Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining

Sara Hajian
Eurecat
Barcelona, Spain
sara.hajian@eurecat.org

Francesco Bonchi
ISI Foundation
Turin, Italy
francesco.bonchi@isi.it

Carlos Castillo
Eurecat
Barcelona, Spain
chato@acm.org

ABSTRACT

Algorithms and decision making based on Big Data have become pervasive in all aspects of our daily lives lives (offline and online), as they have become essential tools in personal finance, health care, hiring, housing, education, and policies. It is therefore of societal and ethical importance to ask whether these algorithms can be discriminative on grounds such as gender, ethnicity, or health status. It turns out that the answer is positive: for instance, recent studies in the context of online advertising show that ads for high-income jobs are presented to men much more often than to women [5]; and ads for arrest records are significantly more likely to show up on searches for distinctively black names [16].

This *algorithmic bias* exists even when there is no discrimination intention in the developer of the algorithm. Sometimes it may be inherent to the data sources used (software making decisions based on data can reflect, or even amplify, the results of historical discrimination), but even when the sensitive attributes have been suppressed from the input, a well trained machine learning algorithm may still discriminate on the basis of such sensitive attributes because of correlations existing in the data. These considerations call for the development of data mining systems which are discrimination-conscious by-design. This is a novel and challenging research area for the data mining community.

The aim of this tutorial is to survey algorithmic bias, presenting its most common variants, with an emphasis on the algorithmic techniques and key ideas developed to derive efficient solutions. The tutorial covers two main complementary approaches: algorithms for discrimination discovery and discrimination prevention by means of fairness-aware data mining. We conclude by summarizing promising paths for future research.

Keywords

Algorithmic bias; Discrimination discovery; Discrimination prevention

KDD '16 August 13-17, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-4232-2/16/08.

DOI: http://dx.doi.org/10.1145/2939672.2945386

1. INTRODUCTION

At the beginning of 2014, as an answer to the growing concerns about the role played by data mining algorithms in decision-making, USA President Obama called for a review of big data collecting and analysing practices. The resulting report[1] concluded that "big data technologies can cause societal harms beyond damages to privacy." In particular, it expressed concerns about the possibility that decisions informed by big data could have discriminatory effects, even in the absence of discriminatory intent, further imposing less favorable treatment to already disadvantaged groups.

In the data mining community, the effort to design discrimination-conscious methods has developed two groups of solutions: (1) techniques for *discrimination discovery from databases* [13] and (2) discrimination prevention by means of *fairness-aware data mining*, developing data mining systems which are discrimination-conscious by-design [8]. Discrimination discovery in databases consists in the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records. Discrimination prevention in data mining consists of ensuring that data mining models automatically extracted from a data set are such that they do not lead to discriminatory decisions even if the data set is inherently biased against protected groups. Different discrimination prevention methods have been proposed considering different data mining algorithms such as naïve bayes models, logistic regression, decision trees, hinge loss, support vector machines, adaptive boosting, classification, and rule and pattern mining. Three approaches are conceivable for discrimination prevention: preprocessing by means of transforming the source data; inprocessing by means of integrating the anti-discrimination constrains in the design of algorithm; postprocessing by means of modifying the results of data mining models.

2. INTENDED AUDIENCE

The tutorial is at researchers interested in the technical aspects behind the societal and ethical problems of discrimination and privacy introduced by data mining and machine learning algorithms. No special knowledge is assumed other than familiarity with algorithmic techniques from a standard computer science background.

3. OUTLINE

The tutorial is structured in three main technical parts, plus a concluding part where we discuss future research

[1] http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

agenda. All three technical parts include both theory and real-world applications.

1. **Introduction, context, and fundamental results:** Motivation and examples of algorithmic bias; sources of algorithmic bias; legal definitions and principles of discrimination; relationship with privacy [1, 4].

