Universidad Autónoma de Querétaro

Facultad de Ingeniería

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

October 6, 2023

Author:*Lic. Ijtsi Dzaya Ramos Morales*

Machine Learning , Dr. Marco Antonio Aceves Fernández

Report 3: Clustering Algorithms

Abstract

This study investigates the performance of K-Means and CLARANS clustering algorithms on a dataset for hepatitis classification. The analysis reveals insights into the alignment between cluster assignments and dataset labels. Findings underscore the importance of parameter tuning to enhance clustering accuracy in both methods.

Introduction

The application of clustering algorithms assumes a critical role in unveiling hidden structures and patterns. This report delves into the utilization of two prominent clustering methodologies, K-means and CLARANS, within the domain of hepatitis detection—a pivotal area in healthcare. While addressing the inherent complexities of the dataset, including missing data, categorical attributes, and disparate numerical scales, the primary focus remains on the exploration of these clustering techniques. By leveraging these algorithms, the aim is to enhance the dataset's utility for advanced medical research and to facilitate more accurate hepatitis detection methods.

Theoretical Foundation

Clustering is an unsupervised machine learning technique that groups similar data points together. It is a widely used technique in data mining, exploratory data analysis, and machine learning.

K-means Clustering

K-means is a widely used clustering algorithm that partitions data into distinct groups, or clusters, based on their similarity.

K-means is highly efficient and works well when clusters are spherical and equally sized. However, it is sensitive to the initial placement of centroids and may converge to local optima.

The elbow method is a technique for determining the optimal number of clusters (K) in the K-means clustering algorithm. It involves calculating the Sum of Squared Distances (SSD) from data points to their assigned centroids.

As K increases, SSD decreases, indicating that data points are closer to their centroids. The elbow method identifies an "elbow point" where the rate of SSD decrease changes significantly. This point represents the ideal K value, balancing clustering complexity and quality.

CLARANS Clustering

Contrasting with K-means, CLARANS (Clustering Large Applications based on Randomized Search) is a robust and scalable clustering algorithm. It employs a randomized search technique to explore the cluster space efficiently. Key characteristics of CLARANS include:

- Medoids: Unlike K-means, which uses centroids, CLARANS employs medoids as cluster representatives. Medoids are actual data points within the cluster, making CLARANS more robust to outliers.
- Robustness: CLARANS can handle noisy data and is less likely to converge to local optima.
- Scalability: CLARANS is well-suited for large datasets, as it avoids the need for computing pairwise distances between all data points.

Materials and methods

1. Data set: A database was obtained from the Kaggle platform, sourced from the UCI Machine Learning Repository. This database encompasses attributes pertinent to the diagnosis of hepatitis C, contributing to a comprehensive understanding of the disease's diagnostic factors.

   The dataset comprises a total of 615 observations, encompassing 14 attributes. These attributes encapsulate laboratory values and de-

mographic information derived from both blood donors and patients diagnosed with hepatitis C. The data is characterized by its numerical nature, except for the "Category" and "Sex" attributes.

The dataset's context underscores its focus on laboratory values, demographics, and diagnostic categorization. The data set originates from the UCI Machine Learning Repository.

2. Software: Python, a versatile programming language, was employed to fulfill the objectives of this project. The analysis involved the utilization of the `numpy`, `pandas`, `matplotlib`,`sklearn`, `seaborn` and `random` libraries, renowned for their capabilities in numerical computations, data manipulation, and graphical analysis.

3. Methods:

   - K-means: It begins by randomly initializing cluster centroids. In the subsequent step, each data point is assigned to the nearest centroid, creating initial clusters. The centroids are then recalculated as the mean of all data points within each cluster. These two steps, assignment, and centroid recalculation, iteratively repeat until convergence. During each iteration, data points may change clusters as centroids shift, leading to a more refined clustering configuration. This process continues until the centroids stabilize, indicating that the clusters have reached a stable configuration or a maximum amount of iterations is reached.

   - Clarans: It initiates by randomly selecting a data point as a medoid within a cluster. Then, employing a randomized search strategy, it evaluates various data points as potential medoids, aiming to improve cluster quality. If a superior medoid candidate is discovered, it replaces the current one. This process iterates, refining cluster arrangements by seeking more compact and well-defined clusters until convergence is achieved.

Results

The results presented herein shed light on various aspects of the analysis.

