ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Integration of Partial Information Hiding Technology with Multi-level Trust Privacy Preserving Data Mining

Rajashree R.Joshi [1], Prof. U.A.Nuli [2]

P.G. Student, Department of Computer Science& Engineering, DYPCET, Kolhapur, Maharashtra, India[1]

Assistant Professor, Department of Computer Science & Engineering, DKTE S's TEI, Ichalkaranji, Maharashtra, India[2]

ABSTRACT: Privacy Preserving Data Mining (PPDM) addresses the problem of developing accurate models about aggregated data without access to precise information in individual data record. A widely studied perturbation-based PPDM approach introduces additive perturbation. Previous solutions of this approach are limited in their tacit assumption of single-level trust on data miners. In this work, the scope of perturbation-based PPDM is expanded to Multilevel Trust (MLT-PPDM). In this setting, the more trusted a data miner is, the less perturbed copy of the data it can access. Under this setting, 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. Preventing such diversity attacks is the key challenge of providing MLT-PPDM services. This challenge is addressed by properly correlating perturbation across copies at different trust levels. This solution allows a data owner to generate perturbed copies of its data for arbitrary trust levels on-demand. This feature offers data owner's maximum flexibility.

KEYWORDS: Privacy preserving data mining, multilevel trust, additive perturbation, Diversity attack, K-anonymity

I. INTRODUCTION

The goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. It consists of various mining types such as privacy preserving data mining, web mining, text mining, spatial mining, edge mining. Dataset contains some private information that data owner does not intend to release. This problem is solved by privacy preserving data mining.

Privacy preserving data mining means extract relevant knowledge from large amount of data and at the same time protect the sensitive information from data miners. Privacy Preserving Data Mining (PPDM) approach, tacitly assumes single-level trust on data miners. This new dimension of Multi-Level Trust (MLT) poses new challenges for perturbation based PPDM [1]. In contrast to the single-level trust scenario where only one perturbed copy is released, now multiple differently perturbed copies of the same data is available to data miners at different trusted levels. The more trusted a data miner is the less perturbed copy it can access; it may also have access to the perturbed copies available at lower trust levels. Moreover, a data miner could access multiple perturbed copies through various other means. The data miner may be able to produce a more accurate reconstruction of the original data than what is allowed by the data owner. This attack is known as a diversity attack. Preventing diversity attacks is the key challenge in solving the MLT-PPDM problem.

Multi-level trust privacy preserving data mining can find many applications. This work takes the initial step to enable MLT-PPDM services. Proposed system will use data perturbation techniques and partial information hiding technology to protect sensitive data. It will preserve privacy for complete databases.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

II. RELATED WORK

Privacy Preserving Data Mining (PPDM) was first proposed in [2] and [3] simultaneously. To address this problem, researchers have since proposed various solutions that fall into two broad categories based on the level of privacy protection they provide. The first category of the Secure Multiparty Computation (SMC) approach provides the strongest level of privacy; it enables mutually distrustful entities to mine their collective data without revealing anything except for what can be inferred from an entity's own input and the output of the mining operation alone [3]. In principle, any data mining algorithm can be implemented by using generic algorithms of SMC. However, algorithms are extraordinarily expensive in practice, and impractical for real use. To avoid the high computational cost, various solutions those are more efficient than generic SMC algorithms have been proposed for specific mining tasks. Solutions to build decision trees over the horizontally partitioned data were proposed in [3]. The second category of the partial information hiding approach trades privacy with improved performance in the sense that malicious data miners may infer certain properties of the original data from the disguised data. Various solutions in this category allow a data owner to transform its data in different ways to hide the true values of the original data while at the same time still permit useful mining operations over the modified data. This approach can be further divided into three categories: (a) data perturbation (which introduces uncertainty about individual values before data is published) [1],[2],[3],[4]. The data perturbation approach includes two main classes of methods: additive [1],[2],[3] and matrix multiplicative schemes. These methods apply mainly to continuous data. Here, additive perturbation approach is used where noise is added to data value.(b) k-anonymity [4],[5](c) retention replacement (which retains an element with probability p or replaces it with an element selected from a probability distribution function on the domain of the elements)

Xiao et al. proposed an algorithm of multi-level uniform perturbation. Multi-level privacy preserving is proposed for additive Gaussian noise perturbation and use a measure based on how closely the original values can be reconstructed from the perturbed data. While [7] presents an algorithm of multi-level uniform perturbation. As a result, neither the solution in [7] can be easily applied to the problem in [1] nor the solution in [1] can be directly applied to the problem in [7]. Secondly, based on Gaussian noise perturbation, the solution is more suitable for high dimensional data, as compared to that in [7].

III. PROPOSED SYSTEM

In the proposed work, Data owner want to provide some information to data miner. But it also wants to protect sensitive data such as name, age, salary. Hence to protect this information following steps are necessary:-
1) First, We need to generate the noise. In order to generate noise, we consider two algorithms such as sequential generation and on-demand generation.

