Density Peaks Clustering with Differential Privacy

Shengna Guo
Renmin University of China
shengnaguo@126.com

Xiaofeng Meng
Renmin University of China
xfmeng@ruc.edu.cn

ABSTRACT

Density peaks clustering (DPC) is a well-known density-based clustering algorithm which finds clusters of arbitrary shapes. However, attackers can deduce sensitive points from the known point when the cluster centers and their sizes are exactly released. To the best of our knowledge, this is the first time that privacy protection has been applied to DPC. In this paper, we provide a density peaks clustering privacy protection(DPCP) model to obtain the clustering results without revealing the data via differential privacy protection. We are inspired by the idea of dividing and conquering algorithm. Firstly, we divide the datasets into relatively independent groups and then add noises. We employ a parallel computation through MapReduce to improve the efficiency. Secondly, according to the principle that the privacy budget can be superimposed in large-scale datasets, we introduce $\epsilon_1+\epsilon_2$-differential privacy protection model and ensure the accuracy via data replication and filter. Finally, we also provide performance analysis and privacy proof of our solution.

Keywords

Differential privacy; Voronoi diagram partition; Clustering; Privacy preserving; Data mining

1. INTRODUCTION

Typical partitioning-based clustering algorithms (the most common is k-means) are not able to detect non-spherical clusters. But the density-based clustering can be done. For the DBSCAN that is the classical density-based clustering algorithm. There are several privacy-preserving algorithms. Such as, Kumar et al. discussed both horizontally and vertically partitioned data. Jinfei et al. oriented to horizontally, vertically and arbitrarily partitioned data and designed a Multiplication Protocol based on Pailler's Additive Homomorphic cryptosystem.

The big data era has been an enormous increase in the multi-dimensional data and diversification data. We need a simple and fast clustering algorithm which can be applied to data sets with various types and shapes. For the above problems, Alex Rodriguez and Alessandro Laio propose an alternative approach. The algorithm has its basis in the assumptions that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density.For each data point i, they compute two quantities: its local density ρ_i and δ_i its distance from points of higher density. To the best of our knowledge, this is the first time that privacy protection has been applied to this clustering process. In this paper, we study the density peaks clusters under differential privacy protection. For instance, it requires to measure distance between any point of objects when computing ρ and δ for each data. Additional, if we directly add noise to the raw data. The model's efficiency and scalability will be limits especially for high-dimensional data.

Therefore, we are inspired by the idea of divide and conquer algorithm and the principle that is the privacy budget can be superimposed in high dimensional data. Our main contributions are summarized as follows[1]:

1) We introduce the idea of Voronoi-diagram partitioning. The original dataset is divided into relatively independent groups. At the same time, in order to prevent errors in the calculation, we use the idea of replication and filtering.

2) We introduce ϵ-differential privacy protection($\epsilon = \epsilon_1 + \epsilon_2$). Because the clustering is determined by the parameters ρ and δ, and these two parameters are all operated on the original dataset.

3) we conduct extensive experiments on three data sets with different dimensions and levels. The experiment results show that our algorithm is effective and accurate.

In this paper, we study the privacy preserving clustering problem and provide DPCP algorithm. We have provided $\epsilon = \epsilon_1+\epsilon_2$-differential privacy preserving model. We provided performance analysis and privacy proof of our solution.

8th Biennial Conference on Innovative Data Systems Research (CIDR '17) January 8-11, 2017 , Chaminade, California, USA.

[1]Acknowledge:This research was partially supported by the grants from the Natural Science Foundation of China (No. 91646203, 61532016, 61532010, 61379050,, 91224008); the National Key R & D Program(No. 2016YFB1000602, 2016YFB1000603); Specialized Research Fund for the Doctoral Program of Higher Education(No. 20130004130001), and the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University(No. 11XNL010).