Data Mining Homework #4

Shumin Guo

Due Date: Nov. 12th, 2010, beginning of the class.

1. (a) Define the clustering problem, and (b) define the measure commonly used to evaluate the quality of clustering results.

   ANSWER:

   Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used.

   The quality of clustering is estimated using a cost function that measures the average dissimilarity between an object and the representative object of its cluster, such as square error.

2. Compare the advantages and disadvantages of (a)K-means and (b) K-medoids for clustering. (c) Discuss a main challenge to both K-means and K-medoids algorithms.

   ANSWER:

   Advantage:

   The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. The algorithm attempts to determine k partitions that minimize the square-error function. It works well when the clusters are compact clouds that are rather well separated from one another. The method is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is $O(nkt)$, where n is the total number of objects, k is the number of clusters, and t is the number of iterations. Normally, k≪n and t≪n. The method often terminates at a local optimum.

   Disadvantage:

   The necessity for users to specify k, the number of clusters in advance can be seen as a disadvantage. The k-means method is not suitable for discovering clusters with nonconvex shapes or clusters of very different size. Moreover, it is sensitive to noise and outlier data points because a small number of such data can substantially influence the mean value.

   In order to reduce the sensitivity to outliers, K-medoids methods, instead of taking the mean value of the objects in a cluster as a reference point, pick actual objects to represent the clusters, using one representative object per cluster.

   The k-medoids method is more robust than k-means in the presence of noise and outliers, because a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the k-means method. Both methods require the user to specify k, the number of clusters.

3. Compare the COBWEB algorithm against other clustering algorithms such as K-means and K-mediods.

   ANSWER:

   COBWEB is a popular and simple method of incremental conceptual clustering. Its input objects are described by categorical attribute-value pairs. COBWEB creates a hierarchical clustering in the form of a classification tree. While on the other hand, K-means and K-mediods methods are used to cluster quantitative data based on distance.

   COBWEB uses a heuristic evaluation measure called category utility to guide construction of the tree. It incrementally incorporates objects into a classification tree in order to get the highest category utility. And a new class can be created on the fly, which is one of big difference between COBWEB and K-means and K-medoids methods.

   COBWEB provides merging and splitting of classes based on category utility, this allows COBWEB to be able to do bidirectional search. For example, a merge can undo a previous split. While for K-means and K-medoids methods, the clustering is usually unidirectional, which means the cluster of a point is determined by the distance to the cluster center. It might be very sensitive to the outliers in the data.

   COBWEB has a number of limitations. First, it is based on the assumption that probability distributions on separate attributes are statistically independent of one another. This assumption is, however, not always true because correlation between attributes often exists. Moreover, the probability distribution representation of clusters makes it quite expensive to update and store the clusters. This is especially so when the attributes have a large number of values because the time and space complexities depend not only on the number of attributes, but also on the number of values for each attribute. Furthermore, the classification tree is not height-balanced for skewed input data, which may cause the time and space complexity to degrade dramatically. And K-means and K-medoids methods don't have such issues as considerations of probabilities and independence. It only take into consideration of distance, but this feature also renders them unproper for high dimensional data sets.

4. Briefly discuss how recent algorithms deal with clustering for large amount of data, or nonsphere shaped clusters.
   ANSWER:
   - Reducing the number of instances to be maintained while maintaining the distribution of regions/clusters.
     CLARA and CLARANS are representative Sampling Methods. CLARA (Clustering LARge Applications) working on samples instead of the whole data, And CLARANS (Clustering Large Applications based on RANdomized Search).
   - Identifying relevant subspaces where clusters possibly exist.
     Grid: STING (STatistical INformation Grid) Statistical parameters of higher-level cells in the grid can easily be computed from those of lower-level cells.
   - Using summarized information to avoid repeated data access.
     BIRCH using Cluster Feature (CF) and CF tree
     CURE (Clustering Using REpresentitives) is another example.
   - Taking advantage of the property of density. If the data is dense in a higher dimensional subspace, it should be dense in some lower dimensional subspaces.
     CLIQUE is a density-based method that can automatically find subspaces of the highest dimensionality such that high-density clusters exist in those subspaces.