2) For obtaining perturbed copy, these generated noise needs to add to the original data.

3) To implement one of the partial information hiding technology using MLT-PPDM to enhance data security and to prevent leakage of sensitive data.

4) After applying Additive perturbation technique and after implementation of partial information hiding technology whichever attributes are generated. We will add this to one another to obtain perturbed copy.

5) Then, multiple differently perturbed copies will available for data miners at different trust level.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

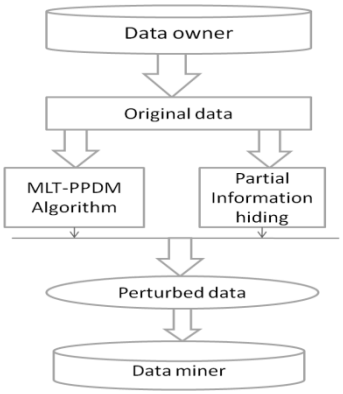

Fig 1: Mechanism of advanced MLT-PPDM

IV. EXPERIMENTAL DATASET

Stock market data which is used in 'Irrational Exuberance' is taken as original data. This data set contains one million tuples with some attributes such as data, price, dividend, earnings, index, and interest. Among these, first 10^2 tuples are taken and conducted the experiments on price, dividend, earnings, index and data.

Minimum	1871.01
Maximum	1871.1
Mean μ_χ	1871.055
Standard deviation	0.03

Table 1: Statistics of the data

V. IMPLEMENTATION

Noise is generated using batch generation and on-demand generation algorithm. Partial information hiding technique is used. There are two batch generation algorithms i.e. parallel and sequential generation algorithm.

5.1 Sequential generation algorithm:-

Input: X, Kx and $\sigma^2 z_1$ to $\sigma^2 zm$
Output: Y_1 to Ym
Construct Z_1
Generate $Y_1 = X + Z_1$
Output Y_1
for i from 2 to m do
    Construct noise ξ
    Generate $Y_i = Y_{i-1} + \xi$
    Output Y_i
end for

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

5.2 On-demand generation algorithms:-

1. Input: X, K_X, $\sigma^2_{Z_L}$ to $\sigma^2_{Z_M}$, and values of Z′:v_1
2. Output: New copies Z″
3. Construct K_Z with K_X and $\sigma^2_{Z_L}$ to $\sigma^2_{Z_M}$
4. Extract $K_{Z'}$,$K_{Z''Z'}$, and $K_{Z''}$ from K_Z
5. Generate Z″ as a Gaussian with mean and variance
6. for i from L + 1 to M do
7. Generate $Y_i = X + Z_i$
8. Output Y_i
9. end for

Kz', Kz"z' and Kz" are extracted from Kz. New vectors Z' and Z" are defined as follows:

$$\mathbb{Z}' = \begin{bmatrix} Z_1 \\ \vdots \\ Z_L \end{bmatrix} \qquad \mathbb{Z}'' = \begin{bmatrix} Z_{L+1} \\ \vdots \\ Z_M \end{bmatrix}.$$

The data owner should generate new noise Z" in such a way that the covariance matrix of Z has corner-wave property, and they are jointly Gaussian. It is sufficient and necessary for the conditional distribution of Z" given that Z' takes any value v1 to be a Gaussian with mean and variance as follows:

$$K_{\mathbb{Z}''\mathbb{Z}'} K_{\mathbb{Z}'}^{-1} v_1, \qquad K_{\mathbb{Z}''} - K_{\mathbb{Z}''\mathbb{Z}'} K_{\mathbb{Z}'}^{-1} K_{\mathbb{Z}''\mathbb{Z}'}^{T}$$

Here Kz' is the covariance matrix of Z'. Kz"z' is the desired covariance matrix between z" and z'. Kz" is the desired covariance matrix between z". Kz' is known to data owner. Kz"z' and Kz" can be extracted from desired covariance matrix Kz. This feature offers data owner's maximum flexibility

5.3 Implementation of partial information hiding:-

It is not clear how to expand the scope of other approaches in the area of partial information hiding technology. To overcome this limitation and further enhance data security, Partial information technology is used to protect sensitive data. There are three Partial information methodologies:-

1. Random rotation based data perturbation
2. K-anonymity
3. Retention replacement

Among these three techniques, K-anonymity is used with PPDM.

Steps for K-anonymity:-

The data holder declares specific attributes and tuples in the original private table (PT) as being eligible for release. The data holder also groups a subset of attributes of PT into one or more quasi-identifiers (QIi). The data holder specifies a minimum anonymity level that computes to a value for *k*. assumes that $k < |\mathbf{PT}|$, which is a necessary condition for satisfying *k*-anonymity.

The algorithm has few steps.

