Chapter 49
Semantically Secure Classifiers for Privacy Preserving Data Mining

Sumana M.
M. S. Ramaiah Institute of Technology, India

Hareesha K. S.
Manipal Institute of Technology, India

Sampath Kumar
Manipal Institute of Technology, India

ABSTRACT

Essential predictions are to be made by the parties distributed at multiple locations. However, in the process of building a model, perceptive data is not to be revealed. Maintaining the privacy of such data is a foremost concern. Earlier approaches developed for classification and prediction are proven not to be secure enough and the performance is affected. This chapter focuses on the secure construction of commonly used classifiers. The computations performed during model building are proved to be semantically secure. The homomorphism and probabilistic property of Paillier is used to perform secure product, mean, and variance calculations. The secure computations are performed without any intermediate data or the sensitive data at multiple sites being revealed. It is observed that the accuracy of the classifiers modeled is almost equivalent to the non-privacy preserving classifiers. Secure protocols require reduced computation time and communication cost. It is also proved that proposed privacy preserving classifiers perform significantly better than the base classifiers.

DOI: 10.4018/978-1-7998-8954-0.ch049

INTRODUCTION

Privacy preserving data mining is essential when useful trends, decisions or patterns are to be discovered from the sensitive data. However, this data could be distributed or centrally available. Mining on the distributed data allows miners to model multiple sites and deduce important conclusions. Let us consider a situation where banks, credit card companies, tax collection agencies hold information about people within a locality. According to the Right to Financial Privacy Act, banks cannot reveal data about their customers to other companies or agencies. Similarly, Data Protection, Privacy and Law do not allow credit card companies to reveal any of their data. But useful inferences such as identifying fraud based on the tax collection, bank transactions and credit card details of an individual. Conclusions as to classify whether a person can be issued a loan, be provided with extra benefits or warned of a further loss or indicate whether the client can subscribe for a term deposit needs to be performed. The proposed privacy preserving classifiers creates classifier model from the data present at 3 different sites and enables any of these sites to make suitable decision. Similar situations can also be seen in hospital sector where hospitals hold the patient information including the type of treatment and its success. Doctors can obtain the private information of an individual and conclude on the type of treatment. Personal data of a patient could be present in bank datasets or insurance dataset where data cannot be revealed to the doctor.

Objectives of the Chapter

1. To build classifier models for the data vertically distributed at multiple sites.
2. To maintain the privacy of the sensitive data during mining and also use them for model building.
3. To construct efficient privacy preserving classifiers with improved performance.

BACKGROUND

The proposed approach allows to privately model classifiers based on the personal data maintained at insurance dataset and the hospital data for a large set of patients identified by name, age and locality without placing the data in a centralized site. As discussed in [(Agrawal & Aggarwal, 2001), (Yehuda & Benny, 2007), (Elisa, Dan, & Wei, 2008)], several approaches in Privacy Preserving data mining have evolved which can be broadly classified into perturbation, anonymization and cryptographic techniques. Perturbation involves transformations on the actual data before mining. This privacy preservation involves transfer of entire datasets as shown in (Jaideep, Hwanjo, & Xiaoqian, 2008) and (Hwanjo, Jaideep, & J, 2006) or partial datasets as mentioned in (Sun, Wei-Song, Biao, & Zhi-Jian, 2014) to single or multiple sites. A detailed survey on the needs and the various form of privacy preserving data mining can be found in (Lei, 2014). The key property of the randomization method is that the original records are not used after the conversion and data mining algorithms need to use the growing distributions of the perturbed data in order to perform the mining process. A symmetric perturbation approach and its reconstruction model that could be used for centralized association mining and classification is discussed in (Shipra, Jayant, & P, 2009).

(Agrawal & Srikant, 2000) Introduced the concept of perturbation in privacy preserving data mining were assorted algorithms are discussed to restructure distributions and learn a decision tree classifier from the perturbed data. Similar approaches of perturbation for privacy preserving association rule mining is