Research Statement

Bishwamittra Ghosh

Ph.D. Candidate
School of Computing
National University of Singapore (NUS)
https://bishwamittra.github.io

My research interest is on the fairness and interpretability in machine learning applied in safety-critical domains. Traditional machine learning, specifically deep learning, is infamous for its unfair predictions towards certain demographic sensitive groups, and for providing uninterpretable black-box predictions. In my dissertation research, I design algorithms to *verify fairness* [1, 2] and *identify the sources of unfairness* [3] and *learn interpretable classifiers* [4, 5, 6]. Prior approaches aimed at these problems are limited by either scalability or accuracy or both. To this end, I closely integrate automated reasoning, formal methods, and statistics with fairness and interpretability in machine learning for scalable and accurate solutions.

My research has thrived through multiple collaborations and internships in industry and academia. Beyond fairness and interpretability, I have collaborated on solving research problems on group testing [7] and social-spacial group queries [8, 9]. We have published our works at premier conferences and journals in artificial intelligence and machine learning (AAAI-2022, 2021, 2020, JAIR-2022, ECAI-2020, AIES-2019) and databases (VLDB-2018, TSAS-2022).

Dissertation Research

Research Thrust 1: Fairness in Machine Learning

Fairness in machine learning centers on detecting and mitigating bias towards different sensitive groups in the data induced by the classifier. In recent years, fairness literature is bestowed with multiple fairness definitions and algorithms to quantify and improve fairness. However, there has been insignificant progress in formally verifying multiple fairness definitions and algorithms in a framework and in identifying the sources of unfairness/bias as a pave way to design better fairness algorithms. My research goal is to assess the fairness claim of a classifier by proposing a formal fairness verification framework. Furthermore, I dive into identifying the sources of unfairness of classifiers.

Probabilistic Fairness Verification

The problem in probabilistic fairness verification is to verify the bias of a classifier given the distribution of input features. The early works on fairness verification focused on measuring fairness metrics of a classifier for a given dataset. Naturally, such techniques were limited in enhancing the confidence of users for wide deployment. Consequently, recent verifiers seek to achieve verification beyond finite dataset and in turn focus on the probability distribution of features. More specifically, the input to the probabilistic fairness verifier is a classifier and the distribution of features, and the output is an estimate of fairness metrics that the classifier obtains given the distribution.

Formal Fairness Verification. In our research, we propose an efficient fairness verification framework for two classes of machine learning classifiers, classifiers represented as Boolean formu-

las [1] and linear classifiers [2]. Based on stochastic satisfiability (SSAT), our proposed verifier, called Justicia, verifies the fairness of Boolean classifiers such as decision trees by solving appropriately designed SSAT formulas. Justicia also extends verification to compound sensitive groups, which are a combination of multiple categorical sensitive features such as race $\in$ {White, Black} and gender $\in$ {male, female}. Because SSAT encoding allows separate quantification to each sensitive feature without restricting the number of features. In experiments, Justicia is more scalable than the existing SMT and sampling-based probabilistic verifiers, and more robust than the sample-based empirical verifiers. We also prove a finite-sample error bound on estimated fairness metrics, which is stronger than the existing asymptotic guarantees.

Tractable Fairness Verification. Linear classifiers have attracted significant attention from researchers in the context of fair algorithms. Existing fairness verifier suffers from two-fold limitations for verifying linear classifiers: (i) poor scalability due to applying SSAT/SMT or sampling-based techniques and (ii) inaccuracy due to ignoring feature correlations. Consequently, we extend Justicia to accurately and scalably verify linear classifiers. In this extension, we propose verification based on novel *stochastic subset-sum problem*, which obtains pseudo-polynomial complexity using dynamic programming. To address feature correlations, we consider a graphical model, particularly a Bayesian Network, that represents conditional dependence (and independence) among features in the form of a DAG (directed acyclic graph). Experimentally, Justicia is more accurate and scalable than existing fairness verifiers for linear classifiers; Justicia can verify group and causal fairness metrics for multiple fairness algorithms. We also demonstrate two novel applications of Justicia as a fairness verifier: (a) detecting fairness attacks, and (b) computing the impact of a subset of features on shifting the incurred bias of the classifiers from the original bias.

Identification of Sources of Unfairness

While fairness metrics globally quantify bias, they cannot detect or explain the sources of bias. To identify the sources of bias and also the effect of affirmative/punitive actions to alleviate/deteriorate bias, it is important to understand *which factors contribute how much to the bias of a classifier on a dataset*. To this end, we follow a feature-attribution approach to understand the sources of bias, where we relate the *influences* of input features towards the resulting bias of the classifier. Particularly, we define and compute *Fairness Influence Function* (FIF) that quantifies the contribution of an individual and a subset of features to the resulting bias [3]. FIFs not only allow practitioners to identify the features to act upon but also to quantify the effect of various affirmative or punitive actions on the resulting bias. Relying on global sensitivity analysis, we instantiate an algorithm, FairXplainer, that uses variance decomposition among the subset of features and a local regressor to compute FIFs accurately, while also capturing the intersectional effects of the features. Our experimental analysis validates that FairXplainer captures the influences of both individual features and higher-order feature interactions, estimates the bias more accurately than existing local explanation methods, and detects the increase/decrease in bias due to affirmative/punitive actions in the classifier.

