Project 3: Clustering

UE 141 on Data Mining Spring 2013

Each team needs to submit a written report and presentation slides by May 1 4pm electronically. The maximum number of members in a team is 2.

The goal of this project is to evaluate your understanding of clustering in real-world applications, and get you familiar with the clustering algorithms in Weka.

You are asked to complete the following three tasks.

- Please discuss clustering in the context of a real-world application. What are the data sets? What is the expected output? How can clustering help decision making in this application?
- Conduct K-Means clustering on a data set you choose. Run the SimpleKmeans algorithm on the data. Try different number of clusters (2,3,4) and report the within cluster sum of squared errors. Visualize the cluster assignments of the above three trials, which number of clusters gives you the best cluster assignment? Save the visualization of the best cluster assignment and put it on the report.
- Conduct hierarchical clustering on one attribute of weather data set. Choose either the temperature or humidity attribute. Remove all the other attributes. Visualize the hierarchy output by the algorithm. Report the order of merging.

Your report should include: 1) A short description of the application and answers to the three questions in part 1, 2) a short description of the data set you use for Weka experiments. Answer the questions in part 2, and 3) answers to the questions in part 3.

Your slides should contain a short overview of your report. You can ignore the details but show the most interesting points.

Note that plagiarism/copying is not allowed and may result in an F in the grades of all the team members. Academic integrity policy can be found at http://www.cse.buffalo.edu/shared/policies/academic.php