Jim C. Bezdek: How big is too big? Clustering in (static) BIG DATA with the Fantastic 4

Abstract. 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 (NxN) (similarity or distance) matrices for relational clustering. Objectives of clustering in static sets of big numerical data are *acceleration* for loadable data and *approximation* for non-loadable data. The Fantastic Four are four basic (aka "naïve") classical clustering methods:

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

This talk describes approximation of literal clusters in non-loadable static data. The method is sampling followed by very fast (usually 1-2% of the overall processing time) non-iterative extension to the remainder of the data with the nearest prototype rule. Three methods of sampling are covered: random, progressive, and MaxiMin. The first three models apply to feature vector data and find partitions by approximately optimizing objective function models with alternating optimization (known as expectation-maximization (EM) for GMD). Numerical examples using various synthetic and real data sets (big but loadable) compare this approach to incremental methods (spH/FCM and olH/FCM) that process data chunks sequentially.

The SAHN models are deterministic, and operate in a very different way. Clustering in big relational data by sampling and non-iterative extension begins with visual assessment of clustering tendency (VAT/iVAT). Extension of iVAT to scalable iVAT (siVAT) for arbitrarily large square data is done with Maximin 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 olHCM; and to CURE. Experiments synthetic data sets of Gaussian clusters, and various real world (big, but loadable) are presented.