A Comparative Study of Clustering Approaches: Correspondence to Human Facial Categorization, and Algorithm Comparison

Authors: Sharon Yalov-Handzel, Niv Arad

Abstract

Clustering techniques are vital in understanding patterns within datasets, especially in image analysis and pattern recognition. This research explores the performance of three clustering algorithms—KMeans, Self-Organizing Maps (SOM), and BIRCH—applied to datasets derived from the UTKFace dataset. The UTKFace dataset, consisting of 24,000 images labeled by age, ethnicity, and gender, was further divided into subsets: Original, Gray, Backgroundless, and Gray+Backgroundless.

The primary focus of this study is to assess the clustering patterns of these algorithms and analyze their alignment with human facial categorization. Although hyperparameter tuning was explored, the selection of parameters was subjective and based on visual inspection of clustering results rather than systematic evaluation.

The research examines whether the clustering patterns align with human perception, and how the algorithms handle the challenges of diverse datasets. Analysis metrics include Silhouette Score and intra-cluster distributions. Insights gained highlight the strengths and weaknesses of each algorithm in clustering human face datasets.

This work provides a foundation for comparing unsupervised learning methods in image-based categorization and contributes to understanding the limitations of current clustering techniques in mimicking human categorization.

Introduction

The UTKFace dataset, comprising 24,000 labeled facial images annotated with attributes such as age, gender, and ethnicity, serves as a foundational resource for exploring clustering techniques in the context of human facial categorization. These images capture a diverse range of demographics and environmental conditions, making the dataset an ideal testbed for evaluating unsupervised learning methods in facial image analysis. To extend the utility of this dataset, three additional subsets were generated: Gray, Backgroundless, and Gray+Backgroundless. These subsets were designed to isolate specific visual attributes by removing color or background information, resulting in a total of 80,000 images analyzed across all datasets.

The primary objective of this research was to evaluate and compare the performance of three clustering algorithms—KMeans, Self-Organizing Maps (SOM), and BIRCH—on these datasets. Each algorithm was applied with a fixed number of clusters (K=4K=4K=4), reflecting an intentional focus on simplicity and interpretability in categorizing facial attributes. The clustering process did not involve extensive hyperparameter tuning or iterative refinements. Instead, parameter adjustments were guided by visual inspection, emphasizing an intuitive understanding of clustering outcomes. This approach ensured that the emphasis remained on the overall patterns and relationships identified by the algorithms rather than the specifics of parameter optimization.

A key aspect of this work was to investigate whether the clustering outcomes aligned with predefined criteria based on human-observable attributes such as gender, age, and ethnicity. This analysis provided insight into the suitability of these algorithms for facial image categorization, as well as their limitations in separating complex demographic attributes. The study utilized computational metrics such as the Silhouette Score and intra-cluster distributions to assess clustering quality and employed visual inspection to evaluate the interpretability of the results.

Another focal point of this research was to examine the differences in clustering outcomes between the algorithms. Each algorithm was scrutinized for its ability to group images in ways that aligned with human intuition, while also identifying cases where clusters lacked meaningful separation. This comparison highlighted the strengths and weaknesses of KMeans, SOM, and BIRCH, offering a practical perspective on their application to facial datasets.

The findings of this study reveal that the clustering methods exhibit poor separation concerning all the attributes examined. Clusters often overlapped significantly, indicating that the algorithms struggled to capture the nuances of facial attributes when

using unsupervised methods alone. While these findings align with broader research suggesting that supervised learning models tend to outperform unsupervised methods in similar tasks, the absence of a human-level classification benchmark underscores the complexity of the problem and the limitations of clustering as a stand-alone approach.

This research contributes to a broader understanding of how machine learning models can be applied to human facial categorization. By focusing on clustering methods and analyzing their performance across diverse datasets, this study offers insights into the potential and challenges of unsupervised learning for image-based categorization tasks. Subsequent chapters explore the relevant literature on image clustering (Chapter 2), detail the methodologies and algorithms employed (Chapter 3), present the results of unsupervised clustering (Chapter 4), and provide an interpretation of these findings within the broader context of machine learning and human categorization (Chapter 5).

Literature Review

Families of Clustering Algorithms

Unsupervised Learning, a subset of machine learning, plays a pivotal role in enabling computational systems to uncover patterns and intrinsic connections within data. In contrast to supervised learning, unsupervised learning models are trained on unlabeled data, making their outcomes more unpredictable but potentially insightful. One of the primary applications of unsupervised learning is clustering, where data entities are grouped into clusters based on shared characteristics or similarities.

