Identifying Biased Subgroups in Ranking and Classification

Identifying Biased Subgroups in Ranking and Classification

Eliana Pastor
eliana.pastor@polito.it
Politecnico di Torino, Italy

Luca de Alfaro
luca@ucsc.edu
UC Santa Cruz, USA

Elena Baralis
elena.baralis@polito.it
Politecnico di Torino, Italy

ABSTRACT

When analyzing the behavior of machine learning algorithms, it is important to identify specific data subgroups for which the considered algorithm shows different performance with respect to the entire dataset. The intervention of domain experts is normally required to identify relevant attributes that define these subgroups.

We introduce the notion of divergence to measure this performance difference and we exploit it in the context of (i) classification models and (ii) ranking applications to automatically detect data subgroups showing a significant deviation in their behavior. Furthermore, we quantify the contribution of all attributes in the data subgroup to the divergent behavior by means of Shapley values, thus allowing the identification of the most impacting attributes.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Mathematics of computing** → **Exploratory data analysis**.

KEYWORDS

Fairness, machine learning, fair AI, bias in machine learning, bias detection, equitable AI.

ACM Reference Format:
Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Identifying Biased Subgroups in Ranking and Classification. In *Workshop on Measures and Best Practices for Responsible AI at ACM KDD 2021 (Responsible AI @ KDD 2021).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Machine learning models and automated-decision making procedures are becoming more and more pervasive, thus a growing interest is arising on the careful understanding of their behavior [11, 25]. A relevant step in the explanation of the outcome of machine learning algorithms is the identification of data subgroups in which the considered algorithm may show a different, and potentially anomalous, behavior. The identification of peculiar behaviors of data subgroups finds important applications in the KDD pipeline, ranging from model validation and testing [7, 21] to the evaluation of model fairness [5, 7]. In particular, societal bias [4] is becoming a growing concern and researchers are increasingly working on measuring and ensuring fairness in machine learning. To this aim, human experts are frequently required to manually identify sensitive attributes (e.g., gender, ethnicity) or problematic data subgroups.

Responsible AI @ KDD 2021, August 14-18, 2021, Virtual Event

Data subgroup	FPR	FNR	Δ_{FPR}	Δ_{FNR}
Entire dataset	0.09	0.70	0.00	0.00
age<25, #prior>3, sex=Male	0.68	0.36	0.59	-0.34
age>45, charge=M, race=Cauc	0.01	1.00	-0.08	0.30

Table 1: Example of patterns in the COMPAS dataset, along with their false-positive rate (FPR), false-negative rates (FNR) and divergence Δ.

The behavior of machine learning algorithms is frequently evaluated by means of global metrics, which consider the performance of the algorithm on a global level, for the entire dataset, or for specific class labels. Differently, in this paper we propose the concept of *divergence* as a measure of the difference in statistics (e.g., false positive rate) between the behavior of the machine learning algorithm on a data subgroup and on the entire dataset. Data subgroups showing a significant deviation in their behavior are automatically identified by our approach, which characterizes data subgroups by means of a combination of attribute values, denoted in the paper as *patterns*, or *itemsets*.

To illustrate the concept of divergence, consider the COMPAS dataset [3], in which a score measuring recidivism risk is assigned to criminal defendants by a proprietary algorithm. In this case, the positive class corresponds to high recidivism scores. Table 1 shows the false-positive (FPR) and false-negative (FNR) rates occurring in the entire dataset (first row) and in the data subgroups characterized by the highest FPR (second row) and FNR (third row) divergence. For example, the pattern (*age>45, charge=M, race=Cauc*) shows a high Δ_{FNR} divergence. Hence, instances belonging to this data subgroup will be wrongly assigned to the negative class with a higher rate with respect to the entire dataset.

