Chapter 2.8
A Partial Optimization Approach for Privacy Preserving Frequent Itemset Mining

Shibnath Mukherjee
Yahoo! Research and Development, India

Aryya Gangopadhyay
University of Maryland Baltimore County, USA

Zhiyuan Chen
University of Maryland Baltimore County, USA

ABSTRACT

While data mining has been widely acclaimed as a technology that can bring potential benefits to organizations, such efforts may be negatively impacted by the possibility of discovering sensitive patterns, particularly in patient data. In this article the authors present an approach to identify the optimal set of transactions that, if sanitized, would result in hiding sensitive patterns while reducing the accidental hiding of legitimate patterns and the damage done to the database as much as possible. Their methodology allows the user to adjust their preference on the weights assigned to benefits in terms of the number of restrictive patterns hidden, cost in terms of the number of legitimate patterns hidden, and damage to the database in terms of the difference between marginal frequencies of items for the original and sanitized databases. Most approaches in solving the given problem found in literature are all-heuristic based without formal treatment for optimality. While in a few work, ILP has been used previously as a formal optimization approach, the novelty of this method is the extremely low cost-complexity model in contrast to the others. They implement our methodology in C and C++ and ran several experiments with synthetic data generated with the IBM synthetic data generator. The experiments show excellent results when compared to those in the literature.

DOI: 10.4018/978-1-61350-323-2.ch2.8

INTRODUCTION

Knowledge discovery from databases (KDD) and knowledge hiding in databases (KHD) are perhaps oxymoron, yet this is indeed the problem faced by the data mining research community these days. Over the past decade, research in the line was focused primarily in discovery of memory efficient and fast algorithms that could discover patterns in large databases in forms of association rules, classification models and clusters of data values. Today with rigorous improvements in the field of these algorithms (Han & Kamber 2006), the primary area has taken quite a leap forward, but has posed some grave problems as well in terms of security and privacy preservation in the knowledge discovery tasks (Dasseni et al., 2001; Evfimienski et al., 2002; Oliviera et al., 2003a, 2003b; Han et al., 2006). A number of cases have been reported in literature where data mining actually has posed threats to discovery of sensitive knowledge and violating privacy. One typical problem is that of inferencing, which means inferring sensitive information from non-sensitive or unclassified data (Oliviera et al., 2002; Clifton, 2001).

Data mining is part of the larger business intelligence initiatives that are taking place in organizations across government and industry sectors, many of which include medical applications. It is being used for prediction as well knowledge discovery that can lead to cost reduction, business expansion, and detection of fraud or wastage of resources, among other things. With its many benefits, data mining has given rise to increasingly complex and controversial privacy issues. For example, the privacy implications of data mining have lead to high profile controversies involving the use of data mining tools and techniques on data related to drug prescriptions. Two major health care data publishers filed a petition to the Supreme Court on whether commercial use of data mining is protected by the First Amendment[1], an appeal to a controversial ruling by the 1st U.S. Circuit Court of Appeals that upheld a 2006 New Hampshire law that banned the usage of doctor's prescription history to increase drug sales.

Privacy implications are a major roadblock to information sharing across organizations. For example, sharing inventory data might reveal information that can be used to gain strategic advantages by competitors. Unless the actual or perceived implications of data mining methods on privacy issues are properly dealt with, it can lead to sub-optimal decision making in organizations, and reluctance to accept such tools by the public in general. For example there could be benefits in sharing prescription data from different pharmacy stores to mine for information such as the use of generic drugs, socio-demographic and geographic analysis of prescription drugs, which will require moving the data from each store or site to a central location, which increases the risks of litigation. In general several potential problems that have been identified for privacy protection make the case for privacy reserving data mining. These include: legal requirements for protecting data (e.g. HIPAA healthcare regulations in the US) Federal register (2002), liability from inadvertent disclosure of data, risk of misuse of proprietary information (Atallah et al., 2003), and antitrust concerns (Vaidya et al., 2006).

Thus it is of growing importance to devise efficient tradeoffs between knowledge discovery and knowledge hiding from databases so that cost to the involved, in general, gets minimized in the process yet the benefit is maximized. The work that will be presented in this article will focus on formulating a model for sanitization of databases against discovery of restrictive associative patterns, while distorting the databases and legitimate pattern discovery as little as possible. To illustrate the problem, consider a classic example given in (Evfimienski et al., 2002; Oliviera et al., 2002). There is a server and several clients, each having its own set of items. The clients want the server to provide them with recommendations based on statistical information about association among items. However the clients do not want the server