Chapter 22
Accurate Classification Models for Distributed Mining of Privately Preserved Data

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

Hareesha K. S.
Manipal Institute of Technology, Udupi, India

ABSTRACT

Data maintained at various sectors, needs to be mined to derive useful inferences. Larger part of the data is sensitive and not to be revealed while mining. Current methods perform privacy preservation classification either by randomizing, perturbing or anonymizing the data during mining. These forms of privacy preserving mining work well for data centralized at a single site. Moreover the amount of information hidden during mining is not sufficient. When perturbation approaches are used, data reconstruction is a major challenge. This paper aims at modeling classifiers for data distributed across various sites with respect to the same instances. The homomorphic 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.

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

DOI: 10.4018/978-1-5225-8897-9.ch022

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.

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 conversed by (Rizvi & Haritsa, 2002) and (Zhang, Wang, & Zhao, 2004). (Latanya, 2002) And (Ashwin, Daniel, & Johannes, 2007) discusses anonymization techniques that can be used for privacy preserving data mining which involves two essential methods: generalization and suppression. (Arik, Assaf, & Ran, 2006) Confers the construction of a decision tree classifier using k-anonymity technique. Anonymization - Based Privacy preserving methods involves several attacks as seen in (Mielikainen, 2004). A privacy preserving distributed Naïve Bayes classifier for horizontally partitioned distributed data is proposed using k-anonymity constraints is mentioned in (Lambodar, 2013). (Elisa, Dan, & Wei, 2008) Clearly mentions that cryptographical approaches provide high level of data privacy compared to the randomization or anonymization approach of privacy preserving data mining.

The theoretical structure for all cryptographic protocols is Secure Multiparty Computation. Yao first developed a provably secure solution for the two-party comparison problem (Yao's Millionaire Protocol) (Yao, 1986). This approach to multiparty computations is discussed in (Goldreich, Micali, & Wigderson, 1987). However, the generic circuit evaluation technique does not work efficiently for large quantities of data. A detailed description of homomorphism is provided by (Benaloh, 1986). These homomorphic properties are used to perform secure computations in (Jaideep, Murat, & Clifton, Privacy-preserving Naïve Bayes classification, 2008), (Chen & Zhong, 2009) and (Yuan & Sheng, 2013). The additive and multiplicative homomorphic properties work well for our techniques. Privacy preserving protocols for back propagation and extreme learning machine for horizontally and vertically partitioned data using cryptography is discussed by Saeed and Ali in (Saeed, 2012). (Blum & Goldwasser, 1984), describes an