ISSN (Online): 2230-7893
www.IJCEM.org

Security of Database for Data Miner with Help of Perturbed Technique

[1]Prasannta Tiwari, [2]Hitesh Gupta

[1]Research Scholar, [2]Asst. Prof.
Dept. of Computer Science and Engineering
PCST, Bhopal, (M.P.) India

Abstract

Today data holder want to utilize and release data to third party for analyzing or researching however, but they not required to disclosing any individual data within its privacy interval to anyone. Here find the answer to what extent confidential information in a perturbed database can be compromised by attackers or snoopers. key of element is preserving privacy and confidentiality of data is the ability to evaluate the extent of all potential disclosure for the data. Several randomized techniques have been proposed for privacy preserving data mining of continuous data. These approaches generally attempt to hide the important data by randomly modifying the important data values using some additive noise and aim to reconstruct the original distribution closely at an aggregate level. The main contribution of this paper lies in the algorithm to accurately reconstruct the community joint density given the perturbed multidimensional stream of data information. Any statistical question about the community can be answered using the reconstructed joint density. In our research objective is to determine whether the distributions of the original and recovered of important data are close enough to each other despite the nature of the noise applied. We are considering an ensemble clustering method to reconstruct the initial data distribution.

Keywords- Perturbation Data, Regenerate of Data, distribution reconstruction, information privacy, random distortion, recovered data.

1. INTRODUCTION

Recently, researchers in the data mining to get more idea about the volume of the information available in perturbed data. Here we mention that databases of two of the largest web resources – National Climatic Data Center and NASA – contain about 600 terabytes of data, which is only about 8% of so-called "deep" web. But some time they along with the availability and the amount of data, the privacy issue have also experienced a big resonance. Different poll among web users reveal that about 85% of people give their preference to a privacy policy. This scenario consider in this paper is that a single party holds a collection of original individual data. Each individual data is associated with one privacy interval [1].

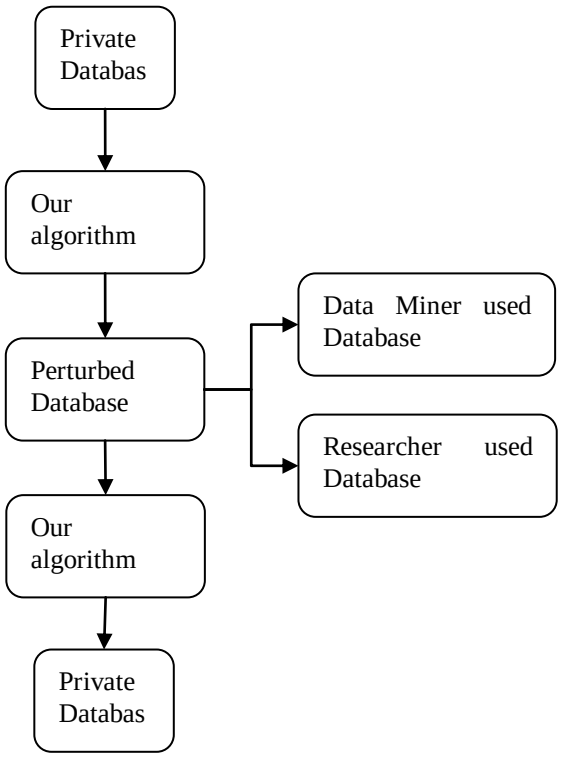

Fig 1. Our Perturbation Model

The data holder can utilize or release data to the third party for analysis; however, he is required not to disclose any individual data within its privacy interval. For example, one company collects its employees' personal information (e.g., income, age, etc.) and needs to release this data set to the third party for analysis. Since each employee has his/her concern on the privacy of their personal data, the company should figure out ways to release data while guaranteeing no individual data can be derived by attackers or snoopers within its privacy interval. The current randomization based privacy preserving data mining approaches [2] seem to fulfill this

need. These approaches generally attempt to hide the sensitive data by randomly modifying the data values using some additive noise. Hence presented a general algorithm for reconstruction of community statistics; it remains to decide on the perturbation function.

One of the examples for the data privacy used in real life is the insurance companies. They do not give access to the original data, the private information of their customers. But instead they can provide some sort of statistics of the data changed in some certain way, without providing the original information of individual customers. But even such "vague" data can be used to identify trends and patterns.

Basically, there are two approaches of data concealment.

1. The first approach is data randomization (perturbation). Usually it conceals the real data by modifying it randomly, superimposing a random noise on it.
2. The approach uses the cryptography techniques to encode the initial information.

