Big data technique shrinks data sets while preserving their fundamental mathematical relationships

December 15 2016, by Larry Hardesty

A new technique devised by MIT researchers can take data sets with huge numbers of variables and find approximations of them with far fewer variables. Credit: Massachusetts Institute of Technology

One way to handle big data is to shrink it. If you can identify a small subset of your data set that preserves its salient mathematical relationships, you may be able to perform useful analyses on it that would be prohibitively time consuming on the full set.

The methods for creating such "coresets" vary according to application, however. Last week, at the Annual Conference on Neural Information Processing Systems, researchers from MIT's Computer Science and Artificial Intelligence Laboratory and the University of Haifa in Israel presented a new coreset-generation technique that's tailored to a whole family of data analysis tools with applications in natural-language processing, computer vision, signal processing, recommendation systems, weather prediction, finance, and neuroscience, among many others.

"These are all very general algorithms that are used in so many applications," says Daniela Rus, the Andrew and Erna Viterbi Professor of Electrical Engineering and Computer Science at MIT and senior author on the new paper. "They're fundamental to so many problems. By figuring out the coreset for a huge matrix for one of these tools, you can enable computations that at the moment are simply not possible."

As an example, in their paper the researchers apply their technique to a matrix—that is, a table—that maps every article on the English version of Wikipedia against every word that appears on the site. That's 1.4 million articles, or matrix rows, and 4.4 million words, or matrix columns.

That matrix would be much too large to analyze using low-rank approximation, an algorithm that can deduce the topics of free-form texts. But with their coreset, the researchers were able to use low-rank approximation to extract clusters of words that denote the 100 most common topics on Wikipedia. The cluster that contains "dress," "brides,"

"bridesmaids," and "wedding," for instance, appears to denote the topic of weddings; the cluster that contains "gun," "fired," "jammed," "pistol," and "shootings" appears to designate the topic of shootings.

Joining Rus on the paper are Mikhail Volkov, an MIT postdoc in electrical engineering and computer science, and Dan Feldman, a lecturer at the University of Haifa and a former postdoc in Rus's group.

The researchers' new coreset technique is useful for a range of tools with names like singular-value decomposition, principal-component analysis, and nonnegative matrix factorization. But what they all have in common is dimension reduction: They take data sets with large numbers of variables and find approximations of them with far fewer variables.

In this, these tools are similar to coresets. But coresets simply reduce the size of a data set, while the dimension-reduction tools change its description in a way that's guaranteed to preserve as much information as possible. That guarantee, however, makes the tools much more computationally intensive than coreset generation—too computationally intensive for practical application to large data sets.

The researchers believe that their technique could be used to winnow a data set with, say, millions of variables—such as descriptions of Wikipedia pages in terms of the words they use—to merely thousands. At that point, a widely used technique like principal-component analysis could reduce the number of variables to mere hundreds, or even lower.

The researchers' technique works with what is called sparse data. Consider, for instance, the Wikipedia matrix, with its 4.4 million columns, each representing a different word. Any given article on Wikipedia will use only a few thousand distinct words. So in any given row—representing one article—only a few thousand matrix slots out of 4.4 million will have any values in them. In a sparse matrix, most of the

values are zero.

Crucially, the new technique preserves that sparsity, which makes its coresets much easier to deal with computationally. Calculations become lot easier if they involve a lot of multiplication by and addition of zero.

The new coreset technique uses what's called a merge-and-reduce procedure. It starts by taking, say, 20 data points in the data set and selecting 10 of them as most representative of the full 20. Then it performs the same procedure with another 20 data points, giving it two reduced sets of 10, which it merges to form a new set of 20. Then it does another reduction, from 20 down to 10.

Even though the procedure examines every data point in a huge data set, because it deals with only small collections of points at a time, it remains computationally efficient. And in their paper, the researchers prove that, for applications involving an array of common dimension-reduction tools, their reduction method provides a very good approximation of the full data set.

That method depends on a geometric interpretation of the data, involving something called a hypercircle, which is the multidimensional analogue of a circle. Any piece of multivariable data can be thought of as a point in a multidimensional space. In the same way that the pair of numbers (1, 1) defines a point in a two-dimensional space—the point one step over on the X-axis and one step up on the Y-axis—a column of the Wikipedia table, with its 4.4 million numbers, defines a point in a 4.4-million-dimensional space.

The researchers' reduction algorithm begins by finding the average value of the subset of data points—let's say 20 of them—that it's going to reduce. This, too, defines a point in a high-dimensional space; call it the origin. Each of the 20 data points is then "projected" onto a hypercircle

centered at the origin. That is, the algorithm finds the unique point on the hypercircle that's in the direction of the data point.

The algorithm selects one of the 20 data projections on the hypercircle. It then selects the projection on the hypercircle farthest away from the first. It finds the point midway between the two and then selects the data projection farthest away from the midpoint; then it finds the point midway between those two points and selects the data projection farthest away from it; and so on.

The researchers were able to prove that the midpoints selected through this method will converge very quickly on the center of the hypercircle. The method will quickly select a subset of points whose average value closely approximates that of the 20 initial points. That makes them particularly good candidates for inclusion in the coreset.

More information: Dimensionality Reduction of Massive Sparse Datasets Using Coresets: arxiv.org/pdf/1503.01663v1.pdf

Provided by Massachusetts Institute of Technology