CS498 – Data Mining Lab 3

Clustering

Submission Requirements

- All reports including text and diagrams are submitted in PDF format.
- Projects are submitted as zip archives
- Assignments are submitted via Blackboard.
- All uploaded files must include lab number, title, and student name(s), for example, cs498-dm-lab1-urbain.pdf
- Submit lab report separately from project archive.

Objective

- Apply lessons learned in class to develop and apply a basic clustering algorithm.

1) Feature Selection

- From your data analysis work in labs 1 and 2, select *at least* 6 distinct attributes in addition to income for clustering. By distinct, I'm refereeing to non-redundant, i.e., non-correlated, feature attributes. For example, don't select both forms of education.

2) Make sure your data is properly normalized

- Numerical attributes, e.g., income, are normalized to values between 0 and 1 using min-max normalization.
- Ordinal attributes, i.e., education, are normalized to a range of 0 to 1.
- Boolean attributes are set to either 0 or 1.
- Categorical attributes are assigned an integer enumerated type, e.g., 1, 2, 3, …

3) Distance measurement

- For categorical and Boolean attributes, match is a distance of 0 (min distance), and non-match is a distance of 1 (max distance).
- For numerical and ordinal attributes, measure distance between attributes as absolute difference.
- Measure data object to data object (or centroid) distance as the sum of absolute differences (Block distance), or as the square root of the sum of squared differences (Euclidean).

3) Clustering

- Implement K-Prototype clustering. As described in class, K-Prototype is an extension of K-Means to accommodate non-numerical, i.e., Boolean and categorical, attributes.
- Evaluate your clustering algorithm by running several trials (~25) without the income attribute as follows (in each case feel free to experiment):
    - K - # clusters: 2, 3, 4, etc.
    - t - # iterations: 5, 10, 50.
    - For each trial, measure total inter- and intra-cluster distance.
        - Measuring intra cluster distance options:
            - Ok: for each cluster determine the two elements assigned to that cluster that are the furthest apart. Sum this distance for each cluster.
            - Better: sum the normalized sum of squared differences between points within each cluster:

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (d(p, c_i))^2$$

        - Measuring intra cluster distance options:
            - Ok: Sum the distances between each cluster centroid.
            - Better: Use the same sum of squared measurement (above) to measure **inter-cluster dissimilarity** by summing the distance between each point within a cluster and the cluster centroids it is *not* a member of.

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (d(p, c_i))^2$$

4) Evaluation and Deliverables

- Tabulate and analyze your results.
    - What parameters (t and k) yielded the best clusters with respect to intra and inter centroid measurements?
    - What was the best parameterization for your clustering algorithm for isolating data objects based on the income attribute?
    - What attributes are the most effective for generating clusters for income (without using the income attribute!)?

Extra credit:

- Run additional trials with different combinations of attributes.
- Devise an automated convergence criteria using intra centroid similarity and inter centroid dissimilarity.
- Implement K-Medoids or Hierarchical Agglomerative clustering.

As in lab 1, you may use Java, SQL, R, Python, Octave/Matlab, or Excel in any combination to complete this assignment. I encourage you to continue experiment and try different approaches.

Submission:
Submit your lab report as a single PDF file and an archive you're your project to Blackboard following the guidelines listed at the beginning of this assignment.

Due: Prior to lab week 5.

Please let me know if anything is not clear.