The Applicability of the Perturbation Model-based Privacy Preserving Data Mining for Real-world Data

Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham
Computer Science Department
University of Texas at Dallas
liliu, muratk, bhavani.thuraisingham@utdallas.edu

Abstract

Perturbation method is a very important technique in privacy preserving data mining. In this technique, loss of information versus preservation of privacy is always a trade off. The question is, how much are the users willing to compromise their privacy? This is a choice that changes from individual to individual. In this paper, we propose an individually adaptable perturbation model, which enables the individuals to choose their own privacy level. Hence our model provides different privacy guarantees for different privacy preferences. We test our new perturbation model by applying different reconstruction methods to the perturbed data sets. Furthermore, we build decision tree and Naive Bayes classifier models on the reconstructed data sets both for synthetic and real world data sets. For the synthetic data set, our experimental results indicate that our model enables the users to choose their own privacy level without reducing the accuracy of the data mining results. For the real world data sets, we got very interesting results, hence we pose the question of whether the perturbation reconstruction model-based privacy preserving data mining is applicable for real-world data?

1. Introduction

Privacy preserving data mining has been intensively investigated since 2000. Perturbation and Randomization methods are very important techniques in this domain. In this paper, we will especially focus on the perturbation technique. This technique is usually used in scenarios where individuals reporting information to a data miner, can perturb the correct data with some kind of known random noise and report the noisy data to the data miner. Since the noise addition method is known by the data miner, data miner could reconstruct the original distribution using different statistical methods and could do data mining on the reconstructed data [2][1][7]. From these works, we can see that introducing noise and reconstructing the original distribution are two important steps in perturbation method. In the Kargupta et al’s work [7], they also challenge perturbation technique by showing point to point estimation of the original data.

1.1. Our contributions

- We have extensively studied the effect of perturbation methods on different data mining tasks and both synthetic data and real world data sets. We have found very interesting results from the experiments using different real world data sets. We would like to pose a question that whether perturbation model-based Privacy Preserving Data Mining (PPDM) is applicable for real world data sets?

- We present a novel two phase perturbation method for numerical data. Our initial experiments indicate that the techniques developed in [7] are not effective against our proposed perturbation method.

- Our proposed two-phase perturbation method provides individually adaptable privacy protection. Different individuals have very different attitudes towards privacy. For example, according to internet user survey reported in [4], “it seems unlikely that a one-size-fits-all approach to online privacy is likely to succeed ”. Our two-phase perturbation method allows individuals to set their privacy level according to their privacy choices.

1.2. Organization of the Paper

The paper is organized as follows: In section 2, we discuss related work. Section 3 introduce a privacy metric system used to measure privacy in our work. Section 4 describes our two-phase perturbation model. Section 5 shows reconstruction results of our model and comparison with
the Bayesian approach [2] and Principal Components Analysis (PCA) based [7] approach. Section 6 describes how the model described in Section 4 could be modified to support individually adaptable perturbation model. In section 7, we present our experimental results conducted by using decision tree and Naive Bayesian data mining techniques on both synthetic data and real world data sets. We discuss the applicability of the perturbation method for the real world data sets. In section 8, we conclude with our analysis of the interesting experiment results and the future work.

2. Related Work

Previous work in privacy-preserving data mining has addressed two issues. In one, the aim is preserving customer privacy by perturbing the data values [2]. The idea is that the perturbed data does not reveal private information, and thus is “safe” to use for mining. The key result is that the distorted data, and information on the distribution of the random data used to distort the data, can be used to generate an approximation to the original data distribution, without revealing the original data values. The distribution is used to improve mining results over mining the distorted data directly, primarily through selection of split points to “bin” continuous data. Later refinement of this approach tightened the bounds on what private information is disclosed, by showing that the ability to reconstruct the distribution can be used to tighten estimates of original values based on the distorted data [1]. The existing perturbation-based approaches apply random noise to the data sets without considering different privacy requirements of the different users. The data distortion approach has been also applied to Boolean association rules [5][10]. Again, the idea is to modify data values such that reconstruction of the values for any individual transaction is difficult, but the rules learned on the distorted data are still valid. One interesting feature of this work is a flexible definition of privacy; e.g., the ability to correctly guess a value of “1” from the distorted data can be considered a greater threat to privacy than correctly learning a “0”. One interesting criticism to the perturbation approach is proposed by Kargupta et al in [7]. By using random matrix properties, Kargupta et. al [7] successfully separates the data from the random noise and subsequently discloses the original data. Recently Huang et al [6] analyzed under what conditions the privacy of the underlying data used in perturbation method could be violated. Their results indicate that when the correlations between the data items are high, the original result can be constructed more easily.

