Fair-Capacitated Clustering

Tai Le Quy
Leibniz University Hannover
Hannover, Germany
tai@l3s.de

Arjun Roy
Leibniz University Hannover
Hannover, Germany
roy@l3s.de

Gunnar Friege
Leibniz University Hannover
Hannover, Germany
friege@idmp.uni-hannover.de

Eirini Ntoutsi
Freie Universität Berlin, Germany
Berlin, Germany
eirini.ntoutsi@fu-berlin.de

ABSTRACT

Traditionally, clustering algorithms focus on partitioning the data into groups of similar instances. The similarity objective, however, is not sufficient in applications where a *fair-representation* of the groups in terms of protected attributes like gender or race, is required for each cluster. Moreover, in many applications, to make the clusters useful for the end-user, a *balanced cardinality* among the clusters is required. Our motivation comes from the education domain where studies indicate that students might learn better in diverse student groups and of course groups of similar cardinality are more practical e.g., for group assignments. To this end, we introduce the *fair-capacitated clustering problem* that partitions the data into clusters of similar instances while ensuring cluster fairness and balancing cluster cardinalities. We propose a two-step solution to the problem: i) we rely on fairlets to generate minimal sets that satisfy the fair constraint and ii) we propose two approaches, namely hierarchical clustering and partitioning-based clustering, to obtain the fair-capacitated clustering. Our experiments on three educational datasets show that our approaches deliver well-balanced clusters in terms of both fairness and cardinality while maintaining a good clustering quality.

Keywords

fair-capacitated clustering, fair clustering, capacitated clustering, fairness, learning analytics, fairlets, knapsack.

1. INTRODUCTION

Machine learning (ML) plays a crucial role in decision-making in almost all areas of our lives, including areas of high societal impact, like healthcare and education. Our work's motivation comes from the education domain where ML-based decision-making has been used in a wide variety of tasks from student dropout prediction [9], forecasting on-time graduation of students [15] to education admission decisions [21]. Recently, the issue of bias and discrimination in ML-based decision-making systems is receiving a lot of attention [28] as there are many recorded incidents of discrimination (e.g., recidivism prediction [20], grades prediction [4, 14]) caused by such systems against individuals or groups or people on the basis of *protected attributes* like gender, race etc. Bias in education is not a new problem. There is already a long literature on different sources of bias in education [24] or students' data analysis [3] as well as studies on racial bias [31] and gender bias [22]. However, ML-based decision-making systems have the potential to amplify prevalent biases or create new ones and therefore, fairness-aware ML approaches are required also for the educational domain.

In this work, we focus on *fairness in clustering*, as in educational activities, group assignments [8] and student team achievement divisions [30] are important tools that help students working together towards shared learning goals. Clustering is an effective solution for partitioning students into groups of *similar instances* [3, 26]. Traditional algorithms, however, focus solely on the similarity objective and do not consider the fairness of the resulting clusters w.r.t. protected attributes like gender. However, studies indicate that students might learn better in diverse groups, e.g., mixed-gender groups [11, 32]. Lately, fair-clustering solutions have been proposed [1, 2, 5, 6], which aim to discover clusters with a fair representation regarding some protected attributes. In this work, we adopt the cluster fairness of [6], called cluster balance, according to which protected groups must have approximately equal representation in every cluster.

In a teaching situation, it is obvious that the size of the groups should be comparable to allow a fair allocation of work among the students. As traditional clustering algorithms do not consider this requirement, clusters of varying sizes might be extracted, reducing the usefulness and applicability of the partitioning for end-users/teachers. This leads to the demand for clustering solutions that also take into account the size of the clusters. The problem is known as *capacitated clustering problem (CCP)* [25] which aims to extract clusters with a limited capacity while minimizing the total dissimilarity in the clusters. Capacitated clustering is useful in quite a few applications such as transferring goods/services from the service providers (post office, stores, etc.), garbage collection and sales force territorial design [27]. To the best of our knowledge, no solution exists that considers both fairness and capacity of clusters on the top of the similarity objective.

To this end, we propose a new problem, the so-called *fair-capacitated clustering* that ensures fairness and balanced cardinalities of the resulting clusters. We decompose the problem into two subproblems: i) the fairness-requirement compliance step that preserves fairness at a minimum threshold of balance score and ii) the capacity-requirement compliance step that ensures clusters of comparable sizes. For the first step, we generate fairlets [6], which are minimal sets that satisfy fair representation w.r.t. a protected attribute. For the second step, we propose two solutions for two different clustering types, namely hierarchical and partitioning-based clustering, that consider the capacity constraint during the merge step (hierarchical approach)/assignment step (partitioning approach). Experimental results, on three real datasets from the education domain, show that our methods result in fair and capacitated clusters while preserving the clustering quality.

2. RELATED WORK

Chierichetti et al. [6] introduced the fair clustering problem with the aim to ensure equal representation for each protected attribute, such as gender, in every cluster. In their formulation, each instance is assigned with one of two colors (red, blue). They proposed a two-phase approach: clustering all instances into fairlets - small clusters preserving the fairness measure, and then applying vanilla clustering methods on those fairlets. Subsequent studies focus on generalization and scalability. Backurs et al. [1] presented an approximate fairlet decomposition algorithm which can formulate the fairlets in nearly linear time thus tackling the efficiency bottleneck of [6]. Rösner and Schmidt [29] generalized the fair clustering problem to more than two protected attributes. A more generalized and tunable notion of fairness for clustering was introduced in Bera et al. [2]. Anshuman and Prasant [5] introduced a fair hierarchical agglomerative clustering method for multiple protected attributes.

