Journal of King Saud University – Computer and Information Sciences (2018) 30, 524–530

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, Chetana Sharma, Pravishti Sharma, Prachi Bharadwaj, K.R. Seeja *

Department of Computer Science & Engineering, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi 110006, India

Received 1 August 2016; revised 14 November 2016; accepted 19 November 2016
Available online 28 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 or misused by the third party data miner. Hence the data need

* 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

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 the 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.

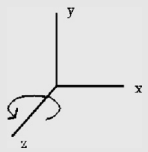
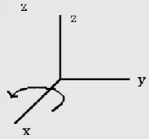
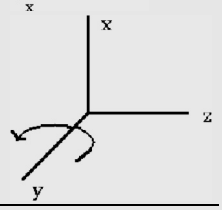
3.3. Proposed method

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 between points remains almost the same. The normalization method used is the *MIN_MAX* method.

Table 1 Three dimensional rotation.

Axis of Rotation	Equations	Rotation Matrix	Spatial Representation
Z-Axis Rotation	$x' = x * \cos\Theta - y * \sin\Theta$ $y' = x * \sin\Theta + y * \cos\Theta$ $z' = z$	$R_z(\theta) = \begin{pmatrix} ccc\cos\theta & -sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$	
X-axis Rotation	$y' = y * \cos\Theta - z * \sin\Theta$ $z' = y * \sin\Theta + z * \cos\Theta x' = x$	$R_x(\theta) = \begin{pmatrix} ccc1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$	
Y-axis Rotation	$z' = z * \cos\Theta - x * \sin\Theta$ $x' = z * \sin\Theta + x * \cos\Theta$ $y' = y$	$R_y(\theta) = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}$	

$$R_{xy} = R_x \, x \, R_y = \begin{pmatrix} cos\theta & 0 & -sin\theta \\ \sin^2\theta & cos\theta & sin\theta\, cos\theta \\ sin\theta\, cos\theta & -sin\theta & \cos^2\theta \end{pmatrix}$$

$$R_{yz} = R_y \, x \, R_z = \begin{pmatrix} \cos^2\theta & -sin\theta\, cos\theta & -sin\theta \\ sin\theta & cos\theta & 0 \\ sin\theta\, cos\theta & -\sin^2\theta & cos\theta \end{pmatrix}$$

$$R_{xz} = R_x \, x \, R_z = \begin{pmatrix} cos\theta & -sin\theta & 0 \\ sin\theta\, cos\theta & \cos^2\theta & sin\theta \\ -\sin^2\theta & sin\theta\, cos\theta & cos\theta \end{pmatrix}$$

Figure 1 Rotation matrices.

3.3.2. Three dimensional rotation transformation (3DRT) algorithm

The proposed procedure to perturb the data matrix D has the following six steps:

Step 1. Select the axes pair for Rotation:

Select an axes pair q €{xy,yz,xz}. Calculate the rotation matrix for q.

$R_q = R_i x R_j$ where i, j € x, y, z and i≠j

Step 2. Group the attributes into triplets:

Group the attributes in k triplets (A_u, A_v, A_w) where u≠v≠w). The triplets are grouped sequentially. After grouping, if one attribute remains then, the last attribute left is grouped with previous two attributes. Similarly, if two attributes remain after grouping then, the last two attributes are combined with the attribute prior to them.

Step 3. Rotate the triplets around axes pairs q in three dimensional plane for different angles of rotation

Perform 3-D Rotation of each triplet V along selected axes pairs q to get the perturbed dataset D_q,

$$D_q : V(A'_u, A'_v, A'_w) = R_q \times V(A_u, A_v, A_w)$$

The values in the rotated datasets are a function of θ where R_q is the rotation matrix about q^{th} axis pair.

Step 4. Find the angle of rotation Θ

For the rotated dataset, do the following:

1. For each triplet, plot variance of difference of original and perturbed data sets v/s θ graph.
2. Derive three inequations for each triplet based on the constraints: Variance $(A_u - A'_u) \geqslant \rho_1$, Variance $(A_v - A'_v) \geqslant \rho_2$, Variance $(A_w - A'_w) \geqslant \rho_3$ where $\rho_1 > 0$, $\rho_2 > 0$ and $\rho_3 > 0$.
3. Find a range for θ that satisfies the security threshold for ρ_1, ρ_2 and ρ_3.. This range is referred as *security range*
4. From the intersection of security range of all triplets obtained in the previous step, choose a random real value α as θ.

