Big Data, Big Problems–Parallelizing Kmeans

Reuben K. McCreanor
Department of Statistics
Duke University
Durham, NC 27708
reuben.mccreanor@duke.edu

Wei (Emily) Shao
Department of Statistics
Duke University
Durham, NC 27708
wei.shao@duke.edu

Abstract

K-means is a data-partitioning algorithm that separates n observations into k partitions. Due to its simplicity of implementation, interpretability of its results, and its ability to categorize data based on desired features, it remains one of the most popular algorithms in fields of statistics and machine learning. However, the key issue with k-means comes from its efficiency and scalability. As the size of datasets continues to increase, implementing k-means on data of any magnitude becomes computationally infeasible. A recently proposed variation of k-means, k-means++, provides a robust method of selecting the initial centers, essentially giving an optimal solution. However, due to the number of passes over the data required in a naive implementation, even clusters a million data points into 100 partitions would be exceedingly slow. In order to combat this, a parallelized version of k-means++ is proposed, k-means $||$. This version uses a sampling factor ℓ to dramatically reduce the number of passes while still arriving at an equivalent solution of partitions. This paper will implement k-means++ and k-means$||$ in both sequential and parallel setting and compare the results both in terms of efficiency and equivalency of partition arrangements.

References

Bahmani, Bahman, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. *Scalable K-means.* Proc. VLDB Endow. Proceedings of the VLDB Endowment 5.7 (2012): 622-33. Web.

7 Appendix