Xinghan Liu
University of Toulouse

A logic of "black box" classifiers

The black box metaphor is widely used in machine learning and artificial intelligence. It refers to some classifier system whose ``inner working'' is opaque to people. To model this epistemic uncertainty, we propose a product modal logic. Its semantics represents a black box as a set of classifiers which are all compatible with the incomplete knowledge that the agent has about the black box. We also show how to represent explanations of black box classifiers in our modal language.