A Data Perturbation Method by Field Rotation and Binning by Averages Strategy for Privacy Preservation

Mohammad Ali Kadampur and Somayajulu D.V.L.N.

Department of Computer Science and Engineering
National Institute of Technology Warangal-506004, A.P. India
ali.kadmpur@gmail, soma@nitw.ac.in
www.nitw.ac.in

Abstract. In this paper a novel technique useful to guarantee privacy of sensitive data with specific focus on numeric databases is presented. It is noticed that analysts and decision makers are interested in summary values of the data rather than the actual values. The proposed method considers that the maximum information lies in association of attributes rather than their actual proper values. Therefore it is aimed to perturb attribute associations in a controlled way, by shifting the data values of specific columns by rotating fields. The number of rotations is determined via using a support function for association rule handling and an algorithm that computes the best-choice rotation dynamically. Final summary statistics such as average, standard deviation of the numeric data are preserved by making bin average replacements for the actual values. The methods are tested on selected datasets and results are reported.

1 Introduction

Privacy is defined as "freedom from unauthorized intrusion" [15]. It is a deterrent against individually identifiable data in the process of knowledge extraction. Data mining technology is used for extracting knowledge from vast quantities of data. However the use of this technology has raised the concern that individual privacy is violated. Therefore the data mining technique must ensure that any information disclosed

1. cannot be traced to an individual; or
2. does not constitute an intrusion.

There are multiple approaches to achieve these goals[15]. Data perturbation is one of the methods for preserving privacy[2][12][15]. In perturbed data bases, if unauthorized data is accessed, the true value is not disclosed. Data perturbation techniques in effect distort the data in different ways before presenting it to the data mining algorithm, thus individually identifiable (private) values are not revealed. The privacy-preserving properties of such databases are a result of the perturbation. In this paper a composite novel method for data perturbation is proposed.

2 Related Work

In order to distort the data and preserve individual privacy, researchers have employed methods such as data encryption[11][13], Data randomization[12][15], Data swapping

C. Fyfe et al. (Eds.): IDEAL 2008, LNCS 5326, pp. 250 – 257, 2008.

[1],Data anonymization[4][6][7],Geometric transformation[2][11] and Nearest Neighbor Data Substitution (NeNDS)[11].The Fast Fourier Transform (FFT) and Wavelet transformation based data perturbation methods also have been reported [20] .

3 Motivation

Motivation for our approach is the observation that maximum information lies in the association of attributes (tuples) rather than the attribute value proper. Therefore it is proposed to break this association of attributes in a controlled and well recorded manner in order to allow only the legitimate users to access the original data. The integrity of the horizontals in the table is broken by shifting the data values of specific columns(fields) by rotating the fields. The number of rotations to be performed is evaluated by considering the internal associations of the data.

Illustration : Consider the following two matrix instances

$$\underset{\text{Matrix-a}}{\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{1n} \\ a_{21} & a_{22} & a_{23} & a_{2n} \\ a_{31} & a_{32} & a_{33} & a_{3n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & a_{mn} \end{bmatrix}} \qquad \underset{\text{Matrix-b}}{\begin{bmatrix} a_{11} & a_{22} & a_{13} & a_{1n} \\ a_{21} & a_{32} & a_{23} & a_{2n} \\ a_{31} & a_{42} & a_{33} & a_{3n} \\ \cdots & a_{m2} & \cdots & \cdots \\ a_{m1} & a_{12} & a_{m3} & a_{mn} \end{bmatrix}}$$

Fig. 1. matrix-a original table, matrix-b perturbed table

Let a_{ij} indicate the atomic value in the i^{th} row and j^{th} column. Let N be an integer indicating number of rotations and R be the number of tuples in the table(matrix). In our approach we try to perturb the i^{th} row by rotating j^{th} column by N times. The column "j" will be chosen depending upon the confidentiality associated with it. The Number of rotations N on j^{th} column is computed as a function of S, the support. N=f(support). The new value of data in the i^{th} row after perturbation in j^{th} column gets changed depending on N and R values in the table. The new perturbed value will be obtained by

$$a_{ij}= a_{i\ (j+N)\ \mathrm{mod}\ R} \tag{1}$$

Choice of number of rotations N on the field is critical to the method of field rotations. Proper value of N is computed by finding association rules and their support values.

3.1 Association Rule

If $I=\{i_1,i_2,i_3,....i_m\}$ is an item set then an association rule is an implication of the form $X\Rightarrow Y$,where $X\subset I$, $Y\subset I$ and $XnY=\phi$ is true[5][15]. The support S is a number