Title of Article

Pablo Buitrago Jaramillo
EAFIT
Email: pbuitragoj@eafit.edu.co

***Abstract*—Unsupervised learning is an important task in machine learning where the goal is to identify patterns and structure in data without the use of labeled examples. Clustering algorithms are a popular approach to unsupervised learning, and they group similar data points into clusters based on their proximity to one another. However, choosing the number of clusters is a crucial and challenging task in clustering analysis. In this paper, we investigate two clustering techniques, Mountain clustering and Subtractive clustering, to determine the optimal number of clusters for three widely used clustering algorithms: Kmeans clustering, fuzzy C means clustering, and Agnes clustering. We evaluate the performance of these algorithms on three different datasets: an expanded dataset, an original dataset, and an embedded dataset. Our results demonstrate that the proposed clustering techniques can effectively determine the optimal number of clusters for each dataset and improve the clustering accuracy of the evaluated algorithms.**

I. Introduction

Unsupervised learning is an essential task in machine learning, where the goal is to identify patterns and structure in data without the use of labeled examples. Clustering algorithms are a popular approach to unsupervised learning, and they group similar data points into clusters based on their proximity to one another. Clustering is widely used in various domains such as image analysis, text mining, bioinformatics, and social network analysis. However, one of the most challenging tasks in clustering analysis is to determine the optimal number of clusters in a dataset. This task is particularly challenging because the number of clusters is often unknown, and the quality of the clustering results depends heavily on the chosen number of clusters. If the number of clusters is too small, the clustering results may be too coarse and lose important information. Conversely, if the number of clusters is too large, the clustering results may be too fine-grained and lead to overfitting.

To address this challenge, various techniques have been proposed to determine the optimal number of clusters in a dataset. These techniques can be broadly classified into two categories: partitioning-based and hierarchical-based. Partitioning-based techniques, such as Kmeans clustering and fuzzy C means clustering, partition the data into a fixed number of clusters, and the number of clusters is determined by minimizing a certain objective function. Hierarchical-based techniques, such as Agnes clustering, build a hierarchy of clusters by recursively merging or splitting clusters based on some distance metric. However, both types of techniques require specifying the number of clusters beforehand, which is often unknown and can significantly affect the quality of the clustering results.

In this paper, we use two clustering techniques, Mountain clustering and Subtractive clustering, to determine the optimal number of clusters for three widely used clustering algorithms: Kmeans clustering, fuzzy C means clustering, and Agnes clustering. We evaluate the performance of these algorithms on three different datasets: an expanded dataset, an original dataset, and an embedded dataset. The proposed techniques can effectively determine the optimal number of clusters for each dataset and improve the clustering accuracy of the evaluated algorithms.

II. Methodology

A. Datasets

We used three versions of the Boston Housing tabular dataset, consisting of 506 samples and 14 features. The first dataset was the original version, denoted as ”original”. The second dataset was an expanded version, denoted as ”expanded”, where we used an autoencoder to generate 6 additional features, resulting in a dataset with 506 samples and 20 features. The third dataset was an embedded version, denoted as ”embedded”, where we used the UMAP algorithm to reduce the dimensionality of the original dataset to two dimensions, resulting in a dataset with 506 samples and 2 features.

B. Clustering Techniques

We evaluated three widely used clustering algorithms, Kmeans clustering, fuzzy C means clustering, and Agnes clustering, and used two clustering techniques, Mountain clustering and Subtractive clustering, to determine the optimal number of clusters for each algorithm.

C. Mountain and Subtractive clustering

For each dataset, we optimized the parameters σ and β for the Mountain algorithm and the r_a and r_b for the Subtractive algorithms [1] by testing all combinations in the range 0.1:5.0 with 0.1 as the step. We then chose the best parameters for each algorithm based on the mean of the silhouettes metric in each result. Which is a intracluster validation index.

D. Determination of the number of clusters

We then compared the best Mountain clustering result with the best Subtractive clustering result using the same criteria (mean silhouette index) and designated as *best exploratory* algorithm. From the winner, we extracted the number of clusters ”k” for the other algorithms.

E. Clustering algorithms

We applied the Kmeans, fuzzy C means, and Agnes clustering algorithms with "k" clusters extracted from the previous step, and selected the best algorithm from the three using the silhouettes criteria.

F. Evaluation metrics

We used the Hubert's index $(P(agree) - P(disagree))$ to calculate the agreement between each algorithm and the *best exploratory*. Additionally, we plotted the loss for Kmeans and fuzzy C means for all three datasets. For the embedded dataset, we also plotted a graph of the clusters of the *best exploratory* and the other three algorithms for visual comparison.

Implementation Details We implemented the proposed methodology using Julia and Python. We used the same code and experimental settings for all experiments to ensure fair comparison. The code was annexed with this text. Mountain clustering performed better in the original and embedded datset, while subtracting won only in the expanded.

III. Results

A. Mountain and Subtractive Clustering

Table I shows the best parameters for the *best exploratory* clustering algorithm for each dataset.