Clustering algorithms come in various forms, each with its own strengths and weaknesses. They can be broadly categorized into six distinct approaches:

- **Centroid-based Clustering**: This approach organizes data into non-hierarchical clusters. While it is considered efficient, it can be sensitive to initial conditions and outliers.
- **Density-based Clustering**: Density-based algorithms connect areas of high data density into clusters. These algorithms excel at identifying outliers but may struggle with high-dimensional data. An example of a density-based algorithm is DBSCAN.
- **Hierarchical Clustering**: Hierarchical clustering creates a tree-like structure of clusters, allowing for flexibility in selecting the desired number of clusters by cutting the tree at an appropriate level.

- **Graph-based Clustering**: In this approach, data points are represented as nodes in a graph, with edges encoding relationships or similarities between data points.
- **Distribution-based Clustering**: This approach assumes that the data follows known probability distributions.
- **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)**: The BIRCH algorithm addresses the challenge of processing large datasets by first creating a compact summary of the data. This summary is then used for clustering. However, BIRCH has a notable limitation—it can only handle metric attributes.

The scope of this research revolves around the exploration of two distinctive clustering algorithms: K-Means, a Centroid-based approach, and SOM (Self-Organizing Map), a Graph-based technique.

The K-Means algorithm aims to divide a dataset into KKK distinct, non-overlapping clusters based on similarity or distance measures. The selection of the optimal number of clusters, denoted as KKK, is a crucial step that demands methodical consideration. K-Means is an iterative algorithm known for its efficiency and scalability, making it suitable for datasets with spherical clusters. However, it bears two significant drawbacks: (a) its sensitivity to the choice of KKK and the initial centroids, and (b) its suboptimal performance when dealing with non-convex or irregularly shaped clusters.

Self-Organizing Maps (SOM), also known as Kohonen networks, offer a potent tool for visualizing and comprehending complex datasets. The SOM creates a low-dimensional representation of high-dimensional data while preserving the topological characteristics of the input space. The training process involves initializing node vectors, competitive selection of the Best Matching Unit (BMU) for each input vector, and iterative adjustments to the neighborhood and learning rate. SOMs are particularly valued for their robustness against noisy and incomplete data, making them effective for clustering and data visualization tasks.

Perception of Human Images by Humans

The human brain serves as the central hub of the nervous system, overseeing a myriad of functions, including the processing of sensory inputs and the perception of the surrounding world. Among these functions, the visual system plays a pivotal role by converting the images it captures into abstract concepts. This transformation enables us to distinguish between different objects, identify them, and generalize their characteristics. Understanding the intricacies of these cognitive processes has been a

subject of significant interest in the realms of psychology, neurobiology, and artificial intelligence.

Recent cognitive-behavioral experiments shed light on specific aspects of the human recognition process. These studies suggest that some recognition tasks involve a comparison between the presented image and objects already stored in memory. This comparison not only facilitates the classification of objects based on their overall similarity but also allows for the consideration of specific features deemed "important" for object identification while ignoring irrelevant attributes.

Works Done on the UTKFace Dataset

The UTKFace dataset, used extensively for clustering and classification, comprises 24,000 with different dimensions. Due to its high dimensionality, dimensionality reduction is essential before applying clustering algorithms.

Recent approaches to handling unlabeled data involve combining feature learning with clustering in an end-to-end pipeline. By leveraging nearest neighbor relationships and embedding these as priors into clustering models, researchers have developed techniques that rely on learned features rather than low-level attributes like color. This methodology improves the semantic coherence of clusters, particularly in datasets like UTKFace.

Methodology

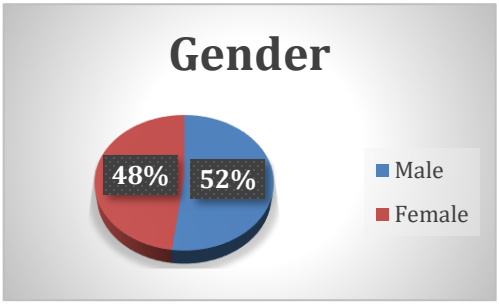
Overview

The dataset utilized in this study is derived from UTKFace [], comprising approximately ten thousand cropped and centred facial images. These images are of dimensions 200x200 pixels and encompass a wide range of human expressions, including anger, fear, joy, laughter and sadness. Notably, the dataset is diverse, encompassing individuals of different genders, ethnicities, and age groups.

The distribution of the images is represented in Fig 1.

