MASTERS PROJECT

Balancing data for mitigating sample and algorithmic biases

Abstract. When Machine Learning (ML) models are used in decision support systems, one must be concerned on whether the errors are concentrated in specific groups of data. There are very sensitive attributes in some domains which should not influence the models' decisions. For instance, race or gender cannot be determinant in offering a loan or determining whether a person will be a repeat offender. This is a necessary concern to prevent possible undesirable biases in ML predictions. In this work we will investigate strategies for mitigating possible sample and algorithmic biases by data balancing strategies.

1 Problem Statement, Materials and Methods

Class imbalance is an issue already well known and investigated in the Machine Learning (ML) literature (Fernández et al., 2018). Faced to this situation, the learning algorithms tend to favor the majority class, in detriment of the minority class (Fernández et al., 2018). Taking COVID-19 studies, for instance, it is far more common to observe mild cases in the contaminated population. But for a better management of hospitals, the interest is to accurately identify patients willing to evolve to serious cases. Therefore, the prediction of negative COVID-19 prognosis is clearly an imbalanced problem (Fernandes et al., 2021).

There are well known strategies for dealing with class imbalance, such as undersampling the majority class and oversampling the minority class. But which strategy should be preferred must regard on the data and domain characteristics. For instance, if the majority class is composed of many examples which are not very representative, a random undersampling may harm the predictive performance for this class too. Some studies also point that class imbalance is not a real issue for problems with well separated classes (Batista and Monard, 2003).

The presence of underrepresented groups can also be problematic in the case of the input attributes. When an emotion recognition dataset has far more observations of males than that of females, the ML model will tend to perform well in the classification of data from male individuals only (Gorrostieta et al., 2019). This creates an undesirable bias, compromising fairness in some ML-based systems. An algorithmic bias can occur whenever sensitive input features from a dataset influence the predictive results when they should not.

This work will investigate and evaluate the use of common strategies for dealing with imbalanced classes taking other qualitative attributes as reference. The idea is to employ standard undersampling and oversampling strategies to balance the data according to these sensitive groups and evaluate their effectiveness in mitigating algorithmic biases. Adaptations might

also be needed and will be investigated accordingly. In addition, we will investigate which data characteristics influence more on the results achieved by such balancing strategies. For instance, if new observations are too redundant compared to the original data the benefits of such balancing can be low. The same might happen if unrealistic data are created. Taking a meta-learning approach, we will try to characterize and understand better the effects of the data mitigating strategies.

These proposals will be evaluated empirically in well known datasets posing bias challenges (Hajian et al., 2016; Mehrabi et al., 2019) from the literature. One example is the COMPAS dataset, used for predicting crime recidivism. These strategies will be experimentally compared to up-to-date fairness-enhancing strategies from the literature (Friedler et al., 2019) as baseline.

2 Research Plan

The following activities are foreseen in the development of this project:

1. *Bibliographic research* on papers related to sample/algorithmic bias and fairness in ML: 1st year;
2. *Study and application of undersampling strategies* to balance data according to sensitive attribute values: 2nd semester;
3. *Study and application of oversampling strategies* to balance data according to sensitive attribute values: 3rd semester;
4. *Evaluation of the proposed strategies* and comparison to appropriate baselines from the literature: 3th and 4th semesters.

3 Contextualization with the Main Project

This project covers the activities related to Section 4.2 from the main project, involving data pre-processing for mitigating data quality issues. Standard solutions from the ML literature will be investigated and compared in novel ways, justifying a study at masters level.

References

Batista, G. E. and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6):519–533.

Fernandes, F. T., de Oliveira, T. A., Teixeira, C. E., de Moraes Batista, A. F., Dalla Costa, G., and Chiavegatto Filho, A. D. P. (2021). A multipurpose machine learning approach to predict covid-19 negative prognosis in são paulo, brazil. *Scientific reports*, 11(1):1–7.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*. Springer.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R., and Kane, J. (2019). Gender de-biasing in speech emotion recognition. In *INTERSPEECH*, pages 2823–2827.

Hajian, S., Bonchi, F., and Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.