The capacitated clustering problem (CCP), a combinatorial optimization problem, was first introduced by Mulvey and Beck [25] who proposed solutions using heuristic and subgradient algorithms. Several approaches exist to improve the efficiency of solutions or CCP approaches for different cluster types. Khuller and Sussmann [17], for example, introduced an approximation algorithm for the capacitated k-Center problem. Geetha et al. [10] improved k-Means algorithm for CCP by using a priority measure to assign points to their centroid. Lam and Mittenthal [19] proposed a heuristic hierarchical clustering method for CCP to solve the multi-depot location-routing problem.

In this work, we introduce the problem of fair-capacitated clustering which builds upon the notions from fair clustering and capacitated clustering. In particular, we build upon the notion of fairlets [6] to extract the minimal sets that preserve fairness. Regarding the CCP we follow the formulation of [25] to ensure balanced cluster cardinalities. To the best of our knowledge, the combined problem has not been studied before and as already discussed, comprises a useful tool in quite a few domains like education.

3. PROBLEM DEFINITION

Let $X \in \mathbb{R}^n$ be a set of instances to be clustered and let $d() : X \times X \rightarrow \mathbb{R}$ be the distance function. For an integer k we use $[k]$ to denote the set $\{1, 2, ..., k\}$. A *k-clustering* $\mathcal{C}$ is a partition of X into k disjoint subsets, $\mathcal{C} = \{C_1, C_2, ..., C_k\}$, called *clusters* with $S = \{s_1, s_2, ..., s_k\}$ be the corresponding cluster centers. The goal of clustering is to find an *assignment* $\phi : X \rightarrow [k]$ that minimizes the objective function:

$$\mathcal{L}(X, \mathcal{C}) = \sum_{s_i \in S} \sum_{x \in C_i} d(x, s_i) \tag{1}$$

As shown in Eq. 1, the goal is to find an assignment that minimizes the sum of distances between each point $x \in X$and its corresponding cluster center $s_i \in S$. It is clear that such an assignment optimizes the similarity but does not consider fairness or capacity of the resulting clusters.

Capacitated clustering: The goal of capacitated clustering [25] is to discover clusters of given capacities while still minimizing the distance objective $\mathcal{L}(X, \mathcal{C})$. The capacity constraint is defined as an upper bound Q_i on the cardinality of each cluster C_i:

$$|C_i| \leq Q_i \tag{2}$$

Clustering fairness: We assume the existence of a binary protected attribute $P = \{0, 1\}$, e.g., gender={male, female}. Let $\psi : X \rightarrow P$ denotes the demographic group to which the point belongs, i.e., male or female.

Fairness of a cluster is evaluated in terms of the balance score [6] as the minimum ratio between two groups.

$$balance(C_i) = \min\left(\frac{|\{x \in C_i | \psi(x) = 0\}|}{|\{x \in C_i | \psi(x) = 1\}|}, \frac{|\{x \in C_i | \psi(x) = 1\}|}{|\{x \in C_i | \psi(x) = 0\}|}\right) \tag{3}$$

Fairness of a clustering $\mathcal{C}$ equals to the balance of the least balanced cluster $C_i \in \mathcal{C}$.

$$balance(\mathcal{C}) = \min_{C_i \in C} balance(C_i) \tag{4}$$

We now introduce the problem of fair-capacitated clustering that combines all aforementioned objectives regarding distance, fairness and capacity.

Definition 1. (Fair-capacitated clustering problem)
We define the problem of (t, k, q)-fair-capacitated clustering as finding a clustering $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$ that partitions the data X into $|\mathcal{C}| = k$ clusters such that the cardinality of each cluster $C_i \in \mathcal{C}$ does not exceed a threshold q, i.e., $|C_i| \leq q$ (*the capacity constraint*), the balance of each cluster is at least t, i.e., $balance(\mathcal{C}) \geq t$ (*the fairness constraint*), and minimizes *the objective function* $\mathcal{L}(X, \mathcal{C})$. Parameters k, t, q are user defined referring to the number of clusters, minimum balance threshold and maximum cluster capacity, respectively.

4. FAIR-CAPACITATED CLUSTERING

4.1 Fairlet decomposition

Traditionally, the vanilla versions of clustering algorithms are not capable of ensuring fairness because they assign the data points to the closest center without the fairness consideration. Hence, if we could divide the original data set into subsets such that each of them satisfies the balance threshold t then grouping these subsets to generate the final clustering would still preserve the fairness constraint. Each fair subset is defined as a fairlet. We follow the definition of fairlet decomposition by [6].

Definition 2. (Fairlet decomposition)
Suppose that $balance(X) \geq t$ with $t = f/m$ for some integers $1 \leq f \leq m$, such that the greatest common divisor $gcd(f, m) = 1$. A decomposition $\mathcal{F} = \{F_1, F_2, ..., F_l\}$ of X is a fairlet decomposition if: i) each point $x \in X$ belongs to exactly one fairlet $F_j \in \mathcal{F}$; ii) $|F_j| \leq f + m$ for each $F_j \in \mathcal{F}$, i.e., the size of each fairlet is small; and iii) for each $F_j \in \mathcal{F}$, $balance(F_j) \geq t$, i.e., the balance of each fairlet satisfies the threshold t. Each F_j is called a fairlet.