Step 5. Perturb the data by rotating at an angle Θ

Calculate the rotated triplet at the angle Θ identified in step 4

i.e. $V'(A'_u, A'_v, A'_w) = R_q \times V(A_u, A_v, A_w)$

Step 6. Choosing the data to be released:

For each of these perturbed data sets, D_{xy}, D_{yz} and D_{xz}, calculate the variance. The data D_q with highest value of variance is selected as the final perturbed data D′.

3.3.3. Pseudo code

```
Input:  Normalized data matrix D,
        number of attributes n
        array of security thresholds S.
Output: Perturbed data matrix D'
  Three_Dimensional_Rotation_Transformation(D,S,n)
    Number of triplets,k ← ⌈n/3⌉;
    For each axes pair q ∈ {xy, yz, xz}
        Range(θq)=ϕ
        Tk(D)← k triplets (Au, Av, Aw ) in D such that 1≤u,  v,w ≤ n
                                                        and u≠v≠w.
        For different values of θ
          For each tk in Tk(D) //create a perturbed dataset
            Compute Dq: V (Au', Av', Aw') = Rq× V (Au, Av, Aw)
                                  where V is a function of θ.
          End_For
          Calculate the variance.
        End_For
        Plot Variance v/s θ graph
        Compute security range of θk,Range(θk),such that Variance(
        Au - Au') ≥ ρk1,Variance(Av - Av')≥ ρk2 and Variance(Aw - Aw') ≥
        ρk3 where ρk∈ Sk
        Range(θq) ← Range(θq) ∩ Range(θk)
        αq← rotation angle such that αq ∈ θq
        Compute Dqα :V (Au', Av', Aw') = Rαq × V (Au, Av, Aw)
        //Dqα denotes the data perturbed along q^th axis with angle α
        Calculate Vq = Variance of D qα
    End_For
    D'← Dqα with maximum Vq
  End_Algorithm
```

3.3.4. Complexity analysis

Let m be the number of objects and n be the number of attributes. The running time of proposed 3DRT algorithm is O(m × n)

4. Experiments

MATLAB Scripts have been used to generate variance-angle graphs and to perturb the dataset. Bank marketing dataset from the UCI repository (Moro et al., 2011, 2014) is used to verify the accuracy and efficiency of the 3DRT algorithm. Bank-Marketing dataset contains the details of 45211 phone calls used for direct marketing campaigns of a Portuguese banking institution. It has 16 input attributes among that seven are numerical and nine are categorical. Output variable is a binary class attribute with values yes or no. Only numeric attributes have been considered for perturbation. They are age, balance, day, duration, campaign, pdays, previous.