We propose a general framework for divergence computation, that allows the automatic identification of problematic data subgroups both in classification and ranking problems. Our approach builds on well-known itemset mining algorithms (e.g., FP-growth [12]) and allows the efficient identification of all the problematic patterns above a (low) frequency threshold. Moreover, given a specific pattern, we exploit the notion of Shapley value [18] to quantify the contribution of each attribute value to the pattern divergence. For example, as will be shown in the following sections, the attribute value *age>45* is contributing the most to the divergent behavior of the pattern discussed before.

2 RELATED WORK

Existing techniques for analyzing data subgroups include both supervised and unsupervised techniques. Supervised techniques rely on domain experts and users to identify the subgroups of interest. Several tools analyze performance over data subgroups specified

by the user of a classification model for validation purposes [2, 13]. Many efforts have been devoted to detecting and mitigating bias in classification tasks. Several approaches evaluate if different treatment or performance occur on groups determined by some sensitive or protected attributes [9, 10, 14, 16]. Recently, researches focused the attention on fairness in rankings [25]. Different works propose measures and mechanisms to audit ranking outputs and mitigate bias over protected groups [6, 22–24]. The analysis of group fairness in both the classification and the ranking tasks generally assumes that the sensitive attributes (e.g. sex, race, age, degree of disability) that define the protected groups are known or specified a priori. We propose an approach for the automatic identification of critical subgroups treated differently by a generic model, be it a classifier or a ranker, without the a priori knowledge of the groups and attributes of interest. We concern ourselves on auditing differences in subgroups, rather than mitigation strategies, which may be application-dependent.

Several works have been proposed to automatically identify critical subgroups in the classification domain [5, 7, 21]. FairVIS [5] audits the fairness of classification models leveraging on a clustering-based subgroup identification technique. Fairness and performance metrics are evaluated on the identified clusters described by a few dominant features obtained via feature entropy. Differently, we exploit frequent pattern mining algorithms to identify frequent critical subgroups. The subgroups are obtained by slicing the attribute domains. Hence, the characterizing features are known and readily interpretable. Errudite [21] is an interactive system that enables data grouping for NLP error analysis using a domain-specific language. Differently from [21], our approach deals with structured data and slices the data by (discrete) attribute values.

Slice Finder [7, 8] automatically detects data slices in which a classification model performs poorly. Similarly to our approach, the data subgroups are identified by slicing via attribute values. Slice Finder defines a top-down lattice search to find the top-k critical slices. The data exploration, based on breadth-first traversal, stops when it reaches a subgroup characterized by a sufficiently large difference in performance which is statistically significant. However, this stopping criterion may prevent finding relevant critical subgroups, because the metrics used for assessing model performance on subgroups are typically non-monotone. Thus, from the critical behavior of a group, we cannot make assumptions on the behavior of its super/sub-groups. We propose a more comprehensive exploration by identifying all the subgroups adequately represented (i.e., above a frequency threshold) in the dataset. We then characterize the subgroups using the notion of Shapley values [18] to estimate the contribution of each attribute value to the subgroup divergence.

We introduced the notion of divergence, and the use of Shapley value to measure the contribution of attributes to divergence, in [17]. This paper extends the definitions to handle ranking systems, as well as general quantitative prediction functions.

3 EXAMPLE DATASETS

As running examples to illustrate the concepts, we use two well-known datasets. The first is the COMPAS dataset [3], which consists of defendants considered for release on parole. For each individual, the dataset contains personal data such as age range, race, gender, and data related to criminal history, such as number of prior offenses. The COMPAS dataset contains also a score that estimates the likelihood that a defendant commits another offense (recidivates) in the next two years. From this score, via comparison with a threshold, we can obtain a binary classification. For each defendant, the ground truth is known, so that the false-positive and false-negative rates of the classification can be computed. We are interested in characterizing the subgroups for which these rates deviate from the average.