By applying fairlet decomposition on the original dataset X, we obtain a set of fairlets $\mathcal{F} = \{F_1, F_2, \ldots, F_l\}$. For each fairlet F_j we select randomly a point $r \in F_j$ as the *center*. For a point $x \in X$, we denote $\gamma : X \rightarrow [1, l]$ as the index of the mapped fairlet. The second step, is to cluster the set of fairlets $\mathcal{F}$ into k final clusters, subject to the cardinality constraint. The clustering process is described below for the hierarchical clustering type (Section 4.2) and for the partitioning-based clustering type (Section 4.3). Clustering results in an assignment from fairless to final clusters: $\delta : \mathcal{F} \rightarrow [k]$. The final fair-capacitated clustering $\mathcal{C}$ can be determined by the overall assignment function $\phi(x) = \delta(F_{\gamma(x)})$, where $\gamma(x)$ returns the index of the fairlet to which x is mapped.

4.2 Fair-capacitated hierarchical clustering

Given the set of fairlets: $\mathcal{F} = \{F_1, F_2, \ldots, F_l\}$, let $W = \{w_1, w_2, \ldots, w_l\}$ be their corresponding weights, where the weight w_j of a fairlet F_j is defined as its cardinality, i.e., number of points in F_j.

Traditional agglomerative clustering approaches merge the two closest clusters, so rely solely on similarity. We extend the merge step by also ensuring that merging does not violate the cardinality constraint w.r.t. the cardinality threshold q.

THEOREM 1. *The balance score of a cluster formed by the union of two or more fairlets, is at least t.*

$$balance(\mathcal{Y}) \geq t, \text{ where } \mathcal{Y} = \cup_{j \leq l} F_j \text{ and } balance(F_j) \geq t$$

PROOF. We use the method of induction to derive the proof. Assume we have a set of fairlets $\mathcal{F} = \{F_1, F_2, \ldots, F_l\}$, in which, $balance(F_j) \geq t, \quad j = 1, \ldots, l$. We first consider the case for any two fairlets $\{F_1, F_2\} \in \mathcal{F}$. We have $balance(F_1) = \frac{f_1}{m_1} \geq t$ and $balance(F_2) = \frac{f_2}{m_2} \geq t$. We denote by $\mathcal{Y}$ is the union of two fairlets F_1 and F_2, then

$$balance(\mathcal{Y}) = balance(F_1 \cup F_2) = \frac{f_1 + f_2}{m_1 + m_2} \tag{5}$$

It holds:

$$\begin{gathered}
\frac{f_1}{m_1} \geq t \quad \text{ or, } \frac{f_1}{m_1 + m_2} \geq \frac{tm_1}{m_1 + m_2} \\
\text{Similarly, } \frac{f_2}{m_1 + m_2} \geq \frac{tm_2}{m_1 + m_2} \\
\implies \frac{f_1}{m_1 + m_2} + \frac{f_2}{m_1 + m_2} \geq \frac{tm_1}{m_1 + m_2} + \frac{tm_2}{m_1 + m_2} \\
\implies \frac{f_1 + f_2}{m_1 + m_2} \geq \frac{t(m_1 + m_2)}{m_1 + m_2} = t
\end{gathered} \tag{6}$$

Therefore, from Eq. 5 and Eq. 6 we get,

$$balance(\mathcal{Y}) \geq t \tag{7}$$

Thus, the statement given in Theorem 1 is true for any cluster formed by the union of any two fairlets. Now we assume that the statement holds true for a cluster formed from i fairlets, i.e, $\mathcal{Y} = \cup_{j \leq i} F_j$, where $1 < i < l$. Then,

$$balance(\mathcal{Y}) = \frac{\sum_{j \leq i} f_j}{\sum_{j \leq i} m_j} \geq t \tag{8}$$

Consider another fairlet $F_{i+1} \in \mathcal{F}$ which is not in the formed cluster $\mathcal{Y}$, $balance(F_{i+1}) = \frac{f_{i+1}}{m_{i+1}} \geq t$. Then, by joining F_{i+1} with the cluster $\mathcal{Y}$ we get the new cluster $\mathcal{Y}^{'}$ such that,

$$balance(\mathcal{Y}^{'}) = \frac{f_{i+1} + \sum_{j \leq i} f_j}{m_{i+1} + \sum_{j \leq i} m_j} \tag{9}$$

Following the steps in Eq. 6, we can similarly show that,

$$\frac{f_{i+1} + \sum_{j \leq i} f_j}{m_{i+1} + \sum_{j \leq i} m_j} \geq t \implies balance(\mathcal{Y}^{'}) \geq t \tag{10}$$

Hence, the theorem holds true for cluster formed with $i + 1$ fairlets if it is true for i fairlets. Since i is any arbitrary number of fairlets, thus the theorem holds true for all cases. □

Theorem 1 shows that for any cluster formed by union of fairlets, the fairness constraint is always preserved. Henceforth, we don't need further interventions w.r.t. fairness.

The pseudocode is shown in Algorithm 2 of Appendix B. In each step, the closest pair of clusters is identified and a new cluster is created only if its capacity does not exceed the capacity threshold q. Otherwise, the next closest pair is investigated. The procedure continues until k clusters remain. The remaining clusters are fair and capacitated according to the correponding thresholds t and q. To compute the proximity matrix (line 1 and line 8), we use the distance between centroids of the corresponding clusters. The function *capacity*(*cluster*) in line 5 returns the size of a cluster.

