Data Mining for Security and Privacy

Johannes Gehrke
Cornell University
Department of Computer Sciences
Ithaca, NY 14853, USA
johannes@cs.cornell.edu

Keywords

Privacy, Security

1. INTRODUCTION

There has been recently a heightened awareness of the importance of data mining to security applications, and data mining has been successfully applied in many scenarios ranging from the study of terrorist networks for homeland defense to the analysis of network audit data for intrusion detection. At the same time, there is a growing concern about the collection and possible misuse of private data about individuals. For the special topic of this issue, I invited experts to write about emerging techniques that build bridges between privacy and security, covering topics from database privacy, privacy-preserving data mining, to intrusion detection, focusing both on recent results and the tools that are necessary to perform research in these areas. I hope that this special issue illustrates the breadt, depth, and excitement of ongoing work in this special application of data mining!

2. CONTENTS OF THIS ISSUE

2.1 Invited Articles

Bhavani Thuraisingham draws the connections between privacy privacy-preserving data mining and the classical inference problem in statistical database systems. She raises the issue that privacy concerns have broad social, political, and legal aspects, and that there is a tension between the necessity of information for national security and the need of personal privacy.

Csilla Farkas and Sushil Jajodia survey the inference problem in statistical databases, where only statistics about *groups* of records are permitted to be revealed and the privacy of individual records needs to be preserved.

Benny Pinkas surveys cryptographic techniques for privacy-preserving data mining. He introduces us to secure multiparty computation techniques that allow servers that do not trust each other to compute functions over the union of local data while ensuring that no server learns anything about the data of the other servers, except the output of the function.

Martin Olivier discusses database privacy from a non-technical angle, and he describes the challenges of balancing data privacy with society's need of access to information.

Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Zhu survey tools for privacy-preserving data mining. They lead the reader on a tour of several basic privacy-preserving operators, and then they show how these operators can be composed to arrive at higher-level data mining algorithms.

Wenke Lee describes data mining techniques for intrusion detection including a framework that transforms audit data into data mining models. As network data arrives in high-speed data streams, model application to new records has to be very efficient, and Wenke describes several approaches on how to speed up model execution.

Alexandre Evfimievski describes techniques for privacy-preserving data mining based on randomization. If designed carefully, randomization can provide a high level of privacy while allowing reconstruction of aggregate properties of the data. The article also discusses issues in quantifying privacy.

2.2 Contributed Articles

Tao Li, Qi Li, Shenghuo Zhu, and Mitsunori Ogihara survey the applications of wavelets in data mining. They divide KDD into data management, data preprocessing, data mining algorithms, and post-processing, and survey recent results from the literature for each of these components.

Luc De Raedt argues for inductive databases as a framework to tightly integrate database systems and data mining. He outlines the components of such an approach including an inductive query language, query processing operators, reasoning about patterns, and efficient implementations.

Victor Lo describes for a new objective function called the true lift for selecting customers as targets in database marketing. The paper proposes both a measure for the true lift, and a procedure that allows estimating the true lift.

2.3 Reports from SIGKDD 2002

The remaining papers in the issue are reports from the SIGKDD 2002 Conference. The first set of articles covers descriptions of the two tasks from the KDD Cup 2002, following by reports from the winning teams and the teams that were awarded honorable mentions. The issue then contains an article describing the KDD 2002 Conference Panel "The Perfect Data Mining Tool". It concludes with reports from five of the workshops associated with SIDKDD 2002.

About the Guest Editor. Johannes Gehrke is an Assistant Professor in the Department of Computer Science at Cornell University. Johannes' research interests are in the areas of data mining, data privacy and security, data stream processing, and novel distributed database technology such as data management for sensor networks and peer-to-peer networks.