Clustering-based Subgroup Detection for Automated Fairness Analysis

Jero Schäfer[1][0000−0001−7727−1181] and Lena Wiese[1][0000−0003−3515−9209]

Institute of Computer Science, Goethe University, Frankfurt am Main, Germany
jeschaef@cs.uni-frankfurt.de, lwiese@cs.uni-frankfurt.de

Abstract. Fairness in Artificial Intelligence is a major requirement for trust in ML-supported decision making. Up to now fairness analysis depends on human interaction – for example the specification of relevant attributes to consider. In this paper we propose a subgroup detection method based on clustering to automate this process. We analyse 10 (sub-)clustering approaches with three fairness metrics on three datasets and identify SLINK as an optimal candidate for subgroup detection.

Keywords: Artificial Intelligence · Fairness · Clustering.

1 Introduction

Nowadays a great variety of AI systems are spread over the digital world and affect the lives of millions of people every day. Over the past years there has been a strive for optimizing performance of such systems by better and faster technologies – but modern ML requires for other inalienable objectives, too. The societal impact of decisions of AI systems has to be considered along the objective of maximizing the prediction accuracy. As a consequence, ML models should also be checked carefully for providing equal treatment of individuals from different ethnics, races, or sexes. Especially, the intersections of sensitive characteristics – or those characteristics not obviously involved in discrimination – make judging the model behavior challenging when facing complex data. Additionally, the huge number of possible subgroups growing exponentially with the number of features inside data makes it infeasible to test the model's behavior towards each subgroup. Thus, automation of fairness testing is required to solve this issue.

We propose two subgroup detection methods based on an unsupervised clustering. The computed clusters serve as subgroups for the fairness evaluation and prototypes for the generation of patterns defining subgroups. Furthermore, we compare different clustering algorithms on their performance to identify subgroups for a fairness assessment of a binary classifier under three common fairness criteria and three fairness-related datasets. In Section 2 we give an overview over related work on automated subgroup fairness and Section 3 introduces the theory for our methods of subgroup detection, that are explained in Section 4. Section 5 describes our experimental setting and discusses the results. Finally, we the summarize key findings and give an outlook into future work in Section 6.

2 Related Work

There has been an uprising number of tools being developed to aid data scientists and developers with the investigation and improvement of their developed ML models. They usually support the model selection or optimization phase by diverse visualizations of data and the model performance. Recently, there evolved an advance towards the assessment of model fairness to meet the demands of society for equality in AI. However, many tools require expert knowledge as they partly rely on user interaction via controls or parameters.

The tools Boxer [5] and Fairkit [7] let the user interactively explore and compare the behavior of multiple models. They opt for the identification of intersectional bias in the model but require the selection of subgroups to investigate for discrimination. The What-If tool [13] yields insights into local and global modal behavior in various scenarios, performs an intersectional analysis for a chosen fairness objective and automatically adapts the model's classification threshold. The framework of Morina et al. [9] comprises a suite of metrics for evaluating and estimating intersectional fairness but also does not discover subgroups automatically. The FairVis tool [2] was designed to identify the intersectional bias of ML models by visualizations. Despite possible user interaction, automatically generated subgroups are suggested and the subgroup performance and fairness of the model are presented. These subgroups are found by a k-means clustering and extracting patterns that describe the makeup of the cluster members. The dominant features in a cluster are ranked by the feature entropy quantifying the feature's uniformity. Our approach uses a similar technique that directly uses the clustering results and the feature entropy with a threshold instead of a ranking.

In contrast, the Divexplorer project [10] provides automatic subgroup detection by frequent-pattern mining. An exhaustive search through possible itemsets (i.e. patterns to match data instances to) is carried out and only itemsets above a support threshold are considered while pruning others from the search tree. The divergence of the model behavior between a subgroup of instances complying to a mined pattern and the full dataset is assessed by an outcome function for classification or ranking tasks that evaluates fairness by the difference between the FPR or FNR of a subgroup and the global rate. The DENOUNCER [8] system discovers subgroups with a low prediction accuracy by pattern graph traversal, applies a support threshold to the attribute-value patterns filtering out insignificant patterns and prunes for the most general patterns.

3 AI Fairness

Generally, one can distinguish *individual* (similarity-based) and *group* (statistical) fairness criteria. Individual fairness refers to the discrimination by the model on an individual level (per instance) and is expressed as the different behavior of the model wrt. similar individuals although they should be treated similarly. This work is focused on the assessment of group fairness of a given classification model that tests whether the model systematically discriminates against a certain subgroup of instances [6]. The subgroups of interest are therefore usually

defined for a set of protected attributes (such as sex or nationality) but generally involve the intersection of multiple protected attributes. Hence we formalize a dataset as $\mathcal{D} = \{x_1, \ldots, x_n\}$ with the set of attributes $\mathcal{A} = \{A_1, \ldots, A_p\}$ that comprises n instances x_i, $\forall i \in \{1, \ldots, n\}$. The active domain of an attribute $A_j \in \mathcal{A}$ is then denoted as $Dom(A_j)$ and describes all the possible values for the feature A_j. The active domain $Dom(\mathcal{D})$ is then the cartesian product of all its attributes' active domains. Thus, each instance $x \in \mathcal{D}$ is from the active domain $Dom(\mathcal{D})$ and we write the value of x for attribute $A_j \in \mathcal{A}$ as $x(A_j)$. To define metrics for measuring group fairness, we first introduce protected attributes and patterns to match instances to certain groups similar to the definitions in [8,10].

