With new algorithms, data scientists could accomplish in days what once took months

October 21 2016, by Larry Hardesty

“The goal of all this is to present the interesting stuff to the data scientists so that they can more quickly address all these new data sets that are coming in,” says Max Kanter MEng ’15. Credit: Massachusetts Institute of Technology

Last year, MIT researchers presented a system that automated a crucial step in big-data analysis: the selection of a "feature set," or aspects of the data that are useful for making predictions. The researchers entered the system in several data science contests, where it outperformed most of the human competitors and took only hours instead of months to perform its analyses.

This week, in a pair of papers at the IEEE International Conference on Data Science and Advanced Analytics, the team described an approach to automating most of the rest of the process of big-data analysis—the preparation of the data for analysis and even the specification of problems that the analysis might be able to solve.

The researchers believe that, again, their systems could perform in days tasks that used to take data scientists months.

"The goal of all this is to present the interesting stuff to the data scientists so that they can more quickly address all these new data sets that are coming in," says Max Kanter MEng '15, who is first author on last year's paper and one of this year's papers. "[Data scientists want to know], 'Why don't you show me the top 10 things that I can do the best, and then I'll dig down into those?' So [these methods are] shrinking the time between getting a data set and actually producing value out of it."

Both papers focus on time-varying data, which reflects observations made over time, and they assume that the goal of analysis is to produce a probabilistic model that will predict future events on the basis of current observations.

Real-world problems

The first paper describes a general framework for analyzing time-varying data. It splits the analytic process into three stages: labeling the data, or categorizing salient data points so they can be fed to a machine-learning system; segmenting the data, or determining which time sequences of data points are relevant to which problems; and "featurizing" the data, the step performed by the system the researchers presented last year.

The second paper describes a new language for describing data-analysis problems and a set of algorithms that automatically recombine data in different ways, to determine what types of prediction problems the data might be useful for solving.

According to Kalyan Veeramachaneni, a principal research scientist at MIT's Laboratory for Information and Decision Systems and senior author on all three papers, the work grew out of his team's experience with real data-analysis problems brought to it by industry researchers.

"Our experience was, when we got the data, the domain experts and data scientists sat around the table for a couple months to define a prediction problem," he says. "The reason I think that people did that is they knew that the label-segment-featurize process takes six to eight months. So we better define a good prediction problem to even start that process."

In 2015, after completing his master's, Kanter joined Veeramachaneni's group as a researcher. Then, in the fall of 2015, Kanter and Veeramachaneni founded a company called Feature Labs to commercialize their data-analysis technology. Kanter is now the company's CEO, and after receiving his master's in 2016, another master's student in Veeramachaneni's group, Benjamin Schreck, joined the company as chief data scientist.

Data preparation

Developed by Schreck and Veeramachaneni, the new language, dubbed Trane, should reduce the time it takes data scientists to define good prediction problems, from months to days. Kanter, Veeramachaneni, and another Feature Labs employee, Owen Gillespie, have also devised a method that should do the same for the label-segment-featurize (LSF) process.

To get a sense of what labeling and segmentation entails, suppose that a data scientist is presented with electroencephalogram (EEG) data for several patients with epilepsy and asked to identify patterns in the data that might signal the onset of seizures.

The first step is to identify the EEG spikes that indicate seizures. The next is to extract a segment of the EEG signal that precedes each seizure. For purposes of comparison, "normal" segments of the signal—segments of similar length but far removed from seizures—should also be extracted. The segments are then labeled as either preceding a seizure or not, information that a machine-learning algorithm can use to identify patterns that indicate seizure onset.

In their LSF paper, Kanter, Veeramachaneni, and Gillespie define a general mathematical framework for describing such labeling and segmentation problems. Rather than EEG readings, for instance, the data might be the purchases by customers of a particular company, and the problem might be to determine from a customer's buying history whether he or she is likely to buy a new product.

There, the pertinent data, for predictive purposes, may be not a customer's behavior over some time span, but information about his or her three most recent purchases, whenever they occurred. The framework is flexible enough to accommodate such different specifications. But once those specifications are made, the researchers' algorithm performs the corresponding segmentation and labeling automatically.

Finding problems

With Trane, time-series data is represented in tables, where the columns contain measurements and the times at which they were made. Schreck and Veeramachaneni defined a small set of operations that can be performed on either columns or rows. A row operation is something like determining whether a measurement in one row is greater than some threshold number, or raising it to particular power. A column operation is something like taking the differences between successive measurements in a column, or summing all the measurements, or taking just the first or last one.

Fed a table of data, Trane exhaustively iterates through combinations of such operations, enumerating a huge number of potential questions that can be asked of the data—whether, for instance, the differences between measurements in successive rows ever exceeds a particular value, or whether there are any rows for which it is true that the square of the data equals a particular number.

To test Trane's utility, the researchers considered a suite of questions that data scientists had posed about roughly 60 real data sets. They limited the number of sequential operations that Trane could perform on the data to five, and those operations were drawn from a set of only six row operations and 11 column operations. Remarkably, that comparatively limited set was enough to reproduce every question that researchers had in fact posed—in addition to hundreds of others that they hadn't.

"Probably the biggest thing here is that it's a big step toward enabling us to represent prediction problems in a standard way so that you could share that with other analysts in an abstraction from the problem specifics," says Kiri Wagstaff, a senior researcher in artificial intelligence and machine learning at NASA's Jet Propulsion Laboratory.

"What I would hope is that this could lead to improved collaboration between whatever domain experts you're working with and the data analysts. Because now the domain experts, if they could learn and would be willing to use this language, could specify their problems in a much more precise way than they're currently able to do."

More information: Label, Segment, Featurize: a cross domain framework for prediction engineering: dai.lids.mit.edu/Pred_eng.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.

Provided by Massachusetts Institute of Technology

Citation: With new algorithms, data scientists could accomplish in days what once took months (2016, October 21) retrieved 24 April 2024 from https://phys.org/news/2016-10-algorithms-scientists-days-months.html