Review Proposal: K-means Clustering Over Time

Jeffery B. Russell
Fourth Year Computer Science Student at RIT
CUBRC Research Assistant
RITlug President

With the ubiquity of data in today's age, machine learning has become a driving force in research and innovation in both the public and private sectors. Due to the vast quantity of data generated by the internet, unsupervised learning has been at the forefront of data science over the past few decades. Clustering has been and will continue to be an essential way of extracting data from large sets of information.

This project aims to break down and analyze two papers that use k-means clustering. In 1967 J. MacQueen wrote the historical article that introduced [1] the original k-means clustering. The paper that originally coined the term k-means [1] is going to be reviewed because it provides useful historical insight into where this field of research started and what influenced it. Since the paper was written in 1967, this opens an interesting area of investigation because we can compare a relatively older research paper on clustering with a new one. This paper can be found on the Project Euclid website [1].

The second paper was written by John Paparrizos and Luis Gravano in 2016 [2] and covers how we can cluster time series data using an algorithm they call k-shape, which was inspired by the k-means algorithm. The paper can be found on the University of Colombia's website[2]. The K-shape paper is of special interest to me because I used during my last co-op while researching Space Situational Awareness (SSA).

Daniel Moore is going to be the peer reviewer for this literature review. Although both papers are over ten pages, they are appropriate for this assignment due to their significance in the field and my prior knowledge of the topic.

References

[1] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967)., 1967.

[2] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. *ACM SIGMOD Record*, 45:69–76, 06 2016.

[1] https://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992

[2] http://web2.cs.columbia.edu/~gravano/Papers/2015/sigmod2015.pdf