Machine Learning Algorithms over Ciel

Brett Lagerwall

University of Cambridge

March 12, 2013

Choice of Execution Engine – The Problem

Many of the existing machine learning frameworks have been designed using the MapReduce paradigm. They are then implemented and executed over Hadoop.

This approach has problems:

- Poor for iterative computations
- No fault tolerance for the driver program
- Limited control flow

Choice of Execution Engine – A Solution

Why not change the execution engine?

Ciel allows for:

- Dynamic control flow
- No limitation on task dependencies.
- Fault tolerance on all nodes

Machine Learning over Ciel

There are two approaches which can be used for implementing machine learning algorithms over Ciel:

- Clean implementation
- Modify existing system

Two key questions need to be answered:

- Which approach leads to better performance?
- Which approach allows for easier implementation?

Clean Implementation – Canopy Clustering

I chose to implement canopy clustering using Ciel.

Research showed that the best strategy for parallelization was as follows:

- Divide the data into n segments – each which runs on a different worker.
- Run the algorithm and find the centres.
- Gather the output from the parallelized computation and re-run the algorithm using this data as the new input.
- This outputs a final list of cluster centres.

Clean Implementation – Canopy Clustering Problem 1

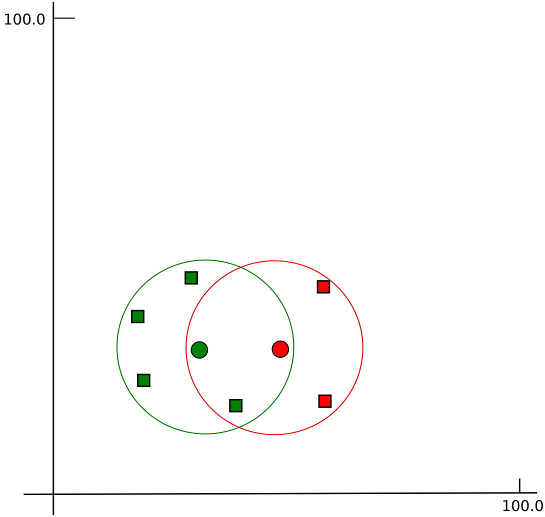

Figure: Problem with this parallelization of canopy clustering.

Clean Implementation – Canopy Clustering Problem 2

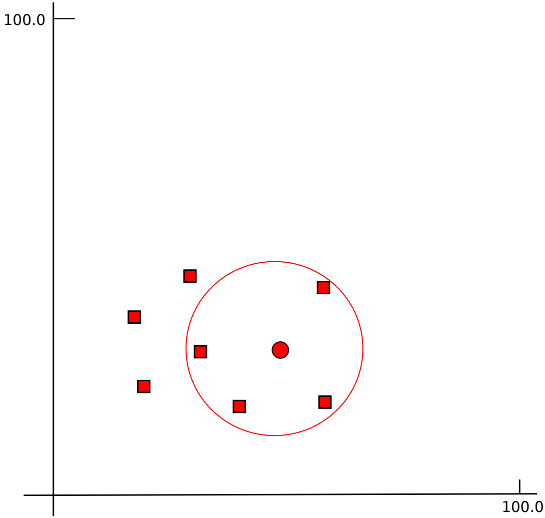

Figure: Problem with this parallelization of canopy clustering.

Clean Implementation – Canopy Clustering Graph

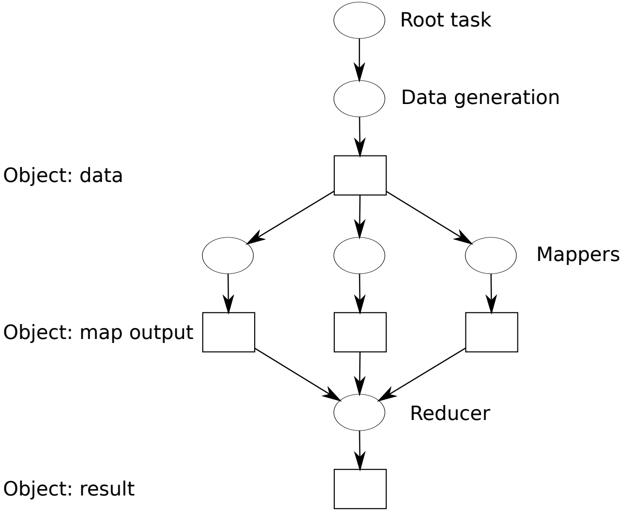

Figure: Canopy clustering dynamic task graph.

Clean Implementation – Comparison with Mahout

It is predicted that the Ciel implementation will outperform the Mahout implementation:

- More even cluster utilization
- Preferential scheduling

Modify Existing System – Ideas

An example system which could be modified is Vowpal Wabbit.

Can look for places where jobs are being sent to Hadoop and modify it to use Ciel.

This leaves the following questions:

- How much Ciel code would need to be written?
- Are jobs in a suitable format for Ciel?
- How homogeneous are Ciel and Hadoop? How easily can they be interchanged as execution engines?

Conclusion

Predictions:

- The clean implementation will have a better performance.
- The "modify existing system" approach will ultimately prove less time consuming and easier to work with when building a machine learning framework.
- In the end, it may come down to deciding what exact benefits we want out of the framework.