IGI PUBLISHING

ITB15225

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

This paper appears in the publication, **Data Mining and Knowledge Discovery Technologies** edited by D. Taniar © 2008, IGI Global

Chapter VIII

Using Cryptography for Privacy-Preserving Data Mining

Justin Zhan, Carnegie Mellon University, USA

Abstract

To conduct data mining, we often need to collect data from various parties. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. How multiple parties collaboratively conduct data mining without breaching data privacy presents a challenge. The goal of this chapter is to provide solutions for privacy-preserving k-nearest neighbor classification, which is one of the data mining tasks. Our goal is to obtain accurate data mining results without disclosing private data. We propose a formal definition of privacy and show that our solutions preserve data privacy.

Introduction

Recent advances in data collection, data dissemination, and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from the point of view of privacy preservation. The term privacy is used frequently in ordinary language, yet there is no single definition of this term. The concept of privacy has broad historical roots in sociological and anthropological discussions about how extensively it is valued and preserved in various cultures (Schoeman, 1984). Historical use of the term is not uniform and there remains confusion over the meaning, value, and scope of the concept of privacy. Privacy refers to the right of users to conceal their personal information and have some degree of control over the use of any personal information disclosed to others (Ackerman, Cranor, & Reagle, 1999). Particularly, in this chapter, the privacy preservation means that multiple parties collaboratively get valid data mining results while disclosing no private data to each other or any party who is not involved in the collaborative computations.

The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. However, there are situations where the sharing of data can lead to mutual benefit. Despite the potential gain, this is often not possible due to the confidentiality issues which arise. It is well documented (Epic, 2003) that the unlimited explosion of new information through the Internet and other media has reached a point where threats against the privacy are very common and they deserve serious thinking.

Let us consider an example. There are several hospitals involved into a multi-site medical study. Each hospital has its own data set containing patient records. These hospitals would like to conduct data mining over the data sets from all of hospitals with the goal of more valuable information would be obtained via mining the joint data set. Due to privacy laws, one hospital cannot disclose their patient records to other hospitals.

How can these hospitals achieve their objective? Can privacy and collaborative data mining coexist? In other words, can the collaborative parties somehow conduct data mining computations and obtain the desired results without compromising their data privacy?

We show that privacy and collaborative data mining can be achieved at the same time. The goal of this chapter is to present technologies to solve privacy-preserving collaborative data mining problems over large data sets with reasonable efficiency.

The contributions of this chapter contain the following: (1) a proposed formal definition of privacy for privacy-preserving collaborative data mining, (2) a solution for k-nearest neighbor classification with vertical collaboration, and (3) the efficiency analysis to show the performance scaling up with various factors such