Matrix Decomposition Techniques for Data Privacy

Jun Zhang
University of Kentucky, USA

Jie Wang
University of Kentucky, USA

Shuting Xu
Virginia State University, USA

INTRODUCTION

Data mining technologies have now been used in commercial, industrial, and governmental businesses, for various purposes, ranging from increasing profitability to enhancing national security. The widespread applications of data mining technologies have raised concerns about trade secrecy of corporations and privacy of innocent people contained in the datasets collected and used for the data mining purpose. It is necessary that data mining technologies designed for knowledge discovery across corporations and for security purpose towards general population have sufficient privacy awareness to protect the corporate trade secrecy and individual private information. Unfortunately, most standard data mining algorithms are not very efficient in terms of privacy protection, as they were originally developed mainly for commercial applications, in which different organizations collect and own their private databases, and mine their private databases for specific commercial purposes.

In the cases of inter-corporation and security data mining applications, data mining algorithms may be applied to datasets containing sensitive or private information. Data warehouse owners and government agencies may potentially have access to many databases collected from different sources and may extract any information from these databases. This potentially unlimited access to data and information raises the fear of possible abuse and promotes the call for privacy protection and due process of law.

Privacy-preserving data mining techniques have been developed to address these concerns (Fung et al., 2007; Zhang, & Zhang, 2007). The general goal of the privacy-preserving data mining techniques is defined as to hide sensitive individual data values from the outside world or from unauthorized persons, and simultaneously preserve the underlying data patterns and semantics so that a valid and efficient decision model based on the distorted data can be constructed. In the best scenarios, this new decision model should be equivalent to or even better than the model using the original data from the viewpoint of decision accuracy. There are currently at least two broad classes of approaches to achieving this goal. The first class of approaches attempts to distort the original data values so that the data miners (analysts) have no means (or greatly reduced ability) to derive the original values of the data. The second is to modify the data mining algorithms so that they allow data mining operations on distributed datasets without knowing the exact values of the data or without direct accessing the original datasets. This article only discusses the first class of approaches. Interested readers may consult (Clifton et al., 2003) and the references therein for discussions on distributed data mining approaches.

BACKGROUND

The input to a data mining algorithm in many cases can be represented by a vector-space model, where a collection of records or objects is encoded as an $n \times m$ object-attribute matrix (Frankes, & Baeza-Yates, 1992). For example, the set of vocabulary (words or terms) in a dictionary can be the items forming the rows of the matrix, and the occurrence frequencies of all terms in a document are listed in a column of the matrix. A

collection of documents thus forms a term-document matrix commonly used in information retrieval. In the context of privacy-preserving data mining, each column of the data matrix can contain the attributes of a person, such as the person's name, income, social security number, address, telephone number, medical records, etc. Datasets of interest often lead to a very high dimensional matrix representation (Achlioptas, 2004). It is observable that many real-world datasets have nonnegative values for attributes. In fact, many of the existing data distortion methods inevitably fall into the context of matrix computation. For instance, having the longest history in privacy protection area and by adding random noise to the data, additive noise method can be viewed as a random matrix and therefore its properties can be understood by studying the properties of random matrices (Kargupta et al., 1991).

Matrix decomposition in numerical linear algebra typically serves the purpose of finding a computationally convenient means to obtain the solution to a linear system. In the context of data mining, the main purpose of matrix decomposition is to obtain some form of simplified low-rank approximation to the original dataset for understanding the structure of the data, particularly the relationship within the objects and within the attributes and how the objects relate to the attributes (Hubert, Meulman, & Heiser, 2000). The study of matrix decomposition techniques in data mining, particularly in text mining, is not new, but the application of these techniques as data distortion methods in privacy-preserving data mining is a recent interest (Xu et al., 2005). A unique characteristic of the matrix decomposition techniques, a compact representation with reduced-rank while preserving dominant data patterns, stimulates researchers' interest in utilizing them to achieve a win-win goal both on high degree privacy-preserving and high level data mining accuracy.

MAIN FOCUS

Data distortion is one of the most important parts in many privacy-preserving data mining tasks. The desired distortion methods must preserve data privacy, and at the same time, must keep the utility of the data after the distortion (Verykios et al., 2004). The classical data distortion methods are based on the random value perturbation (Agrawal, & Srikant, 2000). The more recent ones are based on the data matrix-decomposition strategies (Wang et al., 2006; Wang et al., 2007; Xu et al., 2006).

Uniformly Distributed Noise

The original data matrix A is added with a uniformly distributed noise matrix E_u. Here E_u is of the same dimension as that of A, and its elements are random numbers generated from a continuous uniform distribution on the interval from C_1 to C_2. The distorted data matrix A_u is denoted as: $A_u = A + E_u$.

Normally Distributed Noise

Similar to the previous method, here the original data matrix A is added with a normally distributed noise matrix E_n, which has the same dimension as that of A. The elements of E_n are random numbers generated from the normal distribution with a parameter mean μ and a standard deviation ρ. The distorted data matrix A_n is denoted as: $A_n = A + E_n$.

Singular Value Decomposition

Singular Value Decomposition (SVD) is a popular matrix factorization method in data mining and information retrieval. It has been used to reduce the dimensionality of (and remove the noise in the noisy) datasets in practice (Berry et al., 1999). The use of SVD technique in data distortion is proposed in (Xu et al., 2005). In (Wang et al., 2007), the SVD technique is used to distort portions of the datasets.

The SVD of the data matrix A is written as:

$$A = U\Sigma V^T$$

where U is an $n \times n$ orthonormal matrix, $\Sigma = \text{diag}[\sigma_1, \sigma_2, \cdots, \sigma_s]$ ($s = \min\{m,n\}$) is an $n \times m$ diagonal matrix whose nonnegative diagonal entries (the singular values) are in a descending order, and V^T is an $m \times m$ orthonormal matrix. The number of nonzero diagonal entries of Σ is equal to the rank of the matrix A.

Due to the arrangement of the singular values in the matrix Σ (in a descending order), the SVD transformation has the property that the maximum variation among the objects is captured in the first dimension, as $\sigma_1 \geq \sigma_i$ for $i \geq 2$. Similarly, much of the remaining variation is