1. Construct freq, which is a frequency list containing distinct sequences of values from PT[QI], along with the number of occurrences of each sequence. Each sequence in freq represents one or more tuples in a table.
2. Generalization is performed. The attribute having the most number of distinct values in freq is generalized. Generalization continues until there remains *k* or fewer tuples having distinct sequences in freq.
3. Suppresses any sequences of freq occurring less than *k* times.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

4. Complimentary suppression is performed in step 4 so that the number of suppressed tuples satisfies the *k* requirement.
5. Finally, step 5 produces a table, based on freq such that the values stored as a sequence in freq appear as tuple(s) in a table replicated in accordance to the stored frequency.

VI. EXPERIMENTAL RESULTS

From original data, we have taken only 10 rows and 10 columns for noise generation. And number of trust levels are 10. So they released 10 perturbed copies. These copies are available to the data miners. By combining all these copies, following are some results obtained from sequential generation algorithm.

Sample Results obtained from sequential generation algorithm:-

1870.987	4.4736	0.26	0.4	13.0778
1871.004	4.07976	0.26	0.4	14.12896
1871.04	4.74	0.26	0.4	12.56
1871.05	4.86	0.26	0.4	12.27
1871.298	5.91896	0.26	0.4	7.53116

Table 2: Sample results for Sequential generation

From original data, we have taken only 50 rows and 10 columns for noise generation. Among these, following are some results obtained from on-demand generation algorithm.

Sample Results obtained from on-demand generation algorithm:-

1870.62	5.07	-0.57	0.06	11.65
1870.95	4.49	0.19	0.34	13.09
1871.01	5.53	1.84	-0.33	13.10
1870.82	5.31	0.49	0.63	12.62
1871.11	4.03	0.62	0.64	11.90

Table 3: Sample results for On-demand generation

Sample Results obtained from K-anonymity algorithm:-

```
run:
Loading file... Done.
There are 30 instances in the original file
After deleing records with security code 2, there are 27 records
Datafly - starting...
Trying to acheive 5 anonymity
Suppresed Records: 0
Datafly - ended...
There are 27 instances in the final output.
```

Fig 2: Result for K-anonymity

In this way, we can preserve the privacy for complete and K-anonymous databases.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

VII. SYSTEM EVALUATION

After testing; system's performance is evaluated using the metrics such as precision, recall and F1-measures.For checking performance of the system, this module calculates precision, recall and F1-measure metrics.

The precision is the fraction of retrieved instances that are relevant to the perturbed copy, and the recall is the fraction of relevant instances that have been retrieved. For a binary classification problem the judgment can be defined within a contingency table as depicted in **Table 2**

human judgment			
System judgment	**Yes**		**No**
	Yes	TP	FP
	No	FN	TN

Table 4 : Contingency table

According to the definition in **Table 2,** the precision and recall are calculated using following equations. Where TP (True positives) is the number of instances the system correctly identifies as positives; FP (False Positives) is the number of instances the system falsely identifies as positives; FN (False Negatives) is the number of relevant instances the system fails to identify.

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Comparison of precision, recall and F1-measure for perturbed copy by considering instances with highest relevance score is as shown in following **Fig 2.**

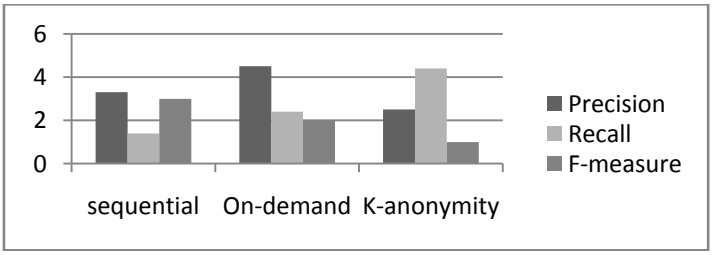

Fig 3: Comparison of Performance metrics over perturbed copy

It can be observed that maximum values for precision, recall and f1-measure are obtained from On-demand generation algorithm. . On-demand gives better results than sequential generation method.

VIII. CONCLUSION

Scope of additive perturbation based PPDM has been expanded to multilevel trust (MLT), by relaxing an implicit assumption of single-level trust in exiting work. MLT-PPDM allows data owners to generate differently perturbed copies of its data for different trust levels. The key challenge lies in preventing the data miners from combining copies at different trust levels to jointly reconstruct the original data more accurate than what is allowed by the data owner.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

This challenge is addressed by properly correlating noise across copies at different trust levels. This solution allows data owners to generate perturbed copies of its data at arbitrary trust levels on-demand. This property offers the data owner maximum flexibility.. K-anonymity is used to enhance data security. Suppression technique is used for K-anonymity. Certain attribute is removed from micro data. In this way, reidentification of individual records is prevented.

REFERENCES

1] Y. Li and M. Chen, "Enabling Multi-Level Trust in Privacy Preserving Data Mining," IEEE transactions on knowledge and data engineering, sept.2012, pp 1598-1612
2] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), 2000,pp.439-450
3]Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Proc. Int'l Cryptology* Conference (CRYPTO), 2000, pp. 36-53
4] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2005
5] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS), vol. 10, 2002,.pp. 557-570