Research Statement

Bishwamittra Ghosh

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

My research interest is in fairness and interpretability in machine learning applied in safety-critical domains. Traditional machine learning, specifically deep learning, is infamous for its unfair/biased prediction towards certain demographic groups, and for providing uninterpretable black-box predictions. To meet these challenges, I design scalable algorithms to verify and understand the bias of machine learning classifiers and to learn interpretable classifiers. In my research, I closely integrate automated reasoning, formal methods, and statistics with fairness and interpretability in machine learning.

In my doctoral dissertation, I have focused on the safety-critical aspects of machine learning: **fairness** and **interpretability**. Particularly, we have proposed algorithmic frameworks for *verifying fairness of classifiers* [1, 2], *identifying sources of unfairness of classifiers* [3], and *designing efficient interpretable classifiers* [4, 5, 6]. Fairness verification provides a formal certificate of whether a classifier achieves the desired level of fairness on specified data distribution. For identifying the sources of unfairness, we compute fairness influence functions of input features as their contribution towards the bias of the classifier. In interpretable machine learning, we efficiently learn interpretable rule-based classifiers alternative to black-box classifiers. In particular, we propose a scalable learning framework through an incremental learning to learn accurate and small interpretable classification rules while enabling learning on large datasets.

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]. Our works have been published 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 Summary

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. In our research, we focus on two fundamental problems in fairness: (1) verification of fairness and (1) identification of sources of unfairness.

Probabilistic Fairness Verification. Given the distribution of input features and a classifier trained on samples from the same distribution, our probabilistic fairness verifier, called Justicia, certifies whether the classifier achieves the desired level of fairness according to a group fairness metric [1]. Being a probabilistic verifier, Justicia robustly estimates the bias of a classifier in contrast to measuring bias on a finite dataset (Figure 1)—one popular notion of bias is statistical parity, which is the difference between the maximum and minimum conditional probability of a binary classifier's positive prediction over all sensitive groups. Based on a stochastic satisfiability (SSAT) formulation, Justicia is the *first fairness verifier for compound sensitive groups* (combination of multiple sensitive features, such as Black-female and White-male), as highlighted in Figure 2. Methodologically, Justicia encodes the classifier as a propositional formula in Conjunctive Normal Form (CNF), and verifies fairness by solving weighted model counting problems, specifically through an SSAT reduction. For example, the maximum conditional probability of the classifier's positive prediction is equal to the solution of an SSAT formula, where the sensitive features are assigned *existential* quantifier to derive their *maximal* assignment such that the satisfying probability of the CNF classifier becomes maximum for random values of non-sensitive features. Similarly, to compute the minimum conditional probability of
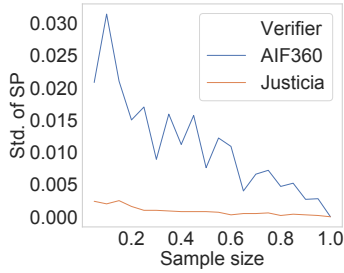
Figure 1: Standard deviation in the estimation of statistical parity (SP) for different sample sizes (sample size = 1 denotes the entire dataset). Probabilistic verifier Justicia is more robust with the variation in sample size than AIF360 estimating statistical parity on a finite dataset.
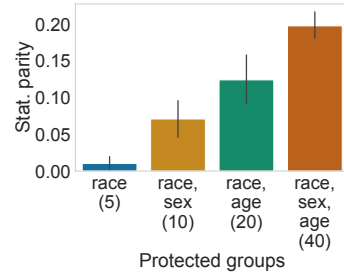

Figure 2: Statistical parity verified by Justicia for different sensitive/protected groups in the Adult dataset. The number within parenthesis in the xticks denotes total compound sensitive groups. The classifier demonstrates more unfairness (higher statistical parity) when considering more groups.

classifier's positive prediction, it suffices to assign sensitive features with *universal* quantifiers to optimize for their *minimal* assignment. Consequently, Justicia benefits from future improvement in SSAT solving, as the modeling of fairness verification into SSAT is separate from solving SSAT formulas. Also, multiple group fairness metrics can be verified in the same SSAT formulation by imposing additional conditionals on the CNF classifier. In experiments, Justicia verifies the fairness of decision tree and CNF classifiers with superior performance in both accuracy and scalability than existing SMT or sampling-based verifiers.

Scalable Fairness Verification with Feature Correlations. In our follow up work on probabilistic fairness verification, we specifically focus on linear classifiers by proposing a verifier FVGM [2]. We improve both the scalability (Figure 3) and accuracy of verification by proposing a stochastic subset-sum problem (S3P) based reduction and enabling feature correlations represented as a Bayesian Network. Theoretically, S3P achieves a pseudo-polynomial complexity based on dynamic programming in contrast to the NP^{PP} hardness of SSAT. Also, S3P reduction for linear classifiers is efficient as it avoids the expensive pseudo-Boolean to CNF encoding in Justicia [1]. Furthermore, FVGM demonstrates significantly higher accuracy than prior verifiers because of considering correlated features. In applications of fairness verifiers, FVGM tests the effectiveness of multiple fairness enhancing and fairness attack algorithms on popular datasets in fairness literature.

