Chapter 18
Disclosure Control of Confidential Data by Applying Pac Learning Theory

Ling He
Saginaw Valley State University, USA

Haldun Aytug
University of Florida, USA

Gary J. Koehler
University of Florida, USA

ABSTRACT

This paper examines privacy protection in a statistical database from the perspective of an intruder using learning theory to discover private information. With the rapid development of information technology, massive data collection is relatively easier and cheaper than ever before. The challenge is how to provide database users with reliable and useful data while protecting the privacy of the confidential information. This paper discusses how to prevent disclosing the identity of unique records in a statistical database. The authors' research extends previous work and shows how much protection is necessary to prevent an adversary from discovering confidential data with high probability at small error.

INTRODUCTION

Statistical organizations, such as the U.S. Census Bureau, collect large amounts of data every year and make it available to the public as statistical databases (SDBs). These organizations have the legal and ethical obligations to maintain the accuracy, integrity and privacy of the information contained in their databases. In order to protect the identity of unique records in SDBs, only limited aggregate queries, such as Sum, Count and Mean, are allowed.

Statistical Disclosure Control (SDC) methods are designed to protect confidential information in a database (minimizing the disclosure risk) while providing the SDB users with reliable and useful data (minimizing the information loss). The goal of disclosure control is to prevent users

DOI: 10.4018/978-1-61350-471-0.ch018

from inferring confidential data on the basis of those successive statistical queries. It is also our research focus.

In contrast to the traditional SDC methods we approach the database security problem from a different perspective: we assume that an adversary regards the true confidential data in the database as an unknown target concept and tries to discover it within a limited number of queries using learning methods.

Probably Approximately Correct (PAC) learning theory is a framework for analyzing machine learning (ML) algorithms (Valiant, 1984). This paradigm focuses on learning algorithms that discover a target concept from examples which are randomly drawn from an unknown but fixed distribution. Given accuracy and confidence parameters, the PAC model bounds the error that the discovered concept may make.

We take as the database administrator's security problem that of determining how to make the adversarial learning task difficult while still providing useful information to legitimate users. That is, we look at the trade-offs between the confidence an adversary can achieve in discovering confidential data, the number of queries he or she must run and the resulting accuracy. We provide a bound that describes the trade-off between the number of queries, accuracy and confidence using PAC learning theory.

Research on privacy of statistical databases has emphasized developing methods to protect privacy without worrying about how much protection is enough. Our results quantify "how much protection one buys when using additive data/query perturbation."

TRADITIONAL APPROACHES FOR DISCLOSURE CONTROL METHODS

A *compromise* of a database occurs when confidential information is disclosed exactly, partially or inferentially in such a way that the user can link the data to an entity. *Inferential disclosure* or *statistical inference* (Más, 2000) refer to the situation that an unauthorized user can infer the confidential data with a high probability by running sequential queries and the probability exceeds a predetermined threshold of disclosure. For example, assume a hospital database has a binary filed called HIV-Status. A user can issue several SUM (HIV-Status) queries against this database. Individually, these queries may not pose a threat, however, when combined the adversary might infer the HIV-Status of a patient (for a full example see Garfinkel, Gopal, & Goes, 2002). This is known as an inference problem, which falls within our research focus.

Adam and Wortmann (1989) classify SDC methods for SDBs into four categories: Conceptual, Query Restriction, Data Perturbation, and Output Perturbation. Perturbations are achieved by applying either an additive or multiplicative technique. An additive technique (Muralidhar, Parsa, & Sarathy, 1999) adds noise to the confidential data. Multiplicative data perturbation (Muralidhar, Batra, & Kirs, 1995) protects the sensitive information by multiplying the original data with a random variable, with mean 1 and a pre-specified variance.

Data shuffling, a perturbation technique, proposed and further studied by Muralidhar and Sarathy (2006) and Muralidhar, Sarathy, and Dandekar (2006) offers a high level of data utility while reducing the disclosure risk by shuffling data among observations. Data shuffling maintains all advantages of perturbation methods and provides a better performance than other data protection methods.

Muralidhar and Sarathy (2008) recently proposed a methodology for generating sufficiency-based non-synthetic perturbed data, which provides the masked data with the same mean vector and covariance matrix as those of the original data, and further prevents the information loss.

Nunez, Garfinkel, and Gopal (2007) developed a hybrid method that combines both data pertur-