Fairly Evaluating and Scoring Items in a Data Set*

Abolfazl Asudeh
University of Illinois at Chicago
asudeh@uic.edu

H. V. Jagadish
University of Michigan
jag@umich.edu

ABSTRACT

We frequently compute a score for each item in a data set, sometimes for its intrinsic value, but more often as a step towards classification, ranking, and so forth. The importance of computing this score fairly cannot be overstated. In this tutorial, we will develop a framework for how to think about this task, and then present techniques for responsible scoring and link these to traditional data management challenges.

PVLDB Reference Format:
Abolfazl Asudeh, H. V. Jagadish. Fairly Evaluating and Scoring Items in a Data Set. *PVLDB*, 13(12): 3445-3448, 2020.
DOI: https://doi.org/3415478.3415566

1. INTRODUCTION

From "the secret trust scores companies use" [1] to the scores used for college admission, we are all constantly being judged by the *scores* automatically generated using *data* about us.

The scores are often derived by combining multiple criteria (aka features or attributes). For instance, a lender may combine attributes such as `payment history`, `salary`, `education`, and `age` to develop a creditworthiness score for each customer. The scores can be generated with different methods, linearly or using a complex function, and be used for different purposes. In *classification*, we use the scores to draw a decision boundary to specify, for example, if a woman is at risk of developing invasive breast cancer over the next 5 years [2]. In *ranking*, the scores are used to sort the entities and, for example, select the top-8 soccer teams for seeding pot 1 in the world cup tournament [3].

The scores are usually assigned either through (i) a process learned by machine learning models using some labeled training data, or (ii) using a weight vector or a procedure designed by human experts. For example, a logistic regression model learns a weight vector that transforms a regularized multi-dimensional feature set into a score that translates to a class label. Conversely, the scoring mechanism used by US News for university ranking is a linear function designed by human experts [4].

*Supported in part by NSF under grants 1741022 and 1934565.

Proceedings of the VLDB Endowment, Vol. 13, No. 12
ISSN 2150-8097.
DOI: https://doi.org/3415478.3415566

Data-informed decisions can be profitable to societies and human beings. For example, China developed an app to help identify high risk individuals in the fight against the Coronavirus [5].

On the other hand, even if it looks promising, data-driven decision making can cause harm. Probably the main reason is that real-life social date is almost always "biased" [6, 7]. No one can miss the extensive recent discussion about race in the context of policing and criminal justice. But similar questions arise in many other domains as well. Take college admission for example. It has been shown that the `GPA` values has gender bias. That is due to grading policies that, for instance, may reduce grades for students with late homework, disruptive behavior, or inattention [8, 9]. As a result, using `GPA` as one of the features for generating the scores and ranking the students without considering the inherent bias in data can lead to gender bias. Evidence of bias has also been reported in predictive policing [10], college admission [11], credit scoring [12], and job interviewing, hiring, and promotion [13], among others.

Such failures and, more generally, bias in data-driven decision making, started the fast-growing area of *fairness* in data science. Existing work has set the ground by providing the necessary terms and definitions [14, 15]. Still, a challenge is that fairness is normative and there is no universal definition; there are trade-offs not only between fairness and other optimization factors, but even between different definitions of fairness [16–18]. As a result, it is impossible even to satisfy all fairness definitions at the same time [19].

This tutorial consists of three parts. First, as outlined in § 2, we will discuss bias in social data and the meaning of fairness. These definitions provide a foundation to describe the problems we seek to address, and some challenges that make our task particularly difficult. Next, in § 3, we will present an overview the work to date towards addressing these problems. We will provide a taxonomy of scoring mechanism design, evaluation tasks based on the scores, and decision types. Discussing various types of interventions, we will survey a group of representative papers. We will conclude the tutorial by viewing the problem domain from the perspective of the database community, in § 4. We will discuss opportunities to leverage techniques originated from solving core data management challenges to make contributions to responsible scoring and algorithmic fairness.

2. DEFINITIONS AND CHALLENGES

A dataset is a collection of tuples $\mathcal{D} = \{t_1, \cdots, t_n\}$, each defined over a set of attributes (aka features) $X = \{x_1, \cdots, x_m\}$ that are used for decision making. In addition, a dataset contains a set of sensitive attributes $S = \{s_1, \cdots, s_{m'}\}$ such as `gender` and `race` that identify the demographic groups. We want decisions based on data to be fair across different demographic groups.

