Race and the Machine

Re-examining Race and Ethnicity in Data Mining

Sophie Chou
MIT Media Lab
75 Amherst St
Cambridge, MA
soph@media.mit.edu

ABSTRACT

The rapid growth in popularity of machine learning and data mining to determine social outcomes shows great potential, but presents many technical and ethical challenges. In particular, the analysis of race and ethnicity, either as a variable or by proxy, can be prone to profiling and discrimination. This paper aims to provoke new directions in analyzing such data. Problems and challenges present themselves along the entire pipeline, including in the acquisition of biased datasets, during feature and class selection for algorithmic classification and segmentation, and in the responsible application of results. Additionally, I discuss potential for computational solutions, both in monitoring existing software and detecting discrimination from datasets and algorithms. I advocate for the treatment of race as a mutable social variable, and subsequent redesign of algorithms.

Keywords

data science, data mining, race, racism, discrimination

1. INTRODUCTION

With the proliferation of social networks and releases of large online datasets in the last decade, data mining human behaviors to create models for all sorts of social applications has been on the rise, and will only continue to grow. Many fields, such as healthcare, education, and governance, all stand to benefit from automation. However, although there lies great potential for data-driven analysis and software, social phenomena are extremely complex and difficult to quantify. Issues of identity—such as race, gender, and orientation—are especially difficult to model. In this vision paper, I encourage more socially conscious views of race in data. Specifically, I advocate for it to be treated as a flexible, variable, and contextual concept, following a constructionist philosophy. I begin in Section 2 and 3 with groundwork definitions of both race and data mining. In Section 4, I explore related work in other fields, and detail the problem statement in Section 5. Sections 6, 7, and 8 serve as a critical examination of existing standards in data mining, and I conclude with potential solutions in Section 9.

2. DEFINING DATA MINING

With the growing availability of large datasets, there has been an increased interest in data-driven methods of analysis and prediction. The terms "data science", "data mining", "data analytics", and "machine learning" are often used interchangeably, and with confusion, to target similar problems.

To make a distinction, in this paper I will refer to data science as the broader umbrella concept, "the set of activities involved in transforming collected data into valuable insights, products, or solutions" [4]; machine learning as the field of computer science that focuses on improving pattern recognition and prediction; data mining as the specific application of existing machine learning techniques to better understand and manipulate specific datasets, and data analytics as the business-facing aspects of the above.

Here, I choose to focus on data mining, for the key reason that it emphasizes the application of algorithms rather than their abstract optimization. However, areas of challenge and redesign are relevant to all related fields.

3. THE POPULAR ACCOUNT OF RACE

Race, a term which seems to expand and shrink in social scope depending on context, can be far more difficult to define. Although the concept of human races might seem intuitive or "natural", it is a politically loaded term subject to differences in philosophy.

Conceptions of race generally boil down into two camps: essentialist versus constructivist perspectives. Common societal perceptions of race view it as a *fixed* and *immutable* variable. Although discussions have broadened in recent years, by and large it is considered a descriptive, rather than prescriptive term. A person is assigned a racial category before birth, chosen from a number of set possible classes, which remains constant throughout their entire life. This is known as an essentialist or primordialist notion of race, also deemed by philosopher Lawrence Blum as the "popular account". Notably, it is distinct from other large-scale human groupings, such as ethnicity and nationality, which are viewed as cultivated [3]. A main contributor to this notion of race is the fact that it includes physical or biological "markers", traits that are prone to genetic explanation (although not necessarily so).

In contrast, social constructionists view race (along with other labels, such as gender) in a less deterministic way.

Instead, it is seen as a category constructed and shaped by existing social structures. Under this school of thought, race is significantly neither biologically nor genetically justified. Instead, it is contextually dependent on time, place, and social surrounding. Key to a constructionist view is the fact that race is not necessarily consistent. In the following, I demonstrate that adopting a constructionist perspective is a better approach for data analysis.

4. RELATED WORK

In "Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics", Harvard researchers Maya Sen and Omar Wasow address current limitations of studying race in the field of political science. Treating race as an immutable, pre-assigned trait limits experimental design—since the variable is determined before birth, considering other effects post-birth risks introducing post-treatment bias, making studies of race difficult [10]. As a solution, Sen and Wasow suggest examining questions involving race in a more constructionist framework—understanding it as a composite variable ("bundle of sticks") so that each aspect may be examined separately in experiment for estimating causal inference.

Data mining faces similar issues in considering race as an immutable variable, challenges that compound when applying results to make predictions based on race. Although predictive algorithms are not necessarily concerned with determining causal links, they are concerned with inference at large—and more importantly, applications. In this paper, I expand on some of the issues specific to data mining in treating race as a static variable.

"Big Data's Disparate Impact", a survey by Barocas and Selbst, examines how existing practices in data mining are prone to discrimination, focusing on legal implications and difficulties with legislation [2]. Here, I choose to highlight how changes in algorithmic design can lead to positive effect—focusing on the power that the scientist and engineer holds.

