# Auto-tuning data science—new research streamlines machine learning

December 20 2017

To solve complex problems, data scientists must shepherd their raw data through a series of steps, each one requiring many human-driven decisions. The last step in the process, deciding on a modeling technique, is particularly crucial. Credit: Massachusetts Institute of Technology

The tremendous recent growth of data science—both as a discipline and

an application—can be attributed, in part, to its robust problem-solving power: It can predict when credit card transactions are fraudulent, help business owners figure out when to send coupons in order to maximize customer response, or facilitate educational interventions by forecasting when a student is on the cusp of dropping out.

To get to these data-driven solutions, though, data scientists must shepherd their raw data through a complex series of steps, each one requiring many human-driven decisions. The last step in the process, deciding on a modeling technique, is particularly crucial. There are hundreds of techniques to choose from—from neural networks to support vector machines—and selecting the best one can mean millions of dollars of additional revenue, or the difference between spotting a flaw in critical medical devices and missing it.

In a paper called "ATM: A distributed, collaborative, scalable system for automated machine learning," which was presented last week at the IEEE International Conference on Big Data, researchers from MIT and Michigan State University present a new system that automates the model selection step, even improving on human performance. The system, called Auto-Tuned Models (ATM), takes advantage of cloud-based computing to perform a high-throughput search over modeling options, and find the best possible modeling technique for a particular problem. It also tunes the model's hyperparameters—a way of optimizing the algorithm—which can have a substantial effect on performance. ATM is now available for enterprise as an open-source platform.

To compare ATM with human performers, the researchers tested the system against users of a collaborative crowdsourcing platform, openml.org. On this platform, data scientists work together to solve problems, finding the best solution by building on each other's work. ATM analyzed 47 datasets from the platform and was able to deliver a

solution better than the one humans had come up with 30 percent of the time. When it couldn't outperform humans, it came very close, and crucially, it worked much more quickly than humans could. While open-ml users take an average of 100 days to deliver a near-optimal solution, ATM can arrive at an answer in less than a day.

## Empowering data scientists

This level of speed and accuracy offers much-needed peace of mind for data scientists, who are often plagued by "what-ifs." "There are so many options," says Arun Ross, professor in the computer science and engineering department at Michigan State University and a senior author on the paper. "If a data scientist chose support vector machines as a modeling technique, the question of whether a neural network or a different model would have resulted in better accuracy always lingers in her mind."

Over the past few years, the problem of model selection/tuning has become the focus of a whole new subfield of machine learning, known as Auto-ML. Auto-ML solutions aim to provide data scientists with the best possible model for a given machine-learning task. There's just one problem: Competing Auto-ML approaches yield different results, and their methods are often opaque. In other words, while seeking to solve one selection problem, the community created another that is even more complex. "The 'what-if' question still remains," says Kalyan Veeramachaneni, a principal research scientist at MIT's Laboratory for Information and Decision Systems (LIDS) and a senior author on the paper. "It simply shifts to, 'what if we used a different Auto-ML approach?'"

The ATM system works differently, using on-demand cloud computing to generate and compare hundreds (or even thousands) of models overnight. To search through techniques, researchers use an intelligent

selection mechanism. The system tests thousands of models in parallel, evaluates each, and allocates more computational resources to those techniques that show promise. Poor solutions fall by the wayside, while the best options rise to the top.

Rather than blindly choosing the "best" one and providing it to the user, ATM displays results as a distribution, allowing for comparison of different methods side-by-side. In this way, Ross says, ATM speeds up the process of testing and comparing different modeling approaches without automating out human intuition, which remains a vital part of the data science process.

## Open-source, community-driven approach

By streamlining the process of model choice, Veeramachaneni and his team aim to allow data scientists to work on more impactful parts of the pipeline. "We hope that our system will free up experts to spend more time on understanding the data, problem formulation, and feature engineering," Veeramachaneni says.

To that end, the researchers are open-sourcing ATM, making it available to enterprises who might want to use it. They have also included provisions that allow researchers to integrate new model selection techniques and thus continually improve on the platform. ATM can run on a single machine, local computing clusters, or on-demand clusters in the cloud, and can work with multiple data sets and multiple users simultaneously.

"A small- to medium-sized data science team can set up and start producing models with just a few steps," Veeramachaneni says. And none of those are followed by a "what-if."

**More information:** ATM: A distributed, collaborative, scalable system

for automated machine learning. cyphe.rs/static/atm.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