The second dataset we consider is the Law School Dataset. The Law School Admission Council conducted a survey across 163 law schools in the United States in 1998 [20]. The resulting Law School Dataset contains information on 21,791 law students such as their entrance exam scores (LSAT), their grade-point average (GPA) collected prior to law school, and their normalized first year average grade (ZFYA), in addition to their race and sex. We use the dataset as prepared by [15]. In this dataset, we study how the average ZFYA score, and the average rank of students after the first year, vary across subgroups.

4 DIVERGENCE

We provide here the definition of *divergence*, which captures the difference between statistical measures computed on individual subgroups, versus the entire dataset, and we illustrate the divergence of various measures on our example datasets.

4.1 Datasets and itemsets

We consider datasets D in tabular form: there is a fixed set A of columns, and a set X of rows; every row x assigns value $x(a)$ to attribute $a \in A$. Thus, A is the schema of the dataset, and the rows X are the *instances*. We assume that every attribute $a \in A$ has a finite domain D_a. Thus, the dataset is discretized. For example, in our COMPAS dataset, an instance is a defendant, and the attributes are age range, gender, and so on.

An *item* α consists in a selection $a = c$ for an attribute $a \in A$ and a value $c \in D_a$. An *itemset* is a set of items, with each item involving a distinct attribute. For instance, in COMPAS, an itemset is {*age>45, race=Caucasian*}. The *support-set* $D(I) = \{x \in D \mid x \models I\}$ of an itemset I consists of the instances that satisfy I; the *support* of I is the fraction of dataset instances in $D(I)$, or $\sup(I) = \frac{|D(I)|}{|D|}$.

4.2 Itemset divergence

We are interested in identifying subgroups of data that behave differently, compared to the overall dataset, with respect to statistical measures. For instance, in classifiers we are interested in identifying data subgroups where the false-positive and false-negative rates differ from the average. In rankings, we are interested in data subgroups where the average rank deviates from the global one. We use itemsets to describe data subgroups, and we use an *outcome function* to capture the statistic of interest.

An *outcome function* $o : X \mapsto \{\bot\} \cup \mathbb{R}$ associates with each instance either a do-not-consider value $\bot$, or a real number. For a (possibly empty) dataset I, we define the *outcome* $o(I)$ *on* I via:

$$o(I) = E\{o(x) \mid x \models I, o(x) \neq \bot\} \ . \tag{1}$$

Itemset	Sup	Δ_{ZFYA}	t
LSAT>41.0, UGPA>3.5, race=White, sex=Female	0.03	0.4115	11.1
LSAT>41.0, UGPA>3.5, race=White	0.07	0.4063	16.8
LSAT>41.0, UGPA>3.5, race=White, sex=Male	0.04	0.4025	13.0
LSAT≤33.0, race=Black, sex=Male	0.02	-1.0257	21.2
LSAT≤33.0, UGPA≤3.0, race=Black, sex=Male	0.01	-1.0049	17.05
LSAT≤33.0, race=Black	0.05	-0.9787	33.3

Table 2: Top-3 itemsets with highest and lowest ZFYA divergence for the Law School Dataset. The support threshold is $s = 0.005$.

Itemset	Sup	Δ_{FPR}	t
age<25, #prior>3, sex=Male	0.02	0.594	6.1
age<25, #prior>3	0.02	0.527	5.7
age<25, stay=1w-3M, race=Afr-Am, sex=Male	0.02	0.306	3.8
	Sup	Δ_{FNR}	t
age>45, charge=M, race=Cauc	0.05	0.302	17.6
age>45, charge=M, #prior=0	0.04	0.302	10.4
age>45, charge=M, #prior=[1,3]	0.03	0.302	14.1

Table 3: Top-3 divergent patterns with respect to FPR and FNR for the COMPAS dataset. The support threshold is $s = 0.0175$.

If I is empty, $o(\emptyset)$ is the outcome of the complete dataset. We then define the *divergence* of I (with respect to outcome function o) to be:

$$\Delta_o(I) = o(I) - o(\emptyset) \ . \tag{2}$$

The divergence of an itemset captures the difference in behavior between the itemset, and the entire dataset, with respect to the outcome function under consideration. We illustrate, via examples, how different outcome functions enable the analysis of raw datasets, as well as the behavior of classifiers and ranking systems.

