Privacy-Preserving Data Mining

Stanley R. M. Oliveira
Embrapa Informática Agropecuária, Brazil

INTRODUCTION

Despite its benefits in various areas (e.g., business, medical analysis, scientific data analysis, etc), the use of data mining techniques can also result in new threats to privacy and information security. The problem is not data mining itself, but the way data mining is done. "Data mining results rarely violate privacy, as they generally reveal high-level knowledge rather than disclosing instances of data" (Vaidya & Clifton, 2003). However, the concern among privacy advocates is well founded, as bringing data together to support data mining projects makes misuse easier. Thus, in the absence of adequate safeguards, the use of data mining can jeopardize the privacy and autonomy of individuals.

Privacy-preserving data mining (PPDM) cannot simply be addressed by restricting data collection or even by restricting the secondary use of information technology (Brankovic & V. Estivill-Castro, 1999). Moreover, there is no exact solution that resolves privacy preservation in data mining. In some applications, solutions for PPDM problems might meet privacy requirements and provide valid data mining results (Oliveira & Zaïane, 2004b).

We have witnessed three major landmarks that characterize the progress and success of this new research area: *the conceptive landmark*, *the deployment landmark*, and *the prospective landmark*. *The Conceptive landmark* characterizes the period in which central figures in the community, such as O'Leary (1995), Piatetsky-Shapiro (1995), and others (Klösgen, 1995; Clifton & Marks, 1996), investigated the success of knowledge discovery and some of the important areas where it can conflict with privacy concerns. The key finding was that knowledge discovery can open new threats to informational privacy and information security if not done or used properly.

The Deployment landmark is the current period in which an increasing number of PPDM techniques have been developed and have been published in refereed conferences. The information available today is spread over countless papers and conference proceedings. The results achieved in the last years are promising and suggest that PPDM will achieve the goals that have been set for it.

The Prospective landmark is a new period in which directed efforts toward standardization occur. At this stage, there is no consensus on privacy principles, policies, and requirements as a foundation for the development and deployment of new PPDM techniques. The excessive number of techniques is leading to confusion among developers, practitioners, and others interested in this technology. One of the most important challenges in PPDM now is to establish the groundwork for further research and development in this area.

BACKGROUND

Understanding privacy in data mining requires understanding how privacy can be violated and the possible means for preventing privacy violation. In general, one major factor contributes to privacy violation in data mining: the misuse of data.

Users' privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many applications (e.g., business, medical analysis, etc), it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected (Culnan, 1993).

Defining Privacy for Data Mining

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. We refer to the former as individual privacy preservation and the latter

as collective privacy preservation, which is related to corporate privacy in (Clifton et al., 2002).

- **Individual privacy preservation:** The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual.
- **Collective privacy preservation:** Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to protect sensitive knowledge that can provide competitive advantage in the business world.

MAIN FOCUS

A Taxonomy of existing PPDM Techniques

The existing PPDM techniques in the literature can be classified into four major categories: data partitioning, data modification, data restriction, and data ownership as can be seen in Figure 1.

Data Partitioning Techniques

Data partitioning techniques have been applied to some scenarios in which the databases available for mining are distributed across a number of sites, with each site only willing to share data mining results, not the source data. In these cases, the data are distributed either horizontally or vertically. In a horizontal partition, different entities are described with the same schema in all partitions, while in a vertical partition the attributes of the same entities are split across the partitions. The existing solutions can be classified into Cryptography-Based Techniques and Generative-Based Techniques.

- **Cryptography-based techniques:** In the context of PPDM over distributed data, cryptography-based techniques have been developed to solve problem of the following nature: two or more

Figure 1. A taxonomy of PPDM techniques.