Artificial data give the same results as real data—without compromising privacy

March 6 2017, by Stefanie Koperniak

Credit: Massachusetts Institute of Technology

Although data scientists can gain great insights from large data sets—and can ultimately use these insights to tackle major challenges—accomplishing this is much easier said than done. Many

such efforts are stymied from the outset, as privacy concerns make it difficult for scientists to access the data they would like to work with.

In a paper presented at the IEEE International Conference on Data Science and Advanced Analytics, members of the Data to AI Lab at the MIT Laboratory for Information and Decision Systems (LIDS) Kalyan Veeramachaneni, principal research scientist in LIDS and the Institute for Data, Systems, and Society (IDSS) and co-authors Neha Patki and Roy Wedge describe a machine learning system that automatically creates synthetic data—with the goal of enabling data science efforts that, due to a lack of access to real data, may have otherwise not left the ground. While the use of authentic data can cause significant privacy concerns, this synthetic data is completely different from that produced by real users—but can still be used to develop and test data science algorithms and models.

"Once we model an entire database, we can sample and recreate a synthetic version of the data that very much looks like the original database, statistically speaking," says Veeramachaneni. "If the original database has some missing values and some noise in it, we also embed that noise in the synthetic version… In a way, we are using machine learning to enable machine learning."

The paper describes the Synthetic Data Vault (SDV), a system that builds machine learning models out of real databases in order to create artificial, or synthetic, data. The algorithm, called "recursive conditional parameter aggregation," exploits the hierarchical organization of data common to all databases. For example, it can take a customer-transactions table and form a multivariate model for each customer based on his or her transactions.

This model captures correlations between multiple fields within those transactions—for example, the purchase amount and type, the time at

which the transaction took place, and so on. After the algorithm has modeled and assembled parameters for each customer, it can then form a multivariate model of the these parameters themselves, and recursively model the entire database. Once a model is learned, it can synthesize an entire database, filled with artificial data.

Outcome and impact

After building the SDV, the team used it to generate synthetic data for five different publicly available datasets. They then hired 39 freelance data scientists, working in four groups, to develop predictive models as part of a crowd-sourced experiment. The question they wanted to answer was: "Is there any difference between the work of data scientists given synthesized data, and those with access to real data?" To test this, one group was given the original data sets, while the other three were given the synthetic versions. Each group used their data to solve a predictive modeling problem, eventually conducting 15 tests across 5 datasets. In the end, when their solutions were compared, those generated by the group using real data and those generated by the groups using synthetic data displayed no significant performance difference in 11 out of the 15 tests (70 percent of the time).

These results suggest that synthetic data can successfully replace real data in software writing and testing—meaning that data scientists can use it to overcome a massive barrier to entry. "Using synthetic data gets rid of the 'privacy bottleneck'—so work can get started," says Veeramachaneni.

This has implications for data science across a spectrum of industries. Besides enabling work to begin, synthetic data will allow data scientists to continue ongoing work without involving real, potentially sensitive data.

"Companies can now take their data warehouses or databases and create synthetic versions of them," says Veeramachaneni. "So they can circumvent the problems currently faced by companies like Uber, and enable their data scientists to continue to design and test approaches without breaching the privacy of the real people—including their friends and family—who are using their services."

In addition, the machine-learning model from Veeramachaneni and his team can be easily scaled to create very small or very large synthetic data sets, facilitating rapid development cycles or stress tests for big data systems. Artificial data is also a valuable tool for educating students—although real data is often too sensitive for them to work with, synthetic data can be effectively used in its place. This innovation can allow the next generation of data scientists to enjoy all the benefits of big data, without any of the liabilities.

More information: "The Synthetic data vault", dai.lids.mit.edu/SDV.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

Citation: Artificial data give the same results as real data—without compromising privacy (2017, March 6) retrieved 28 April 2024 from https://phys.org/news/2017-03-artificial-results-real-datawithout-compromising.html