Multi-Level Trust Privacy Preserving Data Mining to Enhance Data Security and Prevent Leakage of the Sensitive Data

Bourvil and Levi

Abstract—Privacy Preserving Data Mining (PPDM) is commonly utilized for the purpose of extracting related knowledge from large amount of data and simultaneously safeguard the sensitive information from the data miners. The major complication in privacy-sensitive domain is solved through the development of the Multi-Level Trust Privacy Preserving Data Mining (MLT-PPDM) where multiple differently perturbed copies of the single data is available to data miners at different trusted levels. In this scheme, data owners produce perturbed data through various schemes like Parallel generation, Sequential generation and On-demand generation. MLT-PPDM is extremely robust against the diversity attacks. In this work partial information hiding schemes like random rotation perturbation, retention replacement and K-anonymity are incorporated together with MLT-PPDM for the purpose of enhancing the data security and to prevent leakage of the sensitive data. At last, MLT-PPDM approach is improved in order to tackle against the non-linear attacks.

Keywords—K-Anonymity, Diversity Attack, Random Rotation Perturbation, Non-Linear Attack, Multi-Level Trust, Parallel Generation, Sequential Generation, On-Demand Generation.

I. INTRODUCTION

DATA mining (knowledge discovery from data) is defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from large data sets or databases. Advances in hardware technology have increased the capability to store and record personal data about consumers and individuals. This has caused concerns that personal data may be used for a variety of intrusive or malicious purposes. Privacy preserving data mining help to achieve data mining goals without sacrificing the privacy of the individuals and without learning underlying data values. Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure.

II. PROBLEM FORMULATION

Privacy is becoming an increasingly important issue in many data mining applications. A malicious data miner may have access to differently perturbed copies of the same data through various means, and may combine these diverse copies to jointly infer additional information about the original data that the data owner does not intend to release. This is called as Diversity Attack. Every day users are leaving dozens of electronic trails through various activities such as using credit cards, swapping security cards, talking over phones and using email. Additionally, it is a common practice that organizations sell the collected data to other organizations, which use these data for their own purposes. Organizations are extremely dependent on data mining in their every day activities. During the whole process of data mining (from collection of data to discovery of knowledge) these data, which typically contain sensitive individual information such as medical and financial information, often get exposed to several parties including collectors, owners, users and miners. Disclosure of such sensitive information can cause a breach of individual privacy.

Private information can also be disclosed by linking multiple databases belonging to giant data warehouses and accessing web data. An intruder or malicious data miner can learn sensitive attribute values such as disease type (e.g. HIV positive), and income (e.g. AUD 82,000) of a certain individual, through re-identification of the record from an exposed data set. This has triggered the development of many privacy preserving datamining techniques that try to extract the data patterns without directly accessing the original data and guarantees that the mining process does not get sufficient information to reconstruct the original data. Data Perturbation is a popular technique in PPDM and perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data is published. The scope of perturbation-based PPDM is extended to Multi-Level Trust(MLT-PPDM). Even though MLT-PPDM is robust against diversity attacks, partial information hiding methodologies like random rotation based data perturbation, k-anonymity and retention replacement are not supported by MLT-PPDM. Additionally MLT-PPDM considers only linear attacks but more powerful adversaries apply nonlinear techniques to derive original data and recover more information.

A. Data Perturbation

It is the most important technique in MLT-PPDM. It is a category of data modification approaches that protect the sensitive data contained in a dataset by modifying a carefully selected portion of attribute-values pairs of its transactions. The employed modification makes the released values inaccurate, thus protecting the sensitive data, but it also...
achieving preservation of the statistical properties of the dataset. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset. Data perturbation approaches fall into two main categories namely probability distribution approach and the value distortion approach. The probability distribution approach replaces the data with another sample from the same (estimated) distribution or by the distribution itself. On the other hand, the value distortion approach perturbs the values of data elements or attributes directly by some additive or multiplicative noise before it is released to the data miner \[3\].

