Interpretability and Fairness in Machine Learning: A Formal Methods Approach

Bishwamittra Ghosh
School of Computing, National University of Singapore
bghosh@u.nus.edu

Abstract

The last decades have witnessed significant progress in machine learning with applications in different safety-critical domains, such as medical, law, education, and transportation. In high-stake domains, machine learning predictions have far-reaching consequences on the end-users. With the aim of applying machine learning for societal goods, there have been increasing efforts to regulate machine learning by imposing interpretability, fairness, robustness, privacy, etc. in predictions. Towards responsible and trustworthy machine learning, we propose two research themes in our dissertation research: interpretability and fairness of machine learning classifiers. In particular, we design algorithms to *learn interpretable rule-based classifiers*, *formally verify fairness*, and *explain the sources of unfairness*. Prior approaches to these problems are often limited by scalability, accuracy, or both. To overcome these limitations, we closely integrate automated reasoning and formal methods with fairness and interpretability to develop scalable and accurate solutions.

1 Interpretable Rule-based Machine Learning

Interpretable machine learning [Rudin *et al.*, 2022] often employs *rule-based classifiers*, which use a set of rules to represent the decision boundary. One particular notion of interpretability of classification rules is their rule-size: smaller rules with higher accuracy are preferred in medical domains. However, this presents a challenge when dealing with large datasets, as interpretable classification learning becomes a combinatorial optimization problem suffering from poor scalability. To address this issue, we propose an incremental learning framework for rule-based classification on large datasets. Our framework combines maximum satisfiability (MaxSAT) and mixed integer linear programming (MILP) with mini-batch learning.

1.1 Scalability via Incremental Learning

We introduce a new incremental learning framework, referred to as IMLI, which is based on MaxSAT for learning interpretable classification rules in propositional logic [Ghosh *et al.*, 2022b; Ghosh and Meel, 2019]. The framework aims to optimize both the accuracy and interpretability of the classification rules through a joint objective function, and an optimal rule is learned by solving a specially designed MaxSAT query. However, while MaxSAT has made considerable progress in the last decade, it is not scalable to practical classification datasets with thousands to millions of samples. To address this, we incorporate an efficient incremental learning technique that integrates mini-batch learning and iterative rule-learning within the MaxSAT formulation. This results in a framework that learns a classifier by iteratively covering the training data, solving a sequence of smaller MaxSAT queries corresponding to each mini-batch in each iteration. Our experiments demonstrate that IMLI achieves the best balance among prediction accuracy, interpretability, and scalability, with competitive accuracy and interpretability compared to existing interpretable classifiers, and impressive scalability on large datasets where both interpretable and non-interpretable classifiers fail. We also apply IMLI to learn popular interpretable classifiers: decision lists and decision sets.

1.2 Expressiveness via Logical Relaxation

We extend our incremental learning framework to enable the learning of a more relaxed representation of classification rules with higher expressiveness [Ghosh *et al.*, 2020]. Motivated by satisfying M out of N literals, we consider relaxed definitions of OR/AND operators in logic, which allows exceptions in the construction of a clause and in the selection of clauses in a rule. Based on these relaxed definitions, we introduce relaxed logical classification rules, which are inspired by the use of checklists in the medical domain and Boolean cardinality constraints. These rules generalize widely used rule representations, such as CNF, DNF, and decision sets. While the combinatorial structure of these rules results in exponential succinctness, naïve learning techniques are computationally expensive. To overcome this issue, we propose an incremental mini-batch learning procedure, called CRR, which leverages advances in MILP solvers to efficiently learn such rules. Our experimental analysis shows that CRR can generate more accurate and sparser classification rules compared to alternative rule-based classifiers.

2 Fairness in Machine Learning

Fairness in machine learning [Barocas *et al.*, 2017] involves quantifying and mitigating bias towards different sensitive groups in the data that may be introduced by the classifier. Over the past decade, multiple fairness definitions and metrics have been proposed to quantify bias. However, there has been little progress in formally verifying these fairness metrics. Furthermore, fairness metrics only measure the overall bias of a classifier and are unable to detect or explain the sources of bias. Therefore, our research focuses on two key aspects: formally verifying the bias of a classifier and explaining its sources by breaking down bias into individual features and the intersection of multiple features.