4.3 Fair-capacitated partitioning clustering

Partitioning-based clustering algorithms, such as k-Medoids, can be viewed as a distance minimization problem, in which, we try to minimize the objective function in Eq. 1. The vanilla k-Medoids does not satisfy the cardinality constraint since the allocating points to clusters step is only based on the distance among them. Now, if we change the goal of this assignment step to find the "best" data points with a defined capacity for each medoid instead of searching for the most suitable medoid for each point, we can control the cardinality of clusters. We formulate the problem of *assigning points to clusters* subject to a capacity threshold q as a knapsack problem [23].

Let $S = \{s_1, s_2, ..., s_k\}$ be the cluster centers, i.e., medoids, and $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ be the resulting clusters. We change the assignments of points to clusters, using knapsack, in order to meet the capacity constraint q. In particular, we define a flag variable $y_j = 1$ if x_j is assigned to cluster C_i, otherwise $y_j = 0$. Now, we define a value v_j to data point x_j, which depends on the distance of x_j to C_i, with v_j being maximum if C_i is the best cluster for x_j, i.e, the distance between x_j and s_i is minimum. We formulate value v_j of instance x_j based on an exponential decay distance function:

$$v_j = e^{-\frac{1}{\lambda} * d(x_j, s_i)} \tag{11}$$

where $d(x_j, s_i)$ is the Euclidean distance between the point x_j and the medoid s_i. The higher λ is the lower the effect of distance in the value of the points. The point which is closer to the medoid will have a higher value.

Then, the objective function for the assignment step is:

$$\text{maximize} \sum_{j=1}^{n} v_j y_j \tag{12}$$

Now, given $\mathcal{F} = \{F_1, F_2, \ldots, F_l\}$ and $W = \{w_1, w_2, \ldots, w_l\}$ are the set of fairlets and their corresponding weights, i.e., the number of instances in the fairlet, respectively; q is the maximum capacity of the final clusters. Our target is to cluster the set of fairlets $\mathcal{F}$ into k clusters centered by k medoids. We apply the formulas in Eq. 11 and Eq.12 on the set of fairlets $\mathcal{F}$, i.e, each fairlet F_j has the same role as x_j. Then, the problem of assigning the fairlets to each *medoid* in the cluster assignment step becomes finding a set of fairlets with total weights less than or equal to q and the total value is maximized. In other words, we can formulate the cluster assignment step in the partitioning-based clustering as a **0-1 knapsack problem**.

$$\begin{aligned} &\text{maximize} \sum_{j=1}^{l} v_j y_j \\ &\text{subject to} \quad \sum_{j=1}^{l} w_j y_j \leq q \quad \text{and} \quad y_j \in \{0, 1\} \end{aligned} \tag{13}$$

In which, y_j is the flag variable for F_j, $y_j = 1$ if F_j is assigned to a cluster, otherwise $y_j = 0$; v_j is the value of F_j which is computed by the Eq. 11; q is the desired maximum capacity.

The pseudocode of our k-Medoids fair-capacitated algorithm is described in Algorithm 2. In which, for each *medoid* we would search for the adequate points (line 3) by using function $knapsack(p, values, w, q)$ (line 10) implemented using dynamic programming, which returns a list of items with a maximum total value and the total weight not exceeding q. In the main function, line 12, we optimize the clustering cost by replacing *medoids* with *non-medoid* instances when the clustering cost is decreased. This optimization procedure will stop when there is no improvement in the clustering cost is found (lines 19 to 32).

Algorithm 1: k-Medoids fair-capacitated algorithm

```
Input: F = {F_1, F_2, ..., F_l}: a set of fairlets
       W = {w_1, w_2, ..., w_l}: weights of fairlets
       q: a given maximum capacity of final clusters
       k: number of clusters
Output: A fair-capacitated clustering
1  Function ClusterAssignment(medoids):
2      clusters ← ∅;
3      for each medoid s in medoids do
4          candidates ← all fairlets which are not assigned
            to any cluster ;
5          p ← length(candidates) ;
6          w ← weights(candidates) ;
7          for each fairlet_i in candidates do
8              values[i] ← v(fairlet_i) //Eq. 11 ;
9          end
10         clusters[s]←knapsack(p, values, w, q) ;
11     end
12     return clusters;
13 Function main():
14     medoids ← select k of the l fairlets arbitrarily ;
15     ClusterAssignment(medoids) ;
16     cost_best ← current clustering cost;
17     s_best ← null ;
18     o_best ← null ;
19     repeat
20         for each medoid s in medoids do
21             for each non-medoid o in F do
22                 consider the swap of s and o, compute
                    the current clustering cost;
23                 if current clustering cost < cost_best then
24                     s_best ← s;
25                     o_best ← o;
26                     cost_best ← current clustering cost;
27                 end
28             end
29         end
30         update medoids by the swap of s_best and o_best ;
31         ClusterAssignment(medoids)
32     until no improvements can be achieved by any
        replacement;
33 return clusters;
```

5. EXPERIMENTS

In this section, we describe our experiments and the performance of our proposed methods on three educational datasets.

5.1 Experimental setup

Datasets. The datasets are summarized in Table 1.

UCI student performance. This dataset includes the demographics, grades, social and school-related features of students in secondary education of two Portuguese schools [7] in 2005 - 2006. "Gender" is selected as the protected attribute, i.e., we aim to balance gender in the resulting clusters.

