Journal of King Saud University – Computer and Information Sciences (2016) **xxx**, xxx–xxx

King Saud University
Journal of King Saud University – Computer and Information Sciences
www.ksu.edu.sa
www.sciencedirect.com

Privacy preserving data mining with 3-D rotation transformation

Somya Upadhyay [a], Chetana Sharma [a], Pravishti Sharma [a], Prachi Bharadwaj [b], K.R. Seeja [b,*]

[a] *Department of Computer Science & Engineering, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi 110006, India*
[b] *Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi 110006, India*

Received 1 August 2016; revised 14 November 2016; accepted 19 November 2016

KEYWORDS
Data perturbation;
Variance;
Three dimensional rotation;
Privacy preserving;
Data mining

Abstract Data perturbation is one of the popular data mining techniques for privacy preserving. A major issue in data perturbation is that how to balance the two conflicting factors – protection of privacy and data utility. This paper proposes a Geometric Data Perturbation (GDP) method using data partitioning and three dimensional rotations. In this method, attributes are divided into groups of three and each group of attributes is rotated about different pair of axes. The rotation angle is selected such that the variance based privacy metric is high which makes the original data reconstruction difficult. As many data mining algorithms like classification and clustering are invariant to geometric perturbation, the data utility is preserved in the proposed method. The experimental evaluation shows that the proposed method provides good privacy preservation results and data utility compared to the state of the art techniques.

1. Introduction

There are many data mining techniques that have enabled successful extraction of patterns and knowledge from huge amounts of data. Organizations use this information for decision making in order to gain customer satisfaction. While data mining is providing successful advancements in areas like machine learning, statistics and artificial intelligence, it is often associated with the mining of information that can compromise confidentiality. This aspect supports increasing ethical concerns regarding sharing of personal information for data mining activities (Alan, 1999). Privacy preserving data mining (PPDM), techniques transform the data to preserve privacy. PPDM is not only to preserve privacy during mining phase but also needs to consider the privacy issues in other phases of knowledge discovery like data preprocessing and postprocessing (Xu et al., 2014). It addresses the problems faced by an organization or person when the sensitive information lost

* Corresponding author.
E-mail addresses: 23.7saumya@gmail.com (S. Upadhyay), sharma.chetana12@gmail.com (C. Sharma), pravishti21@gmail.com (P. Sharma), prachibhardwaj57@gmail.com (P. Bharadwaj), krseeja@gmail.com, seeja@igdtuw.ac.in (K.R. Seeja).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2016.11.009
1319-1578 © 2016 Production and hosting by Elsevier B.V. on behalf of King Saud University.

or misused by the third party data miner. Hence the data need to be modified so that the third party data miner will not get any idea of the sensitive information. At the same time the utility of the data should be preserved. The aim of data perturbation is to release aggregate information that can be used for mining, without leaking individual information by introducing uncertainty about individual values (Agrawal and Srikant, 2000). It is found that selectively preserving multidimensional geometric information will help to achieve better privacy as well as data utility. Many data mining models like linear classifiers, support vector machine and Euclidean distance based clustering algorithms are invariant to geometric perturbation (Chen and Liu, 2011). This means that the classifiers trained on the geometrically perturbed data and that trained with original data have almost the same accuracy. In this paper a three dimensional geometric rotation of data is proposed to perturb the data before releasing it to the third party data miner.

2. Literature review

Over the past few years, several approaches have been proposed by various research groups for privacy preserving data mining. Initially few basic methods like random addition and multiplication were introduced which were prone to almost all kinds of attacks. Later, some efficient techniques that maintain the balance between data utility and privacy are also proposed. Some of the major approaches (Aggarwal and Philip, 2008) are data perturbation, data swapping, k-anonymization, cryptography based methods, rule hiding methods and secure distributed mining techniques.