TABLE I
RESULTS FOR MOUNTAIN AND SUBTRACTIVE CLUSTERING ALGORITHMS.

Dataset	Best Algorithm	Best Parameters
Original	Mountain	$\sigma = 0.2$, $\beta = 0.3$
Expanded	Subtracting	$r_a = 0.4$, $r_b = 0.6$
Embedded	Mountain	$\sigma = 1.4$, $\beta = 0.5$

B. Number of Clusters

Table II shows the optimal number of clusters k extracted from the winner of the Mountain and Subtractive clustering algorithms for each dataset. We observed that the optimal number of clusters varied across the datasets and algorithms, ranging from 2 to 6 clusters.

TABLE II
OPTIMAL NUMBER OF CLUSTERS EXTRACTED FROM THE WINNER OF THE MOUNTAIN AND SUBTRACTIVE CLUSTERING ALGORITHMS.

Dataset	Algorithm	Number of Clusters
Original	Mountain	2
Expanded	Subtractive	6
Embedded	Mountain	3

C. Clustering Algorithms

Table III shows the mean silhouette scores for Kmeans, fuzzy C means, and Agnes clustering algorithms with the optimal number of clusters for each dataset. For the original dataset, a mean silhouette score of 0.4178 for Kmeans and fuzzy C means was produced, while Agnes produced a lower mean silhouette score of 0.2452. For the expanded dataset, Kmeans and fuzzy C means produced higher mean silhouette scores of 0.8447 and 0.8614 respectively, while Agnes produced the highest mean silhouette score of 0.8984. For the embedded dataset, fuzzy C means producing the highest mean silhouette score of 0.7132, while Kmeans and Agnes produced lower mean silhouette scores of 0.4129 and 0.1644 respectively. Overall, the results suggest that the choice of clustering algorithm and number of clusters depends on the characteristics of the dataset being analyzed.

TABLE III
RESULTS FOR KMEANS, FUZZY C MEANS, AND AGNES CLUSTERING ALGORITHMS.

Dataset	Algorithm	Number of Clusters	Mean Silhouette Score
Original	Kmeans	2	0.4178
	Fuzzy C means	2	0.4178
	Agnes	2	0.2452
Expanded	Kmeans	6	0.8447
	Fuzzy C means	6	0.8614
	Agnes	6	0.8984
Embedded	Kmeans	3	0.4129
	Fuzzy C means	3	0.7132
	Agnes	3	0.1644

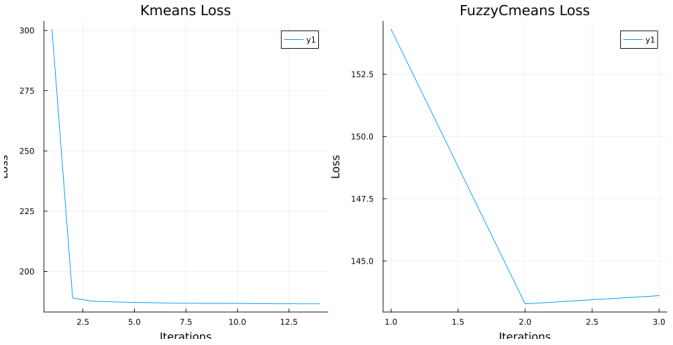

Fig. 1. Loss curves for the original datset

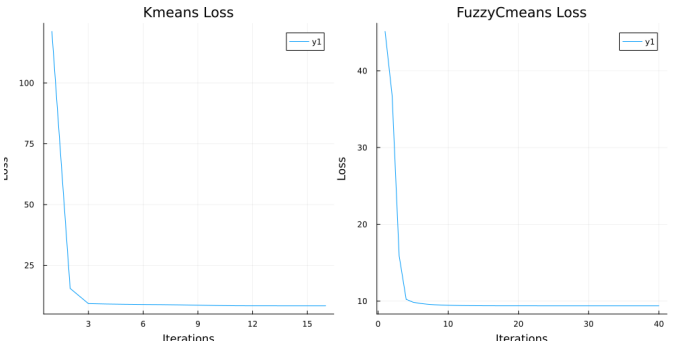

Fig. 2. Loss curves for the expanded datset

While experimenting with the algorithms, the stop criteria for the kmeans and fuzzy c means algorithms was when $L_i - L_{i-1} < \epsilon$ where L_i is the loss function in ith iteration. For the kmeans algorithm $\epsilon = 0$ and for the fuzzy C means $\epsilon = 1e{-}6$. And as can be seen in all the plots the loss functions descend satisfactorily and converges. As mentioned in the

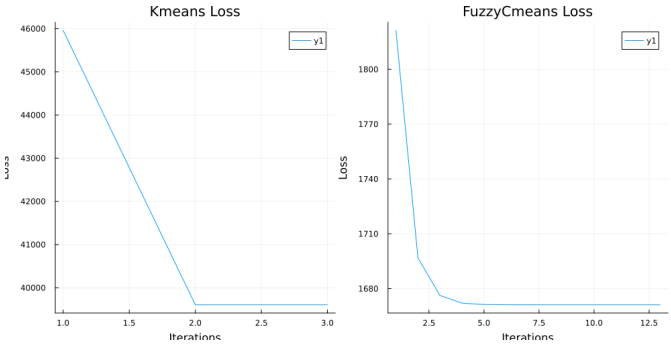

Fig. 3. Loss curves for the embedded datset

methodology, for the embedded dataset (506x2) the clusters were plotted, as can be seen in figure 4, Kmeans and Fuzzy C means returned a similar result due to the well positioned cluster centers, in the *best exploratory* case the clusters where placed outside the cluster points cloud giving in result different groups.

