Secure Building Blocks for Data Privacy

Shuguo Han
Nanyang Technological University, Singapore

Wee Keong Ng
Nanyang Technological University, Singapore

INTRODUCTION

Rapid advances in automated data collection tools and data storage technology have led to the wide availability of huge amount of data. Data mining can extract useful and interesting rules or knowledge for decision making from large amount of data. In the modern world of business competition, collaboration between industries or companies is one form of alliance to maintain overall competitiveness. Two industries or companies may find that it is beneficial to collaborate in order to discover more useful and interesting patterns, rules or knowledge from their joint data collection, which they would not be able to derive otherwise. Due to privacy concerns, it is impossible for each party to share its own private data with one another if the data mining algorithms are not secure.

Therefore, privacy-preserving data mining (PPDM) was proposed to resolve the data privacy concerns while yielding the utility of distributed data sets (Agrawal & Srikant, 2000; Lindell.Y. & Pinkas, 2000). Conventional PPDM makes use of Secure Multi-party Computation (Yao, 1986) or randomization techniques to allow the participating parties to preserve their data privacy during the mining process. It has been widely acknowledged that algorithms based on secure multi-party computation are able to achieve complete accuracy, albeit at the expense of efficiency.

BACKGROUND

In recent years, PPDM has emerged as an active area of research in the data mining community. Several traditional data mining algorithms have been adapted to become privacy-preserving that include decision trees, association rule mining, *k*-means clustering, SVM, Naïve Bayes, and Bayesian network. These algorithms generally assume that the original data set has been horizontally and/or vertically partitioned with each partition privately held by one party. In privacy-preserving data mining algorithms, data are horizontally and/or vertically partitioned in order to spread the data across multiple parties so that no single party holds the overall data.

In this chapter, we focus on current work in privacy-preserving data mining algorithms that are based on Secure Multi-party Computation (Yao, 1986). In secure multi-party computation, there are two models that classify adversarial behaviors (Goldreich, 2002): the semi-honest model and the malicious model. Loosely speaking, a party in a semi-honest model follows the protocol properly but it keeps a record of all the intermediate computations during the execution. After the protocol, a party attempts to compute additional information about other honest parties. A party in the malicious model is allowed to diverge arbitrarily from the protocol. To force a malicious party to follow the protocol, zero-knowledge proofs can be applied. Zero-knowledge proofs as introduced by authors in (Goldwasser, Micali, & Rackoff, 1989) are proofs of the validity of an assertion made by a party without disclosing additional information.

MAIN FOCUS

In this section, we review current work on privacy-preserving data mining algorithms that are based on secure multi-party computation (Yao, 1986).

Privacy-Preserving Decision Trees

In (Lindell & Pinkas, 2000), the authors proposed a privacy-preserving ID3 algorithm based on cryptographic techniques for horizontally partitioned data involving two parties. The authors in (Du & Zhan, 2002) addressed the privacy-preserving decision tree

induction problem for vertically partitioned data based on the computation of secure scalar product involving two parties. The scalar product is securely computed using a semi-trusted commodity server. In the model, a semi-trusted third party helps two parties to compute scalar product; the third party will learn nothing about the parties' private data and is required not to collude with any of them. The authors in (Vaidya & Clifton, 2005a) extended the privacy-preserving ID3 algorithm for vertically partitioned data from two parties to multiple parties using the secure set intersection cardinality protocols.

Privacy-Preserving Association Rule Mining

In (Kantarcioglu & Clifton, 2004), the authors proposed a method to securely mine association rules for horizontally partitioned data involving three or more parties. The method incorporates cryptographic techniques to reduce the information disclosed. The authors in (Vaidya & Clifton, 2002) presented a privacy-preserving association rule mining algorithm for vertically partitioned data using secure scalar product protocol involving two parties. A secure scalar product protocol makes use of linear algebraic techniques to mask private vectors with random numbers. Solutions based on linear algebraic techniques are believed to scale better and perform faster than those based on cryptographic techniques. To extend association rule mining algorithm for vertically partitioned data to multiple parties, the authors in (Vaidya & Clifton, 2005b) proposed a secure set intersection cardinality protocol using cryptographic techniques. The communication and computation complexities of the protocol are O(mn) and $O(mn^2)$, where m and n are the length of private vectors and the number parties respectively.

Privacy-Preserving Clustering

The authors in (Vaidya & Clifton, 2003) presented a method to address privacy-preserving k-means clustering for vertically partitioned data involving multiple parties. Given a sample input that is partially held by different parties, determining which cluster the sample is closest must be done jointly and securely by all the parties involved. This is accomplished by a secure permutation algorithm (Du & Atallah, 2001) and a secure comparison algorithm based on the circuit evaluation protocol (Yao, 1986). The authors in (Geetha Jagannathan & Wright, 2005) proposed a new concept of arbitrarily partitioned data that is a generalization of horizontally and vertically partitioned data. They provided an efficient privacy preserving protocol for k-means clustering in an arbitrarily partitioned data setting. To compute the closest cluster for a given point securely, the protocol also makes use of secure scalar product protocols. The authors in (G. Jagannathan, Pillaipakkamnatt, & Wright, 2006) presented a simple I/O-efficient privacy-preserving k-clustering algorithm. They claimed that cluster centers produced by their algorithm are more accurate than those produced by the iterative k-means algorithm. The algorithm achieved privacy using secure scalar product protocols and Yao's circuit evaluation protocol (Yao, 1986). The authors in (Lin, Clifton, & Zhu, 2005) presented a technique that uses EM mixture modeling to perform clustering on horizontally partitioned distributed data securely. In the protocol, each partition is computed locally based on local data points. The global sum of the partitions from all parties is then computed without revealing the individual values by secure sum.

Privacy-Preserving Support Vector Machine

The authors in (Yu, Vaidya, & Jiang, 2006) proposed a privacy-preserving SVM classification algorithm for vertically partitioned data. To achieve complete security, the generic circuit evaluation technique developed for secure multiparty computation is applied. In another paper (Yu, Jiang, & Vaidya, 2006), the authors securely constructed the global SVM classification model using nonlinear kernels for horizontally partitioned data based on the secure set intersection cardinality protocol (Vaidya & Clifton, 2005b). The authors in (Laur, Lipmaa, & Mielikainen, 2006) proposed secure protocols to implement the Kernel Adaption and Kernel Perception learning algorithms based on cryptographic techniques without revealing the kernel and Gramm matrix of the data.

Privacy-Preserving Naïve Bayes

In (Kantarcioglu & Clifton, 2003), the authors presented a privacy-preserving Naïve Bayes classifier for horizontally partitioned data using secure sum—an instance the Secure Multi-party Computation (Yao, 1986). The