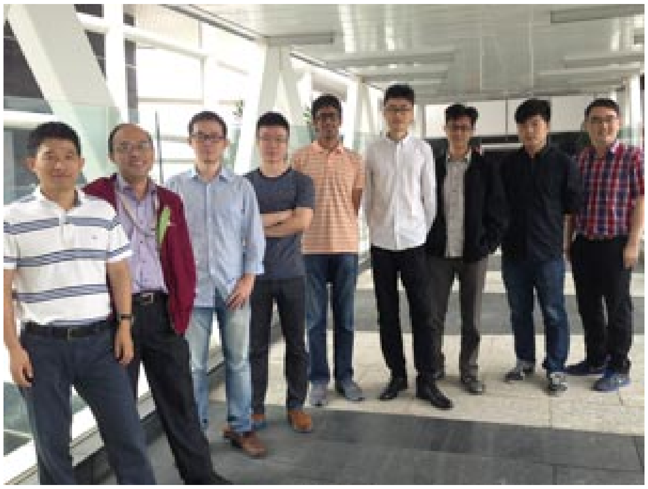
Helping computers learn to tackle big-data problems outside their comfort zones

April 20 2016



Researchers from the A*STAR Institute for Infocomm Research; Peng Xi is third from the left. Credit: A*STAR Institute for Infocomm Research

Imagine combing through thousands of mugshots desperately looking for a match. If time is of the essence, the faster you can do this, the better. A*STAR researchers have developed a framework that could help computers learn how to process and identify these images both faster and more accurately.

Peng Xi of the A*STAR Institute for Infocomm Research notes that the framework can be used for numerous applications, including image segmentation, motion segmentation, data clustering, hybrid system identification and image representation.

A conventional way that computers process data is called representation learning. This involves identifying a feature that allows the program to quickly extract relevant information from the dataset and categorize it—a bit like a shortcut. Supervised and unsupervised learning are two of the main methods used in representation learning. Unlike supervised learning, which relies on costly labeling of data prior to processing, unsupervised learning involves grouping or 'clustering' data in a similar manner to our brains, explains Peng.

Subspace clustering is a form of unsupervised learning that seeks to fit each data point into a low-dimensional subspace to find an intrinsic simplicity that makes complex, real-world data tractable. Existing subspace clustering methods struggle to handle 'out-of-sample', or unknown, data points and the large datasets that are common today.

"One of the challenges of the big-data era is to organize out-of-sample data using a machine learning model based on 'in-sample', or known, observational data," explains Peng who, with his colleagues, has proposed three methods as part of a unified framework to tackle this issue. These methods differ in how they implement representation learning; one focuses on sparsity, while the other two focus on low rank and grouping effects. "By solving the large-scale data and out-of-sample clustering problems, our method makes big-data clustering and online learning possible," notes Peng.

The framework devised by the team splits input data into 'in-sample' data or 'out-of-sample' data during an initial 'sampling' step. Next, the in-sample data is grouped into subspaces during the 'clustering' step, after which the out-of-sample data is assigned to the nearest subspace. These points are then designated as cluster members.

The team tested their approach on a range of datasets including different types of information, from facial images to text—both handwritten and digital—poker hands and forest coverage. They found that their methods outperformed existing algorithms and successfully reduced the computational complexity (and hence running time) of the task while still ensuring cluster quality.

More information: Xi Peng et al. A Unified Framework for Representation-Based Subspace Clustering of Out-of-Sample and Large-Scale Data, *IEEE Transactions on Neural Networks and Learning Systems* (2015). DOI: 10.1109/TNNLS.2015.2490080

Provided by Agency for Science, Technology and Research (A*STAR), Singapore

Citation: Helping computers learn to tackle big-data problems outside their comfort zones (2016, April 20) retrieved 16 August 2024 from https://phys.org/news/2016-04-tackle-big-data-problems-comfort-zones.html