Open University Learning Analytics (OULAD). This is the dataset from the OU Analyse project [18] of Open University, England in 2013 - 2014. Information of students includes their demographics, courses, their interactions with the virtual learning environment, and final outcome. We aim to balance the "Gender" attribute in the results.

MOOC. The data covers students who enrolled in the 16 edX courses offered by the two institutions (Harvard University and the Massachusetts Institute of Technology) during 2012 - 2013 [13]. The dataset contains aggregated records which represent students' activities and their final grades of the courses. "Gender" is the protected attribute.

Baselines. We compare against well-known fairness-aware

Table 1: An overview of the datasets

Dataset	#instances	#attributes	Protected attribute	Balance score
UCI student performance	649	33	Gender (F: 383; M: 266)	0.695
OULAD	4,000	12	Gender (F: 2,000; M: 2,000)	1
MOOC	4,000	21	Gender (F: 2,000; M: 2,000)	1

clustering methods and a vanilla clustering method.

k-Medoids [16]: a traditional partitioning clustering technique that divides the dataset into k clusters so as to minimize clustering cost. Cluster centers are actual instances.

Vanilla fairlet [6]: a fairness-aware clustering approach that i) decomposes the dataset into fairlets and ii) applies a vanilla k-center clustering algorithm [12] to form the final k clusters.

MCF fairlet [6]: Similar to Vanilla fairlet but the fairlet decomposition part is transformed into a *minimum cost flow* (MCF) problem, by which an optimized version of fairlet decomposition in terms of cost value is computed.

Evaluation measures. We report on clustering quality (measured as clustering cost, see Eq. 1), cluster fairness (expressed as cluster balance, see Eq. 4) and cluster capacity (expressed as cluster cardinality).

Parameter selection. Regarding fairness, a minimum threshold of balance t is set to 0.5 for all datasets in our experiments. It means that the proportion of the minority group is at least 50% in each resulting cluster. Regarding the λ factor in Eq. 11, a value $\lambda = 0.3$ is chosen for our experiments from a range of [0.1, 1.0] via grid-search. We evaluate the clustering cost and balance score on a small dataset, i.e., UCI student performance dataset - Mathematics subject w.r.t λ. Theoretically, the **ideal capacity** of clusters is $\left\lceil \frac{|X|}{k} \right\rceil$ where $|X|$ is the population of dataset X, k is the number of desired clusters. However, in many cases, the clustering models cannot satisfy this constraint, especially the hierarchical clustering model. Hence, we set the formula in Eq. 14 to compute the **maximum capacity** q of clusters; ε is a parameter chosen in experiments for each fair-capacitated clustering approach.

$$q = \left\lceil \frac{|X| * \varepsilon}{k} \right\rceil \tag{14}$$

To find the appropriate value of ε, we set a range of [1.0, 1.3] to ensure all the generated clusters have members and evaluate the cardinality of resulting clusters on the UCI student performance (Mathematics subject) dataset. ε is set to 1.01 and 1.2, for k-Medoids fair-capacitated and hierarchical fair-capacitated methods, respectively.

5.2 Experimental results

UCI student performance. When k is less than 4, as shown in Figure 1-a, the clustering quality of our models can be close to that of the vanilla k-Medoids method. However, the clustering cost is fluctuated thereafter due to the effort to maintain the fairness and cardinality of methods. Our *vanilla fairlet hierarchical fair-capacitated* outperforms other competitors in most cases. Vanilla fairlet and MCF fairlet show the worst clustering cost as an effect of the k-Center method. Figure 1-b depicts the clustering fairness. As we can observe, in terms of fairness, *vanilla fairlet hierarchical fair-capacitated* has the best performance when k is less than 10. Contrary to that, by selecting each point for each cluster in the cluster assignment step, the *k-Medoids fair-capacitated* method can maintain well the fairness in many cases. Regarding the cardinality, as illustrated in Figure 1-c, our approaches outperform the competitors when they can keep the number of instances for each cluster under the specified thresholds.

OULAD. Our *MCF fairlet k-Medoids fair-capacitated* approach outperforms other methods in terms of clustering cost, although there is an increase compared to the vanilla k-Medoids algorithm, as we can see in Figure 2-a. Concerning fairness, in Figure 2-b, k-Medoids is the weakest method while others can achieve the highest balance. The balance of *Gender* feature in the dataset is the main reason for this result. All fairlets are fully fair; this is a prerequisite for our methods of being able to maintain the perfect balance. Regarding cardinality, our approaches demonstrate their strength in ensuring the capacity of clusters (Figure 2-c). The difference in the size of the clusters generated by our methods is tiny. This is in stark contrast to the trend of competitors.

MOOC. The results of clustering quality are described in Figure 3-a (Appendix A). Although an increase in the clustering cost is the main trend, our methods outperform the vanilla fairlet and MCF fairlets methods. Regarding clustering fairness, as depicted in Figure 3-b, our approaches can maintain the perfect balance for all experiments. This is the result of an actual balance in the dataset and the fairlets. The emphasis is our methods can divide all the experimented instances into capacitated clusters, as shown in Figure 3-c, which proves their superiority in presenting the results over the competitors regarding clusters' cardinality.