Finally, there exists a small (but growing!) group of publications within the computational community criticizing popular techniques for demographic segmentation and inference. I highlight a few of these papers as examples for improvement.

5. VISION AND PROBLEM STATEMENT

In this paper I propose a shift in the data mining and machine learning community to treat race as a socially constructed variable, instead of one that is static and immutable. Existing methods of labeling race can be prone to stereotyping and profiling, and when applied to social applications, can have unwanted discriminatory effects.

Regardless of personal philosophy, it is more appropriate to treat race in a constructionist sense when dealing with data mining, since it is, by definition, the mining of observed behaviors within a social context. Reconsideration of the way race is examined in data, and how it is presented in results, opens up new avenues for creativity without discounting existing techniques. This consideration is both more experimentally responsible and scientifically sound. Accuracy suffers when behaviors that shift depending on social context are used as a proxy for variables seen as immutable and static. Moreover, harmful racial stereotyping and profiling may occur with the application of such algorithms. In the following, I will outline challenge areas in analyzing (and predicting) with racial data, suggesting examples contrary to the common practice that show improved results, and shine light on the potential for algorithmic and societal accountability through data mining.

6. EXAMINING EXISTING METHODS

Data mining demographics is currently an area of great interest to those in research as well as industry. There exists a large, growing body of work around attempts to infer latent traits based on trails of human behavior, as it is a rich area for machine learning, not to mention profitable to businesses. Predicting current interests and future behavior based on these demographics, either implicitly or explicitly by tagged data, is a natural follow-up.

In considering race in the context of data mining, there are two main concerns regarding discrimination: racial stereotyping in the behavior of the algorithm and racial profiling in its applications. Outright racism, or intentional hatred and ill-doing, requires an element of emotion and human agency that software does not have unless explicitly coded to do so. Both stereotyping and profiling are suspect in algorithmic thinking, which relies primarily on statistical simplifications and extrapolations of world-views.

7. RACIAL STEREOTYPING

Although race categories have held significant weight throughout history, in assignment they have rarely been constant. The number of race categories has been constantly evolving over the past decades, merging and branching with one another, and continue to do so. As politics and populations shift and change, certain ethnic groups and nationalities (Italians, "Eskimos", etc.) become racialized while others lose status or undergo changes in naming designations.

There is a substantial body of work dedicated to the task of detecting underlying demographic attributes as a class label or target using statistical classification. Numerous works seek to detect a person's ethnicity or race based on social media activity. Many algorithms boast high (80% and above) accuracy in prediction [7]. However, almost all existing works view race as a fixed, static, number of categories (target classes), usually limiting the view to today's census categories or simply black and white.

According to the U.S. census, self-reporting of race has been valid since 1960, and since 2000, one could choose several racial categories to describe one's identity. Even in the now expanded list of choices given, a significant number of the population (about 6%, corresponding to several millions of people) chose to report a category not listed [1].

That leads to two important implications: first, the realization that viewing race as a static variable, as in an essentialist matter, does not hold when data is taken either cross-generationally or cross-culturally. Concepts of race are local and time-dependent phenomena and need to be modeled accordingly. Second, that there is a significant population of people who fail to be correctly categorized even if all categories, historically and currently, are taken into consideration.

In "Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment", Nguyen et. al examine limitations with viewing age and gender as fixed variables. Through a crowdsourcing experiment, they are

able to show that not all usage of gendered language indicates the twitter user's underlying gender. More importantly, they found that disagreements amongst both human annotators and machine methods indicate signal about a person's social behavior, instead of the common assumption that it is merely noise. For example, a male user who was incorrectly labeled as female used gendered language to tweet to a group of female friends, indicating in-group behavior [6].

Performing racial (along with other demographic) classification without the pretense that race is performative and adaptive risks relying on harmful stereotype. In data mining, it is common to examine linguistic features (such as prototypical words and hashtags), semantic features (such as prototypical topics) and network features (such as friends or followers on social networks) in order to form a model. Depending on the bias of the initial training data, and subsequent overfitting, this can easily result in a handful of speech and social patterns that evoke stereotype for a racial group (or any other demographic breakdown, as Ruths and Cohen demonstrated with political affiliation) [5]. Moreover, even in a randomly sampled training set, the assumption that race has uniform aggregate behavior leaves room for only one core identity within a racial group, marginalizing those whose identities are less common, and overlooking identity switches within the same individual.

8. RACIAL PROFILING

Another concern of mining racial data is that it might lead in unintentional profiling once applied. Using machine intelligence to generate outcomes for areas of high consequence such as policing, criminal sentencing, medical treatment and diagnoses, and employment opportunities is especially prone to risk. All of these fields are subject to bias if examining race as a variable directly, as the effect being observed is often an aggregate of different features: social, political, economic, physical, geographical.