There exist a lot of cases when we need to obtain the information on the initial data. For instance, companies, selling their product in online stores, might be interested in finding out the range of customer age/salary their product should target to. Since this information is not available in its initial state, a company needs to deal with the perturbed/encrypted data. The main goal of this article is to evaluate the initial distribution of the data using a so called ensemble clustering method, and then to compare its efficiency to other methods of data reconstruction.

2. RELATED WORK

In this paper they propose new method for the obtaining the original data distribution the Ensemble Method for Clustering. This method is considered, discussed the Ensemble Method and its core the Voting Algorithm in more details. The contribution of this paper is to develop robust and efficient method for the data distribution reconstruction

In the area of matrix multiplicative perturbation, distance based preserving data perturbation [3, 4, 5] has gain a lot of attention because it guarantees better accuracy. The transformed data is used as input for many important data mining algorithms, such as k-mean classification [6], k-nearest neighbor classification [7] and distance based clustering [8], and the corresponding output is exactly as same as the result of analyzing the original data. However the security issue of how much the privacy loss has caused researchers' concern. [9] Studied that how well an attacker can recover the original data from the transformed data and prior information. They proposed three different attack techniques based on prior information. [10] Made further study. They proposed a closed-form expression for the privacy breach probability and indicated that even with a small number of known inputs; the attack can achieve a high privacy breach probability.

Either additive perturbation or matrix multiplicative perturbation has the potential possibility of being attacked. [11] considered a combination of matrix multiplicative and additive perturbation: $Y = M(X \square + R)$ this method makes it better to hide the original data. They also discussed a known I/O attack technique, and pointed out that $\hat{} M$, an estimate of M , can be produced using linear regression and then X is estimated.

Mohammad's [12] method is only applicable to building privacy-preserving decision tree. The two additive perturbation algorithms they proposed expand its application to security mine patients' information. The original data is pre-mined by the government officials to get the "patterns", and then after being added noise, the data is adjusted properly to keep the clusters similar to the ones in the original data.

The academic researchers only need to mine the perturbed data directly without any extra work, so the step of reconstructing the original data distribution with its high computation cost and the step of modifying mining algorithm are both not needed any more. To protect privacy better, they address the application of their algorithms to a two-step model: $Y = M(X \square + R)$ which is not fit for building decision tree, but fit for statistical analysis. The first step of it gets the perturbed data by their algorithms, and the second step protects Euclidean distance of the perturbed data. In this way, computation cost is minimized and privacy is better preserved. their experimental results have shown that this model not only has a higher degree of accuracy, but also guarantees that its privacy security is as good as, if not better than, the other models.

3. PROPOSED TECHNIQUE

a. Data Perturbation

Data perturbation techniques can be grouped into two main categories, which call the value distortion technique and probability distribution technique. The value distortion technique perturbs data elements or attributes directly by either some other randomization procedures. On the other hand, the probability distribution technique considers the private database to be a sample from a given population that has a given probability distribution. In this case, the perturbation replaces the original database

by another sample from the same (estimated) distribution or by the distribution itself.

There has been extensive research in the area of statistical databases (SDB) on how to provide summary statistical information without disclosing individual's confidential data. The privacy issues arise the summary statistics are derived from data of very few individuals. A popular disclosure control method is data perturbation, which alters individual data in a way such that the summary statistics remain approximately the same. However, problems in data mining become somewhat different from those in SDBs. Data mining techniques, such as clustering, classification, predication and association rule mining are essentially relying on more sophisticated relationships among data records or data attributes, but no just simple summary statistics. This research paper specifically focuses on data perturbation for privacy preserving data mining. Some important perturbation approaches in SDBs are also covered for sake of completeness.

Multiplicative Perturbation:

Two basic forms of multiplicative noise have been studied in the statistics community. One multiplies each data element by a random number that has a truncated Gaussian distribution with mean one and small variance. The other takes a logarithmic transformation of the data first, adds multivariate Gaussian noise, then takes the exponential function exp(.) of the noise-added data. Neither of these perturbations preserves pair wise distance among data records.

$$P = \frac{i-1}{i} P_{i-1} + \frac{1}{i} \Pi^{i}_{\max}$$

Where i=0 to i-1

Where the data is multiplied by a randomly generated matrix – in effect, the data is projected into a lower dimensional random space. This technique preserves distance on expectation. These authors observed that the distance preserving nature of random rotation enables a third party to produce exactly the same data mining results on the perturbed data as if on the original data. However, they did not analyze the privacy limitations of random rotation. The privacy issues of distance preserving perturbation (including rotation) by studying how well an attacker can recover the original data from the transformed data and prior information.

