System that replaces human intuition with algorithms outperforms human teams

October 16 2015, by Larry Hardesty

Big-data analysis consists of searching for buried patterns that have some kind of predictive power. But choosing which "features" of the data to analyze usually requires some human intuition. In a database containing,

say, the beginning and end dates of various sales promotions and weekly profits, the crucial data may not be the dates themselves but the spans between them, or not the total profits but the averages across those spans.

MIT researchers aim to take the human element out of big-data analysis, with a new system that not only searches for patterns but designs the feature set, too. To test the first prototype of their system, they enrolled it in three data science competitions, in which it competed against human teams to find predictive patterns in unfamiliar data sets. Of the 906 teams participating in the three competitions, the researchers' "Data Science Machine" finished ahead of 615.

In two of the three competitions, the predictions made by the Data Science Machine were 94 percent and 96 percent as accurate as the winning submissions. In the third, the figure was a more modest 87 percent. But where the teams of humans typically labored over their prediction algorithms for months, the Data Science Machine took somewhere between two and 12 hours to produce each of its entries.

"We view the Data Science Machine as a natural complement to human intelligence," says Max Kanter, whose MIT master's thesis in computer science is the basis of the Data Science Machine. "There's so much data out there to be analyzed. And right now it's just sitting there not doing anything. So maybe we can come up with a solution that will at least get us started on it, at least get us moving."

Between the lines

Kanter and his thesis advisor, Kalyan Veeramachaneni, a research scientist at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), describe the Data Science Machine in a paper that Kanter will present next week at the IEEE International Conference on

Data Science and Advanced Analytics.

Veeramachaneni co-leads the Anyscale Learning for All group at CSAIL, which applies machine-learning techniques to practical problems in big-data analysis, such as determining the power-generation capacity of wind-farm sites or predicting which students are at risk for dropping out of online courses.

"What we observed from our experience solving a number of data science problems for industry is that one of the very critical steps is called feature engineering," Veeramachaneni says. "The first thing you have to do is identify what variables to extract from the database or compose, and for that, you have to come up with a lot of ideas."

In predicting dropout, for instance, two crucial indicators proved to be how long before a deadline a student begins working on a problem set and how much time the student spends on the course website relative to his or her classmates. MIT's online-learning platform MITx doesn't record either of those statistics, but it does collect data from which they can be inferred.

Featured composition

Kanter and Veeramachaneni use a couple of tricks to manufacture candidate features for data analyses. One is to exploit structural relationships inherent in database design. Databases typically store different types of data in different tables, indicating the correlations between them using numerical identifiers. The Data Science Machine tracks these correlations, using them as a cue to feature construction.

For instance, one table might list retail items and their costs; another might list items included in individual customers' purchases. The Data Science Machine would begin by importing costs from the first table

into the second. Then, taking its cue from the association of several different items in the second table with the same purchase number, it would execute a suite of operations to generate candidate features: total cost per order, average cost per order, minimum cost per order, and so on. As numerical identifiers proliferate across tables, the Data Science Machine layers operations on top of each other, finding minima of averages, averages of sums, and so on.

It also looks for so-called categorical data, which appear to be restricted to a limited range of values, such as days of the week or brand names. It then generates further feature candidates by dividing up existing features across categories.

Once it's produced an array of candidates, it reduces their number by identifying those whose values seem to be correlated. Then it starts testing its reduced set of features on sample data, recombining them in different ways to optimize the accuracy of the predictions they yield.

"The Data Science Machine is one of those unbelievable projects where applying cutting-edge research to solve practical problems opens an entirely new way of looking at the problem," says Margo Seltzer, a professor of computer science at Harvard University who was not involved in the work. "I think what they've done is going to become the standard quickly—very quickly."

More information: "Deep Feature Synthesis: Towards Automating Data Science Endeavors." groups.csail.mit.edu/EVO-Desig … te/DSAA_DSM_2015.pdf

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.