Homework # 5 (Due Tuesday, October 17, 2000)

Note: For questions 1 and 2 you can refer to CURE and to Scatter/Gather, both on the class WEB page.

1. Scalability Issues

Hierarchical clustering algorithms require at least O(m^2) time and thus, are impractical to use directly on larger data sets. One possible technique for reducing the time required is to sample the dataset. For example, if K clusters are desired, then $\sqrt{m}$ points are sampled from the m points, and a hierarchical clustering algorithm will produce a hierarchical clustering in O(m) time. K clusters can be extracted from this hierarchical clustering by looking at the clusters on the K^{th} level of the dendogram. The remaining points can then be assigned in linear time, by using a variety of strategies. For example, the centroids of the K clusters can be calculated and then each of the m - $\sqrt{m}$ points can be assigned to the closest cluster.

For each of the following types of data or clusters, discuss briefly if a) sampling will cause problems and b) what those problems are. Assume that the sampling technique randomly chooses points from the total set of m points and that the unmentioned characteristics of the data or clusters are as optimal as possible. In other words, focus only on problems caused by the particular characteristic mentioned. Finally, assume that K is very much less than $\sqrt{m}$.

1) Data with small and large clusters.

2) Data with a small number of outliers, i.e., atypical points, which **cannot** be discarded.

3) Data with widely different densities.

4) Data with a large percentage of noise points (>95%) and small dense clusters.

2. Scalability Issues

Another approach to making hierarchical clustering more computationally practical is to use a partitioning approach. First the n data points are partitioned into p almost equal sized partitions of size n/p, where $p > K$ (K is the number of desired clusters). Each group is then separately clustered into $q > 1$ clusters using a hierarchical approach. (Each of the new clusters has (n/p)/q points.) This procedure is repeated, using the clusters as the new "points" and using the same clustering approach until the desired number of clusters is reached.

Comment on the strengths and weakness of this approach. You may want to consider some of the examples given in problem 1.

3. Scalability Issues

In the CURE conclusions, the statement is made that sampling and partitioning are a way "to reduce the size of large input data sets without reducing the cluster quality." Given your answers to questions 1 and 2, what do you think of this statement?

4. Noise and outliers

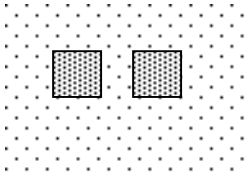
a) One approach to eliminating noise and outliers during hierarchical clustering is to periodically eliminate small clusters. (CURE follows this approach and gives a justification.) Comment on this approach. In particular identify when this approach would work well and when it might break down or be inappropriate.

b) Consider the technique that DBSCAN uses to handle outliers and/or noise. Identify when this approach would work well and when it might break down or be inappropriate.

5. Behavior of K-means, MIN and MAX

The outlined regions in the following diagrams are the desired clusters. Indicate for each diagram whether MIN, MAX, and K-means will find the desired clusters and briefly describe your reasoning. As usual, *K* is the number of desired clusters.

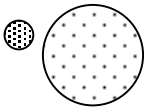
i) different sizes, $K = 3$

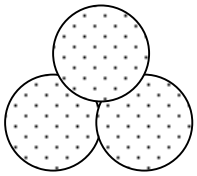


ii) noise, $K = 2$

Note: the noise is really random not a uniform grid, as shown.

iii) concave shapes with holes, $K = 2$



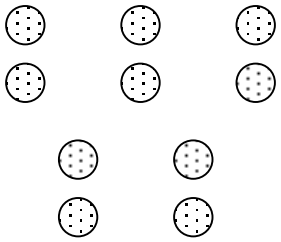


iv) Different densities, $K = 2$ (Same number of points in both clusters.)

v) no separation, $K = 3$

6. Wards method and K-means.

On the following data set, will K-means (with random initialization) or Wards method do better? Why?



7. Cure, DBSCAN, and Chameleon. (Write brief answers for the following.)

a) For DBSCAN, both border points and noise point have fewer than MinPts within a radius of Eps? What is the difference?

b) What would be the effect on DBSCAN if border points were treated as noise points?

c) What type of clusters does DBSCAN find, i.e., well separated, center-based, contiguous or density-based?

d) CURE uses a subset of points from a cluster as the "representative" points of the cluster. For DBSCAN, what type of points could be considered to be "representative" points?

e) Briefly discuss the situations when DBSCAN would fail. Draw a 2-D example of one such situation.

f) In CURE, what effect does the α parameter have on the shape of the clusters that are found?

g) Why does Chameleon sparsify the proximity graph before applying METIS to partition the graph?

h) Chameleon tries to "model the data" and only merge clusters which will preserve self-similarity of the sub-clusters. What kinds of similarity does Chameleon try to preserve?

i) Briefly discuss the situations when Chameleon would fail. Draw a 2-D example of one such situation, if possible.

j) Suppose that you are asked to cluster data that consists of the geographical locations of the outbreak of a serious disease. (Assume that the data you are given is the location of the reporting hospital and the number of affected patients at that location.) Which, if any, of these three methods (CURE, DBSCAN, Chameleon) would you use? You can specify none, some or all of the methods, but justify your answer.