2.1 Data-Driven Bias

A potential advantage of algorithmic decision making is that it removes human bias and only looks at the data dispassionately. But "an algorithm is only as good as the data it works with" [7]. Social data is almost always biased as it inherently reflects historical biases and stereotypes [6]. Data collection and representation methods often introduce additional bias. Using biased data without paying attention to societal impacts can create a *feedback loop*, and even increase discrimination in society. There are two main perspectives on bias and the types of bias definitions: statistical/algorithmic and societal bias. Biases have been looked at for a long time in statistical community [20] but social data presents different challenges [6,7,15]. For social data, the term bias refer to demographic disparities in the sampled data that compromises its representativeness and are objectionable for societal reasons [6,15]. At a high level, viewing a dataset $\mathcal{D}$ as a table of rows $\{t_1, \cdots, t_n\}$ and columns $\{x_1, \cdots, x_n\}$, bias can exist with regard to rows and the columns:

- *Bias on rows* (Population bias): Systematic distortions in the number of tuples from different demographic groups. Failing to include enough samples from minority (sub)groups (a.k.a. the lack of coverage [21]) in a dataset used for building or training a scoring strategy makes it "difficult" to provide accurate analytical results for the minorities [22]. In other words, population bias in training data can result in models that perform differently across different groups.
- *Bias on columns* (Behavioral bias): differences in behavior (value distribution) of attributes across different demographic groups. This is also known as proxy bias, due to the high correlation with sensitive attributes of "proxy" attributes in the dataset. Lower salaries for employees of female sex or higher arrest rate for people of color are some examples of biased attributes. Biased values in the data can directly transform to bias in the algorithms' outcomes.

In addition to biases and inaccuracies occurring at the source of the data, bias can also be introduced during the data collection, including linking bias, content production bias, and temporal bias [6].

2.2 Fairness and Stability

There has recently been much work towards defining fairness [14, 15, 23]. Fairness, at a high level, is partitioned into individual fairness, which deals with discrimination against individuals, and group fairness, which considers parity over different demographic groups. While some works such as [24] study individual fairness, considering the social implications, most attention has been on group fairness. Kearns et al. [25, 26] proposed the notion of rich subgroup fairness to bridge between group fairness and individual fairness. Probably the more popular notion of fairness is based on model *independence* or *demographic parity* [14, 15, 23], also referred to by terms such as statistical parity [24], and disparate impact [7]. Model independence simply requires the sensitive characteristic to be statistically independent of the score [15]. There is a similarity between this model and diversity [27]. In addition to independence, fairness can be defined using the notions of *separation* and *sufficiency* [15]. Considering a target variable (true label in classification) for every tuple in a supervised learning setting, the separation model allows correlation between the score and a sensitive attribute to the extent that it is justified by the target variable. Fairness measures such as *predictive equality*, *Equal opportunity*, and *Equalized odds* follow the separation model. Sufficiency model requires independence of a target variable and a sensitive attribute conditional to the scores. In other words, a score satisfies sufficiency if the sensitive attribute and target variable are clear from the context. *Predictive parity* is an example fitting into this model.

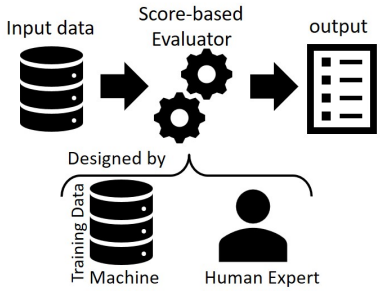

Figure 1: The architecture of evaluation systems based on scoring.

In addition to fairness, we want a scoring to be stable with respect to changes in the parameters used for scoring. A scoring is stable if small changes in its parameters do not change the outcomes based on the scores [28]. For example, consider a scoring function for employees of a company that combines multiple performance criteria X, such as `sales` and `customer satisfaction`, using a weight vector θ in the form of $f_\theta = \theta^\top X$. The weight vector can be designed by human experts or learned using a machine learning method. Suppose the company would like to promote the top-k employees. The scoring is stable if weight vectors similar to θ also generate the same top-k as of θ. Formally speaking, a scoring is stable if a large portion of vectors in the vicinity of the current scoring-parameters vector also generate the same outcome. In addition to changes in the scoring parameters, stability has also been defined in form of changes in the input data and its distribution in learning settings [29]. Stability from this perspective is similar to the concept of *robustness* in machine learning [30].

