Confusion Matrix Design for Downstream Decision-making

YIDING FENG, Hong Kong University of Science and Technology, Hong Kong, China
WEI TANG, Chinese University of Hong Kong, Hong Kong, China

We initiate the study of *confusion matrix design*. In this problem, an algorithm designer needs to generate a machine learning model (for a classification task from contexts to labels) which makes predictions for a population of downstream decision makers. The prediction accuracy of the machine learning model is characterized by its confusion matrix, which is a stochastic matrix where each entry encodes the probability of predicting the true label to another label. Each downstream decision maker faces a separate optimization task and will decide his binary action based on his own context, realized prediction given his context, and the confusion matrix selected by the algorithm designer. Decision makers are heterogeneous, as they may hold different contexts. Both the decision makers and the algorithm designer will obtain utilities that are determined by the actions the decision makers take, and their true labels. The goal of the algorithm designer is to design a public confusion matrix that is used for all decision makers subject to some feasibility constraints in order to maximize her net utility.

Viewing the design of the confusion matrix as the design of an information structure (aka., signaling scheme) that carries information about the underlying true outcome, we can study the above problem through an information design framework (Bergemann and Morris 2016, Kamenica and Gentzkow 2011). We consider two common constraints that are naturally motivated by the machine learning literature for the feasibility of the designed confusion matrix:

- Post-processing constraints: the designer is initially given an algorithm and can only generate the predictions by post-processing the predictions from the given algorithms. This constraint is motivated by the scenario when retraining the machine model is costly and impractical.
- Receiver Operating Characteristic (ROC) constraints: the designer can indeed train the machine model but is subject to the inherent uncertainty involved when making predictions/inferences about a population based on the limited training data. This constraint also captures the scenarios where the designed machine prediction needs to additionally satisfy certain properties, e.g., privacy, fairness, etc.

We are interested in the algorithmic aspects of the designer's optimal confusion matrix design problem. Since the designer in our setting is facing a population of downstream decision-makers, the designer's problem closely relates to the *public persuasion* in the information design literature (Dughmi and Xu 2017, Guo and Shmaya 2019, Xu 2020). The key difference here is that the heterogeneity in our setting is from the agents' differing prior beliefs. With this connection, we show that, for the post-processing constraints, under a mild condition on the decision-makers' prior beliefs, there exists a convex programming-based algorithm that can compute the optimal confusion matrix. In addition to the computational result, we also obtain analytical structural results for the special case when the designer's net utility is outcome-independent. Specifically, for the outcome-independent net utility, the optimal confusion matrix can be characterized by a linear program. Furthermore, when the outcome space is binary, we provide a geometric characterization of the optimal confusion matrix. For the ROC constraints, we also show that there exists a convex programming-based algorithm that can compute the optimal confusion matrix. When the designer has the social-aware net utility, we show that the optimal confusion matrix must be binding on the ROC constraint, and either the Type I error is equal to 0 or the Type II error is equal to 0. [1]

REFERENCES

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016.

Shaddin Dughmi and Haifeng Xu. Algorithmic persuasion with no externalities. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 351–368, 2017.

Yingni Guo and Eran Shmaya. The interval structure of optimal disclosure. *Econometrica*, 87(2):653–675, 2019.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Haifeng Xu. On the tractability of public persuasion with no externalities. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2708–2727. SIAM, 2020.

[1] A preliminary one-page abstract of this paper has been accepted at The 16th Innovations in Theoretical Computer Science (ITCS 2025), and a full version of this work can be found at https://arxiv.org/abs/2402.12562