Improving K-Means Algorithm

Masum Billal

November 4, 2019

Abstract

K-Means is a popular clustering algorithm that reduces sum of minimum distances from data points to centers. In this paper, we aim to improve the accuracy over the so called *k-means++* algorithm. We modify the algorithm in both choosing initial centers and calculating new center to improve the value of *potential*[1] *function*. Experiments show that this improves the potential function for pretty much every dataset, both in terms of accuracy and convergence speed.

1 Introduction

Widely regarded as the most popular clustering techniques, K-Means remains a humble interesting topic in machine learning as well as computational geometry. Roughly the problem is: given a set of points $\mathbb{X}$ in $\mathbb{R}^d$. Find a set of centers $\mathcal{C}$ such that the function (we already used the name potential function)

$$\phi_{\mathcal{C}}(\mathbb{X}) = \sum_{x \in \mathbb{X}} \min_{c \in \mathcal{C}} (||c - x||^2)$$

is minimum. We simply want to minimize $\phi_{\mathcal{C}}(\mathbb{X})$ as much as possible. In short, we can denote this using ϕ alone when it is clear what $\mathcal{C}$ and $\mathbb{X}$ are.

Default K-Means algorithm starts with random centers and then converge based on minimum distances of the centers from the data points. New centers are calculated based on the centroid. This is known as Lloyd's algorithm. This is done until no more change is possible. K-Means++ takes it one step further by choosing the initial centers carefully. Only the first center is chosen at random. Then the rest of the $k - 1$ centers are chosen using D^2 *weighting* as Arthur and Vassilvitskii call it[1]. However, it seems very surprising that no one has worked on choosing the centroid in a better way.

We will first describe K-Means and K-Means++ algorithms. Then we will discuss our ideas and experimental results.

References

[1] David Arthur and Sergei Vassilvitskii. `k-means++`: The Advantages of Careful Seeding. 2006 − 13.

[1] Also known as inertia, which is sum of minimum distances from a particular data point to centers. We follow [1] notations as they are pretty convenient and seem to be mathematically sensible.