Algorithmic bias: from discrimination discovery to fairness-aware data mining (KDD'16 Tutorial)

Slides: http://francescobonchi.com/algorithmic_bias_tutorial.html

Francesco Bonchi[1,2] Carlos Castillo[2] Sara Hajian[2]
[1]ISI Foundation
Turin, Italy
[2]Eurecat, Technological Center of Catalonia
Barcelona, Spain
francesco.bonchi@isi.it chato@acm.org sara.hajian@eurecat.org

Abstract

Algorithms and decision making based on Big Data have become pervasive in all aspects of our daily (offline and online) lives, as they have become essential tools in personal finance, health care, hiring, housing, education, and policies. Data and algorithms determine the media we consume, the stories we read, the people we meet, the places we visit, but also whether we get a job, or whether our loan request is approved. It is therefore of societal and ethical importance to ask whether these algorithms can be discriminative on grounds, such as gender, ethnicity, marital or health status. It turns out that the answer is positive: for instance, recent studies have shown that Google's online advertising system displayed ads for high-income jobs to men much more often than it did to women; and ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity.

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

1 Target audience, prerequisites and importance

The tutorial is aimed 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 will be assumed other than familiarity with algorithmic techniques from a standard computer science background.

This topic is one of the important and challenging topics to the KDD community, and has recently received significant attention from mainstream media. At the beginning of 2014, as an answer to the growing concerns about the role played by data mining algorithms in decision-making, US President Obama called for a 90-day 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.

This is an emerging research area with plenty of open questions, and in dire need of theoretical results as well as practical tools, for researchers and practitioners.

2 Outline

The tutorial is structured in three main technical parts, plus a concluding part where we will discuss future research agenda. All three technical parts will include both theory and real-world applications.

1. **Introduction, context, and fundamental results**

   1.1 Motivation and examples of algorithmic bias [1, 2, 3].

   1.2 Sources of algorithmic bias [4, 5, 6].

   1.3 Legal definitions and principles of discrimination [7, 8, 9].

   1.4 Relationship with privacy [10, 11, 12, 13].

2. **Discrimination discovery**

   2.1 Measures of discrimination [7, 14, 15].

   2.2 Data analysis techniques for discrimination discovery: multidisciplinary approaches in the economic, legal, and statistical domains [7].

   2.3 Data mining approaches for discrimination discovery: classification rule mining [16, 17, 18], K-NN classification [19], bayesian networks [20], probabilistic causation [21], privacy attack strategies [22].

   2.4 Case studies: crime suspect dataset [23], scientific project evaluation [24].

   2.5 Online discrimination discovery: search and price discrimination [25], AdFisher [2], discrimination in online ad delivery [1].

   2.6 Tools: DCUBE [26], FairTest Testing Toolkit [27].

3. **Fairness-aware data mining**

   3.1 Pre-processing approaches: correcting the training data [28, 29, 30, 31]; the relation between accuracy and fairness [31, 32].

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

3.2 In-processing approaches: naïve bayes models [33], logistic regression [34, 35, 40], decision trees [36], hinge loss [34], support vector machines [34, 35], adaptive boosting [35], classification [41].

3.3 Post-processing approaches: classification [12, 33, 36, 37], rule and pattern mining [10, 17].

3.4 Simultaneous discrimination prevention and privacy protection: k-anonymity [11, 13], t-closeness [13], differential privacy [12]

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

3 Related tutorials

To the best of our knowledge, no tutorials on this topic have been held at recent computer science conferences. Nevertheless, it is worth mentioning that at ICML'15 a workshop took place that is related to the topic of this tutorial: *Fairness, Accountability, and Transparency in Machine Learning (FAT ML 2015).*

4 Tutors' short bio

Dr. Francesco Bonchi (http://www.francescobonchi.com) 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, where he was leading the Web Mining Research group.

He will be **PC Chair** of the 16th IEEE International Conference on Data Mining **(ICDM 2016)** to be held in Barcelona in December 2016. He is member of the ECML PKDD Steering Committee, Associate Editor of the newly created IEEE Transactions on Big Data (TBD), of the IEEE Transactions on Knowledge and Data Engineering (TKDE), the ACM Transactions on Intelligent Systems and Technology (TIST), Knowledge and Information Systems (KAIS), and member of the Editorial Board of Data Mining and Knowledge Discovery (DMKD). He has been program co-chair of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010). Dr. Bonchi has also served as program co-chair of 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), and the 4th International Workshop on Knowledge Discovery in Inductive Databases (KDID 2005). He is co-editor of the book "Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques" (Chapman & Hall/CRC Press).

He gave a tutorial at KDD 2014 on *"Correlation Clustering: from Theory to Practice"* (slides: http://www.francescobonchi.com/CCtuto_kdd14.pdf).

Dr. Carlos Castillo (http://chato.cl/research/) is the Director of Research for Data Mining at Eurecat. His current research focuses in social computing, particularly the application of web mining methods to social media during disasters and humanitarian crises. 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. Carlos received his Ph.D from the University of Chile, and was a senior scientist at Yahoo! Research, and a principal scientist at Qatar Computing Research Institute. He has served in the PC or SPC of all major conferences in his area (WWW, WSDM, KDD, SIGIR, CIKM) and is part of the editorial committee of ACM Transactions on the Web and Internet Research. He is Program Committee Co-Chair of ACM Digital Health 2016, and was Program Committee Co-chair of WSDM 2014, co-organized the Adversarial Information Retrieval Workshop and Web Spam Challenge 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.

He gave a tutorial at KDD 2012 on *"Information and Influence Spread in Social Networks"* (which was the basis of a book: http://dx.doi.org/10.2200/S00527ED1V01Y201308DTM037) and a tutorial at SDM 2014 on *"Leveraging Social Media and Web of Data to Assist Crisis Response Coordination"* (https://www.siam.org/meetings/sdm14/leveraging.php).