- **Additive Noise**

  This approach adds to the original data, \(X\), some random noise, \(Z\), to obtain the perturbed copy, \(Y\) as \(Y = X + Z\). This approach tries to preserve data privacy by adding random noise, while making sure that the random noise still preserves the “signal” (original data) from the data so that the patterns can still be accurately estimated.

- **Multiplicative Noise**

  This method is based on generating random numbers that have a truncated Gaussian distribution with mean one and small variance, and multiplying each element of the original data by the noise.

B. **Parallel Generation**

In this method the components of noise \(Z\), i.e., \(Z_1\) to \(Z_M\), are generated simultaneously based on the probability distribution function.

**Algorithm 1: Parallel Generation**

1. **Input:** \(X\), \(K_X\), and \(\sigma_{Z_1}\) to \(\sigma_{Z_M}\)
2. **Output:** \(Y\)
3. **Construct** \(K_Z\) with \(K_X\) and \(\sigma_{Z_1}\) to \(\sigma_{Z_M}\)
4. **Generate** \(Z\) with \(K_Z\)
5. **Generate** \(Y = H_X + Z\)
6. **Output** \(Y\)

**Algorithm 1** serves as a baseline algorithm for the next two algorithms.

C. **Sequential Generation**

This method sequentially generates \(M\) independent noise \(Z_1\), and \((Z_i - Z_{i-1})\) for \(i\) from 2 to \(M\). The large memory requirement of Algorithm 1 motivates to seek for a memory efficient solution. Instead of parallel generation, sequentially generating noise \(Z_1\) to \(Z_M\), each of which a Gaussian vector of \(N\) dimension.

**Algorithm 2: Sequential Generation**

1. **Input:** \(X\), \(K_X\), and \(\sigma_{Z_1}\) to \(\sigma_{Z_M}\)
2. **Output:** \(Y_1\) to \(Y_M\)
3. **Construct** noise \(Z_1\) \(\sim N(0, \sigma_{Z_1}K_X)\)
4. **Generate** \(Y_1 = X + Z_1\)
5. **Output** \(Y_1\)
6. **for** \(i\) from 2 to \(M\) **do**
7. **Construct** noise \(Z_i\) \(\sim N(0, \sigma_{Z_i} - \sigma_{Z_{i-1}})K_X\)
8. **Generate** \(Y_i = Y_{i-1} + Z_i\)
9. **Output** \(Y_i\)
10. **end for**

The main disadvantage of the batch generation approach is that it requires a data owner to foresee all possible trust levels a priori. So On-demand generation methodology is proposed.

D. **On Demand Generation**

As opposed to the batch generation, new perturbed copies are introduced on demand in this scenario.

**Algorithm 3: On Demand Generation**

1. **Input:** \(X\), \(K_X\), \(\sigma_{Z_1}\) to \(\sigma_{Z_M}\), and values of \(Z'\)
2. **Output:** New copies \(Z''\)
3. **Construct** \(K_Z\) with \(K_X\) and \(\sigma_{Z_1}\) to \(\sigma_{Z_M}\)
4. **Extract** \(K_Z', K_Z''\) from \(K_Z\)
5. **Generate** \(Z''\) as a Gaussian with mean and variance
6. for i from L + 1 to M do
7. Generate Yi = X + Zi
8. Output Yi
9. end for

Assume L (L < M) existing copies of Y1 to YL, so that the data owner, upon requests, generates additional M−L copies of YL+1 to YM. Among three techniques on-demand generation offers data owners maximum flexibility where data owners generate perturbed copies of the data at arbitrary trust levels on-demand.

III. PROBLEM SOLUTION

A malicious data miner may have access to differently perturbed copies of the same data through various means, and may combine these diverse copies to jointly infer additional information about the original data that the data owner does not intend to release. This is called as diversity attacks and MLT-PPDM is robust against such attacks. To further enhance the data security and to prevent the loss of sensitive information efficiently, MLT-PPDM is integrated with partial information hiding techniques like K-anonymity, retention replacement and random rotation perturbation [4].

