DATA INTERCHANGE BIASES AND ITS IMPACT ON ALGORITHMIC FAIRNESS

GEVORG GHALACHYAN

Keywords – Artificial intelligence, data interactions biases, socio-econometric metrics, selection bias, Simpson's paradox, behavioral bias.

INTRODUCTION

With the popularity of artificial intelligence and machine learning in the past years, and their wide spread in different applications, fairness and safety constraints became a significant problem for researchers and engineers. ML is used in courts for assessing the probability whether a defendant commits a new crime. It is also used in broad medical fields, in children welfare systems, autonomous vehicles. All these applications do have a direct effect in our everyday lives and can harm our society once not engineered and designed correctly, that is with respective considerations to fairness. Odonse el al[1] has shown a list of applications and ways these systems affect our lives with their inherent biases. Some of them are

- AI chatbots,
- flight routing,
- employment matching,
- immigration support automated legal aid algorithms,
- search and advertising placement algorithms, etc.

Howard and Borenstein[2] discuss some examples of how real-world biases creep into robotic systems, which include bias in voice recognition, face recognition applications, and search engines. Thus, it is significantly important for

[1] **Osonde A** Osoba and William Welser IV. 2017. An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation.

[2] **Ayanna Howard and Jason Borenstein.** 2018. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics 24, 5 (2018), 1521–1536.

engineers and researchers to be worried about the upstream applications and their huge potential harmful effects when they model an algorithm or a system.

Types of Biases

Biases are in many forms and shapes, some of which lead to unfairness to different downstream learning tasks. In Suresh and Guttag[1], authors speak about bias sources in machine learning along with their descriptions and categorizations in order to make motivation for future solutions of each sources of bias shown in the paper. In Olteanu et al[2], the authors prepared a full list of different biases and their respective definitions which exist in different cycles - data origins to collection and processing. Now we reiterate some most important sources of bias in these two papers, also add some new work from other different research papers. Also, we introduce a different categorization of these definitions - according to the algorithm, data, and user interaction loop.

Data to Algorithm Biases

Here we talk biases in data, which might result biased algorithmic outcomes, when used by machine learning training algorithms.

1. **Measurement Bias**. Measurement, or reporting, is a type of bias that arises from the way we choose, utilize, and measure the features[3]. Such bias was observed and found inside the recidivism risk prediction system COMPAS, when friend/family arrests and prior arrests were used as input variables to measure level of crime and riskiness. It on its own can be viewed as mismeasured proxies. This is partially due to the fact that minorities are policed and controlled more frequently, thus they do have higher arrest rates. Yet, we should not conclude that as people from minority groups have higher arrest rates they are therefore more dangerous because there is some difference in how these groups are controlled and assessed.

2. **Omitted Variable Bias.** Omitted variable bias[4] happens when one or more very important features are left out of the statistical model. An example is when someone designs a model to predict, with quite high accuracy, the quarterly percentage rate at which the customers stop subscribing to some service, but late observes that the vast majority of users now cancel the

---

[1] **Harini Suresh and John V Guttag**. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019)

[2] **Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman**. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016)

[3] **Harini Suresh and John V Guttag.** 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).

[4] **Kevin A Clarke.** 2005. The phantom menace: Omitted variable bias in econometric research. Conflict management and peace science 22, 4 (2005), 341–352.

subscription without receiving warnings from the designed model. So, imagine that a reason for canceling one's subscriptions is the appearance of a strong competitor in market which now offers the very same solution, but for half the price you offer. The appearance of the new competitor was a thing that the model was not quite ready for; thus, it is considered an omitted variable.

3. **Representation Bias**. Representation bias happens from the way we sample from a population in the data collection process. Non-representative samples often lack the diversity of the population, e.g. missing subgroups or other anomalies. Lack of ethno-geographical diversity in datasets like ImageNet[1] resulted in enormous bias to Western cultures (see Figure 1).

