Qeios

Peer Review

Review of: "Fixed-Sized Clusters k-Means"

Farid Saberi-Movahed[1]

1. Graduate University of Advanced Technology, Kerman, Iran

This paper introduces Fixed-Sized Clusters k-Means, an adaptation of the standard k-Means algorithm designed to produce clusters of predetermined sizes. The approach modifies the traditional assignment step by using the Hungarian algorithm to assign data points to fixed cluster slots, thereby ensuring balanced clusters while still minimizing the mean square error. The paper discusses the algorithm's convergence properties and computational complexity, and it demonstrates the method's practical application through a seating plan example, showing that the algorithm can efficiently handle datasets with up to around 5000 points.

I believe this paper will be of value to readers who are keen on dimensionality reduction methods and the application of clustering problems. The paper presents a unique adaptation of the k-Means algorithm that incorporates fixed cluster sizes. This innovation allows for balanced clustering, which is particularly useful in applications where equal or near-equal distribution among clusters is required. However, the paper has several issues that must be addressed by the authors.

1- How does the computational complexity of the Fixed-Sized Clusters k-Means algorithm compare with that of standard k-Means and other balanced clustering methods, especially when scaling beyond 5000 data points?

2- The method relies on the Hungarian algorithm for the assignment phase. Could the authors discuss potential limitations or performance bottlenecks of using the Hungarian algorithm, particularly in the context of larger or more complex datasets?

3- The paper mentions a convergence result by relating it to constrained k-Means. How robust and reliable is the convergence of the proposed algorithm in practice, and are there specific cases or data distributions where convergence might be slow or problematic?

doi.org/10.32388/XBHPT8

4- The algorithm updates the edge weights dynamically based on the current centroids. How does this dynamic update affect the overall performance and stability of the clustering, and are there scenarios where it might lead to suboptimal assignments?

5- In the seating plan application, the algorithm was tested with a specific dataset (22 persons). Can the authors provide further experimental validation or comparative analysis against other seating arrangement approaches to illustrate the effectiveness and versatility of their method?

6- Since the algorithm tackles a two-objective optimization problem (minimizing MSE and balancing cluster sizes), how do the authors address potential trade-offs between these conflicting goals? Is there any sensitivity to tuning parameters that control the balance between minimizing the mean square error and ensuring equal cluster sizes?

Declarations

Potential competing interests: No potential competing interests to declare.