2.3 Challenges

The following are challenges towards responsible scoring:

- *Lack of representative data.* A major challenge is that social data is often limited to what is called "found data" [6, 21]. That is, analyses are done with data that has been acquired independently, possibly through a process on which the data scientist has limited, or no, control. Collecting more data is often challenging and, hence, we are restricted to the biased found data.
- *Unknown values for sensitive attributes.* A challenge in designing scores is to mitigate disparate impact without practicing *disparate treatment* – not to explicitly use the sensitive attributes such as gender or race in the scores. In other words, how to be fair without asking the the demographic information of users?
- *Lack of universal definitions.* As explained in § 2.2, there is not universal definition for terms such as fairness as those are often application-specific and identified based on societal norms.
- *Trade-offs and impossibility theorems.* Trade-offs between different objectives is a major challenge in responsible scoring design. This includes the trade-off between evaluation performance (e.g. model accuracy in classification) and fairness [16] as well as the trade-off between different definitions fairness [18, 19].

3. CURRENT SOLUTIONS

3.1 Score-based Evaluation

Scoring is the key component in the architecture of score-based evaluators (Figure 1 [31]). The scores are computed by combining attributes X either through a learning process (using some training data), or using a weight vector assigned by human experts. Learning methods require that there be labeled data, and assume that there is some known ground truth. In contrast, an expert-specified method does not require any labeled data. Also, it has recently been recognized that *"for predicting social outcomes, AI is not substantially better than manual scoring using just a few features"* [32]. Such methods, however, may be ad-hoc.

Two tasks that use scores to support decisions are classification and ranking[1]. Classification is often attacked by first solving, for example, a regression problem to summarize the data in a single score [15]. The scores are then turned into discrete labels by discretizing them into buckets. In the case of binary classification, a threshold t is used that when the score is larger than t the classifier outputs one (e.g. accept) and zero (reject) otherwise.

The scores can be used to evaluate an object in a non-competitive setting by only looking at its score value. For example, whether or not an individual is classified as high risk for having breast cancer only depends on that individual herself. In addition, the evaluations can be competitive. For example, consider a company that would like to give a raise to k percent of its employee or a student that would like to apply to the top-20 universities in the US. In such settings, knowing the absolute score of an entity is not enough for identifying the decision about it, but rather we need the score relative to that of other entities.

3.2 Responsible Scoring Interventions

Interventions to achieve responsible scoring fall in two main categories: *pre-processing* techniques and *algorithm modification* [29][2]. Pre-processing techniques are mainly designed for scores learned using training data. These methods attempt to rectify bias in training data that causes unfairness in model outcomes. They do so by modifying the data so that any learning algorithm applied to it will generate fair outcomes. The second category of techniques blends fairness into the scoring algorithm design. They change the problem formulation to either ensure fairness as hard constraints, or to trade-off between scoring performance and these terms. In the following, we first review some of the techniques that fall in the first category. We will then acknowledge some representative algorithm modification interventions for ranking and classification.

3.2.1 Pre-processing and Data Investigation

Algorithmic decisions are unfair because social data is biased. Removing bias from the data will remove unfairness. Given a specific dataset, pre-processing techniques modify data according to specific needs [29]. Many pre-processing techniques focus on removing behavioral bias. These include removing biased attributes, adding derived attributes, removing problematic tuples, *massaging* the data (changing class labels), *re-weighting* (assigning weights to tuples), and *re-sampling* [33–35]. Unlike other pre-processing approaches that use statistical correlations [36] formulates the problem as a causal database repair problem, proving sufficient conditions for fair classifiers in terms of admissible variables.

[21,37,38] study *coverage* over a dataset to ensure that there are enough representatives in the dataset for demographic subgroups (e.g. Hispanic Female). Specifically, [21] uses "patterns" to represent the subgroups in form of attribute-value combinations and aims to find the ones for which there are not "enough" instances in the dataset. It also recommends a small number of additional data points to resolve "problematic" uncovered patterns.

Besides pre-processing, it is necessary to provide tools that help to investigate datasets. For example [39] provides task-specific information about a dataset, in the form of a set of visual widgets, as a flexible "nutritional label".