Definition 1. ***Pattern.*** *Let $\mathcal{D}$ a dataset and $A = \{A_1, \ldots, A_q\} \subseteq \mathcal{A}$ a non-empty subset of the dataset attributes. Then, a tuple of attribute values $P = (a_1, \ldots, a_q) \in Dom(A)$ is a* pattern *over dataset $\mathcal{D}$. An instance $x \in Dom(\mathcal{D})$* satisfies *such a pattern P if the respective attribute values of x match the attribute values of P: If $\forall A_j \in A : x(A_j) = a_j$, then $x \vDash P$*

A classification model $\hat{M}$ can be trained on a dataset $\mathcal{D}$ labeled by the true classes $\mathcal{Y} = \{y_1, \ldots, y_z\}$ by $\mathcal{M} : \mathcal{D} \mapsto \mathcal{Y}$ to learn predicting the class $\hat{y} = \hat{M}(x)$ of any new input instance x. The model $\hat{M}$ serves as an approximation of the real mapping $M : Dom(\mathcal{D}) \mapsto \mathcal{Y}$ on $Dom(D)$. We also call $\hat{M}$ the predictor and in the following assume a binary classification model, i.e. $\mathcal{Y} = \{0, 1\}$, where $y = 1$ corresponds to the positive or favorable class label and $y = 0$ to the negative or unfavorable class label. In case of a score $\hat{M}(x) \in [0, 1]$ estimating the probability of an instance to belong to the favorable class a t-threshold rule [3] rule can be used to discretize the prediction as $\hat{y} = 1$ if $\hat{M}(x) \geq t$ or $\hat{y} = 0$ otherwise.

A pattern $P = (a_1, \ldots, a_q)$ partitions dataset $\mathcal{D}$ into two disjoint subgroups based on the protected attribute values. This partitions are the *protected* (P satisfied) and *unprotected* (P not satisfied) subgroups $\mathcal{D}_P = \{x \in \mathcal{D} \mid x \vDash P\}$ and $\mathcal{D}_{\bar{P}} = \{x \in \mathcal{D} \mid x \nvDash P\} = \mathcal{D} \setminus \mathcal{D}_P$, respectively. The different behavior regarding the prediction of the favorable or unfavorable class label on the subgroups by a classification model $\hat{M}$ is tested for fairness violations. The probabilities of a model $\hat{M}$ to predict the positive or negative class label are denoted as $\mathbb{P}(\hat{y} = 1)$ and $\mathbb{P}(\hat{y} = 0)$, respectively. Given a pattern P over dataset $\mathcal{D}$ the probability for instances from the protected subgroup to be predicted the class label $c \in \{0, 1\}$ is written as $\mathbb{P}(\hat{y} = c \mid x \in \mathcal{D}_P)$. Furthermore, the probability for a correct or wrong prediction of class c given the groundtruth class label g and one of the subgroups is expressed by $\mathbb{P}(\hat{y} = c \mid y = g, x \in \mathcal{D}_P)$. Notations for the probabilities of true classes (according to mapping M) and the unprotected group are analogous.

3.1 Subgroup Fairness Metrics

There exist various subgroup fairness metrics that mostly rely on the rates computed from confusion matrices [6,12] such as the positive predictive value (PPV) or TPR to estimate the chances for $\mathcal{D}_P$ and $\mathcal{D}_{\bar{P}}$. We do not focus on any specific group fairness metric but consider multiple of them as there are various,

sometimes opposing opinions on the justification of equal treatment. Instead, we define in the following three common fairness criteria that we will use in our evaluation to ensure a broad comparison.

Statistical parity (Def. 2) is a fairness definition based on the predicted outcome $\hat{y} = \hat{M}(x)$ and is satisfied if $\mathcal{D}_P$ has the same probability of getting a positive prediction ($\hat{y} = 1$) from the model as $\mathcal{D}_{\bar{P}}$ [12]. A fair classifier predicts the favorable label with a probability independent from the protected attribute values but the bias against instances belonging to multiple protected groups might be magnified [11]. Subgroup fairness based on equal opportunity (Def. 3) is achieved if the TPR of $\mathcal{D}_P$ is equal to the TPR of $\mathcal{D}_{\bar{P}}$. Regardless of their subgroup membership, the chance for each individual $x \in \mathcal{D}$ to get a positive prediction if they actually belong to the favorable class should be the same. As a consequence of Eq. 2, every individual from the subgroup should have the same probability of being assigned the unfavorable class label if they actually belong to the favorable class. The equalized odds subgroup fairness (Def. 4) is a generalization of equal opportunity as it requires the equality of the TPRs and FPRs of the both subgroups. In addition to the equal chance of getting a correct positive prediction also the chance of being incorrectly assigned the favorable class label has to be equal between the protected and unprotected subgroups.

