Fair Labeled Clustering

A Preprint

Mirza Abbas Uddin
United International University
United City, Madani Ave, Dhaka 1212
muddin201315@bscse.uiu.ac.bd

Sanjana Hossain Sonali
United International University
United City, Madani Ave, Dhaka 1212
ssonali201423@bscse.uiu.ac.bd

Irtesam Mahmud Khan*
Lecturer, Dept. of CSE
United International University
United City, Madani Ave, Dhaka 1212
mahmud@cse.uiu.ac.bd

January 1, 2024

1 Introduction

The advent of machine learning algorithms in critical decision-making processes has a significant interest in developing fair variants of these algorithms. This paper addresses the expanding field of fair clustering, where recent efforts have focused on achieving group fairness – ensuring proportional group representation within each cluster. The primary focus has been on proportional group representation within clusters. However, our work extends this perspective by examining the downstream application of clustering results.

In many practical scenarios, a decision-maker utilizes a clustering algorithm, examines each cluster's center, and assigns an appropriate outcome or label to the cluster. For instance, in hiring processes, clusters may represent groups of job applicants, and the decision-maker assigns outcomes such as 'hire' or 'reject' to each cluster. Our approach aims to ensure group fairness by maintaining proportional group representation within each label rather than insisting on it within every cluster, as in traditional group fair clustering.

We introduce algorithms tailored to this problem, demonstrating their efficiency compared to NP-hard counterparts in group fair clustering. Additionally, we explore an alternative scenario where the decision-maker can freely assign labels to clusters regardless of their positions in the metric space. This setting exhibits intriguing transitions from computationally challenging to manageable based on additional constraints.

To validate the efficacy of our algorithms, we conduct experiments on real-world datasets, providing empirical evidence of their effectiveness in addressing fairness concerns in clustering applications.

2 Literature Review

The authors (1) proposed a new definition of fairness at the label level, which is more suitable for decision making applications. They consider two settings: **labeled clustering with assigned labels (LCAL)** and **labeled clustering with unassigned labels (LCUL)**. For the LCAL setting, they showed that the problem can be solved in polynomial time for a constant number of labels. For the special case of two labels, they developed a faster algorithm with running time O(n(log n + k)). For the LCUL setting, they gave a detailed characterization of the hardness under different constraints and show a randomized algorithm that always achieves an optimal clustering and satisfies the fairness constraints in expectation for a natural specific form of constraints. They also conducted experiments on real-world datasets that show the effectiveness of their algorithms. In particular, they showed that their algorithms provided fairness at a lower cost than fair clustering and that they indeed scaled to large datasets. The authors conducted experiments on two real-world datasets to evaluate the performance of their algorithms: Adult and CreditCard. They used the following metrics:

- **Price of Fairness (PoF)**: The ratio of the cost of the fair solution to the cost of the color-blind solution.
- **Proportional violation**: The smallest relaxation of the fairness constraints needed to make the solution feasible.

The authors compared their algorithms to the following baselines:

- **Nearest center**: Assigns each point to its closest center.
- **Fair clustering**: Solves the fair clustering problem.

LCAL experiments: The authors first conducted experiments on the LCAL setting, where the center labels are assigned based on their position. They found that their algorithm outperformed fair clustering on both PoF and proportional violations. For example, on the Adult dataset, their algorithm achieved a PoF of 1.0059 (0.59%), while fair clustering achieved a PoF of 1.7 (70%). The authors also showed that their algorithm is more scalable than fair clustering. They found that their algorithm was able to solve the LCAL problem for the Adult dataset in just a few seconds, while fair clustering took over an hour to solve the problem.
LCUL experiments: The authors then conducted experiments on the LCUL setting, where the center labels are free to be selected. They compared their algorithm to two baselines:

- **Nearest center with random assignment (NCRA)**: Assigns each point to its closest center with a random label assignment.
- **Fair clustering (FC)**: Assigns each point to a center in a way that satisfies the fairness constraints.