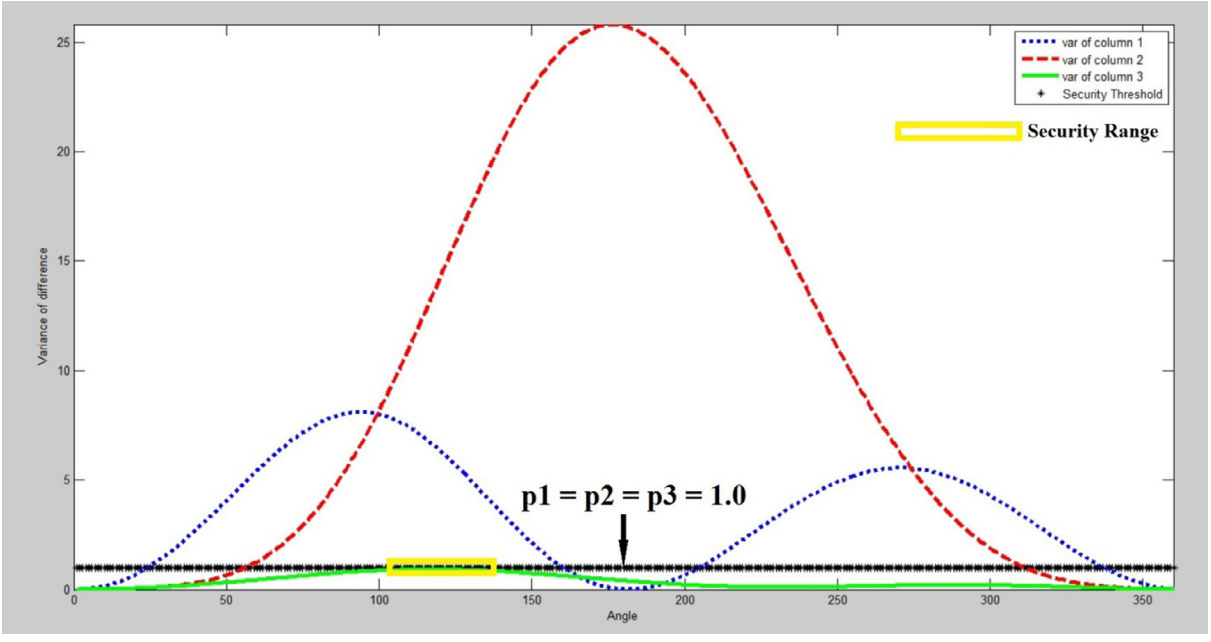


Figure 2 Variance v/s angle for t_1.

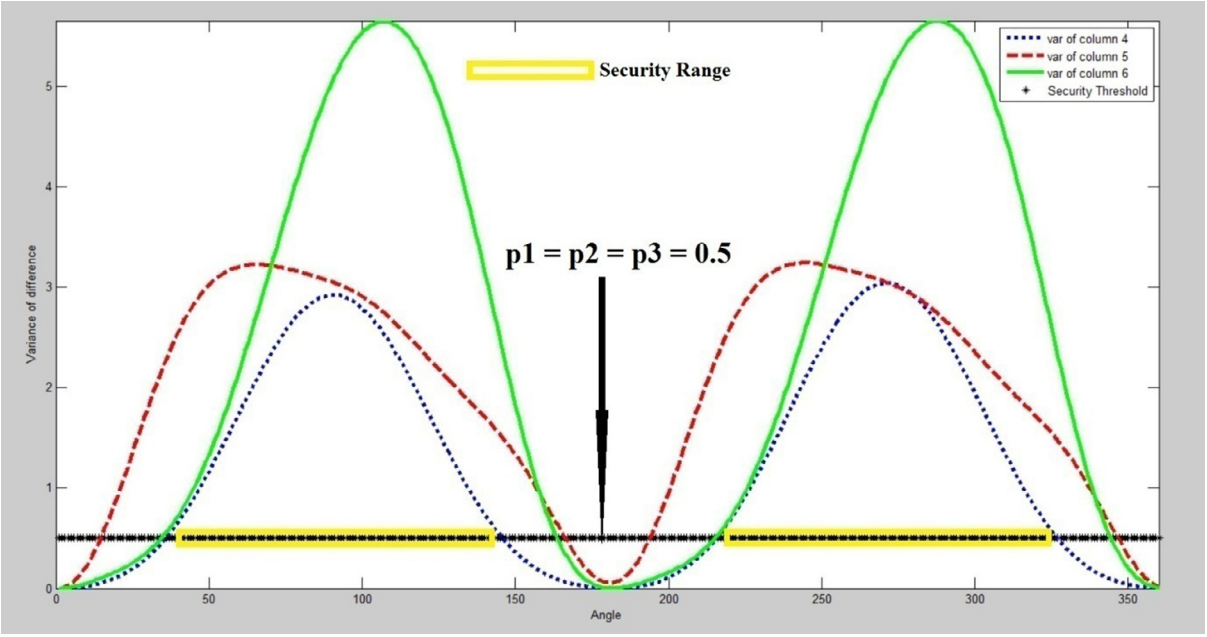


Figure 3 Variance v/s angle for t_2.

