Similarities and differences between K-Means, PAM, and Clara

Joao Cláudio Macosso (422575)

1. Introduction

K-means, PAM (Partitioning Around Medoids), and CLARA (Clustering Large Applications) are popular unsupervised learning algorithms used for clustering analysis. Despite the similarities between these methods, they do also have some dissimilarities. In this article present similarities and differences between these methods focusing on approach to clustering, computational cost, and appropriate use cases.

2. Similarities and differences

All the three algorithms are unsupervised learning algorithms that do not require target variable in the dataset, In addition to that, they are all partitional clustering algorithms, which means that they assign each data point to a single cluster. Furthermore, they are all iterative algorithms that aim to minimize a defined objective function that measures the similarity between data points in a cluster and the dissimilarity between different clusters.

Similarities and differences

2.1. Construction of clusters

In K-means each cluster is represented by its centroid (mean) of the data points assigned to it. The algorithm assigns each data point to the closest centroid and updates the centroids until there is convergence.

On the other hand, with PAM each cluster is represented by a medoid, which is the most centrally located data point in a cluster. At the initial stage the algorithm assigns random data points as medoids, and then iteratively swaps a medoid with a non-medoid point to improve the clustering.

To reduce the computational cost of PAM, the data can be split into multiple samples and run PAM on each sample, which is known as CLARA algorithm which the extension of PAM algorithm with less computational cost as it works on a sample of the data rather than the full dataset.

2.2. Use cases

K-means is more suitable for datasets with a moderate number of clusters and well-separated data points. PAM on the other hand is more appropriate for datasets with noisy or high-dimensional data and where the number of clusters is not obvious. However, when faced with large dataset that might not be feasible to run PAM algorithm on the entire dataset, CLARA algorithm should be considered.

2.3. Efficiency

K-means can easily converge quickly even on large datasets. However, the optimal convergence is not guaranteed. On the other Hand, PAM is slower than k-means due to the distance calculation, but it is more robust to outliers and can produce better results when faced with noisy data. Furthermore, CLARA, can be slower than PAM as it needs to generate multiple samples of the data, but it can handle large datasets and can produce more reliable results than k-means or PAM.

2.4. Computational cost

K-means has the lowest computational cost when compared to PAM and CLARA. However, it can get stuck in local optima, which might not be the global optima. The algorithm scales linearly with the number of data points and the number of clusters. PAM on the other hand, has a higher computational cost than k-means, because it computes distances for each data points and the medoids for each iteration. While CLARA has the highest computational cost between these three algorithms, because it requires running on multiple samples.

3. Summary

To sum up, despite similarities between K-means, PAM, and CLARA, there are also significant differences between them, more specifically when it comes to how clusters are constructed, computational cost, efficiency, and appropriate use cases. K-means is fast and suitable for datasets with a moderate number of well-separated groups within the dataset, while PAM and CLARA are slower but more robust to datasets with a lot of noise. Furthermore, PAM is appropriate for noisy or high-dimensional datasets where the number of clusters is not known in advance, while CLARA is more suitable for large datasets where robust clustering results are desired.