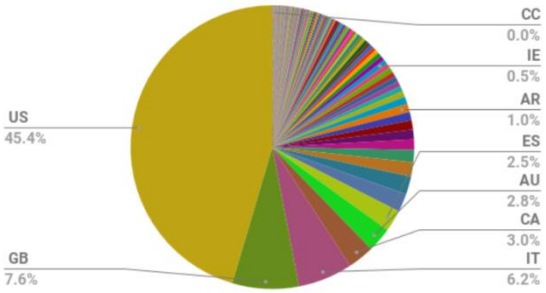

Figure 1: Percentage of each country (two-letter ISO code) of ImageNet dataset. Bias towards Western countries like US, Great Britain and Italy are obvious.

4. **Aggregation Bias**. Aggregation bias (sometimes called ecological fallacy) happens when a false conclusion is drawn about individuals while observing the whole population. An example of such bias can be seen in clinical aid system. Consider type II diabetes patients who have significant morbidity differences between genders and ethnicities. More specifically, levels of HbA1c, which are widely used to monitor and diagnose type II diabetes, differ complexly across ethnicities and genders. Thus, any model that ignores per individual differences will probably not be well-suited for all gender and ethnic groups inside the population. This is even true when they are represented in equal proportions in the training set. All general assumptions for subgroups within the population will result in aggregation bias. It occurs so often that similar cases are looked as phenomena. Here are couple of them:

[1] **Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al** . 2015. Imagenet large scale visual recognition challenge. International journal of computer vision 115, 3 (2015), 211–252

a. **Simpson's Paradox.** Simpson's paradox is an aggregation bias that happens while analyzing heterogeneous data[1]. The paradox is defined as an association found in aggregated data which reverses or disappears when the data is disaggregated into the underlying subgroups (see Figure 2). A better-known example occurred during the gender bias law-suit against UC Berkeley in university admissions[2]. After the analysis of the graduate school admissions data, it was obvious that there was bias to women, a smaller fraction of them were admitted to graduate programs than their male counterparts. Yet, when they separated admissions data and analyzed it over the departments, female applicants had equality and sometimes even a slight advantage over males. The paradox happened as females tended to apply to university departments with lower admission rates for all the genders. Simpson's paradox has been observed in a variety of domains, including astronomy, psychology, biology, and computational social science.

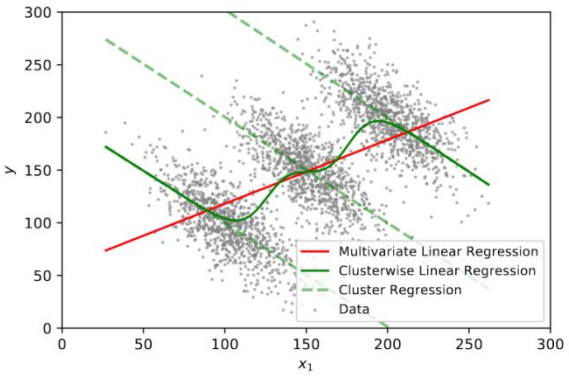

Figure 2: Visual representation of Simpson's Paradox

b. **Modifiable Areal Unit Problem** is another type of statistical bias of geospatial analysis, thar arises during the modeling of data at different levels of spatial aggregation[3]. It results in different mini-trends learned when data is being aggregated at different spatial scales.

---

[1] **Colin R Blyth.** 1972. On Simpson's paradox and the sure-thing principle. J. Amer. Statist. Assoc. 67, 338 (1972), 364–366.

[2] **Peter J Bickel, Eugene A Hammel, and J William O'Connell.** 1975. Sex bias in graduate admissions: Data from Berkeley. Science 187, 4175 (1975), 398–404.

[3] **C. E. Gehlke and Katherine Biehl.** 1934. Certain effects of grouping upon the size of the correlation coefficient in census tract material. J. Amer. Statist. Assoc. 29, 185A (1934), 169–170. https://doi.org/10.2307/2277827

5. **Sampling Bias**. Sampling bias is like representation bias, and it arises because of non- random sampling of the subgroups. As a result of sampling bias, the mini-trends estimated for a population might not generalize to the entire data collected from another population. For the example, let's consider the example in Figure 2 and Figure 3. Figure 2 represents data collected from a study of three subgroups, which are uniformly sampled. Let's suppose that the next time the same study was conducted, one of the sub-groups was non-randomly sampled more frequently than the others (Figure 3). The positive trend we found by a regression model in the first case almost entirely disappears (solid red line in Figure 3), yet the subgroup trends (dashed with green lines) are unaffected.

