Cryptographic Approach For Privacy Preservation in DM

Brijal Patel[1], Dimple Kanani[2]
[1, 2] IT department
[1, 2] Vadodara Institute of Technology, Vadodara, India

Abstract- *Data governs the present world. Data mining has become a vital tool among many industries and organizations for effective decision making. Data mining involves extracting patterns from large data sets to form an understandable structure for future use. In recent years, sharing of data between organizations has taken place for their development. Even though sharing of data has an advantage, data which contains sensitive attributes pose a threat of exposing private information. Thus privacy of the sensitive attributes is of a greater concern. This work aims at preserving privacy of sensitive attributes by using two modified cryptographic techniques namely, Rail Fence and Vigenere Cipher algorithms. The unique feature of this approach is that the encryption matrix is formed from the original data itself. From the key matrix, modified data . From our experimental results it is evident that the original data cannot be inferred from the modified.*

Keywords- non-computer professional, Data Structure, algorithm

I. 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. Privacy preserving data mining encryption methods:

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.

II. ARCHITECTURE OF SYSTEM

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.

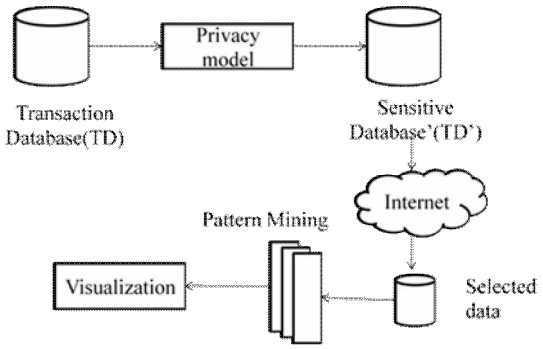

Figure 1.

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.

III. ENCRYPTION AND DECRYPTION SCHEME

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 andstorage and fast retrival of items.The main use of hash function is maps n keys to n integers associated with an identifier TID. Let A be the set of items. An association rule is implication of the form $A \Rightarrow B$, where $A \subseteq I$ and $B \subseteq I$ and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with support S, where S is the percentage of transactions in D that contain $A \cup B$. This is taken to be the probability $P(A \cup B)$. The rule $A \Rightarrow B$ 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, Support $(A \Rightarrow B) = P (A \cup B)$ 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.

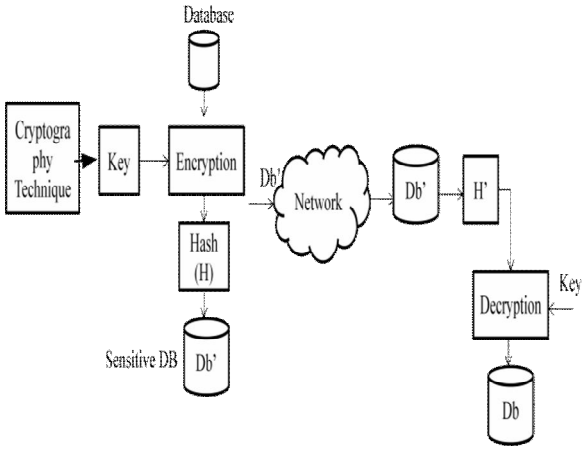

Figure 2.

IV. 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, 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.

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] Brijal Patel H ,Ankur N shah, "Privacy Preserving in DM using min-max normalization and Noise addition", International journal of advanced Engineering and research Development, Vol. 2, Issue 10, October-2015.