Refining Initial points for K-Means clustering

Technical update(s) in the ML-related component of the method

In my paper we saw the method for refining initial points for k means clustering. The method discussed in the paper is scalable and very powerful when we're working on very large datasets. The procedure is motivated by the observation that subsampling can provide guidance regarding the location of the modes of the joint probability density function assumed to have generated the data. By initializing a general clustering algorithm near the modes, not only are the true clusters found more often, but it follows that the clustering algorithm will iterate fewer times prior to convergence.

We got great results on synthetic as well as real world data. Refiner solutions were between 2.34 (d=3) and 6.44 (d=50) times closer to the true Gaussian means than solutions from the random initial point and between 1.09 (d=3) and 4.80 (d=50) times closer than solution computed from "Refined (J=1)" initial point for synthetic data. When we used the method on small real world dataset like Image Segmentation data set from UCI ML Repository. The amount of information gained on average by the solutions computed from the refined point was 2.6222 time that of the solution computed over the random initial point. on average, solutions computed from the refined initial points (J=10) reduced distortion by 44.41% over solutions computed from random initial points. Furthermore, the average distortion is decreased by 9%. When we used method on large real world dataset like Reuters Information Retrieval Data Set. On average the distortion of a solution obtained by starting from a refined initial point was about 80% of the corresponding distortion obtained by clustering from the corresponding randomly chosen initial starting point. Computational results on the Reuters database of newswire stories in 300 dimensions indicate a drop in distortion by about 20%. . one cannot guard against the possibility of points from the tails appearing in the subsample.

We have to overcome the problem that the estimate is fairly unstable due to elements of the tails appearing in the sample. Due to random subsampling we got noise in our data points in skew distributions and on these noisy estimates smoothing should be done in a very optimal way. We can find a new way for smoothing so that noise can be removed from samples and all the similar data points which produce less variance could be refined together. We could

research for the implementation for this method in other clustering methods also like Hierarchical clustering, Agglomerative clustering etc as this method could help a lot to reduce time complexity for huge databases . As we know a clustering approach that is to estimate the density and attempt to find the maxima (“bumps”) of the estimated density function. But density estimation in high dimensions is difficult , as is bump hunting so we can also use some technique to find density estimation in large dimensions.