Preventing Diversity Attacks in Privacy Preserving Data Mining

Sailaja.R.J.L.¹, P.Dayaker²

¹M.Tech, Computer Science Engineering, MLRIT, Hyderabad, Andhra Pradesh, India
²Assistant Professor, Department of CSE, MLRIT, Hyderabad, Andhra Pradesh, India

Abstract -- Data mining exposes dataset to the researchers. However, it is possible to establish sensitive information from the given dataset. To overcome this problem Privacy Preserving Data Mining (PPDM) came into existence. PPDM ensures that the datasets are generated in such a way that they can not disclose identity of the entities present in the dataset. Before data is published, perturbation of data is made in order to preserve privacy in the data. Existing PPDM systems assumed single level trust on data miners. Recently Li et al. relaxed this by introducing Multilevel Trust in PPDM. They believe that if the data miner is more trusted less perturbation is required. However, malicious data miners can establish identity information by combining multiple perturbed copies. It prevents malicious attacks and it is a true PPDM. In this paper we implement that multilevel trust based PPDM which enables data owners to have freedom to choose the level of privacy needed. Based on this trust level perturbations are made. We built a prototype application that demonstrates the proof of concept. The empirical results revealed that the proposed approach is effective and offers flexibility to data owners.

Index Terms -- Data mining, privacy preserving data mining, random perturbation, multilevel trust

I. INTRODUCTION

Data mining is a process of extracting trends in the historical data. In other words it provides business intelligence that helps in taking good decisions which will lead to profits. However, data owner is supposed to give his data to data miner for the purpose of extracting business intelligence or discovering actionable knowledge from the data. When data miner has malicious intentions, he can use the data for knowing sensitive information and take advantage of it for monetary gains. To avoid this problem before giving data to data miner, it is perturbed in such a way that sensitive information is encoded or altered in order to ensure that the privacy of data is preserved. This feature is known as Privacy Preserving Data Mining (PPDM). There were many researches on the PPDM [1], [2], [3], [4], [5]. All these existing PPDM schemes assume single level trust on data miner. It does mean that the data owner generates only one perturbed copy. However, in the real world different trust levels might be required. It is the motivation behind exploring multilevel trust. For instance a government organization might have internal and external miners. Moreover, it also wants to give data to public. In this case obviously it needs to generate multiple copies of data with different perturbation applied based on the trust level of the data miners and general public. There is a problem in this approach too. When adversaries are able to get internal and external miner copies, they can establish the identity information from the data. Therefore it is very challenging to implement multiple level trust based PPDM.

When compared to single level trust scenario, many perturbed copies are required by the data owner to ensure non-disclosure of sensitive details. The number of perturbed copies depend on the trust level of the data miner. If the miner is trusted more, then it is likely that the number of perturbations of data to be published is less. However, from multiple diverse and perturbed copies the miner might produce original information accurately. This is the problem with the approach. Preventing such diversity attacks is a challenging task in multi-level trust based PPDM. Random Gaussian noise is added to overcome this problem.

Recently Li et al. [6] proposed a multi-level trust based PPDM that solves the problem of diversity attacks. The key challenge in the construction of the scheme is to ensure that the adversaries can’t establish original data from diversity of multiple perturbed copies. This problem is overcome by properly correlating perturbed copies with different trust levels. In this paper we implement the scheme proposed by Li et al. [6] and make a provision to end users to choose levels trust and number of perturbed copies required which do not allow diversity attacks. The reminder of the paper is structured as follows. Section II reviews related literature. Section III describes the proposed multilevel trust based PPDM. Section IV presents experimental results while section V concludes the paper.

II. PRIOR WORKS
Privacy preserving data mining has been around for many years. First of all it was introduced in [7] and [8]. The problem of PPDM has attracted many researchers. The existing work on this concept is categorized into two types. The first one is SMC (Secure Multipart Computation) which can provide highest level of privacy to data. It enables data miners to perform mining operations on data without disclosing sensitive information to data miners [7], [9]. This is achieved by implementing algorithms using genetic algorithms pertaining to SMC [10]. However, in practice, these algorithms are very expensive and not feasible. To overcome this drawback many solutions came into existence that are more efficient than SMC. The solutions are nothing but data mining techniques such as building decision trees [7], association rule mining [9], clustering through K-means algorithm [11] and other data mining algorithms like frequent pattern mining [12].

For privacy preserving data mining in collaborative fashion secure coprocessor was used for privacy preserving. The second category is the approach known as partial information hiding which ensures good privacy and also performance. Many solutions came into this category. For instance k-anonymity [13], [14], [15], [16], [17], retention replacement [18], [19], [20] and perturbation of data [4], [3], [2], [1], [21] are some of the examples. The data perturbation methods are of two classes. Thy are known as additive [5], [1], [3], [8] and other scheme is known as matrix multiplicative schemes [2], [22]. These methods are suitable for numeric data. To solve the problem of PPDM, in [23] a new adversary model is proposed.

Some of the existing protocols have drawback of leaking information [24]. Private and threshold set is introduced in [25] and they are algorithms that are equity based. Reanonymization is the concept studied to solve the problem of PPDM. This Study is made in [26], [27], and [28]. IN this paper we study the anonymization of data many times for privacy preserving data mining with multiple levels of trust computations. Our multilevel trust based approach is effective as it relaxes the assumption of single level trust on data miners.

