A Secure Scheme for Privacy Preserving Data Mining Using Matrix Encoding

Jiawei Han and Micheline Kamber

Department of Computer Applications, Coimbatore Institute of Technology, Coimbatore, Pincode- 641014, Tamilnadu, India
Department of Physics, Coimbatore Institute of Technology, Coimbatore, Pincode- 641014, Tamilnadu, India.

A B S T R A C T

Organizations are generating data in large amount which is being subjected to analysis to extract useful information out of them. Also there is multiple storages on disks and transfer of data before applying data mining algorithms on them. Preserving the privacy of the dataset as well the mined results has become very important for the data owners. In this paper we suggest a novel scheme to protect the data at rest before it moves for analysis. It uses a special form of encoding using the concept of matrix theory. The encoded data matrix is assessable to only the authorized data miner who has the key for the encoder matrix. The goal is to protect the dataset without affecting the mining results. Experiments were conducted on synthetic and real datasets and it proves that this simple method is effective in preserving the privacy of data against the usual methods of data encryption.

© 2015 AENSI Publisher All rights reserved.


INTRODUCTION

Huge amount of data is generated in various forms in organizations. Many of them are into in-house data analytics where data is subjected to various analysis techniques to discover hidden knowledge. Data get stored on disks, moved to back-ups or uploaded to servers for analysis. It then becomes necessary to protect the database before it moves for analysis. Data mining discovers hidden patterns from large datasets by applying techniques and methods from areas of machine learning, computer science, mathematics and statistics (Jiawei Han and Micheline Kamber, 2006). Huge amount of data is generated in various forms in organizations. Many of them are into in-house data analytics where data is subjected to various analysis techniques to discover hidden knowledge. Data get stored on disks, moved to back-ups or uploaded to servers for analysis. It then becomes necessary to protect the database before it moves for analysis. Data mining discovers hidden patterns from large datasets by applying techniques and methods from areas of machine learning, computer science, mathematics and statistics (Jiawei Han and Micheline Kamber, 2006).

Explosive progress in networking, storage, and processor technology has led to the creation of ultra-large databases that record unprecedented amount of transactional information. The main problem is that with the availability of non-sensitive information or unclassified data, one is able to infer sensitive information that is not supposed to be disclosed. Despite its benefits in various areas such as marketing, business, medical analysis, bioinformatics and others, data mining can also pose a threat to privacy. The database security if not done or used properly. This paper focuses on protecting the database before it goes for mining in a centralized environment. In paper we suggest a novel scheme for encoding the database using matrix termed as matrix encoding. It can be very well extended to a distributed environment also. Many researchers proposed many methods for privacy preserving mining for both centralized and distributed databases. The state of the art in the area of privacy preserving data mining techniques is discussed by the authors in (Agrawal, D. and C. Aggarwal, 2001). This paper also describes the different dimensions of preserving data mining techniques such as data distribution, data modification techniques, data mining algorithms, data or rule hiding and approaches for privacy preserving data mining techniques. The survey of basic paradigms and notations of secure multiparty computations and their relevance to the field of privacy preserving data mining are presented by the authors in (Verykios, V.S., 2004). They also discussed the issue of efficiency and demonstrate the difficulties involved in constructing highly efficient...
protocols. In (Lindell, Y. and B. Pinkas, 2009), the authors proposed a framework for evaluating privacy preserving data mining algorithms and based on their framework one can assess the different features of privacy preserving algorithms according to different evaluation criteria. In (Elisa Bertino, 2005), Secure mining of association rules over horizontally partitioned database using cryptographic technique to minimize the information shared by adding the overhead to the mining process is presented. In (Kantarcioglu, M. and C. Clifton, 2002), authors proposed an enhanced scheme two-phase privacy preserving in distributed data mining. They presented two protocols to increase the security against collusion in the communication environment with or without trusted party. A new algorithm which is the modification of the existing algorithm and based on a semi-honest model with negligible collision probability is proposed in (Chin-Chen Chang, 2006). They also used cryptography techniques to preserve the privacy. Privacy preserving in data mining by using cryptographic role based access control is presented in (Mahmoud Hussein, 2008). They proposed a new solution by integrating the advantages of the first approach which protects the privacy of the data by using an extended role based access control approach and the second approach which uses cryptographic techniques with the view of minimizing loss of information and privacy. In (Lalanthika Vasudevan, 2008), authors addresses the problem of association rule mining in vertically partitioned database by using cryptography based approach and also presents the analysis of security and communication.

Proposed Framework:

In this paper we propose a novel scheme to protect the dataset. Figure 1 shows the architecture of the proposed system. There are two phases in this model. First phase is termed the Encoding Phase where the data owner encodes the dataset. The second phase is the Decoding Phase, where only the authorized data miner can perform the decoding of the dataset and apply the mining algorithm. The dataset prepared for analysis has attributes that are sometimes very sensitive and the feature selection process also selects the relevant attributes for data analysis. This dataset cannot be exposed to any unauthorized personnel in the organization. The proposed system enables the data owner to protect the dataset from being accessed by malicious intruders. Only authorized personnel who has the key to unlock the analysis dataset can perform the data mining operations. This scheme can be very easily adopted in any organization as a internal security and privacy measure.

