Privacy Preserving DBSCAN Algorithm for Clustering

K. Anil Kumar and C. Pandu Rangan

Department of Computer Science and Engineering
Indian Institute of Technology - Madras
Chennai - 600036, India
kumark@cse.iitm.ernet.in, rangan@iitm.ernet.in

Abstract. In this paper we address the issue of privacy preserving clustering. Specially, we consider a scenario in which two parties owning confidential databases wish to run a clustering algorithm on the union of their databases, without revealing any unnecessary information. This problem is a specific example of secure multi-party computation and as such, can be solved using known generic protocols. However there are several clustering algorithms are available. They are applicable to specific type of data, but DBSCAN [4] is applicable for all types of data and the clusters obtained by DBSCAN are similar to natural clusters. However, DBSCAN [4] algorithm is basically designed as an algorithm working on a single database. In this paper we proposed a protocols for how the distances are measured between data points, when the data is distributed across two parties. By using these protocols we propose the first novel method for running DBSCAN algorithm operating over vertically and horizontally partitioned data sets, distributed in two different databases in a privacy preserving manner.

1 Introduction

A key problem that arises with large collection of information is that of confidentiality. The need for privacy is due to law or can be motivated by business interests. However, there are situations where sharing of data that can lead to mutual gain. Recently, more emphasis has been placed on preserving the privacy of user-data aggregations, ex. databases of personal information. Access to these collection is, however, enormously useful. It is from this balance between privacy and utility that the area of privacy preserving data mining [1,7] emerged. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private after the mining process. In this paper we are concentrating particularly on privacy preserving clustering algorithms.

Consider the following scenario, Let parties X and Y own private databases $D_X$ and $D_Y$. The parties wish to get clusters on their joint data, apply a clustering algorithm (ex: DBSCAN [4]) to the joint database $D_X \cup D_Y$ without revealing any unnecessary information about their individual databases. i.e. the only information learned by $X$ about $D_Y$ is that which can be learned from the output
of the clustering algorithm, and vice versa. We are giving two kinds of solutions for the above problem, one with the help of trusted third party (TTP) and another TTP not available.

There are several applications of clustering. Any application of clustering where there are privacy concerns is a possible candidate for our privacy-preserving clustering algorithm. For example, suppose network traffic is collected at two ISPs, and the two ISPs want to cluster their joint network traffic without revealing their individual traffic data. Out DBSCAN for horizontally partitioned algorithm can be used to obtain joint clusters while respecting the privacy of the network traffic at the two ISPs. Another example, two organizations, an internet marketing company and an on-line retail company, have datasets with different attributes for a common set of individuals. These organizations decides to share their data for clustering to find the optimal customer targets so as to maximize using each other’s data without learning anything about the attribute value of each other? For this problem to solve our DBSCAN for vertically partitioned algorithm is used.

2 Related Work

Recently privacy preserving data mining has been a very active area of research. Privacy preserving data mining introduced by Rakesh Agrawal and Ramakrishnan Srikanth [1]. Lindall and Pinkas [17] arouse as a solution to this problem by allowing parties to cooperate in the extraction of knowledge without any of the cooperating parties having to reveal their individual data to any other party. Initial focus in this area was on construction of decision trees from distributed datasets. There is also a significant body of research on privacy preserving mining of association rules [10]. We will focus only on privacy preserving clustering. Privacy preserving clustering has been previously addressed Jha et al [6] gave a distributed privacy preserving algorithm for k-means clustering when the data is horizontally partitioned. Oliveira and Zaiane [8] gave a privacy preserving solution for clustering based on perturbation. In their work they used geometric data transformation methods for perturbing the data and they give solution for hierarchical and partition based clustering on perturbed data. Vaidya and Clifton [11] introduced the problem of privacy preserving clustering when the data is vertically partitioned. Prasad and Rangan [9] proposed a solution, privacy preserving BIRCH. Their algorithm works for vertically partitioned, large databases. Jagannathan and Wright [5] introduced the notion of arbitrarily partitioned databases and gave a privacy preserving solution of k-means clustering. In this paper we present a privacy preserving protocol for DBSCAN clustering which works for large databases.

3 Preliminaries

In this section we briefly describe the primitives that we used in our work for clustering the distributed data securely.