Privacy-Preserving Estimation

Mohammad Saad Al-Ahmadi
King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia

Rathindra Sarathy
Oklahoma State University, USA

INTRODUCTION

Data mining has evolved from a need to make sense of the enormous amounts of data generated by organizations. But data mining comes with its own cost, including possible threats to the confidentiality and privacy of individuals. This chapter presents a background on privacy-preserving data mining (PPDM) and the related field of statistical disclosure limitation (SDL). We then focus on privacy-preserving estimation (PPE) and the need for a data-centric approach (DCA) to PPDM. The chapter concludes by presenting some possible future trends.

BACKGROUND

The maturity of information, telecommunications, storage and database technologies, have facilitated the collection, transmission and storage of huge amounts of raw data, unimagined until a few years ago. For raw data to be utilized, they must be processed and transformed into information and knowledge that have added value, such as helping to accomplish tasks more effectively and efficiently. Data mining techniques and algorithms attempt to aid decision making by analyzing stored data to find useful patterns and to build decision-support models. These extracted patterns and models help to reduce the uncertainty in decision-making environments.

Frequently, data may have sensitive information about previously surveyed human subjects. This raises many questions about the privacy and confidentiality of individuals (Grupe, Kuechler, & Sweeney, 2002). Sometimes these concerns result in people refusing to share personal information, or worse, providing wrong data.

Many laws emphasize the importance of privacy and define the limits of legal uses of collected data. In the healthcare domain, for example, the U.S. Department of Health and Human Services (DHHS) added new standards and regulations to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) to protect "*the privacy of certain individually identifiable health data*" (HIPAA, 2003). Grupe et al. (2002, Exhibit 1, p. 65) listed a dozen privacy-related legislative acts issued between 1970 and 2000 in the United States.

On the other hand, these acts and concerns limit, either legally and/or ethically, the releasing of datasets for legitimate research or to obtain competitive advantage in the business domain. Statistical offices face a dilemma of legal conflict or what can be called "war of acts." While they must protect the privacy of individuals in their datasets, they are also legally required to disseminate these datasets. The conflicting objectives of the Privacy Act of 1974 and the Freedom of Information Act is just one example of this dilemma (Fienberg, 1994). This has led to an evolution in the field of statistical disclosure limitation (SDL), also known as statistical disclosure control (SDC).

SDL methods attempt to find a balance between data utility (valid analytical results) and data security (privacy and confidentiality of individuals). In general, these methods try to either (a) limit the access to the values of sensitive attributes (mainly at the individual level), or (b) mask the values of confidential attributes in datasets while maintaining the general statistical characteristics of the datasets (such as mean, standard deviation, and covariance matrix). *Data perturbation* methods for microdata are one class of masking methods (Willenborg & Waal, 2001).

Data Mining vs. Statistical Analysis

Statisticians and researchers conduct surveys and collect datasets that are considered to be large when they contain a few hundred records (Hand, 1998). Traditional statistical techniques are the main (and the most suitable) tools for analyzing these datasets to make inferences and estimate population parameters. When the size of datasets is large, traditional statistical analysis techniques may not be the appropriate tools (Hand, 1998, 2000; Hand, Blunt, Kelly, & Adams, 2000). First, traditional statistical analysis may be inappropriate because almost any small difference in a large dataset becomes statistically significant. Second, large datasets may suggest that data was not collected for inference (parameter estimation) about the population. Third, in businesses, a significant amount of data is generated because of unplanned activities (e.g., transactional databases) and not from planned activities (e.g., experiment or survey designs). Therefore, for large datasets, data mining becomes more appropriate.

Examples of large datasets are abundant. MarketTouch, a company located in Georgia, USA, supports direct marketers with data and analytical tools (DMReview.com, 2004). It has a six-terabyte database called Real America Database (RADBÒ), which provides information about more than 93 million households and 200 million individuals. It is updated monthly with more than 20 million records.

Statistical agencies also experience this phenomenon of rapidly growing datasets. The US Census Bureau (Census, 2001) reported that the Census 2000 data consist of "*information about the 115.9 million housing units and 281.4 million people across the United States.*" These large sizes suggest the need for analytical tools that are suitable for large datasets, and again, data mining tools naturally come into play. Consequently, the Bureau provides programs with data mining capabilities such as DataFerrett (Federated Electronic Research, Review, Extraction and Tabulation Tool), which can be used to analyze and extract data from TheDataWeb - a repository of datasets that cover more than 95 subject areas.

Motivation for Privacy-Preserving Data Mining (PPDM)

Data mining techniques may lead to more significant threats to privacy and confidentiality than statistical analysis. Domingo-Ferrer and Torra (2003) make a connection between SDL methods and some data-mining AI (artificial intelligence) tools and suggest that disclosure and re-identification threats can be magnified.

DM tools can be used to aggregate or combine masked copies of a specific original dataset to reverse masking and re-build the original dataset, which raises a *confidentiality* issue. This is particularly true when unsophisticated SDL techniques are used and many masked copies are released. DM tools can also be used to enforce data integrity and consistency in distributed datasets by re-identifying different records belonging to the same individual raising a *privacy* issue.

These concerns about privacy and confidentiality when DM tools are used have led to the birth of privacy-preserving data mining (PPDM). The main goal of PPDM is to find useful patterns and build accurate models from datasets without accessing the individuals' precise original values in records of datasets (Agrawal & Srikant, 2000).

Related Work in Privacy-Preserving Data Mining (PPDM)

Similar to the classification of data mining (DM) techniques proposed by Berry and Linoff (2004), privacy-preserving data mining (PPDM) techniques can be classified as: (a) directed PPDM techniques: privacy-preserving estimation and privacy-preserving classification, and (b) undirected PPDM techniques: privacy-preserving association rules and privacy-preserving clustering.

Directed PPDM techniques try to model the relationship between a dependent variable and other (independent) variables in masked datasets. *Estimation* deals with continuous dependent variables and *classification* with categorical or binary dependent variables. The models obtained from the masked data using directed PPDM techniques must be the same (or similar) to that from the original dataset at the aggregate level, while protecting the privacy and confidentiality at the individual level.

In undirected PPDM, there is no concept of a dependent variable. Instead, the goal is to find unknown patterns and rules. *Clustering* is used to discover (and usually profile) homogenous subsets of data records and often used as a preprocessing tool (to segment the customer base, for example) before applying other DM technique (Berry & Linoff, 2004). *Association rules* are used to discover which items go together (are associated). Again, the goal of PPDM is to obtain similar