SECURE CIVIC DATA BY PERTURBETED DATA METHOD USING WITH CLUSTERING

Bipul Ranjan and Malti Nagle

Email id: ranjanbipul53@gmail.com

Abstract— a key element in preserving privacy With the wide placement of public cloud computing infrastructures and confidentiality of sensitive data is the ability to evaluate the extent of all potential disclosure for such data using clouds to host data query facilities has become an appealing solution for the rewards on scalability and cost-saving.. 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. That sensitive data owner does not want to move to the cloud unless the data confidentiality and query privacy are guaranteed. Several randomized techniques have been proposed for privacy preserving data mining of continuous data. This paper propose the clustering base data perturbation method to provide secure and effective range query services for protected data in the cloud. These approaches generally attempt to hide the sensitive data by randomly adapting the data values using some preservative noise and aim to rebuild the original distribution closely at an aggregate level. Secured query service should still provide effective query processing and significantly reduce the in-house workload to fully realize the assistances of cloud computing. The main contribution of this paper deceits in the algorithm to accurately perturbation and reconstruct the civic joint density given the perturbed multidimensional stream data information. 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. Extensive experiments have been conducted to show the advantages of this approach on Efficiency and security. As the tool for the algorithm implementations we chose the "language of choice in industrial world" – MATLAB.

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

I. INTRODUCTION

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 9% of so-called "deep" web. The dramatic growth of the Internet during the past decade has resulted in the tremendous amount of information. Hosting data-intensive data in the cloud is increasingly popular because of the unique advantages in scalability and cost-saving. The service owners can conveniently scale up or down the service with the cloud Infrastructures, and only pay for the hours of using the servers. This is an attractive feature because the workloads of query services are highly dynamic, and it will be expensive and inefficient to serve such dynamic workloads with in-house infrastructures [2]. But along with the availability and the amount of data, the privacy issue has also experienced a big resonance. 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. 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. 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. 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. 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. 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.

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. 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.

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. 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.

II. 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.

The reconstruction of gene regulatory networks [3] from gene Expression data has been the subject of intense research activity. The more recent availability of higher-throughput sequencing platforms, combined with more precise modes of genetic perturbation, presents an opportunity to formulate more robust and comprehensive approaches to gene network inference. A variety of models and methods have been developed to address different aspects of this important problem. However, these techniques are often difficult to scale, are narrowly focused on particular biological and experimental platforms, and require experimental data that are typically unavailable and difficult to ascertain. Here, they propose a step-wise framework for identifying gene-gene regulatory interactions that expand from a known point of genetic or chemical perturbation using time series gene expression data. This novel approach sequentially identifies non-steady state genes post-perturbation and incorporates them into a growing series of low-complexity optimization problems. The governing ordinary differential equations of this model are rooted in the biophysics of stochastic molecular events that underlie gene regulation, delineating roles for both protein and RNA-mediated gene regulation. They show the successful application of our core algorithms for network inference using simulated and real datasets.

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. 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. 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.
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)”. 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. 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

The more recent concept of differential privacy [4], introduced by the cryptographic community, is an approach that provides a rigorous definition of privacy with meaningful privacy guarantees in the presence of arbitrary external information, although the guarantees may come at a serious price in terms of data utility. The protection of privacy of individual-level information in GWAS databases has been a major concern of researchers following the publication of ‘‘an attack’’ on GWAS. Traditional statistical methods for confidentiality and privacy protection of statistical databases do not scale well to deal with GWAS data, especially in terms of guarantees regarding protection from linkage to external information. Building on such notions, they proposed new methods to release aggregate GWAS data without compromising an individual’s privacy. They extend the methods developed in for releasing differentially-private v2-statistics by allowing for arbitrary number of cases and controls, and for releasing differentially-private allelic test statistics. They assess the performance of the proposed methods through a risk-utility analysis on a real data set consisting of DNA samples collected by the Welcome Trust Case Control Consortium and compare the methods with the differentially-private release mechanism. They also provide a new interpretation by assuming the controls’ data are known, which is a realistic assumption because some GWAS use publicly available data as controls.

