25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

Explainable algorithm evaluation from lessons in education

Sevvandi Kandanaarachchi [a] and **Kate Smith-Miles** [b]

[a] *CSIRO Data61, Clayton, Melbourne, Australia*
[b] *School of Mathematics and Statistics, University of Melbourne, Australia*
Email: sevvandi.kandanaarachchi@data61.csiro.au

Abstract: How can we evaluate a portfolio of algorithms to extract meaningful interpretations about them? Suppose we have a set of algorithms. These can be classification, regression, clustering or any other type of algorithm. And suppose we have a set of problems that these algorithms can work on. We can evaluate these algorithms on the problems and get the results. From these results, can we explain the algorithms in a meaningful way? The easy option is to find which algorithm performs best for each problem and find the algorithm that performs best on the greatest number of problems. But, there is a limitation with this approach. We are only looking at the overall best! Suppose a certain algorithm gives the best performance on hard problems, but not on easy problems. We would miss this algorithm by using the "overall best" approach. How do we obtain a salient set of algorithm features?

To find an answer to this question we turn to social sciences. Methodologies in social sciences focus on explanations as opposed to accurate predictions (Shmueli 2010). As such, quantitative models in social sciences only have a handful of parameters which have meaningful interpretations. Explanations are often linked with causality. Miller (2019) presents an argument for linkages with social sciences stating that "the field of explainable artificial intelligence can build on existing research, and relevant papers from philosophy, cognitive psychology/science, and social psychology, which study these topics". We propose such a linkage.

Item response theory (IRT) is a methodology in educational psychometrics that is used to design, analyse and score test questions and questionnaires. It is used to measure abilities and attitudes such as political preferences and stress proneness. Participants take a test and IRT is used to uncover the ability of participants and the discrimination and difficulty of test questions. For example, difficult test items generally yield lower scores than easy test items. Similarly, students with high ability obtain higher scores compared to students with low ability. Thus, IRT parameters are given causal interpretations. We propose a novel mapping of the traditional IRT framework modified to the algorithm evaluation domain. Using this new mapping, we elicit a richer suite of characteristics including algorithm consistency, difficulty limit and anomalousness that describe important aspects of algorithm performance. The explainable interpretations discussed above get translated to the algorithm evaluation setting as follows: problems with high difficulty generally result in low performance values; algorithms with high difficulty limits can handle harder problems; algorithms that are consistent give similar results irrespective of the problem difficulty; anomalous algorithms behave in an unusual fashion by giving better results to harder problems compared to easier problems.

We call our framework AIRT – Algorithmic IRT. The word airt is an old Scottish word which means "to guide". In addition to general algorithm metrics, AIRT has visual capabilities. We can visualise the problem space in terms of problem difficulty. For each algorithm, AIRT produces a performance curve, which shows the performance of an algorithm across the problem space. These curves can be used to find similarities and differences between algorithms. From these curves AIRT finds strengths and weaknesses of algorithms. Algorithm strengths/weaknesses can be visualised as part of the problem space, which shows us the type of problems the algorithm is good at. The R package airt makes this framework available.

REFERENCES

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007

Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3), 289–310. https://doi.org/10.1214/10-STS330

Keywords: *Algorithm evaluation, item response theory, algorithm portfolios*