A quantitative approach to Consistency Theorem in clustering

Zehui Li (Data Science)
School of Mathematical Science & Computer Science,
Supervisor: Dr. Yves van Gennip
University of Nottingham

Introduction

Various Clustering Algorithms are usually studied independently, however, in 2003, Kleinberg[1] published an influential paper to build system for studying clustering algorithm as a whole. In that paper, he proposed three properties for clustering: Scale-invariance, Richness and Consistency, and prove that no clustering algorithm can satisfy three of them at the same time. In this project, we continue to study this general system for clustering, we start by reviewing the work of Kleinberg's work, then focus our study on the consistency property. This paper mainly has four contributions:

- Provide the proofs for three of the theorems in Kleinberg's Paper
- Describe the potential problem with consistency property
- Show that Clustering Algorithm without Consistency property has "Partial Consistency" under Γ−transformation through simulation.
- Use Support Vector Machine and other Learning Algorithm to show the use case of Partial Consistency

Clustering Algorithm

Clustering analysis can be defined as a process of segmenting the data points into several subsets, the goal of clustering is to make the data within a cluster to be similar (with small dissimilarity) to each other, while the data points in distinct clusters to be different (with large dissimilarity). Figure 1 shows an example of applying the **K-medoid** clustering algorithm to a dataset $\{(x_1, y_1), (x_2, y_2) \dots (x_i, y_i)\}$.

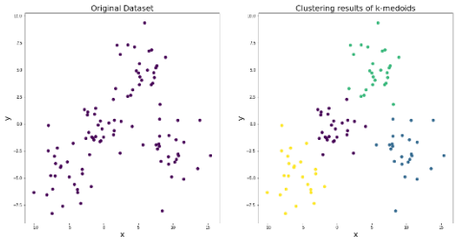

Figure 1: K-medoid on real clustering algorithm

K-medoid will seek to minimize the "**within cluster distance**". Suppose we have k clusters $\{C_1, C_2 \dots C_k\}$ in the final results, then **within cluster distance** w is defined as the following

$$w = \frac{1}{2}\sum_{i=1}^{k}\sum_{n\in c_i}\sum_{m\in C_i}(x_m - x_n)^2 + (y_m - y_n)^2$$

Clustering algorithm fall into three categories[2]: combinatorial algorithms, mixture modelling, and model seeking, the algorithms in each category follows different underlying principal. K-medoid algorithm we discussed above belongs to combinatorial algorithms. One thing good about Kleinberg's framework is that it can be applied to all the clustering algorithms regardless of these principal.

Framework: Scale-Invariance and Richness

In Kleinberg's framework, every clustering algorithm can be represented by a function f, the input of this function are the a set S consisting of n data points, and the pairwise distance among them. Pairwise distance is represented by a distance function d, $d(i,j)$ denote the distance between i and j. f take S and d as input then output a partition Γ, omitting the Set S, we have $\Gamma = f(d)$. Scale-Invariance and Richness is defined in the following way.

Scale-Invariance: f satisfy Scale-Invariance ⇔ For any given distance function d and any $\alpha > 0$, $f(d) = f(\alpha \cdot d)$

Richness: f satisfy Richness ⇔ For any given partition Γ of S, $\exists d$ such that $f(d) = \Gamma$

Framework: Γ-transformation and Consistency

The third property defined in this framework is called consistency. It is more subtle than the first two properties, and to define it, we have to define Γ-transformation. Γ-transformation is a special perturbation manner towards the original Dataset.

Γ-transformation: Given partition $\Gamma = \{C_1, C_2 \dots C_m\}$ on data set S, d' is a Γ-transformation of d ⇔ For any points $i, j \in C_k$, $d'(i,j) \le d(i,j)$; for $i \in C_k, j \notin C_k$ $d'(i,j) \ge d(i,j)$

It may seem very complex at the first glace, but essentially, it is creating a new data set by squashing together the points within the same cluster, and move away points in one cluster from the other one. Figure 2 is an example of legitimate *Γ-transformation.*

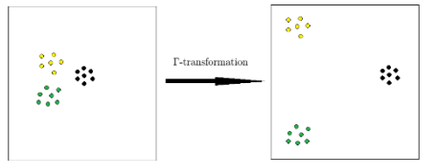

Figure 2 : An example of Γ-transformation.

Then consistency property simply requires that if we apply cluster function f to this new dataset , we can still got the same partition.

Consistency : f satisfy consistency ⇔ Given that d' is of distance function d, $f(d) = f(d')$

These three properties together reflect our expectation to clustering algorithms, although Kleinberg prove that no clustering algorithm can have three properties at the same time, knowing a given clustering algorithm satisfy one or two of these property can still give us much help when using the clustering algorithm.

Rand Index

Rand Index measure the difference between two partition on the same data set. Given a dataset $S = \{1,2 \dots n\}$ containing n points, and two partition of S $\Gamma = \{C_1, C_2 \dots\}$ and $\Gamma' = \{C_1', C_2' \dots\}$, where $C_1, C_2 \dots$ and $C_1', C_2' \dots$ are non-overlap subsets.