Summary of the results. In general, fairness is well maintained in all of our experiments. When the data is fair, in case of OULAD and MOOC datasets, our methods achieve a perfect fairness. In terms of cardinality, our methods are able to maintain the cardinality of resulting clusters within the maximum capacity threshold, which is significantly superior to competitive methods. The fair-capacitated partitioning based method is better than the hierarchical approach since it can determine the capacity threshold closest to the *ideal capacity*. Regarding the clustering cost, the hierarchical approach has an advantage over other methods by outperforming its competitors in most experiments.

6. CONCLUSION AND OUTLOOK

In this work, we introduced the fair-capacitated clustering problem that extends traditional clustering, solely focusing on similarity, by also aiming at a balanced cardinality among the clusters and a fair-representation of instances in each cluster according to some protected attributes like gender or race. Our solutions work on the fairlets derived from

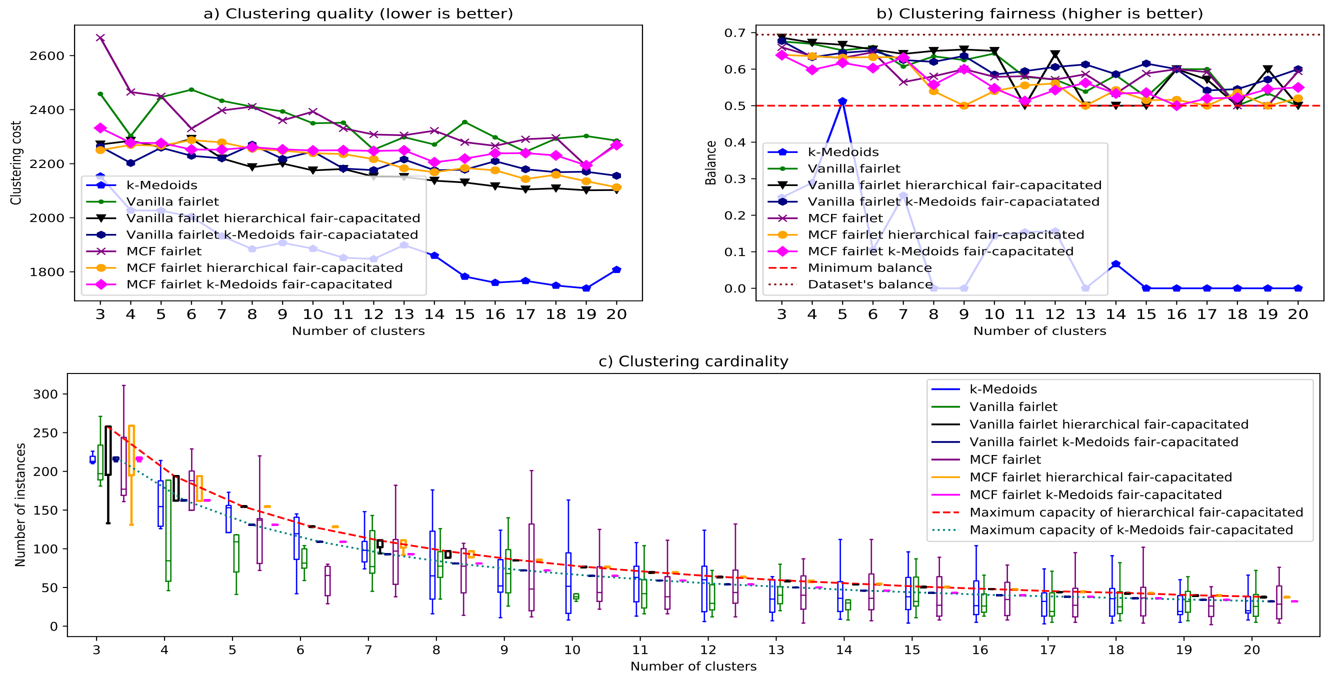

Figure 1: Performance of different methods on UCI student performance dataset

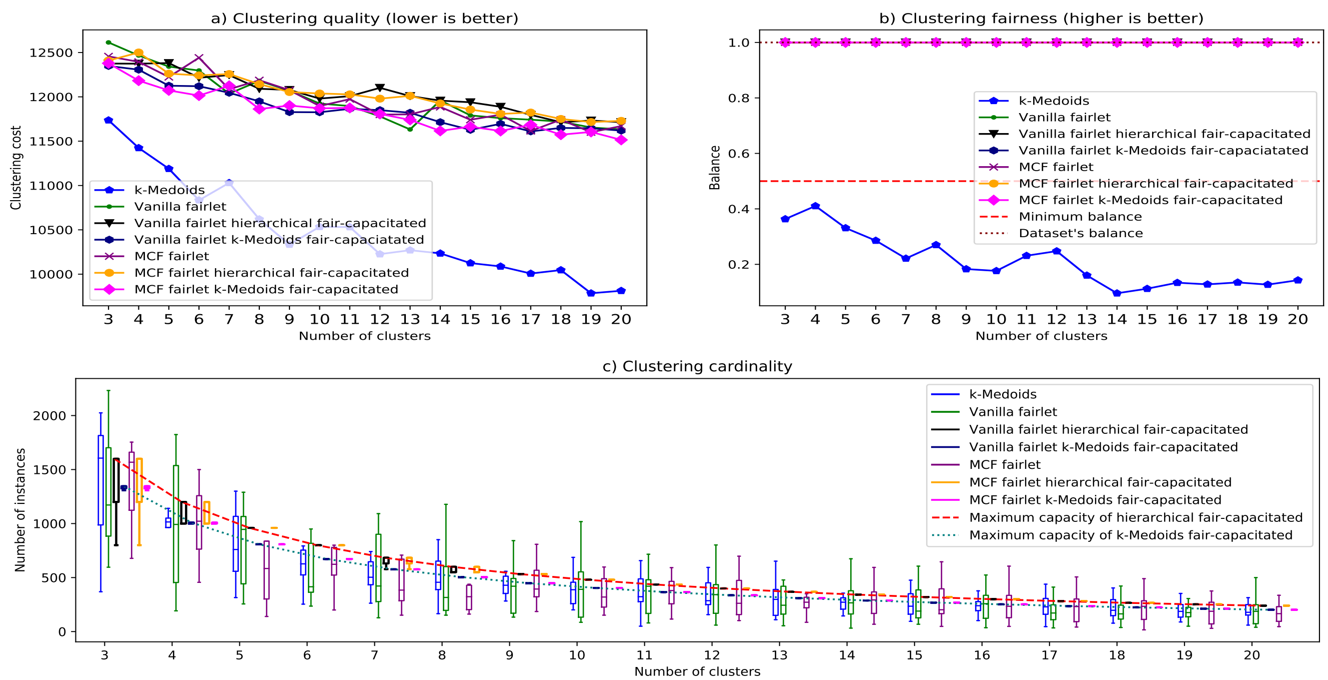

Figure 2: Performance of different methods on OULAD dataset

the original instances: the hierarchical-based approach takes into account the cardinality requirement during the merging step, whereas the partitioning-based approach takes into account the cardinality of the final clusters during the assignment step which is formulated as a knapsack problem. Our experiments show that our methods are effective in terms of fairness and cardinality while maintaining clustering quality. In the future, we plan to extend our approach for more than one protected attributes as well as to further investigate what fair group assignments means in educational settings.

Acknowledgements

The work of the first author is supported by the Ministry of Science and Education of Lower Saxony, Germany, within the PhD program "LernMINT: Data-assisted teaching in the MINT subjects". The work of the second author is supported by the Volkswagen Foundation under the call "Artificial Intelligence and the Society of the Future".

7. REFERENCES

[1] A. Backurs, P. Indyk, K. Onak, B. Schieber, A. Vakilian, and T. Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR, 2019.

[2] S. Bera, D. Chakrabarty, N. Flores, and M. Negahbani. Fair algorithms for clustering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[3] S. Bharara, S. Sabitha, and A. Bansal. Application of learning analytics using clustering data mining for students' disposition analysis. *Education and Information Technologies*, 23(2):957–984, 2018.

[4] K. Bhopal and M. Myers. The impact of covid-19 on a level students in england. *SocArXiv*, 2020.

