

(1.1) How big is too big? Clustering in BIG DATA with the Fantastic 4

Abstract. What is big data? For this talk "big" refers to the number of samples (n) and/or number of dimensions (p) in static sets of feature vector data; or the size of (similarity or distance) matrices for relational clustering. Objectives of clustering in static sets of big numerical data are acceleration for loadable data and feasibility for non-loadable data.

Three ways currently in favor to achieve these objectives are (i) streaming (online) clustering, which avoids the growth in (n) entirely; (ii) chunking and distributed processing; and (iii) sampling followed by very fast (usually 1-2% of the overall processing time) noniterative extension to the remainder of the data. Kernel-based methods are mentioned, but not covered in this talk.

This talk describes the use of sampling followed by non-iterative extension that extend each of the "Fantastic Four" to the big data case. Three methods of sampling are covered: random, progressive, and minimax. The last portion of this talk summarizes a few of the many acceleration methods for each of the Fantastic Four.

WHICH ARE? Four classical clustering methods have withstood the tests of time. I call them the Fantastic Four:

- Gaussian Mixture Decomposition (GMD, 1898)
- Hard c-means (often called "k-means," HCM, 1956)
- Fuzzy c-means (reduces to HCM in the limit, FCM, 1973)
- SAHN Clustering (principally single linkage (SL, 1909))

The first three models apply to feature vector data. All three define good clusters as part of extrema of optimization problems defined by their objective functions, and in this talk, alternating optimization (known as expectation-maximization (EM) for GMD) is the scheme for approximating solutions. Approximate clustering with HCM, FCM and GMD based on literal clustering of a sample followed by non-iterative extension is discussed. Numerical examples using various synthetic and real data sets (big but loadable) compare this approach to incremental methods (spH/FCM and oIH/FCM) that process data chunks sequentially.

This portion of the talk concludes with a "recommendation tree" for when to use the various c-means models.

The SAHN models are deterministic, and operate in a very different way. Clustering in big relational data by sampling and non-iterative extension proceeds along these lines. Visual assessment of clustering tendency (VAT/iVAT) builds and uses the minimal spanning tree (MST) of the input data. Extension of iVAT to scalable iVAT (siVAT) for arbitrarily large square data is done with minimax sampling, and affords a means for visually estimating the number of clusters in the literal MST of the sample. siVAT then marries quite naturally to single linkage (SL), resulting in two offspring: (exact) scalable SL in a special case; and clusiVAT for the more general case. Time and accuracy comparisons of clusiVAT are made to crisp versions of three HCM models; HCM (k-means), spHCM and oIHCM; and to CURE.

Experiments synthetic data sets of Gaussian clusters, and various real world (big, but loadable) are presented.