CS 5433 - Big Data Management

Spring 2018

Big Data Group Project – Phase II

Group: 3

Baath, Harpinder

Polavarapu,Tejaswanth

Shankara, Akshay

Dandamudi, Madhu Kiran

1. Tasks and deadline:

- ~~Understanding how MLib library works in Spark~~ ~~03/28~~
- ~~Working with feature selection with respect to the data~~ ~~03/30~~
- ~~Converting feature into chromosome representation~~ ~~04/04~~
- ~~Implementing machine learning algorithms in map reduce~~ ~~04/06~~
  ~~to calculate predictive accuracy~~
- Implementing Selection method in Mapreduce   04/10
- Implementing Crossover and mutation methods in Mapreduce   04/13
- Integrating all the above methods in Mapreduce   04/21
- Testing and debugging   04/21 - 04/26

PS: The tasks which are completed have a strikethrough

2. Describe the output you expect to show and types of graphs you expect to produce.

- The final output we expect, is to provide the **set of features** that are **predicted** to be important by the Genetic Algorithm running under several **ML algorithms** to find the accuracy(fitness) for the off-spring/population. We would also include the accuracy measured relatively by the model respectively. (MODEL - #of features - #accuracy measured)

- We expect the results for **best features to be changing** with respect to the **change in the accuracy prediction algorithm** in the model and we predict that this happens **because of the prediction strategy** that each ML algorithm uses for accuracy prediction.

- Once, we obtain the **best feature set**, we want to calculate the **Mean Absolute Deviation** for the minimal feature set. This would help us understand the how much **difference of deviation** is present between the mean of the predictions and the predicted values. This acts as a **factor of comparison** of the models.

- We plan to plot a graph of comparison with the labels "**Predication Model**" and "**# of minimal features**". This graph could provide a simplistic understanding of how the **Genetic Algorithms Fitness Function Calculator** effect its prediction results as a diagrammatic illustration