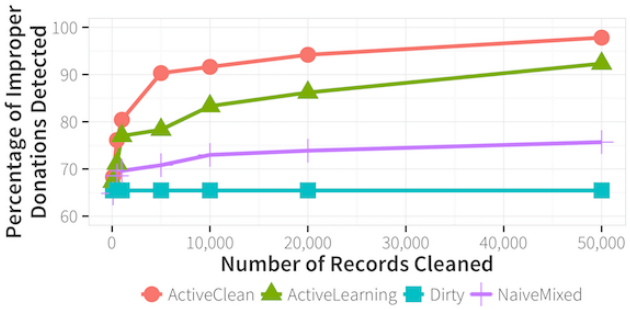
A data-cleaning tool for building better prediction models

31 August 2016, by Kim Martineau



Tested on a dirty, real-world data set, ActiveClean (in red), was able to clean just 5,000 records to bring the researchers' prediction model to a 90 percent accuracy level. The next best technique, called active learning (in green), had to clean 50,000 records to achieve comparable results. The most common data-cleaning method -- trial-and-error (in purple) -- provided minimal model improvement. Credit: Eugene Wu

Big data sets are full of dirty data, and these outliers, typos and missing values can produce distorted models that lead to wrong conclusions and bad decisions, be it in healthcare or finance. With so much at stake, data cleaning should be easier.

That's the inspiration for software developed by computer scientists at Columbia University and University of California at Berkeley that hands much of the dirty work over to machines. Called ActiveClean, the system analyzes a user's prediction model to decide which mistakes to edit first, while updating the model as it works. With each pass, users see their model improve.

"Dirty data is pervasive and prevents people from doing useful things," said Eugene Wu, a computer science professor at Columbia Engineering and a member of the Data Science Institute. "This is our first step towards automating the data-cleaning process."

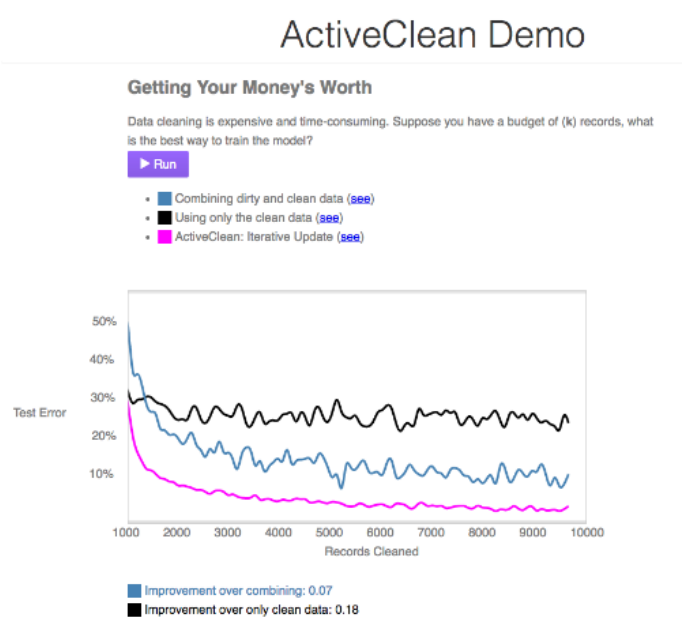
The team will present its research on Sept. 7 in New Delhi, at the 2016 conference on Very Large Data Bases. Wu helped develop ActiveClean as a postdoctoral researcher at Berkeley's AMPLab and has continued this work at Columbia.

Big data sets are still mostly combined and edited manually, aided by data-cleaning software like Google Refine and Trifacta, or custom scripts developed for specific data-cleaning tasks. The process consumes up to 80 percent of analysts' time as they hunt for dirty data, clean it, retrain their model, and repeat the process. Cleaning is largely done by guesswork.

"Will it help or hurt the model? You have no idea," said Wu. "Data scientists either clean everything, which is impossible for huge datasets, or clean random subsets and hope for the best."

In the process, statistical biases can be introduced that skew models into producing misleading results. Those mistakes may not be caught until weeks later, as the researchers learned in an earlier survey of industry data scientists.

"Most of these errors are subtle enough that the analysis will go through," said one consultant from a large database vendor. "Usually it's only caught weeks later after someone notices something like, "Well, the Wilmington branch cannot have $1 million sales in a week."



In the above online demo, researchers show how ActiveClean rapidly improves model accuracy by focusing on the data errors and inconsistencies most likely to skew the user's model. Credit: Eugene Wu

ActiveClean tries to minimize mistakes like these by taking humans out of the most error-prone steps of data cleaning: finding dirty data and updating the model. Using machine learning, the tool analyzes a model's structure to understand what sorts of errors will throw the model off most. It goes after those data first, in decreasing priority, and cleans just enough data to give users assurance that their model will be reasonably accurate.

The researchers tested ActiveClean on Dollars for Docs, a database of corporate donations to doctors that journalists at ProPublica compiled to analyze conflicts of interest and flag improper donations.

ActiveClean's results were compared against two baseline methods. One edited a subset of the data and retrained the model. The other used a popular prioritization algorithm called active learning that picks the most informative labels for ambiguous data. The algorithm improves the model without bothering, as ActiveClean does, whether the labels are accurate.

Nearly a quarter of ProPublica's 240,000 records had multiple names for a drug or company. Left uncorrected these inconsistencies could lead journalists to undercount donations by large companies, which were more likely to have such inconsistencies.

With no data cleaning, a model trained on this dataset could predict an improper donation just 66 percent of the time. ActiveClean, they found, raised the detection rate to 90 percent by cleaning just 5,000 records. The active learning method, by contrast, required 10 times as much data, or 50,000 records, to reach a comparable detection rate.

"As datasets grow larger and more complex, it's becoming more and more difficult to properly clean the data," said study coauthor Sanjay Krishnan, a graduate student at UC Berkeley. "ActiveClean uses machine learning techniques to make data cleaning easier while guaranteeing you won't shoot yourself in the foot."

ActiveClean is a free, open-source tool released in August. Download it here.

Watch how it works: ActiveCleanDemo

Provided by Columbia University School of Engineering and Applied Science