The authors found that their algorithm outperformed both NCRA and FC on proportional violations. For example, on the CreditCard dataset, the average proportional violation for their algorithm was 0.02, while the average proportional violation for NCRA and FC were 0.15 and 0.06, respectively. The authors also showed that their algorithm is more scalable than fair clustering in the LCUL setting. They found that their algorithm was able to solve the LCUL problem for the CreditCard dataset in just a few minutes, while fair clustering took over a day to solve the problem. Overall, the authors' experiments showed that their algorithms outperform other fair clustering algorithms on PoF, proportional violations, and scalability. Additionally, the authors evaluated the scalability of their fair clustering algorithms on the Census1990 dataset and found that their algorithms are highly scalable, even for large datasets. For example, it took their algorithms less than 90 seconds to solve the LCAL/LCUL problems for 500,000 data points. In contrast, the fair clustering algorithm of would take around 30 minutes to solve a similar problem on the same dataset. Their scalability results have important implications for the deployment of fair machine learning algorithms in real-world applications.

3 Data Preparation

3.1 Dataset Description

The Adult dataset (2), also known as the "Census Income" dataset, aims to predict whether an individual's income exceeds $50K/year based on census data. It is a multivariate dataset in the social science domain, primarily used for classification tasks. The dataset comprises categorical and integer features, totaling 14 attributes with 48,842 instances. Table 1 summarised some of the key features.

Feature	**Type**	**Description**
age	Integer	Represents the age of individuals.
workclass	Categorical	Indicates the income source, including private, self-employed, government employment, etc.
fnlwgt	Integer	Represents the final weight, a census-adjusted weight.
education	Categorical	Denotes the education level, ranging from preschool to doctorate.
education-num	Integer	Represents the numerical encoding of education levels.
marital-status	Categorical	Specifies the marital status of individuals.
occupation	Categorical	Represents the type of occupation, with some missing values.
relationship	Categorical	Describes the relationship status.
race	Categorical	Represents the race or ethnicity of individuals.
sex	Binary	Denotes the gender as female or male.

Table 1: Description of Features in the Dataset

3.2 Data Cleaning and Preprocessing

In this dataset exploration, a comprehensive analysis of the dataset and its basic statistical characteristics was conducted to lay the groundwork for subsequent knowledge discovery processes. Initially, a check for missing values was performed, revealing discrepancies primarily in the 'Workclass' and 'Occupation' columns. To address this, rows with missing values were selectively removed to ensure data integrity. The subsequent step involved transforming categorical variables into a suitable format for analysis.

Several columns, such as 'Workclass', 'MaritalStatus', 'Occupation', 'Relationship', 'Race', 'Sex', and 'NativeCountry', underwent one-hot encoding to facilitate meaningful insights during the analysis. The 'Education' column, representing ordinal data, was systematically mapped to numerical values for a more cohesive representation. Additionally, whitespaces were uniformly removed from all entries in the dataset to enhance consistency.

Challenges were encountered during the handling of missing values, particularly in deciding whether to impute or remove them. In this case, a strategic decision was made to remove rows with missing values, given the specific context of the data and the potential impact on subsequent analyses.

This meticulous dataset preparation not only addressed missing values but also transformed categorical and ordinal variables into a format conducive to meaningful analysis. The challenges faced were navigated with careful consideration of the dataset's characteristics, ensuring the integrity of the data for knowledge discovery endeavors.

4 Knowledge Discovery

In this section, the application of clustering algorithms for the dataset is presented. The objective is to group similar instances together based on certain features, providing insights into inherent patterns within the data.

- **Data Standardization:** To ensure uniformity in the scale of numerical features, standardization was applied to columns such as Age, FinalWeight, Education, EducationNum, CapitalGain, HoursPerWeek, and Income. The `StandardScaler` from scikit-learn was utilized for this task.
- **K-Means Clustering:** K-Means clustering, a widely-used algorithm for partitioning data into distinct groups, was applied to the standardized numerical columns. The dataset was segmented into two clusters (`num_clusters = 2`). The `KMeans` implementation from scikit-learn with 10 initializations was employed.
- **K-Modes Clustering:** For categorical columns like Education and Income, K-Modes clustering was employed. This algorithm is suitable for clustering categorical data. The dataset was divided into two clusters using the `KModes` implementation from the `kmodes` library.
- **K-Prototypes Clustering:** To handle mixed data types (numerical and categorical), K-Prototypes clustering was utilized. The algorithm was applied to the entire dataset with a focus on specific columns (Education, EducationNum, CapitalGain, HoursPerWeek) designated as categorical. The `KPrototypes` implementation from the `kmodes` library was utilized for this purpose.
- **Discussion:** The clustering process revealed distinctive patterns in the data. K-Means clustering primarily considers numerical features, providing insights into the distribution of individuals based on age, education, and income. On the other hand, K-Modes and K-Prototypes clustering consider categorical information, shedding light on patterns related to education levels and income categories.