In this paper we consider the first approach – the data randomization. If we have the initial data set of N independent variables $X=\{x_1, x_2 \ldots x_N\}$. In order to perturb the data we consider N independent random values $Y= \{y_1, y_2 \ldots y_N\}$ and the perturbed data set will be given as X'=X+Y. In this case it is impossible to reconstruct initial values exactly but it is possible to recover the initial data distribution with some certain precision. There also is some loss of information during the previous distribution reconstruction process. However, the reconstruction algorithms offered in different papers (including this one) are able to recover the original data pattern. Which algorithm one should use, is a matter of a precision and an efficiency of the method.

b. Perturbation-Invariant Classification Models

The classification models that is invariant to geometric data perturbation with our algorithms. The model quality $Q(M_X, Y)$ is the classification accuracy of the trained model tested on the test dataset.

KNN Classifiers:

A k-Nearest-Neighbor (kNN) classifier determines the class label of a point by looking at the labels of its k nearest neighbors in the training dataset and classifies the point to the class that most of its neighbors belong to. Since the distance between any pair of points is not changed with our algorithm, the k nearest neighbors is not changed and thus the classification result is not changed either. This approach preserves data covariance instead of the pair-wise distance among data records. Proposed algorithm based perturbation method which recursively partitions a data set into smaller subset such that data records in each subset are more homogeneous after each partition; the private data in each subset are than perturbed using the subset average. The relationship between-attributes are expected to be preserved.

The basic problem considered in this paper can be abstracted as the following: we have the set of distracted data set. Our task is to obtain the original data distribution based on the present distorted data. Again, as it was mentioned, we reconstruct only distribution, not the actual values of individual records of the dataset. Before announcing the method to be used in this paper, let us define the concept of clustering, since it is "a mile-stone" of the background theory implemented in algorithms described later. We consider a set of data points each having a set of attributes.

The main goal of clustering is to divide data into groups called clusters, such that data points in one cluster would be more similar to one another and respectively, data points in separate clusters would be less similar to one another. The similarity can be measured based on Euclidean Distance (in case attributes are continuous). After obtaining results our goal is to the compare effectiveness of the methods suggested. We also try the

algorithm for the different types of perturbations such as product and exponential, as well as for various kinds if