4.3 Attribute divergence

In many cases, one can take the outcome of an instance to be one of the quantitative attributes of the instance itself. For the Law School Dataset, the simplest choice consists in taking $o(x) = ZFYA(x)$, setting the outcome equal to the normalized first-year average of each student. Table 2 lists the three itemsets with greatest positive and negative divergence, among those with support at least $s = 0.005$, which corresponds to about 100 students. The table reports also the t-value of the divergence, computed according to Welch's t-test. We use a support limit both to provide a termination criterion for the divergence computation algorithm, as discussed in Section 4.6, and to exclude itemsets with such small support that the analysis is affected by statistical fluctuations. From the results, we see that the itemset with greatest positive divergence is {LSAT>41.0, UGPA>3.5, race=White, sex=Female}, for which the ZFYA-divergence is 0.41. The itemset with the greatest negative divergence is {LSAT≤33.0, race=Black, sex=Male}, for which the ZFYA score is on average lower by 1.03 compared with the dataset average. In Section 5, we will see how to analyze the contribution of each of the three items LSAT≤33.0, race=Black, and sex=Male, to the divergence of this itemset.

4.4 Classifier divergence

Divergence can also be applied to analyze classifier behavior. Given a classifier, let $p(x) \in \{\text{T}, \text{F}\}$ be the predicted value for an instance x, and let $t(x)$ be the true value (ground truth). In a classifier, it is often of interest to study the divergence of the false-positive rate (FPR) and false-negative rate (FNR). The variation of these rates across data subgroups gives an indication of how the subgroups are advantaged, or disadvantaged, by classifier errors. To capture the divergence of the false-positive rate, we use the outcome function:

$$o(x) = \begin{cases} \bot & \text{if } t(x) = \text{T} \\ 0 & \text{if } t(x) = \text{F and } p(x) = \text{F} \\ 1 & \text{if } t(x) = \text{F and } p(x) = \text{T} \end{cases}$$

for $x \in X$. Here, the outcome $\bot$ is used to exclude from the statistic the true-positives, so that the outcome $o(I)$ of an itemset I is its FPR. Outcome functions for capturing the FNR, true-positive rate, and so on, can be similarly defined.

In Table 3 we report the top-3 divergent patterns with respect to FPR and FNR, for a minimum support of $s = 0.0175$, equivalent to about 100 instances. An itemset with positive divergence for FPR is an itemset consisting of defendants that are incorrectly predicted to recidivate at a rate higher than the average for the dataset. We see that young males, with prior crimes, are the defendants most often falsely predicted to recidivate. Conversely, old Caucasian males are the most frequent instances incorrectly predicted *not* to recidivate.

4.5 Ranking divergence

In a ranking system, every instance x has a *rank* $i(x) \in \mathbb{N}_{>0}$, where $i = 1$ is the top rank. It is natural to define the outcome function o via a *rank valuation function* $\gamma : \mathbb{N}_{>0} \mapsto \mathbb{R}$, where $\gamma(i)$ represents the value, to an instance, of being ranked in position i. We define the outcome of instance $x \in X$ via:

$$o(x) = \gamma(r(x)) \tag{3}$$

The outcome $o(I)$ of an itemset I would then correspond to the average value an instance in I receives from being ranked.

As an example, consider admissions to a university. If applicants are ranked, and the top k admitted, we can take $\gamma(i) = 1$ for $i \leq k$ and $\gamma(i) = 0$ otherwise. The outcome $o(I)$ corresponds to the admission rate of I, that is, the fraction of applicants in I that are admitted, and the divergence $\Delta_o(I)$ would then represent how more, or less, likely applicants in I are to be admitted, compared with the general population. Notice that the use of a rank-value function γ, rather than simply the rank, is key to capturing the impact of the ranking on instances in top-k admissions.

