Privacy Preserving Data Mining Using Random Rotation Based Data Perturbation Technique

AshaKathariya1Pankajkawadkar2
1Department of Computer science & Engineering 2H.O.D. computer science dept., PIES, Bhopal, Patel Institute Of Engineering & Science, Bhopal, India

Abstract---To preserve privacy of data, privacy preserving data mining is the study of valid mining patterns and models which mask private information. There are many privacy preserving data mining techniques which have been studied. One crucial concept about existing data mining privacy preserving techniques is suitable and designed for static databases and not suitable for data streams. Recently, data streams are introduced as new type of data which are different from traditional static data. Various features of data streams are: with time, data distribution changes constantly; data is having time preferences; amount of data is extensive; flow of data with fast speed; requirement of immediate response. When has been modified, it would be necessary to recompute whole database, so it leads to more computation time and inability to respond the user fastly. Further, it is observed that accuracy of data is decreases when transformation is carried out on data. So, there has been need to develop the system which preserve privacy along with accuracy. So privacy preserving on data stream mining is very crucial issue.

Keywords: Privacy, Data Streams, K-means clustering

1. INTRODUCTION

In recent years, data mining is shown as a powerful data analysis tool and has made remarkable contributions in many areas and has wide applications viewpoint. With the development of database technology and network technology, a large number of fulldata, which contains much individual privacy information, has been amassed in various fields, such as patient’s condition information, preferences to customer, personal background information, account information, etc. Once the information leaked, it will be unsafe to individual. If they give the actual data directly to the prospectors, it will predictably produce private information disclosure. As the field of data mining technology extending, privacy disclosure problem becomes worse, e.g., using the attention of phases of industry and social, so, how to do datamining under the circumstances of privacy preserving becomes a hot spot in datamining, so privacy preserving datamining (PPDM) is introduced.

Securing against unauthorized accesses has been a long-term objective of database security, the government, research statistical agencies and research community. Solutions to such a difficulty require combining several techniques and mechanisms. In a situation where data have different sensitivity stages, this data may be classified at different levels, and it has made available only to those subjects with an appropriate consent.

Technique of Clustering is a well-known and fundamental problem in statistics and engineering, namely, how to arrange a set of measurements into a number of clusters. Clustering is an significant area of application for a variety of areas including datamining, vector quantization and statistical data analysis. The problem has been framed in various ways in the pattern recognition, machine learning, optimization and statistics literature. The fundamental clustering problem is that of assemble together data items that are related to each other. Given a set of data items, clustering algorithms can group analogous items together. Clustering has many applications, such as analysis of customer behavior, targeted marketing, forensics, and bioinformatics.

By mining sensitive characteristics from the original database, Reconstruction-based approaches generate privacy aware database. These approaches have been generated and side effects in database, an heuristic approach. Reconstruction-based methods perturb the original data to achieve privacy preserving. The perturbed data will meet the conditions of privacy. First, an attacker cannot determine which data from the issue of the original data. Secondly, the altered data still preserving some statistical properties of the original data, namely, some of the information derived from the partial data are equivalent to data acquired from the original information. Perturbing the data or preserving privacy is very fertile technique used by many researchers. It is also capable to reconstruct the distribution at a cumulative level and perform the mining.

@IJAERD-2015, All rights Reserved 231
There are three types of data perturbation approaches: Rotation Perturbation, Projection Perturbation, and Geometric Data Perturbation.

II. ROTATION PERTURBATION

In order to preserve data privacy, in rotation perturbation method, the data matrix is multiplied by a random rotation matrix before publishing. There is one advantage of this approach: it can preserve the geometric properties of the data matrix, so even categories of classifiers which are based on the geometric properties of the data can achieve comparable accuracy on the perturbed data that on the original data. Suppose original data is represented by a d×N matrix, and N records represented as Xd×N, the rotation perturbation of the dataset X will be defined as G(X)=RX^T, where R is a random rotation orthogonal matrix. A key feature of rotation transformation is preserving the Euclidean distance, geometric shape, hyperplane, multi-dimensional space, and inner product.

PCA (Principal Component Analysis) is a technique that is used to decompose the multidimensional data into lower dimensions. PCA assumes that all the inconsistency in process should be used in the analysis so it becomes difficult to distinguish the crucial variable from the less vital. Subsequently, Principle Component Analysis replaces the original variables of a dataset with the smallest number of uncorrelated variables called the principle component [3].

PCA is a standard tool in modern fields from neuroscience to computer graphics. Since it is a simple, on-parametric method for extracting relevant information from mystifying datasets. With minimal effort, PCA provides a road map for how to reduce a composite dataset to a lower dimension so that the essential structures that remain underlie it. Principal Component Analysis (PCA) is a suitable transforming the multidimensional data into lower dimensions. It is a standard tool in modern data analysis. PCA assumes that all the inconsistency in process should be used in the analysis so it becomes challenging to distinguish the important variable from the less important variable. PCA is most appropriate for usual distributions (where linear Principle Component Analysis approach provides the best possible solution). Accordingly, Principal Component Analysis replaces the original variables of a dataset with the smallest number of uncorrelated variables called as the “principal components”. If the original dataset of dimension D contains highly associated variables, then there is no operational dimensionality exist as, d < D, that explains all the data. The presence of only a few components of D makes which is easier to label each dimension with an intuitive meaning. Furthermore, its more effective to operate on fewer variables in subsequent analysis.