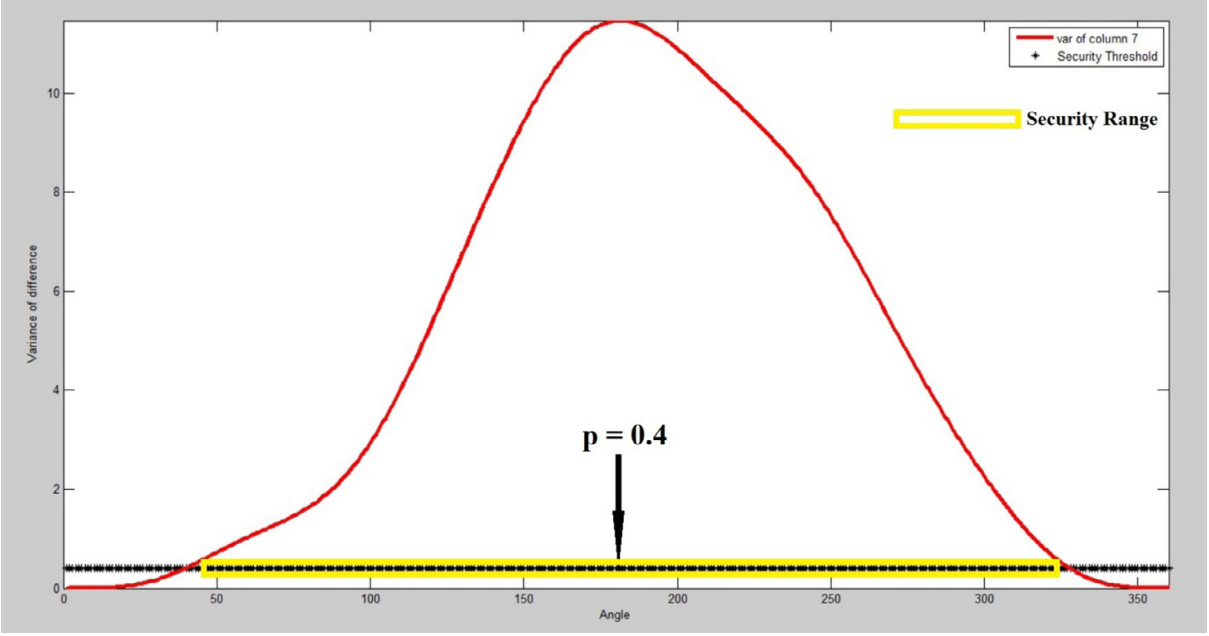


Figure 4 Variance v/s angle for t_3.

Categorical attributes can be protected using standard encryption algorithm DES.

Data set is normalized using MIN_MAX and convert the attribute values into the range [0,5]. The seven numeric attribute values are grouped into triplets as: t_1 = [attribute$_1$, attribute$_2$, attribute$_3$], t_2 = [attribute$_4$, attribute$_5$, attribute$_6$] and t_3 = [attribute$_5$, attribute$_6$, attribute$_7$].

Each triplet is assigned a security threshold as:

Table 2 Variance comparison for different pairs of axes-Bank Dataset.

Axes of rotation	Variance
y-z	2.9687 * e + 006
x-z	4.0039 * e + 005
x-y	1.077 * e + 003

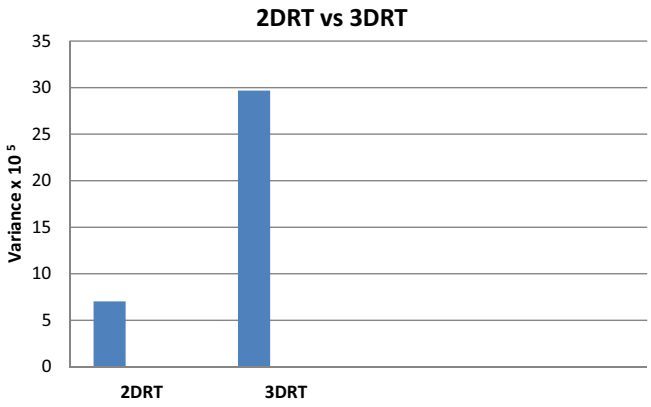


Figure 5 Privacy preserving capability comparison.

$S_1 = (1.0, 1.0, 1.0), \quad s_2 = (0.5, 0.5, 0.5), \quad s_3 = (0.4).$

Each triplet thus obtained is rotated about an axis pair (xy, yz, xz) with the aim to increase the amount of perturbation to elements without affecting spatial distances. For each rotation, corresponding rotation matrix is calculated as described in Fig. 1. Then for different values of Θ, calculate the rotated triplet $V' = V \times R_q$ and plot a graph between angle Θ and variances of the difference.