### III. MULTILEVEL TRUST BASED PPDM

This section provides details about the PPDM scheme which will support multi-level trust to protect privacy of data which is given to data miners for discovering knowledge. The existing solutions use single level trust on data miners. It does mean that those systems use a single perturbed copy of data. In this paper we support multilevel trust based PPDM that helps data owners to make multiple perturbed copies of the data. For each data miner based on the trust level the data owner generates different perturbed copy.

#### Privacy Goal

The design goal of the multilevel trust based PPDM is that the data owner should be allowed to built distinct perturbed data for each data miner. The data owner distributes such copies to data miners. M is assumed to be the number of copies generated by data owner for ease of analysis. Based on the trust level of data miner, the privacy of perturbed copy can be computed as follows.

\[
D(X, \hat{X}(Y_i)) = \frac{1}{N} \text{Tr}(E(\hat{X}(Y_i) - X)(\hat{X}(Y_i) - X)^T)
\]

\[
= \frac{\sigma^2}{\sigma^2_i + 1/N} \text{Tr}(K_X)
\]

This will help data owner to control the data easily. This is achieved by setting the privacy of perturbed copies. However, this has a drawback when adversaries launch attack by combining multiple perturbed copies and generate the original data. To overcome this problem Noise concept is introduced. With this data owner produces another copy of perturbed copy which will ensure more security. The adversaries can’t have diversified attacks. To achieve this noise covariance matrix as part of corner-wave property is used.

#### Batch Generation

When the data owner knows the number of data miners, trust levels he generates corresponding number of perturbed copies with privacy goal in mind. This process is known as batch generation. Two algorithms are proposed for noise generation in parallel and sequentially. The parallel generation algorithm is as shown in fig. 1.
This algorithm generates noise components simultaneously. The algorithm 2 is presented in figure 2 which generates noise sequentially.

This algorithm generates noise components sequentially. It is a memory efficient solution. The drawback of the parallel generation is that the data owner should have known all the possible trust levels required by all data miners priori. Therefore this is not flexible. For on demand requirement, another algorithm is proposed as presented in figure 3.

This algorithm provides flexibility to data owner as it can generate perturbed copies of data on demand. This will achieve privacy goal specified and also based on multilevel trust based. More details about the technical details of multi-level trust based PPDM can be found in [29].

IV. EXPERIMENTAL RESULTS

In the first set of experiments, algorithm 3 is used as it gives flexibility to data owner. The experiments are made on multiple perturbed copies which is the best case attack scenario. We assume that all perturbed copies are accessible to data miners to have such scenario. The perturbed copies are releases one by one therefore their availability also is influenced by this. Experiments are made based on two sensitive columns such as Age and Income. The experimental results are compared with that of IN (independent noise scheme).
Fig. 4. Comparisons of average normalized estimation error of the independent noise scheme (denoted as IN) and our scheme (denoted as Ours) on the data Age

As shown in the above figure represents horizontal axis represents number of perturbed copies while vertical axis represents normalized estimation error.

Fig. 5. Comparisons of average normalized estimation error of the independent noise scheme (denoted as IN) and our scheme (denoted as Ours) on the data Income

As shown in the above figure represents horizontal axis represents number of perturbed copies while vertical axis represents normalized estimation error.

Fig. 6. The corresponding histogram of the estimation error when M = 5

As shown in the above figure 6 represents the horizontal axis represents X while vertical axis represents Proportion.

Fig. 7. The cumulative histogram of the estimation error when M = 5

As shown in the above figure 7 represents the horizontal axis represents X while vertical axis represents Proportion.

Fig. 8. The corresponding histogram of the estimation error when M = 10

As shown in the above figure 8 represents the horizontal axis represents X while vertical axis represents Proportion.
Fig. 9. The cumulative histogram of the estimation error when M =10
As shown in the above figure 9 represents the horizontal axis represents X while vertical axis represents Proportion.

Fig. 10. The corresponding histogram of the estimation error when M =20
As shown in the above figure 10 represents the horizontal axis represents X while vertical axis represents Proportion.

V. CONCLUSION
In this paper we proposed a new approach for privacy preserving data mining. The existing perturbation based PPDM models assume single level trust on data miners. In this work we focus on multilevel trust based PPDM which provides more flexibility to data owner in choosing the level of privacy to data. A challenge in doing so is that malicious data miners can use multiple copies of perturbed data in order to establish the original data. This kind of attack is prevented by using noise correlation matrix across the copies to deny the attackers not to have diversity option. We build a prototype application that demonstrates the proof of concept pertaining to multi-level trust based PPDM. The empirical results revealed that the proposed solution is robust and effective.

REFERENCES


AUTHORS

Sailaja Returi She is pursuing M.Tech (CSE)in MLRIT, Hyderabad, AP, INDIA. She has received B.Tech Degree in Computer Science and Engineering. Her main research interest includes Datamining and Neworking.

P.Dayaker He is currently with the Department of Computer Science and Engineering, MLRIT, Andhra Pradesh, India. He is having 7 years of teaching experience. His main research interest includes Cloud Computing and Data Mining.