Project 2: Clustering Algorithms

Due: Code should be submitted by Nov. 12. Report is due on Nov. 13 at 11:00am.

The time series microarray data sets are at:
http://www.cse.buffalo.edu/~jing/cse601/fa12/docs/cho.txt and
http://www.cse.buffalo.edu/~jing/cse601/fa12/docs/iyer.txt). A short description of the two datasets can be found at http://www.cse.buffalo.edu/~jing/cse601/fa12/docs/README.txt

Complete the following tasks:

- Implement three clustering algorithms, each of which belongs to one category of the approaches introduced in class (partitional, hierarchical, density-based, mixture model, spectral) to find clusters of genes which exhibit similar expression profiles. Compare the three methods and discuss their pros and cons.
- Implement one clustering ensemble approach to combine the clustering results of the three clustering algorithms. Compare the clustering ensemble approach with each base clustering algorithm on the two datasets.
- Set up a single-node hadoop cluster on your machine and implement MapReduce K-means. Compare with non-parallel K-means on the given datsets. Follow the instructions at http://www.cse.buffalo.edu/~jing/cse601/fa12/docs/setup.pdf

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

- Choose an external index and compare the clustering results from different clustering algorithms with an external index (the ground truth clusters are provided in the data sets).
- Choose an internal index and compare the clustering results.

Your final submission should include the following:

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

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