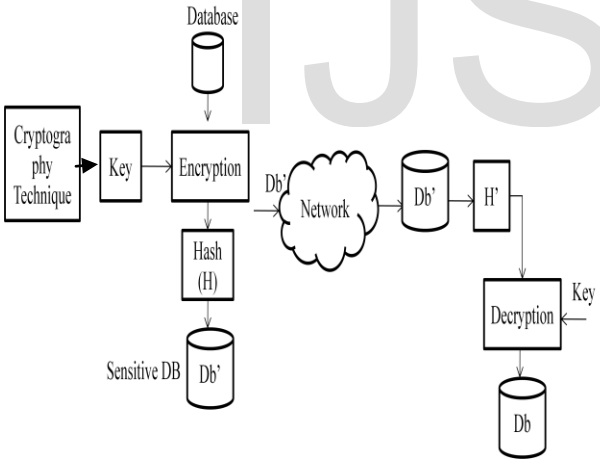

Fig.2 encryption and decryption

VII. ASSOCIATION RULE MINING

The association rule mining finds interesting association or correlation relationships among a large set of data items. The association rules are considered interesting if they safety both a minimum support and a minimum confidence. Let I= $\{i_1, i_2, \ldots, i_m\}$ be a set of items. Let D be the task-relevant data, be a set of database transaction where each transaction T is a set of items such that T⊆I. Each transaction T is associated with an identifier TID. Let A be the set of items. An association rule is implication of the form A⇒B, where A⊆I and B⊆I and A∩B=ϕ. The rule A⇒B holds in the transaction set D with support S, where S is the percentage of transactions in D that contain A∪B. This is taken to be the probability P(A∪B). The rule A⇒B has confidence C in the transaction set D if C is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability P (A/B). That is,

$$\text{Support } (A \Rightarrow B) = P(A \cup B)$$

$$\text{Confidence } (A \Rightarrow B) = P(A \cap B)$$

A set of items is referred to as an item set (pattern). An item set that contains k-items is a k-item set. For example the set {Butter, Bread} is a 2-itemset. An item set satisfies minimum support if the occurrence frequency of the item set is greater than or equal to the product of minimum support and the total number of transactions in D. The number of transactions required for the item set to satisfy the minimum support count. If an item set satisfies the minimum support, then it is said to be frequent item set.

VIII. CONCLUSION AND FUTURE WORKS

In this paper, we study the problem of privacy-preserving mining of frequent patterns on an encrypted outsourced TDB. We proposed an encryption scheme to avoid attacks in cloud databases and also help to secure it as it is often difficult to physically secure all access to networks. Unlike previous works, such as [11] and [12], we formally proved that our method is robust against an adversarial attack based on the original items and their exact support. In addition for mining, it is must to cluster different datasets to derive different patterns. Our experiments based on both large real and artificial datasets yield strong evidence in favor of the practical applicability of our approach.

IX. REFERENCES

[1]. Samarati P., Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression. IEEE Symp. On Security and Privacy, 1998.

[2]. Murat Kantarcioglu, Chris 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, 2004.

[3]. P. K. Prasad and C. P. Rangan, "Privacy preserving birch algorithm for clustering

over arbitrarily partitioned databases," in *Proc. Adv. Data Mining Appl.*, 2007, pp. 146–157.

[4]. W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "Security in outsourcing of association rule mining," in *Proc. Int. Conf. Very Large Data Bases*, 2007, pp. 111–122.

[5]. Lalanthika Vasudevan, S.E. Deepa Sukanya, N. Aarthi," Privacy Preserving Data Mining Using Cryptographic Role Based Access Control Approach", in Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I IMECS 2008.

[6]. Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo," A Survey on Privacy Preserving Approaches in Data Publishing", in First International Workshop on Database Technology and Applications,2009.

[7]. Tai, P. S. Yu, and M. Chen, "K-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining," in *Proc. Int. Knowledge Discovery Data Mining*, 2010, pp. 473–482.

[8]. Nishant Doshi," A novel approach for cryptography technique on Perturbed data for Distributed Environment", in International Journal on Cryptography and Information Security (IJCIS),Vol.2, No.3, September 2012.

[9]. Shweta Shrma Hitesh Gupta Priyank Jain," A Study Survey of Privacy Preserving Data Mining", in International Journal of Research & Innovation in Computer Engineering , Vol 2, Issue 2 , April 2012.

[10]. Alex Gurevich, Ehud Gudes," Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation".

[11]. Shariq J. Rizvi, Jayant R. Haritsa, Maintaining Data Privacy in Association Rule Mining.

[12]. Rahat Ali Shah1, Sohail Asghar," Privacy preserving in association rules using genetic algorithm".

[13]. Chunhua Su_ and Kouichi Sakurai*,"*A Distributed Privacy-Preserving Association Rules Mining Scheme Using Frequent-Pattern Tree".

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

[15]. S.M. Mahajan and A. K. Reshamwala," Data Mining Ethics in Privacy Preservation - A Survey", in International Journal of Computer Theory and Engineering, Vol. 3, No. 4, August 2011.