  The clusters generated by these algorithms can be further analyzed to understand the characteristics of each group. It is essential to interpret the results in the context of the dataset's domain and the specific features used for clustering.

  The clustering models were trained with a straightforward configuration to maintain transparency. Further refinement and tuning can be explored based on the specific objectives and requirements of the analysis.

5 Results and Comparison

5.1 Paper (1) Results

In their experiments, the authors evaluated the performance of their algorithms using commodity hardware with Python 3.6, NumPy, and the Scikit-learn library. They applied their algorithms to a collection of datasets from the UCI repository, including the Adult dataset. The evaluation involved setting upper and lower proportion bounds for fairness constraints, measuring the price of fairness (PoF), and calculating proportional violations for color proportionality constraints. For the LCAL experiments on the Adult dataset, the authors compared their labeled fair clustering algorithm

(LCAL) to fair clustering and an unfair baseline. They observed that their algorithm achieved a much smaller PoF and was more robust to variations in the number of clusters.

For the LCUL experiments, the authors compared their labeled fair clustering algorithm (LFC) to two baselines: Nearest Center with Random Assignment (NCRA) and Fair Clustering (FC). The evaluation considered the average values of proportional violations for color, points/label, and center/label over 50 runs.

5.2 Our Results

In contrast, our experiments focused on evaluating clustering quality using the silhouette score. The silhouette score measures how well-defined clusters are within the data. For the K-Means, KModes, and KPrototypes clustering algorithms, you obtained silhouette scores of 0.2529, -0.0377, and 0.3522, respectively.

5.3 Comparison

While the evaluation metrics differ between the paper and our results, it is notable that silhouette scores indicate the quality of clustering. The positive silhouette score for K-Means and KPrototypes suggests reasonably well-defined clusters, whereas the negative score for KModes may indicate overlapping clusters. However, it's essential to consider the context and goals of our analysis when interpreting these scores.

In conclusion, the paper's focus on fairness constraints and proportional violations provides insights into the trade-offs between fairness and clustering quality. our evaluation using silhouette scores complements this perspective by assessing the inherent cluster structure in the data.

6 Conclusion

In this study, we explored the performance of clustering algorithms with a focus on fairness and silhouette-based evaluation. The comparison between the paper's results and our evaluation using silhouette scores provides a comprehensive view of the trade-offs inherent in clustering tasks. The paper's (1) emphasis on fairness constraints and proportional violations sheds light on the challenges of achieving equitable clustering outcomes. Their labeled fair clustering algorithm demonstrated improved fairness and robustness compared to alternative approaches, as evidenced by a smaller price of fairness (PoF) in LCAL experiments. In our evaluation, silhouette scores revealed insights into the inherent cluster structures produced by K-Means, KModes, and KPrototypes. While silhouette scores are not inherently linked to fairness metrics, they provide a valuable perspective on the quality and separability of clusters within the data. In conclusion, the study presents a nuanced understanding of clustering algorithms, considering both fairness constraints and traditional clustering quality measures. Future work may explore hybrid approaches that integrate fairness-aware clustering with silhouette-based evaluations to achieve well-defined and equitable clusters.

Appendix

The source code, implementation details, and additional materials for this project can be found in the GitHub repository: https://github.com/mirzaaa101/Data-Mining-Comparative-Study/blob/main/code.ipynb.
Feel free to explore the repository for a more in-depth understanding of the algorithms, datasets, and experiments conducted in this study.

References

[1] S. A. Esmaeili, S. Duppala, J. P. Dickerson, and B. Brubach, "Fair labeled clustering," p. 327–335, 2022.

[2] B. Becker and R. Kohavi, "Adult," 1996. DOI: https://doi.org/10.24432/C5XW20.