IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Dimensionality Reduction for Privacy Preserving Data Mining using Random Projection Perturbation Approach in Outsourced Environment

Vijayalakshmi Pasupathy [1], YamunaDevi S [2]

Asst. Professor, Department of CSE[1,2]

Abstract: Data perturbation is a popular technique in privacy-preserving data mining. A major challenge in data perturbation is to balance privacy protection and data utility, which are normally considered as a pair of conflicting factors. We discuss that selectively preserving the task/model specific information in perturbation will help achieve better privacy guarantee and better data utility. One type of such information is the multidimensional geometric information, which is implicitly utilized by many data mining models. To preserve this information in data perturbation, we propose the Random Projection Data Perturbation (RPP) method. In this paper, we describe several aspects of the RPP method. Methods of dimensionality reduction provide a way to understand and visualize the structure of complex data. Recently, they have been proposed for ensuring that a given data, in a lower space, are protected against privacy threats, and meanwhile expose many of the useful and interesting properties of the original data. Dimensionality reduction methods assume that the data records are represented as vectors in a multidimensional space where each dimension represents a single attribute.

Keywords: Privacy-preserving Data Mining, Data Perturbation, Random Projection Perturbation, Privacy Evaluation, Data Mining Algorithms

1. INTRODUCTION

With the rise of cloud computing, service-based computing is becoming the major paradigm (Amazon, n.d.; Google, .d.). Either to use the cloud platform services or to use existing services hosted on clouds, users will have to export their data to the service provider. Since these service providers are not within the trust boundary, the privacy of the outsourced data has become one of the top-priority problems . As data mining is one of the most popular data intensive tasks, privacy preserving data mining for the outsourced data has become an important enabling technology for utilizing the public computing resources. Different from other settings of privacy preserving data mining such as collaboratively mining private datasets from multiple parties, this paper will focus on the following setting: the data owner exports data to and then receives a model (with the quality description such as the accuracy for a classifier) from the service provider. This setting also applies to the situation that the data owner uses the public cloud resources for large-scale scalable mining, where the service provider just provides computing infrastructure.

We present a new data perturbation technique for privacy preserving outsourced data mining in this paper. A data perturbation procedure can be simply described as follows. Before the data owners publish their data, they change the data in certain way to disguise the sensitive information while preserving the particular data property that is critical for building meaningful data mining models. Perturbation techniques have to handle the intrinsic tradeoff between preserving data privacy and preserving data utility, as perturbing data usually reduces data utility. Several perturbation techniques have been proposed for mining purpose recently, but these two factors are not satisfactorily balanced.

In this paper, we propose a new *multidimensional* data perturbation technique: random projection data perturbation that can be applied for several categories of popular datamining Models with better utility preservation and privacy preservation.

1.1. Data Privacy vs. Data Utility

Perturbation techniques are often evaluated with two basic metrics: the level of preserved privacy guarantee and the level of preserved data utility. Data utility is often task/model-specific and measured by the quality of learned models. An ultimate goal for all data perturbation algorithms is to maximize both data privacy and data utility, although these two are typically representing conflicting goals in most existing perturbation techniques. Data privacy is commonly measured by the difficulty level in estimating the original data from the perturbed data. Given a data perturbation

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

technique, the more difficult the original values can be estimated from the perturbed data, the higher level of data privacy this technique provides.

The level of data utility typically refers to the amount of critical information preserved after perturbation. More specifically, the critical information should be task or model oriented. For example, decision tree and k-Nearest-Neighbor (kNN) classifier for classification modeling typically utilize different sets of information about the datasets: decision tree construction primarily concerns the related column distributions; the kNN model relies on the distance relationship which involves all columns. Most of existing perturbation techniques do not explicitly address that the critical information is actually task/model-specific.
We are able to provide better quality guarantee on both privacy and model accuracy.

1.2. Contributions and Scope

We have developed the random projection data perturbation approach to privacy preserving datamining. In contrast to other perturbation approaches ([2];[3];[4];[5]), our method exploits the task and model specific multidimensional information about the datasets and produces a robust data perturbation method that not only preserves such critical information well but also provides a better balance between the level of privacy guarantee and the level of data utility. The contributions of this paper can be summarized into three aspects.