As another example, returning to our Law School Dataset, assume that at the end of their first year, students internship applications are displayed to internship hosts sorted according to the first-year average grade (ZFYA) of a student. Assume that the benefit of being ranked in position i to the student is proportional

Itemset	Sup	Δ_γ	t
LSAT>41.0, UGPA>3.5, race=White, sex=Female	0.03	0.0206	8.7
LSAT>41.0, UGPA>3.5, race=White	0.07	0.0196	13.0
LSAT>41.0, UGPA>3.5, race=White, sex=Male	0.04	0.0189	9.9
LSAT≤ 33.0, race=Black, sex=Male	0.02	-0.0283	25.6
LSAT≤ 33.0, UGPA≤ 3.0, race=Black, sex=Male	0.01	-0.0280	21.0
LSAT≤ 33.0, UGPA≤ 3.0, race=Black	0.03	-0.0278	31.4

Table 4: Top-3 itemsets with highest and lowest divergence for the Law School Dataset, for the internship example, with $\gamma(i) = i^{-0.1}$. The support threshold is $s = 0.005$.

to $\gamma(i) = i^{-0.1}$; this type of relation between rank and benefit is common in search applications. Table 4 gives the itemsets with top and bottom divergence with respect to this benefit. We see that the itemset that derives the most benefit out of internships would be {LSAT>41.0, UGPA>3.5, race=White, sex=Female}, and the one deriving the least benefit would be {LSAT≤33.0, race=Black, sex=Male}.

4.6 Computing the divergence

We compute the outcome (1) of itemsets, and hence their divergence, by extending frequent-pattern mining algorithms [1, 12, 19].

The input to the algorithms is a *support threshold* s: only itemsets I with $\sup(I) \geq s$ are considered. The support threshold is chosen according to the minimum size of the data subgroup that one is interested in investigating. For each itemset I above the support threshold, its outcome $o(I)$ is computed by tallying, during itemset extraction, (a) the sum of all instance outcomes in I that are not $\perp$, and (b) the number of such instances, and then taking the ratio.

By extending frequent-pattern mining algorithms, we obtain algorithms that compute the divergence of all itemsets above the support size extremely efficiently. Even when there are 10^5 itemsets above the support threshold, they can be identified and their divergence computed in a matter of a dozen seconds or so.

5 ITEM CONTRIBUTION TO DIVERGENCE

Once one has identified the itemsets for which the relevant classification or rank statistics most deviate from the average, it is of interest to determine the contributions of the individual items to the divergence of the itemset. For instance, we see from Table 3 that the defendants most likely to be incorrectly predicted to recidivate are those in {age<25, #prior>3, sex=Male}; which one of the three items age<25, #prior>3, and sex=Male, is the most important? Of which fraction of the divergence 0.594 is each item responsible?

To answer this question, we use the notion of Shapley values [18]. In game theory, Shapley values are a way of distributing the value $v(1, 2, \ldots, N)$ of a team of N players, to each of the players $1, 2, \ldots, N$, in such a way that the sum of the player's values is equal to the team's value. We define the contribution $\Delta(\alpha \mid I)$ of item I to the divergence $\Delta(I)$ as the Shapley value of α in I, for value Δ:

$$\Delta(\alpha \mid I) = \sum_{J \subseteq I \setminus \{\alpha\}} \frac{|J|!(|I| - |J| - 1)!}{|I|!} \left[\Delta(J \cup \alpha) - \Delta(J)\right]. \quad (4)$$

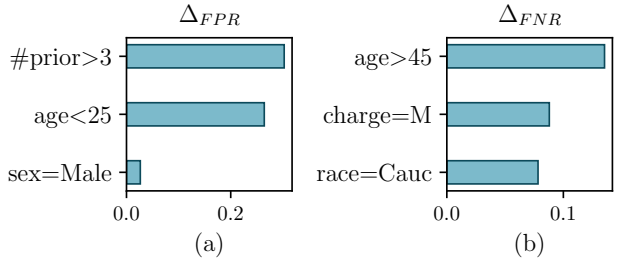

Figure 1: Contributions of individual items to the divergence of the COMPAS frequent patterns having greatest FPR and FNR divergence in Table 3.