Research Thrust 2: Interpretable Machine Learning

In interpretable machine learning, rule-based classifiers are particularly effective in representing the decision boundary using a set of rules. The interpretability of rule-based classifiers is generally related to the size of the rules, where smaller rules with higher accuracy are preferable in practice. As such, interpretable classification learning becomes a combinatorial optimization problem

suffering from poor scalability in large datasets. To this end, we propose an *incremental learning framework* to extend interpretable classification to large datasets by wrapping MaxSAT (maximum satisfiability) and MILP (mixed integer linear programming) solving in mini-batch learning.

Scalability via Incremental Learning We propose an incremental learning framework, called IMLI [4, 5], based on MaxSAT for synthesizing interpretable classification rules expressible in proposition logic. IMLI considers a joint objective function to optimize the accuracy and the interpretability of classification rules and learns an optimal rule by solving an appropriately designed MaxSAT query. Despite the progress of MaxSAT solving in the last decade, the straightforward MaxSAT-based solution cannot scale to practical classification datasets containing thousands to millions of samples. Therefore, we incorporate an efficient incremental learning technique inside the MaxSAT formulation by integrating mini-batch learning and iterative rule-learning. The resulting framework learns a classifier by iteratively covering the training data, wherein in each iteration, it solves a sequence of smaller MaxSAT queries corresponding to each mini-batch. In our experiments, IMLI achieves the best balance among prediction accuracy, interpretability, and scalability. For instance, IMLI attains a competitive prediction accuracy and interpretability w.r.t. existing interpretable classifiers and demonstrates impressive scalability on large datasets where both interpretable and non-interpretable classifiers fail. As an application, we deploy IMLI in learning popular interpretable classifiers such as decision lists and decision sets.

Expressiveness via Logical Relaxation We extend our incremental learning framework to learn a more relaxed representation of classification rules obtaining higher expressiveness [6]. Elaborately, we consider relaxed definitions of standard OR/AND operators in Boolean logic, which allow exceptions in the construction of a clause and also in the selection of clauses in a rule. Building on these relaxed definitions, we introduce relaxed logical classification rules *motivated by the popular usage of checklists in the medical domain and Boolean cardinality constraints in logic*. Relaxed logical classification rules generalize widely employed rule representations including CNF, DNF, and decision sets. While the combinatorial structure of these rules offers exponential succinctness, the naïve learning techniques are computationally expensive. To this end, we propose an incremental mini-batch learning procedure, called CRR, that employs advances in MILP solvers to efficiently learn such rules. Our experimental analysis demonstrates that CRR can generate more accurate and sparser classification rules compared to the alternative rule-based models.

Future Research Plans

My long-term research plan is to continue designing efficient and scalable algorithms for machine learning while prioritizing its trustworthiness in safety-critical applications. I plan to work in a collaborative environment, understand problems arising in real-world applications, and solve them with advancements in machine learning and formal methods. In the following, I discuss several research themes.

Fairness and Interpretability As a Service. I envision machine learning as an alternate decision-maker of the human in the future, with applications in law, education, transportation, etc. In the high-stake and safety-critical domains, end-users expect higher transparency from black-box algorithms. Hence, achieving fairness, interpretability, robustness, and privacy are significant challenges in front of current machine learning models. While traditional machine learning such as deep learning is data-hungry, certifying and verifying properties such as fainress and interpretability

will be challenging in complex models and in the presence of large data. From this vantage point, I aim to design efficient algorithms for the fairness and interpretability of deep models, transformer-based natural language processing (NLP), and computer vision.

Counting and Optimization Problems. My past research has been centering on formulating fairness and interpretability in machine learning as counting and optimization problems; and our proposed algorithms based on formal methods and incremental solving result in both higher scalability and better accuracy. In the future, I plan to apply these techniques in solving similar counting and optimization problems, even in areas beyond machine learning.

References

[1] B. Ghosh, D. Basu, and K. S. Meel, "Justicia: A stochastic SAT approach to formally verify fairness," in *Proc. of AAAI*, 2021.

[2] B. Ghosh, D. Basu, and K. S. Meel, "Algorithmic fairness verification with graphical models," in *Proc. of AAAI*, 2022.

[3] B. Ghosh, D. Basu, and K. S. Meel, ""How biased is your feature?": Computing fairness influence functions with global sensitivity analysis (under review)," 2022.

[4] B. Ghosh, D. Malioutov, and K. S. Meel, "Efficient learning of interpretable classification rules," in *Proc. of JAIR*, 2022.

[5] B. Ghosh and K. S. Meel, "IMLI: An incremental framework for MaxSAT-based learning of interpretable classification rules," in *Proc. of AIES*, 2019.

[6] B. Ghosh, D. Malioutov, and K. S. Meel, "Classification rules in relaxed logical form," in *Proc. of ECAI*, 2020.

[7] L. Ciampiconi, B. Ghosh, J. Scarlett, and K. S. Meel, "A MaxSAT-based framework for group testing," in *Proc. of AAAI*, 2020.

[8] B. Ghosh, M. E. Ali, F. M. Choudhury, S. Hasan, T. Sellis, and J. Li, "The flexible socio spatial group queries," in *Proc. of VLDB*, 2018.

[9] S. H. Apon, M. E. Ali, B. Ghosh, and T. Sellis, "Social-spatial group queries with keywords," *ACM Transactions on Spatial Algorithms and Systems (TSAS), 2021*, 2021.