CMSC 603 – High Performance Distributed Systems

Virginia Commonwealth University, Fall 2018

Due date: December 2, 2018

J. Maillo, I. Triguero, F. Herrera, A MapReduce-based k-Nearest Neighbor Approach for Big Data Classification, IEEE Trustcom/BigDataSE/ISPA, pp. 167-172, 2015

The k-Nearest Neighbor classifier is one of the most well-known methods in data mining because of its effectiveness and simplicity. Due to its way of working, the application of this classifier may be restricted to problems with a certain number of examples, especially, when the runtime matters. However, the classification of large amounts of data is becoming a necessary task in a great number of real-world applications. This topic is known as big data classification, in which standard data mining techniques normally fail to tackle such volume of data.

This assignment consists of implementing the KNN classifier in Hadoop or Spark following the recommendations for the MapReduce implementation detailed in the research article. You may also propose different alternatives to compute the KNN as long as it is performed in a distributed way using the API from the frameworks. Finally, conduct the following experiments and write a small report including:

1. Evaluate the performance (runtime) of the Hadoop or Spark implementations using the datasets provided (small and medium) in the first assignment.
2. Evaluate the performance (runtime) of the Hadoop or Spark implementations as compared with the sequential and parallel version implemented in the first assignment using the datasets provided (small and medium). Accuracies must be the same!
3. Evaluate the scalability of the Hadoop and Spark implementations using bigger datasets provided in the KEEL dataset repository. You may run the experiments in the maple.cs.vcu.edu server with up to 56 cores.

Create a .zip file containing the report, source code, .jar files, and datasets, together with the instructions to run the experiments, and upload it into the blackboard assignment.