A. Random Rotation Based Perturbation in MLT-PPDM

• Agglomerative Hierarchical Clustering

The process of grouping a set of data points into classes of similar objects is called clustering. This bottom-up strategy of clustering starts by treating each data object as a cluster and then recursively merges these clusters based on the minimum distance between them until all objects are in a single cluster. At each step, a cluster is represented by the mean value of all the objects in the cluster. A tree structure called dendrogram is commonly used to represent the process of hierarchical clustering.

- Rotational Data Perturbation

‘λ’. After cluster rotation (θ) the coordinates are given by, Rotational data perturbation is performed by pair-wise attribute distortion of all objects within a cluster using the rotational transformation matrix. Cluster displacement is a translational perturbation performed by first displacing a set of clusters by a displacement factor ‘λ’ from the mean of the entire dataset, so that no two clusters overlap after the subsequent cluster rotation process. For this purpose, the displacement vector between every cluster centroid Gi and the mean of the entire dataset G must be multiplied by a factor

\[ \begin{align*}
\alpha’irp &= \alpha’irp + (λ-1) gipd = (\alpha’irp \cos \theta - \alpha’irq \sin \theta) + (λ-1) gipd \\
\alpha’irq &= \alpha’irq + (λ-1) giqd = (\alpha’irq \cos \theta + \alpha’irp \sin \theta) + (λ-1) giqd
\end{align*} \]

where gipd and giqd are the changes in cluster centroid

Figure 3: On Demand Generation

Figure 4: Mechanism of Advanced MLT-PPDM
Algorithm 4
Data Perturbation by Iterative Cluster Rotation for Hierarchical Clustering

Input: Dataset D containing ‘m’ n-dimensional objects.
Output: Perturbed dataset consisting of ‘m’ n-dimensional objects.

Method
1. Create ‘m’ clusters corresponding to each data object.
2. Repeat
3. Distance(Cp,Cq) = minimum(distance(Ci,Cj)), for all i, j forming cluster pairs.
4. Merge the clusters Cp and Cq.
5. Update the cluster centroid of merged cluster.
6. For each cluster Ci, calculate the maximum distance di of an object in Ci from its centroid Gi.
7. For each pair of clusters Ci and Cj, calculate \( \lambda_{ij} > 2 * \max\{d_i,d_j\} / (G_i - G_j) \);
8. Displace the clusters by \( \lambda \) around the mean of the entire dataset if \( \lambda > 1 \), where \( \lambda = \max\{\lambda_{ij}\} \), for all pair of clusters Ci and Cj.
9. For each object in cluster Ci, apply rotational transform about centroid Gi to each pair of attributes p and q, where \( \theta_{pq} \) is constant for all objects in Ci;
10. Until, only 1 cluster remains

Figure 5: Random Rotation based Perturbation in MLT-PPDM

B. K-Anonymity Clustering Method for Effective Data Privacy Preservation

The k-anonymity model is a simple and practical approach for data privacy preservation. In the k-anonymity model, the quasi-identifier feature set consists of features in a table that potentially reveals private information, possibly by joining with other tables. In the proposed K-anonymity method, CBK(L,K)-anonymity algorithm is used to solve the problem of privacy information leakage in data publishing. The main idea is anonymizing the data set by K-clustering based on influence matrix of background knowledge, it can make anonymous data effectively resist background knowledge attack and homogeneity attack, and can solve diversity of sensitive attribute. [6].

