

STA663 Term Report Scaleable K-Means++

Shuaiqing Zhang & Tongrong Wang

Department of Biostatistics & Bioinformatics

Abstract

K-means has remained one of the very important unsupervised learning methodology since it is first used by James MacQueen in 1967. A poorly selected initialization of this method would compromise the method's efficiency. However, with the approaching of the era of big data, the recently proposed initialization algorithm, K-means++, becomes overly computationally intensive because of its inherent sequential nature. A revised version of it called K-meansII was addressed to improve efficiency in both sequential and parallel settings for large-scaled data. The superiority is demonstrated by a simulation written below.

Keywords: K-means, Initialization, Large-scaled Data

1. Introduction

1.1. Introduce k-means and k-means++

1.2. Parallel version of k-means++

2. Algorithm

2.1. Algorithm 1

2.2. Algorithm 2

2.3. Parallel implementation

3. Experimentence

3.1. Data

3.2. Baseline

4. Result

4.1. Computational cost

4.2. Running time

4.3. Trading-off between quality and running time

5. Full analysis of the algorithm

6. Reference