Research Statement

Bishwamittra Ghosh

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

The last decades have witnessed significant progress in machine learning with a host of applications of algorithmic decision-making in different safety-critical domains, such as college admission, recidivism prediction, employee hiring, and clinical processes. In these domains, the decisions by machine learning classifiers have far-reaching consequences, potentially influencing human lives in various ways. Consequently, fairness and interpretability of the deployed classifiers are of paramount importance to the end-users, who expect to have a fair decision from an algorithm and an interpretable decision that they can understand. My research vision is to improve the safety-critical properties of machine learning, such as fairness and interpretability, and enhance the applicability of machine learning for social good.

Research Summary

Towards **fair and interpretable machine learning**, my research so far has focused on fairness verification, identification of sources of the unfairness of classifiers, and designing efficient interpretable classifiers. Fairness verification provides a formal certificate of whether a classifier has achieved the desired level of fairness on specified data distribution. In identifying the sources of unfairness, we propose to compute fairness influence functions of input features as their contribution towards the bias/unfairness of the classifier. In interpretable machine learning, we focus on designing interpretable rule-based classifiers alternative to black-box classifiers. In particular, we propose a scalable learning framework to generate accurate and small interpretable rules while enabling learning on large datasets. In the technical aspects of my research, I have closely integrated formal methods with machine learning and have designed frameworks that provide improved scalability and performance guarantees.

My first research direction is on **fairness in machine learning**. Classifiers relying on mere accuracy-centric learning objectives may become unfair towards certain demographic groups in the data. To this end, my research addresses the fairness verification problem of group and causal fairness metrics and multiple fairness-enhancing and fairness-attacking algorithms. In fairness verification, our goal is to robustly estimate the unfairness of a classifier given the distribution of features. We propose a formal approach for fairness verification of linear classifiers [1] and classifiers represented as logical formulas [2]. Furthermore, we verify fairness by considering feature correlations represented as a Bayesian network. We have compared our verifier with state-of-the-art fairness verifiers and demonstrated higher accuracy and scalability of our approach. Next, we focus on computing fairness influence functions to quantify the contribution of input features on the incurred bias of the classifier on a dataset [3]. In particular, we compute individual and inter-sectional influences to capture the dynamics of correlated features in the resultant bias. Based on global sensitivity analysis, our approach computes influences more accurately than the state-of-the-art.

My second research direction is on **interpretable machine learning** with a goal of efficiently learning interpretable rule-based classifiers. Rule-based classifiers, such as decision trees, decision sets, and decision lists, are popular interpretable classifiers, and they can explain the inner working of black-box classifiers, such as neural networks. The challenge in interpretable rule-learning is to trade-off between the predictive accuracy and the generation of the smallest classification rule to favor interpretability; thus, the underlying combinatorial optimization problem cannot scale to large datasets. To this end, we propose an incremental learning framework wrapping traditional optimization solvers such as MaxSAT [4, 5] and MILP [6] to learn small and accurate classification rules in datasets containing a million samples.

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 the post-hoc explainability of black-box classifiers [7, 8], group testing [9], and social-spacial group queries [10, 11]. Our works have been published at premier conferences in artificial intelligence and machine learning (AAAI×3, AIES, ECAI) and databases (VLDB, TSAS).

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 plan to design efficient algorithms for the fairness and interpretability of deep models, transformer-based natural language processing (NLP), and computer vision. To this end, I develop better approximate algorithms combining formal methods (SAT/SMT), machine learning, and statistics.

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 apply these techniques in solving similar counting and optimization problems, even in areas beyond machine learning.

References

[1] **Bishwamittra Ghosh**, Debabrota Basu, and Kuldeep S. Meel. Algorithmic fairness verification with graphical models. In *Proceedings of AAAI*, 2022.

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

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

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

[5] **Bishwamittra Ghosh**, Dmitry Malioutov, and Kuldeep S. Meel. Efficient learning of interpretable classification rules (under review). 2022.

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

[7] **Bishwamittra Ghosh** and Daniel Neider. A formal language approach to explaining RNNs. In *arXiv:2006.07292*, 2020.

[8] Daniel Neider and **Bishwamittra Ghosh**. Probably approximately correct explanations of machine learning models via syntax-guided synthesis. In *arXiv:2009.08770*, 2020.

[9] Lorenzo Ciampiconi, **Bishwamittra Ghosh**, Jonathan Scarlett, and Kuldeep S. Meel. A MaxSAT-based framework for group testing. In *Proceedings of AAAI*, 2020.

[10] **Bishwamittra Ghosh**, Mohammed Eunus Ali, Farhana Murtaza Choudhury, Sajid Hasan, Timos Sellis, and Jianxin Li. The flexible socio spatial group queries. In *Proceedings of PVLDB*, 2018.

[11] Sajid Hasan Apon, Mohammed Eunus Ali, **Bishwamittra Ghosh**, and Timos Sellis. Social-spatial group queries with keywords. *TSAS*, 2021.