Chapter 44
Association Rule Hiding in Privacy Preserving Data Mining

S. Vijayarani Mohan
Department of Computer Science, Bharathiar University, Coimbatore, India

Tamilarasi Angamuthu
Department of MCA, Kongu Engineering College, Erode, India

ABSTRACT

This article describes how privacy preserving data mining has become one of the most important and interesting research directions in data mining. With the help of data mining techniques, people can extract hidden information and discover patterns and relationships between the data items. In most of the situations, the extracted knowledge contains sensitive information about individuals and organizations. Moreover, this sensitive information can be misused for various purposes which violate the individual's privacy. Association rules frequently predetermine significant target marketing information about a business. Significant association rules provide knowledge to the data miner as they effectively summarize the data, while uncovering any hidden relations among items that hold in the data. Association rule hiding techniques are used for protecting the knowledge extracted by the sensitive association rules during the process of association rule mining. Association rule hiding refers to the process of modifying the original database in such a way that certain sensitive association rules disappear without seriously affecting the data and the non-sensitive rules. In this article, two new hiding techniques are proposed namely hiding technique based on genetic algorithm (HGA) and dummy items creation (DIC) technique. Hiding technique based on genetic algorithm is used for hiding sensitive association rules and the dummy items creation technique hides the sensitive rules as well as it creates dummy items for the modified sensitive items. Experimental results show the performance of the proposed techniques.

DOI: 10.4018/978-1-7998-8954-0.ch044

INTRODUCTION

Data mining has become an important technology in the past decade because of its ability to extract hidden knowledge and identifying patterns and trends from the large volume of data. There are many advantages of data mining and it is used for various applications such as businesses, marketing, medical, production, sales, science and technology. Even though, people can benefit through data mining techniques, there is a big disadvantage in data mining technology i.e. the risk to data privacy. For example, with the help of data mining techniques, we will be able to infer sensitive information which includes personal information or even patterns from non-sensitive information or unclassified data which violates the privacy of an individual. This should be protected, and all the data mining tasks should be performed in a secured way. This situation has created and raised the necessity of development of new privacy preserving data mining techniques (Aris et al., 2010).

Privacy preserving data mining is relatively a new research area in the data mining community, counting approximately a decade of existence. It investigates the side effects of data mining methods that originate from the penetration into the privacy of individuals and organizations. Since the pioneering work of Agrawal et al. (2000) and Lindell et al. (2000), several approaches have been proposed in the research literature for the offering of privacy in data mining.

Privacy preserving data mining discovers several applications in surveillance which is obviously supposed to be "privacy-breaching" applications. The solution is to propose techniques (Sweeney, 2005) which continue to be efficient, without negotiating security. In (Sweeney, 2005), a number of methods have been conversed for bio-surveillance, identity theft and facial de-identification. Most techniques for privacy computations apply some type of alteration on the data in order to execute the privacy preservation. Naturally, such techniques decrease the granularity of demonstration in order to decrease the privacy. This diminution in granularity results in some failure of efficiency of data management or mining algorithms.

The majority of the proposed approaches can be classified along two principal research directions: (i) data hiding approaches and (ii) knowledge hiding approaches. The first direction collects methodologies that investigate how the privacy of raw data, or information, can be maintained before the course of mining the data. Many approaches of this category aim at the removal of confidential or private information from the original data prior to its disclosure and operate by applying techniques such as perturbation, sampling, generalization or suppression, transformation, etc. to generate a sanitized counterpart of the original dataset. The ultimate goal is to enable the data holder to receive accurate data mining results when it is not provided with the real data or adhere to specific regulations pertaining to micro data publication (e.g., as is the case of publishing patient-specific data).

The second direction of approaches involves methodologies that aim to protect the sensitive data mining results (i.e., the extracted knowledge patterns) rather than the raw data itself, which were produced by the application of data mining tools on the original database. This direction of approaches mainly deals with distortion and blocking techniques that prohibit the leakage of sensitive knowledge patterns in the disclosed data, as well as with techniques for downgrading the effectiveness of classifiers in classification tasks, so that the produced classifiers do not reveal any sensitive knowledge.

There are two different types of privacy concerns such as input privacy and output privacy in data mining. In input privacy, the data is manipulated in which the data mining results are not affected or affected minimally. In the literature, the works related to input privacy are in Evfimievski et al. (2002, 2003), Kantarcioglu et al. (2002), Rizvi et al. (2002), Vaidya et al. (2002). Another method for input