2. **Discrimination discovery:** Measures of discrimination; data analysis techniques for discrimination discovery (multidisciplinary approaches in the economic, legal, and statistical domains); data mining approaches for discrimination discovery including classification rules, K-NN, bayesian networks, probabilistic causation and privacy attack strategies; case studies in crime suspect dataset and scientific project evaluation; online discrimination discovery including search, ad delivery, and price discrimination; tools including DCUBE and FairTest [13, 12, 14, 2].

3. **Fairness-aware data mining:** Pre-processing approaches including correcting the training data; in-processing approaches including naïve bayes models, logistic regression, decision trees, hinge loss, support vector machines, adaptive boosting and classification; post-processing approaches including classification, rule and pattern mining; simultaneous discrimination prevention and privacy protection considering k-anonymity, t-closeness and differential privacy models [3, 8, 11, 7, 6, 10, 9].

4. **Challenges and directions for future research.**

4. SUPPORT MATERIAL

We developed a mini-website for the tutorial: http://francescobonchi.com/algorithmic_bias_tutorial.html. It contains the tutorial slides and a full list of references.

5. INSTRUCTORS

Sara Hajian is a Researcher at Eurecat. She received her Ph.D. degree from Computer Engineering and Maths Department of the Universitat Rovira i Virgili (URV). Her research interests include data mining methods and algorithms for social media and social network analysis, privacy-preserving data mining and algorithmic bias. The results of her research on algorithmic bias featured in Communications of ACM journal [15].

Francesco Bonchi is Research Leader at the ISI Foundation, Turin, Italy, where he leads the "Algorithmic Data Analytics" group. Before he was Director of Research at Yahoo Labs in Barcelona. He is PC Chair of the 16th IEEE International Conference on Data Mining (ICDM 2016). He has also been PC Chair of ECML PKDD 2010 and of several workshops on privacy-aware data mining including f the first and second ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD 2007 and 2008), the 1st IEEE International Workshop on Privacy Aspects of Data Mining (PADM 2006). He is co-editor of the book "Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques" (Chapman & Hall/CRC Press). Homepage: http://www.francescobonchi.com.

Carlos Castillo is Director of Research for Data Science at Eurecat. Carlos is an active researcher with more than 70 papers in top-tier international conferences and journals, including an upcoming book on Big Crisis Data, a book on Information and Influence Propagation, and a monograph on Adversarial Web Search. He has been PC Chair of ACM Digital Health 2016 and WSDM 2014, organized the Adversarial Information Retrieval Workshop in 2007 and 2008, the ECML-PKDD Discovery Challenge in 2010 and 2014, the Web Quality Workshop from 2011 to 2014, and the Social Web for Disaster Management Workshop in 2015. Homepage: http://chato.cl/research/.

6. ACKNOWLEDGMENTS

This tutorial is partially supported by the EU H2020 innovation action programme, TYPES project (grant No 653449) and Catalonia Trade and Investment Agency (ACCIÓ).

7. REFERENCES

[1] S. Barocas and A. D. Selbst. Big data's disparate impact. *SSRN Pre-Print 2477899*, 2014.

[2] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti. Exposing the probabilistic causal structure of discrimination. *arXiv:1510.00552*, 2015.

[3] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *DMKD*, 21(2):277–292, 2010.

[4] B. Custers, T. Calders, B. Schermer, and T. Zarsky, editors. *Discrimination and Privacy in the Information Society.* Springer, 2013.

[5] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *Proc. Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *ITCS*, 2012.

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

[8] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *TKDE*, 25(7):1445–1459, 2013.

[9] S. Hajian, J. Domingo-Ferrer, and O. Farràs. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *DMKD*, 28(5-6):1158–1188, 2014.

[10] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. Discrimination- and privacy-aware patterns. *DMKD*, 29(6):1733–1782, 2015.

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

[12] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *KDD*, 2011.

[13] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, 2008.

[14] S. Ruggieri, S. Hajian, F. Kamiran, and X. Zhang. Anti-discrimination analysis using privacy attack strategies. In *ECML-PKDD*, 2014.

[15] N. Savage. When computers stand in the schoolhouse door. *Commun. ACM*, 59(3):19–21, Feb. 2016.

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