STA663 Term Report Scaleable K-Means++

Shuaiqing Zhang & Tongrong Wang

Department of Biostatistics & Bioinformatics

Abstract

K-means has remained one of the very import unsupervised learning methodology since it is first used by James MacQueen in 1967. A poorly selected initialization of this method would compromised the method efficient. However, with the approaching of the are of big data, the recently proposed initialization algorithm, K-means++, become overly computational intensive because of its inherent sequential nature. An revised version of it called K-meansII were addressed to improve the efficient 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