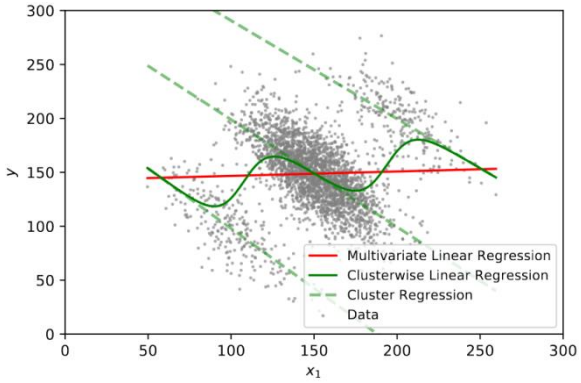

Figure 3: A showcase of the sampling bias

6. **Longitudinal Data Fallacy**. Researchers who analyze temporal data should use longitudinal analysis in order to track cohorts to learn their behavior over time. Instead of that, temporal data is quite often modeled with cross-sectional analysis, that combines diverse cohorts at a very single time point. The heterogeneous cohorts may bias the cross-sectional analysis, which leads to conclusions other than longitudinal analysis. For example, analysis of Reddit data[1] showed that the average length of comment decreased over time. Yet, data represented a cross-sectional snap of the population, that in reality contained various cohorts which joined Reddit in different months. When we disaggregated the data by cohorts, the comment length within each cohort increased over time.

[1] **Samuel Barbosa, Dan Cosley, Amit Sharma, and Roberto M. Cesar-Jr**. 2016. Averaging Gone Wrong: Using Time- Aware Analyses to Better Understand Behavior. (April 2016), 829–841

7. **Linking Bias**. Linking bias happens when the network features of user activities, connections, or in-between interactions misinterpret and differ from the real behavior of the users[1]. In Mehrabi et al[2] authors show that social networks are biased toward low-degree nodes when they only consider the links of the network and do not consider the content and behavior of users. Wilson et al[3] also show that the user interactions are notoriously different from social media link patterns that are created on features, e.g. method of interaction or time. The biases and differences in the networks are a result of many factors, such as network sampling, that can change the network measures to cause different types of problems.

Algorithm to User Biases

Algorithms temper user behavior. All biases in algorithms may show biases in user behavior. Here we speak about biases which are a result of algorithmic outcomes and as a consequence affect user behavior.

1. **Algorithmic Bias**. Algorithmic bias happens when the bias is not in the input features and is added solely by the algorithm[4] The choices of algorithmic design, like use of certain optimization methods, regularizations, fitting regression models on the entire data or thinking subgroups, and the standard use of biased estimators in models, all contribute to biased decisions which can bias the predictor of the models.

2. **User Interaction Bias**. User Interaction bias is a statistical phenomenon that can not only be seen on the web-pages but also be triggered from two different sources—the UI and by the user itself by showing his/her self-biased behavior and interaction. Such bias is influenced by other types, like presentation and ranking biases.

    a. **Presentation Bias**. This is a result of how the information is presented. For example, on the network users can click on content that they see, thus the seen content is clicked, while everything other has no click. And that could be the reason that the user won't see the whole information on the internet.

---

[1] **Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman**. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016)

[2] **Ninareh Mehrabi, Fred Morstatter, Nanyun Peng, and Aram Galstyan.** 2019. Debiasing Community Detection: The Importance of Lowly-Connected Nodes. arXiv preprint arXiv:1903.08136 (2019).

[3] **Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao.** 2009. User interactions in social networks and their implications. In Proceedings of the 4th ACM European conference on Computer systems. Acm, 205–218

[4] **Ricardo Baeza-Yates.** 2018. Bias on the Web. Commun. ACM 61, 6 (May 2018), 54–61

b. **Ranking Bias**. The idea that highly ranked results are usually the most important and relevant results in obtaining of more clicks than other URLs. This bias affects search engines systems and applications of crowdsourcing.