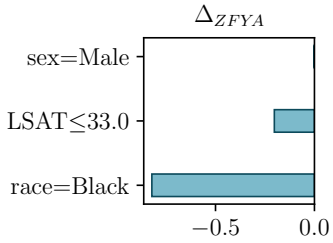

Figure 2: Contributions of individual items to the divergence of the frequent patterns having lowest ZFYA divergence for the Law School Dataset ($s = 0.005$).

Note that, if an itemset is above the support threshold, all its subsets also are, so all divergences of itemsets in (4) can be computed by our algorithm.

Figure 1 reports the influence of individual items to the itemsets with greatest FPR and FNR divergence. We see that sex=Male is responsible only for a small fraction of the FPR divergence of {age<25, #prior>3, sex=Male}, while age<25 and #prior>3 have effects of similar magnitude.

Figure 2 reports the results of a similar analysis for the itemset with lowest ZFYA divergence in the Law School Dataset. We see that race=Black is the predominant factor, with LSAT≤33.0 giving a minor contribution, and sex=Male a negligible one. The predominant role of race stands out as a warning signal, indicating that this negative rank divergence merits further investigation.

6 CONCLUSIONS

In this paper we presented a method for finding data subgroups that behave differently from the overall dataset in classification, ranking, or other automated prediction systems. At the core of the approach is the notion of *divergence*, which quantifies the difference in behavior, and which can be efficiently determined via data mining approaches. We believe our approach may provide a useful building block in strategies to mitigate algorithmic bias.

ACKNOWLEDGMENTS

This work has been partially supported by the SmartData@PoliTO center on Big Data and Data Science.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.

[2] TensorFlow Model Analysis. 2018. Introducing TensorFlow Model Analysis: Scaleable, Sliced, and Full-Pass Metrics. https://medium.com/tensorflow/introducing-tensorflow-model-analysis-scaleable-sliced-and-full-pass-metrics-5cde7baf0b7b.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[5] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

[6] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[7] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Automated Data Slicing for Model Validation: A Big data - AI Integration Approach. *IEEE Transactions on Knowledge and Data Engineering* (2019). https://doi.org/10.1109/TKDE.2019.2916074

[8] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Slice Finder: Automated Data Slicing for Model Validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 1550–1553.

[9] Cynthia Dwork and Christina Ilvento. 2018. Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*.

[10] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1918–1921.

[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. https://doi.org/10.1145/3236009

[12] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining Frequent Patterns without Candidate Generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein (Eds.). ACM, 1–12. https://doi.org/10.1145/342009.335372

[13] Minsuk Kahng, Dezhi Fang, and Duen Horng Chau. 2016. Visual exploration of machine learning results using data cube analysis. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–6.

[14] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 2569–2577. http://proceedings.mlr.press/v80/kearns18a.html

[15] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems (NIPS)*. 4069–4079.

[16] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. 2019. Auditing and Achieving Intersectional Fairness in Classification Problems. arXiv:1911.01468 [cs.LG]

[17] Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD/PODS '21)*. Association for Computing Machinery, New York, NY, USA, 1400–1412. https://doi.org/10.1145/3448016.3457284

[18] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[19] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2018. *Introduction to Data Mining (2nd Edition)* (2nd ed.). Pearson.

[20] Linda F. Wightman. 1998. In *LSAC research report series*. https://eric.ed.gov/?id=ED469370

[21] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, Reproducible, and Testable Error Analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 747–763. https://doi.org/10.18653/v1/P19-1073

[22] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.

[23] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

[24] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.

[25] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in Ranking: A Survey. *arXiv preprint arXiv:2103.14000* (2021).