In the area of matrix multiplicative perturbation, distance based preserving data perturbation [5], [6] 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 [7], k-nearest neighbor classification [8] and distance based clustering [9], 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. [10] 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. [11] 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. [12] 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.

With the wide deployment of public cloud computing infrastructures [13], using clouds to host data query services has become an appealing solution for the advantages on scalability and cost-saving. However, some data might be sensitive that the data owner does not want to move to the cloud unless the data confidentiality and query privacy are guaranteed. On the other hand, a secured query service should still provide efficient query processing and significantly reduce the in-house workload to fully realize the benefits of cloud computing. We propose the RASP data perturbation method to provide secure and efficient range query and kNN query services for protected data in the cloud. The RASP data perturbation method combines order preserving encryption, dimensionality expansion, random noise injection, and random projection, to provide strong resilience to attacks on the perturbed data and queries. It also preserves multidimensional ranges, which allows existing indexing techniques to be applied to speedup range query processing. The kNN-R algorithm is designed to work with the RASP range query algorithm to process the kNN queries. We have carefully analyzed the attacks on data and queries under a precisely defined threat model and realistic security assumptions. Extensive experiments have been conducted to show the advantages of this approach on efficiency and security.

Mohammad’s [14] 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.

III. PROPOSED TECHNIQUE

In this paper we consider the first approach – the data randomization. However, the reconstruction algorithms offered in different papers (including this one) are able to recover the original data pattern. If we have the initial data set of N independent variables X={x_1, x_2… x_N}. In order to perturb the data we consider N independent random values Y= {y_1, y_2… y_N} and the perturbed data set will be given as X'=X+Y. Which algorithm one should use, is a matter of a precision and an efficiency of the method. 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.

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:

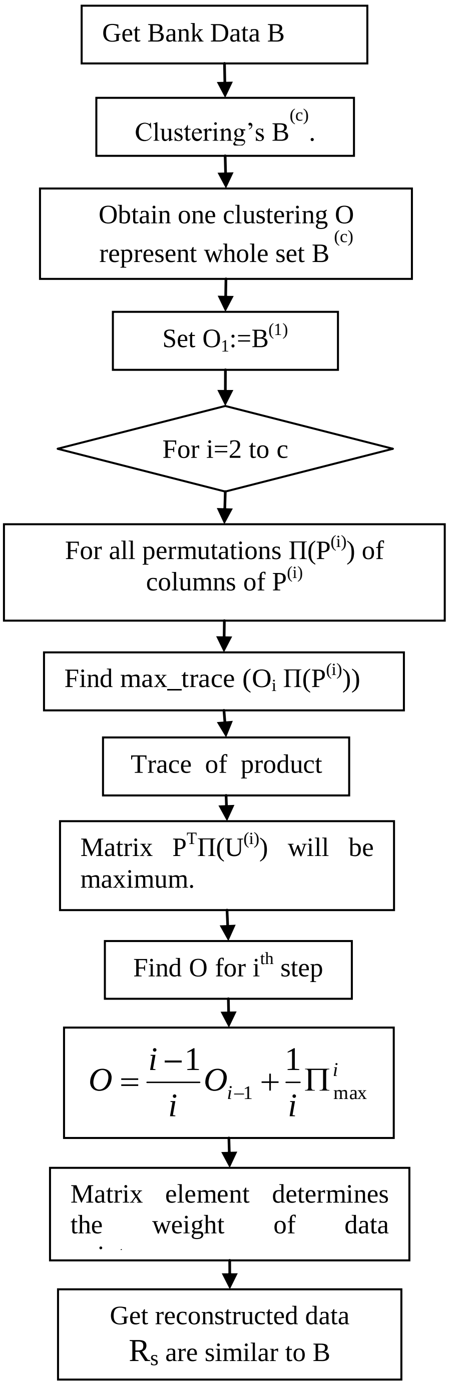
A k-Nearest-Neighbor (kNN) classifier determines the class label of a point by looking at the labels of its k nearest neighbors in the training dataset and classifies the point to the class that most of its neighbors belong to. We consider a set of data points each having a set of attributes. Since the 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.