Fairness Influence Functions. Our next research quest is to *quantify the influence of different features on the incurred bias of a classifer on a dataset* [3]. To this end, we propose a framework of Fairness Influence Function (FIF), and compute it as a scaled difference of the conditional variances in the classifier's prediction. Based on global sensitivity analysis (GSA), 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 (Figure 4), and detects the increase/decrease in bias due to fairness 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 through a set of rules comprising input features. Examples of such classifiers include decision trees, decision lists, and decision sets. The interpretability of rule-based classifiers is in general related to the size of the rules, where smaller rules are considered more interpretable. To learn such a classifier, the brute-force direct approach is to consider an optimization problem that tries to learn the smallest classification rule achieving close to maximum accuracy. This optimization problem is computationally intractable due to its combinatorial nature and thus, the problem is not scalable in large datasets. To this end, our research
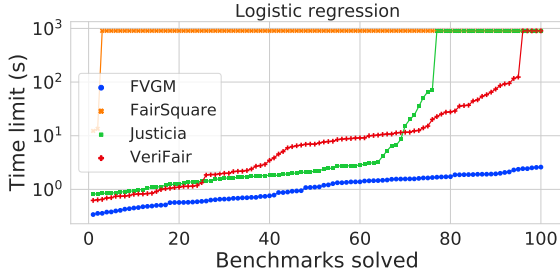
Figure 3: A cactus plot to present the scalability of different fairness verifiers in solving 100 fairness verification benchmarks. FVGM achieves the best scalability results by solving all benchmarks with 1 to 2 orders of runtime improvement.
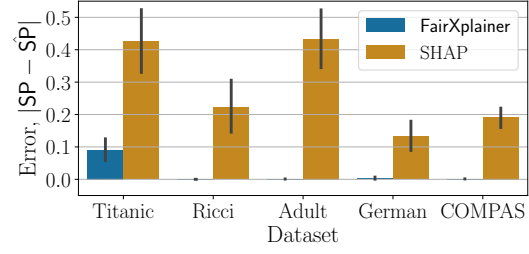
Figure 4: Comparing FairXplainer and local explanation method SHAP on the estimation error of statistical parity, computed as the sum of FIFs of all subset of features [3]. Lower values on the Y-axis denote a better result. FairXplainer has significantly less error than SHAP.

goal is to study the triangular relationship among the accuracy, interpretability, and scalability of learning rule-based classifiers.

Incremental Learning of Interpretable Classfication Rules. We propose an incremental learning framework, called IMLI [4, 5], based on maximum satisfiability (MaxSAT) for synthesizing 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 (Figure 5) 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.

(a) Samples: $32,561$
Features: 94

(b) Samples: $107,696$
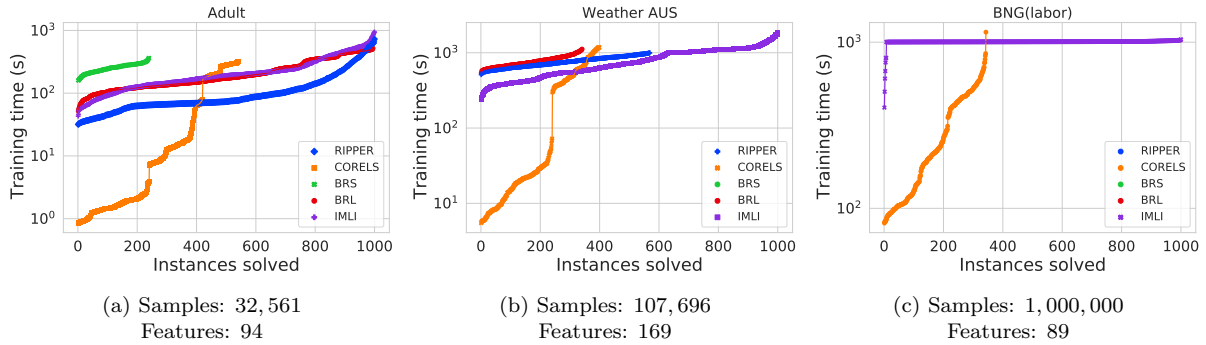Features: 169

(c) Samples: $1,000,000$
Features: 89

Figure 5: Results on the scalability of IMLI compared to existing interpretable classifiers, presented using cactus plots for datasets with varied dimensions. As datasets become large, IMLI becomes the only interpretable classifier to classify all of 1000 classification instances for a dataset. The success of IMLI is attributed to its incremental learning wrapping a MaxSAT-based formulation.

Classification Rules in Relaxed Logical Form. We extend our incremental learning framework to learning more relaxed representation of classification rules [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 definition, we introduce relaxed-CNF classification rules motivated by the popular usage of checklists in the medical domain. Relaxed-CNF generalizes widely employed rule representations including CNF, DNF, and decision sets. While the combinatorial structure of relaxed-CNF rules offers exponential succinctness, the naïve learning techniques are computationally expensive. To this end, we propose a incremental mini-batch learning procedure, called CRR, that employs advances in the Mixed-Integer Linear Programming (MILP) solvers to efficiently learn relaxed-CNF rules. Our experimental analysis demonstrates that CRR can generate relaxed-CNF rules, which are more accurate and sparser 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 future, with applications in law, education, transportation etc. In these 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 in nature, certifying safety-critical properties will be challenging in complex models and 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] <u>B. Ghosh</u>, D. Basu, and K. S. Meel, "Justicia: A stochastic SAT approach to formally verify fairness," in *Proc. of AAAI*, 2021.

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

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

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

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

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

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

[8] <u>B. Ghosh</u>, 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, <u>B. Ghosh</u>, and T. Sellis, "Social-spatial group queries with keywords," *ACM Transactions on Spatial Algorithms and Systems (TSAS), 2021*, 2021.