A HYBRID APPROACH FOR PRIVACY PRESERVING DATA MINING

Tatshini G.¹, Shaili Jha¹ and R. Devika²
¹B.Tech Computer Science Engineering, SASTRA University, Thanjavur, India
²Computer Science Engineering, School of Computing, SASTRA University, Thanjavur, India
E-Mail: shaili_jha@yahoo.co.in

ABSTRACT

Privacy preservation has become a major issue during data publishing. The simplest solution is not to disclose the information. This would be of no use since it will hinder the process of data analysis. Instead, the data can first be modified so that it can guarantee privacy and, at the same time, it retains sufficient utility and can be released to other parties safely. The existing system uses a hybrid approach of sampling and generalization for data modification. An alternative system using crossover and mutation of genetic algorithm for data modification, without the use of a trusted third party is proposed. In addition, shearing technique and double pass matrix based encryption technique are implemented to improve the confidentiality of the data.

Keywords: privacy preservation, crossover, matrix based encryption, shearing, privacy preserving data publishing (PPDP).

1. INTRODUCTION

Data publishing is the process of making data, collected by various institutions such as hospitals, financial institutions, government etc. public, so that it can be used to find useful patterns for the purpose of data mining. But the growing concern with this is the violation of the individual’s privacy. When seen in isolation, the individual’s identity may not be found out. But this can be linked with other publicly available documents like voter list and the sensitive information of an individual may be compromised. One approach to overcome this is to apply transformation techniques to the sensitive attributes in a manner that the data published will be useful for analysis while still protecting the privacy. A term closely related to this is privacy preserving data mining (PPDM) that refers to development of data mining techniques that have privacy concerns incorporated. These involve data hiding, rule hiding etc. Data perturbation is a specific class of data hiding techniques.

In a typical scenario, there are three actors involved. They are: Record Owner, Data Publisher and Data Recipient. The interactions between them are depicted below:

![Figure-1. Representation of PPDM.](image)

2. METHODOLOGY

The data to be perturbed are given to the Data publisher as depicted. The Record owner is given the freedom to apply the three proposed techniques namely, Crossover, Double pass Matrix based Encryption and Shearing, to the numerical attributes. Crossover is applied to a quasi identifier that has low tolerance to deviation, Double pass Matrix based Encryption to the unique identifier and Shearing to any other numerical quasi identifier. Double pass Matrix based Encryption and Shearing techniques are applied not to whole value of the attribute as such. Instead, a part of the value of the attribute that might reveal sensitive information is taken into consideration. This ensures that data analysis can still be carried out in an efficient manner.

2.1 Crossover

Crossover is one of the operators applied in genetic algorithms. It is a twostep process. After reproduction, members of the resulting population are mated at random. These are then converted into binary strings. After this, an integer position is chosen at random and the bits at that position are swapped. This gives rise to two new strings.
Figure-2. A typical example for Crossover.

Mutation is another operator of genetic algorithm. In this, a position is chosen at random in the string and its bit value is changed. They are depicted below:

Figure-3. A typical example of Mutation.

A combination of these two techniques is adopted in the proposed system. Two attribute values (field chosen by the user to which crossover is to be applied) are selected in a sequential manner and then converted into binary values. Then the bits are paired up. A number is generated and the bits in that particular position are swapped. Mutation is applied in the following three cases:

a) When the number of records are odd in number, mutation alone is applied to the first record

b) When the bits chosen for crossover have the same value and crossing over will not make any change to the original data

c) When the value of the attribute exceeds the threshold. In such a case, the higher order bits are mutated (unlike the other two cases wherein lower order bits are taken into consideration) until the attribute’s value is brought back into the specified range

2.2 Double pass matrix based encryption

Encryption is a cryptographic technique in which the message (referred to as the plaintext) is encoded using an algorithm, converting it into an indecipherable cipher text. The encryption algorithm typically involves the usage of an encryption key that defines how the message is to be encoded.

