Project 2: Clustering Algorithms

Due: Code and Report should be submitted by 5pm on Oct. 28. Demo is on Oct. 29.

The gene data sets are at: http://www.cse.buffalo.edu/~jing/cse601/fa13/docs/cho.txt and http://www.cse.buffalo.edu/~jing/cse601/fa13/docs/iyer.txt). A short description of the two datasets can be found at http://www.cse.buffalo.edu/~jing/cse601/fa13/docs/README.txt

Complete the following tasks:

1. Implement three clustering algorithms to find clusters of genes that exhibit similar expression profiles: K-means, Hierarchical Agglomerative clustering with Single Link (Min), and one from (density-based, mixture model, spectral). Compare these three methods and discuss their pros and cons.

For each of the above tasks, you are required to validate your clustering results using the following methods:

- Choose an external index (Rand Index or Jaccard Coefficient) and compare the clustering results from different clustering algorithms. The ground truth clusters are provided in the data sets.
- Choose an internal index (Silhouette or Correlation) and compare the clustering results.
- Visualize data sets and clustering results by Principal Component Analysis (PCA). You can implement PCA yourself or use any existing implementation or package.

2. Set up a single-node Hadoop cluster on your machine and implement MapReduce K-means. Compare with non-parallel K-means on the given data sets. Try to improve the running time. For single-node Hadoop set up, follow the instructions at http://www.cse.buffalo.edu/~jing/cse601/fa13/docs/setup.pdf

Your final submission should include the following:

- Codes: Three clustering algorithms, and MapReduce K-means algorithm.
- Report: Describe the flow of all the implemented algorithms. Compare the performance of these approaches using visualization, external, and internal index on the two given data sets. State the pros and cons of each algorithm and any findings you get from the experiments.

The details about Demo will be released at **5pm on Oct. 28**. Please note:

- Two new data sets will be given to check your implemented clustering algorithms and validation measures. The data format is the same with the data sets we already provided.
- During the demo, you will be asked to adopt specific parameter setting and run your code.

Note that copying code/report from another group or source 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