Definition 2. ***Statistical parity.*** *Let $\mathcal{D}$ a dataset. A classifier $\hat{M}$ satisfies* statistical parity *wrt. a pattern P over $\mathcal{D}$ if:*

$$\mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_P) = \mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_{\bar{P}}) \tag{1}$$

Definition 3. ***Equal opportunity.*** *Let $\mathcal{D}$ a dataset and M the groundtruth mapping. A classifier $\hat{M}$ satisfies* equal opportunity *wrt. a pattern P over $\mathcal{D}$ if*

$$\mathbb{P}(\hat{y} = 1 \mid y = 1,\ x \in \mathcal{D}_P) = \mathbb{P}(\hat{y} = 1 \mid y = 1,\ x \in \mathcal{D}_{\bar{P}}). \tag{2}$$

Definition 4. ***Equalized odds.*** *Let $\mathcal{D}$ a dataset, M the groundtruth mapping and $g \in \{0, 1\}$. A classifier $\hat{M}$ satisfies* equalized odds *wrt. pattern P over $\mathcal{D}$ if*

$$\mathbb{P}(\hat{y} = 1 \mid y = g,\ x \in \mathcal{D}_P) = \mathbb{P}(\hat{y} = 1 \mid y = g,\ x \in \mathcal{D}_{\bar{P}}) \tag{3}$$

Commonly, the strict equality of fairness definitions is relaxed to accept also similar chances for predictions by $\hat{M}$ as fair, e.g., by ϵ-differential fairness definitions [4,9]. We prefer a simpler relaxation as provided by the "AI Fairness 360" toolkit [1] that relies on the difference between the probabilities for $\mathcal{D}_P$ and $\mathcal{D}_{\bar{P}}$ (Table 1). The fairness of $\hat{M}$ wrt. stat. parity and eq. opportunity is calculated as the difference between the probabilities given $x \in \mathcal{D}_{\bar{P}}$ or $x \in \mathcal{D}_P$. As equalized odds (Eq. 3) requires the equality of two probabilities, the average of the probability differences denoted as F_{aod} in Table 1 is calculated. Each metric $F \in \{F_{spd}, F_{eod}, F_{aod}\}$ yields a value in $[-1, 1]$. If $F = 0$, the evaluated classifier $\hat{M}$ is considered perfectly fair wrt. P and the fairness definition as the confusion matrix rates to estimate the probabilities coincide for $\mathcal{D}_P$ and $\mathcal{D}_{\bar{P}}$ (i.e., under the same conditions the prediction is independent of the membership in $\mathcal{D}_P$ or $\mathcal{D}_{\bar{P}}$). A value $F > 0$ corresponds to discrimination against individuals in $\mathcal{D}_P$ or favoritism of individuals in $\mathcal{D}_{\bar{P}}$ by $\hat{M}$ and $F < 0$ indicates the opposite.

Definition	Fairness Metric
Statistical parity	$F_{spd} = \mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_{\bar{P}}) - \mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_P)$
Eq. opportunity	$F_{eod} = \mathbb{P}(\hat{y} = 1 \mid y = 0, x \in \mathcal{D}_{\bar{P}}) - \mathbb{P}(\hat{y} = 1 \mid y = 0, x \in \mathcal{D}_P)$
Equalized odds	$F_{aod} = \frac{1}{2}\big[\mathbb{P}(\hat{y} = 1 \mid y = 0, x \in \mathcal{D}_{\bar{P}}) - \mathbb{P}(\hat{y} = 1 \mid y = 0, x \in \mathcal{D}_P)$ $+ \mathbb{P}(\hat{y} = 1 \mid y = 1, x \in \mathcal{D}_{\bar{P}}) - \mathbb{P}(\hat{y} = 1 \mid y = 1, x \in \mathcal{D}_P)\big]$

