International Journal of Engineering & Technology, 7 (2) (2018) 884-886

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i2.12202
Research paper

The role of user defined function in privacy preserving data mining

G. Manikandan [1] *, A.Vamsi Krishna [1], P. Lakshmana Sarvagna [1]

[1] School of Computing, SASTRA Deemed University, Thanjavur, India
**Corresponding author E-mail: manikandan@it.sastra.edu*

Abstract

Data mining is used to retrieve the plugged in information from the huge data warehouse. Data mining techniques use wide variety of tools for extracting the required knowledge from the hefty data warehouse. Many organizations trust on the extracted knowledge for strategical decision making. In the downside these techniques also reveals some private information as part of the conversion process. Experts rely on various privacy preserving approaches to prevent the data disclosure. This paper primarily focuses on the use of a User-defined Function to maintain data privacy. The output of the proposed approach is compared with the 3-Dimensional geometric transformations such as Translation, Scaling and Shearing. From the experimental outcome it is evident that the proposed approach results in a minimal misclassification error when compared with the other data transformations.

Keywords*:Clustering; Misclassification error; Privacy; Scaling; Shearing; Translation.*

1. Introduction

Data is a vital element in the information technology domain. Every organization requires immense quantity of data for their experimental analysis. Usually the organization’s store their data in various storage medium like grid and cloud. Many organizations employ a wide variety of Data Mining techniques for effective strategic decision-making. Data mining plays a vital role in many fields like Medical Diagnosis, Weather Forecasting and Image Analysis. With advanced tools it is possible to extract a huge quantum of data in a smaller time period. One major disadvantage of data mining is that it also reveals some private information as part of the mining process. Different methods are already proposed by the researchers to address this privacy problem. The idea of privacy preserving approaches is to provide solution to the data miners without revealing any additional information which violates the user privacy. Usually the sensitive attribute is suppressed before sharing the data with the data miner. In spite of that, it is possible to infer sensitive information based on the other attributes known as the Quasi-Identifier attributes. To prevent privacy leakage, the data in the original database is modified using a sanitization mechanism. This sanitized data should exhibit the same behavior as the original data.

The primary objective of the privacy preserving algorithms is to create a modified data with the same characteristics as the original data. To ensure privacy, techniques like randomization, anonymization and secure multi-party computation are available in the literature. Randomization alters the data by adding some random noise before sharing. k-anonymization makes use of generalization and suppression. Generalization involves replacing a value with less specific but semantically reliable value. Suppression eliminates some rows or columns which results in the dimensionality reduction. In secure multi-party computation, Data Owners contribute to a protocol that generates valid data mining results without revealing the data to other parties.

In this paper we have proposed a user defined function to sanitize the original data. The output of this function is compared with the 3-dimensional geometrical transformations namely Translation, Scaling and Shearing. From the study, it is inferred that the user defined function results in a minimal misclassification error when compared with the geometrical transformations.

The rest of the paper is organized as follows. In Section 2, the existing privacy preserving data mining techniques are discussed. Section 3 provides the proposed user defined function. Section 4 presents the experimental results. Some conclusions are drawn in Section 5.

2. Literature survey

Geometric Data Perturbation (GDP) method using data partitioning and three-dimensional rotations was proposed [1]. In this method, attributes are divided into groups of three and each group of attributes is rotated about a 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.

In [2], a hybrid approach for achieving privacy during the mining procedure was proposed. The proposed approach is a two step process. In the first step, original data is sanitized using a geometric data transformation and in the second step min-max normalization is used to generate a normalized data. k-means clustering algorithm was used for the validation purpose.

This paper yields a complete review on Privacy-Preserving Data Mining techniques such as data partition, data modification and data restriction. These methods were used to prevent the data access from unauthorized users [3]. The authors listed a variety of methods, that can be used for ensuring data privacy and analyzed the merits and demerits of each technique.

In this paper, the authors proposed a new algorithm based on k-anonymity for privacy-preserving data publishing [4]. For privacy preservation k-anonymity is integrated with pattern-based multidimensional suppression (kPB-MS). This approach uses feature selection algorithm for dimensionality reduction.