3.2.2 Scoring Design and Algorithm Modification

The types of algorithm modifications for responsible scoring depend on the evaluation task and how those are created.

Modifying learning algorithms for achieving fairness in score-based classification has extensively been studied [40–42]. For example, [40] adds fairness as a regularization term in the optimization of logistic regression. Zafar et al. [42] observe that fairness constraints are non-convex and propose a convex approximation for the purpose of optimization. Zemel et al. [41] propose a combination of preprocessing and algorithm modification. They formulate fairness as the problem of finding an unbiased representation of data that is good for classification.

Expert-designed scores in the form of $f_\theta = \theta^\top X$ are commonly used for ranking (and sometimes for classification). A major issue with such scores is that the assignment of weights is ad-hoc. Designing fair score-based rankings has been studied in [43]. It proposes a query answering system that helps experts choose weight vectors that lead to greater fairness – for a set of user-defined fairness requirements. Given a user-defined weight vector θ, it returns the most similar vector to θ whose output (ranking) satisfies the fairness requirements. The expert can choose the system suggestion, or use it to explore different weights before finalizing their scoring function. [28] aims at obtaining stable evaluations based on ranking. The size of the region, in parameter space, that produces an observed ranking identify its stability. The intuition behind this is that if only a few weight choices can produce an output, it may suggest that the output was engineered or "cherry-picked". Besides measuring the stability of a given ranking, [28] designs a GET-NEXT operator that returns the next stable ranking upon calling it. It also provides an unbiased samplers from the weight-vector space that enables Monte-carlo methods for responsible scoring [44]. [45] combines [43] and [28] in system for responsible ranking.

4. OPPORTUNITIES

Fairness has become a big topic for the ML/AI research community. However, the construction of the ML model is only one step in the Big Data ecosystem. We must address all parts of this ecosystem to ensure fairness. In the following we highlight some of the many contributions the database community can make in the general area of algorithmic fairness:

- *Input data preparation.* Mitigating bias through the pipeline of data preparation is a necessary step towards algorithmic fairness. The database community has a lot to offer here, given its expertise in *data discovery, cleaning, and integration*. Removing bias from the input data can be viewed as a special case of data cleaning where the goal is to replace, modify, or delete problematic tuples or values that cause bias.
- *Data representation.* Representation choices are critical design decisions, traditionally approached with performance as the central objective. These decisions can also impact fairness. For example bucketization choices can lead to very different analysis results.
- *Data investigation.* Data scientists require tools to investigate bias in the data. Topics such as data profiling, context, and provenance have an important role to play in designing such tools.
- *Algorithm design.* Topics such as ranking and top-k queries are well-studied in the database community. Utilizing such techniques for responsible scoring is a promising direction.
- *Result presentation.* How results are presented can also introduce bias. The database community has studied biased framing, cherry-picking [46], and such other spin methods. This work can be continued to understand implications for fairness.
- *Integrating to databases.* Last but not least, an important step is to fully implement fairness concepts and requirements in the database engine and to add declarative functions to SQL.

[1]Note that in some contexts ranking or classification is done without scoring. Our focus are the evaluations based on scores.

[2]Post-processing techniques are another category of interventions that minimally change the evaluations to satisfy fairness. Since such methods do not change the scores, those are out of our scope.

5. REFERENCES

[1] J. Mason. The secret trust scores companies use to judge us all. The Wall Street Journal, April 6, 2019.

[2] The breast cancer risk assessment tool. bcrisktool.cancer.gov, (accessed March 2020).

[3] World cup 2018 seeding: Pots, procedure & all you need to know ahead of the draw. GOAL.COM, 12/1/2017.

[4] How u.s. news calculated the 2020 best graduate schools rankings. bit.ly/39HjnGQ, 3/11/2019.

[5] P. Mozur, R. Zhong, and A. Krolik. In coronavirus fight, china gives citizens a color code, with red flags. The New York Times, March 1, 2020.

[6] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[7] S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[8] R. Buddin. Gender gaps in high school gpa and act scores. *ACT Research & Policy*, 2014.

[9] C. Roth. Women job applicants punished for higher grades, study finds. WOSU Public Media, Mar 26, 2018.

