Volume 3, No. 10, October 2012

Journal of Global Research in Computer Science

ISSN-2229-371X

RESEARCH PAPER

Available Online at www.jgrcs.info

RECONSTRUCTION OF PERTURBED DATA USING K-MEANS

Prasannta Tiwari*[1] and Hitesh Gupta[2]
*M-Tech (Computer science and engineering) PCST Bhopal, (M.P.) India
prasanntakaur@yahoo.com
[2]Asst. Professor, PCST Bhopal, (M.P.) India
Gupta_hitesh@sify.com

Abstract: A key element in preserving privacy and confidentiality of sensitive data is the ability to evaluate the extent of all potential disclosure for such data. In other words, we need to be able to answer to what extent confidential information in a perturbed database can be compromised by attackers or snoopers. Several randomized techniques have been proposed for privacy preserving data mining of continuous data. These approaches generally attempt to hide the sensitive data by randomly modifying the 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 data information. Any statistical question about the community can be answered using the reconstructed joint density. There have been many efforts on the community distribution reconstruction. Our research objective is to determine whether the distributions of the original and recovered 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. As the tool for the algorithm implementations we chose the “language of choice in industrial world” – MATLAB.

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

INTRODUCTION

The dramatic growth of the Internet during the past decade has resulted in the tremendous amount of information. In order to get some idea about the volume of the information available today 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 along with the availability and the amount of data, the privacy issue has also experienced a big resonance. Despite whether the private data is being retrieved for malicious (i.e. obtaining information about credit card number or bank information) or for official (i.e. information on online activity of individuals gathered by federal government) reasons, people are concerned about keeping the private information undisclosed. Different poll among web users reveal that about 85% of people give their preference to a privacy policy. The scenario we consider in this paper is that a single party (data holder) holds a collection of original individual data. Each individual data is associated with one privacy interval [1]. 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. The perturbed individual data is expected to be dissimilar with its original one (or lie out of its privacy interval), hence the individual’s privacy is assumed to be preserved. Hance 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. The first approach is data randomization (perturbation). Usually it conceals the real data by modifying it randomly, superimposing a random noise on it. The second 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 (since customers do not want their personal information to be available for public), 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.

In this paper we consider the first approach – the data randomization. If we have the initial data set of N independent variables $X=\{x_1, x_2 \dots x_N\}$. In order to perturb the data we consider N independent random values $Y= \{y_1,$

$y_2 \dots\ y_N\}$ 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.

RELATED WORK

Privacy-preserving is one of the mostly considerable topics in data mining. Respectively, there exist a lot of references and literature on this extensive subject. Although there exist different categories for the privacy-preserving data mining algorithms (such as ones based on a so called distributed framework and data-swapping approaches), our prime interest is still the random perturbation of data. In such approach is considered: additional random noise modulates the data, such that the individual data values are distorted preserving the original distribution properties if considering the dataset as a whole. After applying random noise, the perturbed data is used to extract the patterns and models. The randomized value distortion technique for learning decision trees and association rule learning are examples of this approach.

There are many different algorithms dealing with the randomly perturbed data sets. One of the mostly used algorithms is so called an expectation-maximization (EM) algorithm considered in. It is also remarked in that the method, based on the Bayesian approach, suggested in “does not take into account the distribution of the original data (which could be used to guess the data value to a higher level of accuracy)”. Compared to the method used in, EM Algorithm provides more robust evaluation of initial distribution and less information loss even in case of large number of data points. Another method for the privacy-preserving data mining considered in is the association rule analysis.

In this paper they propose new method for the obtaining the original data distribution – the Ensemble Method for Clustering. This method is considered and discussed in. The next section describes the Ensemble Method and its core – the Voting Algorithm in more details. The main contribution of this article 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. He 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: $Y = M(X + R)$ This method makes it better to hide the original data. They also discussed a known I/O attack technique, and pointed out that ^ 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 we 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, we address the application of our algorithms to a two-step model: $Y = M(X + R)$ which is not fit for building decision tree, but fit for statistical analysis. The first step of it gets the perturbed data by our algorithms, and the second step protects Euclidean distance of the perturbed data. In this way, computation cost is minimized and privacy is better preserved. Our 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.

PROPOSED TECHNIQUE

Perturbation-Invariant Classification Models:

The classification models that are 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 are 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).

We implement the voting algorithm in MATLAB. 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 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.

PROPOSED ALGORITHM

The basis of the our algorithm is the Voting and k-Nearest-Neighbor (kNN) classifier. The algorithm itself is based on the following idea:

```
Get data set S
      let us have a set of m clustering's $S^{(m)}$.
     Obtain one clustering P represent whole set $S^{(m)}$
      optimally.
    Set $P_1$:=$S^{(1)}$
        for i=2 to m
          For all permutations $\Pi(C^{(i)})$ of columns of $C^{(i)}$
          find max_trace ($P_i$ $\Pi(C^{(i)})$ )
         Permutation $\Pi^{i}_{max}$ for which the trace of product
         Matrix $P^T\Pi(U^{(i)})$ will be maximum.
         Find P for i-th step using recursive formula:
$P = \frac{i-1}{i} P_{i-1} + \frac{1}{i} \Pi^{i}_{\max}$
         $P_m$ is the optimal clustering.
        $C^{(i)}$ is the fuzzy clustering matrix where columns are
         clusters and rows are the data points.
         Each matrix element determines the weight of data
          point belonging to the certain cluster.
         Get reconstructed data set $RData_s$
```

That is for each row, the sum of all elements will be equal one (except for cases when point does not belong to any cluster – so called noise. In this case all elements in correspondent row will be zeros).

Notice that in step 2 of our algorithm we consider k! Permutations of k columns of clustering matrix $C^{(i)}$. For the number of clusters greater than 8-9, our algorithm will become computationally expensive. For such cases there are some other techniques not considered in this paper.

IMPLEMENTATION AND EXPERIMENT

This research will use MATLAB as the environment for the algorithm implementations. For the given dataset we are considering clustering techniques: k-means method with k-Nearest-Neighbor (kNN) classifier algorithm. 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 incidence matrices (initial and clustered ones).

Then we considered our algorithm. To obtain the set of clustering to be used in the voting algorithm, we ran k-means clustering algorithms, 19 times each, with varying parameters for each run. Parameters for the methods were chosen in the following way:

We ran k-means for four consequent numbers of centroids, all around some K number, which was chosen as the one for which the k-means method was issuing clustering with largest correlation to the original data distribution. That is, first 4 runs were performed for K-2 centroids, the next 4 runs for K-1 centroids and so on.

a. Find area A enclosing all points in dataset.
b. EPS α≈ √A, where α is (roughly) the ratio of the average and maximum densities.

For our case a≈ 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 40 clustering, which we were 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 Kmax of clusters among all clustering's as the universal one. Then we extended the number of clusters in the clustering to the given number Kmax. Since P optimal clustering is a fuzzy one (that is the belonging of point to a cluster is weighted), we choose the cluster where the point has maximal weight. After finding optimal clustering P for the given set, we calculated the incidence matrix and found the correlation between it and the original incidence matrix.

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) we propose the brand 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 use clustering algorithms, to cluster the perturbed dataset, which is to find the original partitions. After this we provide 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 correlation coefficients for all clustering methods.

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]. Giannella C and Liu K. "On the Privacy of Euclidean Distance Preserving Data Perturbation." *Compute Science-Cryptography and Security*. 2009

[11]. 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.

[12]. 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.