The similarity can be measured based on Euclidean Distance (in case attributes are continuous). 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.

This paper 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). We implement the voting algorithm in MATLAB. After obtaining results our goal is to the compare effectiveness of the methods suggested. 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.

IV. PROPOSED ALGORITHM

The basis of proposed algorithm is the Voting and k-Mean classifier. The algorithm itself is based on the following idea:


Get Bank Data B
Clustering's $B^{(c)}$.
Obtain one clustering O represent whole set $B^{(c)}$
Set O_1:=$B^{(1)}$
For i=2 to c
For all permutations $\Pi(P^{(i)})$ of columns of $P^{(i)}$
Find max_trace (O_i $\Pi(P^{(i)})$)
Trace of product
Matrix $P^T\Pi(U^{(i)})$ will be maximum.
Find O for ith step
$O = \frac{i-1}{i} O_{i-1} + \frac{1}{i}\Pi^{i}_{\max}$
Matrix element determines the weight of data
Get reconstructed data R_s are similar to B

Where we getting data from Bank records

Get data set B

$$O = \frac{i-1}{i} O_{i-1} + \frac{1}{i}\Pi^{i}_{\max}$$

O_m is the optimal clustering. Each matrix element determines the weight of data point belonging to the certain cluster. Were P $^{(i)}$ is the fuzzy clustering matrix where columns are clusters and rows are the data points. 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.
The 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.

1. Find area A enclosing all points in dataset.
2. 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 finding optimal clustering P for the given set, we calculated the incidence matrix and found the correlation between it and the original incidence matrix.

V. CONCLUSION

As the 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. And another important issue in this area, paper consider the possibility of original data distribution restoration from the available perturbed dataset. 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 overcome this problem 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. This paper propose the clustering base data perturbation method to provide secure and effective range query services for protected data in the cloud. These approaches generally attempt to hide the sensitive data by randomly adapting the data values using some preservative noise and aim to rebuild the original distribution closely at an aggregate level. To examine our proposition, we consider the two-dimensional dataset, where the data points are grouped into four elliptic-shaped partitions. Now, given the perturbed dataset we use clustering algorithms, to cluster the perturbed dataset, which is to find the original partitions. To perturb data, we apply our algorithm, therefore masking real values of data points. 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.

VI. REFERENCES

[1]. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. K. andAndy Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A berkeley view of cloud computing," Technical Report, University of Berkerley, 2009.

[2]. 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.

[3]. Mahdi Zamanighomi, Mostafa Zamanian, Michael Kimber, Zhengdao Wang, "Gene regulatory network inference from perturbed time-series expression data via ordered dynamical expansion of non-steady state actors", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Sep. 29, 2014.

[4]. Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavkovic, Caroline Uhler, "Scalable privacy-preserving data sharing methodology for genome-wide association studies", Journal of Biomedical Informatics, Elsevier 133–141, 2014.

[5]. Yang, W. J. "Privacy protection by matrix transformation." IEICE Transactions on Information and Systems, E92-D(4), 740-741 2009.

[6]. 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

[7]. 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.

[8]. 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.

[9]. 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.

[10]. 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.

[11]. Giannella C and Liu K. “On the Privacy of Euclidean Distance Preserving Data Perturbation.” Compute 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]. Huiqi Xu, Shumin Guo, Keke Chen, “Building Confidential and Efficient Query Services in the Cloud with RASP Data Perturbation”, IEEE TKDE, December 2012.

[14]. 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.