CLASSIFICATION USING SPARK-ENABLED SWARM INTELLIGENCE ALGORITHMS

Kendrick Dahlin
North Dakota State University

Firefly Algorithm

Swarm intelligence (SI) describes a collection of models that imitate the behavior of natural phenomena such as birds or ants. Individual entities all act upon the same principles within a group, responding to others in the group and their environment to swarm to a best solution. This behavior is de-centralized and self-organized.
The Firefly Algorithm (FA) [2] is a swarm intelligence algorithm modeled after the flashing light emitted by fireflies. A firefly is most attracted to the most intense light they observe. An intensity of another firefly is inversely proportional to the distance, and proportional to the brightness of the firefly.

Algorithm 1 Firefly Algorithm

```
Objective function f(x),    x = (x_1, ..., x_d)^T
Generate initial population of fireflies x_i (i = 1, 2, ..., n)
Light intensity I_i at x_i is determined by f(x_i)
Define light absorption coefficient γ
while t < MaxGeneration do
    for i = 1 to allfireflies do
        for j = 1 to allfireflies do
            if I_j > I_i then
                Move firefly i towards j in d-dimension
                Attractiveness varies with distance r via exp[−γr^2]
                Evaluate new solutions and update light intensity
            end if
        end for
    end for
    Rank the fireflies and find the current best
end while
Postprocess results and visualization
```

Fig. 1: Big fancy graphic.

Apache Spark

On large scales, many methods of processing data are either computationally expensive or insufficient. Apache Spark is a “multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters."[**Spark**] Spark utilizes parallel processing to segment programs into sub-tasks, and run these sub-tasks simultaneously.
A key feature of Spark is the Resilient Distributed Dataset (RDD). An RDD is a “collection of elements partitioned across the nodes of the cluster than can be operated on in parallel." An RDD can be created by parallelizing data in the driver function. In our experiments we utilize an RDD to parallelize elements of the FA.

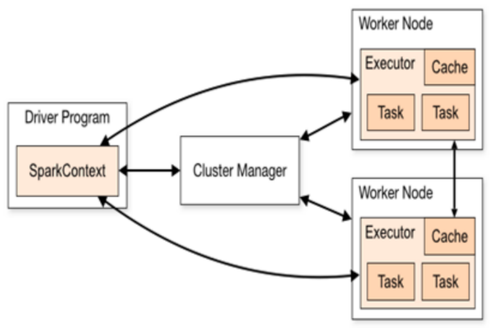


Fig. 2: Apache Spark Framework [1]

Implementation

We implemented four variations of driver programs that ran the Firefly Algorithm in parallel using Apache Spark. Each program measures time from the beginning of the driver program to the end, including pre-processing data and measuring accuracy.

1. **Parallel Data** Python program that creates RDD from data.
2. **Parallel Particles** Python program that creates RDD from particles initialized outside of FA.
3. **Measured Execution Time** Instead of measuring time of entire driver function, only measures parallel parts.
4. **Java** Identical to program 1, but in Java.

Algorithm 2 Driver Program

```
Begin spark session
Read data
Split data into X and y
Transform y into integer values
Standard scale X
rdd = sc.parallelize()
weights = rdd.mapPartitions(lambda x: firefly(x)).collect()
model = avg(weights)
```

Fig. 3: Driver Program Pseudo Code

Results

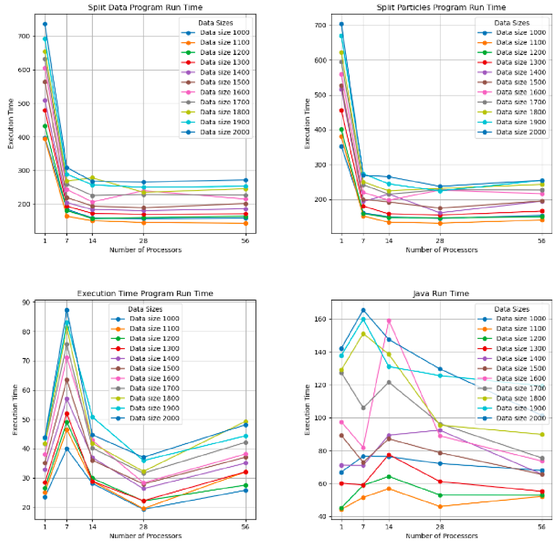
The two metrics we utilized to measure scalability were speedup and scaleup. [**Gramma**]. Speedup measures the difference in time of as the number of nodes are increased.

$$speedup = \frac{T_1}{T_n} \tag{1}$$

where T_1 is the time it takes for an algorithm to be run using one node, and T_n is the time it takes for an algorithm to be run using n nodes.
Scaleup measures the simultaneous increase in data size and nodes by the same ratio.

$$scaleup = \frac{T_{sn}}{T_{Rsn}} \tag{2}$$

where T_{sn} is the running time for data size s with n nodes and T_{Rsn} is the running time for data size $R * n$ and $R * n$ nodes.



Results Cont.

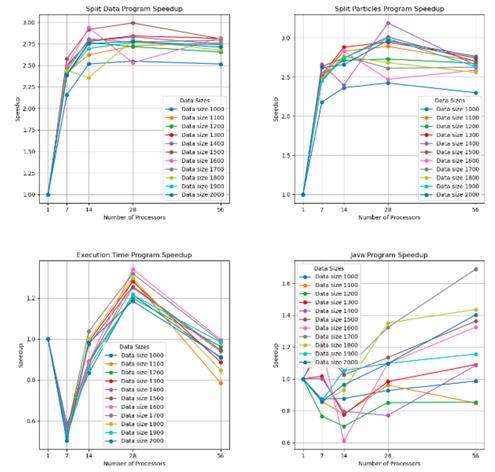


Fig. 5: Speedup

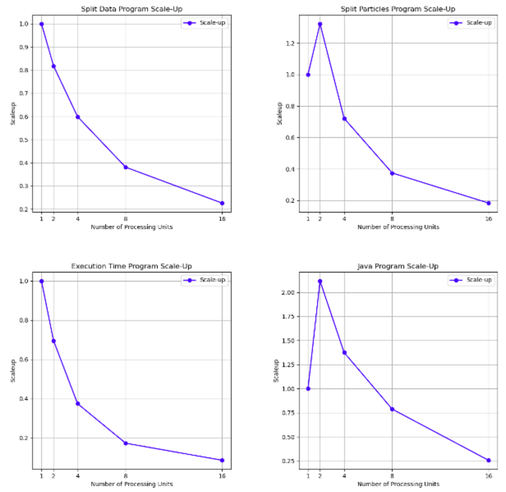


Fig. 6: Scaleup

Acknowledgements

I would like to thank my research advisor Dr. Simone Ludwig, and mentor for this project, Aaron Misquith.

References

[1] Apache Spark. *Spark Overview*. Accessed: 2024-07-25. 2024. URL: https://spark.apache.org/docs/2.1.0/.

[2] Xin-She Yang. “Firefly Algorithms for Multimodal Optimization”. In: 2010.