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

Joao Cláudio Macosso (422575)

1. Intrroduction

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.

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.

1.1. 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.

1.2. Computational cost:

K-means: It has a lower computational cost than PAM and CLARA but can get stuck in local optima. The algorithm scales linearly with the number of data points and the number of clusters.

PAM: It has a higher computational cost than k-means, as it requires calculating distances between each data point and the medoids in each iteration. However, it can be more robust to outliers than k-means.

CLARA: It has the highest computational cost among the three algorithms, as it generates multiple samples of the data and runs PAM on each sample. However, it can handle large datasets and can produce more reliable results than k-means or PAM.

1.3. Efficiency:

K-means: It can converge quickly, even on large datasets, but the convergence is not guaranteed to be optimal.

PAM: It is slower than k-means due to the distance calculation, but it is more robust to outliers and can produce better results.

CLARA: It can be slower than PAM due to generating multiple samples of the data, but it can handle large datasets and can produce more reliable results than k-means or PAM.

1.4. Appropriate use cases:

K-means: It is suitable for datasets with a moderate number of clusters and well-separated data points.

PAM: It is suitable for datasets with noisy or high-dimensional data and where the number of clusters is not known in advance.

CLARA: It is suitable for large datasets where running PAM on the entire dataset is not feasible and where robust clustering results are desired.

2. Conclusion

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 snoise. 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.