A quantitative approach to Consistency Theorem in Clustering

G14PJA

Mathematics 3rd Year Project

Autumn 2018/19

School of Mathematical Sciences

University of Nottingham

Zehui Li

Supervisor: Dr. Yves van Gennip

Project code: XX P99

Assessment type: Review

I have read and understood the School and University guidelines on plagiarism. I confirm that this work is my own, apart from the acknowledged references.

Abstract

Contents

1 Introduction

Clustering analysis can be defined as a process of segmenting the data points into several subsets, or clusters, with the goal of making the data points within a cluster to be similar to each other, while the data points in distinct clusters to be different. Clustering has been widely used in many fields, such as pattern recognition, bio-informatics and image processing, however, most of the study toward the uniform notion of clustering only stop at the very general level. The algorithms to achieve the clustering task are called clustering algorithm: depending on the definition of the clusters and the way to find the clusters, these clustering algorithms differ from each other significantly. In 2003, Kleinberg [1] published a highly influential paper, in which he set up a general framework to study clustering algorithms as a whole, and proposed three properties that the clustering algorithms could have.

Clustering algorithms fall into three categories [2]: combinatorial algorithms, mixture modelling, and model seeking, the algorithms in each category follows different underlying principal. The advantage of Kleinberg's framework is that it can be applied to all these clustering algorithms regardless of these principal. The core of the framework - three properties proposed by Kleinberg - are called scale invariance, richness and consistency respectively. Scale invariance states that if the distance(dissimilarity) between the data points is multiplied by a positive number, the clustering algorithm should partition the data into the same clusters as before. Richness requires that for any given partition of the data points, it will be possible to come up with a pair wise distance between the data points, so that the clustering algorithm can produce the given partition. Finally, a clustering algorithm satisfy the consistency theorem if we decrease distances between the data points within a cluster, increase the distance between the cluster, the algorithm should produce the same partition. The most important conclusion from Kleinberg's paper is that there is no clustering algorithm which could satisfy three properties at the same time.

Following the impossibility theorem proposed by Kleinberg, 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 (we will explain why this theorem is not sensible in more details in the following section). In this paper, we start from reviewing the work of Kleinberg, adding the missing proof to statements made in his paper, and do simulations on the computer to show the validity of these statements. The rest of the paper will focus on the study of consistency theorem - we come up with a quantitative framework to investigate the consistency property of clustering algorithm. and identify the highly separable distribution of partition results given legitimate perturbation. Finally, we tried several machine learning method to capture the relation between the extend of perturbation and change of partition results. It turns out the methods can capture the model with a very high accuracy and f-measure. At the end of the paper, we try to construct a quantity to measure the extend of perturbation and explore correlation between the the measurement and the change of partition.

1.1 Related work

1.2 Our contribution

2 Review of Kleiberg's work

2.1 Missing Proof

- Work of Kleinberg: three theorems
- Recent Works around Kleinberg's theorems and Refined Consistency theorem
- (Potentially) SVM and Decision Tree model?

2.2 Simulation

- come up with quantitative framework to investigate the consistency theorem of clustering algorithm

- Identify the highly separable distribution of rand index score for clustering results

2.3 Why consistent theorem is flawed?

3 Missing proof from the Kleinberg's Work

4 Framework

4.1 Generate data points by perturbation

4.2 Measurement of partition result

- Rand Index and other measurements

5 Classification with SVM and other methods

6 Conclusions

A Raw data

Material that needs to be included but would distract from the main line of presentation can be put in appendices. Examples of such material are raw data, computing codes and details of calculations.

B Calculations for section ??

In this appendix we verify equation (**??**).

References

[1] Jon M Kleinberg. An impossibility theorem for clustering. In *Advances in neural information processing systems*, pages 463–470, 2003.

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.