In [5], the use of normalization techniques like Min-Max normalization, Z-Score normalization, and Decimal Scaling methods for preserving privacy were analyzed with respect to privacy and accuracy. To justify the output of Min-Max normalization k-means clustering algorithm was used.

Privacy protection is achieved by using a Geometric data perturbation (GDP) method [6]. This approach is compared with other multidimensional methods such as Random projection perturbation. From the analysis it is found that GDP results in better privacy than the other methods.

In [7], a technique that makes use of both randomization and cryptographic techniques to provide privacy was proposed. The proposed method securely constructs RDTs (Random Decision Tree) for both horizontally and vertically partitioned data sets. The proposed protocol was compared with the other available protocols using parameters like computation and communication cost.

An integrated approach using data perturbation and normalization was proposed to ensure better data privacy in a distributed environment. Shearing based composite data transformation method is used for performing privacy-preserving clustering. For validation purpose, k-means algorithm is used with the original data and the distorted data [8-9].

Data perturbation technique is used by the data owner to modify the original data. The modified data is encrypted using a cryptographic algorithm and the resultant cipher text is given as a result to the user query. From the experimental analysis it is proved that the proposed system results in better privacy and security [10].

2.1. 3D-translation

Translation is defined as repositioning of an object along a straight line path from one coordinate location to another coordinate location . This is a straightforward extension of the 2D translation transformation as shown in equation 1.1.

$$(x'\ y'\ z'\ 1) = (x\ y\ z\ 1) * \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ T_x & T_y & T_z & 1 \end{pmatrix}$$

$$\begin{aligned} x' &= x + T_x \\ y' &= y + T_y \\ z' &= z + T_z \end{aligned} \qquad (1.1)$$

2.2.3D scaling

Scaling is used to change the size of an object. Scaling can be achieved by multiplying the original coordinates of the object with the scaling factor to get the desired result. In 3D-Scaling operation, three coordinates are used. Let us assume that the original coordinates are (x,y,z) and scaling factors are (S_x, S_y, S_z) respectively, and the scaled coordinates are (x',y',z'). The formula used for 3D-Scaling is shown in equation 1.2

$$(x'\ y'\ z'\ 1) = (x\ y\ z\ 1) * \begin{pmatrix} S_x & 0 & 0 & 0 \\ 0 & S_y & 0 & 0 \\ 1 & 0 & S_z & 0 \end{pmatrix}$$

$$= [x.S_x\ \ y.S_y\ \ z.\ S_z] \qquad (1.2)$$

2.3. 3D-shearing

A transformation that slants the shape of an object is called the shear transformation. The formula used for 3D-Shearing is shown in equation 1.3

$$Sh = \begin{pmatrix} 1 & Sh^{y}{}_{x} & Sh_x{}^{z} & 0 \\ Sh_y{}^{x} & 1 & Sh_y{}^{z} & 0 \\ Sh_z{}^{x} & Sh_z{}^{y} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\begin{aligned} x' &= x + Sh^{y}{}_{x}\,.y + Sh_x{}^{z}\,.z \\ y' &= Sh_y{}^{x}\,.x + y + Sh_y{}^{z}\,.z \\ z' &= Sh_y{}^{x}\,.x + Sh_z{}^{y}\,.y + z \end{aligned} \qquad (1.3)$$

3. Proposed system

3.1. User defined function

Data privacy can be achieved by changing Original Data into Modified Data with the help of the function shown in equation 1.5. The constants which are used in this function are assigned so as to obtain the value of Modified Data in the given range. User defined function is derived from two dimensional Discrete cosine transform. The DCT transforms a signal from a spatial representation into a frequency representation. This DCT-II is most commonly used form of Discrete cosine transform.

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \qquad k = 0, \ldots, N-1. \qquad (1.4)$$

So, the User defined function used in this work is derived from the equation 1.4.

$$ModifiedAge = abs\{(OriginalAge)*$$
$$\cos[(pi/33)(originalAge+1/2)*55)]\}+1 \qquad (1.5)$$

4. Experimental results

For experimental purpose the Indian Liver Dataset from UCI repository is used. This data set contains information about 416 liver patient records and 167 non liver patients. The data set was collected from north east of Andhra Pradesh, India. This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90". This data set contains 10 attributes namely age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. The age attribute is used as the input for the proposed system.