We study whether random projection perturbation [6] can be an alternative component in geometric data perturbation, based on the formal analysis of the effect of multiplicative perturbation to model quality. We use the Gaussian mixture model [6] to show in which situations the multiplicative component can affect the model quality. It helps us understand why the rotation component is a better choice than other multiplicative components in terms of preserving model accuracy.

2. RELATEDWORK

A considerable amount of work on privacy preserving data mining methods have been reported in recent years ([2]; [3]; [7] ;[8]; [9]; [10]).

The most relevant work about perturbation techniques for data mining includes the random noise addition methods ([3]; [9]), the condensation-based perturbation [1], rotation perturbation ([11]; [4]) and projection perturbation [5]. In addition, k-anonymization [13] can also be regarded as a perturbation technique, and there is a large body of literatures focusing on the k-anonymity model [12].

Noise Additive Perturbation The typical additive perturbation technique [3] is column-based additive randomization. This type of techniques relies on the facts that 1) Data owners may not want to equally protect all values in a record, thus a column-based value distortion can be applied to perturb some sensitive columns. 2) Data classification models to be used do not necessarily require the individual records, but only the column value distributions[3] with the assumption of independent columns. The basic method is to disguise the original values by injecting certain amount of additive random noise, while the specific information, such as the column distribution, can still be effectively reconstructed from the perturbed data.

Condensation-based Perturbation The condensation approach [2] is a typical multi-dimensional perturbation technique, which aims at preserving the covariance matrix for multiple columns. Thus, some geometric properties such as the shape of decision boundary are well preserved. Different from the randomization approach, it perturbs multiple columns as a whole to generate the entire "perturbed dataset". As the perturbed dataset preserves the covariance matrix, many existing data mining algorithms can be applied directly to the perturbed dataset without requiring any change or new development of algorithms.

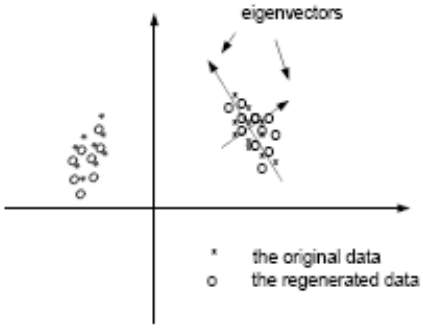

Fig.1 Condensation approach

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Rotation Perturbation preserves Euclidean distance and inner product of data points preserves geometric shapes such as hyperplane and hyper curved surfaces in the multidimensional space.

Random Projection Perturbation Random projection perturbation refers to the technique of projecting a set of data points from the original multidimensional space to another randomly chosen space.

In this paper we have completely analyzed about the random projection perturbation for outsourced environment.

3. PRELIMINARIES

In this section, we first give the notations and then define the components in random projection perturbations. The datasets discussed in this paper are all numerical data.

3.1 Training Dataset

Training dataset is the part of data that has to be exported/published in privacy-preserving data classification or clustering. A classifier learns the classification model from the training data and then is applied to classify the unclassified data.

Suppose that X is a training dataset consisting of N data rows (records) and *d* columns (attributes, or dimensions). For the convenience of mathematical manipulation, we use $X_{d\times N}$ to denote the dataset, i.e., $X = [x_1 \ldots x_N]$, where x_i is a data tuple, representing a vector in the real space R_d. Each data tuple x_i belongs to a predefined class if the data is for classification modeling, which is indicated by the class label attribute y_i. The data for clustering do not have labels. The class label can be nominal (or continuous for regression), which is public, i.e., privacy-insensitive. All other attributes containing private information needs to be protected. Unclassified dataset could also be exported/published with privacy-protection if necessary.

If we consider X is a sample dataset from the d-dimension random vector $[X_1, X_2, \ldots, X_d]^T$, we use bold X_i to represent the random variable for the column *i*.