Figure 1 shows the final result of the database preprocessing. It contains no null values, no categorical attributes and normalized attributes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 615 entries, 0 to 614
Data columns (total 13 columns):
 #    Column   Non-Null Count  Dtype
---   ------   --------------  -----
 0    Age      615 non-null    float64
 1    ALB      615 non-null    float64
 2    ALP      615 non-null    float64
 3    ALT      615 non-null    float64
 4    AST      615 non-null    float64
 5    BIL      615 non-null    float64
 6    CHE      615 non-null    float64
 7    CHOL     615 non-null    float64
 8    CREA     615 non-null    float64
 9    GGT      615 non-null    float64
 10   PROT     615 non-null    float64
 11   Sex_f    615 non-null    float64
 12   Sex_m    615 non-null    float64
dtypes: float64(13)
memory usage: 62.6 KB
```

Figure 1: Dataset after preprocessing

Figure 2 presents the result of the elbow method after being applied to the k-means algorithm that was created. It was tested with k values from 2 to 20.

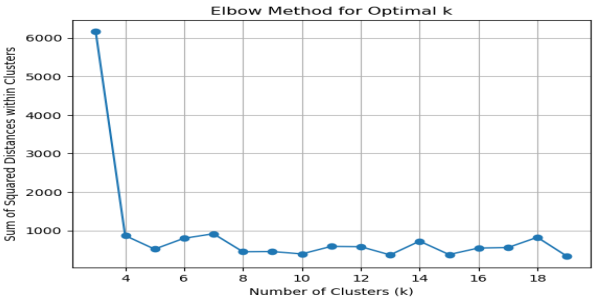

Figure 2: Elbow method results for K-means

Figure 3 displays the ARI for the clusters' results vs the original labels for the k-means algorithm results.

```
Adjusted Rad Score
-0.06589633816510787
```

Figure 3: Adjusted Rad Score for K-means

Figure 4 presents the result of the elbow method after being applied to the clarans algorithm that was created. It was tested with k values from 2 to 20.

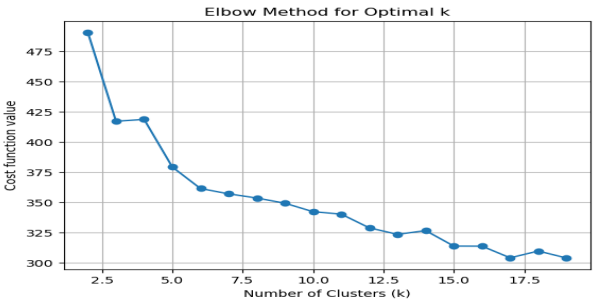

Figure 4: Elbow method results for Clarans

Figure 3 displays the ARI for the clusters' results vs the original labels for the k-means algorithm results.

```
Adjusted Rad Score
0.01313010858108128
```

Figure 5: Adjusted Rad Score for Clarans

Discussion

In this section, the results of the clustering algorithms' analysis are discussed.

K-Means

Based on the analysis presented in Figure 2, it is observed that the optimal number of clusters, as determined by the elbow method, appears to be either 4 or 5. However, to assess the degree of similarity between the clusters formed by the K-means algorithm and the original labels in the dataset, the Adjusted Rand Score was calculated. The result yielded a negative value, indicating that the clustering performed by the algorithm does not align well with the original labels. Several factors may contribute to this discrepancy. Firstly, the simplicity of the K-means algorithm, which halts iteration after a predefined number of steps, can lead to suboptimal cluster assignments. Additionally, the initialization of centroids is sensitive, potentially resulting in centroids that are closely positioned, further affecting clustering quality. Further exploration and refinement of the clustering approach may be required to achieve a closer alignment between the algorithm's results and the true data structure.

CLARANS

Figure 4 presents the results of applying the CLARANS clustering algorithm to the dataset. Unlike the K-means method, the elbow method does not reveal a distinct optimal value for k. Nevertheless, based on a visual approximation, a value of k equal to 6 was chosen for clustering. However, an important evaluation metric to consider is the Adjusted Rand Index (ARI), which quantifies the similarity between the clusters generated by the algorithm and the original dataset labels. In the case of CLARANS with k=6, the ARI score was found to be 0.01, which indicates a low level of alignment between the clusters and the true labels. Several factors may contribute to this observed discrepancy. CLARANS, known for its robustness and scalability, may uncover different underlying patterns that are not necessarily tied to the provided labels. The choice of parameters, such as the number of iterations and the number of nearest neighbors, can significantly influence the algorithm's performance. Further exploration and fine-tuning of these parameters may be necessary to enhance the alignment between the resulting clusters and the original dataset labels

Conclusion

The results found led to a better understanding of the performance of K-Means and CLARANS algorithms. K-Means' simplicity and sensitivity to initialization hindered accurate clustering, while CLARANS' ability to uncover diverse patterns and sensitivity to parameter choices impacted its alignment with dataset labels. These findings underscore the importance of parameter fine-tuning to enhance clustering accuracy in both methods, allowing for a more precise capture of the underlying data structure and alignment with true labels.

References

[1] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651-666, 2010. (IEEE)

[2] C. C. Aggarwal and C. K. Reddy, *Data clustering: Algorithms and applications*. Chapman and Hall/CRC, 2014. (Book)

[3] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*, pp. 25-71, Springer, 2006. (Book Chapter)

[4] M. A. Aceves Fernández, "Inteligencia Artificial para Programadores con Prisa," 2023.

[5] R. Lichtinghagen, F. Klawonn, and G. Hoffmann, "HCV data," 2020. DOI: `https://doi.org/10.24432/C5D612`.