This gives a range for Θ such that minimum security threshold requirement for each attribute is satisfied by Θ. The variance vs angle graphs for the three triplets rotated along yz-axis are shown Figs. 2–4.

Security range for θ for t1 is 105–135 degrees, t2 is degrees 40–145 and 220–325 and t3 is 40–330. From the intersection of such ranges of θ, an angle α = 125 degrees is chosen which gives a decent level of maximized variance difference between original and perturbed data. Substituting the value of α in V′ for each triplet the transformed database D'_q is obtained. For Bank Database, the variance ρ after rotation about different axes pairs is given in Table 2.

From Table 2, it is found that for the bank dataset, y-z pair gives the best privacy measure as the variance corresponds to y-z pair is the highest. Therefore, the perturbed dataset corresponding to y-z pair is considered as final perturbed data D′.

5. Results and discussion

To verify the accuracy and efficiency of the proposed perturbation method, the result analysis has been done on two aspects – privacy preservation capability and mining accuracy.

5.1. Privacy preservation capability

The privacy preserving capability has been measured using variance metric. The variance is a numerical value used to indicate how widely items in a group vary. Here variance is used to measure the dissimilarity between original data and perturbed data. The variance is calculated by the formula:

$$\text{Variance}(x_1, x_2 \ldots\ldots x_n) = \frac{1}{n}\sum_{i=1}^{n}(|x_i - \mu|)^2$$

where μ denotes mean.

The variance-based privacy metric for the perturbed dataset D′ is defined as:

$$\rho = \text{Variance}\left(\frac{\text{Variance}(\text{col} - \text{col})}{\text{Variance}(\text{col})}\right)$$

where col represents column of original data matrix and col′ represents the corresponding column of perturbed data matrix.

The privacy preserving capability of the proposed 3DRT is compared with that of 2DRT (Oliveira and Zaane, 2004) and is shown in Fig. 5. It is found that 3DRT is four times better than 2DRT.

5.2. Data mining accuracy

In order to evaluate the effect of data perturbation in data mining capability, four classification models namely Naïve Bayes Classifier, Decision Table(rule based classifier), Ibk(k-nearest neighbor classifier) and J48(Decision tree classifier) are selected. Weka 3.6.9 is used to obtain data mining results

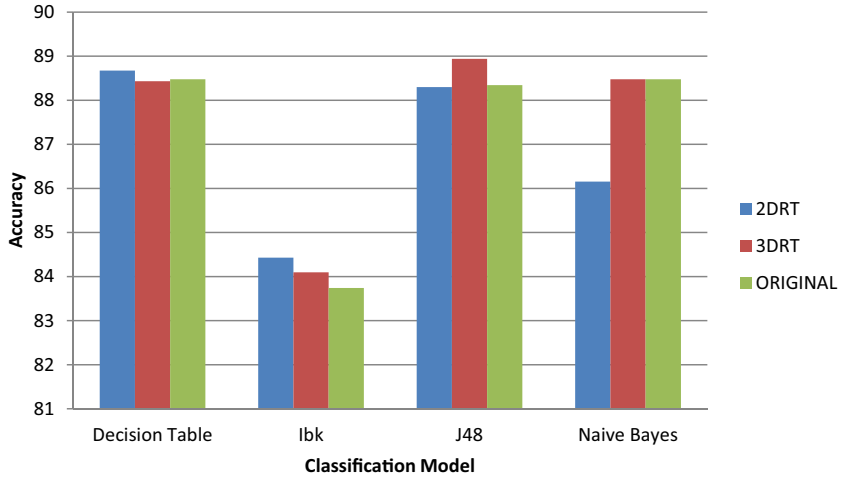


Figure 6 Data mining accuracy comparison.

for these classification models. The selected measure is classification accuracy and it is defined as the ratio of correctly classified instances to total number of instances. 10-fold cross validation is used for testing the classifiers. The 2-D perturbed, 3-D perturbed and original datasets were mined using the four classification algorithms and the results are shown in Fig. 6. It is found that the classification accuracy on perturbed data is almost equal to that of original data.