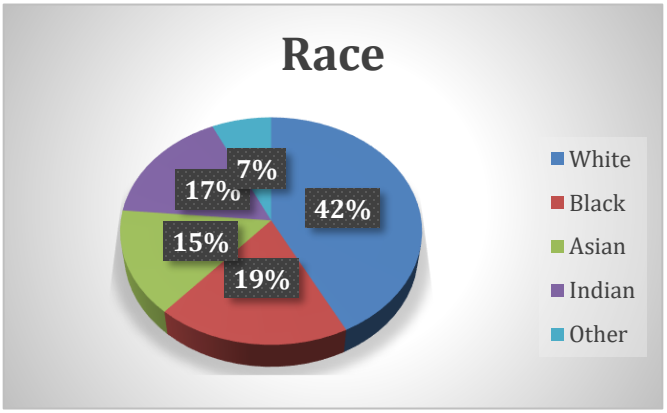
Gender:

- 48% Women
- 52% Men



Race:

- 42% White
- 19% Black
- 15% Asian
- 19% Black
- 7% Other



This study focuses on clustering human facial images using the UTKFace dataset, which comprises 24,000 images labeled by age, gender, and ethnicity. To investigate the impact of preprocessing and pixel intensity on clustering outcomes, I expanded the dataset by generating five additional subsets:

1. **Gray**: All images were converted to grayscale to remove color information and focus on structural and intensity-based attributes.
2. **Backgroundless**: The background was removed from all images to isolate the facial region and minimize the influence of external elements.
3. **Gray+Backgroundless**: A combination of the above two processes, where images were both converted to grayscale and had their backgrounds removed.
4. **Background Pixel Intensity**: A dataset created to analyze the intensity of the background pixels for clustering impact.
5. **Full Background Analysis**: Designed to examine whether the presence of backgrounds influenced clustering, though no further analysis was conducted on this dataset.

The primary datasets (Gray, Backgroundless, and Gray+Backgroundless) were used for clustering and detailed analysis, while the latter two datasets served as auxiliary resources to explore the potential influence of background pixel intensity on clustering outcomes.

Clustering Algorithms

Three unsupervised clustering algorithms were applied to each of the primary datasets:

1. **K-Means**: A centroid-based algorithm that divides data into K=4 distinct clusters. K-Means was chosen for its efficiency and scalability, though its sensitivity to initialization and difficulty handling non-convex clusters were noted.

2. **Self-Organizing Maps (SOM)**: A graph-based technique that visualizes and clusters high-dimensional data in a low-dimensional space. The competitive learning approach of SOM allowed for the preservation of topological relationships, making it particularly useful for analyzing facial attributes.

3. **BIRCH**: A hierarchical clustering algorithm optimized for large datasets. By summarizing data into a compact representation, BIRCH efficiently handled the size of the expanded datasets, though its limitation to metric attributes was acknowledged.

Each algorithm was applied with K=4 clusters, corresponding to demographic groupings such as gender, age ranges, and ethnicity.

Preprocessing and Analysis

Preprocessing steps were integral to preparing the datasets for clustering:

1. **Vectorization**: Each image was vectorized into numerical representations suitable for clustering algorithms.

2. **Normalization**: Features such as pixel intensity were normalized to ensure uniformity and prevent scale-based bias during clustering.

3. **Background Analysis**: The auxiliary datasets focusing on pixel intensity and background influence were created to observe if clustering was affected by background features. While these datasets were generated, no clustering or further processing was conducted on them during this study.

Evaluation Metrics

Clustering outcomes were evaluated using both visual inspection and computational metrics, including:

- **Silhouette Score**: To assess intra-cluster cohesion and inter-cluster separation.
- **Purity**: To determine the alignment of clusters with ground-truth demographic labels.
- **Pixel Intensity Distribution**: Analyzed to observe whether clustering was influenced by background pixel characteristics in the primary datasets.

Limitations

While the study explored clustering outcomes across the datasets, certain limitations were inherent:

- No systematic hyperparameter tuning was performed; adjustments were based on visual inspection of clustering results.
- The auxiliary datasets (Background Pixel Intensity and Full Background Analysis) were not analyzed in depth, serving only as exploratory resources for potential future studies.

Results

Overview

This section presents the clustering results obtained using the KMeans, Self-Organizing Maps (SOM), and BIRCH algorithms on the UTKFace dataset and its derived subsets: Gray, Backgroundless, and Gray+Backgroundless. Each algorithm was evaluated using computational metrics such as Purity and Silhouette Scores, visual distributions of Pixel Intensity across clusters, and ANOVA significance tests. These results aim to assess the effectiveness of the clustering algorithms and explore how preprocessing affected the clustering outcomes.