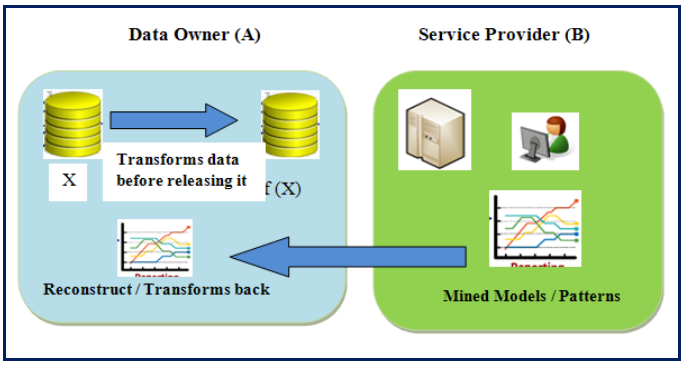

Fig.2 Applying Random Projection data perturbation to outsourced data

3.2 Framework and Threat Model for Applying Random Projection Data Perturbation

We study random projection data perturbation under the following framework (Figure 2). The data owner wants to use the data mining service provider (or the public cloud service provider). The outsourced data needs to be perturbed first and then sent to the service provider. Then, the service provider develops a model based on the perturbed data and returns it to the data owner, who can use the model either by transforming it back to the original space or perturb new data to use the model. In the middle of developing models at the service provider, there is no additional interaction happening between the two parties. Therefore, the major costs for the data owner incur in optimizing perturbation parameters that can use a sample set of the data and perturbing the entire dataset.

We take the popular and reasonable honest-but-curious service provider approach for our threat model. That is, we assume the service provider will honestly provide the data mining services. However, we also assume that the provider might look at the data stored and processed on their platforms. Therefore, only well-protected data can be processed and stored on such an untrusted environment.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

4. RANDOM PROJECTION PERTURBATION

4.1 Definition

Random Projection (RP) aims to protect the original data values, whilst preserving the data utility, by projecting data objects in n-dimensional space into a lower p dimensional
space, where $p < n$, capturing as much of the variation of the data as possible.

The RP can be defined by
$Y = XR;$
Where R is an n x p RP matrix onto p-subspace such that each column is orthogonal and the elements rij have zero mean and unit variance. Let A be a matrix whose columns are linearly independent vectors, then the projection of matrix X into the subspace of the columns of A is known to be $R = A(A^TA)^{-1}A^T$. Note that even though A still embeds X into the lower dimensional space, it is no longer an isometry in general.

This approach is fundamentally based on the result of Johnson-Lindenstrauss lemma [14] which says that any n points subset of Euclidean space can be embedded into a random subspace of $p = O(\log n = \in^2)$ dimensions without distorting the pair-wise distances by more than a factor of $(1 \pm \in)$, for any $0 < \in < 1$. This implies that there is a transformation $T : R^n \rightarrow R^p$ such that the distances between the points are approximately preserved.

Let x and y be two points in the higher dimension, Rn, T(x) and T(y) be their images in the lower dimension, Rp, there exists $\in > 0$ such that the distance between x and y and their images T(x) and T(y) is bounded by

$$(1 - \in) \| x - y \| \leq \| T(x) - T(y) \| \leq (1 + \in) \| x - y \|.$$

By using such a transformation, it would be possible to change the original form of data whilst maintaining the distance properties by a small error $\in$. However, since the pair wise distances are not strictly preserved but rather maintained with some distortion $\in$, the accuracy of data mining model may still be negatively affected. Assume that data points of the original data are represented as column vectors in matrix X, i.e. X is an n x m matrix, define a perturbation model that preserves the inner product as

$$Y = \frac{1}{\sqrt{p\sigma}} XR;$$

where each entry r_{ij} of R is independent and identically distributed chosen from a distribution with mean $\mu = 0$ and standard deviation σ . It has been proved that $E[R^TR] = n\sigma^2 I$, where n is the number of rows of matrix R, and I is the identity matrix. The values of the original data X can be estimated as $E[Y^TY] = X^TX$ since the entries of the random matrix are independent and identically distributed.

