Reliable and fair machine learning for risk assessment

Pereira Barata, A.P.

Summary

In Chapter 1, we introduce the current movement towards trustworthy AI, as well as its application in risk assessment activities, which is the motivation of this work. In particular, we further elaborate on the practical use-cases within the Inspectorate of the Netherlands, in which reliable and fair models are required, given the high-risk nature of the domain. Our aim is to promote a paradigm-shift towards data-driven approaches —via machine learning techniques— to be used by the agents of the Inspectorate (i.e., inspectors).

Explicitly, the focus of this thesis is on classification models. Data generated by the Inspectorate is often tabular. As such, we focus on tree-based learning architectures. Moreover, two real-world data traits detrimental to classifier learning are considered: (1) data quality, viz missingness and noise; and (2) bias in the data. The goal is to generate, via learning techniques which combat these drawbacks, adequately-performing classifiers (measured in AUC) which are both (1) reliable, and (2) fair, respectively addressing each data trait. As such, we formulate the PS as follows.

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

Thereafter, we decompose the PS into three tractable RQs. Below, these are stated, together with the main results gathered from answering them.

When dealing with missing data, one of three distinct mechanisms of missingness occurs: (1) MCAR; (2) MAR; or (3) MNAR (see List of Abbreviations). Depending on the missing mechanism, the efficiency of missing-data handling techniques varies. This is measured in the performance of the downstream task (i.e., classification performance) which is also dependent on the selected learner.

On the one hand, under a non-MCAR scenario, an adequate approach is to encode missingness; i.e., the missing-indicator method. On the other hand, imputation is preferred under MCAR. Although real-world data are rarely MCAR (thus justifying a missing-indicator approach), the assumption over the missing mechanism is not always true.

Although testing for MCAR is possible, the result is not a guaranteed truth. With this uncertainty in mind, the first RQ is formulated accordingly.

> **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?*

In Chapter 2, the MCAR scenario is studied. In a controlled environment, missing data is artificially generated for different proportions of missingness; several imputation and missing-indicator methods are deployed. Distinct classifiers are then constructed via distinct learning algorithms. The resulting performance of each learner-data handling pair is retrieved.

The results show that imputation methods provide a superior classification performance, compared to the missing-indicator method, under the MCAR scenario. Yet and most importantly, for the classifiers learned via decision tree-based gradient boosting, the differences in performance derived from the two distinct data-handling techniques becomes negligible. Hence, the answer to RQ1 is that the missing-indicator method, in conjunction with a decision tree-based learner —particularly via gradient boosting— should be used regardless of the missing mechanism.

Noise in data deteriorates the classification performance and may present itself as either feature noise or (class) label noise, of which the latter is more detrimental. Under label noise, three noise-generating mechanisms exist: (1) NCAR; (2) NAR; and (3) NNAR. Handling noise in data generally entails generating noisy-sample detection scores, traditionally addressed by leveraging classifiers learned on the noisy data, which is an endeavour in itself.

Within the scenario of the Inspectorate, noisy data may represent misconduct; e.g., companies manipulating waste transportation reports to lower the costs associated with each waste type. Therefore, it is of importance to detect these misconducts and simultaneously learn classifiers from the available data which contain them. Hence, the second RQ is a compound one and decomposed into RQ2(a) and RQ2(b).

> **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, we introduce the term *crosslier*. It denotes a sample with disharmonious feature values. Concretely, crossliers are a special case of outlier with respect to some overarching category feature. They are samples which exhibit label noise with respect to the category feature, and potentially feature noise, relating to the NNAR mechanism. To detect crossliers, we propose the EXPOSE method.

The EXPOSE method evaluates samples in a CV-manner, such that all samples in the data are evaluated. Each training set is used to produce a well-calibrated classifier via Platt scaling. The classifier is then deployed on the corresponding test samples. From the classifier output $f(x)$, the crosslier score is $-\log_2[f(x)]$. By evaluating the performance of our method in a controlled setup, we validate EXPOSE. Then, we answer RQ2(a).

Chapter 4 follows logically, utilising the core principal of EXPOSE —and its established validity— towards classifier learning with noisy data. We term our compound method DENOISE. The DENOISE method entails two steps.

First, well-calibrated probabilities are computed for each sample following the EXPOSE method. Second, the probabilities are used to generate individual sample weights, such that the weight is the log-odds of the output sample probability. Under a logistic loss function applied via a gradient boosting decision tree learner, an adequately-performing noise-resilient classifier is produced. In a controlled experimental environment, we validate our method, thereby answering RQ2(b).

Learning from biased data leads to biased models. Even when the conditions of the data gathering processes are ideal, bias in data may still occur due to historical factors (e.g., gender wage gap). To address this issue, different fair machine learning techniques can be used. The purpose of these techniques is to generate models of which the output is independent of some sensitive attribute, such as gender; i.e., fair models. Moreover, several measures of fairness exist, of which the strong demographic parity is analogous to the AUC performance measure. Generally, the greater the fairness of a model, the lesser its predictive performance: here we speak of the performance-fairness trade-off.

Regarding the Inspectorate, biased data may represent a form of bias given some distinct criterion. For example, ships sailing under specific country flags are deemed of higher risk and hence more targetable than others, which may result in confirmation bias. Moreover, since the country flag is easily mutable, it enables companies to bypass the inspection selection protocol. From these considerations, we formulate our RQ3.

> **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, we propose a fair decision tree learning algorithm via strong demographic parity. We do so by defining a compound splitting criterion, termed SCAFF —splitting criterion AUC for fairness— which is tunable with respect to the performance-fairness trade-off. It leverages several sensitive attributes concurrently, of which the values may be multicategorical.

SCAFF is defined as a weighted linear combination of (a) the traditional AUC classification performance, and (b) the strong demographic parity scaled to the range of the AUC. We term the scaled fairness measure *sensitive* AUC. The closer the sensitive AUC is to 0.5, the greater the fairness of the model. By incorporating an orthogonality parameter $\Theta \in [0, 1]$ implemented as an elastic net-like weight to the performance and the fairness terms, the performance-fairness trade-off is tunable. In the case of $\Theta = 0$, a traditional (potential) non-fair classifier is generated, and increasing Θ augments the fairness of the final model. By comparing SCAFF to other fair splitting criteria in a controlled experiment, we validate our approach and answer RQ3.

The conclusion of the thesis is that, under the current movement towards trustworthy AI in Europe, shifting the current risk assessment paradigm to a more data-driven methodology is a delicate yet feasible venture via reliable and fair machine learning. Given the high-risk nature of risk assessment activities, and the characteristics of the data generated by them, we believe that technical methods can ensure the adequacy of the final learned models. In particular, the issues associated with learning a classification model from biased and low-quality data can be successfully addressed, producing adequate models.