Table 1. Subgroup Fairness Metrics

4 Automatic Subgroup Detection

Assigning instances of a dataset to meaningful groups that mirror high similarities between the instances is challenging. Clustering algorithms provide unsupervised techniques to compute such a grouping $\mathcal{C}$, called *clustering*, that assigns the instances $x \in \mathcal{D}$ to clusters $C_1, \ldots, C_k$. Each pair $x, y \in C_i$ for $i \in \{1, \ldots, k\}$ shares some similarity as defined by the type of clustering, the algorithm parameters and the similarity/distance measure. For example, a centroid-based clustering expresses similarity by cluster membership wrt. the proximity to the computed centroids representing the clusters and a density-based clustering distinguishes dense regions of instances, that are considered the clusters, from the sparse regions, which are marked as containing outliers. With the notion of a clustering, we automatically evaluate the fairness of a classification model with our previous fairness metrics in two ways. A clustering $\mathcal{C}$ specifies a set of clustering-based patterns $P^{\mathcal{C}}$ (Def. 5) defining $\mathcal{D}_P$ and $\mathcal{D}_{\bar{P}}$ according to the established clusters. The fairness of a classifier $\hat{M}$ can then be evaluated for the subgroups of instances that comply to $\mathcal{C}$. To this end, the cluster labels of the instances $x \in \mathcal{D}$ are added as an artificial attribute $A_{\mathcal{C}}$ to $\mathcal{D}$. For each pattern $r = P_i^{\mathcal{C}} \in P^{\mathcal{C}}$ we can assess the fairness of $\hat{M}$ with any of the mentioned fairness metrics by comparing the treatment of the clustering-based protected and unprotected subgroup $\mathcal{D}_r = \{x \in \mathcal{D} \mid x \vDash P_i^{\mathcal{C}}\}$ and $\mathcal{D}_{\bar{r}} = \{x \in \mathcal{D} \mid x \nvDash P_i^{\mathcal{C}}\}$, respectively. The metric values are aggregated over all pairs of subgroups $\mathcal{D}_r$ and $\mathcal{D}_{\bar{r}}$ as defined by the patterns (i.e. over all clusters).

Definition 5. ***Clustering-based pattern.*** *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ a clustering of dataset $\mathcal{D}$ with an attribute set $\mathcal{A} \cup \{A_{\mathcal{C}}\}$ that was extended by the attribute $A_{\mathcal{C}}$ of the clustering labels of $\mathcal{C}$. We call $P_i^{\mathcal{C}} = (i)$ a* clustering-based pattern *over $\mathcal{D}$ and denote the set of all clustering-based patterns over $\mathcal{D}$ as $P^{\mathcal{C}} = \{P_i^{\mathcal{C}}\}_{i=1}^{k}$. An instance $x \in \mathcal{D}$ satisfies $P_i^{\mathcal{C}}$ if x belongs to cluster $C_i \in \mathcal{C}$.*

Furthermore, our system extracts more general patterns from $\mathcal{C}$ to perform the subgroup fairness analysis. We use the cluster feature entropy [2] to identify dominant features in $C_1, \ldots, C_k$ from which patterns are extracted. The cluster feature entropy $H_{i,j}$ quantifies the distribution of values for attribute A_j in cluster C_i and is calculated for each cluster and feature separately. An entropy value $H_{i,j}$ close to zero indicates a single dominant value at attribute A_j in C_i

whereas high values indicate a frequent occurrence of multiple values. A uniform distribution of all values across the cluster has maximal entropy.

Definition 6. ***Normalized feature entropy.*** *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ a clustering of dataset $\mathcal{D}$ with attributes $\mathcal{A} = \{A_1, \ldots, A_p\}$. We define the* normalized cluster feature entropy *for cluster $C_i \in \mathcal{C}$ and feature $A_j \in \mathcal{A}$ where $N_i = |C_i|$ and $N_{i,j,v} = |\{x \in C_i \mid x(A_j) = v\}|$ as*

$$H_{i,j} = -\frac{1}{\log_2 |Dom(A_j)|} \cdot \sum_{v \in Dom(A_j)} \frac{N_{i,j,v}}{N_i} \cdot \log_2 \left(\frac{N_{i,j,v}}{N_i} \right) \tag{4}$$

However, it is impossible to define an appropriate global entropy threshold t when using the definition of Cabrera et al. [2] as it does not account for the varying sizes of the active domains $\mathcal{A}$. As a consequence, it often fails to classify non-dominant features with larger active domains as such in clusters without a clear dominant value but potentially multiple of them when t was tuned for smaller active domains. To improve their definition, we also normalize the entropy by the logarithm $\log_2(|Dom(A_j)|)$ of the number of possible values for feature A_j in Def. 6. This ensures entropy values between 0 and 1 such that t can be picked independently of the size of the active domain of a feature.

For example, consider the three value distributions (frequencies) $a = [0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.8]$, $b = [0.1, 0.1, 0.1, 0.1, 0.6]$ and $c = [0.25, 0.25, 0.5]$, respectively. The first distribution a clearly shows a dominant feature in the cluster (80%), b also expresses a single value making up most of the feature but less significantly (60%), and c represents a scenario where still one value occurs more often than others but the feature is not really dominant. The feature entropies are $H_a \approx 1.32$, $H_b \approx 1.77$ and $H_c = 1.5$, respectively. Other than expected, the H_b does not reflect a dominant feature and, in contrast, H_c is lower although the distribution c does not dominate the feature as much as b. Instead of this misleading values, one can normalize them to obtain $\frac{H_a}{\log_2 9} \approx 0.42$, $\frac{H_b}{\log_2 5} \approx 0.76$ and $\frac{H_c}{\log_2 3} \approx 0.95$ accounting for the feature diversity.