Geometry of data gets perturbed by random projection, but not too much:

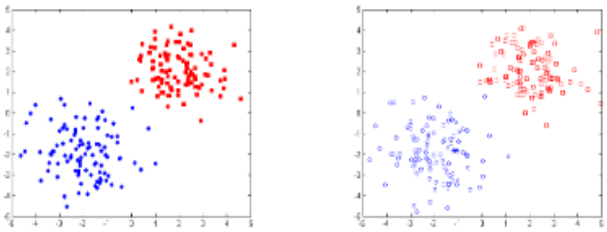

Fig 3. (a) Original Data (b) RP Data (schematic)

4.2 Attacks to Dimensionality Reduction

The preservation of privacy in dimensionality reduction seems better than other data anonymisation and randomization methods, there are still some major challenges including measuring the level of uncertainty in the perturbed data and ensuring the resilience of the perturbed data against data disclosure. For most data randomization techniques, if more is known about the original data, then the probability of breaching the privacy model is high as these techniques are usually dependent on a transformation basis to map the data. This implies that the perturbed data, in most cases, contain

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

much of the statistical properties which can then be exploited by privacy attacks to estimate the transformation matrix and thus recover the original data. Therefore, the success of these attacks basically depends on how much information is still embedded in the data and how this information is available to the attacker. The quantification of uncertainty in dimensionality reduction models can be evaluated by assuming that prior knowledge about the original data is available to the attacker. The prior knowledge can be used within the inference process to effectively estimate the original data. For example, one can consider a scenario when the attacker knows some original data points, their images in the perturbed data and their distances from a point under attack. That is, the disclosure may occur by measuring the distance from the attacked point to the other known points and minimizing the sum of squared errors using some heuristic methods [15].

Another possible attack scenario can be described when a sample of the original data or the distribution from where the original data are drawn is available to the attacker. In this case, the attacker can estimate the original data by examining the relationship between the principle eigenvectors of the known sample and the principle eigenvectors of the perturbed data. Intuitively, a large sample size will give the attacker a better recovery because large sample sizes tend to minimize the probability of errors, and thereby maximize the accuracy of estimating the original data. The attacker would attempt to find a transformation that composes a set of the eigenvectors obtained from both the known sample and the perturbed data and then project the data onto these eigenvectors such that the principle directions of the perturbed data are aligned as much as possible with principle directions of the known sample. The robustness of the attack basically depends on the estimation of the covariance matrix [16].

4.3 $\in$ - Distortion Mapping

Projecting data into a lower dimensional space usually results in some distortion of the distance relationships. It is very rare to find a mapping between two spaces of interest in which distances are exactly preserved. Obviously, we often allow the mapping to alter the distances in some fashion but hopefully with restricted damages as much as possible. The metric space is a set of points with a global function that measures the degree of closeness or distance of pairs of points in this set [17].

5. EVALUATION OF PRIVACY AND INFORMATION LOSS

The success of any distance-based data mining depends significantly on finding a metric that reflects reasonably well the important relationships between the objects. The metric is usually defined by the distance measured from one object to another in the space holding these objects. Therefore, to minimize data distortion, we need a transformation that can preserve the distance between all points and allow useful patterns to be easily discovered from the perturbed data. It is critically important to measure both privacy and utility using certain criteria. Otherwise, maximizing utility may lead to privacy violations as these two factors are often mutually contradictory.

Evaluation of privacy is a challenging task since it depends on many factors including what is already known (prior knowledge) to the attacker and the nature of the technique used to perturbed the data. In general, the privacy breach can be described in terms of how well the original data values can be estimated or reconstructed from the perturbed data. It is inversely proportional to the level of protection offered by the perturbation technique.

In PPDM, most methods depend on data randomisation in order to sanitise the original data values using additive or multiplicative noise. However, a key weakness of data randomization methods is that the perturbed data, in most cases, contain much of the statistical proprieties which can then be exploited by privacy attacks. Therefore, the success of these attacks mainly depends on how information is still embedded in the data and how this information is available to the attacker.

6. CONCLUSION

The choice of which approach to use to perturb the data is crucial, but essentially the perturbation method should not compromise either privacy or utility. PCA provides very good data utility, it is vulnerable to some distance based privacy attacks since the location of the original data points can be estimated when some prior knowledge is available to the attacker RP, SVD and DCT approaches cause more distortion to the data, and therefore, better privacy would be achieved. However, the large size of distortion negatively affects the utility of the data, and thus they seem inefficient, especially if the analysis utilizes the distance between data objects. Using RP causes some distance distortion, specially at the low dimensions, but, interestingly, the accuracy is highly competitive at the higher dimensions.

