Chapter 40
A Perturbation Method Based on Singular Value Decomposition and Feature Selection for Privacy Preserving Data Mining

Mohammad Reza Keyvanpour
Alzahra University, Iran

Somayyeh Seifi Moradi
Ports and Maritime Organization, Iran

ABSTRACT

In this study, a new model is provided for customized privacy in privacy preserving data mining in which the data owners define different levels for privacy for different features. Additionally, in order to improve perturbation methods, a method combined of singular value decomposition (SVD) and feature selection methods is defined so as to benefit from the advantages of both domains. Also, to assess the amount of distortion created by the proposed perturbation method, new distortion criteria are defined in which the amount of created distortion in the process of feature selection is considered based on the value of privacy in each feature. Different tests and results analysis show that offered method based on this model compared to previous approaches, caused the improved privacy, accuracy of mining results and efficiency of privacy preserving data mining systems.

INTRODUCTION

Data mining or knowledge discovery is a process that analyzes voluminous digital data in order to discover hidden but effective patterns from digital data (Ashrafi, Taniar, & Smith, 2005). In other words, this is a powerful tool for data analysis, with the goal of accurate and efficient identification of hidden and valuable patterns in the data, can facilitate the process of decision making, improve the allocation of resources, reduce costs and the exploitation of opportunities. Data mining is tip-top described as the union of historical and recent developments in statistics, artificial intelligence, and machine learning. These methods are then used together to study information and find previously

DOI: 10.4018/978-1-4666-8111-8.ch040

hidden trends or patterns within (Daly, & Taniar, 2004). Data mining applications have extremely altered the strategic decision-making procedures of organizations (Tjioe & Taniar, 2005). Hence, the various applications of this scope are used by various governmental, industrial, commercial, medical, financial, and scientific due to several advantages. In fact, wide range of data mining applications has made it an important field of research (Keyvanpour, Javadieh, & Ebrahimi, 2011).

As privacy is an issue of individual perception, an infallible and general solution to this dichotomy is infeasible. However, there are measures that can be undertaken to raise privacy protection (Wahlstrom, Roddick, Sarre, Estivill-Castro, & de Vries, 2009). Accordingly in recent years due to increasing concerns related to privacy, data mining methods are faced with a serious challenge which is to preserve the privacy of sensitive data. This method is under attack from privacy advocates because of a misunderstanding about what it really is and a credible concern about how it's generally done (Vaidya & Clifton, 2004). The organizations from one side should publish their customized information so as to access the benefits of data mining and on the other hand, are not unwilling to share their data due to preserving the privacy. The occurrence of such problems in data collection can be undesirable for data mining methods success as to achieve its goals (Seifi & Keyvanpour, 2012).

Hence, a new aspect of in the development of data mining is the approaches which are related to the concerns about privacy, in particular, in regard to this issue that data mining methods can produce accurate models without access to precise information of given records and to access valid results of the data mining (Clifton, Kantarcioglu, & Vaidya, 2002). In response to such anxieties, the data mining researches started to work on methods which preserved privacy along with data mining. As a result of this research, various approaches of privacy preserving data mining (PPDM) approaches are defined.

Data modification is one of the most popular approaches of privacy preserving data mining, especially for applications that require data owners to publish their personal and sensitive data. In this way, the data prior to publication are changed through certain methods so as to hide sensitive information (Keyvanpour & Seifi, 2010).

Approaches based on the data modification usually have good efficiency in terms of calculation but possess a few guarantees in preserving privacy and create balance with difficulty between ensuring privacy and data utility (important information and patterns existing in the data which should be preserved during data modification so that the accuracy of the data mining results in one level should be acceptable). As a result, the main challenge of the data modification based methods is to create a good and fair balance between privacy and data utility (Liu, Giannella, & Kargupta, 2006).

Recently, one of the most effective approaches to meet the challenges in privacy preserving data mining is the use of methods based on dimension reduction. The above methods operate based on this idea that they first identify worthless information in the dataset and then eliminate these worthless data so as to be perturbed. On the other side, since in the data mining applications, the eliminated parts are considered as noise, in many cases, the use of these methods can produce better results in terms of accuracy compared to mining on the original dataset (Xu, Zhang, Han, & Wang, 2006). One of the dimension reduction based methods which is used in PPDM is a Singular Value Decomposition (SVD) method (Keyvanpour & Seifi, 2010).

Generally, there are two general approaches regarding dimension reduction area: The feature extraction approach and feature selection approach. Strategies for feature extraction are the