The other approach uses cryptographic tools to build data mining models. This approach treats privacy-preserving data mining as a special case of secure multi-party computation (SMC). We do not discuss this approach any further here, please refer to [3] for details.

3. Privacy Metrics

In the work [1], Agrawal and Aggarwal have proposed a privacy measure based on differential entropy. We briefly repeat the ideas here. The differential entropy \( h(A) \) of a random variable \( A \) is defined as follows:

\[
h(A) = - \int_{\Omega_A} f_A(a) \log_2 f_A(a) da
\]

where \( \Omega_A \) is the domain of \( A \). Actually \( h(A) \) is a measure of uncertainty inherent in the value of \( A \) in the statistics. Agrawal and Aggarwal [1] based on this, proposed that the privacy measure inherent in the random variable \( A \) as \( \Pi(A) \).

\[
\Pi(A) = 2^{h(A)}
\]

For example, a random variable \( U \) distributed uniformly between 0 and \( a \) has privacy \( \Pi(U) = 2^{\log_2(a)} = a \). Thus if \( \Pi(A) = 1 \), then \( A \) has as much privacy as a random variable distributed uniformly in an interval of length 1. Furthermore if \( f_B(x) = 2f_A(2x) \), then \( B \) offers half as much privacy as \( A \). This can be easily illustrated as, a random variable uniformly distributed over [0, 1] has half as much privacy as a random variable uniformly distributed over [0, 2]. In [1] Agrawal and Aggarwal have also defined conditional privacy and information loss. For more detail please refer the original work [1].

We choose this privacy measure in our work to quantify the privacy in our individually adaptable model in section 6.

4. Two Phase Perturbation Model

General perturbation method is based on introducing noise without significantly changing the distribution of the original data. Later different statistical techniques are applied to the perturbed data to reconstruct the original distribution. Information loss versus privacy preservation is always a trade off in this method. The extent to which we perturb the original data will dramatically affect the data mining result and subsequently contribute to the potential risk of privacy disclosure. So choosing the appropriate level of perturbation is not trivial. A very common noise addition technique, first proposed in [2], is described below.

- Let \( x_1, x_2, ..., x_n \) be the original values of a one-dimensional distribution as realization of \( n \) independent identically distributed (iid) random variables, each has the same distribution as the random variable \( X \).
- Let \( y_1, y_2, ..., y_n \) be the random values used to distort the original data, \( y_i \) is the realization \( n \) independent
identically distributed (iid) random variables, each has the same distribution as the random variable Y. In the experiments, we use Uniform and Gaussian distributions for random variable Y, described as below:

- Uniform distribution: The random variable has a uniform distribution over an interval $[-\alpha, +\alpha]$. The mean of the random variable is 0.
- Gaussian distribution: The random variable has a normal distribution with mean $\mu = 0$ and standard deviation $\sigma$.

Given, $x_1 + y_1, x_2 + y_2, ..., x_n + y_n$ (perturbed data set) and cumulative probability distribution $F_{Y'}$ (noise), estimate probability distribution $F_X'$ (of original data).

We can see that noise addition procedure described above is only one step, just add the noise to the original data; and then apply reconstruction algorithm to estimate the original distribution. We call this model one phase perturbation model. Recent application of signal processing techniques (see for example [7]) offer many filters to remove the white noise. Furthermore, the signal processing method discussed in [7] poses a challenge to the random perturbation technique described above. The question is that under what conditions the random perturbation techniques sufficiently preserve privacy? One of the key insights from the work of Huang et al. [6] is that when the correlations between the perturbed data points are high, the original data can be reconstructed easily. For example [6], it is easy to see that in the method described above,

$$Cov(x_1 + y_1, x_2 + y_2) = Cov(x_1, x_2) \text{ for } i \neq j.$$  

Here we propose a two-phase perturbation model to create perturbed data points that are not correlated even if $Cov(x_i, x_j)$ is high. First we divide the domain of the $W$ into predetermined intervals. After generating the $w_i = x_i + y_i$, we calculate the predetermined interval $[l_k, k_{k+1})$ which $w_i$ falls. Instead of using $w_i$ during the reconstruction phase, we use a $w'_i$ that is generated uniformly from $[l_k, k_{k+1})$. Since $w'_i$'s are created as iid, $Cov(w'_i, w'_j) = 0$. Clearly, this second phase gets rid of the correlation between the perturbed data points. If the intervals used for sampling are chosen small enough, the second phase do not effect the cumulative distribution of $W$. Therefore it may not degrade the reconstruction phase. Figure 1 shows the processes of both one-phase and two-phase perturbation models. In the next section, we explore the effectiveness of the various reconstruction methods on our two phase perturbation model.