A trade-off between privacy and accuracy need to be determined so that the data owner can choose an appropriate lower dimension and transform the data to that dimension.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

The Random Projection Perturbation is Linear, Cheap, universal. Oblivious to data distribution. Target dimension doesn't depend on data dimensionality and it is tractable to analysis.

REFERENCES

[1] Vijayalakshmi Pasupathy and N.Priya, Dimensionality Reduction for Privacy Preserving Data Mining using Random Projection Perturbation Approach in Outsourced Environment.
[2] Aggarwal, C. C. and Yu, P. S. (2004), A condensation approach to privacy preserving data mining, in 'Proceedings of Intternational Conference on Extending Database Technology (EDBT)', Vol. 2992, Springer, Heraklion, Crete, Greece, pp. 183–199.
[3] Agrawal, R. and Srikant, R. (2000), Privacy-preserving data mining, in 'Proceedings of ACM SIGMOD Conference', ACM, Dallas, Texas.
[4] Chen, K. and Liu, L. (2005), A random rotation perturbation approach to privacy preserving data classification, in 'Proceedings of International Conference on Data Mining (ICDM)', IEEE, Houston, TX.
[5] Liu, K., Kargupta, H. and Ryan, J. (2006), 'Random projection-based multiplicative data perturbation for privacy preserving distributed data mining', IEEE Transactions on Knowledge and Data Engineering (TKDE) 18(1), 92–106.
[6] McLachlan, G. and Peel, D. (2000), Finite Mixture Models, Wiley.
[7] Clifton, C. (2003), Tutorial: Privacy-preserving data mining, in 'Proceedings of ACM SIGKDD Conference'.
[8] Agrawal, D. and Aggarwal, C. C. (2002), On the design and quantification of privacy preserving data mining algorithms, in 'Proceedings of ACM Conference on Principles of Database Systems (PODS)', ACM, Madison, Wisconsin.
[9] Evfimievski, A., Srikant, R., Agrawal, R. and Gehrke, J. (2002), Privacy preserving mining of association rules, in 'Proceedings of ACM SIGKDD Conference'.
[10] Vaidya, J. and Clifton, C. (2003), Privacy preserving k-means clustering over vertically partitioned data, in 'Proceedings of ACM SIGKDD Conference'.
[11] Oliveira, S. R. and Zaiane, O. R. (2010), 'Privacy preserving clustering by data transformation', Journal of Information and Data Management (JIDM) 1(1).
[12] Fung, B. C.,Wang, K., Chen, R. and Yu, P. S. (2010), 'Privacy-preserving data publishing: A survey on recent developments', ACM Computer Survey .
[13] Sweeney, L. (2002), 'k-anonymity: a model for protecting privacy', International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5).
[14] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics, 26:189 - 206, 1984
[15] W. Navidi, W. S. Murphy, and W. Hereman. Statistical methods in surveying by trilateration. Computational statistics & data analysis, 27(2):209, 227,1998.
[16] K. Liu, C. Giannella, and H. Kargupta. An attacker's view of distance preserving maps for privacy preserving data mining. In J. Frnkranz, T. Scheer, and M. Spiliopoulou, editors, Knowledge Discovery in Databases: PKDD 2006, volume 4213 of Lecture Notes in Computer Science, pages 297{308. Springer, Berlin, Heidelberg, 2006
[17] B. Mendelson. Introduction to topology. Dover Publications, New York, USA, 3nd edition, 1990.
[18] L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. Journal of the American Statistical Association, 104(485):209{219, 2009.
[19] Sadun, L. (2001), Applied Linear Algebra: the Decoupling Principle, Prentice Hall.
[20] Stewart, G. (1980), 'The efficient generation of random orthogonal matrices with an application to condition estimation', SIAM Journal on Numerical Analysis 17.
[21] Sweeney, L. (2002), 'k-anonymity: a model for protecting privacy', International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5).
[22] Teng, Z. and Du,W. (2009), 'A hybrid multi-group approach for privacy-preserving data mining', Knowledge and Information Systems 19(2).
[23] Vempala, S. S. (2005), The Random Projection Method, American Mathematical Society.