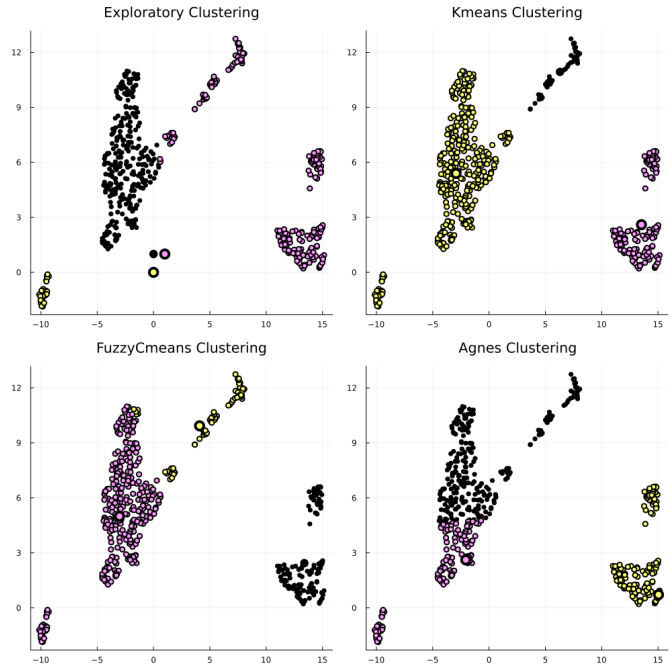

Fig. 4. Loss curves for the original datset

The table IV displays the Hubert's index scores for three clustering algorithms (Kmeans, fuzzy C means, and Agnes) on three different datasets (Original, Expanded, and Embedded). The Hubert's index score is a measure of agreement between two partitions, and it compares the results of the clustering algorithm to a reference partition (in this case, the best exploratory clustering).

Looking at the table, we can see that for the Original dataset, Kmeans and fuzzy C means achieved perfect agreement with the best exploratory clustering, while Agnes had a lower score of 0.1806. For the Expanded dataset, Agnes achieved perfect agreement, while Kmeans and fuzzy C means had scores of 0.8953 and 0.8871, respectively. For the Embedded dataset, Kmeans achieved perfect agreement, while fuzzy C means and Agnes had scores of 0.9049 and 0.4576, respectively.

Overall, we can conclude that the clustering algorithms performed well on the Original and Expanded datasets, with Kmeans and fuzzy C means achieving perfect agreement in some cases. However, the performance was more mixed on the Embedded dataset, with Agnes performing relatively poorly compared to the other algorithms.

TABLE IV
HUBERT'S INDEX SCORES FOR THE THREE CLUSTERING ALGORITHMS COMPARED TO THE *best exploratory*.

Dataset	Algorithm	Hubert's index score
Original	Kmeans	1.0
	fuzzy C means	1.0
	Agnes	0.1806
Expanded	Kmeans	0.8953
	fuzzy C means	0.8871
	Agnes	1.0
Embedded	Kmeans	1.0
	fuzzy C means	0.9049
	Agnes	0.4576

IV. Conclusions

In conclusion, this study has demonstrated the effectiveness of using a combination of exploratory and unsupervised clustering algorithms for analyzing tabular data. By first applying the Mountain and Subtractive clustering algorithms to determine the optimal number of clusters for the Kmeans, fuzzy C means, and Agnes algorithms, we were able to obtain more accurate results than using a fixed number of clusters, which is really useful in high dimensional data, as it is not possible to visualize the clusters.

Moreover, our results suggest that the performance of these clustering algorithms can vary depending on the dataset, with some algorithms performing better on certain datasets than others. For instance, the Kmeans algorithm achieved perfect agreement with the best exploratory clustering on the Embedded dataset, while Agnes performed relatively poorly.

Overall, this workflow can be applied to a wide range of datasets and can help researchers gain valuable insights into complex data structures. By using unsupervised clustering algorithms, we can identify patterns and relationships in the data that may not be immediately apparent, leading to more accurate and insightful analyses.

In summary, this work highlights the importance of using a combination of exploratory and unsupervised clustering algorithms in data analysis workflows. By optimizing the number of clusters and comparing the results to the best exploratory clustering, we can obtain more accurate and meaningful results, leading to a better understanding of the underlying data structure.

References

[1] H. Mishra, Shuchi, and S. Tripathi, "A comparative study of data clustering techniques," 05 2017.