


Crowdsourcing big-data analysis

 news.mit.edu/2017/crowdsourcing-big-data-analysis-1030

Larry Hardesty | MIT News
Office

In the analysis of big data sets, the first step is usually the identification of “features” — data points with particular predictive power or analytic utility. Choosing features usually requires some human intuition. For instance, a sales database might contain revenues and date ranges, but it might take a human to recognize that average revenues — revenues divided by the sizes of the ranges — is the really useful metric.

MIT researchers have developed a new collaboration tool, dubbed FeatureHub, intended to make feature identification more efficient and effective. With FeatureHub, data scientists and experts on particular topics could log on to a central site and spend an hour or two reviewing a problem and proposing features. Software then tests myriad combinations of features against target data, to determine which are most useful for a given predictive task.

In tests, the researchers recruited 32 analysts with data science experience, who spent five hours each with the system, familiarizing themselves with it and using it to propose candidate features for each of two data-science problems.

The predictive models produced by the system were tested against those submitted to a data-science competition called Kaggle. The Kaggle entries had been scored on a 100-point scale, and the FeatureHub models were within three and five points of the winning entries for the two problems.

But where the top-scoring entries were the result of weeks or even months of work, the FeatureHub entries were produced in a matter of days. And while 32 collaborators on a single data science project is a lot by today’s standards, Micah Smith, an MIT graduate student in electrical engineering and computer science who helped lead the project, has much larger ambitions.

FeatureHub — like its name — was inspired by GitHub, an online repository of open-source programming projects, some of which have drawn thousands of contributors. Smith hopes that FeatureHub might someday attain a similar scale.

“I do hope that we can facilitate having thousands of people working on a single solution for predicting where traffic accidents are most likely to strike in New York City or predicting which patients in a hospital are most likely to require some medical intervention,” he says. “I think that the concept of massive and open data science can be really leveraged for areas where there’s a strong social impact but not necessarily a single profit-making or government organization that is coordinating responses.”

Smith and his colleagues presented a paper describing FeatureHub at the IEEE International Conference on Data Science and Advanced Analytics. His coauthors on the paper are his thesis advisor, Kalyan Veeramachaneni, a principal research scientist at

MIT's Laboratory for Information and Decision Systems, and Roy Wedge, who began working with Veeramachaneni's group as an MIT undergraduate and is now a software engineer at Feature Labs, a data science company based on the group's work.

FeatureHub's user interface is built on top of a common data-analysis software suite called the Jupyter Notebook, and the evaluation of feature sets is performed by standard machine-learning software packages. Features must be written in the Python programming language, but their design has to follow a template that intentionally keeps the syntax simple. A typical feature might require between five and 10 lines of code.

The MIT researchers wrote code that mediates between the other software packages and manages data, pooling features submitted by many different users and tracking those collections of features that perform best on particular data analysis tasks.

In the past, Veeramachaneni's group has developed software that automatically generates features by inferring relationships between data from the manner in which they're organized. When that organizational information is missing, however, the approach is less effective.

Still, Smith imagines, automatic feature synthesis could be used in conjunction with FeatureHub, getting projects started before volunteers have begun to contribute to them, saving the grunt work of enumerating the obvious features, and augmenting the best-performing sets of features contributed by humans.
