Clustering with Over-Representation Constraint

Camilo Cedeno-Tobon, Kevin Ko, Simran Shinh, Irina Wang

September 28, 2019

1 Introduction

There are a myriad of clustering algorithms, all taking into account different constraints to yield the desired results. One particular concern in clustering is fairness and over-representation. Having one particular entity that comprises the majority of a cluster is often undesirable and at times dangerous. Despite the dangers that this situation poses, the over-representation constraint has been largely unexplored. In this project, we seek to validate and potentially improve the clustering algorithm that Google Research scholars Ahmadian, Epasto, Kumar, and Mahdian have created that simultaneously addresses the over-representation constraint and minimizes the classical clustering cost.

2 Original Algorithm

The original model consists of the foundation for solving the k-clustering problem which contains a set of D points to be clustered into at most k clusters. The terminology used for a regular k-clustering problem denotes points as clients and facilities as cluster centers. This paper introduces a third parameter to the classic k-clustering problem: each point has a color, and there exists a constraint on the representation of each color in each cluster.

The initial approach to solving the α-capped k-center clustering problem involves first finding a linear program relaxation of the problem and then modifying the fractional result to obtain an integral solution. In the process, approximation and fairness are both sacrificed to yield a simpler algorithm (the bicriteria algorithm) with an improved additive 2 violation, compared to an additive 4 violation given by the Bera et al. [1] and Bercea et al. [2] algorithms. As a result, for each color and facility pair, the cap for the allowed number of clients is exceeded by at most 2 clients (hence additive 2 violation).

3 Proposed Experimentation

Data sets: the same public data sets in the paper (page 6), along with other data sets from the UCI library – Adult (A US census record data set from 1994), Diabetes (A data set about a study with diabetes patients), Bank (A data set about a marketing campaign of a Portuguese bank) will be used to test the algorithm.

We will first implement the algorithms outlined in the paper, and test them against their own results by emulating their parameters. Next, we will test additional parameters to see if the results hold, and also test them over additional data sets. We will also attempt to generalize the algorithm for fair k-median clustering.

4 Project Proposal

In real life, there are risks in having clusters that are too homogeneous in nature. For example, in Congress, it is dangerous to create committees with one dominant ideology/political view. Similarly, homogenous clusters in social networks can be bad because it could increase the political polarization we see today. By implementing the clustering algorithms that prevent over-representation in clusters and testing them on data from the real world, we can have more accurate causal inferences.

5 References

https://www.kdd.org/kdd2019/accepted-papers/view/clustering-without-over-representation