Reliable and fair machine learning for risk assessment

Pereira Barata, A.P.

Chapter 6

Conclusions

In the final chapter, we answer the three RQs in Section 6.1. Then, we address the problem statement and clearly identify the research results together with their conclusions in Section 6.2. Lastly, in Section 6.3, research directions are proposed with the intent of furthering reliable and fair data-driven risk assessment applications.

6.1 Answers to the Research Questions

Below, the RQs are reiterated as formulated in Chapter 1. Each question is answered separately.

> **RQ1:** *Given data with missing values, which (a) missing data-handling technique and (b) learning algorithm should be jointly selected such that, regardless of the missing mechanism, the detriment to the downstream task performance is minimal when compared to the non-missing (unavailable) case?*

When dealing with real-world data, a non-MCAR scenario is traditionally assumed. Thus, a viable option is to use the missing-indicator method, encoding the missingness itself [Lipton et al., 2016]. However, this assumption does not always hold; in such cases, the missing-indicator method may even deteriorate the performance of the (downstream) classification task, when compared to an imputation approach.

It is established in Section 2.6 that, under MCAR, the differences in the downstream classification task performance between (a) imputation, and (b) missing-indicator are negligible, if the appropriate learner is used. Specifically via feature selection protocols, it is possible to learn classification models of which the performances are statistically *indistinguishable* across the two different methods used to handle missingness.

Given that real-world data are seldom MCAR and that, even under the MCAR mechanism, the performance decrease can be made indistinguishable between imputation and missing-indicator, the answer to RQ1 is, therefore, that the missing-indicator method, in conjunction with decision tree-based learners —particularly via gradient boosting— is an adequate solution. In doing so, the detriment to classification performance should be minimal, whether under MCAR or non-MCAR.

One limitation to our answer is that, in our experimental design, both training and test sets are identically distributed with respect to the missing mechanism. Further studies could be conducted to assess classifier performance under a *slightly* different, yet most impactful scenario, in which the training set (i.e., the available training data) and test set (representing the model deployment) are *differently* distributed with respect to the mechanism of missingness. For example, if a classifier were to be learned from data under MCAR, coupled with a missing-data handling technique, what would the expected performance be if the test data were either *non-missing* or non-MCAR?

> **RQ2:** *Given data with label noise, how can noisy-samples be (a) adequately detected, and (b) used to learn a well-performing model?*

In Chapter 3, the notion of *crosslier* is introduced. Crossliers are anomalous instances with respect to a domain-sensible category. These instances may be misconducts as their characteristics position them farther from their category cluster, across the decision boundary towards one or more other categories. With respect to the Inspectorate, waste category crossliers were presented as potential misconducts.

By learning well-calibrated probabilistic classifiers, it is possible to use the output probabilities of trained models towards a set of new samples, where lowest posterior probability indicates a most probable crosslier. To achieve this, we propose the EXPOSE method. Our method addresses the entire dataset in a CV manner, generating well-calibrated sample probabilities for the test sets. As a result, all samples have comparable crosslier scores with which crosslier detection may be performed. This process answers RQ2(a).

Yet, we denote that (as established in Chapter 1), data within the Inspectorate represent the *administrative* reality. What might be considered misconduct, could in fact be simply a data quality issue (e.g., an entry error). The distinction is, however, dependent on which reality the data represents. For this reason, special care is required when considering the real-world *meaning* of a sample with a high crosslier score. Ultimately, a case-by-case approach should be taken if deployment is to be reliable in the Inspectorate.

The amount of noise in observations can be quantified so long as there is access to the posterior probabilities of those samples. After quantification using the EXPOSE method, the probabilities may be manipulated into weights reflective of the *clean* target distribution. In Chapter 4, we proposed a viable weighting scheme in which we consider the weight of an observation as the log-odds of its posterior probability. This proposed learning method is coined DENOISE. We showed that the resulting models learned on noisy data with this sample weighting scheme are well-performing. This result answers RQ2(b).

Although we are confident on the performance of our method, we must denote that the manner by which noise was artificially generated is a limitation: instead of a random univariate approach to noise generation, a more complex approach could have been deployed. For example, selecting features of which the importance for the final classifier is high. Further studies should be conducted, where learning performance could be assessed not only in terms of noise probability, but also in terms of the generation of noise itself.

> **RQ3:** *How can we, from biased data, learn a model tunable with respect to the performance-fairness trade-off such that the selection of the trade-off point is made intuitive for the relevant stake-holders?*

In Chapter 5, a decision tree learning framework is leveraged by incorporating a threshold-agnostic fair classification splitting criterion termed SCAFF. The splitting criterion is formulated as a weighted linear combination of (a) the AUC towards the class label and (b) the strong demographic parity, implemented as the AUC towards the sensitive attribute. To provide tunability with respect to performance-fairness trade-off, an orthogonality parameter $\Theta \in [0, 1,]$ is part of the splitting criterion.

