Chapter 212
Distributed Privacy Preserving Clustering via Homomorphic Secret Sharing and Its Application to (Vertically) Partitioned Spatio-Temporal Data

Can Yildizli
Sabanci University, Turkey

Thomas Brochmann Pedersen
Sabanci University, Turkey

Yucel Saygin
Sabanci University, Turkey

Erkay Savas
Sabanci University, Turkey

Albert Levi
Sabanci University, Turkey

ABSTRACT

Recent concerns about privacy issues have motivated data mining researchers to develop methods for performing data mining while preserving the privacy of individuals. One approach to develop privacy preserving data mining algorithms is secure multiparty computation, which allows for privacy preserving data mining algorithms that do not trade accuracy for privacy. However, earlier methods suffer from very high communication and computational costs, making them infeasible to use in any real world scenario. Moreover, these algorithms have strict assumptions on the involved parties, assuming involved parties

DOI: 10.4018/978-1-61350-323-2.ch212

will not collude with each other. In this paper, the authors propose a new secure multiparty computation based k-means clustering algorithm that is both secure and efficient enough to be used in a real world scenario. Experiments based on realistic scenarios reveal that this protocol has lower communication costs and significantly lower computational costs.

INTRODUCTION

Massive amounts of data are collected for various reasons by many organizations with the hope that data mining technology will extract useful knowledge from the collected data and turn it into something beneficial for the organization. In fact, data mining technology proved its success in numerous areas such as business intelligence, life-sciences, and security. On the other hand, the popularity of data mining was about to pave the way to its demise. Part of the reason for that is the launch of large scale projects related to homeland security. Some projects were actually stopped since they failed to meet privacy concerns. According to a recent article in Computer World by Vijayan (2007) "The chairman of the House Committee on Homeland Security, has asked Department of Homeland Security Secretary Michael Chertoff to provide a detailed listing of all IT programs that have been canceled, discontinued or modified because of privacy concerns". In addition to that, the Chairman also asked for information about the measures being taken to address privacy issues (Vijayan, 2007). As a result of increased privacy concerns, data mining researchers focused on developing techniques that would enable data mining while preserving the privacy of individuals and started a popular branch of research named "privacy preserving data mining" (Agrawal & Srikant, 2000). Protocols based on statistics and cryptography were proposed for privacy preserving classification, clustering, and pattern mining in centralized and distributed environments. However, privacy preserving data management, in general, is still an ongoing research topic, and efficient, as well as provably secure, methods without strong assumptions are yet to be proposed.

In this work, we propose a new secure multiparty computation algorithm for distributed privacy preserving k-means clustering. Our algorithm is both more efficient and more secure than the current state of the art secure k-means clustering algorithm of Vaidya and Clifton (2003). In this protocol we avoid the computationally heavy public key encryption. Instead we use secret sharing as the underlying cryptographic primitive. The main contributions of this work can be listed as:

- We show that our protocol outperforms the state of the art protocol by Vaidya and Clifton (2003). Backed by experiments we show that our protocol has a much lower computational overhead due to the fact that we replace computationally expensive public key encryption operations with additive secret sharing.
- As a case study we apply our technique on a trajectory data set obtained in the context of the GeoPKDD project (http://www.geopkdd.eu/).
- To the best of our knowledge, this is the first work which implements and tests privacy preserving clustering in a realistic setting. We run the protocols on a real dataset of trajectories in a novel testing platform. The test platform is a combination of simulation and real execution, which enables a detailed comparison of the protocols in a controlled environment.
- We take full advantage of the security model, which we share with (Vaidya & Clifton, 2003).

The work presented in this paper extends the work done by Kaya et al. (2007) and Doganay et