Researchers develop new approach for better big data prediction

December 12 2016

Researchers at Columbia University, Princeton and Harvard University have developed a new approach for analyzing big data that can drastically improve the ability to make accurate predictions about medicine, complex diseases, social science phenomena, and other issues.

In a study published in the December 13 issue of *Proceedings of the National Academy of Sciences* (*PNAS*), the authors introduce the Influence score, or "I-score," as a statistic correlated with how much variables inherently can predict, or "predictivity", which can consequently be used to identify highly predictive variables.

"In our last paper, we showed that significant variables may not necessarily be predictive, and that good predictors may not appear statistically significant," said principal investigator Shaw-Hwa Lo, a professor of statistics at Columbia University. "This left us with an important question: how can we find highly predictive variables then, if not through a guideline of statistical significance? In this article, we provide a theoretical framework from which to design good measures of prediction in general. Importantly, we introduce a variable set's predictivity as a new parameter of interest to estimate, and provide the I-score as a candidate statistic to estimate variable set predictivity."

Current approaches to prediction generally include using a significance-based criterion for evaluating variables to use in models and evaluating variables and models simultaneously for prediction using cross-validation or independent test data.

"Using the I-score prediction framework allows us to define a novel measure of predictivity based on observed data, which in turn enables assessing variable sets for, preferably high, predictivity," Lo said, adding that, while intuitively obvious, not enough attention has been paid to the consideration of predictivity as a parameter of interest to estimate. Motivated by the needs of current genome-wide association studies (GWAS), the study authors provide such a discussion.

In the paper, the authors describe the predictivity for a variable set and show that a simple sample estimation of predictivity directly does not provide usable information for the prediction-oriented researcher. They go on to demonstrate that the I-score can be used to compute a measure that asymptotically approaches predictivity. The I-score can effectively differentiate between noisy and predictive variables, Lo explained, making it helpful in variable selection. A further benefit is that while usual approaches require heavy use of cross-validation data or testing data to evaluate the predictors, the I-score approach does not rely as much on this as much.

"We offer simulations and an application of the I-score on real data to demonstrate the statistic's predictive performance on sample data," he said. "These show that the I-score can capture highly predictive variable sets, estimates a lower bound for the theoretical correct prediction rate, and correlates well with the out of sample correct rate. We suggest that using the I-score method can aid in finding variable sets with promising prediction rates, however, further research in the avenue of sample-based measures of predictivity is needed."

The authors conclude that there are many applications for which using the I-score would be useful, for example in formulating predictions about diseases with high dimensional data, such as gene datasets, in the social sciences for text prediction or financial markets predictions; in terrorism, civil war, elections and financial markets.

"We're hoping to impress upon the scientific community the notion that for those of us who might be interested in predicting an outcome of interest, possibly with rather complex or high dimensional data, we might gain by reconsidering the question as one of how to search for highly predictive variables (or variable sets) and using statistics that measure predictivity to help us identify those variables to then predict well," Lo said. "For statisticians in particular, we're hoping this opens up a new field of work that would focus on designing new statistics that measure predictivity."

More information: Adeline Lo et al. Framework for making better predictions by directly estimating variables' predictivity, *Proceedings of the National Academy of Sciences* (2016). DOI: 10.1073/pnas.1616647113

Provided by Columbia University