IRM PRESS
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com

ITB15304

This chapter appears in the book, ***Computer Security, Privacy, and Politics: Current Issues, Challenges, and Solutions*** by **R. Subramanian** © 2008, IGI Global

Chapter V

Privacy Preserving Data Mining:
Taxonomy of Existing Techniques

Madhu V. Ahluwalia,
University of Maryland, Baltimore County (UMBC), USA

Aryya Gangopadhyay,
University of Maryland, Baltimore County (UMBC), USA

Abstract

This chapter gives a synopsis of the techniques that exist in the area of privacy preserving data mining. Privacy preserving data mining is important because there is a need to develop accurate data mining models without using confidential data items in individual records. In providing a neat categorization of the current algorithms that preserve privacy for major data mining tasks, the authors hope that students, teachers and researchers can gain an understanding of this vast area and apply the knowledge gained to find new ways of simultaneously preserving privacy and conducting mining.

Introduction

Data mining, also referred to as knowledge discovery in databases (KDD), has been embraced with much enthusiasm by researchers and market analysts due to its promise to reveal information useful for scientific and technical research, business intelligence, and decision-support. A multitude of tools and techniques to facilitate knowledge discovery have therefore been developed and used increasingly. In the post 9/11 era, interest in data mining techniques has escalated due to their usefulness in counter-terrorism activities. However, the revelation of private information that may occur with data mining is unconstitutional. Laws have been enforced to forbid the U.S. Department of Defense to conduct data mining unless deemed necessary for security purposes. It also is mandatory that all operations of U.S. government agencies involving data mining ensure individual privacy. This conflict has given birth to a novel research direction known as the privacy preserving data mining.

Privacy preserving data mining entails two notions: 1) extracting or mining knowledge from large amounts of data and 2) performing data mining in such a way that data privacy is not compromised. This is a daunting task in an information age where we generate data with every move that we make. One challenge is the ease with which unauthorized parties can deduce confidential information from released data sources. Also known as the inference problem, this difficulty is discussed at length by Samarati (2001) and Sweeney (2002). Another major challenge is nailing down the concept of privacy. Whose privacy ought to be protected, an organization's or an individual's (Clifton, Kantarcioglu, & Vaidya, 2002)? How do we measure privacy (Pinkas, 2002)? What kind of adversarial models do we deal with (Gangopadhyay & Ahluwalia, 2006; Pinkas, 2002)? In addition to these issues that complicate the development of models and algorithms to preserve privacy, there are other legal, commercial, governmental, philosophical, ethical, and personal perspectives that need to be incorporated into the definition of privacy. However, this makes it even more difficult to address privacy concerns and provide a universally satisfactory resolution to the problem. Finally, all privacy-enhancing technologies influence the outcome of data mining to some extent. Depending on the modifications made to the data or the accuracy of information obtained from subjects who are unwilling to divulge their personal data due to privacy concerns, knowledge discovery tools might taint the results so that they exhibit lower accuracy or, sometimes even worse, false knowledge. Therefore, it is necessary to strike a balance between the need to privatize data and to retain the meaningfulness of the data mining results.

As mentioned earlier, there is no dearth of tools and techniques to achieve the twin goals of sufficient privacy of input data and sufficient utility of mining results in the data mining community today. There is, however, a lack of literature that pro-