Number 116 January 2019

ERCIM NEWS

www.ercim.eu

Special theme:

Transparency in Algorithmic Decision Making

Research and Society:
Ethics in Research

KEYNOTE

RESEARCHCH AND SOCIETY

This section about "Ethics in research has been coordinated by Claude Kirchner (Inria) and James Larrus (EPFL)

SPECIAL THEME

The special theme "Transparency in Algorithmic Decision Making" has been coordinated by Andreas Rauber (TU Wien and SBA), Roberto Trasarti and Fosca Giannotti (ISTI-CNR).

Introduction to the special theme

RESEARCH AND INNOVATION

This section features news about research activities and innovative developments from European research institutes

ANNOUCEMENTS, IN BRIEF

The AI Black Box Explanation Problem

by Riccardo Guidotti, Anna Monreale and Dino Pedreschi (KDDLab, ISTI-CNR Pisa and University of Pisa)

Explainable AI is an essential component of a "Human AI", i.e., an AI that expands human experience, instead of replacing it. It will be impossible to gain the trust of people in AI tools that make crucial decisions in an opaque way without explaining the rationale followed, especially in areas where we do not want to completely delegate decisions to machines.

On the contrary, the last decade has witnessed the rise of a black box society [1]. Black box AI systems for automated decision making, often based on machine learning over big data, map a user's features into a class predicting the behavioural traits of individuals, such as credit risk, health status, etc., without exposing the reasons why. This is problematic not only for lack of transparency, but also for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions [2].

Machine learning constructs decision-making systems based on data describing the digital traces of human activities. Consequently, black box models may reflect human biases and prejudices. Many controversial cases have already highlighted the problems with delegating decision making to black box algorithms in many sensitive domains, including crime prediction, personality scoring, image classification, etc. Striking examples include those of COMPAS [L1] and Amazon [L2] where the predictive models discriminate minorities based on an ethnic bias in the training data.

The EU General Data Protection Regulation introduces a right of explanation for individuals to obtain "meaningful information of the logic involved" when automated decision-making takes place with "legal or similarly relevant effects" on individuals [L3]. Without a technology capable of explaining the logic of black boxes, this right will either remain a "dead letter", or outlaw many applications of opaque AI decision making systems.

It is clear that a missing step in the construction of a machine learning model is precisely the explanation of its logic, expressed in a comprehensible, human-readable format, that highlights the biases learned by the model, allowing AI developers and other stakeholders to understand and validate its decision rationale. This limitation impacts not only information ethics, but also accountability, safety and industrial liability [3]. Companies increasingly market services and products with embedded machine learning components, often in safety-critical industries such as self-driving cars, robotic assistants, and personalised medicine. How can companies trust their products without understanding the logic of their model components?

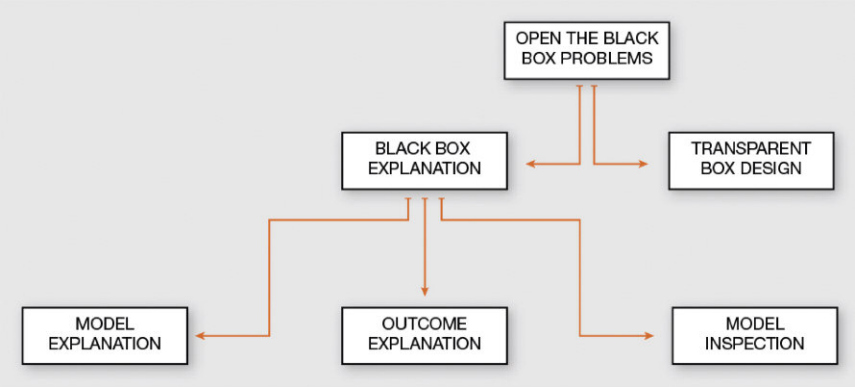

Figure 1: Open the black box problems taxonomy.

At a very high level, we articulated the problem in two different flavours:

- *eXplanation by Design* (XbD): given a dataset of training decision records, how to develop a machine learning decision model together with its explanation;
- *Black Box eXplanation* (BBX): given the decision records produced by a black box decision model, how to reconstruct an explanation for it.

In the XbD setting the aim is to provide a transparent machine learning decision model providing an explanation of the model's logic by design. On the other hand, the BBX problem can be resolved with methods for auditing and finding an explanation for an obscure machine learning model, i.e., a black box for which the internals are unknown. In fact, only the decision behaviour of the black box can be observed. As displayed in Figure 1, the BBX problem can be further decomposed into:

- model explanation when the explanation involves the whole (global) logic of the black box classifier;
- outcome explanation when the target is to (locally) understand the reasons for the decision of a given record;
- model inspection when the object is to understand how internally the black box behaves changing the input by means of a visual tool.

