Density Peaks Clustering with Differential Privacy

Guo Shengna
School of Information, Renmin University of China, Beijing
shengnaguo@126.com

Meng Xiaofeng
School of Information, Renmin University of China, Beijing
xfmeng@ruc.edu.cn

ABSTRACT

Density peaks clustering (DPC) is a latest and well-known density-based clustering algorithm which offers advantages for finding clusters of arbitrary shapes compared to others algorithm. However, the attacker can deduce sensitive points from the known point when the cluster centers and sizes are exactly released in the cluster analysis. To the best of our knowledge, this is the first time that privacy protection has been applied to DPC. In this paper, we provide density peaks clustering privacy protection(DPCP) model to obtain the clustering results without revealing the data via differential privacy protection, in which the privacy protection is achieved by add Laplace noise to local density ρ and distance δ. However, the computation complexity will reaches O(n) and have an inaccurate clustering results when adding noise to the data set directly. Therefore, we are inspired by the idea of divide and conquer algorithm. Firstly, we divide the data set into relatively independent groups by Voronoi diagram and then adding noises. We employ a parallel computing by MapReduce to improve the efficiency. Secondly, according to the principle that is the privacy budget can be superimposed in high dimensional data. We introduces $\epsilon_1+\epsilon_2$-differential privacy protection model and ensure the accuracy of the calculation via data replication and filter. Where ϵ_1 and ϵ_2 to protect ρ and δ respectively. Finally, through a lot of experiments, we also provide performance analysis and privacy proof of our solution.

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 divesification 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 its distance δ_i 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 value 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 data set is divided into relatively independent grouping. Meanwhile, in order to prevent errors in the calculation of ρ and δ, we use the idea of replication and filtering.

2) We introduce $\epsilon = \epsilon_1+\epsilon_2$-differential privacy protection, in which ϵ_1 and ϵ_2 to protect ρ and δ, respectively. Because the clustering is determined by parameters ρ and δ, and these two parameters are all operated on the original data.

3) We conduct extensive experiments on three data sets with different dimensions and levels. The experimental 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.

DOI: 10.1145/1235

[1]Acknowledgments:This research was partially supported by the grants from the Natural Science Foundation of China (No. 91646203, 61532016, 61532010, 61379050); the National Key Research and Development Program of China (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).