6. Conclusion

This paper presents a novel privacy preserving data transformation technique that can be used with different types of data mining models. In the proposed 3DRT technique the data are rotated twice along two different axes. This increases the variance, making the data more resilient to attacks. Moreover the perturbation of data is not affecting much the data mining capability of the data mining model because of preservation of Euclidean distances. The experimental results show that the data mining accuracy of the original and perturbed data are nearly same and has high variance, which shows its high privacy preserving capability.

References

Aggarwal, C.C., Philip, S.Y., 2008. A general survey of privacy-preserving data mining models and algorithms. In: Privacy-Preserving Data Mining. Springer, US, pp. 11–52.

Agrawal, R., Srikant, R., 2000. Privacy-preserving data mining. ACM Sigmod. Record 29 (2), 439–450. ACM.

Alan, J.B., 1999. Data Mining, the Internet, and Privacy. International WEBKDD'99 Workshop San Diego, CA, USA.

Chen, K., Liu, L., 2005. Privacy preserving data classification with rotation perturbation, Fifth IEEE International Conference on Data Mining (ICDM'05), p. 4.

Chen, K., Liu, L., 2008. A survey of multiplicative perturbation for privacy-preserving data mining. In: Privacy-Preserving Data Mining. Springer, US, pp. 157–181.

Chen, K., Liu, L., 2009. Privacy-preserving multiparty collaborative mining with geometric data perturbation. IEEE Trans. Parallel Distrib. Syst. 20 (12), 1764–1776.

Chen, K., Liu, L., 2011. Geometric data perturbation for privacy preserving outsourced data mining. Knowl. Inf. Syst. 29 (3), 657–695.

Fienberg, S. E., & McIntyre, J. (2004, June). Data swapping: Variations on a theme by dalenius and reiss. In International Workshop on Privacy in Statistical Databases (pp. 14–29). Springer, Berlin Heidelberg.

Gionis, A., Tassa, T., 2009. K-Anonymization with minimal loss of information. IEEE Trans. Knowl. Data Eng. 21 (2), 206–219.

Huang, Z., Du, W., & Chen, B. (2005, June). Deriving private information from randomized data. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (pp. 37–48). ACM.

Li, X.B., Sarkar, S., 2006. A tree-based data perturbation approach for privacy-preserving data mining. IEEE Trans. Knowl. Data Eng. 18 (9), 1278–1283.

Liew, C.K., Choi, U.J., Liew, C.J., 1985. A data distortion by probability distribution. ACM Trans. Database Syst. (TODS) 10 (3), 395–411.

Lin, K.P., Chen, M.S., 2011. On the design and analysis of the privacy-preserving SVM classifier. IEEE Trans. Knowl. Data Eng. 23 (11), 1704–1717.

Liu, K., Kargupta, H., Ryan, J., 2006. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans. Knowl. Data Eng. 18 (1), 92–106.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. In Proceedings of European Simulation and Modelling Conference-ESM'2011 (pp. 117–121). Eurosis.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22-31. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing(accessed 25.01.2016).

Oliveira, S. R., & Zaane, O. R. (2004). Data perturbation by rotation for privacy-preserving clustering., Technical Report TR04-17, Department of Computing Science, University of Alberta, Edmonton, AB, Canada.

Pinkas, B., 2002. Cryptographic techniques for privacy-preserving data mining. ACM Sigkdd Explor. Newslett. 4 (2), 12–19.

Sweeney, L., 2002. K-anonymity: a model for protecting privacy. Int. J. Uncertainty, Fuzz. Knowl. Based Syst. 10 (05), 557–570.

Vaidya, J., Shafiq, B., Fan, W., Mehmood, D., Lorenzi, D., 2014. A random decision tree framework for privacy-preserving data mining. IEEE Trans. Dependable Secure Comput. 11 (5), 399–411.

Verykios, V.S., Elmagarmid, A.K., Bertino, E., Saygin, Y., Dasseni, E., 2004. Association rule hiding. IEEE Trans. Knowl. Data Eng. 16 (4), 434–447.

Xu, L., Jiang, C., Wang, J., Yuan, J., Ren, Y., 2014. Information security in big data: privacy and data mining. IEEE Access 2, 1149–1176.

Yang, Z., Wright, R.N., 2006. Privacy-preserving computation of Bayesian networks on vertically partitioned data. IEEE Trans. Knowl. Data Eng. 18 (9), 1253–1264.