Weekly Journal

Leila Uy

July 26, 2021

1 Work Update

At this point in time, Jishnu and I are focused on the writing and official output creation of our project. We decided to split the work up where I start the writing portion and Jishnu focuses on running the code to get our results and create the time complexity model.

1.1 Time Complexity Model

If we do a mass test of different data sizes, we can plot the points where x is the input size and y is the time it takes. We can then use our these plots to find the best fit model. We can use these to create estimations for bigger data sets.

1.2 Big O Notation

Last meeting I brought up the possibility of calculating Big O (upper bound) notation for showing the improvement in time. I attempted the process and got this:

Notation:
k = num. of clusters
p = num. of processors
n = input size (num. of rasters)
$F(n)$ = parallel k-means clustering
$G(n)$ = serial k-means clustering

$$O(n) = 50[11 + \frac{n}{p} + 13 + F(n) + 3 + \frac{n(20)}{p} + 2]$$

$$\iff O(n) = 50[29 + F(n) + \frac{21}{p}]$$

$$\iff O(n) = 50(F(n)) + \frac{1050}{p} + 1450$$

$$\approx F(n)$$

Of course it's not completely accurate becuase it assumes that certain function calls are constant time complexity, but in the estimation it doesn't matter because when comparing the serial and parallel, it doesnt play a big role. The main difference is the addition of the value p and the change from $F(n)$ to $G(n)$. When using time compexity we remove constant values and p is constant so we are left with comparing $F(n)$ to $G(n)$. So we have to find the time complexity of the two different functions. So if we want to go the route of time complexity, we have to find it for each function. I think time complexity models will suffice anyway.

1.3 Writing

I'll send out a link for a Google Doc later this week that you can comment/edit.

2 Literature Review

One of the articles I read this week [1] was about a package in R that can be used to find the time complexity of a function. The package is called GuessCompx and it estimates the time and memory complexities of the algorithm and fits it to one of seven complexities. The problem is that parallel programming is out of the scope of the package so that turned out to be a dead end...

The other article I read this week [2] was studying ecoregion classification and finding the best suited classification system to use. An article by Thompson et al. compared three classification systems and concluded that choosing a clasification system is dependent on the ultimate goal and the methodologies used in developing the classification system. This does not really effect us because the important part about our classification system is choosing important environmental variables, but it provided me more context on other ecoregion classification systems.

References

[1] Marc Agenis-Nevers, Neeraj Dhanraj Bokde, Zaher Mundher Yaseen, and Mayur Kishor Shende. An empirical estimation for time and memory algorithm complexities newly developed r package. *Multimedia tools and applications*, 80(2):2997, 2021.

[2] Robert S Thompson, Sarah L Shafer, Katherine H Anderson, Laura E Strickland, Richard T Pelltier, Patrick J Bartlein, and Michael W Kerwin. Topographic, bioclimatic, and vegetation characteristics of three ecoregion classification systems in north america: Comparisons along continent-wide transects. *Environmental management*, 34(S1):S125–S148, 2004.