Dr. Sara Hajian is a Researcher at Eurecat Technology Center, Barcelona, Spain. She received her Ph.D. degree from Computer Engineering and Maths Department of the Universitat Rovira i Virgili (URV) in June 2013. She received her M.Sc. degree in Computer Science from Iran University of Science and Technology (IUST) in 2008. She also had been a member of APA-IUTcert, an academic research and development center in the area of Network Security Vulnerabilities and Incident Handling (2008-2010). Her research interests are data mining methods and algorithms, social media and social network analysis, privacy-preserving data mining and publishing, and algorithmic bias (discovery and prevention of discrimination). She has been a visiting student at the Knowledge Discovery and Data Mining Laboratory (KDD-Lab), a joint research group of the Information Science and Technology Institute of the Italian National Research Council (CNR) in Pisa and the Computer Science Department of the University of Pisa (2011). She has been a visiting scientist at Yahoo! Labs in Barcelona (2013-2014).

5 Tutorial materials

Slides and videos of the tutorial are available at http://francescobonchi.com/algorithmic_bias_tutorial.html

References

[1] L. Sweeney. Discrimination in online ad delivery. *Commun. ACM*, 56(5), pp.44-54, 2013.

[2] A. Datta, M.C. Tschantz, and A. Datta. Automated experiments on Ad privacy settings. In *PETS*, pp.92-112, 2015.

[3] A.E. Roth and E. Peranson. The effects of the change in the NRMP matching algorithm. *JAMA*, 278(9):729-732, 1997.

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

[5] M. Hardt. How big data is unfair: Understanding sources of unfairness in data driven decision making. 2014.

[6] T. Calders and I. I. Zliobaite. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society* (eds. B. H. M. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky), volume 3 of Studies in Applied Philosophy, Epistemology and Rational Ethics, pp. 4357. Springer, 2013.

[7] A. Romei, S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(05),2013.

[8] B. Custers, T. Calders, B. Schermer and T. Z. Zarsky (eds.). *Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases.* Studies in Applied Philosophy, Epistemology and Rational Ethics 3. Springer, 2013.

[9] R. Gellert, K. D. Vries, P. D. Hert, and S. Gutwirth. A comparative analysis of anti-discrimination and data protection legislations. In *Discrimination and Privacy in the Information Society* (eds. B. H. M. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky), volume 3 of Studies in Applied Philosophy, Epistemology and Rational Ethics, pp. 43-57. Springer, 2013.

[10] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. Discrimination-and privacy-aware patterns. In *Data Mining and Knowledge Discovery*, 29(6), 2015.

[11] S. Hajian, J. Domingo-Ferrer, and O. Farras. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*, 28(5-6), pp.1158-1188, 2014.

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. Fairness through awareness. In *ITCS 2012*, pp. 214-226. ACM, 2012.

[13] S. Ruggieri. Using t-closeness anonymity to control for non-discrimination. *Transactions on Data Privacy*, 7(2), pp.99-129, 2014.

[14] I. Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.

[15] D. Pedreschi, S. Ruggieri, and F. Turini. A study of top-k measures for discrimination discovery. In *SAC*, pp. 126-131, 2012.

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

[17] D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *SDM*, 2009.

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

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

[20] K. Mancuhan and C. Clifton. Combating discrimination using bayesian networks. In *Artificial Intelligence and Law*, 22(2), 2014.

[21] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti. Exposing the Probabilistic Causal Structure of Discrimination. *arXiv preprint arXiv:1510.00552*, 2015.

[22] S. Ruggieri, S. Hajian, F. Kamiran and X. Zhang. Anti-discrimination Analysis Using Privacy Attack Strategies. In *PKDD*, 2014.

[23] F. Kamiran, A. Karim, S. Verwer and H. Goudriaan. Classifying socially sensitive data without discrimination: an analysis of a crime suspect dataset. In *ICDMW*, pp. 370-377, 2012.

[24] A. Romei, S. Ruggieri and F. Turini. Discrimination discovery in scientific project evaluation: A case study. *Expert Systems with Applications*, 40(15), 2013.

[25] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting price and search discrimination on the internet. In *Hotnets*, pp.79-84, 2012.

[26] S. Ruggieri, D. Pedreschi and F. Turini. DCUBE: Discrimination discovery in databases. In *SIGMOD*, pp. 1127-1130, 2010.

[27] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.P. Hubaux, M. Humbert, A. Juels, and H. Lin. Discovering Unwarranted Associations in Data-Driven Applications with the FairTest Testing Toolkit. *arXiv preprint arXiv:1510.02377*, 2015.

[28] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. In *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 25(7), 2013.

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

[30] I. Zliobaite, F. Kamiran and T. Calders. Handling conditional discrimination. In *ICDM*, pp. 992-1001, 2011.

[31] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pp. 259-268, 2015.

[32] I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

[33] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277-292, 2010.

[34] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. Fairness Constraints: A Mechanism for Fair Classification. *arXiv preprint arXiv:1507.05259*, 2015.

[35] B. Fish, J. Kun and A.D. Lelkes. A Confidence-Based Approach for Balancing Fairness and Accuracy. *arXiv preprint arXiv:1601.05764*, 2015.

[36] F. Kamiran, T. Calders and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, pp. 869-874, 2010.

[37] F. Kamiran, A. Karim, and X. Zhang. Decision Theory for discrimination-aware classification. In *ICDM*, pp. 924-929, 2012.

[38] S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer and F. Giannotti. Injecting discrimination and privacy awareness into pattern discovery. In *ICDMW*, pp. 360-369, 2012.

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

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

[41] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. Learning fair representations. In *ICML*, pp. 325-333, 2013.