The steps followed in Double pass matrix based encryption are as follows:

a) The sensitive information is first extracted from the specific record and stored in an array

b) A number is generated at random (range of random number chosen in a way that it does not exceed the size of the database) and similarly the sensitive information of the record indicated by the random number is extracted and stored in another array

c) The matrix to be used for the process of encryption is generated next. It is a 10 x n matrix (‘n’ depends on the number of digits of the attribute decided on them being the sensitive information) generated in the following manner:

- The sensitive information is split up into individual digits
- A row is generated based on the individual digit, in a serial manner, by incrementing it by one, until all the digits from 0-9 are covered.
  - For example: If the digit extracted is 6, then the row generated would be 6 7 8 9 0 1 2 3 4 5
- The process is repeated n times

d) The position of the digit to be encrypted is found in the particular row. This value is used to encrypt the original data. This is shown in the following example:

Consider the following banking database:

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Name</th>
<th>Age</th>
<th>Gender</th>
<th>Account number</th>
<th>Pin code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>X</td>
<td>23</td>
<td>M</td>
<td>1798000015058</td>
<td>613401</td>
</tr>
<tr>
<td>2.</td>
<td>Y</td>
<td>25</td>
<td>F</td>
<td>1798000016330</td>
<td>500016</td>
</tr>
<tr>
<td>3.</td>
<td>Z</td>
<td>34</td>
<td>M</td>
<td>1798000067451</td>
<td>625006</td>
</tr>
<tr>
<td>4.</td>
<td>A</td>
<td>40</td>
<td>F</td>
<td>1798000040001</td>
<td>600028</td>
</tr>
<tr>
<td>5.</td>
<td>B</td>
<td>27</td>
<td>M</td>
<td>179800032897</td>
<td>530032</td>
</tr>
<tr>
<td>6.</td>
<td>C</td>
<td>55</td>
<td>M</td>
<td>179800021456</td>
<td>613004</td>
</tr>
</tbody>
</table>
Assume the account number is to be encrypted and the last five significant digits are taken as the significant digits. Then, another account number is taken at random and its significant digits are used as a key for encrypting the account number in question. This is done in the following format:

**Account number to be encrypted:**
17980000015058  Significant digits: 15058

**Random number generated:** 6

**Account number chosen as key:**
1798000021456  Significant digits: 21456

**e)** The entire process is repeated twice to ensure that the original sensitive data cannot be traced back. Hence the name double pass matrix based encryption. Using the same example as reference, the second pass is shown below:

**Account number to be encrypted:**
1798000005713  Significant digits: 05713

**Random number generated:** 2

**Account number chosen as key:**
1798000016330  Significant digits: 16330

A similar approach is adopted in the proposed system. The formula used for shearing is:

\[
X' = X + (\text{Shearing\_factor} \times X)
\]

Where \(X'\) is the modified data and \(X\) is the original data. Shearing\_factor is the random number generated in the range \([2-9]\)

The steps are as follows:

a) The specified digits (which contain sensitive information) are extracted from the field and stored in a variable

b) A random number is generated between \([2-9]\)

c) This is multiplied with the extracted number and added to the original data and its average is taken

d) Modulo operation with \(10^n\) (where \(n\) denotes the number of digits extracted from the original data) is performed for normalization in cases where in the number of digits exceeds that of the original data

e) The original data is replaced with this modified data

### 2.3 Shearing

Shearing is one of the graphics-based transformation techniques. In this, distortion of the object happens in a way that it appears as if the object were comprised of internal layers that had been caused to slide over each other. It may deform the shape of the object along either one of the axes: x-axis, y-axis or both.

#### 2.3.1 Matrix form of shearing along X axis in 3D:

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
S_{h_{zx}} & S_{h_{zy}} & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

#### 2.3.2 Matrix form of shearing along Y axis in 3D:

\[
\begin{bmatrix}
1 & S_{h_{zy}} & S_{h_{yx}} & 0 \\
S_{h_{zx}} & 1 & S_{h_{yx}} & 0 \\
S_{h_{zx}} & S_{h_{zy}} & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

A User Interface has been created to facilitate the application of the proposed technique on the desired database.
As it can be seen, the user can choose the path of the specified database by clicking on the Browse button. All the tables present in the given database will be displayed and the user can accordingly choose the required table. On selecting the table, all the fields in it will become available to help the user choose on which specified field he/she would like to apply the specific technique to.

