Chapter 1.10
Data Mining and Privacy

Esma Aïmeur
Université de Montréal, Canada

Sébastien Gambs
Université de Montréal, Canada

INTRODUCTION

With the emergence of Internet, it is now possible to connect and access sources of information and databases throughout the world. At the same time, this raises many questions regarding the privacy and the security of the data, in particular how to mine useful information while preserving the privacy of sensible and confidential data. *Privacy-preserving data mining* is a relatively new but rapidly growing field that studies how data mining algorithms affect the privacy of data and tries to find and analyze new algorithms that preserve this privacy.

At first glance, it may seem that data mining and privacy have orthogonal goals, the first one being concerned with the discovery of useful knowledge from data whereas the second is concerned with the protection of data's privacy. Historically, the interactions between privacy and data mining have been questioned and studied since more than a decade ago, but the name of the domain itself was coined more recently by two seminal papers attacking the subject from two very different perspectives (Agrawal & Srikant, 2000; Lindell & Pinkas, 2000). The first paper (Agrawal & Srikant, 2000) takes the approach of randomizing the data through the injection of noise, and then recovers from it by applying a reconstruction algorithm before a learning task (the induction of a decision tree) is carried out on the reconstructed dataset. The second paper (Lindell & Pinkas, 2000) adopts a cryptographic view of

DOI: 10.4018/978-1-61350-323-2.ch1.10

the problem and rephrases it within the general framework of secure multiparty computation.

The outline of this chapter is the following. First, the area of privacy-preserving data mining is illustrated through three scenarios, before a classification of privacy-preserving algorithms is described and the three main approaches currently used are detailed. Finally, the future trends and challenges that await the domain are discussed before concluding.

BACKGROUND

The area of privacy-preserving data mining can still be considered in its infancy but there are already several workshops (usually held in collaboration with different data mining and machine learning conferences), two different surveys (Verykios *et al.*, 2004; Výborný, 2006) and a short book (Vaidya, Clifton & Zhu, 2006) on the subject. The notion of privacy itself is difficult to formalize and quantify, and it can take different flavours depending on the context. The three following scenarios illustrate how privacy issues can appear in different data mining contexts.

- **Scenario 1:** A famous Internet-access provider wants to release the log data of some of its customers (which include their personal queries over the last few months) to provide a public benchmark available to the web mining community. How can the company anonymize the database in such a way that it can guarantee to its clients that no important and sensible information can be mined about them?
- **Scenario 2:** Different governmental agencies (for instance the Revenue Agency, the Immigration Office and the Ministry of Justice) want to compute and release some joint statistics on the entire population but they are constrained by the law not to communicate any individual information on citizens, even to other governmental agencies. How can the agencies compute statistics that are sufficiently accurate while at the same time, safeguarding the privacy of individual citizens?
- **Scenario 3:** Consider two bioinformatics companies: Alice Corporation and Bob Trust. Each company possesses a huge database of bioinformatics data gathered from experiments performed in their respective labs. Both companies are willing to cooperate in order to achieve a learning task of mutual interest such as a clustering algorithm or the derivation of association rules, nonetheless they do not wish to exchange their whole databases because of obvious privacy concerns. How can they achieve this goal without disclosing any unnecessary information?

When evaluating the potential privacy leak caused by a data mining process, it is important to keep in mind that the adversary may have some side information that could be used to infringe this privacy. Indeed, while the data mining process by itself may not be directly harmful, it is conceivable that associated with the help of linking attacks (derived from some *a priori* knowledge), it may lead to a total breakdown of the privacy.

MAIN FOCUS

The privacy-preserving techniques can generally be classified according to the following dimensions:

- *The distribution of the data*. During the data mining process, the data can be either in the hands of a single entity or distributed among several participants. In the case of distributed scenarios, a further distinction can be made between the situation where the attributes of a single record are split