Randomized SUP: A clustering algorithm for large-scale data

Ting-Li Chen

Academia Sinica

Abstract

The self-updating process (SUP) is a clustering algorithm which iteratively updates every data point according to its neighboring points. SUP has been shown to be particularly competitive in clustering (i) data with noise and (ii) data with a large number of clusters. However, the algorithm relies on the pairwise similarities between data points, which becomes computationally inefficient for large-scale data. We will present a randomized approach to overcome the computational difficulty. At each iteration, relatively small portions of data are considered for location updates. The Law of Large Numbers guarantees that the result of the randomized updating process converges to that of the original SUP when the number of data points becomes large. Simulations as well as real data will be presented to show the clustering performance and the computational efficiency of the proposed randomized algorithm.

Keywords：SUP, clustering, randomized algorithm, large scale data