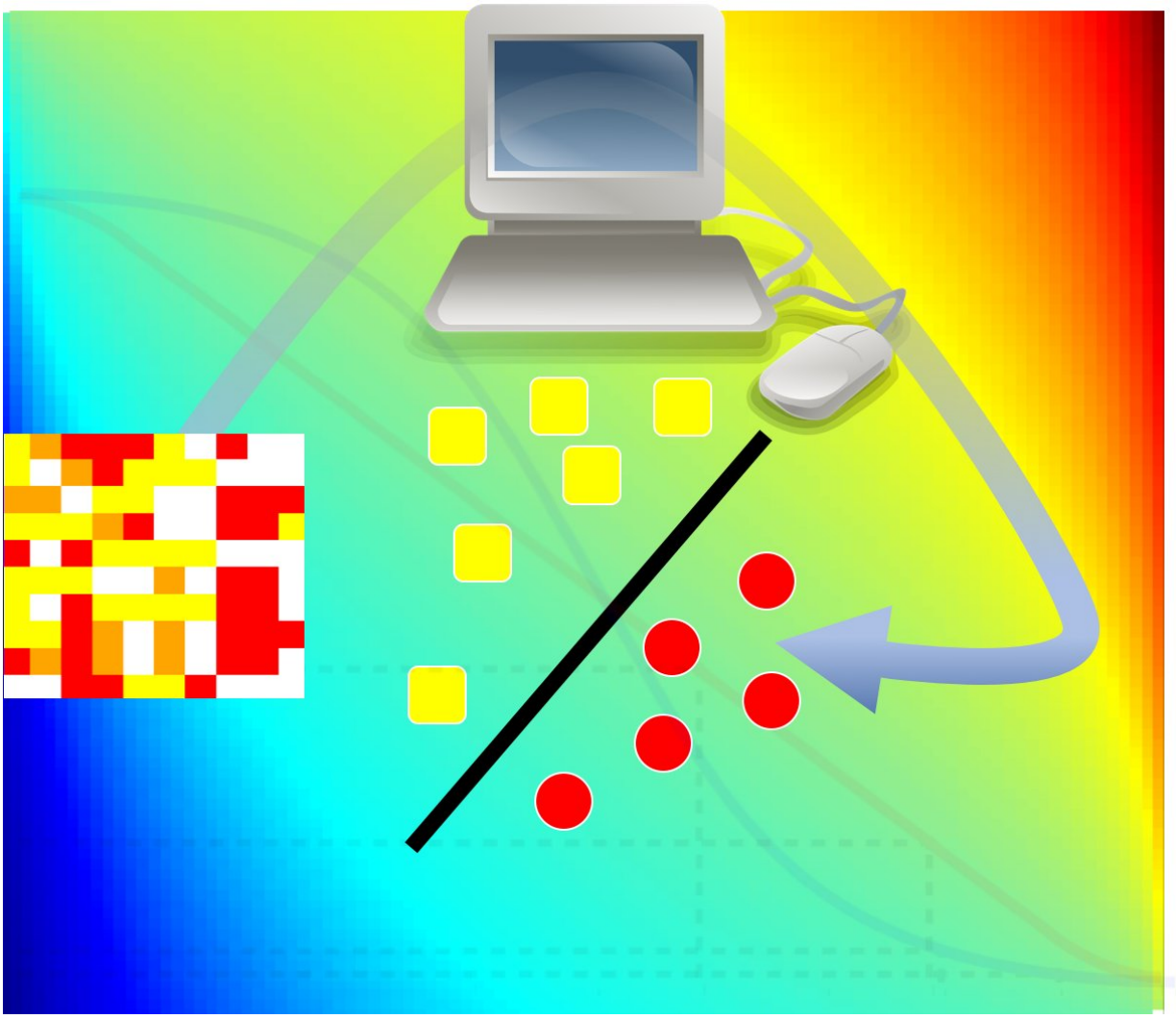
How accurate is your AI?

March 14 2018



The new AI evaluation method looks at the input data itself to find if the 'accuracy' of the AI can be trusted. Credit: Kyoto University / JB Brown

As AI's role in society continues to expand, J B Brown of the Graduate School of Medicine reports on a new evaluation method for the type of AI that predicts yes/positive/true or no/negative/false answers.

Brown's paper, published in *Molecular Informatics*, deconstructs the utilization of AI and analyzes the nature of the statistics used to report an AI program's ability. The new technique also generates a probability of the performance level given evaluation data, answering questions such as: What is the probability of achieving accuracy greater than 90%?

Reports of new AI applications appear in the news almost daily, including in society and science, finance, pharmaceuticals, medicine, and security.

"While reported statistics seem impressive, research teams and those evaluating the results come across two problems," explains Brown. "First, to understand if the AI achieved its results by chance, and second, to interpret applicability from the reported performance statistics."

For example, if an AI program is built to predict whether or not someone will win the lottery, it may always predict a loss. The program may achieve '99% accuracy', but interpretation is key to determine the accuracy of the conclusion that the program is accurate.

But herein lies the problem: in typical AI development, the evaluation can only be trusted if there is an equal number of positive and negative results. If the data is biased toward either value, the current system of evaluation will exaggerate the system's ability.

So to tackle this problem, Brown developed a new technique that evaluates performance based on only the input data itself.

"The novelty of this technique is that it doesn't depend on any one type

of AI technology, such as deep learning," Brown describes. "It can help develop new evaluation metrics by looking at how a metric interplays with the balance in predicted data. We can then tell if the resulting metrics could be biased."

Brown hopes this analysis will not only raise awareness of how we think about AI in the future, but also that it contributes to the development of more robust AI platforms.

In addition to the accuracy metric, Brown tested six other metrics in both theoretical and applied scenarios, finding that no single metric was universally superior. He says the key to building useful AI platforms is to take a multi-metric view of evaluation.

"AI can assist us in understanding many phenomena in the world, but for it to properly provide us direction, we must know how to ask the right questions. We must be careful not to overly focus on a single number as a measure of an AI's reliability."

Brown's program is freely available to the general public, researchers, and developers.

More information: J. B. Brown, Classifiers and their Metrics Quantified, *Molecular Informatics* (2018). DOI: 10.1002/minf.201700127

Provided by Kyoto University