5. Original Distribution Reconstruction

We apply two major reconstruction techniques to the perturbed data set. The first technique is Bayesian inference base approach proposed by Agrawal et al [2]. The second technique is the principal component analysis (PCA) based approach proposed by Kargupta et al [7]. Please refer to the original work for the algorithms’ details.

For accurate comparison, we use the same experimental set up which are also used in [2] and [7]. Let $X$ be the original distribution. We create the original data set by creating 10K records from the triangle shape distribution from $[0, 1]$. Let $Y$ be noise distribution where we use two different alternatives:

- Gaussian Distribution: We use a normal distribution with mean $\mu = 0$ and standard deviation $\sigma$. We used three different $\sigma$ values (e.g., 0.25, 0.3 and 0.4) to see how it effects the result.
- Uniform Distribution: We use a uniform distribution over an interval $[-\alpha, +\alpha]$, e.g. $[-0.5, 0.5]$.

One-phase Perturbation Model

First we applied two algorithms, Agrawal et al [2] and Kargupta et al [7] algorithms to one-phase perturbation data set up. According to our experiments, both algorithms can estimate the distribution close to the original one. The results also indicate that when we use Gaussian noise, the variance $\sigma^2$ can dramatically affect the results. In Kargupta et al’s work [7], they have defined the term Signal-to-Noise Ratio (SNR) to quantify the relative amount of noise added to actual data, and also have stated that when $SNR$ is above 1.3, their algorithm produces good results. Our experimental results also confirm this statement [8].

Two-phase Perturbation Model

We keep the original data set and the first step perturbed data set $W$ (the same data sets used in the above experiments). We create the data $W^\prime$ as follow:

1. We divide $[-0.5, 1.5]$ (the domain of $W$) into 80 intervals.
2. Given the each perturbed data point $w_i$, we sample uniformly from the corresponding interval where $w_i \in [l_k, k_{k+1})$.

Figure 1. Two phase perturbation model
The experimental results have shown that Agrawal et al.'s algorithm can still successfully reconstruct the original distribution, but Kargupta et al.'s algorithm failed to do so. Please refer to [8] for details.

Analysis of the Two-step Perturbation Based Approach

Clearly, our two step algorithm does not significantly change the underlying cumulative distribution of the $W$. Therefore Bayesian inference based reconstruction approach could still achieve good reconstruction results. On the other hand, the PCA based approach described in [7] does not achieve successful results. Note that in the original columns from perturbed data matrix $Q$, upper and lower bounds of the eigenvalues of covariance matrix of $V$, denoted as $\lambda_{\text{max}}$ and $\lambda_{\text{min}}$, respectively are used. Although we obtain the sample matrix $Q'$ row by row to keep the mapping, our re-sampling from the corresponding intervals introduce a bias, denoted as $T$. So in our case, $U + V + T = Q'$. We still can calculate the bounds of $\lambda_{\text{max}}$ and $\lambda_{\text{min}}$ respectively, to separate $V$ from $Q'$, but now we only have the matrix $Q'$ instead of $Q$ to map the actual data and noise data with columns corresponding to the eigenvectors, and eventually get the estimated matrix $P'$, which should be our estimate of the original data set. $Q'$ has a small bias $T$; this $T$ does not change the variance of $(V + T)$ much, but it significantly affects the eigenvalues and changes the results dramatically. From another point of view, by re-sampling from the corresponding intervals, we reduce the correlation between perturbed data points. Based on the results showed in [6], it is normal that PCA-based reconstruction techniques do not work well when the correlation between the data points is low.

6. Individually Adaptable Perturbation Model

Our two-stage perturbation model described in Section 4 could be easily modified to incorporate different individual privacy preferences. Note that for each $w_i$, we are sampling a random point from the interval where $w_i \in [l_k, l_{k+1})$. Clearly this sampling could be done using different intervals for different users. For example, a user who is more cautious could use the interval $[l_{k-1}, l_{k+2}]$ for generating $w_i'$ (i.e., twice the original size). Using this fact, we can describe our individually adaptable perturbation method as follows:

1. User $i$ first adds a random noise $y_i$ to the original data value $x_i$ to get $w_i$.
2. Based on privacy preference, user choose an interval with an arbitrary length $[l_i, l_j]$ such that $w_i \in [l_i, l_j]$. $w_i'$ is created by sampling uniformly from the interval $[l_i, l_j]$.
3. $w_i'$ value is sent to the data miner.

