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 2002, Kleinberg [1] published a highly influential paper, in which he set up a general framework for clustering algorithms as a whole, and proposed three properties that the clustering algorithms could have. The most important conclusion from this paper is that it is impossible for any clustering algorithm to satisfy three properties at the same time.

Clustering algorithms fall into three categories: combinatorial algorithms, mixture modelling, and model seeking, the algorithms in each category follows different underlying principal. The framework proposed by Kleinberg can be applied to all the clustering algorithm 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 is said to satisfy the consistency theorem if we perturb the data points in the following way: pair wise distances within a cluster are decreased, and pair wise distances between the cluster is increased, the algorithm should produce the same partition.

2 MathBackground

2.1 Related Work

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

2.2 Our contribution

- 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

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

[2] N. L. Alling and N. Greenleaf, *Foundations of the Theory of Klein Surfaces*, Lecture Notes in Mathematics Vol. 219 (Springer, Berlin, 1971).

[3] J. W. Barrett and R. A. Dawe Martins, "Non-commutative geometry and the standard model vacuum", *J. Math. Phys.* **47**, 052305 (2006). (arXiv:hep-th/0601192)

[4] R. Bott and L. W. Tu, *Differential Forms in Algebraic Topology* (Springer, New York, 1982).

[5] B. S. DeWitt, "Quantum theory of gravity. I. The canonical theory", *Phys. Rev.* **160**, 1113–1148 (1967).

[6] pictures D. Giulini, "3-manifolds in canonical quantum gravity", PhD Thesis, University of Cambridge (1990).

[7] A. Hatcher, *Algebraic Topology* (Cambridge University Press, Cambridge, 2002), Proposition 1.40 and Exercise 1.3.24.

[8] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge University Press, Cambridge, 1973).

[9] K. Krasnov and J. Louko, "$\mathrm{SO}_0(1, d+1)$ Racah coefficients: Type I representations", *J. Math. Phys.* **47**, 033513 (2006). (arXiv:math-ph/0502017)

[10] P. Langlois, "Imprints of spacetime topology in the Hawking-Unruh effect", PhD Thesis, University of Nottingham (2005). (arXiv: gr-qc/0510127)

[11] E. Poisson, "The motion of point particles in curved spacetime", *Living Rev. Relativity* **7** 6 (2004), URL : `http://www.livingreviews.org/lrr-2004-6` (cited on 31 August 2006). (arXiv: gr-qc/0306052)

[12] E. Poisson, "The gravitational self-force", in *Proceedings of the 17th International Conference on General Relativity and Gravitation* (Dublin, Ireland, July 18–23,

2004), edited by P. Florides, B. Nolan and A. Ottewill (World Scientific, Singapore, 2005) 119–141. (arXiv:gr-qc/0410127)

[13] J. A. Wheeler, “Geons”, *Phys. Rev.* **97**, 511–536 (1955).

[14] J. A. Wolf, *Spaces of Constant Curvature*, 5th edition (Publish or Perish, Wilmington, 1984).

[15] Website `http://www.ligo.caltech.edu/`, visited 14 August 2007.