A textbook example of this type of misevaluation can be found in medicine. In the case of sickle cell disease, race is often falsely interpreted. Although both blacks and whites can carry the sickle-cell gene, because the population of individuals with sickle-cell in the United States is primarily black, it has come to be treated as a black disease. In reality, sickle cell develops in malarial regions, but in the US, the number of blacks from a malarial region of Africa is far greater than whites from a malarial region of Europe, resulting in a skewed statistic. Even today, many screenings for sickle-cell are targeted towards blacks, resulting in over-diagnosis in one population, and under-diagnosis in the other [8]. In medical practice and research, when doctors and scientists use race as a proxy for medical traits, they are engaging in racial profiling.

Data mining that looks at race as a variable risks masking true causal links in a similar way. Consequences can extend beyond healthcare. Although organizations cannot practice racial discrimination in hiring, they can use health results (such as the sickle-cell test) to deny blacks or other minorities employment for certain jobs [8]. Medical software that employs data mining, then, has a potential to further the disparate impact. The pitfalls of using race as a proxy feature are not limited to healthcare data only. The same issues arise in mining census data, employment data, housing data, or in fact, any information source from which race is listed and considered.

Instances where race is not explicitly stated can also be prone to bias if other variables implicitly serve as a proxy for race. Recently, concern has been raised about the fairness of evidence-based sentencing, or "punishment profiling", which uses statistical analysis and data mining to help determine the gravity of incarceration. Although specifics vary by state, traits commonly examined such as unemployment data, neighborhood, and marital status can be heavily correlated with race [11]. This can result in a disporportionate number of sentences in certain populations, regardless of the crime committed itself. The aspiration for scientifically grounded sentencing can lead to harmful discrimination if the compound effects of race are left unacknowledged.

9. POTENTIAL SOLUTIONS

Revisiting existing techniques in data mining from a constructionist perspective requires, more than anything, a shift in interpretation of results. Instead of viewing a classification of a person from her digital traces as indicative of her "real" identity, regarding it as a snapshot of an evolving identity in time avoids maligning communities and individuals. Under the example of Nguyen et al., this analysis can also lead to great social insight [6] ! On the side of data collection, providing better surveys which include the option for self-reporting can be a more sensitive and descriptive approach. These self-described categories can also provide valuable insight to outliers. In cases where race is to be examined to make predictions, taking the approach of Sen and Wasow to treat it as an aggregate variable helps pinpoint specific causal links and avoid false proxies.

Most importantly, algorithmic accountability does not end at program design. It can also take the form of close monitoring of software after it has been released in the wild. Even the most careful of designs can prove to be biased once put into production, as realized by researcher Latanya Sweeney, as she analyzed discrimination in Google ads [12]. Although data mining does comes under scrutiny for being potentially discriminatory, it can also be a great ally in detecting patterns of bias in datasets, as detailed by Ruggieri in "Data Mining for Discrimination Discovery" [9]. Applying such algorithms periodically to data generated and processed by ones that make predictions with race data can be a useful measure of fairness.

10. CONCLUSION

In this paper, I advocate for re-examination of the way race is treated as both a feature and a label in data mining. I believe that a shift to more socially constructive and contextial notions of race can improve algorithmic accuracy and avoid stereotype. Significantly, this can help avoid potentially discriminative software, which can harbor grave consequences for marginalized communities. An foray into new methods for treating race and ethnicity as socially constructed proves to be rich for expansion, while existing methods can easily be adapted to relay more socially conscious results. Any science, including data science, that examines human behavior in society is first and foremost a "social science", and care should be taken to evaluate results as such.

11. REFERENCES

[1] "us census looking at big changes in how it asks about race and ethnicity". http://www.pewresearch.org/fact-tank/2014/03/14/u-s-census-looking-at-big-changes-in-how-it-asks-about-race-and-ethnicity/. Accessed: 2015-08-08.

[2] S. Barocas and A. D. Selbst. Big data's disparate impact. *Available at SSRN 2477899*, 2014.

[3] L. A. Blum. *I'm Not a Racist, But–: The Moral Quandary of Race*. Cornell University Press, 2002.

[4] S. Chou, W. Li, and R. Sridharan. Democratizing data science. 2014.

[5] R. Cohen and D. Ruths. Classifying political orientation on twitter: It's not easy! In *ICWSM*, 2013.

[6] D.-P. Nguyen, R. Trieschnigg, A. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. de Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. Association for Computational Linguistics, 2014.

[7] M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[8] M. Root. The use of race in medicine as a proxy for genetic differences. *Philosophy of Science*, 70(5):1173–1183, 2003.

[9] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):9, 2010.

[10] M. Sen and O. Wasow. 2016. "race as a 'bundle of sticks': Designs that estimate effects of seemingly immutable characteristics.". *Annual Review of Political Science*, 2016.

[11] S. B. Starr. "sentencing, by the numbers". http://www.nytimes.com/2014/08/11/opinion/sentencing-by-the-numbers.html. Accessed: 2015-09-11.

[12] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.