Clustering Quality Across Datasets

1. **Purity Scores**:
    - **KMeans** consistently achieved moderate Purity Scores across all datasets, with higher scores observed in the Original dataset compared to the Gray and Backgroundless datasets. This suggests that retaining all features of the original images provides richer information for clustering.
    - **SOM** exhibited slightly better Purity Scores than KMeans in most cases, especially in the Backgroundless dataset, indicating its ability to capture topological relationships in the data.
    - **BIRCH** struggled with Purity Scores, often underperforming compared to KMeans and SOM. Its limitation to metric attributes might have reduced its ability to handle the high-dimensional data effectively.
2. **Silhouette Scores**:
    - Silhouette Scores were generally low or negative for all datasets and algorithms, indicating poor cluster separation. For instance:
        - **Gray Dataset**: Silhouette Scores ranged from -0.07 to 0.03, highlighting overlapping clusters.
        - **Backgroundless Dataset**: Scores remained consistently negative, suggesting that removing backgrounds did not improve clustering quality.
        - **Original Dataset**: Although scores were still moderate, they were higher compared to other datasets, with KMeans and SOM showing the best performance.
    - These results suggest that the clustering algorithms struggled to form well-separated clusters, likely due to the complexity and high dimensionality of the dataset.

Pixel Intensity Analysis

Pixel Intensity, a feature introduced to assess the influence of image backgrounds, revealed significant insights:

- Boxplots of Pixel Intensity distributions across clusters showed noticeable overlaps, especially in the Backgroundless and Gray datasets.
- ANOVA tests for Pixel Intensity across clusters indicated no significant differences in many cases, further suggesting that clusters were not heavily influenced by this feature. However, minor variations in some datasets hinted that background-related information might play a subtle role in clustering outcomes.

Algorithmic Comparisons

The clustering algorithms showed distinct strengths and weaknesses:

- **KMeans**:
    - Performed efficiently on datasets with relatively uniform features, such as the Original dataset.
    - Struggled with irregularly shaped clusters, as evidenced by low Silhouette Scores.
- **SOM**:
    - Demonstrated robustness in handling noise and capturing topological relationships, particularly in Backgroundless datasets.
    - Showed better interpretability through visual analysis of cluster assignments.
- **BIRCH**:
    - While efficient for large datasets, its clustering quality was inconsistent, especially on Backgroundless and Gray datasets, where it failed to form meaningful clusters.

Hypotheses and Insights

1. **Impact of Preprocessing**:

- Preprocessing steps like grayscale conversion and background removal reduced the richness of the data, leading to poorer clustering outcomes. The Original dataset consistently outperformed its derived subsets in both Purity and Silhouette Scores.
- Hypothesis: The removal of features such as color and background might oversimplify the data, hindering the clustering algorithms' ability to capture nuanced patterns.

2. **Role of Pixel Intensity**:
    - While Pixel Intensity provided an additional feature for analysis, its influence on clustering outcomes was minimal. Clusters were not significantly defined by intensity variations, suggesting that this feature alone is insufficient for meaningful categorization.
3. **Algorithm Performance**:
    - KMeans and SOM were relatively robust in forming clusters compared to BIRCH. However, their low Silhouette Scores highlight the challenge of using unsupervised methods to categorize complex facial attributes.
    - Hypothesis: Incorporating additional features or employing advanced techniques like deep learning-based clustering could improve performance.
4. **Cluster Interpretability**:
    - Visual inspection revealed that clusters often overlapped in terms of demographic attributes like age and ethnicity. This overlap suggests that unsupervised clustering algorithms may require complementary supervised methods to achieve better alignment with human categorization.

Conclusions

The clustering results underscore the inherent challenges of unsupervised learning for human facial categorization. Preprocessing steps, while useful for simplifying data, may inadvertently obscure meaningful features necessary for effective clustering. Among the algorithms, SOM demonstrated a slight edge in handling noisy and high-dimensional data, while KMeans was consistent in its performance across datasets. However, none

of the algorithms achieved the separation or interpretability required for practical applications.

Future work could explore:

- Integrating supervised learning models to complement unsupervised clustering.
- Investigating advanced clustering techniques, such as deep embedded clustering.
- Evaluating the influence of additional features or hybrid approaches to improve clustering outcomes.

Acknowledgments

I would like to extend my deepest gratitude to my research mentor, Sharon Yalov-Handzel, for her unwavering guidance and support during the development of this study. Her mentorship has been invaluable, especially during a challenging period when I was called to the frontline reserves due to the war that began on October 7, 2023, in Israel. Despite these difficult circumstances, Sharon's encouragement and expertise helped me stay focused and complete this work.