[10] A. D. Selbst. Disparate impact in big data policing. *Ga. L. Rev.*, 52:109, 2017.

[11] J. L. Santos, N. L. Cabrera, and K. J. Fosnacht. Is "race-neutral" really race-neutral?: Disparate impact towards underrepresented minorities in post-209 uc system admissions. *J. High. Educ*, 81(6):605–631, 2010.

[12] M. F. Vidal and J. Menajovsky. Algorithm bias in credit scoring: What's inside the black box? *CGAP blog*, 2019.

[13] P. T. Kim. Data-driven discrimination at work. *Wm. & Mary L. Rev.*, 58:857, 2016.

[14] I. Žliobaitė. Measuring discrimination in algorithmic decision making. *DATA MIN KNOWL DISC*, 31(4):1060–1089, 2017.

[15] S. Barocas, M. Hardt, and A. Narayanan. Fairness and machine learning: Limitations and opportunities. fairmlbook.org, 2019.

[16] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *SIGKDD*. ACM, 2017.

[17] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *FAT**, 2018.

[18] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR, abs/1609.05807*, 2016.

[19] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *CoRR, abs/1609.07236*, 2016.

[20] J. Neyman and E. S. Pearson. Contributions to the theory of testing statistical hypotheses. *Statistical Research Memoirs*, 1936.

[21] A. Asudeh, Z. Jin, and H. Jagadish. Assessing and remedying coverage for a given dataset. In *ICDE*, 2019.

[22] R. A. Baeza-Yates. Big data or right data? In *AMW*, 2013.

[23] A. Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *FAT**, 2018.

[24] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *ITCS*, pages 214–226, 2012.

[25] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. An empirical study of rich subgroup fairness for machine learning. In *FAT**, pages 100–109, 2019.

[26] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*, pages 2564–2572, 2018.

[27] M. Drosou, H. Jagadish, E. Pitoura, and J. Stoyanovich. Diversity in big data: A review. *Big data*, 5(2):73–84, 2017.

[28] A. Asudeh, H. Jagadish, G. Miklau, and J. Stoyanovich. On obtaining stable rankings. *PVLDB*, 12(3):237–250, 2018.

[29] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *FAT**, 2019.

[30] J. Steinhardt. *Robust Learning: Information Theory and Algorithms*. PhD thesis, Stanford University, 2018.

[31] A. Asudeh, H. Jagadish, and J. Stoyanovich. Towards responsible data-driven decision making in score-based systems. *Data Engineering*, 42(3):76–87, 2019.

[32] A. Narayanan. How to recognize ai snake oil www.cs.princeton.edu/~arvindn/talks. Technical report, MIT-STS-AI-snakeoil.pdf, 2019.

[33] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[34] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *SIGKDD*, 2015.

[35] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *NIPS*, pages 3992–4001, 2017.

[36] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD*, pages 793–810, 2019.

[37] Z. Jin, M. Xu, C. Sun, A. Asudeh, and H. Jagadish. MithraCoverage: A system for investigating population bias for intersectional fairness. *SIGMOD*, 2020.

[38] Y. Lin, Y. Guan, A. Asudeh, and J. H. V. Identifying insufficient data coverage in databases with multiple relations. *PVLDB*, 13(11):2229–2242, 2020.

[39] C. Sun, A. Asudeh, H. Jagadish, B. Howe, and J. Stoyanovich. MithraLabel: Flexible dataset nutritional labels for responsible data science. In *CIKM*, 2019.

[40] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML PKDD*, pages 35–50. Springer, 2012.

[41] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013.

[42] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. *CoRR, abs/1507.05259*, 2015.

[43] A. Asudeh, H. Jagadish, J. Stoyanovich, and G. Das. Designing fair ranking schemes. In *SIGMOD*, 2019.

[44] A. Asudeh and H. Jagadish. Responsible scoring mechanisms through function sampling. *CoRR, abs/:1911.10073*, 2019.

[45] Y. Guan, A. Asudeh, P. Mayuram, H. Jagadish, J. Stoyanovich, G. Miklau, and G. Das. MithraRanking: A system for responsible ranking design. In *SIGMOD*, 2019.

[46] A. Asudeh, H. Jagadish, Y. Wu, and C. Yu. On detecting cherry-picked trendlines. *PVLDB*, 13(6):939–952, 2020.