We are focusing on the open challenge of constructing a global meaningful explanation for a black box model in the BBX setting by exploiting local explanations of why a specific case has received a certain classification outcome. Specifically, we are working on a new local-first explanation framework that works under the assumptions of: (i) local explanations: the global decision boundaries of a black box can be arbitrarily complex to understand, but in the neighbourhood of each specific data point there is a high chance that the decision boundary is clear and simple; (ii) explanation composition: there is a high chance that similar records admit similar explanations, and similar explanations are likely to be composed together into more general explanations. These

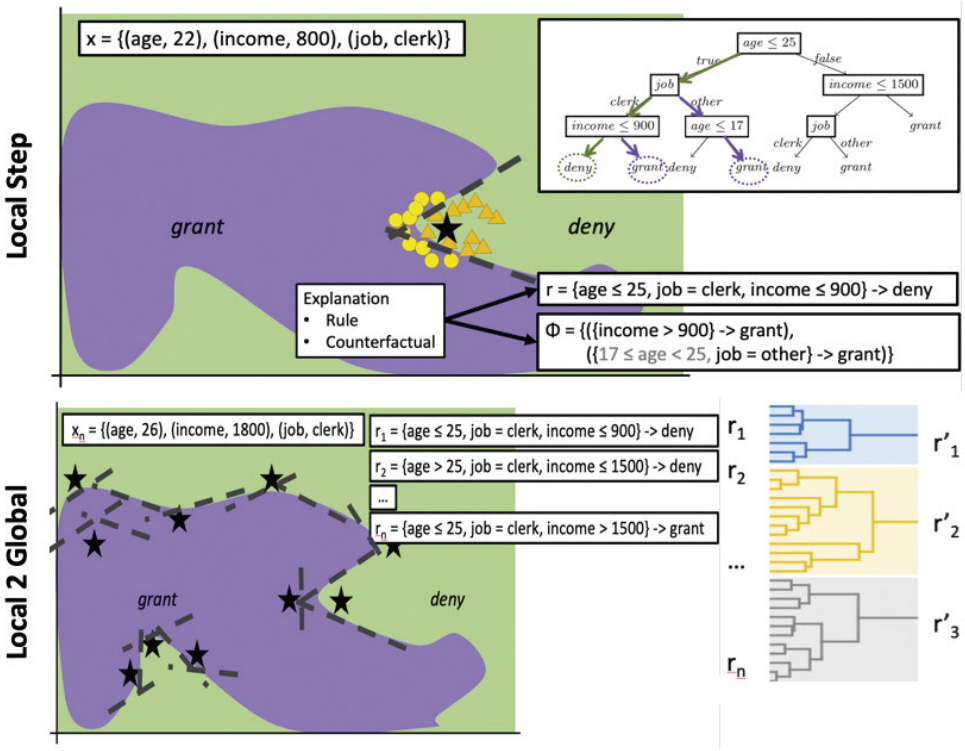

Figure 2: Local-first global explanation framework.

assumptions suggest a two-step, local-first approach to the BBX problem:

- Local Step: for any record x in the set of instances to explain, query the black box to label a set of synthetic examples in the neighbourhood of x which are then used to derive a local explanation rule using an interpretable classifier (Figure 2 top).
- Local-to-Global Step: consider the set of local explanations constructed at the local step and synthesise a smaller set by iteratively composing and generalising together similar explanations, optimising for simplicity and fidelity (Figure 2 bottom).

The most innovative part is the Local-to-Global (L2G) Step. At each iteration, L2G merges the two closest explanations e1, e2 by using a notion of similarity defined as the normalized intersection of the coverages of e1, e2 on a given record set X. An explanation e covers a record x if all the requirements of e are satisfied by x, i.e., boundary constraints, e.g. age > 26. L2G stops merging explanations by considering the relative trade-off gain between model simplicity and fidelity in mimicking the black box. The result is a hierarchy of explanations that can be represented by using a dendrogram (a tree-like diagram, Figure 2 bottom right).

We argue that the L2G approach has the potential to advance the state of art significantly, opening the door to a wide variety of alternative technical solutions along different dimensions: the variety of data sources (relational, text, images, etc.), the variety of learning problems (binary and multi-label classification, regression, scoring, etc.), the variety of languages for expressing meaningful explanations. With the caveat that impactful, widely adopted solutions to the explainable AI problem will be only made possible by truly interdisciplinary research, bridging data science and AI with human sciences, including philosophy and cognitive psychology.

This article is coauthored with Fosca Giannotti, Salvatore Ruggieri, Mattia Setzu, and Franco Turini (KDDLab, ISTI-CNR Pisa and University of Pisa).

Links:

[L1] https://kwz.me/hd9
[L2] https://kwz.me/hdf
[L3] http://ec.europa.eu/justice/data-protection/

References:

[1] F. Pasquale: "The black box society: The secret algorithms that control money and information", Harvard University Press, 2015.
[2] R. Guidotti, et al.: "A survey of methods for explaining black box models", ACM Computing Surveys (CSUR), 51(5), 93, 2018.
[3] J. Kroll, et al.: "Accountable algorithms", U. Pa. L. Rev., 165, 633, 2016.

Please contact:

Riccardo Guidotti, ISTI-CNR, Italy
+39 377 9933326,
riccardo.guidotti@isti.cnr.it

Anna Monreale, University of Pisa, Italy
+39 328 2598903,
anna.monreale@unipi.it

Dino Pedreschi, University of Pisa, Italy
+39 348 6544616,
dino.pedreschi@unipi.it