Title: Optimization of the K-means Clustering Algorithm through Initialized Principal Direction Divisive Partitioning

Author: Bruce James

Mentor: Dr. Jacob Kogan

Abstract: Clustering is an invaluable procedure to the unsupervised task of data analysis, with a cursory list of applications including genomics, bioinformatics, and signal processing. In the setting of text mining, the inherent irregularity and size of the data place demands on traditional methods. To handle this complexity, high-dimensional vector spaces, equipped with some distance-like function, have become the standard environment for large document sets, and typically, documents are converted into very sparse vectors in a finite dimensional space, remanding the resulting collection of salient features to an algorithm of one's choice, such as the classic k-means clustering algorithm. Due to various sizes of the feature space, different algorithms offer a trade-off between accuracy and computational efficiency, and work is done to reduce the dimensionality of the feature vectors. This study investigates the linear algebra and optimization techniques, which underwrite the Principal Direction Divisive Partitioning (PDDP) algorithm, described as a top-down hierarchical technique, to be used as a plug-in to the k-means algorithm. K-means reliance on initial random partitioning builds computational cost into document analysis, detracting from its stand-alone utility. Using a PDDP initialized partition to seed k-means, computational efficiency and the resulting clusters will be compared to the k-means procedure without PDDP.

Keywords: cluster analysis, data mining, text mining, k-means, principle directive divisive partitioning, hierarchical divisive clustering