Topics in Selective Classification

Andrea Pugnana

Scuola Normale Superiore
Piazza dei Cavalieri 7, Pisa, Italy
andrea.pugnana@sns.it

Introduction

In recent years, continuous advancements of Artificial Intelligence (AI) have allowed for unprecedented performances of predictive systems. This has brought to the widespread deployment of these methods in many fields, supporting and often automating decision-making. As algorithms are more and more used in socially-sensitive domains, there is a pressing demand for a trustworthy AI (Wing 2021). For instance, the EU Regulatory framework proposal on AI (European Parliament and the Council 2021) rules that high-risk AI systems will be subject to strict obligations before deployment. One such obligation is to ensure "a high level of robustness, security and accuracy". High-risk AI systems include several domains, such as healthcare, justice, hiring, and credit scoring. In many of these cases, AI is typically framed as a probabilistic binary classifier (see, e.g., (Dastile, Çelik, and Potsane 2020) for credit scoring), and the predictions are used to score or rank people.

A pressing question is how to make AI systems robust to challenges in real-life scenarios.

Contributions so far

A first issue is that most AI systems are based on simple associations. This can be detrimental in environments where training data distribution differs from the test one - a problem known as distribution shift - as the model performance can drop. One possible approach to make AI systems more robust is to embed **causality** in the classification learning setting, as it captures cause-effect relationships that are invariant from the specific context under analysis (Zhang et al. 2013). Due to the multi-disciplinary nature of the subject, we reviewed the current state of the art of causality literature to map the potential next steps in this area. As a first result, we published (Nogueira et al. 2022). In particular, I focused on the causal inference part.

Another common issue is that the predictive performance of classifiers is typically not homogeneous over the data distribution. Identifying sub-populations with low performance could be helpful, e.g. for debugging and monitoring purposes, especially in high-risk scenarios. A direction towards improving robustness and accuracy in this context is to lift from the canonical framework of binary classification to **selective classification** one. Selective classification (also known as classification with reject option, or learning to defer) (Chow 1970) extends a classifier with a selection function (reject option/strategy) to determine whether or not a prediction should be accepted. This mechanism allows the AI system to abstain on those instances where the classifier is more uncertain about the class to predict, introducing a trade-off between performance and coverage (the percentage of cases where the classifier does not abstain). The reject option has been extensively studied from a theoretical side (El-Yaniv and Wiener 2010; Franc and Průša 2019). However, state-of-the-art practical approaches and tools are model-specific, e.g., they are tailored to DNNs, as, e.g., in the case of SelectiveNet (Geifman and El-Yaniv 2019) and Self-Adapting Training (SAT) (Huang, Zhang, and Zhang 2020), and focused/experimented mainly on image datasets. During my second year of PhD, we started focusing more on the selective classification framework. Therefore, as a first contribution, I developed a model-agnostic heuristics that is able to lift any (probabilistic) classifier to a selective classifier. The approach exploits both a cross-fitting strategy and results from quantile estimation to build the selective function. I tested our algorithm on several real-world datasets, showing improvements concerning existing state-of-the-art methodologies. The paper describing this work is currently under review.

Another open issue in the selective classification scenario regards the performance metrics. The canonical choice is to use distributive loss functions - where the loss is defined for every prediction in isolation - such as accuracy over accepted instances (selective accuracy). However, there are cases where other measures are more informative, e.g. whenever the classes to predict are imbalanced. A popular choice in this context is Area Under the ROC Curve (AUC). AUC is a metric about the ranking induced by a classifier, for which the loss is determined on pairs of instances. I provided the theoretical and empirical evaluation to ensure the trade-off between AUC improvements and coverage. Also this work is currently under review.

Next steps

In the remaining part of my PhD, I plan to investigate other open issues related to selective classification. As a first con-

tribution, I want to provide an empirical benchmark of existing selective classification methods, as their empirical evaluation in the literature is limited. For instance, the claimed state-of-the-art (Huang, Zhang, and Zhang 2020) is tested only on image data, ignoring tabular and text ones. A comprehensive empirical evaluation of current methods would benefit selective classification researchers by providing a benchmark for a proper comparison and reproducibility of existing methods.