On clicking the Preserve Privacy button, the specified techniques are applied to the selected fields. The changes are accordingly made in the database. In addition, a sample of the original data and the modified data are displayed in the user interface for the user to view the differences.
4. ANALYSIS

Performance Analysis is done to evaluate the loss of data in applying the hybrid approach. Bias in mean (BIM) and Bias in Standard deviation (BIS) for each technique is found.

\[
BIM = \frac{\text{average of perturbed data} - \text{average of original data}}{\text{average of original data}}
\]

\[
BIS = \frac{\text{standard deviation of perturbed data} - \text{standard deviation of original data}}{\text{standard deviation of original data}}
\]

Lower the value of BIM and BIS better the quality of the technique.

Here is the value run on sample data and the graph for those values.
The values of BIM and BIS keep fluctuating because it depends on the values of the attributes present in the database. In spite of this, the value is low thus ensuring negligible data loss and high level of privacy.

**BIM and BIS values for encryption**

![Graph plotted against BIM and BIS values vs different sizes of dataset for Crossover.](image)

**Figure-9.** Graph plotted against BIM and BIS values vs different sizes of dataset for Crossover.

<table>
<thead>
<tr>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Dataset_size</td>
<td>BIM</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>0.126056</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>-6.03E-05</td>
</tr>
<tr>
<td>11</td>
<td>25000</td>
<td>0.000325</td>
</tr>
<tr>
<td>12</td>
<td>50000</td>
<td>0.34567</td>
</tr>
<tr>
<td>13</td>
<td>100000</td>
<td>-3.23E-02</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure-10.** Table representing BIM and BIS values for encryption on different sizes of dataset.

![Graph plotted against BIM and BIS values vs different sizes of dataset for Encryption.](image)

**Figure-11.** Graph plotted against BIM and BIS values vs different sizes of dataset for Encryption.

The values of BIM and BIS keep fluctuating because it depends on the values of the attributes present in the database. In spite of this, the value is low thus ensuring negligible data loss and high level of privacy.

**BIM and BIS values for Shearing:**

![Table representing BIM and BIS values for Shearing on different sizes of dataset.](image)

**Figure-12.** Table representing BIM and BIS values for Shearing on different sizes of dataset.
The values of BIM and BIS keep fluctuating because it depends on the values of the attributes present in the database. In spite of this, the value is low thus ensuring negligible data loss and high level of privacy.

5. CONCLUSIONS

This paper explains the hybrid approach for privacy preservation during the process of data publishing. Three major techniques are applied, crossover, double pass matrix based encryption and shearing. In crossover, two attribute values (field chosen by the user to which crossover is to be applied) are selected in a sequential manner and then converted into binary values. The bits are paired up. A number is generated at random and the bits in that particular position are swapped. In addition, mutation is applied if needed. During double pass matrix based encryption, the significant digits of the chosen attribute are extracted, a corresponding matrix is generated in a sequential manner, and the data is encrypted by finding its analogous position in the matrix. This process is repeated twice. In shearing, the significant digits are extracted and sheared (similar to the shearing process of graphics) by multiplying it with a random number. An average of the sheared and the original data is taken and the original data is replaced with the modified data. Further modification that can be implemented for the proposed system in the future is that the technique can be altered to include all data types.

REFERENCES


International Journal of Advanced Research in Computer and Communication Engineering. 2(8).

Sridhar Mandapati, Dr. Raveendra Babu Bhogapathi and Ratna Babu Chekka. 2013. A Hybrid Algorithm for Privacy Preserving in Data Mining. I.J. Intelligent Systems and Applications, 2013, 08, 47-53 Published Online July 2013 in MECS.

Li Liu, Murat Kantarcioglu, Bhavani Thuraisingham. 2007. The applicability of the perturbation based privacy preserving data mining for real-world data. Published Online.


