Privacy and Confidentiality Issues in Data Mining

Yücel Saygin
Sabanci University, Turkey

INTRODUCTION

Data regarding people and their activities have been collected over the years, which has become more pervasive with widespread usage of the Internet. Collected data usually are stored in data warehouses, and powerful data mining tools are used to turn it into competitive advantage. Besides businesses, government agencies are among the most ambitious data collectors, especially in regard to the increase of safety threats coming from global terrorist organizations. For example, CAPPS (Computer Assisted Passenger Prescreening System) collects flight reservation information as well as commercial information about passengers. This data, in turn, can be utilized by government security agencies. Although CAPPS represents US national data collection efforts, it also has an effect on other countries. The following sign at the KLM ticket desk in Amsterdam International Airport illustrates the international level of data collection efforts: "Please note that KLM Royal Dutch Airlines and other airlines are required by new security laws in the US and several other countries to give security customs and immigration authorities access to passenger data. Accordingly, any information we hold about you and your travel arrangements may be disclosed to the concerning authorities of these countries in your itinerary." This is a very striking example of how the confidential data belonging to citizens of one country could be handed over to authorities of some other country via newly enforced security laws. In fact, some of the largest airline companies in the US, including American, United, and Northwest, turned over millions of passenger records to the FBI, according to the *New York Times* (Schwartz & Maynard, 2004).

Aggressive data collection efforts by government agencies and enterprises have raised a lot of concerns among people about their privacy. In fact, the Total Information Awareness (TIA) project, which aims to build a centralized database that will store the credit card transactions, e-mails, Web site visits, and flight details of Americans was not funded by Congress due to privacy concerns. Although the privacy of individuals is protected by regulations, such regulations may not be enough to ensure privacy against new technologies such as data mining. Therefore, important issues need to be considered by the data collectors and the data owners. First of all, the data itself may contain confidential information that needs to be secured. More importantly, the privacy of the data owners needs to be taken into consideration, depending on how the data will be used. Privacy risks increase even more when we consider powerful data-mining tools that can be used to extract confidential information. We will try to address the privacy and confidentiality issues that relate to data-mining techniques throughout this article.

BACKGROUND

Historically, researchers have been more interested in the security aspect of data storage and transfer. Data security is still an active area of research in relation to new data storage and transfer models, such as XML. Also, new media of data transfer, such as wireless, introduce new challenges.

The most important work on data security that relates to the privacy and confidentiality issues in data analysis is statistical disclosure control and data security in statistical databases (Adam & Wortmann, 1989). The basic idea of a statistical database is to let users obtain statistical or aggregate information from a database, such as the average income and maximum income in a given department. This is done while limiting the disclosure of confidential information, such as the salary of a particular person. The concept of k-anonymity, which ensures that the obtained values refer to at least k individuals (Sweeney, 2002), was proposed as a measure for the degree of confidentiality in aggregate data. In this way, identification of individuals can be blocked up to a certain degree. However, successive query results over the database may pose a risk to security, since a combination of intelligently issued queries may infer confidential data. The general problem of adversaries being able to obtain sensitive data from non-sensitive data is called the inference problem (Farkas & Jajodia, 2003).

MAIN THRUST

Data mining is considered to be a tool for analyzing large collections of data. As a result, government and industry seek to exploit its potential by collecting extensive information about people and their activities. Coupled with its power to extract previously unknown knowledge, data mining also has become one of the main targets of privacy advocates. In reference to data mining, there are two main concerns regarding privacy and confidentiality:

(1) Protecting the confidentiality of data and privacy of individuals against data mining methods.
(2) Enabling the data mining algorithms to be used over a database without revealing private and confidential information.

We should note that, although both of the tracks seem to be similar to each other, their purposes are different. In the first case, the data may not be confidential, but data-mining tools could be used to infer confidential information. Therefore, some sanitization techniques are needed to modify the database, so that privacy concerns are alleviated. In the second track, the data are considered confidential and are perturbed before they are given to a third party. Both of the approaches are necessary in order to achieve full privacy of individuals and will be discussed in detail in the following subsections.

Privacy Preserving Data Mining

In the year 2000, Agrawal and Srikant, in their seminal paper published in ACM SIGMOD proceedings, coined the term *privacy preserving data mining*. Based on this work, privacy-preserving data mining can be defined as a technology that enables data-mining algorithms to work on encrypted or perturbed data (Agrawal & Srikant, 2000). We think of it as mining without seeing the actual data. The authors have pointed out that the collected data may be outsourced to third parties to perform the actual data mining. However, before the data could be handed over to third parties, the confidential values in the database, such as the salary of employees, needs to be perturbed. The research results are for specific data-mining techniques; namely, decision tree construction for classification. The main idea is to perturb the confidential data values in a way that the original data distribution could be reconstructed but not the original data values. In this way, the classification techniques that rely on the distribution of the data to construct a model can still work within a certain error margin that the authors approve.

Another interesting work on privacy-preserving data mining was conducted in the US at Purdue University (Kantarcioglu & Clifton, 2002). This work was in the context of association rules in a distributed environment. The authors considered a case in which there were multiple companies that had their own confidential local databases that they did not want to share with others. When the datasets of the individual companies have the same schema, then this means that the data are distributed horizontally. If the schema of the local databases are complementary to each other, then this means that the data are vertically distributed. The Purdue research group considered both horizontal and vertical data distribution for privacy-preserving association rule mining. In both cases, the individual association rules together with their statistical properties were assumed to be confidential. Also, secure multi-part computation techniques were employed, based on specialized encryption practices, to make sure that the confidential association rules were circulated among the participating companies in encrypted form. The resulting global association rules can be obtained in a private manner without each company knowing which rule belongs to which local database.

Both of the seminal works on privacy-preserving data mining (Agrawal & Srikant, 2000; Kantarcioglu & Clifton, 2002) have shaped the research in this area. They show how data mining can be done on private data in a centralized and distributed environment. Although a specific data-mining model was the target in both of these papers, these authors initiated ideas that could be applied to other data-mining models.

Privacy and Confidentiality Against Data Mining

Data mining and data warehousing technology have a great impact on data analysis and business decision making, which motivated the collection of data on practically everything, ranging from customer purchases to navigation patterns of Web site visitors. As a result of data collection efforts, privacy advocates in the US are now raising their voices against the misuse of data, even if it was collected for security purposes. The following quote from the *New York Times* article demonstrates the point of privacy against data mining (Markoff, 2002):

> *Pentagon has released a study that recommends the government to pursue specific technologies as potential safeguards against the misuse of data-mining systems similar to those now being considered by the government to track civilian activities electronically in the United States and abroad.*

This shows us that even the most aggressive data collectors in the US are aware of the fact that the data-mining tools could be misused, and we need a mechanism to protect the confidentiality and privacy of people.

The initial work by Chang and Moskovitz (2000) points out the possible threats imposed by data mining tools. In this work, the authors argue that even if some of the data values are deleted (or hidden) in a database, there is still a threat of the hidden data being recovered. They show that a classification model could be constructed using

Related Content