[5] A. Chhabra and P. Mohapatra. Fair algorithms for hierarchical agglomerative clustering. *arXiv preprint arXiv:2005.03197*, 2020.

[6] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5036–5044, 2017.

[7] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. *EUROSIS-ETI*, 2008.

[8] M. Ford and J. Morice. How fair are group assignments? a survey of students and faculty and a modest proposal. *Journal of Information Technology Education: Research*, 2(1):367–378, 2003.

[9] J. Gardner, C. Brooks, and R. Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 225–234, 2019.

[10] S. Geetha, G. Poonthalir, and P. Vanathi. Improved k-means algorithm for capacitated clustering problem. *INFOCOMP*, 8(4):52–59, 2009.

[11] D. Gnesdilow, A. Evenstone, J. Rutledge, S. Sullivan, and S. Puntambekar. Group work in the science classroom: How gender composition may affect individual performance. 2013.

[12] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38:293–306, 1985.

[13] HarvardX. HarvardX Person-Course Academic Year 2013 De-Identified dataset, version 3.0, 2014.

[14] S. Hubble and P. Bolton. A level results in england and the impact on university admissions in 2020-21. *House of Commons Library*, 2020.

[15] S. Hutt, M. Gardner, A. L. Duckworth, and S. K. D'Mello. Evaluating fairness and generalizability in models predicting on-time graduation from college applications. *International Educational Data Mining Society*, 2019.

[16] L. Kaufman and P. J. Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125, 1990.

[17] S. Khuller and Y. J. Sussmann. The capacitated k-center problem. *SIAM Journal on Discrete Mathematics*, 13(3):403–418, 2000.

[18] J. Kuzilek, M. Hlosta, and Z. Zdrahal. Open university learning analytics dataset. *Scientific data*, 4:170171, 2017.

[19] M. Lam and J. Mittenthal. Capacitated hierarchical clustering heuristic for multi depot location-routing problems. *Int. J. Logist. Res. Appl.*, 16(5):433–444, 2013.

[20] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016.

[21] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich. Implications of ai (un-) fairness in higher education admissions: the effects of perceived ai (un-) fairness on exit, voice and organizational reputation. In *FAT* '20*, pages 122–130, 2020.

[22] T. Masterson. An empirical analysis of gender bias in education spending in paraguay. *World Development*, 40(3):583–593, 2012.

[23] G. B. Mathews. On the partition of numbers. *Proceedings of the London Mathematical Society*, 1(1):486–490, 1896.

[24] M. Meaney and T. Fikes. Early-adopter iteration bias and research-praxis bias in the learning analytics ecosystem. In *Companion Proceeding of the 9th International Conference on Learning Analytics & Knowledge (LAK'19), Fairness and Equity in Learning Analytics Systems Workshop*, pages 14–20, 2019.