Selective classification is an approach for increasing the trustworthiness of the AI system. Another approach to increase trust consists of explaining AI models. This branch of research is known as Explainable AI (XAI), and it aims to make comprehensible to humans the complex mechanisms that drive an AI output (Guidotti et al. 2019). However, little attention has been devoted to studying how to embed explainability methods into selective classification. Since selective classification is relevant, especially for sensitive contexts, I aim at building a selective function that could be directly explained to humans: generally, the mechanism ruling the abstention is whether the score of the selective function is below a certain threshold. However, such an "explanation" is not sufficiently transparent, as the confidence function is often obtained through a not interpretable model (a black box model). Adding explanations to rejections allows for understanding and characterizing the areas where the classifier is not confident enough, which can help build better classifiers. I plan to study different ways to make **selective classification explainable**:

- explain the selective function directly through a surrogate (explainable) model. Such a feature might help explain why certain instances are being rejected;
- modify local explanation methods to account for the rejection option. Intuitively, allowing the model to abstain means enlarging the decision boundary, making local explanation methods easier and more robust.

As a further topic, I also plan to investigate the fairness implications of selective classification. Recent works (Jones et al. 2021) showed how increasing abstention might decrease accuracy over units belonging to socially sensitive groups. Other works, such as (Schreuder and Chzhen 2021), enforced fairness and rejection constraints while optimizing for accuracy. As a result, they provide a computationally efficient post-processing algorithm for both fairness and rejection option constraints. Starting from existing literature, my goal is to design **fair selective classification** as a new framework that aims to answer the following questions:

- what does it mean to be fair whenever the model has the option to abstain?
- how can we define a fair selective classifier?
- does a fair classifier automatically turn into a fair selective classifier if paired with a standard confidence function? If not, how can we build it?

Finally, I plan to investigate how to make **selective classification robust** to distribution shifts. More specifically, these are the questions I will try to address:

- can specific selective classification features, such as the confidence function, directly mitigate (or amplify) the concerns deriving from shifts?
- can we bridge selective classification with causal domain adaptation to develop a new framework?

Acknowledgements. Work partially supported by the XAI Project, from the European Union's Horizon 2020 Excellent Science European Research Council (ERC) programme under grant agreement No. 834756, and by SoBigData, European Union's Horizon 2020 research and innovation programme under grant agreement No. 871042.

References

Chow, C. K. 1970. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16(1): 41–46.

Dastile, X.; Çelik, T.; and Potsane, M. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Appl. Soft Comput.*, 91: 106263.

El-Yaniv, R.; and Wiener, Y. 2010. On the Foundations of Noise-free Selective Classification. *J. Mach. Learn. Res.*, 11: 1605–1641.

European Parliament and the Council. 2021. Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *2021/0106(COD)*.

Franc, V.; and Průša, D. 2019. On discriminative learning of prediction uncertainty. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 1963–1971. PMLR.

Geifman, Y.; and El-Yaniv, R. 2019. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 2151–2159. PMLR.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5): 93:1–93:42.

Huang, L.; Zhang, C.; and Zhang, H. 2020. Self-Adaptive Training: beyond Empirical Risk Minimization. In *NeurIPS*.

Jones, E.; Sagawa, S.; Koh, P. W.; Kumar, A.; and Liang, P. 2021. Selective Classification Can Magnify Disparities Across Groups. In *ICLR*. OpenReview.net.

Nogueira, A. R.; Pugnana, A.; Ruggieri, S.; Pedreschi, D.; and Gama, J. 2022. Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1449.

Schreuder, N.; and Chzhen, E. 2021. Classification with abstention but without disparities. In *UAI*, volume 161 of *Proceedings of Machine Learning Research*, 1227–1236. AUAI Press.

Wing, J. M. 2021. Trustworthy AI. *Commun. ACM*, 64(10): 64–71.

Zhang, K.; Schölkopf, B.; Muandet, K.; and Wang, Z. 2013. Domain Adaptation under Target and Conditional Shift. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, 819–827. JMLR.org.