Algorithm 4: CBK(L,K)-Anonymity Algorithm
Input: data set S, parameter k, influence matrix \( M[S] \) based on background knowledge, threshold \( c \)
Output: a CBK(L,K)-anonymous table S’

If ( |S| < k )
return S
End if
r = a selected tuple randomly form S
While ( S > k )
t = the furthest distance tuple with r from S
S = S - {t}
if t[disease] < c
S = S - {t} \( \epsilon = \epsilon \cup \{t\} \)
else
\( \epsilon = \epsilon \cup \{t\} \)
end if
\( \epsilon = \epsilon \cup \{t\} \)
\( \epsilon = \epsilon \cup \{t\} \)
\( \epsilon = \epsilon \cup \{t\} \)
end while
End while
While ( S \#0 )
t = a selected tuple randomly form S
S = S - {t}
\( \epsilon = \epsilon \cup \{t\} \)
if t[disease] < c
\( \epsilon = \epsilon \cup \{t\} \)
else
\( \epsilon = \epsilon \cup \{t\} \)
end if
\( \epsilon = \epsilon \cup \{t\} \)
end if
End while
End while

tik should be promoted generalization hierarchies or suppress directly
end if
\( \epsilon = \epsilon \cup \{t\} \)
end if
End while
End while
Generalize and output a CBK (L,K)-anonymous table S’ according to \( \epsilon \)

C. Handling of Non-Linear Attack in MLT-PPDM

In this method the rotation matrix R, the random translation vector t, the noise level \( \sigma^2 \), and the minimum privacy guarantee are generated as the output. If the privacy guarantee is lower than an anticipated threshold, the data
owner can select not to release the data.

**Algorithm 5: Finding a Resilient Perturbation** $(\mathcal{X}_d\chi_N, w, \Phi, m)$

**Input:** $\mathcal{X}_d\chi_N$: the original dataset, $w$: weights of attributes in privacy evaluation, $\Phi$: the expected privacy guarantee in terms of distance-inference attack, $m$: the number of iterations.

**Output:** $\mathcal{R}_t$: the selected rotation matrix, $\psi$: the random translation, $\sigma^2$: the noise level, $p$: privacy quality.

1. Calculate the covariance matrix $C$ of $X$; $p = 0$; and randomly generate the translation $\psi$;
2. For each iteration do
   - Randomly generate a rotation matrix $R$;
   - Swapping the rows of $R$ to get $R_1$, which maximizes $\min\{1/w_i(Cov(R_1X-X))\}$;
   - $p_0 =$ the privacy guarantee of $R_1$, $p_1 = 0$; if $p_0 > p$ then
     - Generate $O$ with ICA;
     - Scale the columns in $O$ with the maximum/minimum values of original columns;
     - $\{((1), (2), \ldots, (d)) = \text{Argmin}((1), (2), \ldots, (d))/\Sigma \Delta PDF(X_i; O(i))\}$
     - $p_1 = \min\{1/w_k(VOD(X_k, O(k))\}$
     - End if
   - If $p < \min(p_0; p_1)$ then
     - $p = \min(p_0, p_1)$, $\mathcal{R}_t = R_1$;
     - End if
   - End for
3. $p_2 =$ the privacy guarantee to the distance-inference attack with the perturbation $\mathcal{G}(X) = R_1X + \psi + \Delta$.
4. Tune the noise level $\sigma^2$, so that $p_2 > \Phi$

**D. Retention Replacement in MLT-PPDM**

Retention replacement is based on the Member Migration (MM) technique. Basically, the technique first groups tuples in the original dataset $D$ into separate groups based on the similarity of values on a given quasi-identify attribute set and then performs a MM operation between each group pair where there is at least one group having the number of tuples less than $k$. If a tuple $t$ in group $A$ migrates to group $B$, values of $t$ have to change to ones of those in group $B$ with respect to the quasi-identify attribute set.

**IV. CONCLUSION**

The Multi-Level Trust in Privacy-Preserving Data Mining when integrated with partial information hiding methodologies help to find the right balance between maximum analysis results and keep the inferences that disclose private information about organizations or individuals at a minimum. Thus random rotation based data perturbation and K-anonymity are incorporated with MLT-PPDM to significantly enhance the data accuracy and to prevent the leakage of the sensitive data.

**REFERENCES**


[3] Charu C. Aggarwal and Philip S. Yu, “Privacy-Preserving Data Mining Models and Algorithms”, IBM T.J. Watson Research Center, USA and University of Illinois at Chicago, USA.


