Progress report

Zehui Li

31 October 2018

Statement of Research topic

Following the impossibility theorem proposed by Kleinberg [1], numerous relaxation methods on the axiomatic system are proposed in recent years. In particular. relaxation of consistency theorem is the focus of this paper. The consistency theorem proposed by Kleinberg, while reasonable and simple, it give a relative strict restriction on clustering algorithms - it require the clustering algorithm to give the same partition results even when the data set is perturbed profoundly. Following the idea of Vincent(2018) [2], we will relax the consistent theorem by adding constraints on it, but instead of using 'natural number of clusters', two quantitative variables are used to describe the extend of perturbation, and rand index is used to measure the difference of two partition results. If an inequality relation or equality relation between the two measurement of perturbation and the clustering results can be found, consistent theorem will be re-defined using this relation, which result in a more general results than refined consistency theorem in Vincent(2018) [2].

Summary of progress

To properly measure the degree of perturbation, two metrics are defined as following. But before two metrics, the definition of distance d between to clusters $C1, C2$ are given (if the coordinates of the data is given):

$$d(C1, C2) = D(\overline{X_1}, \overline{X_2}) \tag{1}$$

where the $\overline{X_i}$ is the average points for the cluster C_i. In the case of the coordinate of the data points is not given, we define the

$$d(C1, C2) = Hausdorff(C_1, C_2) \tag{2}$$

where Hausdorff is the Hausdorff distance beween two sets.

First metric M_1 and second quantity M_2 is defined upon this distance. M_1 is used to measure the changes of the ratio between the different clusters, and M_2 is used to measure the changes of ratio between

the points within a cluster.(Formal definition is a too long to state here). At the same time, we have **rand index** to measure the difference between two partition results.

Because the relation between three quantities is not clear, we first rely on the computer simulation to identify the potential trend. To implement this simulation, we will first perturb the data points, then record the measures of each perturbation as one three dimensional point - $(M_1, M_2, randIndex)$.After numerous data points are generated, plot of these data points can provide some insights for defining the relation between the extend of perturbation and difference between the two partition results. Furthermore, statistical method can help to capture this relation. Apart from the simulation, we are also try to analyse the properties of these three metrics from theoretical perspective.

Plan for rest of the project

Reasoning behind this kind of relation is the foremost part for this project, if the simulation result do confirm the existence of some relation, much work will need to be done Firstly, we need to compare the simulation results for multiple clustering algorithm, and then summarize these relations into rigorous mathematical statement. Secondly, we need to prove this numerical relation using the properties of clustering algorithm and the definition of three quantities.

On the other hand, if these relation dose not exist, there are two directions to go The first is to stick to this approach and select other quantities which could measure the degree of perturbation and changes of partition results. The second is to find way to add other constraints to consistency theorem.

References

[1] J. M. Kleinberg. *An impossibility theorem for clustering.* In Advances in Neural Information Processing Systems, pages 463470, 2002.

[2] VincentCohen-Addad VarunKanade: FrederikMallmann-Trenn; *ClusteringRedemptionBeyond theImpossibilityofKleinbergsAxioms.* In Advances in Neural Information Processing Systems, 2018.