3. **Popularity Bias**. More popular items tend to be shown more. Yet, metrics of popularity can be manipulated, e.g. by fake reviews or social bots[1] . As an example, such bias can be seen in search engine systems or recommendation providing systems when popular objects are presented more to the users. But such presentation might not be a result of quality; moreover, it can be due to any other biased factors.

4. **Emergent Bias**. Emergent bias happens as a consequence of interaction and use with real-life users, after change in cultural values, population, or societal knowledge, and usually sometime after the design is completed[2]. It is more likely to be found in user interfaces, as interfaces reflect the characteristics, capacities, and habits of exemplary users by design. This type of bias itself can be divided into more subtypes.

5. **Evaluation Bias**. It arises when the final model is evaluated. Evaluation bias includes the use of disproportionate and inappropriate baselines and benchmarks for evaluation. Such benchmarks are often used in for facial recognition systems that are biased to gender and skin color, and may ba an example for such bias.

User to Data Biases

A lot of data sources that is used for training ML models are generated by users. All inherent biases that users possess may be reflected in the data that their actions generate. Moreover, when user action is affected by any algorithm, all the biases present in those algorithms could introduce a new bias in the process of data generation. Furtherly we list several quite important types of user-to-data biases.

1. **Historical Bias**. Historical bias is the one already existing and socio-demographical issues in the world, and can go from the generation process of data even with a perfect sampling and process of feature selection. An example of such bias can be a 2018 image search result where keyword searching for female CEOs always resulted in fewer female images of CEOs because of the fact that only five percent of Fortune 500 CEOs were female - which did cause

[1] **L. Introna and H. Nissenbaum.** 2000. Defining the Web: the politics of search engines. Computer 33, 1 (Jan 2000), 54–62

[2] **Batya Friedman and Helen Nissenbaum.** 1996. Bias in Computer Systems. ACM Trans. Inf. Syst. 14, 3 (July 1996), 330–347.

the keyword search results to be biased to male CEOs. Those search results were definitely reflecting our reality, but whether the search algorithms should reflect this reality or not is an concern worth thinking of.

2. **Population Bias**. Population bias happens when demographics, statistics, representatives, and user-relating characteristics are varying in the user-based population of the platform from the original population. Population bias generates non-representative data. An example of such bias can come from different demographics of user on different social media platforms, e.g. women are more likely to use Facebook, Pinterest, Instagram, while men are more active in some online forums like Reddit and Twitter. More of such examples and descriptive statistics related to social media platform use among young adults according to ethnicity, sex, race, and parental education background can be found in research[1].

3. **Self-Selection Bias**. It is a subtype of the sampling or selection bias where subjects of the research do select themselves. An example of this can be observed in a user-opinion poll that measures enthusiasm for a political candidate, whereas the most amused supporters are more propable to complete the poll.

4. **Social Bias**. Social bias arises when others subjects' actions affect our judgment. An example is a case where users want to rate or review a shopping item with a comparatively low score, but while influenced by some high ratings, we start changing our thinking of scoring, thinking that perhaps we are too harsh.

5. **Behavioral Bias**. Behavioral bias comes from different users' behavior along contexts, platforms, or different data. An example of it can be observed in Miller et al[2], where they show how differences of emoji representations in platforms result in various reactions and behavior from users and sometimes even leads to communication chaos.

6. **Temporal Bias**. Temporal bias represents differences in behaviors and populations over time. An example is observed in Twitter; people that talk about some topic start using a hashtag(#) at a point to draw attention, then carry on the discussion of the event without considering the hashtag.

---

[1] **Eszter Hargittai.** 2007. Whose Space? Differences among Users and Non-Users of Social Network Sites. Journal of Computer-Mediated Communication 13, 1 (10 2007), 276–297.

[2] **Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht.** 2016. "Blissfully Happy" or "Ready to Fight": Varying Interpretations of Emoji. In Tenth International AAAI Conference on Web and Social Media.