- Let $x = |S^*|$, where $|S^*| = \{(i,j) | i,j \in C_l \text{ and } i,j \in C_k'\}$
- Let $y = |S^{**}|$, where $|S^{**}| = \{(i,j) | i \in C_{l_1}, j \in C_{l_2} \text{ and } i \in C_{k_1}', j \in C_{k_2}'\}$

Then Rand Index, Rd is defined as:

$$Rd = \frac{x + y}{\binom{n}{2}}$$

Rand Index (Cont.)

$\binom{n}{2}$ is the total number of possible choices of pair. Rand index can be interpreted as the probability that Γ and Γ' will agree on a randomly chosen pair. The range of Rd is $[0,1]$, when $Rd(\Gamma, \Gamma') = 1$, Γ and Γ' are exactly the same. In the simulation below, we will use normalized Rand Index to decide if two Clusters are the same, and this normalized criteria is called **Adjusted Rand Index (ARI)**[3].

Simulation results: Partial Consistency

Consistency requires that the clustering algorithm have to produce the same partition on all datasets under Γ-transformation. But through the simulation, we find that although some clustering algorithms don't satisfy consistency theorem, it still shows the similar properties. We start from the dataset d, then create multiple datasets through Γ-transformation, what we found is that in many cases, clustering algorithm will produce the same partition on these new datasets. We apply this simulation on two algorithms K-medioids and complete linkage, Figure 3 shows the distribution of ARI of K-medioids, we use the sum of Gaussian to estimate the density function.

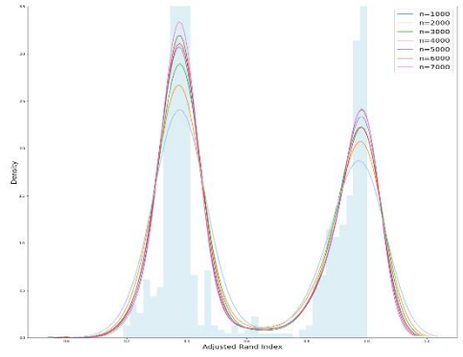

Figure 3: Distribution of ARI for k-medoids

As the number of samples increase, the density function of ARI tend to a fixed function , with the heavy tail at $ARI = 1$. Single linkage algorithm have the similar property, Figure 4 shows the estimated distribution of ARI for Complete-linkage algorithm, but this time, we use Gamma distribution to estimate the density function.

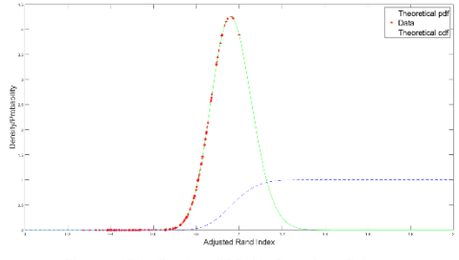

Figure 4: Distribution of ARI for Complete-linkage

We summarize phenomena of skewed distribution as Partial Consistency, and make the assumption that many clustering algorithms will have this Partial Consistency. Due to the limit of space, we don't provide the definition of Partial Consistency, but it is a very loose statement about the behavior of clustering algorithm on these perturbed dataset.

Discussion

We can think of partial consistency as a very weak version of consistency theorem, all the clustering algorithms satisfying original consistency will naturally satisfy partial consistency. The future study could investigate what are the reasons behind this Partial Consistency. There are also other versions of consistency property, one recently published paper [4] focus on this problem as well. In that paper, they propose another consistency theorem called **Refined Consistency**, it basically states that if the Γ-transformation change the "natural number of clusters" of original dataset, the clustering algorithm is still consistent even if it produce different partition on the new dataset. In that case, for example, if Complete-linkage algorithm satisfy Refined Consistency, then we could reasonably deduce that Γ-transformation, in many cases, will not change the "natural number of clusters"

We also try to use learning algorithms to classify the Γ-transformation into two class: the first class will result in the change of clustering results, and the other will keep the structure of the datasets. Without any hyperparameter tuning and feature selection (we just use the perturbed distance function as the feature), we can achieve model with 70% accuracy and recall. Figure 5 shows the accuracy of various model.

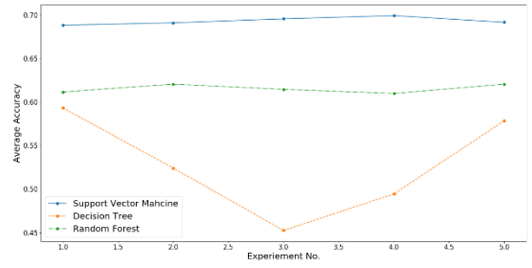

Figure 5: Performance of classifier

Conclusion

In this project, we have a complete review to the work of Kleinberg, then take a quantitative approach to study Γ-transformation and consistency. During the simulation, we make use of ARI to decide if two partition are the same , then we identify the skewed distribution of ARI under Γ-transformation. In the future study, more theoretical analysis is required to confirm the assumptions we made in this paper.

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..

[3] Lawrence Hubert and Phipps Arabie. *Comparing partitions*. Journal of classication, 2(1):193{218, 1985.

[4] Cohen-Addad, V., Kanade, V. and Mallmann-Trenn, F., 2018. *Clustering Redemption–Beyond the Impossibility of Kleinberg's Axioms*. In Advances in Neural Information Processing Systems (pp. 8526-8535).