Researcher tackles some of the biggest bottlenecks holding back the data science industry

February 25 2015, by Eric Brown

Kalyan Veeramachaneni. Credit: David Sella

When Kalyan Veeramachaneni joined the Any Scale Learning For All (ALFA) group at MIT's CSAIL as a postdoc in 2010, he worked on large-scale machine-learning platforms that enable the construction of models from huge data sets. "The question then was how to decompose a learning algorithm and data into pieces, so each piece could be locally

loaded into different machines and several models could be learnt independently," says Veeramachaneni, currently a research scientist at ALFA.

"We then had to decompose the learning algorithm so we could parallelize the compute on each node," says Veeramachaneni. "In this way, the data on each node could be learned by the system, and then we could combine all the solutions and models that had been independently learned."

By 2013, once ALFA had built multiple platforms to accomplish these goals, the team started on a new problem: the growing bottleneck caused by the process of translating the raw data into the formats required by most machine-learning systems.

"Machine-learning systems usually require a covariates table in a column-wise format, as well as a response variable that we try to predict," says Veeramachaneni. "The process to get these from raw data involves curation, syncing and linking of data, and even generating ideas for variables that we can then operationalize and form."

Much of Veeramachaneni's recent research has focused on how to automate this lengthy data prep process. "Data scientists go to all these boot camps in Silicon Valley to learn open source big data software like Hadoop, and they come back, and say 'Great, but we're still stuck with the problem of getting the raw data to a place where we can use all these tools,'" Veeramachaneni says.

Veeramachaneni and his team are also exploring how to efficiently integrate the expertise of domain experts, "so it won't take up too much of their time," he says. "Our biggest challenge is how to use human input efficiently, and how to make the interactions seamless and efficient. What sort of collaborative frameworks and mechanisms can we build to

increase the pool of people who participate?"

GigaBeats and BeatDB

One project in which Veeramachaneni tested his automated data prep concepts was ALFA's GigaBeats project. GigaBeats analyzes arterial blood pressure signals from thousands of patients to predict a future condition. With GigaBeats, numerous steps are involved to prepare the data for analysis, says Veeramachaneni. These include cleaning and conditioning, low pass filters, and extracting features by applying signal-level transformations.

Many of these steps involve human decision-making. Often, domain experts know how to do it, but sometimes it's up to the computer scientist. In either case, there's no easy way to go back and revisit those human interventions when a choice made later in the pipeline does not result in the expected level of predictive accuracy, says Veeramachaneni.

Recently, ALFA has built some novel platforms that automate the process, shrinking the prep time from months to a few days. To automate and accelerate data translation, while also enabling visibility into earlier decision-making, ALFA has developed a "complete solution" called BeatDB.

"With BeatDB, we have tunable parameters that in some cases can be input by domain experts, and the rest are automatically tuned," says Veeramachaneni. "From this, we can learn how decisions made at the low-level, raw representation stage can impact the final predicted accuracy efficacy. This deep-mining solution combines all layers of machine learning into a single pipeline and then optimizes and tunes with other machine-learning algorithms on top of it. It really enables fast discovery."

Now that ALFA has made progress on integrating and recording human input, the group is also looking for better ways to present the processed data. For example, when showing GigaBeats data to medical professionals, "they are often much more comfortable if a better representation is given to them instead of showing them raw data," says Veeramachaneni. "It makes it easier to provide input. A lot of our focus is on improving the presentation so we can more easily pull their input into our algorithms, clean or fix the data, or create variables."

A crowdsourcing solution

While automating ALFA's machine-learning pipelines, Veeramachaneni has also contributed to a number of real-world analytics projects. Recently, he has been analyzing raw click data from massive open online courses (MOOCs) with the hopes of improving courseware. The initial project is to determine stop-out (drop-out) rates based on online click behavior.

"The online learning platforms record data coming from the interaction of hundreds of thousands of learners," says Veeramachaneni. "We are now able to identify variables that can predict stop-out on a single course. The next stage is to reveal the variables of stop-out and show how to improve the course design."

The first challenge in the MOOC project was to organize the data. There are multiple data streams in addition to clickstream data, and they are usually spread over multiple databases and stored in multiple formats. Veeramachaneni has standardized these sources, integrating them into a single database called MOOCdb. "In this way, software written on top of the database can be re-used," says Veeramachaneni.

The next challenge is to decide what variables to look at. ALFA has explored all sorts of theories about MOOC behavior. For example, if a

student is studying early in the morning, he or she is more likely to stay in the course. Another theory is based on dividing the time spent on the course by how many problems a student gets right. But, Veeramachaneni says, "If I'm trying to predict stop-out, there's no algorithm that automatically comes up with the behavioral variables that influence it. The biggest challenge is that the variables are defined by humans, which creates a big bottleneck."

They turned to crowdsourcing "to tap into as many people as we can," says Veeramachaneni. "We have built a crowdsourcing platform where people can submit an idea against problems such as stop-out," says Veeramachaneni. "Another set of people can operationalize that, such as writing a script to extract that variable on a per student basis."

This research could apply to a number of domains where analysts are trying to predict human behavior based on captured data, such as fraud detection, says Veeramachaneni. Banks and other companies are increasingly analyzing their transaction databases to try to determine whether the person doing the transaction is authentic.

"One variable would be how far the transaction happened from the person's home, or how the amount compares to the total that was spent by the person over the last year," says Veeramachaneni. "Coming up with these ideas is based on very relatable data with which we can all identify. So crowdsourcing could be helpful here, too."

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: Researcher tackles some of the biggest bottlenecks holding back the data science