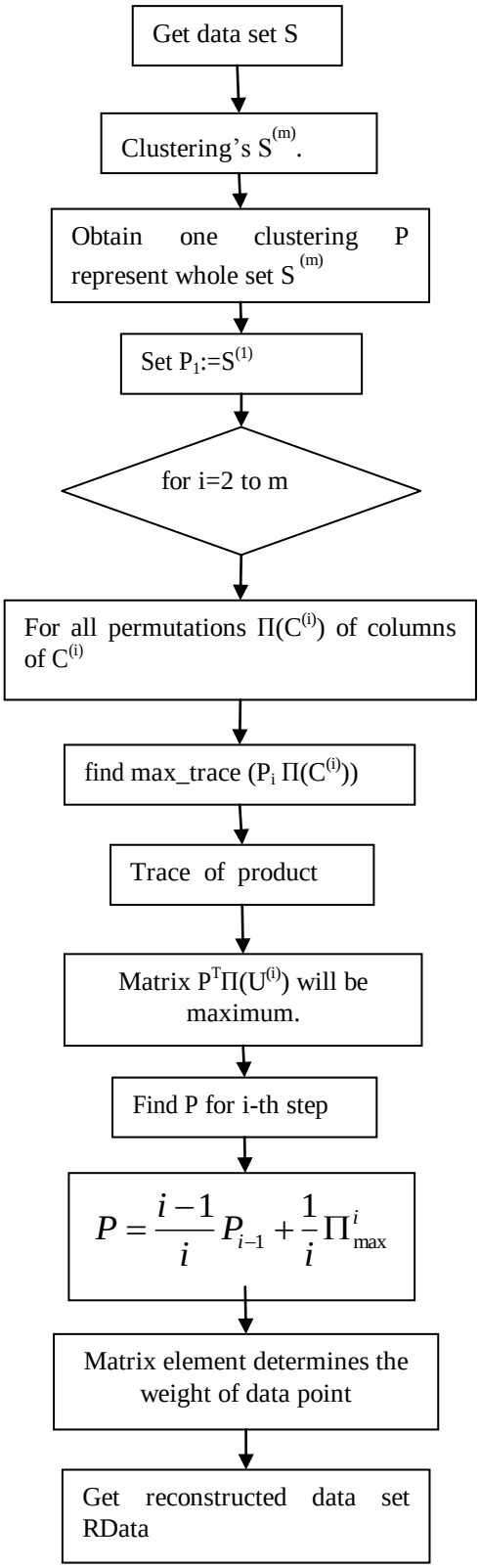

Fig 2. Algorithms Flow Chart

distributions (normal, uniform). For instance, if X is the initial dataset matrix and Y is matrix consisting if random noise, then in case of product perturbation the perturbed dataset.

4. PROPOSED ALGORITHM

The basis of the algorithm Clustering's $S^{(m)}$ is obtain one cluster P which is represented by whole set $S^{(m)}$ the whole algorithm show in fig 2 and perturbed model show in fig 1.

5. IMPLEMENTATION AND EXPERIMENT

This research will use MATLAB as the environment for the algorithm implementations. For experiment here taken 14*300 data of patient from national hospital. the given dataset we are considering clustering techniques: k-means method with k-Nearest-Neighbor (kNN) classifier algorithm.

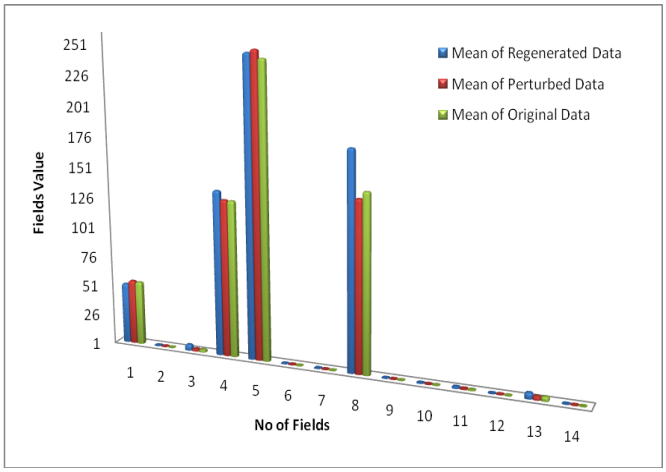

Fig 3. Show result of mean of perturbed data, mean of original data and mean recovered data

We taken the perturbed dataset, we applied our algorithms mentioned above (we selected the parameters for which the methods were issuing the best results). As the measure of the quality we considered the correlation between the initial distribution and the one obtained due to the clustering. Namely, we were calculating the correlation between mean of original, recovered and perturbed data (initial and clustered ones).

Here we considered our algorithm to obtain the set of clustering to be obtain from our algorithm; here k-means clustering algorithms run 19 times each, with varying parameters for each run. Parameters for the methods were chosen in the following way:

No of Field	Mean of Recovered Data	Mean of Perturbed Data	Mean of Original Data
1	51.52	54.88	54.32
2	1.19	0.60	0.69
3	6.02	3.00	3.16
4	138.88	131.90	131.61
5	250.55	253.30	246.66
6	0.14	0.15	0.15
7	1.21	0.96	0.98
8	182.67	143.62	149.78
9	0.25	0.33	0.33
10	1.21	1.04	1.05
11	2.36	1.75	1.61
12	0.37	0.66	0.66
13	6.25	4.16	4.75
14	0.62	0.51	0.54

Table 1. Show result in tabular form of our research, mean of perturbed mean, mean of original data and mean recovered data.

1. Find area A enclosing all points in dataset.
2. EPS $\alpha \approx \sqrt{A}$, where α is (roughly) the ratio of the average and maximum densities.
For our case $a \approx 0.09$.

$$MinPts = \frac{2\pi * EPS^2 * N}{A}$$

Here N is the total number of data points.

After running these methods we will be taken set of 14 clustering, which using as the input for our Algorithm.
Another challenging issue in our experiment was the varying number of clusters in each clustering produced by methods, while our Algorithm requires equal number of clusters in each clustering. To overcome this problem we were taking the maximal number of clusters $S^{(m)}$ among all clustering's as the universal one. Then we extended the number of clusters in the clustering to the given number $S^{(m)}$. After finding optimal clustering P for the given set, we calculated the mean of result data and found the correlation between it and the original mean.