III. RESULTS AND DISCUSSION

In order to evaluate the clustering accuracy, series of trials were performed over various sliding window sizes (w). Our evaluation approach focused on the inclusive quality of generated clusters after data perturbation. Experiment was based on following steps:

1. In MOA framework, Setup each dataset as stream.
2. To evaluate measures and cluster membership matrix, define sliding window (w) over the dataset.
3. Alter all the occurrences in sliding window by applying proposed data perturbation method to protect the sensitive characteristic value.
4. To find the clusters for performance evaluation, K-Means clustering algorithm is used. Our selection was influenced by (a) K-Means isone of the best known clustering Algorithm and it is also scalable.
   (b) The number of clusters found from original and perturbed dataset was taken as the number of clusters.

Make Comparison that how closely each cluster in the perturbed data matches its corresponding cluster in the original dataset. By computing the F-measure, we expressed the quality of the generated clusters.

To measure accuracy while protecting sensitive data, experiments were performed. Here we have presented two different results, one is analogous to clustering accuracy in terms of membership matrix which was manually plagiaristic from clustering result and another represent the equivalent graph for F1_P (precision) and F1_R (Recall) measures.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Total Instances</th>
<th>Instances Processed</th>
<th>Attributes Protected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account Management</td>
<td>42210</td>
<td>43k</td>
<td>Balance, Age</td>
</tr>
</tbody>
</table>

Table 1.1: Dataset configuration to determine accuracy based on Membership Matrix

@IIAERD-2015, All rights Reserved 232
To determine the accuracy of our proposed method, Table 1.1 shows datasets configuration. To determine set of 3 and 5 clusters using K-Means clustering algorithm, we configured each dataset.

Table 1.2, 1.3 shows the membership matrix acquired while clustering the perturbed attributes of Account Management dataset respectively. Each matrix representing 3 and 5 clusters scenario for true dataset and de-compose dataset. True dataset clustering provides information about no. of instances are actual classified in each cluster whereas perturb dataset clustering showing result of accurate assignments after attributes data perturbation and percentage of accuracy achieved.

### Table 1.2: Resultant accuracy of 5 Clusters

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Attributes</th>
<th>No. of Cluster</th>
<th>Stream Data</th>
<th>K-Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank Management</td>
<td>Age</td>
<td>5</td>
<td>85.21%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Balance</td>
<td></td>
<td>89.39%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duration</td>
<td></td>
<td>86.81%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Balance</td>
<td>3</td>
<td>82.96%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duration</td>
<td>3000</td>
<td>83.64%</td>
<td></td>
</tr>
</tbody>
</table>

### Table 1.3: Resultant accuracy of 3 Clusters

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Attributes</th>
<th>No. of Cluster</th>
<th>Stream Data</th>
<th>K-Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank Management</td>
<td>Age</td>
<td>3</td>
<td>88.13%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Balance</td>
<td>2000</td>
<td>92.60%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duration</td>
<td></td>
<td>89.42%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>3</td>
<td>85.59%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Balance</td>
<td>3000</td>
<td>91.22%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duration</td>
<td></td>
<td>89.64%</td>
<td></td>
</tr>
</tbody>
</table>
For each modified attribute, Results are presented in terms of graphs. Here each graph comprises the measure we obtained when original data is processed without applying privacy preserving method and K-Means is applied in order to evaluate both cases by keeping number of clusters fix (K=5, K=3) when data is undergone through our proposed privacy preserving method. In defined sliding window size, instances are processed. Here we representing the accuracy of our method by calculating the precision of individual cluster. F1_R measure determine the recall of system, which take into account the clustering measure provided with MOA framework. We focused on two important measures F1_R and F1_P. F1_P measure determine the precision of system by considering the precision of individual cluster. F1_R measure determine the recall of system, which take into account the recall of each cluster.

Fig 1.1: Accuracy on attribute Age in Bank Management with 5-Cluster
Fig. 1.2: Accuracy on attribute Balance in Bank Management with 5-Cluster

Fig. 1.3: Accuracy on attribute Duration in Bank Management with 5-Cluster
Fig. 1.4: Accuracy on attribute Age in Bank Management with 5-Cluster
Fig. 1.5: Accuracy on attribute Balance in Bank Management with 5-Cluster

Fig. 1.6: Accuracy on attribute Duration in Bank Management with 5-Cluster
Fig. 1.7: Accuracy on attribute Age in Bank Management with 3-Cluster
Fig. 1.8: Accuracy on attribute Balance Bank Management with 3-Cluster

Fig. 1.9: Accuracy on attribute Duration in Bank Management with 3-Cluster
Fig. 1.10: Accuracy on attribute Age in Bank Management with 3-Cluster
IV. CONCLUSION

While presenting on a publicly accessible place like internet, the proposed method can be used to hide sensitive information. The proposed privacy preserving prototype has been successfully implemented in Java under Windows 7 operating system and evaluated using Massive Online Analysis (MOA). The arrived results were more substantial and promising. Additionally, the proposed model can be used to multi party cooperative clustering development. Some of the results of earlier works have been shown, accuracy sometimes suffers as a result of security. However in the proposed method, the accuracy has been conserved and in some cases, the accuracy was almost equal to that of original data set.

REFERENCES


[10] Haisheng Li;” Study Of Privacy Preserving Data Mining” ;Third International Symposium On Intelligent Information Technology And Security Informatics; 978-0-7695-4020-7/10 $26.00 © 2010 Ieee


2013