Evolutionary approaches to big-data problems

January 15 2015, by Eric Brown

Una-May O'Reilly. Credit: David Sella

The AnyScale Learning For All (ALFA) Group at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) aims to solve the most challenging big-data problems—questions that go beyond the scope of typical analytics. ALFA applies the latest machine learning and evolutionary computing concepts to target very complex problems that involve high dimensionality.

"People have data coming at them from so many different channels these days," says ALFA director Una-May O'Reilly, a principal research scientist at CSAIL. "We're helping them connect and link the data between those channels."

The ALFA Group has taken on challenges ranging from laying out wind farms to studying and categorizing the beats in blood pressure data in order to predict drops and spikes. The group is also analyzing huge volumes of recorded click data to predict MOOC-learning behavior, and is even helping the IRS protect against costly tax-evasion schemes.

ALFA prefers the challenge of working with raw data that comes directly from the source. It then investigates the data with a variety of techniques, most of which involve scalable machine learning and evolutionary computing algorithms.

"Machine learning is very useful for retrospectively looking back at the data to help you predict the future," says O'Reilly. "Evolutionary computation can be used in the same way, and it's particularly well suited to large-scale problems with very high dimensions."

In the past, machine learning was challenged by the lack of sufficient data to infer predictive models or classification labels, says O'Reilly. "Now we have too much data, so we have scalable machine learning to try to process a vast quantity of data exemplars," she says. "We also need to improve machine learning's capability to cope with the additional variables that come with extremely high dimensional problems."

O'Reilly has a particular interest in ALFA's other major tool: evolutionary computing. "Taking ideas from evolution, like population-based adaptation and genetic inheritance, and bringing them into computational models is really effective," she says. "In engineering, we often use evolutionary algorithms like covariance-matrix adaptation or

discrete-valued algorithms for optimization. Also, one can parallelize evolutionary algorithms almost embarrassingly easily, which allows it to handle a lot of the latest data-knowledge discovery problems."

Within the evolutionary field, O'Reilly is especially interested in genetic programming, or as she defines it, "the evolution of programs." "We distribute the genetic programming algorithms over many nodes and then factor the data across the nodes," she explains. "We take all the independent solutions we can compute in parallel and bring them together. We then eliminate the weaker ones and collectively fuse the stronger ones to create an ensemble. We've shown that ensemble based models are more accurate than a single model based on all the data."

Laying out wind farms

One of ALFA's most successful projects has been in developing algorithms to help design wind farms. The problem is marked by very high dimensionality, especially when hundreds of turbines are involved, says O'Reilly.

"One can see great efficiency gains in optimizing the placement of turbines, but it's a really complex problem," she says. "First, there are the parameters of the turbine itself: its energy gain, its height, and its proximity cone. You must find out how much wind is required for the site and then acquire the finer detailed information about where the wind is coming from and in what quantities. You have to factor in the topographical conditions of the land and the way the wind sweeps through it."

The most difficult variable is the wake effect of one turbine on the turbines behind it, says O'Reilly. "We have to do very complex flow modeling to be able to calculate the loss behind each turbine."

ALFA discovered how to apply parallelized evolutionary algorithms that could scale up for wind farms of a thousand plus turbines. "We were able to scale to lay out turbines on a bigger scale than anyone had ever done before," says O'Reilly.

More recently, ALFA has been building a generative template for site design. "Now, we're using evolutionary concepts to develop a program that can lay out any set of turbines on any site," she says. "We're building a design process rather than the site design itself."

GigaBEATS: Making sense of blood pressure data

Many of the same evolutionary and machine-learning concepts used to lay out a wind farm can also be applied to gleaning insights from clinical data. ALFA is attempting to elicit useful information from the growing volume of physiological data collected from medical sensors. The data include measurement of everything from sleep patterns to ECG and blood pressure.

"It's hard for clinicians to understand such a high volume of data," says O'Reilly. "We're interested in taking signal-level information and combining it with machine learning to make better predictions."

Researchers tend to collect a small amount of data from a small sample, and do a study that takes over 18 months, says O'Reilly. "We want to take that 18 months and reduce it to hours," she says.

ALFA is working on a project called GigaBEATS that extracts knowledge from very large sets of physiological data. Initially, the project has studied blood-pressure data taken from thousands of patients in critical care units.

"We are examining the microscopic characteristics of every beat," says

O'Reilly. "Eventually, we will aggregate those characteristics in terms of historical segments that allow us to predict blood pressure spikes."

The ALFA group has created a database called BeatDB that collects not only the beats of the waveforms, but "a set of properties or features of every beat," says O'Reilly. BeatDB has already stored a billion blood pressure beat features from more than 5,000 patients.

"For every beat we describe a time-series set of morphological features," explains O'Reilly. "Once we establish a solid set of fundamental data about the signals, we can provide technology as services on top of that, allowing new beats to be added and processed."

Because BeatDB enables beats to be aggregated into segments, physicians can better decide how much history is needed to make a prediction. "To predict a blood pressure drop 15 minutes ahead, you might need hours of patient data," says O'Reilly. "Because the BeatDB data is organized, and leverages machine learning algorithms, physicians don't have to compute this over and over again. They can experiment with how much data and lead time is required, and then check the accuracy of their models."

Recently, O'Reilly has begun to use the technology to explore ECG data. "We're hoping to look at data that might be collected in context of the quantified self," says O'Reilly, referring to the emerging practice of wearing fitness bracelets to track one's internal data.

"More and more people are instrumenting themselves by wearing a Fitbit that tells them whether they're tired or how well they sleep," says O'Reilly. "Interpreting all these bodily signals is similar to the GigaBEATS project. A BeatDB-like database and cloud-based facility could be set up around these signals to help interpret them."