Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (35^{th}cycle)

A Study on Data Imbalance: Using Metrics on Input Data to Foresee Bias and Fairness in Classification Outcomes

Mariachiara Mecati
******

Supervisors:
Prof. Marco Torchiano, Supervisor
Prof. Antonio Vetrò, Co-Supervisor

Doctoral Examination Committee:
Prof. Michael Littman, Referee, Brown University
Prof. Andrea Marrella, Referee, Sapienza Università di Roma
Prof. Letizia Tanca, Politecnico di Milano
Prof. Tania Cerquitelli, Politecnico di Torino
Prof. Fulvio Corno, Politecnico di Torino

Politecnico di Torino
2023

A Study on Data Imbalance: Using Metrics on Input Data to Foresee Bias and Fairness in Classification Outcomes

Mariachiara Mecati

Data has become a fundamental element of our society in conjunction with the increasing adoption of automation software in a variety of organizational and production processes, and especially of automated decision-making (ADM) systems, which may affect multiple aspects of our lives. Indeed, when software makes decisions that allocate resources or opportunities, might disparately impact people based on personal traits (for example, gender, ethnic group, etc.) and thus might systematically (dis)advantage certain social groups; for these reasons, bias in software systems is a serious threat to human rights. One of the potential causes of unfairness lies in the quality of the data used to train ADM systems. In particular, bias in input data is a relevant socio-technical issue that emerged in recent years, and it still lacks a commonly accepted solution: the "bias in-bias out" problem is one of the most significant risks of discrimination, which encompasses technical fields, as well as ethical and social perspectives. Among the causes of bias, one of the most relevant issues is represented by data imbalance, that is, an unequal distribution of data between the classes of an attribute.

We enrich the current body of research on this topic by proposing a risk assessment approach based on the measurement of data imbalance, which is derived from the principles outlined in ISO standards for software quality and risk management. We look at data imbalance in a given dataset as a potential risk factor for detecting discrimination caused by ADM systems: specifically, we aim to evaluate whether it is possible to identify the risk of bias in a classification output by measuring the level of (im)balance of protected attributes in training data. After that, we investigate the issue of data imbalance more and more thoroughly: we define a methodology to identify imbalance thresholds in input data to achieve desired levels of algorithmic fairness; then, we study imbalance on intersectional protected attributes and on the combination of the target variable with protected attributes.

To conduct our studies, we selected a set of indexes of balance (Gini, Simpson, Shannon, Imbalance ratio) and we first assess their capability to detect (im)balance

in synthetic attributes. Then, we tested their ability to identify unfair classification outcomes in large datasets belonging to different application domains, that is, their capacity to foresee a certain level of discrimination risk –which depends on the context, the dataset's domain, and the choice of the measures. Specifically, we applied the indexes of balance to protected attributes in the training sets, while we computed the unfairness by applying different fairness criteria to the same protected attributes in the test sets. In subsequent studies we tested our approach on a large number of data mutations with different classification tasks and on a variety of combinations of balance-unfairness-algorithm in order to identify specific imbalance thresholds. Lastly, we investigated whether measures of balance on intersectional attributes are helpful to detect unfairness in classification outcomes, and whether the computation of balance on the combination of a target variable with protected attributes improves the detection of unfairness.

The results show that our approach is suitable for the proposed goal, thus the balance measures can properly detect unfairness of software output. Indeed, a negative correlation holds between balance and unfairness measures, as low levels of balance in protected attributes are related to high levels of unfairness in the output; in addition, we found that measures of balance on intersectional protected attributes are helpful to detect unfairness in classification outcomes. However, the choice of the index has a relevant impact on the detection of discriminatory outcomes, and thus on the threshold to consider as risky. Overall, to increase the generalizability of our findings, it would be recommended to extend our studies on a wider number of datasets as well as indexes of balance, for instance by considering measures for non-categorical attributes. Given the different behaviors of the balance measures in detecting possible unfairness risks, we elaborated specific pragmatic recommendations for their application.

We believe that our approach for assessing the risk of discrimination should encourage to take more conscious and appropriate actions, as well as to prevent adverse effects caused by the "bias in-bias out" problem. Especially, we hope that our findings on data imbalance will improve the identification and assessment of discrimination risks in ADM systems.