Reliability Assessment Methods for eXplainable Artificial Intelligence

Sara Narteni
Doctoral Program in Artificial Intelligence - Industry 4.0 (37.th cycle)

Abstract

The rapid evolution of machine learning (ML) algorithms in recent years, along with the availability of powerful technological infrastructures supporting them, is generating more and more interest in adopting artificial intelligence (AI) solutions in many domains. These also include safety-critical ones, such as healthcare, smart mobility or cybersecurity, which deserve a special attention. Trustworthy AI (TAI) framework has been introduced to individuate the requirements that AI systems should fulfill during all their design, development, and deployment stages, in order to be lawful, ethical, and robust. Among these, eXplainable AI (XAI) and Reliable AI (RAI) emerge as two key properties that must be guaranteed in pursuing a trustworthy AI solution. XAI collects a broad set of techniques providing insights to the logic behind AI-based decision-making systems. On the other hand, RAI consists in ensuring that such systems work with adequate performance both at the model level and at a system level. Current research on trustworthy AI recognizes the important role of XAI in reliability assessment, since a lack of the first compromises the latter. However, a completely unified vision on XAI and RAI is still little investigated.

In this context, this dissertation focuses on studying how to integrate reliability in rule-based binary classification methods, thus attempting to establish a link between XAI and RAI. The focus is posed on rule-based classifiers of the *if-then* kind, since they constitute, at least in principle, one of the most simple forms of XAI. Nevertheless, this is not always true, since rule-based classifiers often reveal to be sensitive to the complexity of the data under analysis, providing decision rules that are not so straightforward to comprehend, especially by non-experts of a given application domain. Therefore, the preliminary contribution of this dissertation is the introduction of innovative *rule similarity* metrics that allow to compare and extract knowledge from sets of rules. Specifically, three new metrics are introduced: syntactic rule similarity and Bag of Words similarity, which, through different mechanisms, both compare rules in terms of their syntax, and geometrical rule similarity, which accounts for how rules are geometrically related to each other in the feature space.

The research then moves to designing *rule-based safety regions*, i.e., regions

of the feature space that ensure a guaranteed performance of the rule-based classifier on a target class. Such target depends on the application scenario, and typically represents a safe situation, intended as the absence of conditions that can cause harms to humans/environments: for example, the absence of collisions in autonomous driving, or the absence of a pathology in healthcare applications. The final objective is that safety regions, representing the subset of inputs that will most probably lead to a correct performance of the rule-based classifier, can serve as monitoring tools over the inputs, possibly triggering alerts when these fall outside their boundaries.

Yet pursuing the same objective, two different solutions to this problem are researched.

Initially, a heuristic approach is explored, where the regions are designed starting from the rules themselves - opportunely synthesized by their feature and value ranking properties - and optimizing their thresholds in a grid-search-like mode, until achieving the minimal statistical error on the desired class. These methods, called *reliability from inside*, *reliability from outside* and *rules with zero error*, look at identifying the safety regions by exploiting the properties of the rules for the target class, the non-target class, and a combination of target class rules specifically trained with zero error constraint, respectively. Despite the heuristic approach, experiments show promising results when tested in some safety-critical applications such as the collision avoidance in vehicle platooning scenarios, the prediction of physical fatigue, and also the detection of adversarial ML attacks.

Afterwards, a more formal solution is sought by relying on a well-established statistical framework, widely used in ML uncertainty quantification: *conformal prediction* (CP), which allows to probabilistically guarantee the error rate under a desired level, by assigning prediction sets to data samples, instead of the typical point predictions provided by classifiers. The key aspect of CP is the design of a score function that encodes the behavior of a classifier, assigning larger values to encode a worse agreement between a point and a candidate label, and vice versa. In this respect, a *novel score function*, called CONFIDERAI, that allows to apply CP theory to rule-based binary classifiers is proposed, accounting for both the geometrical structure (distance of points to rule boundaries, and rule overlaps), and the predictive performance of decision rules (i.e., rule relevance). Leveraging on the results provided through CONFIDERAI, the *conformal critical set* has been defined as the subset of points in the input space where the prediction set is solely composed by points belonging to the target class, and whose error is bounded by the CP. It is shown that this definition provides a new labelling of the dataset, which reveals useful for the generation of new rules that have improved performance on the target class with respects to the original ruleset. Extensive experimentation on both toy datasets/rulesets and real-world applications shows good results, highlighting the relevance of this contribution at the intersection of explainability and reliability.

Finally, these techniques are tested in some applications of interest in Industry 4.0, also deriving from research projects I contributed to.