Definition 7. ***Entropy-based pattern.*** *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ a clustering of dataset $\mathcal{D}$ with attribute set $\mathcal{A} = \{A_1, \ldots, A_p\}$ and threshold $t \geq 0$. We then define an* entropy-based pattern *over $\mathcal{D}$ as $P_i^t = (a_1, \ldots, a_q) \in Dom(A^i)$ where $A^i = \{A_j \in \mathcal{A} \mid H_{i,j} < t\}$ contains the* dominant features *of cluster $C_i \in \mathcal{C}$ and $a_j \in Dom(A_j)$ is the most frequent or* dominant value *of $A_j \in A^i$ in C_i. The set of entropy-based patterns for all clusters of $\mathcal{C}$, $\{P_i^t\}_{i=1}^k$, is denoted as P^t.*

From all dominant features of a cluster, one pattern is created for each cluster and used for the fairness metric calculation: all features A_j with $H_{i,j} \leq t$ are collected in a subset of attributes $A^i = \{A_j \in \mathcal{A} \mid H_{i,j} < t\}$ for each $C_i \in \mathcal{C}$. An entropy-based pattern P_i^t (Def. 7) is then derived from each C_i comprising the most frequent value of each dominant attribute $A_j \in A^i$. In contrast to a clustering-based pattern $P_i^{\mathcal{C}}$, that is satisfied by an instance $x \in \mathcal{D}$ if x belongs to C_i, an entropy-based pattern P_i^t is satisfied by x according to Def. 1. Hence, an entropy-based pattern is evaluated wrt. specific attribute values as determined

by $\mathcal{C}$ and t whereas clustering-based patterns are evaluated value-agnostic and specific attribute values are considered only implicitly via the computation of $\mathcal{C}$. For a hard partitional clustering, where each instance belongs to exactly one of the pairwise disjoint clusters, it holds that $\mathcal{D}_{P_i} \cap \mathcal{D}_{P_j} = \emptyset$ for each pair of clustering-based patterns $P_i, P_j \in P^{\mathcal{C}}$. However, two entropy-based patterns P_i^t and P_j^t extracted from different clusters $C_i, C_j \in \mathcal{C}$ might coincide as the clusters might have the same dominant features and values, i.e., $A^i = A^j$ and $P_i^t = P_j^t$. For our evaluation, we remove the duplicate entropy-based patterns before the subgroup fairness assessment and report the duplication rate.

5 Results

We evaluated our system on three fairness-related datasets: ProPublica's COMPAS dataset [1], the South German Credit dataset [2] and the Medical Expenditure Panel Survey [3] dataset of the year 2015 (panel 19). We refer to them as COMPAS, Credit and MEPS19, respectively. The COMPAS data ($n = 6172$, $p = 7$) was taken as provided in the FairVis [2] repository incl. predictions. We trained LightGBM classifiers (gradient boosting decision tree) for the Credit ($n = 1000$, $p = 20$) and MEPS dataset ($n = 15830$, $p = 40$). Our system first preprocessed the data by encoding categorical features and min-max normalizing before clustering. To analyze the suitability of clustering models for our subgroup detection task, we compared the proposed method for the following (subspace) clustering techniques of different types: k-Means, DBSCAN, OPTICS, Spectral Clustering, SLINK, Ward, BIRCH, SSC-BP, SSC-OMP, and EnSC. We tested a small set of parameter values individually on each algorithm and dataset. The fairness metrics producing the best combination of fairness violation indication and clustering performance was then reported. We selected the run that maximizes the product of silhouette score $S_{\mathcal{C}}$ and mean absolute error of the clustering-induced subgroup prediction accuracy $Acc_{\hat{M}}(\mathcal{D}_{P_i^{\mathcal{C}}})$, $i = 1, \ldots, k$, as compared to the global accuracy $Acc_{\hat{M}}(\mathcal{D})$: $\arg\max_{\mathcal{C}} \left(S_{\mathcal{C}} \cdot \frac{1}{|\mathcal{C}|} \cdot \sum_{P_i^{\mathcal{C}} \in P^{\mathcal{C}}} \left| Acc_{\hat{M}}(\mathcal{D}_{P_i^{\mathcal{C}}}) - Acc_{\hat{M}}(\mathcal{D}) \right| \right)$

The experimental results are shown in Table 2 - 4. Each table represents a dataset and both subgroup detection methods with one row for the best run. The columns display the mean (Avg), standard deviation (Std) and absolute mean (Abs) values across all clustering- or entropy-induced subgroups as measured by the fairness metrics. The clustering model with the highest absolute mean is highlighted for each fairness criterion by bold numbers for the three reported quantities. On the COMPAS dataset (Table 2) the best results were obtained for SLINK throughout all fairness metrics and both subgroup detection methods. Especially the entropy-based subgroups detected by SLINK clearly outperformed in the absolute mean of the metric values when compared to the other clustering

