

1163-03-86

**Mehmet M. Dalkilic\*** (dalkilic@iu.edu), 700 North WoodLawn Dr, Office 3056, Bloomington, IN 47408, and **Kurban Hasan** (hakurban@gmail.com), 700 North Woodlawn Dr, Bloomington, IN 47408. *Teaching an Old Dog New Tricks: Making EM work with Big Data using Heaps.*

Contemporary data mining algorithms are easily overwhelmed with truly big data. While parallelism, improved initialization, and ad hoc data reduction are commonly used and necessary strategies, we note that (1) continually revisiting data and (2) visiting all data are two of the most prominent problems—especially for iterative learning techniques like expectation-maximization algorithm for clustering (EM-T). To the best of our knowledge, there is no freely available software that specifically focuses on improving the original EM-T algorithm in the context of big data. We demonstrate the utility of CRAN package DCEM that implements an improved version of EM-T which we call EM\* (EM star). DCEM provides an integrated and minimalistic interface to EM-T and EM\* algorithms, and can be used as either (1) a stand-alone program or (2) a pluggable component in existing software. We show that EM\* can both effectively and efficiently cluster data as we vary size, dimensions, and separability. (Received August 11, 2020)