Once the data miner receives the $w_i'$ values, Bayesian based reconstruction approach described in Section 5 could be used to reconstruct the original data distribution. Clearly if the length of the chosen intervals is large, the given data will look more like a random sample. If the interval is chosen small $w_i'$ will be very closed to the $w_i$. It is clear that by adjusting the interval length user could adjust their desired privacy. At the same time, if the number of users who are choosing to have a large interval to sample from is small, cumulative distribution function $F_W$ and $F_{W'}$ will not be too much different. In order to test the effectiveness of our individually adaptable perturbation method, we have conducted extensive experiments described in next section.

In our model, customers can choose different interval lengths to modify their privacy level. This privacy level could be measured using the metrics described in section 3.

7. Data Mining Experiment Results

Experiment Setup

The ultimate goal of privacy preserving data mining is to obtain accurate data mining patterns without any privacy violations. To test on this point, we build decision trees and Naive Bayes classifiers on both synthetic and real world data sets.

For comparison, we used the synthetic data which was used in [2]. The data set has nine attributes and five data mining functions. We have chosen three real world data sets: Income Data, Haberman Survival Data and Liver Data for our experiments from University of California, Irvine, machine learning database repository. In each data set, there are many attributes, and for training data, each instance has its class label.

When and how to reconstruct the original distribution will effect the data mining results. In [2] Agrawal et al. proposed three alternatives, Global, ByClass and Local. Global method reconstructs the original distribution for each attribute using the complete perturbed training data. ByClass method, for each attribute, first split the training data by class, then reconstructs the distributions separately for each class. Local method, when building a decision tree classifier, repeats the ByClass method at each node. It is obvious that Global method will not give good results, and Local method depends on what data mining technique is used, and it is computationally expensive due to repeated reconstruction process. So in our experiments we used ByClass method to reconstruct the original distribution.

Data Mining Experimental Results on Synthetic Data
For the synthetic data, we have generated 100,000 records, 66,000 of those records are used for training, and 34,000 of those records are used for testing purposes. We used WEKA [11] data mining software to run decision tree and Naive Bayes classifiers on our reconstructed data. We have compared the different performance of these two classifiers, and also our two phase model with the results reported in [2]. In our experiments, we assumed that there are four different user types with different privacy requirements. Table 1 shows the four different cases with different percentage of people with different attitudes towards privacy. Compared to the normal people, cautious people may want to preserve more privacy. So when we are sampling the perturbed data, we use a smaller interval length for normal people; for cautious people we use twice the interval length; and so on. Based on the four different groups, we generate four different data sets, named case 1 to case 4, and with different percentages are assigned to different categories. Since different attributes have different lengths in our data set, we set the intervals for each type by calculating domain-size divided by the number of intervals. For example, in case 4, 100K perturbed record data set is created as follows: we sample 70K perturbed data records using 200 intervals, then we sample 20K records using 100 intervals, and then we sample 5K records using 50 intervals, the last 5K records are sampled using 25 intervals.

We apply the privacy metrics described in section 3 to four different cases. We used the mutual information estimation technique given in [9] to calculate the privacy loss, shown in table 1. We can see case 1 has the most privacy loss, case 2 has less, and so on, then case 4 has the least privacy loss.

Due to the space limitations, we only show one of our experiments here. Figure 2 shows the data mining accuracy of four cases under our proposed two phase perturbation model, one phase perturbation model and the original data. Our results indicate that the differences between the outcomes of our two phase perturbation model and one phase perturbation model are not significant. Among these five functions we tested, function 2 and 3 are not easy to learn. (Please see the discussion in [2]). Although the models build using the reconstructed data have good prediction accuracy, there are not as good as the ones built from the original data. Clearly privacy is not free. Not surprisingly, the performance of decision tree classifiers is better than the Naive Bayes classifiers.

Applicability of Reconstruction Method in PPDM for Real-world Data

In this section we discuss the data mining results obtained from real world data sets. The surprising results are shown in table 2. Using the results obtained from original data set as the base line, we can see that the results obtained directly from perturbed data set mostly have lower accuracy, except for Haberman data set. We build the classifier from reconstructed data, and perform three different tests to compare data mining accuracy. In the first case, we test the classifier on the reconstructed test data; in the second case, we test the classifier on original test data; and in the third case, we test the classifier on the perturbed test data.

