Chapter 49
Privacy-Preserving Hybrid K-Means

Zhiqiang Gao
Engineering University of PAP, China

Yixiao Sun
Official College of PAP, China

Xiaolong Cui
Engineering University of PAP, China

Yutao Wang
Engineering University of PAP, China

Yanyu Duan
Engineering University of PAP, China

Xu An Wang
Engineering University of PAP, China

ABSTRACT

This article describes how the most widely used clustering, k-means, is prone to fall into a local optimum. Notably, traditional clustering approaches are directly performed on private data and fail to cope with malicious attacks in massive data mining tasks against attackers' arbitrary background knowledge. It would result in violation of individuals' privacy, as well as leaks through system resources and clustering outputs. To address these issues, the authors propose an efficient privacy-preserving hybrid k-means under Spark. In the first stage, particle swarm optimization is executed in resilient distributed datasets to initiate the selection of clustering centroids in the k-means on Spark. In the second stage, k-means is executed on the condition that a privacy budget is set as ε/2t with Laplace noise added in each round of iterations. Extensive experimentation on public UCI data sets show that on the premise of guaranteeing utility of privacy data and scalability, their approach outperforms the state-of-the-art varieties of k-means by utilizing swarm intelligence and rigorous paradigms of differential privacy.

DOI: 10.4018/978-1-5225-7113-1.ch049

1. INTRODUCTION

Nowadays, big data is ubiquitous and abundant as the booming growth of cloud computing and mobile Internet (Xia et al., 2016; Li, Taniar & Indrawan-Santiago, 2017). However, it poses a rising challenge on individuals' raw data when data-mined or released by untrustworthy data analyzers. Individual privacy is always faced with threatens from potential malicious attackers (Khan & Al-Yasiri, 2016; Sander, Teh & Sloka, 2017; Brocardo, Rolt, Dias, Custodio & Traore, 2017). Furthermore, with massive deployment of cloud computing and increasing demand of big data services, traditional data mining methods are in urgent requirement to be optimized and security-enhanced (Fu, Huang, Ren, Weng & Wang, 2017; Xiong et al., 2017). Consequently, privacy-preserving data mining (PPDM) as well as privacy-preserving data releasing (PPDR) have become extremely challenging problems. Overall, the research direction of privacy-preserving techniques can be illustrated in Table 1.

As the most commonly used clustering method, k-means (Lloyd, 1982; Yamada et al., 2017; Ma, 2017) has the prominent characteristics of fast convergence and low execution complexity. Since the proposal of PPDM, privacy-preserving k-means has attracted lots of attention from various fields. Specifically, Su et al. (2016) systematically investigated the concept of differential privacy data mining and proposed a composite k-means algorithm which integrates interactive and non-interactive methods. Ren et al. (2017) proposed a DPLK-means algorithm which improved the selection of the initial center points to each subset while the added noise reduced the performance of clustering. Additionally, regarding the modes of horizontal, vertical and arbitrary data storage, large amounts of privacy-preserving data mining schemes are specifically designed accordingly. Xing et al. (2017) provided a privacy preserving k-means containing two privacy-preserving algorithms without disclosing private information in clusters.

From another perspective, multiparty k-means is developed by conforming to such privacy-preserving protocols as secure multiparty computation (SMC) (Samet & Miri, 2007; Upmanyu, Namboodiri, Srinathan & Jawahar, 2010). Guided by SMC, multi-sourced data can be shared by several parties and each party independently produces k clusters securely. Meanwhile, all data are coordinated in a privacy-preserving manner. Doganay et al. (2008) studied the privacy of k-means clustering protocols and highlighted the situation where data is shared within two and more participants respectively. Miyajima et al. (2017) explored to combine reinforcement learning (RL) with SMC and proposed learning methods with SMC for RL. However, the most promising scheme of homomorphic encryption (Chen, 2015; Jain et al., 2017) is still immature and inevitably results in overwhelming computing expense. In a nutshell, early

Table 1. Existing research direction of privacy-preserving techniques

Direction	Paradigm	Feature
PPDR	k-anonymity (Sweeney, 2002), l-diversity (Machanavajjhala, Kifer, Gehrke & Venkitasubramaniam, 2007), t-closeness (Li, Li & Venkatasubramanian, 2007)	Based on background knowledge; Managed by a centralized data curator; Unable to provide strictly mathematical guarantee.
PPDM	Differential privacy (Dwork, McSherry, Nissim & Smith, 2006)	Strong privacy guarantee; Centralized and decentralized model.
	SMC (Samet & Miri, 2007; Miyajima et al., 2017)	Computation overheads; Strict limitation on involved parties.
	Homomorphic encryption (Chen, 2015; Jain, Rasmussen & Sahai, 2017)	Computation overheads; Far from large-scaled production.