7. **Content Production Bias**. Content Production bias comes from lexical, structural, semantic, and sometimes even from syntactic differences in the contents that users generate. An example of this type of bias is the languages used in post-soviet countries to generate content by different age groups.

Conclusion

After discussing different types of data interactions biases, we conclude

- There are 3 main subtypes of data interaction biases – data to algorithm, algorithm to users, users to data.
- All 3 types are in a loop, which means that a bias has an everlasting effect on the data-intensive system.
- Feedback loops are created to find the bias and its quantitative measure.
- Bias mitigation mechanisms are used for more complete machine learning systems. We plan to research this topic furtherly.

Գևորգ Ղալաչյան, Տվյալների փոխանակման շեղումները և դրա ազդեցությունը ալգորիթմի արդարության վրա - Մեքենայական ուսուցման համակարգ կառուցելու համար անհրաժեշտ են տվյալներ, որոնք օգտագործելով կկառուցվի համակարգը, ալգորիթմ, որը պատմական փաստերի հիման վրա կկայացնի որոշումներ, և օգտագործողներ, որոնք համակարգի ազդեցության առաջնային կրողն են: Շեղումներ կարող են առաջանալ նման համակարգերի կառուցման տարբեր փուլերում՝ ինքնուրույն, ինչպես նաև խմբային՝ որպես երևույթների փոխազդեցության հետևանք, խանգարելով համակարգի կառուցմանը, ինչպես նաև հետագա կատարելագործմանը, պահպանմանը: Հոդվածում քննարկում ենք այն շեղումները, որոնք հանդիպում են տվյալների, ալգորիթմի և օգտագործողների միջև փոխազդեցության հետևանքով՝ զույգ առ զույգ: Տվյալներից ալգորիթմ շեղումը մարդկանց կողմից պատմական կողմնակալության ազդեցությունն է, որը փոխանցվում է ալգորիթմին ուսուցման ընթացքում, ալգորիթմից օգտվողներ շեղումը տեղի է ունենում, երբ ալգորիթմում առկա է գիտելիքի անհամապատասխանություն օգտվողների բազմության հետ, և օգտվողներից ալգորիթմ շեղումը տվյալների հավաքագրման թերացումներն են օգտատերերի գործողությունների և փոխազդեցությունների պատճառով: Մենք ներկայացրել ենք շեղումների ենթատեսակները, անդրադարձել դրանց ներկայացմանը այլ հեղինակների կողմից, ցույց տվել օրինակների միջոցով: Նաև քննարկել ենք շեղման որոշ ենթատեսակներ ալգորիթմի արդարության տեսանկյունից. որոշումների անհավասարակշռությունը սոցիալական խմբերում կարող են տարբեր-

վել, հետևաբար և առաջացնել անհավասարություն կամ ունենալ այլ սոցիալ-տնտեսական հետևանքներ։

Геворг Калачян, Ошибки обмена данных и их влияние на алгиритмическую справедливость - Для создания машинного обучения необходимы 3 вещи — данные, на основе которых будет построена система; алгоритм, который будет принимать решения на основе исторической справки, и пользователей, заинтересованных сторон, на которых система будет оказывать влияние. В этой статье мы обсудим предубеждения, которые могут возникнуть во время взаимодействия этих троих в паре. Смещение данных по отношению к алгоритму считается историческим предубеждением со стороны людей, которые переходят к алгоритму на этапе обучения, смещение алгоритма по отношению к пользователям происходит как несоответствие между популяцией пользователей и алгоритмическими знаниями, а смещение пользователей по отношению к данным — это дефекты в процессе сбора данных, основанные на при действиях и взаимодействиях пользователя. Мы интерпретируем каждый подтип смещения, делаем ссылку на предыдущее исследование и показываем его на примере. Также мы обсуждаем некоторые виды смещения в контексте алгоритмической справедливости, поскольку дисбаланс решений может проявляться в каждой группе по-разному и, следовательно, вызывать неравенство и его социально-демографические последствия.

Ուղարկվել է խմբագրություն 22.02.2022թ.
Գրախոսվել է 22.02.2022թ.