Research Statement

Bishwamittra Ghosh
Scientist
Institute of High Performance Computing (IHPC), A*STAR, Singapore
https://bishwamittra.github.io

My research is on fairness and explainability in machine learning applied in safety-critical domains. Traditional machine learning, particularly deep learning, is known for unfair predictions towards marginalized sensitive groups and for generating black-box predictions. In my research, I design algorithmic framework to *formally quantify fairness in machine learning* [1, 2], *explain the sources of unfairness* [3], and *learn explainable rule-based classifiers* [4, 5, 6]. Prior approaches to these problems are often limited by scalability, accuracy, or both. To address the limitations, I closely integrate automated reasoning, formal methods, and statistics with fairness and explainability to develop scalable and accurate solutions.

During my PhD, I have research collaborations and internships in academia and industry. In addition to fairness and explainability, I have collaborative research on group testing [7], social-spatial group queries [8, 9], and hypergraph core decomposition [10]. We publish our research in leading conferences and journals in artificial intelligence and machine learning: AAAI (2022, 2021, 2020), JAIR (2022), FAccT (2023), ECAI (2020), and AIES (2019); and databases: VLDB (2023, 2018) and TSAS (2022).

Research Thrust 1: Fairness in Machine Learning

Fairness in machine learning focuses on quantifying and mitigating the bias or unfairness of the prediction of the classifier towards different sensitive groups in the data. To quantify bias in algorithmic decision-making, multiple fairness metrics have been proposed based on societal norms and believes. However, there has been insignificant progress in *formally quantifying existing fairness metrics*. In addition, fairness metrics measure the overall bias of a classifier, but they cannot *explain the sources of bias*. Therefore, our research focuses on two key aspects: formally quantifying bias of a classifier and explaining its sources.

Probabilistic Fairness Quantification

In probabilistic fairness quantification, we formally quantify the bias of a classifier given the distribution of input features—essentially beyond a finite dataset. We propose two approaches to the problem: a general approach for finite classifiers encoded as Boolean formulas [1] and a specific approach for linear classifiers [2].

Fairness Quantification via SSAT. The key idea in quantifying group fairness metrics is to compute the maximum (resp. minimum) probability of predictions of the classifier across all sensitive groups—the probability of selecting White-male vs. Black-female candidates in job applications. We propose a stochastic satisfiability (SSAT) based framework, called Justicia [1], for computing such probabilities. More specifically, the maximum probability becomes the solution of an existential-random (ER)-SSAT formula—we encode the classifier as a Boolean formula, the feature distribution via random Boolean variables, and compute the maximum conditional probability of the satisfaction of the formula for existentially quantified sensitive features. In the presence of multiple sensitive features resulting in exponentially many sensitive groups, SSAT efficiently finds the most (resp. least) favored group by the classifier, thanks to the progress in satisfiability (SAT) solving, and particularly in weighted model counting problem. In experiments, Justicia is more scalable in the fairness quantification of tree-based classifiers than existing SMT or sampling methods.

Tractable Fairness Quantification with Feature Correlation. We extend Justicia to consider feature correlations for an accurate fairness quantification [2]. We consider a Bayesian network to represent the conditional distribution of features—the SSAT formula grows with the complexity of the Bayesian network, calling for a more scalable solution. Therefore, we demonstrate a tractable fairness quantification for linear classifiers by proposing a stochastic subset sum problem, which admits an efficient dynamic programming solution with pseudo-polynomial complexity. Experimentally, Justicia becomes more accurate and scalable than existing fairness verifiers for linear classifiers.

Explaining Fairness: Identifying Sources of Bias

We combine both explainability and fairness in machine learning and propose a framework for explaining fairness. We formalize *fairness influence functions* (FIFs) to quantify the contribution of an individual feature and the intersection of multiple features to the resulting bias of the classifier [3]. Based on global sensitivity analysis, we propose a model-agnostic framework, called FairXplainer, to estimate FIFs. The key idea is to represent fairness metrics using the variance of predictions and apply variance decomposition to compute FIFs. In experiments, FIFs are highly correlated with fairness interventions and demonstrate a higher granular explanation of unfairness through intersectional influences, unlike existing local explainability methods. In addition, FairXplainer approximates bias via FIFs with lower error than prior methods across classifiers such as neural networks and SVMs.

Research Thrust 2: Explainable Rule-based Machine Learning

We learn classifiers explainable by design, such as rule-based classifiers. In rule-based classifiers, such as decision lists and decision sets, the decision boundary is explained using a set of rules relating input features to class prediction. The explainability of such classifiers depends on the size of the rules—smaller rules with higher accuracy are preferred in practice, such as by practitioners in the medical domain. Thus, explainable classification learning becomes a combinatorial optimization problem suffering from poor scalability in large datasets. We propose an incremental learning framework for rule-based classification by combining the progress in maximum satisfiability (MaxSAT) and mixed integer linear programming (MILP).

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 quantify 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," in *Proc. of FAccT*, 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.

[10] N. A. Arafat, A. Khan, A. K. Rai, and B. Ghosh, "Neighborhood-based hypergraph core decomposition," *Proc. of VLDB*, 2023.