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29
[3] https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-183

Algorithm	Statistical Parity			Equal Opportunity			Equalized Odds		
	Avg	Std	Abs	Avg	Std	Abs	Avg	Std	Abs
k-Means	−0.0074	0.3017	0.2513	0.0592	0.3100	0.2454	−0.0071	0.2877	0.2315
DBSCAN	−0.1446	0.1914	0.1992	−0.0880	0.1516	0.1491	−0.1242	0.1763	0.1733
OPTICS	−0.0592	0.1778	0.1580	0.0102	0.1790	0.1439	−0.0466	0.1651	0.1415
Spectral	0.0328	0.3558	0.3081	0.1388	0.3755	0.3198	0.0254	0.3305	0.2749
SLINK	**0.0343**	**0.4057**	**0.3580**	**0.2474**	**0.4373**	**0.4217**	**0.0767**	**0.3457**	**0.2821**
Ward	−0.0376	0.2711	0.2233	0.0163	0.2487	0.1859	−0.0316	0.2460	0.1984
BIRCH	−0.1032	0.1744	0.1710	−0.0506	0.1376	0.1274	−0.0810	0.1449	0.1411
SSC-OMP	−0.0160	0.2115	0.1725	0.0205	0.1679	0.1386	−0.0167	0.1802	0.1460
SSC-BP	−0.0993	0.2078	0.1949	−0.0490	0.1699	0.1467	−0.0851	0.1863	0.1712
EnSC	−0.1356	0.1978	0.2018	−0.0798	0.1538	0.1470	−0.1172	0.1808	0.1765
k-Means	−0.0985	0.2323	0.2022	−0.0459	0.2076	0.1700	−0.0918	0.2269	0.1891
DBSCAN	−0.2115	0.1860	0.2265	−0.1332	0.1417	0.1630	−0.1985	0.1875	0.2143
OPTICS	−0.1517	0.1830	0.1850	−0.0842	0.1509	0.1498	−0.1263	0.1737	0.1580
Spectral	−0.0601	0.2997	0.2496	0.0290	0.3288	0.2477	−0.0581	0.2728	0.2188
SLINK	**−0.0158**	**0.4009**	**0.3487**	**0.1725**	**0.4312**	**0.3761**	**0.0138**	**0.3524**	**0.2783**
Ward	−0.1211	0.2274	0.2081	−0.0748	0.1620	0.1517	−0.1190	0.2148	0.1920
BIRCH	−0.1860	0.1854	0.2069	−0.1010	0.1417	0.1423	−0.1665	0.1794	0.1885
SSC-OMP	−0.0781	0.1920	0.1664	−0.0305	0.1525	0.1295	−0.0724	0.1677	0.1456
SSC-BP	−0.0993	0.2128	0.1861	−0.0566	0.1502	0.1249	−0.0914	0.1968	0.1629
EnSC	−0.1839	0.1842	0.2117	−0.1167	0.1358	0.1500	−0.1713	0.1795	0.1959

Table 2. Clustering- (top) & Entropy-based (bottom) Subgroup Fairness COMPAS

Algorithm	Statistical Parity			Equal Opportunity			Equalized Odds		
	Avg	Std	Abs	Avg	Std	Abs	Avg	Std	Abs
k-Means	−0.0096	0.2210	0.2062	0.0120	0.0738	0.0628	−0.0665	0.1838	0.1556
DBSCAN	−0.0059	0.2911	0.2097	0.0578	0.1562	0.1000	−0.0876	0.2957	0.1928
OPTICS	−0.0051	0.2616	0.2029	0.0253	0.0524	0.0317	0.0035	0.1379	0.0988
Spectral	−0.0323	0.3207	0.2480	0.0761	0.3091	0.1340	−0.1124	0.3216	0.2560
SLINK	**−0.0120**	**0.4338**	**0.3123**	**0.3691**	**0.5306**	**0.4103**	**0.0006**	**0.4249**	**0.3332**
Ward	−0.0006	0.1744	0.1442	0.0055	0.0980	0.0655	−0.0320	0.1569	0.1232
BIRCH	−0.0002	0.1787	0.1509	0.0051	0.0967	0.0632	−0.0361	0.1580	0.1291
SSC-OMP	0.0032	0.0461	0.0375	0.0033	0.0277	0.0232	−0.0011	0.0279	0.0229
SSC-BP	−0.0117	0.1286	0.0898	−0.0003	0.0428	0.0299	−0.0097	0.0514	0.0359
EnSC	−0.0217	0.1377	0.1070	−0.0034	0.0298	0.0227	−0.0239	0.0824	0.0677
k-Means	−0.0558	0.2130	0.1940	−0.0094	0.0591	0.0469	−0.0969	0.2065	0.1694
DBSCAN	−0.0171	0.2516	0.1717	0.0907	0.2258	0.1434	−0.0967	0.2829	0.1814
OPTICS	−0.0878	0.3209	0.2763	−0.0122	0.0694	0.0575	**−0.2370**	**0.3103**	**0.3179**
Spectral	−0.0504	0.3193	0.2469	0.0631	0.3119	0.1282	−0.1268	0.3190	0.2617
SLINK	**−0.0108**	**0.4339**	**0.3135**	**0.3605**	**0.5374**	**0.4017**	−0.0293	0.4155	0.3033
Ward	−0.0164	0.1915	0.1682	0.0080	0.1180	0.0786	−0.0845	0.2022	0.1661
BIRCH	−0.0354	0.2032	0.1775	0.0023	0.1177	0.0754	−0.0989	0.2122	0.1783
SSC-OMP	0.0407	0.0612	0.0499	0.0048	0.0173	0.0151	0.0223	0.0324	0.0240
SSC-BP	−0.0307	0.1802	0.1326	−0.0010	0.0352	0.0253	−0.0824	0.2285	0.1469
EnSC	−0.0407	0.1384	0.0857	−0.0001	0.0333	0.0232	−0.0684	0.1997	0.1152

