A simple solution for big data

A SISSA algorithm published in Science simplifies the categorization of data

June 27, 2014

Categorizing and representing huge amounts of data (we're talking about peta- or even exabytes of information) synthetically is a challenge of the future. A research paper from the International School for Advanced Studies (SISSA) in Trieste, published in *Science*, proposes an efficient procedure to face up to this challenge.

Experts use the expression *big data* to indicate huge amounts of information, such as those (photos, videos, texts, but also other more technical types of data) shared at any time by billions of people on computers, smartphones and other electronic devices. The present-day scenario offers unprecedented perspectives: tracking flu epidemics, monitoring road traffic in real time, or handling the emergency of natural disasters, for example. For us to be able to use these huge

amounts of data, we have to understand them and before that we need to categorize them in an effective, fast and automatic manner. One of the most commonly used systems is a series of statistical techniques called Cluster Analysis *(CA)*, which is able to group data sets according to their "similarity". Two researchers from SISSA devised a type of CA based on simple and powerful principles, which proved to be very efficient and capable of solving some of the most typical problems encountered in this type of analysis.

Data sets can be imagined as "clouds" of data points in a multidimensional space. These points are generally differently distributed: more widely scattered in one area and denser in another. CA is used to identify the denser areas efficiently, grouping the data in a certain number of significant subsets on the basis of this criterion. Each subset corresponds to a category.

"Think of a database of facial photographs ", explains Alessandro Laio, professor of Statistical and Biological Physics at SISSA. "The database may contain more than one photo of the same person, so CA us used to group all the pictures of the same individual. This type of analysis is carried out by automatic facial recognition systems, for example".
"We tried to devise a more efficient algorithm than those currently used, and one capable of solving some of the classic problems of CA", continues Laio.

More in detail…

"Our approach is based on a new way of identifying the centre of the cluster, i.e., the subsets", explains Alex Rodrigez, co-author of the paper. "Imagine having to identify all the cities in the world, without having access to a map. A huge task", says Rodriguez. "We therefore identified a heuristic, that is, a simple rule or a sort of shortcut to achieve the result".

To find out if a place is a city we can ask each inhabitant to count his "neighbours", in other words, how many people live within 100 metres from his house. Once we have this number, we then go on to find, for each inhabitant, the shortest distance at which another inhabitant with a greater number of neighbours lives. "Together, these two data", explains Laio, "tell us how densely populated is the area where an individual lives and the distance between individuals who have the most neighbours. By automatically cross-checking these data, for the entire world population, we can identify the individuals who represent the centres of the clusters, which correspond to the various cities". "Our algorithm performs precisely this kind of calculation, and it can be applied to many different settings", adds Rodriguez.

The performance of the procedure proved to be optimal: "we tested our mathematical model on the Olivetti Face Database, an archive of facial photographs, obtaining highly satisfactory results. The system recognised most individuals correctly, and never produced 'false positive' results", comments Rodriguez. "This means that in some cases it failed to recognise a subject, but it never once confused one individual with another. Compared to other similar methods, ours was particularly effective in eliminating outliers, that is, those data points that are so very different from the others that they tend to skew the analysis".

USEFUL LINKS:

- **Original paper's abstract on *Science* (http://goo.gl/zt03g7)**

IMAGES:

- **Credits: SISSA**

Contact:

Communication Office:
pressroom@sissa.it
Tel: (+39) 040 3787557 | (+39) 340-5473118, (+39) 333-5275592
via Bonomea, 265
34136 Trieste

More information about SISSA: www.sissa.it