Assignment 3: KNN on Hadoop

Steven M. Hernandez
Department of Computer Science
Virginia Commonwealth University
Richmond, VA USA
hernandezsm@vcu.edu

I. INTRODUCTION

In this project we look at developing a MapReduce implementation in Hadoop to compute k nearest neighbors to compare with previously developed implementations such as the sequential program, threaded program and most recently the CUDA implementation which ran on the GPU.

II. DISCUSSION

Unfortunately, there appears to be some issue with my implementation when comparing the accuracy from previous assignments. Additionally, I was unable to complete all aspects of the project to run on the server, thus runtime include only running on my machine which overall defeats the purpose of the distributed framework provided by Hadoop.

TABLE I
RESULT

Dataset	Time (ms)	Accuracy (%)	Speedup
Sequential	8433	49.59	N/A
CPU (1 thread)	8446	49.59	0.998
CPU (2 threads)	4402	49.59	1.916
CPU (4 threads)	2270	49.59	3.175
CPU (8 threads)	1192	49.59	7.075
CPU (16 threads)	695	49.59	12.13
CPU (2048 threads)	351	49.59	24.03
GPU	244	49.59	34.56
Hadoop	50,432	41.71	0.167

III. CONCLUSION

Hadoop and the MapReduce paradigm provides methods to distribute work across a cluster for distributed computing. However, with this ability, there is quite a bit of complexity in implementation. The system is very verbose which in my opinion makes development hard.