5. Apply the K-means algorithm for the following 1-dimensional points and k=2:1,2,3,4,6,7,8,9. Use 1 and 4 as the starting centroids. You should compute the aggregate dissimilarity for each iteration.
   ANSWER:
   With the initial centers of 1 and 4:
   $\Rightarrow$: Clusters: {1,2} And {3,4,6,7,8,9}; meanC1 = 1.5, meanC2 = 4.83. Aggregate dissimilarity = $0^2 + 1^2 + 1^2 + 0 + 2^2 + 3^2 + 4^2 + 5^2 = 56$.
   $\Rightarrow$: Clusters: {1,2,3} And {4,6,7,8,9}; meanC1=2, meanC2=6.8. Aggregate dissimilarity = $2.75 + 0.689 + 1.369 + 4.71 + 27.43 = 36.95$.
   $\Rightarrow$: Clusters: {1,2,3,4} And {6,7,8,9}; meanC1=2.5, meanC2=7.5. Aggregate dissimilarity = 6+6.96=12.96.
   $\Rightarrow$: Clusters: {1,2,3,4} and {6,7,8,9}. There is no change compared with the last iteration, so we will end here. And the Aggregate dissimilarity = 5+5 = 10.

6. Apply the K-medoids algorithm for the following 1-dimensional points and k=2:1,2,3,4,6,7,8,9. Use 1 and 4 as the starting medoids. To make your job easier, assume that the algorithm is smart to choose the best replacement for each iteration. This will let you do the work using the smallest number of iterations. You should compute the aggregate dissimilarity for each iteration.
   ANSWER:
   With the initial centers of 1 and 4:
   $\Rightarrow$: Clusters {1,2} and {3,4,6,7,8,9}; Aggregate dissimilarity = 1+55= 56.
   $\Rightarrow$ Replace 1 by 2: AggDis = 56. (good)
   $\Rightarrow$ Replace 1 by 3: AggDis = 59.
   $\Rightarrow$ Replace 1 by 6|7|8|9: AggDis is large.
   $\Rightarrow$ Replace 4 by 2|3: AggDis is large.
   $\Rightarrow$ Replace 4 by 6: AggDis = 20. Medoids: 2, 6.
   $\Rightarrow$ Replace 4 by 7: AggDis = 12. Medoids: 2, 7.
   $\Rightarrow$ Replace 4 by 8: AggDis = 12. Medoids: 2, 8.
   $\Rightarrow$ Replace 4 by 9: AggDis is large.
   So, for smallest aggregate dissimilarity, the medoids can be combinations (See Table 1)of 2(3) and 7(8) with aggregate dissimilarity as 12.

Medoid1	Medoid2	DisAgg
2	7	12
2	8	12
3	7	12
3	8	12

Table 1: Possible Modoids.

7. Apply the bottom-up hierarchical algorithm for the following 1-dimensional points and k=2:1,2,3,4,6,7,8,9. You should consider (a) using single linkage, and (b) complete linkage.
For each of (a) and (b), draw the dendrogram. When there is a tie, choose clusters with ”smaller means” before choosing clusters with larger means.
ANSWER:
Single linkage uses nearest neighbor technique, this method define distance between clusters as the distance between the closest pair of objects, where only pairs consisting of one object from each group is considered. The distance $D(r, s)$ can be computed as:

D(r,s) = Min{d(i,j) : Where object i is in cluster r and object j is cluster s}

At each stage of hierarchical clustering, the clusters r and s , for which D(r,s) is minimum, are merged.

The complete linkage, also called farthest neighbor, clustering method is the opposite of single linkage. Distance between groups is now defined as the distance between the most distant pair of objects, one from each group. In the complete linkage method, D(r,s) is computed as

D(r,s) = Max{d(i,j) : Where object i is in cluster r and object j is cluster s}

At each stage of hierarchical clustering, the clusters r and s , for which D(r,s) is minimum, are merged.
Please see the dendrogram for single linkage at Figure 1 and dendrogram for complete linkage at Figure 2.

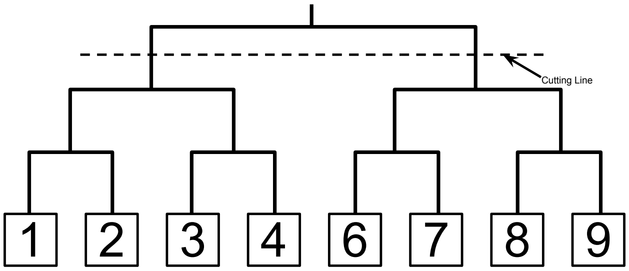

Figure 1: Dendrogram for Single Linkage.

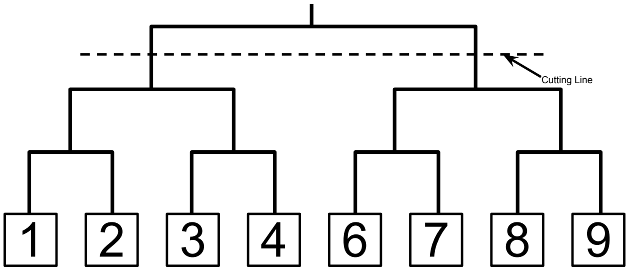

Figure 2: Dendrogram for Single Linkage.

Marks available: 10 marks for each question.