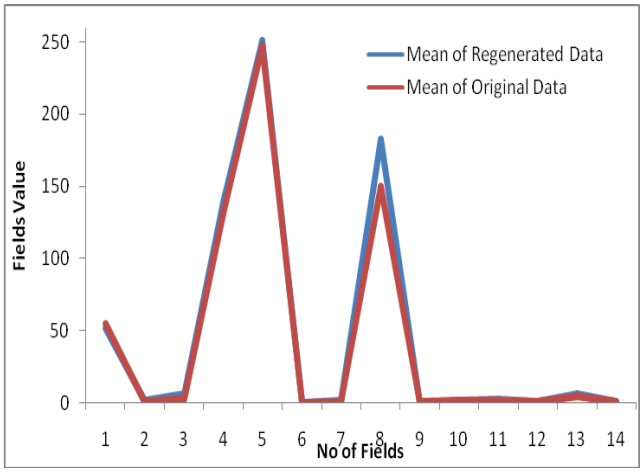

Fig 4. Algorithms Flow Chart

6. CONCLUSION

As the important issue in this area, we consider the possibility of original data distribution restoration from the available perturbed dataset. In addition to the several other techniques available (such as Bayes Rule and Expectation Maximization based techniques) but we propose the new one, which is based on the recently invented approach concerning the merging several different clustering's into optimal one.

To examine our proposition, we consider the two-dimensional dataset, where the data points are grouped into four elliptic-shaped partitions. To perturb data, we apply our algorithm, therefore masking real values of data points. Now, given the perturbed dataset we will use clustering algorithms, to cluster the perturbed dataset, which is to find the original partitions.

After this we will be try to the our algorithm with the set of forty clustering's obtained from running k-means with varying parameters and obtain one optimal clustering. As the measure of the efficiency of the original data distribution restoration we consider the correlation between the original and restored incidence matrices. We calculate the mean of original dataset, perturbed dataset after recovered dataset. From our experimental result it is clear that our algorithm recovered more near to original dataset.

REFERENCES

[1]. Z. Huang,W. Du, and B. Chen. "Deriving private information from randomized data". In *Proceedings of theACM SIGMOD Conference on Management of Data*.Baltimore, MA, 2005.

[2]. H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. "On the privacy preserving properties of random data perturbation techniques". In *Proc. of the 3rd Int'l Conf. on Data Mining*, pages 99–106, 2003.
[3]. Yang, W. J. "Privacy protection by matrix transformation." *IEICE Transactions on Information and Systems*, E92-D(4), 740-741 2009.
[4]. Chen, K. and Liu, L. "Privacy preserving data classification with rotation perturbation." In *Proceedings of the Fifth IEEE International Conference on Data Mining*. Houston, TX, 589-592. 2005
[5]. Liu, K. Kargupta, H. and Ryan, J. "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining*." IEEE Transactions on knowledge and Data Engineering*, 18(1), 92-106. 2006
[6]. Su, C. H., Zhan, J. and Sakurai. K. "Importance of Data Standardization in Privacy-Preserving K-Means Clustering." In the *Proceedings of International Workshops on Database Systems for Advanced Applications*. Brisbane, QLD, Australia, 276-286, 2009.
[7]. Chong, Z. H, Ni, W. W., Liu, T. T. and Zhang, Y. "A privacy-preserving data publishing algorithm for clustering application." *Computer Research and Development*, 47(12), 2083-2089, 2010.
[8]. Raaele Giancarlo, Giosue Lo Bosco, Luca Pinello. "Distance functions, clustering algorithms and microarray data analysis." In *Proceedings of the 4th International Conference on Learning and Intelligent Optimization. Venice*, Italy, 125-138, 2010.
[9]. Liu, K., Giannella, C. and Kargupta, H. "A survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods." In: *Privacy-Preserving Data Mining: Models and Algorithms*. 2008.
[10]. Preeti Jain, Vijay Kumar Trivedi, "A Novel Technique for Data Hiding in Audio by using DWTS", International Journal of Computational Engineering & Management IJCEM, Vol. 15 Issue 4, July 2012.
[11]. Giannella C and Liu K. "On the Privacy of Euclidean Distance Preserving Data Perturbation." *Computer Science-Cryptography and Security*. 2009
[12]. Chen, K., Sun, G. and Liu, L. "Towards attackresilient geometric data perturbation." In *Proceedings of the 2007 SIAM International Conference on Data Mining*. Minneapolis, MN. 2007.
[13]. Mohammad, A. K. and Somayajulu, D.V.L.N. "A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining." *Journal of Computing*, 2(1), 2151-9617, 2010.