2.1 Probabilistic Fairness Verification

The problem of probabilistic fairness verification is to verify the resulting bias of a classifier given the distribution of input features. Early work on fairness verification focused on quantifying the bias of a classifier for a specific dataset. Such techniques were limited in terms of increasing confidence for wide deployment. Consequently, recent probabilistic verifiers aim to achieve verification beyond a finite dataset and instead focus on the distribution of features. Specifically, the input to the probabilistic fairness verifier is a classifier and the distribution of features, and the output is a quantification of fairness metrics that the classifier obtains given the distribution.

Formal Fairness Verification. We propose two approaches to probabilistic fairness verification: a general approach that verifies a finite classifier by encoding it into a Boolean formula [Ghosh *et al.*, 2021] and a more tailored approach for linear classifiers [Ghosh *et al.*, 2022a]. Based on stochastic satisfiability (SSAT), our proposed verifier, called Justicia, verifies the fairness of classifiers such as decision trees by solving appropriately designed SSAT formulas. In contrast to prior methods, Justicia extends verification to compound sensitive groups by combining multiple categorical sensitive features. In experiments, Justicia is more scalable than existing SMT and sampling-based probabilistic verifiers and more robust than dataset-centric empirical verifiers.

Tractable Fairness Verification with Feature Correlation. Linear classifiers have received significant attention from researchers in the context of fair algorithms. Existing fairness verifiers suffer from two-fold limitations while verifying linear classifiers: (i) poor scalability due to the use of SSAT, SMT, or sampling-based techniques, and (ii) limited accuracy due to ignoring feature correlations. To alleviate both limitations, we extend Justicia with a novel stochastic subset-sum problem-based encoding that verifies linear classifiers by dynamic programming, obtaining pseudo-polynomial complexity. To incorporate feature correlations, we consider a probabilistic graphical model, specifically a Bayesian Network, to represent the conditional dependence and independence among features using directed acyclic graphs. Experimentally, Justicia is more accurate and scalable than existing fairness verifiers for linear classifiers while verifying multiple group and causal fairness metrics. We also demonstrate two novel applications of Justicia as a fairness verifier: (a) detecting fairness attacks and fairness improvement algorithms, and (b) computing the impact of feature subsets on shifting the incurred bias of the classifiers from the original bias.

Ongoing and Future Work: Explaining Sources of Bias In our ongoing work, we combine both research themes: interpretability and fairness and propose a framework to explain the sources of unfairness, *essentially beyond sensitive features*. Fairness metrics quantify bias in a global sense, but they cannot identify or explain the sources of bias. To understand the sources of bias, it's necessary to determine *which factors contribute how much to the bias of a classifier on a dataset*. We use a feature-attribution approach to explain the sources of bias, which relates the influences of input features to the resulting bias of the classifier. We formalize Fairness Influence Function (FIF) to quantify the contribution of an individual feature and the intersection of multiple features to the resulting bias [Ghosh *et al.*, 2023]. We build an algorithm called FairXplainer, which estimates FIFs by decomposing the variance of the classifier's prediction among all subsets of features, using global sensitivity analysis. In experiments, FairXplainer captures the influences of both individual and intersectional features across various datasets and classifiers, approximates bias more accurately using FIFs than existing local explanation methods, and demonstrates a higher correlation of FIFs with fairness interventions. In future, we aim to develop improved fairness enhancing and attack algorithms based on Justicia and FairXplainer.

References

[Barocas *et al.*, 2017] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 1, 2017.

[Ghosh and Meel, 2019] Bishwamittra Ghosh and Kuldeep S. Meel. IMLI: An incremental framework for MaxSAT-based learning of interpretable classification rules. In *AIES*, 2019.

[Ghosh *et al.*, 2020] Bishwamittra Ghosh, Dmitry Malioutov, and Kuldeep S. Meel. Classification rules in relaxed logical form. In *ECAI*, 2020.

[Ghosh *et al.*, 2021] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. Justicia: A stochastic SAT approach to formally verify fairness. In *AAAI*, 2021.

[Ghosh *et al.*, 2022a] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. Algorithmic fairness verification with graphical models. In *AAAI*, 2022.

[Ghosh *et al.*, 2022b] Bishwamittra Ghosh, Dmitry Malioutov, and Kuldeep S. Meel. Efficient learning of interpretable classification rules. In *JAIR*, 2022.

[Ghosh *et al.*, 2023] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. "How biased are your features?": Computing fairness influence functions with global sensitivity analysis. In *FAccT*, 2023.

[Rudin *et al.*, 2022] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.