Research Statement

Bishwamittra Ghosh

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

My research interest is on fairness and interpretability in machine learning, specifically as it applies to safety-critical domains. Traditional machine learning, particularly deep learning, is known for producing unfair predictions towards certain sensitive demographic groups and for generating uninterpretable black-box predictions. In my dissertation research, I design algorithms to *verify fairness* [1, 2], *explain the sources of unfairness* [3], and *learn interpretable rule-based classifiers* [4, 5, 6]. Prior approaches to these problems are often limited by scalability, accuracy, or both. To overcome these limitations, I closely integrate automated reasoning, formal methods, and statistics with fairness and interpretabilit to develop scalable and accurate solutions.

My research has flourished through multiple collaborations and internships in industry and academia. In addition to my work on fairness and interpretability, I have collaborated on research problems related to group testing [7] and social-spatial group queries [8, 9]. Our work has been published in leading 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 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.

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. However, such techniques were limited in terms of increasing confidence for wide deployment. Consequently, recent verifiers aim to achieve verification beyond a finite dataset and instead focus on the probability 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 [1] and a more tailored approach for linear classifiers [2]. 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.

Explaining Fairness Metrics: Identifying the Sources of Bias

Fairness metrics can 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 the *Fairness Influence Function* (FIF) to quantify the contribution of an individual feature and the intersection of multiple features to the resulting bias [3]. 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.

Research Thrust 2: Interpretable Rule-based Machine Learning

Interpretable machine learning often employs rule-based classifiers, which use a set of rules to represent the decision boundary. The interpretability of such classifiers depends on the size of the rules: smaller rules with higher accuracy are preferred in practice. However, this presents a challenge when dealing with large datasets, as interpretable classification learning becomes a combinatorial optimization problem that suffers from poor scalability. To address this issue, we propose an incremental learning framework for interpretable rule-based classification on large datasets. Our framework combines maximum satisfiability (MaxSAT) and mixed integer linear programming (MILP) with mini-batch learning.

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. 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. Finally, we apply IMLI to learn popular interpretable classifiers such as decision lists and decision sets.

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, as described in [6]. Specifically, we consider relaxed definitions of the standard OR/AND operators in propositional logic by allowing 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 motivated 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. However, the combinatorial structure of these rules results in exponential succinctness, and na"ive 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.

Future Research Plans

My long-term research plan is focused on designing efficient and scalable algorithms for machine learning, with a particular emphasis on their trustworthiness in safety-critical applications. To achieve this goal, I plan to work in a collaborative environment, where I can better understand the challenges arising in real-world applications and use advances in machine learning and formal methods to solve them. In pursuit of this vision, I have identified several key research themes that will guide my work.

Fairness and Interpretability As a Service. I believe that in the future, machine learning will serve as an alternative decision-maker to humans in various domains, including law, education, and transportation. However, in high-stakes and safety-critical applications, black-box algorithms are expected to provide higher transparency. Therefore, it is crucial to achieve fairness, interpretability, robustness, and privacy in machine learning models. However, traditional machine learning models, such as deep learning, are often data-hungry, making it challenging to certify and verify properties such as fairness and interpretability in complex models and large datasets. With this in mind, my research aims to develop efficient algorithms for ensuring the fairness and interpretability of deep learning models, transformer-based natural language processing (NLP), and computer vision.

Counting and Optimization Problems. In my previous research, I focused on formulating fairness and interpretability in machine learning as counting and optimization problems. I developed algorithms based on formal methods and incremental solving, which resulted in both higher scalability and better accuracy. Building on this work, I plan to extend these techniques to solve counting and optimization problems 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 are your features?”: Computing fairness influence functions with global sensitivity analysis (under review),” 2023.

[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.