Table 3. Clustering- (top) & Entropy-based (bottom) Subgroup Fairness Credit

results with absolute mean values from $\approx$ 0.28-0.42. The eq. opportunity mean values revealed a skew towards the detection of mainly discriminated subgroups or subgroups with a higher degree of discrimination than the degree of favorization for the other subgroups by SLINK in both detection methods. The other fairness metrics were balanced between discriminated and favored subgroups as indicated by the mean metric values close to zero. The subspace clustering algorithms showed no improvement over the conventional clustering algorithms and the spectral clustering also performed quite good on the COMPAS dataset. The duplication rate of the entropy-induced subgroups was between 0 and 0.5.

Table 3 displays again outstanding results (Abs $\approx$ 0.30-0.41) of SLINK across both detection methods and all fairness criteria. Only OPTICS achieved a higher absolute mean value of 0.3179 under eq. odds and entropy-induced subgroups on the Credit dataset. For the SLINK clustering, again we observed an even more significant skew towards the detection of discriminated subgroups over favored ones for the eq. opportunity criterion. The OPTICS clustering, in contrast, showed a skew towards the favored subgroups with an average eq. odds value of -0.2370 across the clustering-induced subgroups. The spectral clustering also

Algorithm	Statistical Parity			Equal Opportunity			Equalized Odds		
	Avg	Std	Abs	Avg	Std	Abs	Avg	Std	Abs
k-Means	−0.0279	0.3180	0.2050	0.2343	0.3675	0.4002	0.0862	0.3058	0.2630
DBSCAN	−0.0566	0.4154	0.3009	0.5054	0.0395	0.5054	0.1176	0.2836	0.2573
OPTICS	0.1323	0.0002	0.1323	**0.5806**	**0.0118**	**0.5806**	0.3097	0.0058	0.3097
Spectral	−0.6962	0.4504	0.7958	−0.2681	0.4133	0.4830	**−0.5098**	**0.3938**	**0.6090**
SLINK	**−0.7121**	**0.4276**	**0.8031**	−0.1229	0.4810	0.4834	−0.4814	0.3629	0.5461
Ward	0.0029	0.2636	0.1810	0.2988	0.3222	0.4111	0.1311	0.2565	0.2549
BIRCH	−0.0336	0.2998	0.2068	0.2503	0.3432	0.4003	0.0930	0.2835	0.2625
SSC-OMP	0.1175	0.0585	0.1306	0.4805	0.2388	0.5339	0.2613	0.1299	0.2904
SSC-BP	0.1174	0.0623	0.1321	0.4805	0.2415	0.5350	0.2613	0.1321	0.2913
EnSC	0.1127	0.0662	0.1293	0.4429	0.2386	0.4927	0.2419	0.1311	0.2700
k-Means	0.0082	0.3082	0.2031	0.3665	0.3484	0.4679	0.1574	0.3043	0.2981
DBSCAN	−0.0866	0.3894	0.2709	0.5370	0.0027	0.5370	0.1264	0.2910	0.2661
OPTICS	0.1322	0.0001	0.1322	**0.5826**	**0.0002**	**0.5826**	0.3106	0.0001	0.3106
Spectral	−0.7765	0.2673	0.7902	−0.2653	0.4100	0.4788	**−0.5528**	**0.2827**	**0.5818**
SLINK	**−0.7906**	**0.2507**	**0.8043**	−0.1163	0.4895	0.4900	−0.5222	0.2675	0.5262
Ward	0.0306	0.2562	0.1889	0.3955	0.3076	0.4796	0.1832	0.2535	0.2930
BIRCH	−0.0344	0.3370	0.2373	0.3123	0.3866	0.4807	0.1148	0.3271	0.3180
SSC-OMP	0.1310	0.0016	0.1310	0.5341	0.0010	0.5341	0.2656	0.1119	0.2866
SSC-BP	0.1408	0.0102	0.1408	0.5400	0.0085	0.5400	0.2571	0.1404	0.2893
EnSC	0.1405	0.0092	0.1405	0.5411	0.0083	0.5411	0.2962	0.0061	0.2962