There are two major data perturbation approaches namely probability distribution approach and data value distortion approach. In probability distribution approach (Liew et al., 1985), the data are replaced with another sample from the same distribution. In data value distortion, data elements are perturbed by either additive noise, multiplicative noise or some other randomization procedures. Noise Additive Perturbation perturbs the dataset by the addition of noise. Generally the Gaussian distribution is used to generate the noise value. The more the correlation of noises is similar to the original data, the more the preservation of privacy. Principal Component Analysis (PCA) and Bayes Estimate (BE) techniques have been extensively studied to estimate the reconstruction aversion of randomization techniques (Huang et al., 2005). Other methods of perturbation include multiplicative perturbation (Chen and Liu, 2008), rotation perturbation (Huang et al., 2005; Chen and Liu, 2011) and multi-dimensional perturbation (Chen and Liu, 2005). In another approach (Oliveira and Zaane, 2004) logarithmic transformation is applied to the data first, and then a predefined multivariate Gaussian noise is added and then took the antilog of the noise-added data.

In data swapping (Fienberg and McIntyre, 2004) the database is transformed by swapping values of sensitive attributes among records and hence create uncertainty about the sensitive data. k-Anonymity model (Sweeney, 2002; Gionis and Tassa, 2009) uses data generalization and suppression methods and the data are released only if the information for each person contained in the release cannot be distinguished from at least (k-1) other people. In kd-tree based perturbation method (Li and Sarkar, 2006) data are partitioned recursively into smaller subsets and the sensitive data in the subsets are perturbed using the subset average. A privacy preserving distributed data mining technique based on multiplicative random projection matrices (Liu et al., 2006) is proposed to preserve the statistical characteristics of data while improving the privacy level. Cryptographic techniques (Pinkas, 2002) are also proposed for privacy preserving data mining. Chen et al. propose a multiparty collaborative privacy preserving mining method (Chen and Liu, 2009) that securely unifies multiple geometric perturbations that are preferred by different parties using concept of keys. In Association Rule Hiding approach (Verykios et al., 2004) the database is transformed to hide the sensitive rules. New data mining algorithms like random decision tree (Vaidya et al., 2014), modified Bayesian network (Yang and Wright, 2006) and SVM classifier (Lin and Chen, 2011) specially for PPDM are also proposed.

This paper aims to take forward the work done in (Oliveira and Zaane, 2004) where two dimensional rotations have been used as a method for data modification in order to preserve privacy. In the proposed approach the attributes are divided in groups of three and then rotation perturbation is applied such that the data preserve their internal Euclidean distances.

3. Materials and methods

3.1. Min–Max normalization

The normalization method used is the MIN_MAX method. This method maps the value of an attribute v lies between the range min and max to a new value v' which lies between the range *newmin* and *newmax*.

$$v' = (v - min/(max - min)) \times (newmax - newmin) + newmin$$

Here to standardize the data, all the attributes values are mapped between a range 0.0 and 5.0

3.2. Three dimensional rotation (3DR)

In 2DR the axis of rotation is always perpendicular to the xy plane, i.e., the Z axis. In 3DR the axis of rotation can have any spatial orientation. i.e., X-axis or Y-axis or Z-axis depending on the underlying plane. The rotation matrices, equations and spatial representations for each of the axes of rotation are listed in Table 1.

In double rotation the data are rotated twice along different axes for better data perturbation i.e., three axes pairs xy, yz and xz. Using the associative nature of matrix, the rotation matrices R_{xy}, R_{yz} and R_{xz} can be calculated as shown in Fig. 1.

3.3. Proposed method

In this paper a 3-dimensional rotation transformation (3DRT) approach is proposed which distorts the data by rotating three attributes at a time along two different axes without compromising the mining results.

3.3.1. Pre-processing

The data matrix D is assumed to have only numeric attributes. The data matrix before perturbation needs to be normalized to standardize it so that during rotation the Euclidean distance

Please cite this article in press as: Upadhyay, S. et al., Privacy preserving data mining with 3-D rotation transformation. Journal of King Saud University – Computer and Information Sciences (2016), http://dx.doi.org/10.1016/j.jksuci.2016.11.009