# Refining Initial Points for K-Means Clustering

Paul S. Bradley Usama M. Fayyad Microsoft Research, Microsoft Corporation, One Microsoft Way Redmond, WA 98052

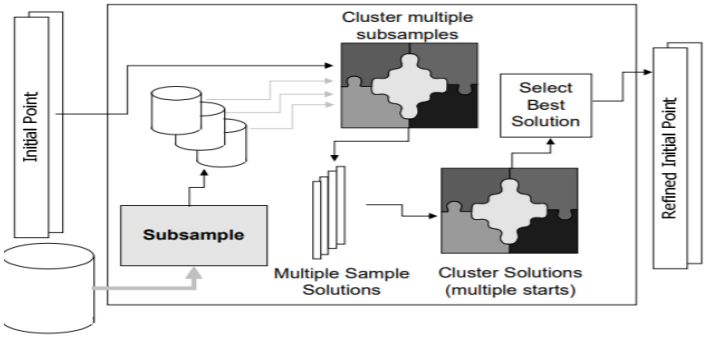
Practical approaches to clustering use an iterative procedure (e.g. K-Means, EM) which converges to one of numerous local minima. It is known that these iterative techniques are especially sensitive to initial starting conditions. We present a procedure for computing a refined starting condition from a given initial one that is based on an efficient technique for estimating the modes of a distribution. The refined initial starting condition allows the iterative algorithm to converge to a "better" local minimum. The method can be used for large clustering problems.

## 1 Background

Clustering is an important area of application for a variety of fields including data mining, statistical analysis, statistical data analysis , compression , and vector quantization.[1] The fundamental clustering problem is that of grouping together data items which are similar to each other. The most general approach to clustering is to view it as a density estimation problem . [2]We assume that in addition to the observed variables for each data item, there is a hidden, unobserved variable indicating the "cluster membership" of the given data item. Hence the data is assumed to arrive from a mixture model and the mixing labels (cluster identifiers) are hidden.In general, a mixture model M having K clusters $C_i$, i=1,. . . ,K, assigns a probability. The KMeans algorithm finds locally optimal solutions minimizing the sum of the L2 distance squared between each data point and its nearest cluster center.[3] . The KMeans algorithm finds locally optimal solutions minimizing the sum of the L2 distance squared between each data point and its nearest cluster center ("distortion") which is equivalent to a maximizing the likelihood given the assumptions listed above.[1] There are various approaches to solving the problem of determining (locally) optimal values of the parameters given the data. Iterative refinement approaches, which include EM and K-Means, are the most effective.

## 2 Refining Initial Conditions

We address the problem of initializing a general clustering algorithm, but limit our presentation of results to K-Means. Since no good method for initialization exists, we compare against the defacto standard method for initialization i.e. randomly choosing an initial starting point. A solution of the clustering problem is a parameterization of each cluster model. This parameterization can be performed by determining the modes of the joint probability density of the data and placing a cluster centroid at each mode. Hence one clustering approach is to estimate the density and attempt to find the maxima of the estimated density function. Density estimation in high dimensions is difficult. So what do we do is that we take multiple subsamples, say J, are drawn and clustered independently producing J estimates of the true cluster locations. To avoid the noise associated with each of the J solutions, we employ a "smoothing" procedure. However, to "best" perform this smoothing, one needs to solve the problem of grouping the K*J points (J solutions, each having K clusters) into K groups in an "optimal" fashion.




## 3 The Algorithm

The refinement algorithm initially chooses J small random sub-samples of the data, $S_i$ , i=1,. . . ,J. The subsamples are clustered via K-Means with the proviso that empty clusters at termination will have their initial centers re-assigned and the sub-sample will be re-clustered. The sets $CM_i$ , i=1,. . . ,J are these clustering solutions over the sub-samples which form the set CM. CM is then clustered via K-Means initialized with $CM_i$ producing a solution $FM_i$. The refined initial point is then chosen as the $FM_i$ having minimal distortion over the set CM.

### 3.1 Algorithm Refine

Algorithm Refine( SP, Data, K, J)
0. CM = $\phi$
1.*For* $i = 1, \ldots, J$
*Let* $S_i$ be a small random subsample of Data, Let $CM_i$ = KMeansMod(SP, $S_i$ , K), CM = CM $\cup$ $CM_i$
2. FMS =$\phi$
3.*For* $i = 1, \ldots, J$
*Let* $FM_i$ = KMeans($CM_i$ , CM, K), Let FMS = FMS $\cup$ $FM_i$
4. Let FM = ArgMinDistortion($FM_i$ ,CM )
5. Return (FM)

## 4 Results on Data

Synthetic data was created for dimension d = 2, 3, 4, 5, 10, 20, 40, 50 and 100. For a given value of d, data was sampled from 10 Gaussians (hence K=10).
Methodology:-The goal of this experiment is to evaluate how close the means estimated by classic K-Means are to the true Gaussian means generating the synthetic data. We compare 3 initializations:
1. No Refinement: random starting point chosen uniformly on the range of the data.
2. Refinement (J=10): a starting point refined from (1) using our method. The size of the random subsamples being 10 percent of full dataset size and the number of subsamples taken being 10.
3. Refinement (J=1): same as 2 but over a single random subsample of size 10 percent.

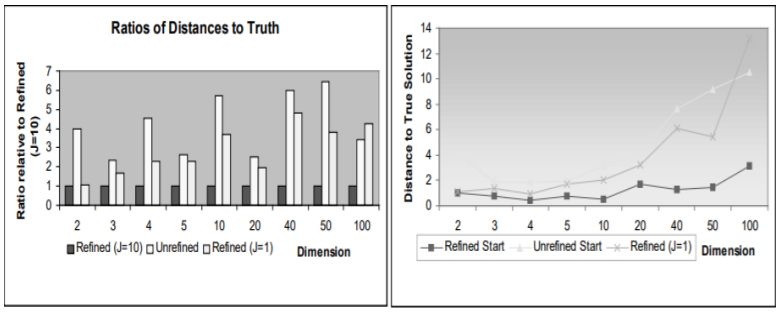



Figure 1: Comparing performance as dimensionality increases

Results-: 1. For dimensions 2-50, the refinement method (Refined (J=10)) always did better than the random starting point (Unrefined) and the point refined over 1 subsample (Refined (J=1)).
2. For dimension 100, in 9 of the 10 independent trials our refinement method did better than the random starting point.
3. Refiner solutions are 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.

[1] O. L. Mangasarian and W. N. Street. Clustering via concave minimization. *Advances in Neural Information Processing Systems 9*, 1997.

[2] D. W. Scott. Multivariate density estimation. *New York: Wiley*, 1992.

[3] S. Z. Selim and M. A. Ismail. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1984.