PC-2016/17 Course Project Template

Tommaso Ceccarini
E-mail address
tommaso.ceccarini1@stud.unifi.it

Federico Schipani
E-mail address
federico.schipani@stud.unifi.it

Abstract

KMeans algorithm is one of the most popular method for clustering analysis. In our work we provide a CUDA implementation that use an Nvidia GPU to solve the clustering problem. We also provide a performance analysis with the purpose of compare the performance of our parallel CUDA implementation with a sequential implementation written in C language.

1. Introduction

KMeans algorithm is one of the most popular method for clustering analysis. The purpose of the cluster analysis is that of divide data into meaningfull group, called cluster. The resultant cluster should then capture the structure of the data.

KMeans methods attempt to do this by evaluating a similarity measure according to the mean value of the data that are cointained in the clusters. So, given a set of observation $(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N})$ where each observation is a $P - dimensional$ real vector, k-means clustering aims to partition the N observation into $K(\leq N)$ sets $\boldsymbol{S} = \{S_1, S_2, \ldots, S_K\}$ so as to minimize the within-cluster sum of squares. In other words, its objective is to find:

$$\arg\min_{\boldsymbol{S}}$$

$$\sum_{i=1}^{k} \sum_{x \in S_i} ||\boldsymbol{x} - \boldsymbol{\mu}_i||^2 [1]$$

References

[1] Wikipedia. K-means clustering — wikipedia, the free encyclopedia, 2017. [Online; accessed 23-January-2017].