By analysing the performance-fairness trade-off curve, an appropriate value of Θ can be selected according to the application domain requirement(s). In addition, multiple multicategorical sensitive attributes may be addressed simultaneously by minimising the maximal sensitive AUC across all sensitive attributes as the term in the splitting criterion. Through experimentation and comparison with other fair splitting criteria, we validated our method for various datasets and sensitive attribute scenarios. Our proposed SCAFF method, specifically via the orthogonality parameter Θ, answer RQ3.

While the implementation of our method easily extends to bagging and boosting frameworks, the computational costs associated with either of the extensions are not equivalent. Concretely, the computation of AUC in the boosting framework needs to follow the traditional and more time complex approach: unlike with bagging where each tree node can be represented as either a positive or negative class prediction, nodes in boosted trees contain samples each with their sample-specific prediction score.

6.2 Answer to the Problem Statement

We may now give an answer to the PS based on the answers to the RQs provided above.

> **PS:** *How can machine learning methods advance data-driven risk assessment by the Inspectorate in a reliable and fair manner?*

In our research, we focused on two aspects inherent to the development and deployment of high-risk AI: reliability and fairness. In particular, we highlighted classification models towards risk assessment under the current EU movement towards trustworthy AI. In line with our research, we first address reliability and thereafter fairness.

At the start, we transposed the aspect of reliability into that of two important data quality issues: *missingness* and *noise*. For *missingness*, as described in Chapter 2, we experimented with different methods which handle missing data, by addressing RQ1. Based on our results, we may conclude that with real-world data, a missing-indicator method in conjunction with a decision tree-based learner is a viable solution to address the problem of missing data.

With respect to *noise*, discussed in Chapter 3, we considered *noise in data* as potential real-world misconduct. Hence, we proposed a method to detect it by addressing RQ2(a). Moreover in Chapter 4 we leveraged this noise detection towards a novel method of learning adequately-performing models from noisy data by addressing RQ2(b).

With respect to *fairness*, we claim that countering biased data is crucial in risk assessment. In Chapter 5, we consolidated this claim by answering RQ3. Our main result is proposing a decision tree learning algorithm which takes into account bias in data to produce a classifier that performs adequately and is easily-tunable in terms of the performance-fairness trade-off.

We contribute to reliable and fair machine learning methods for risk assessment by the Inspectorate via Chapters 2– 5. On top of that, we reinforce the principle given in Chapter 1 towards trustworthy AI: *the expertise of domain experts must not be replaced by automation*, but rather *enhanced* by it. To note, while we state that we contribute towards reliability and fairness, we wish to make very explicit that the road to data-driven solutions for the problem of risk assessment has merely started being paved.

The opportunities for continuation are ample, and present their own set of stimulating challenges: not only in terms of the actual implementation of the solutions presented in this thesis into the daily operations of the Inspectorate, but also in terms of the myriad of other data-related issues that this thesis did not cover. As such, we follow with suggestions for potential future research in this domain.

6.3 Future Research

Dealing with real-world data in the inspection domain is (1) a sensitive and laborious undertaking and (2) a stimulating research area. One straightforwards research direction is to procure a joint solution to the individual issues derived from the three investigations (RQs 1–3). To put it differently, it would be advantageous to develop learners which are able to simultaneously address (1) missingness, (2) noise, and (3) fairness, while remaining highly performing. This could be achieved by applying the missing-indicator method, in conjunction with the incorporation of our sample weighing scheme into our fair tree learner.

Another prominent research direction, specifically with application in the Inspectorate, may be to broaden the scope of how data is represented, prior to learning. In other words, while the classification problems remain the same (e.g., identifying misconduct in different sub-fields), there is still opportunity for enhancing the feature representation/embedding process which, in turn, might promote superior model performance. A common approach to adequately model the complex interactions between individuals in a dataset, builds on concepts from the relatively young field of network science [Barabási, 2016]. It is widely accepted in the literature that networks are the *de facto* data architectures towards modelling the behaviour and dynamics of real-world systems, as further corroborated by our most recent work towards automated and fair ship targetting [de Bruin et al., 2022]. Therefore, we believe that more applications should be tractable following such approaches.

Lastly, another problem of relevance to the Inspectorate relates to the following. By definition, historical inspection data is neither an independent nor identically distributed sample of the entire population. In other words, since the function of the inspectors of the ILT is to select for inspection the cases which are of highest risk, the selection will generate data samples of which the distribution in feature space is not representative of the entire pool of cases. This makes it difficult to learn a classifier that distinguishes between more or less risky samples adequately due to this under-representation of feature space, which may lead to inspection blind-spots; i.e., regions in feature space which are considered by the inspectors and, consequently, the classifier to be of non-interest when in fact they pertain to risk behaviour. One way to address this issue would be to deploy active sampling methods such that, rather than targetting samples of which the true (unseen) label most probably indicates risk, samples would be selected towards increasing the generalisation of a given model. Another approach would be to leverage the study of co-domain adaptation (or covariate shift), under which the assumption is that the training and deployment distributions are different between each other.

Ultimately, our goal is to enact tangible change in the way the Inspectorate operates. For this purpose, however, action by the responsible agents, and not solely the machine learners, is required. The shift towards a data-driven Inspectorate has only just begun.