Work 4

Carlos Jiménez, Sheena Lang, Zachary Parent, and Kacper Poniatowski

December 10, 2024

1 Introduction

2 Data

This section describes the datasets used in our study and details the preprocessing steps applied to prepare them for analysis.

3 Dataset Selection and Characteristics

This section outlines our dataset selection criteria and analyzes the characteristics of the chosen datasets for evaluating clustering algorithms.

3.1 Dataset Selection

To evaluate clustering algorithms, we selected three datasets with diverse characteristics in size, attribute types, class distribution, and missing data patterns. This diversity allows for a comprehensive analysis of algorithm performance under varying conditions.

The datasets vary in size, including a small dataset (Hepatitis), a medium-sized dataset (Vowel), and a large dataset (Mushroom), enabling an assessment of scalability and computational efficiency. They also include diverse attribute types: nominal, numerical, and mixed, ensuring the algorithms are tested across different data representations.

Class distribution was another consideration. Although clustering is unsupervised, class distributions serve as baselines for validating cluster quality. We included a balanced dataset (Vowel), a slightly imbalanced one (Mushroom), and a heavily imbalanced one (Hepatitis). Lastly, missing data patterns were addressed, with Hepatitis containing missing values to test robustness, while Vowel and Mushroom datasets are complete, allowing performance comparisons under ideal conditions.

3.2 Dataset Characteristics

The Hepatitis dataset includes 155 instances with a mix of 13 nominal and 6 numerical attributes. It represents a binary classification problem (survive vs. die), but the classes are imbalanced, with the majority class comprising 79.35% of the instances. Approximately 6.01% of the values are missing.

The Mushroom dataset consists of 8,124 instances with 22 nominal attributes. It represents a binary classification problem (edible vs. poisonous) with nearly balanced classes and contains no missing values.

The Vowel dataset contains 990 instances with 3 nominal and 10 numerical attributes. It represents a multi-class classification problem with 11 balanced classes and no missing values.

3.3 Selection Criteria

We selected the Hepatitis, Mushroom, and Vowel datasets to provide complementary characteristics for evaluating clustering algorithms. Our selection criteria focused on:

- **Dataset size variation** (small, medium, large)
- **Attribute type diversity** (nominal, numerical, mixed)

3.4 Data Preprocessing

This section outlines the preprocessing steps applied to prepare the three datasets (*Hepatitis*, *Mushroom*, and *Vowel*) for clustering analysis.

We began by replacing all missing values, denoted as `?`, with `NaN` to facilitate appropriate imputation techniques. The class label column, which was required only for post-clustering evaluation, was temporarily removed during preprocessing. To account for the varying characteristics of the datasets, we differentiated between categorical and numerical features using a heuristic-based approach. Columns were classified as categorical if their data type was `object` or if the proportion of unique values was below 5% of the dataset's size. All other columns were treated as numerical.

Numerical features were processed by imputing missing values with the column mean to preserve the overall data distribution, followed by rescaling to the [0, 1] range using Min-Max scaling. This scaling ensured that all features contributed equally during distance-based clustering. Categorical features were handled by imputing missing values with the most frequent category (mode), thus preserving the dominant data patterns. Binary categorical features were encoded using label encoding, while non-binary categorical features were one-hot encoded, creating separate columns for each category. This encoding strategy was selected based on recommendations for clustering algorithms like K-Means, where numerical representations avoid introducing arbitrary ordinal relationships.

After processing, the numerical and categorical features were concatenated to form the final dataset, and the class labels were label-encoded for use in evaluation metrics such as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). The preprocessing pipeline involved imputing missing values (mean for numerical and mode for categorical features), applying Min-Max scaling to numerical features, and using label encoding (binary categories) and one-hot encoding (non-binary categories) for categorical features. This approach ensures that the preprocessing pipeline is tailored to the unique requirements of clustering algorithms, while maintaining data integrity and minimizing potential biases.

4 Methods

5 Results and Analysis

Full results, in tabular and graphical form, are available in the appendix (Section 7).

6 Conclusion

Clustering is a core machine learning technique used in many applications, but choosing the right clustering algorithm and hyperparameters is not always straightforward. In this report, we have evaluated several clustering algorithms and their hyperparameters on three distinct datasets, exploring the strengths and weaknesses of each method. Here we present our key findings and practical implications of the results.

6.1 Key Findings

6.2 Practical Implications

6.3 Limitations and Future Work

6.4 Final Remarks

Each clustering algorithm demonstrates distinct strengths and weaknesses, suggesting that the choice of algorithm should be primarily driven by specific application requirements and data characteristics. While K-means and Fuzzy C-means offer good general-purpose solutions, specialized algorithms like OPTICS and Spectral clustering provide advantages for specific data distributions. Future work should focus on developing hybrid approaches that can combine the strengths of multiple algorithms while mitigating their individual weaknesses.

7 Appendix