Practical Privacy-Preserving MapReduce Based K-means Clustering over Large-scale Dataset

ABSTRACT

Clustering is one major task of exploratory data mining and statistical data analysis, which has been ubiquitously adopted in many domains, including healthcare, social network, image analysis, pattern recognition, etc. Meanwhile, the rapid growth of big data involved in today's data mining and analysis also introduces challenges for clustering over them in terms of volume, variety, and velocity. To efficiently manage large-scale datasets and support clustering over them, public cloud infrastructure is acting the major role for both performance and economic consideration. Nevertheless, using public cloud services inevitably introduces privacy concerns. This is because not only many data involved in data mining applications are sensitive by nature, such as personal health information, localization data, financial data, etc, but also the public cloud is an open environment operated by external third-parties [1]. For example, a promising trend for predicting an individual's disease risk is clustering over existing patients' health records [2], which contain sensitive patient information according to the Health Insurance Portability and Accountability Act (HIPAA) Policy [3]. Therefore, appropriate privacy protection mechanisms must be placed when outsourcing sensitive datasets to the public cloud for clustering.

EXISTING SYSTEM

The problem of privacy-preserving K-means clustering has been investigated under the multi-party secure computation model [4]–[9], in which owners of distributed datasets interact for clustering without disclosing their own datasets to each other. In the multi-party setting, each party has a collection of data and wishes to collaborate with others in a privacy preserving manner to improve clustering accuracy. Differently, the dataset in clustering outsourcing is typically owned by a single entity, who aims at minimizing the local computation by delegating the clustering task to a third-party cloud server. In addition, existing multi-party designs always rely on powerful but expensive cryptographic primitives (e.g., secure circuit evaluation, homomorphic encryption and oblivious transfer) to achieve collaborative secure computation among multiple parties, and are inefficient for large-scale datasets. Thus, these multi-party designs are not practical for privacy-preserving outsourcing of clustering. Another line of research that targets at efficient privacy-preserving clustering is to use distancepreserving data perturbation or data transformation to encrypt datasets [10], [11]. Nevertheless, utilizing data perturbation and data transformation for privacy-preserving clustering may not achieve enough privacy and accuracy guarantee [12], [13]. For example, adversaries who get a few unencrypted data records in the dataset will be able to recover rest records protected by data transformation [12]. Recently, the outsourcing of K-means clustering is studied in ref [14] by utilizing homomophic encryption and order preserving index. However, the homomophic encryption utilized in [14] is not secure as pointed out in ref [15]. Moreover, due to the cost of relative expensive homomophic encryption, ref [14] is efficient only for small datasets, e.g., less than 50,000 data objects. Another possible candidate to achieve privacy-preserving K-means clustering is to extend existing privacy-preserving K-nearest neighbors (KNN) search schemes [16]–[18]. Unfortunately, these privacy-preserving KNN search schemes are limited by the vulnerability to linear analysis attacks [16], the support up to two dimension data [17], or accuracy loss [18]. In addition, KNN is a single round search task, but K-means clustering is an iterative process that requires the update of clustering centers based on the entire dataset after each round of clustering. Considering the efficient support over large-scale datasets, these update processes also need to be outsourced to the cloud server in a privacy-preserving manner.

DISADVANTAGES

- The homomophic encryption utilized in [14] is not secure.
- Privacy-preserving KNN search schemes are limited by the vulnerability to linear analysis attacks [16], the support up to two dimension data [17], or accuracy loss [18].

PROPOSED SYSTEM

In this paper, we propose a practical privacy-preserving K-means clustering scheme that can be efficiently outsourced to cloud servers. Our scheme allows cloud servers to perform clustering directly over encrypted datasets, while achieving comparable computational complexity and accuracy compared with clusterings over unencrypted ones. We also investigate secure integration of MapReduce into our scheme, which makes our scheme extremely suitable for cloud computing environment. Thorough security analysis and numerical analysis carry out the performance of our scheme in terms of security and efficiency. Experimental evaluation over a 5 million objects dataset further validates the practical performance of our scheme.

ADVANTAGES

- Our scheme allows cloud servers to perform clustering directly over encrypted datasets.
- Security analysis and numerical analysis carry out the performance of our scheme in terms of security and efficiency

SYSTEM REQUIREMENTS

H/W System Configuration:-

Processor	-	Pentium –IV
RAM	-	4 GB (min)
Hard Disk	-	20 GB
Key Board	-	Standard Windows Keyboard
Mouse	-	Two or Three Button Mouse
Monitor	-	SVGA

S/W System Configuration:-

Operating System	:	Windows 7 or 8 32 bit
Application Server	:	Tomcat5.0/6.X
Backend coding	:	Java
Tool	:	Virtual Box
Environment	:	Ubuntu
Technology	:	Hadoop