Table 4. Clustering- (top) & Entropy-based (bottom) Subgroup Fairness MEPS19

performed well as measured by stat. parity and eq. odds for both detection methods. For the Credit dataset the observed performance of the subspace clustering algorithms was weaker than the other algorithms. We observed a duplication rate of 0 for all reported trials except for the SCC-OMP model (single duplication).

The MEPS19 dataset (Table 4) yielded more variety regarding the best performance. We observed for both detection methods the best performance in stat. parity for SLINK (≈ 0.80), in eq. opportunity for OPTICS (≈ 0.58) and in eq. odds for spectral clustering (≈ 0.60 and 0.58). The spectral clustering produced similarly good results as the SLINK clustering for each of the detection methods. In contrast to the other two datasets, our experiments showed more shift towards favored or discriminated subgroups as detected by any of the computed clusterings for the MEPS19 dataset. Furthermore, we observed (nearly) equality of the absolute value of the mean and the absolute mean metric value for selected models. This indicates that few or no favored subgroups were detected by the clustering algorithm if the skew occurred towards discriminated subgroups and vice versa. This behavior was often observed for the subspace clustering algorithms, that again did not perform differently than the conventional algorithms. Only for SSC-BP, we observed a duplication rate of 0.4 whereas the other algorithms had maximally one collision on the entropy-induced subgroups.

6 Conclusion & Future Work

In our research we have proposed two techniques to identify subgroups in data to perform a subgroup fairness analysis on. The experimental results proved the ability of our clustering- and entropy-based approach to detect subgroups in datasets on which a given classifier violates common fairness criteria, namely statistical parity, equal opportunity and equalized odds. We found a strong overall performance when employing the SLINK clustering algorithm in our subgroup detection methods as it identifies subgroups with a high deviation from what would be considered fair. Future research could investigate on relationships between classification models and the subgroup detection performance or extend

the proposed subgroup detection. We currently work on the integration of our method into a graphical, web-based tool allowing users to perform an automatic subgroup fairness analysis on their dataset and classifier in a user-friendly manner. With the help of the fairness analysis tool, we want to provide deeper insights into the composition of the detected subgroups straightforward to users.

References

1. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias (10 2018), https://arxiv.org/abs/1810.01943
2. Cabrera, A.A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., Chau, D.H.: FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. 2019 IEEE Conference on Visual Analytics Science and Technology (VAST) (10 2019)
3. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G., Cosentini, A.C.: A clarification of the nuances in the fairness metrics landscape. Scientific Reports **12**(1), 1–21 (2022)
4. Foulds, J.R., Islam, R., Keya, K.N., Pan, S.: An Intersectional Definition of Fairness. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE). pp. 1918–1921. IEEE (2020)
5. Gleicher, M., Barve, A., Yu, X., Heimerl, F.: Boxer: Interactive comparison of classifier results. In: Computer Graphics Forum. vol. 39, pp. 181–193. Wiley Online Library (2020)
6. Hertweck, C., Heitz, C.: A Systematic Approach to Group Fairness in Automated Decision Making. In: 2021 8th Swiss Conference on Data Science (SDS). pp. 1–6. IEEE (2021)
7. Johnson, B., Brun, Y.: Fairkit-learn: A Fairness Evaluation and Comparison Toolkit. 44th International Conference on Software Engineering Companion (ICSE '22 Companion) (2022)
8. Li, J., Moskovitch, Y., Jagadish, H.: DENOUNCER: Detection of Unfairness in Classifiers. Proceedings of the VLDB Endowment **14**(12), 2719–2722 (2021)
9. Morina, G., Oliinyk, V., Waton, J., Marusic, I., Georgatzis, K.: Auditing and Achieving Intersectional Fairness in Classification Problems. arXiv preprint arXiv:1911.01468 (2019)
10. Pastor, E., de Alfaro, L., Baralis, E.: Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence. In: Proceedings of the 2021 International Conference on Management of Data. pp. 1400–1412 (2021)
11. Teodorescu, M.H., Morse, L., Awwad, Y., Kane, G.C.: Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. MIS Quarterly **45**(3) (2021)
12. Verma, S., Rubin, J.: Fairness Definitions Explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 1–7. IEEE (2018)
13. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J.: The what-if tool: Interactive probing of machine learning models. IEEE transactions on visualization and computer graphics **26**(1), 56–65 (2019)