Table 1: Comparison Table

Original Age	3D Translation	3D Scaling	3D Shearing	User Defined Function
17	20	51	119	14
24	27	72	168	20
27	29	54	135	24
30	32	60	150	25
31	33	62	155	28
42	44	84	210	36
48	50	96	240	41
50	52	100	250	53
55	57	110	275	44
64	66	128	320	63

The output obtained by applying the user defined function is shown in Table 1.From the table it can be inferred that the user defined function results in an optimal modified data when compared with the other geometrical data transformations. Figure 1 is the graphical representation of the same.

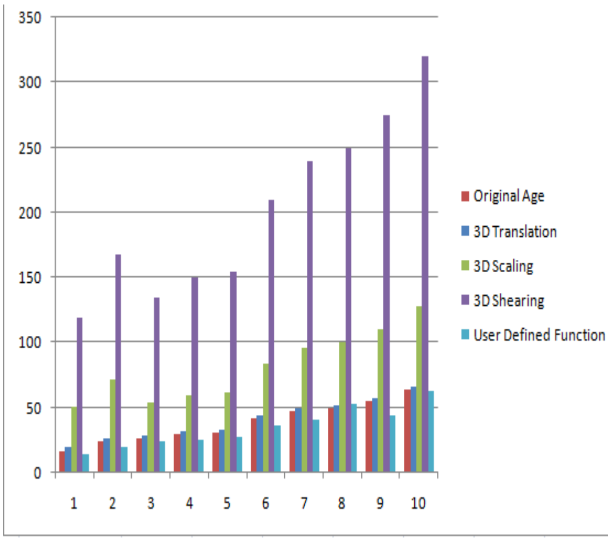

Fig. 1: Comparison Graph.

5. Conclusion

In this paper, a user defined function is created from the two dimensional discrete cosine transform for privacy preservation. From the experimental analysis it is inferred that the proposed scheme results in an optimal modified data. The modified data generated from the user defined function maintains data similarity i.e, the deviation between the original and the modified data is considerably small.

References

[1] Upadhyay, S. et al., “Privacy-preserving data mining with 3-D rotation transformation”, Journal of King Saud University – Computer and Information Sciences,pp.,(2016).

[2] G. Manikandan, N. Sairam, C. Saranya and S. Jayashree, "A Hybrid Privacy Preserving Approach in Data Mining", Middle-East Journal of Scientific Research, pp.581-85, 2013

[3] TamannaKachwala, Dr. L. K. Sharma, “A Literature analysis on Privacy Preserving Data Mining”, International Journal of Innovative Research in Computer and Communication Engineering, pp.2838-42, 2015.

[4] Aristos Aristodimou, Athos Antoniades, Constantinos S. Pattichis, “Privacy-preserving data publishing of categorical data through k-anonymity and feature selection”, Healthcare Technology Letters, pp.16-21, 2016.

[5] C.Saranya, G.Manikandan, “A Study on Normalization Techniques for Privacy Preserving Data Mining”, International Journal of Engineering and Technology, pp. 2701-04, 2013.

[6] Chen, K., Liu, L., “Geometric data perturbation for privacy-preserving outsourced data mining”. Knowledge Information Systems, pp.657-95, 2011.https://doi.org/10.1007/s10115-010-0362-4.

[7] Vaidya, J., Shafiq, B., Fan, W., Mehmood, D., Lorenzi, D., “A random decision tree framework for privacy-preserving data mining”, IEEE Transaction Dependable Secure Computing”, pp.399-11, 2014.

[8] G.Manikandan, N.Sairam, S.Jayashree, C.Saranya, “Achieving Data Privacy in a Distributed Environment Using Geometrical Transformation”, Middle East Journal of Scientific Research, pp.107-11, 2013.

[9] G.Manikandan, N.Sairam, R.Sudan and B.Vaishnavi "Shearing based Data Transformation Approach for Privacy Preserving Clustering " Third International Conference on Computing Communication and Networking Technologies (ICCCNT-2012), SNS College of Engineering, Coimbatore,2012.

[10] Kamakshi, P. and Dr. A.VinayaBabu, “ Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed Data”, Journal of Computing, pp.115-19, 2010.