1. Introduction

- This chapter makes an introduction to clustering and the K-means algorithm, and their necessity
- There has been a steep increase in available data which requires a computationally efficient method that can work in an unsupervised fashion
- K-means is the most popular clustering method, despite being so old, due to its simplicity, performance and proven results

2. Improvements

- This section contain recent research results into specific aspects of the K-means algorithm

2.1. Execution time optimization

- Enhanced K-means is a simple improvement, targeted for datasets with a large amount of clusters; it aims to reduce the number of times distance is calculated from a point to all means by keeping an additional data structure
- MapReduce is a framework for splitting calculations over a cluster of commodity hardware, and provides the best solution for scaling the algorithm
- GPU's are special circuits that can run simple operations on thousands of threads, leading to the fastest implementation of K-means

2.2. Seed selection

- Seed selection is a weak point of the classic algorithm due to outliers and local optima
- A simple heuristic is to sort the dataset by distance to origin, split into k portions and choose the middle point from each
- K-d trees are a data structure that helps organize multidimensional data; a density estimation can be made over initial data for a more robust initialization method
- ROBIN chooses initial means based on the Local Outlier Factor

3. Variants

- This chapter introduces some different approaches for the K-means algorithm
- Bisecting K-means is an algorithm that applies K-means with 2 clusters at each step, bisecting the dataset
- Genetic Algorithms are efficient heuristic search methods that have been proven to deal well with local optima
- Two variants of genetic K-means are presented, one which also finds the optimal number for k
- GA K-means obtains good performance but may not scale so well

4. Conclusions

- K-means remains a popular clustering method with over 30k publications on Google Scholar each year since 2010
- The popularity is well deserved, as the algorithm is a useful clustering tool that can deal with almost any dataset