Neural Computing and Applications (2019): https://doi.org/10.1007/s00521-019-04673-0

Rethinking k-means clustering in the age of massive datasets: A constant-time approach

Shareable link: https://rdcu.be/b3w47

Olukanmi, P
Nelwamondo, Fulufhelo V
Marwala, T

ABSTRACT:

We introduce a highly efficient *k*-means clustering approach. We show that the classical central limit theorem addresses a special case (k = 1) of the *k*-means problem and then extend it to the general case. Instead of using the full dataset, our algorithm named *k*-means-lite applies the standard *k*-means to the combination C (size nk) of all sample centroids obtained from n independent small samples. Unlike ordinary uniform sampling, the approach asymptotically preserves the performance of the original algorithm. In our experiments with a wide range of synthetic and real-world datasets, *k*-means-lite matches the performance of *k*-means when C is constructed using 30 samples of size 40 + 2k. Although the 30-sample choice proves to be a generally reliable rule, when the proposed approach is used to scale *k*-means++ (we call this scaled version *k*-means-lite++), *k*-means++' performance is matched in several cases, using only five samples. These two new algorithms are presented to demonstrate the proposed approach, but the approach can be applied to create a constant-time version of any other *k*-means clustering algorithm, since it does not modify the internal workings of the base algorithm.