

Abstract

The interest of collecting and exploring large amounts of data to extract valuable knowledge has become critical to commercial companies and governmental organizations, which is the motivation of data mining. The current tendency is that two or several organizations share their datasets and give them as inputs to the process of data mining in order to have more effective results. This raised a real problem of privacy that most of these data relate to individuals and their personal information. The very active research area of privacy preserving data mining aims to extract useful information from data coming from multiple sources, while preserving these data against disclosure or loss. Clustering is also a more exploratory data mining task which the aim is to classify items described by features into groups, according to some similarities in a given context of application and poses the same problem of privacy when data come from different sources. K-means is one of the algorithms of clustering and the most widely used. Most of works in privacy preserving clustering are developed on the k-means algorithm by applying the model of secure multi-party computation. The ways in which data are shared or distributed on these parties may be different. The first solution of preserving privacy in k-means algorithm was proposed by Jaideep Vaidya and Chris Clifton in 2003 on vertically partitioned dataset. Thus, approaches allowing solving the problem on a vertical, horizontal and even arbitrary partitioned dataset were proposed, but the preservation of privacy is still not complete. The major problem is to reveal the minimum of information during the execution of the algorithm, especially in k-means iterations, which poses a real challenge for secure multi party computation. This paper consists in drawing up a panorama of all works of preserving privacy in k-means clustering algorithm, classifies the various approaches according to the used distribution dataset, while presenting the weaknesses and strengths of each solution. The permanent growth of the data and the need emerging to explore them requires a real thinking to effectively protect them, especially that these data become increasingly individualized.

Key words: privacy preservation, k-means clustering algorithm, secure multi-party computation, data distribution.