![Architecture of Proposed System](image)

**Fig. 1: Architecture of Proposed System.**

**Encoding Phase:**

Here the dataset is first partitioned into sub-datasets. Any kind of partition horizontal or vertical can be applied to reduce the dataset into manageable units according to the memory and computation resources available. Then each partition is converted into a matrix form using any numerical encoding method. The algorithmic steps for matrix encoding used in this scheme are given below.

**Algorithm for Matrix encoding and decoding:**

Input: \{D_i | i = 1 to num_partitions\}

Output= \{X_i\} // encoded matrices

**Encoding**

1: Choose a Quadratic Form (QF) of n variables to form the encoded matrix
2: Derive the induced diagonal matrix from QF, E.
3: Encode as \(X = E.D\) //matrix multiplication
4: Encrypt the QF //key

**Decoding**

5: Decrypt the QF
6: Identify the matrix and its inverse
7: Perform \(D = X.E^{-1}\)
8: Apply mining algorithm on D
Here an arbitrary non-singular matrix is chosen for encoding the given dataset. For example, consider a matrix

\[
\begin{bmatrix}
5 & -1 & 4 \\
-6 & 5 & 7 \\
11 & 8 & 0 \\
3 & 2 & 0 \\
\end{bmatrix}
\]

An arbitrary non-singular matrix A is chosen for decoding the data matrix. The inverse of which is used for decoding.

\[
A = \begin{bmatrix}
1 & 5 & -1 \\
-2 & 11 & 7 \\
1 & 5 & 2 \\
\end{bmatrix}
\]

whose inverse is given by, \(A^{-1} = \begin{bmatrix}
5 & -5 & 4 \\
1 & -1 & 9 \\
-1 & 0 & -1 \\
\end{bmatrix}\)

Decoding Phase:

In the decoding phase, the inverse of matrix is used to derive the original dataset as described below.

![Matrix Image]

Experimental Results:

The characteristics of the dataset used are shown in Table 1. Here two synthetic and two real datasets from FIMI repository are used for studying the encoding time of the data matrix. The experiments were focused on studying the encoding time for the dataset. The storage requirements were also studied using the horizontal and vertical partitions. All the experiments were run on a 4GB machine with 400GB hard disk running Windows 7 Home Premium Edition. The proposed method is tested against the known content based encryption adopted in all database servers. Figure 2 shows the time taken for encoding each partition using the matrix based approach and the usual content encryption. There is significant difference decrease in time for the proposed method because matrix operation in memory is faster compared to encryption and decryption. Also the storage requirements in terms of partition size are manageable as shown in Table 1.

Table 1: Dataset Characteristics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Distinct Items</th>
<th>#N</th>
<th>T_{run}</th>
</tr>
</thead>
<tbody>
<tr>
<td>TD100K</td>
<td>1000</td>
<td>100000</td>
<td>31</td>
</tr>
<tr>
<td>TD1000</td>
<td>1000</td>
<td>1000</td>
<td>19</td>
</tr>
<tr>
<td>BMSWebView1</td>
<td>267</td>
<td>59602</td>
<td>267</td>
</tr>
<tr>
<td>BMSWebView2</td>
<td>3340</td>
<td>77512</td>
<td>161</td>
</tr>
</tbody>
</table>

Table 2: Partition Size for Datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>M</th>
<th>N</th>
<th>#P_Size</th>
<th>Memory/Partition (KB)</th>
<th>Total Memory(MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS 1</td>
<td>500</td>
<td>267</td>
<td>120</td>
<td>521.48</td>
<td>61.11</td>
</tr>
<tr>
<td>DS 4</td>
<td>500</td>
<td>161</td>
<td>155</td>
<td>314.45</td>
<td>47.59</td>
</tr>
<tr>
<td>DS 3</td>
<td>500</td>
<td>31</td>
<td>200</td>
<td>60.54</td>
<td>11.82</td>
</tr>
<tr>
<td>DS 2</td>
<td>500</td>
<td>19</td>
<td>1</td>
<td>37.11</td>
<td>0.072</td>
</tr>
</tbody>
</table>

Figure 2 Dataset vs encoding Time

![Graph Image]

Fig. 2: Performance of proposed system with Standard content encoding.
Conclusion:
The experimental results prove that the time taken for encoding is significantly less for larger matrices. Also the two way protection scheme enables to protect the dataset from intruders. There is no one standard metric to quantify privacy. Here we assume the adversary-knowledge model. What knowledge will an intruder possess in order to gain access of the original dataset? Anytime the dataset is in encoded form. Let us assume that the intruder is able to gain access to the encoder matrix. This is possible only if the intruder is able to know the matrix itself. Here we encrypt the matrix form and access of the decryption key is given to only authorized persons. This method is a simple effective solution for protecting the dataset for data mining. As further extension we are to study how the model behaves when big data in distributed environment is given as input.

REFERENCES


