PHYS.ORG

'Data smashing' could unshackle automated discovery

October 3 2014

(Phys.org) —A little-known secret in data mining is that simply feeding raw data into a data analysis algorithm is unlikely to produce meaningful results, say the authors of a new Cornell study.

From recognizing speech to identifying unusual stars, new discoveries often begin with comparison of data streams to find connections and spot outliers. But most data comparison algorithms today have one major weakness – somewhere, they rely on a human expert to specify what aspects of the data are relevant for comparison, and what aspects aren't. But experts aren't keeping pace with the growing amounts and complexities of big data.

Cornell computing researchers have come up with a new principle they call "data smashing" for estimating the similarities between streams of arbitrary data without human intervention, and without access to the data sources. Hod Lipson, associate professor of mechanical engineering and of computing and information science, and Ishanu Chattopadhyay, a former postdoctoral associate with Lipson now at the University of Chicago, have described their method in Royal Society Interface, Oct. 1.

Data smashing is based on a new way to compare data streams. The process involves two steps. First, the data streams are algorithmically "smashed" to "annihilate" the information in each other. Then, the process measures what information remains after the collision. The more information remains, the less likely the streams originated in the same source.

Data smashing principles may open the door to understanding increasingly complex observations, especially when experts do not know what to look for, according to the researchers.

The authors demonstrated the application of their principle to data from real-world problems, including the disambiguation of electroencephalograph patterns from epileptic seizure patients, detection of anomalous cardiac activity from heart recordings, and classification of astronomical objects from raw photometry.

In all cases and without access to original domain knowledge, the researchers demonstrated performance on par with the accuracy of specialized algorithms and heuristics devised by experts.

More information: "Data smashing: uncovering lurking order in data." Ishanu Chattopadhyay and Hod Lipson. *J. R. Soc. Interface* December 6, 2014 11 101 20140826; DOI: 10.1098/rsif.2014.0826 1742-5662

Provided by Cornell University

Citation: 'Data smashing' could unshackle automated discovery (2014, October 3) retrieved 27 April 2024 from https://phys.org/news/2014-10-unshackle-automated-discovery.html