Chapter 21
Privacy Preserving Distributed K-Means Clustering in Malicious Model Using Verifiable Secret Sharing Scheme

Sankita Patel
S. V. National Institute of Technology, India

Mitali Sonar
Shankersinh Vaghela Bapu Institute of Technology, India

Devesh C. Jinwala
S. V. National Institute of Technology, India

ABSTRACT

In this article, the authors propose an approach for privacy preserving distributed clustering that assumes malicious model. In the literature, there do exist, numerous approaches that assume a semi honest model. However, such an assumption is, at best, reasonable in experimentations; rarely true in real world. Hence, it is essential to investigate approaches for privacy preservation using a malicious model. The authors use the Pederson's Verifiable Secret Sharing scheme ensuring the privacy using additively homomorphic secret sharing scheme. The trustworthiness of the data is assured using homomorphic commitments in Pederson's scheme. In addition, the authors propose two variants of the proposed approach - one for horizontally partitioned dataset and the other for vertically partitioned dataset. The experimental results show that the proposed approach is scalable in terms of dataset size. The authors also carry out experimentations to highlight the effectiveness of Verifiable Secret Sharing scheme against Zero Knowledge Proof scheme.

DOI: 10.4018/978-1-5225-8897-9.ch021

INTRODUCTION

Maintaining privacy has always been crucial for an information system that collects large amounts of data pertaining to their customers. 'Data mining' or 'knowledge discovery research' undermines extracting potentially useful information from a raft of data. On the darker side, the user-friendliness of data mining results jeopardizes the privacy of the data. This can be countered by integrating privacy-preserving mechanisms in data mining tools. Due to the increasing need of distributed databases in business environment, need for Privacy Preserving Distributed Data Mining (PPDDM) becomes imperative. In distributed databases, the dataset may be horizontally or vertically partitioned. In horizontal partitioning of dataset, the parties have different number of objects each having same number of attributes; whereas in vertical partitioning of dataset, the parties have same number of objects but with partial set of attributes with them.

PPDDM applications can be characterized in two models viz. the *corporate model* and the *World Wide Web* model (Clifton, 2002). In *corporate* model, we assume that data is created and held by the participating parties whereas in the *World Wide Web* model individuals provide the data in electronic form themselves. In this paper, our focus is on investigating the privacy concerns associated with the corporate model, when it is necessary to share the data. The privacy policy and law prevents the parties from over pooling their data or revealing it to each other, due to the confidentiality of records. In such cases, classical data mining solutions cannot be used. Rather it is necessary to find a solution that enables the parties to collaboratively compute the desired data mining algorithms on the union of their databases, without ever pooling or revealing their data. The goal of our study is to propose an approach that enables multiple parties to collaboratively perform data mining in corporate model without jeopardizing the privacy of their data. We focus on detection of malicious behaviour by the parties so as to assure the trustworthiness of data. In particular, we focus on clustering application of data mining.

Among the two main approaches of Privacy Preserving Data Mining (PPDM) viz. the *Randomization based* and the *Cryptography based* approach, the latter provides higher level of privacy (Oliveira, 2003; Pederson, 2007). However, the cryptography-based approach is expensive in terms of computational and communication overheads and so the existing protocols proposed are not scalable with respect to dataset size and number of parties (Pederson, 2007). Therefore, the chief concern in designing such protocols must be on minimizing the overheads incurred in their design and implementation. In this paper, we address this issue. We focus on K-Means clustering algorithm of data mining and propose cryptography based approach for distributed K-Means clustering.

The existing approaches in cryptography-based category are classified into three categories viz. the Oblivious Transfer based, the homomorphic encryption based and the secret sharing based. Among these, the oblivious transfer based approach incurs considerable overheads in terms of computation and communication. The homomorphic encryption based approach achieves better efficiency as compared to oblivious transfer based approach. Still, the use of classical public key ciphers imposes higher computational overhead. Due to this, the homomorphic encryption based approach is not scalable with respect to dataset size. In practical scenario, large datasets exist and scalable technique for privacy preservation is imperative. Recently, researchers have come up with the solutions that incorporate the secret sharing schemes in data mining tools for preserving privacy (Doganay, 2008; Upmanyu, 2010). Secret sharing based approach avoids use of costly public key operations over a large field by using linear primitive operations over a small field in computation while providing privacy. Due to this, the secret sharing based approach preserves privacy with lesser computational cost as compared to the homomorphic encryption based approach. Hence, in this paper, we focus on the secret sharing based approach.