In the first case, all the results have higher accuracy than what obtained from original data sets. In the second and third cases, two data sets get very bad data mining accuracy, the Haberman data set get the accuracy similar to what obtained from perturbed data. We have to ask the questions: Is the reconstruction method applicable for real world data sets? When we rebuild the original distribution of each attribute of the data, we certainly keep the original distribution of each attribute, but for the instances, we do not keep the mapping of each instance in the reconstruction process. We noticed that the synthetic data used in [2], every attribute

<table>
<thead>
<tr>
<th>Attitude</th>
<th>Interval</th>
<th>Case1</th>
<th>Case2</th>
<th>Case3</th>
<th>Case4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paranoid</td>
<td>25</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Extra cautious</td>
<td>50</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
<td></td>
</tr>
<tr>
<td>Cautious</td>
<td>100</td>
<td>0%</td>
<td>10%</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>Normal</td>
<td>200</td>
<td>100%</td>
<td>90%</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>Privacy loss</td>
<td>N/A</td>
<td>0.2902</td>
<td>0.2809</td>
<td>0.2788</td>
<td>0.2705</td>
</tr>
</tbody>
</table>

Table 1. Privacy measure of different data sets used for data mining

Figure 2. Data mining accuracy
is uniformly distributed in a data range. In this case, after reconstruction, the instances of the rebuild data set are similar to the original data set due to the special distribution, every attribute is uniformly distributed. This is not easy to find in the real world. In this work we have only tested on three real world data sets. These data sets may have their limitations. So at this point we can not simply say, the reconstruction method is not applicable for real world data sets. We still need to do more research work to find out.

<table>
<thead>
<tr>
<th>Data</th>
<th>Orig.</th>
<th>Perturb.</th>
<th>Recon. case1</th>
<th>Recon. case2</th>
<th>Recon. case3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td>83.77</td>
<td>78.39</td>
<td>91.43</td>
<td>26.30</td>
<td>40.16</td>
</tr>
<tr>
<td>Haberm</td>
<td>71.89</td>
<td>77.78</td>
<td>99.67</td>
<td>75.49</td>
<td>75.49</td>
</tr>
<tr>
<td>Liver</td>
<td>79.42</td>
<td>77.10</td>
<td>97.68</td>
<td>25.22</td>
<td>23.48</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Orig.</td>
<td>Perturb.</td>
<td>Recon. case1</td>
<td>Recon. case2</td>
<td>Recon. case3</td>
</tr>
<tr>
<td>Income</td>
<td>79.72</td>
<td>77.89</td>
<td>88.74</td>
<td>24.06</td>
<td>28.12</td>
</tr>
<tr>
<td>Haberm</td>
<td>74.84</td>
<td>76.14</td>
<td>99.67</td>
<td>75.49</td>
<td>75.49</td>
</tr>
<tr>
<td>Liver</td>
<td>79.97</td>
<td>78.26</td>
<td>99.71</td>
<td>24.35</td>
<td>24.35</td>
</tr>
</tbody>
</table>

Table 2. Data mining accuracy

8. Conclusions and Future Work

Due to the different privacy needs of different individuals, the one-size-fits-all approach is not realistic in many privacy preserving data mining tasks. In order to address this problem, we propose a new perturbation method for privacy preserving data mining that can provide individually adaptable privacy protection. Our method enables users to choose different privacy levels without significant data mining performance degradation.

Also recent advances in reconstruction methods have showed that some of the perturbation techniques can be effectively attacked to violate privacy. Using one of the best known attacks against perturbation methods, we have shown that our method is not susceptible to such attacks. In order to confirm the effectiveness of our method, we have done extensive experiments under different data mining scenarios. Our results indicate that if only small number of people choose to have high levels of privacy (i.e., more perturbation), we can still find useful data mining results.

Similar to the previous work in this area, our method does not address the multiple attribute case. Obvious solution of adding independent random noise to each attribute may not offer good privacy protection for high dimensional data, since outliers can be easily detected. Although existing methods could be used by adding a random noise using a multivariate distribution, construction process could require huge amount of data. As a future work, we plan to investigate a different approach. Instead of trying to come up with a noise addition method that can be used for general data mining tasks, we plan to develop noise addition method specific to different data mining demands.

Reconstruction is a very important step for perturbation based PPDM approaches. We have found some problems when apply this technique to real world data sets as we discussed in section 7. In our other research work, we have proposed PPDM methods which skip this reconstruction step and map the data mining functions instead. We will investigate different data mining methods in this direction and believe this is very promising for perturbation based privacy preserving data mining approaches.

References