[25] J. M. Mulvey and M. P. Beck. Solving capacitated clustering problems. *European Journal of Operational Research*, 18(3):339–348, 1984.

[26] Á. A. M. Navarro and P. M. Ger. Comparison of clustering algorithms for learning analytics with educational datasets. *IJIMAI*, 5(2):9–16, 2018.

[27] M. Negreiros and A. Palhano. The capacitated centred clustering problem. *Computers & operations research*, 33(6):1639–1663, 2006.

[28] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al. Bias in data-driven artificial intelligence systems-an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[29] C. Rösner and M. Schmidt. Privacy preserving clustering with constraints. *arXiv preprint arXiv:1802.02497*, 2018.

[30] M. Tiantong and S. Teemuangsai. Student team achievement divisions (stad) technique through the moodle to enhance learning achievement. *International Education Studies*, 6(4):85–92, 2013.

[31] N. Warikoo, S. Sinclair, J. Fei, and D. Jacoby-Senghor. Examining racial bias in education: A new approach. *Educational Researcher*, 45(9):508–514, 2016.

[32] Z. Zhan, P. S. Fong, H. Mei, and T. Liang. Effects of gender grouping on students' group performance, individual achievements and attitudes in computer-supported collaborative learning. *Computers in Human Behavior*, 48:587–596, 2015.

APPENDIX

A. MOOC DATASET

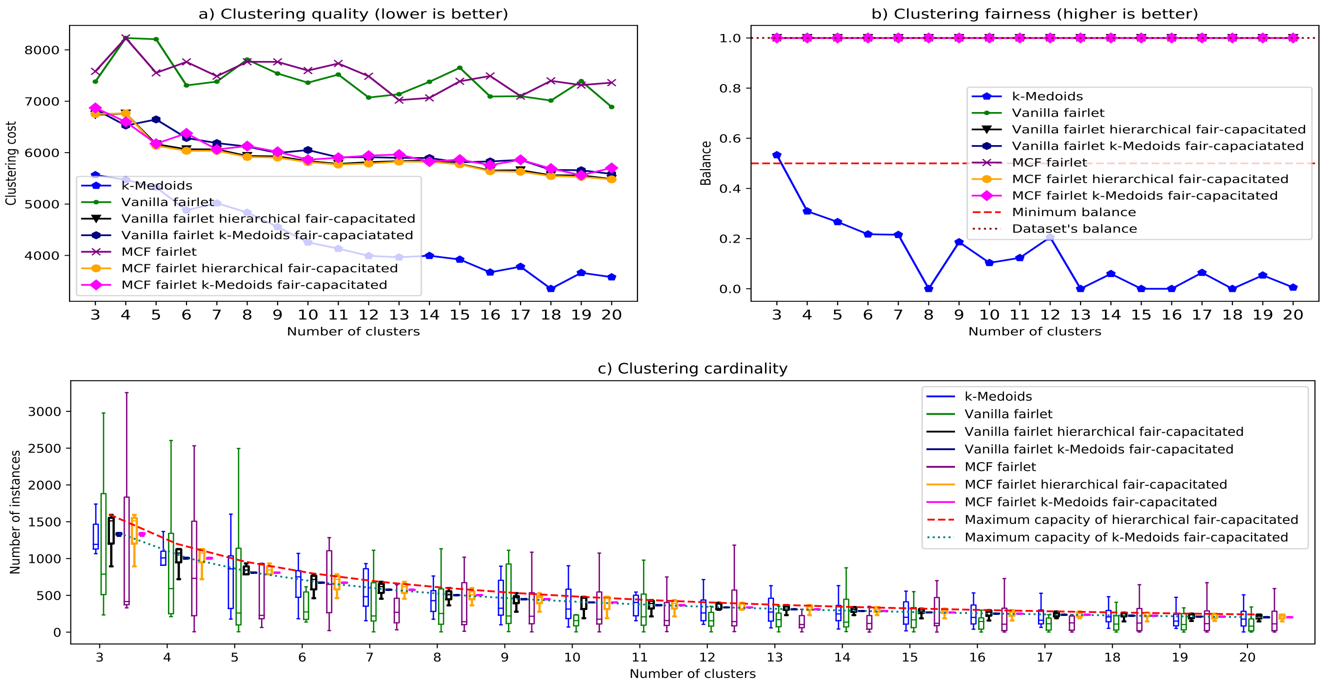

Figure 3: Performance of different methods on MOOC dataset

B. HIERARCHICAL FAIR-CAPACITATED ALGORITHM

Algorithm 2: Hierarchical fair-capacitated algorithm

Input: $\mathcal{F} = \{F_1, F_2, \ldots, F_l\}$: a set of fairlets
q: a given maximum capacity of final clusters
$W = \{w_1, w_2, \ldots, w_l\}$: weights of fairlets
k: number of clusters

Output: A fair-capacitated clustering

compute the proximity matrix ;
$clusters \leftarrow \mathcal{F}$ //each fairlet F_j is considered as cluster ;
repeat
  $cluster_1, cluster_2 \leftarrow$ the closest pair of clusters ;
  if $capacity(cluster_1) + capacity(cluster_2) \leq q$ **then**
    $newcluster \leftarrow$ merge$(cluster_1, cluster_2)$;
    update $clusters$ with $newcluster$;
    update the proximity matrix ;
  else
    continue;
  end
until k *clusters remain*;
return $clusters$;