Streaming-KNORA:
An ensemble learning approach for distributed streaming machine learning

A dissertation submitted to The University of Manchester for the degree of Master of Science in the Faculty of Engineering and Physical Sciences

2016

MinChun Lin

School of Computer Science
Abstract

In recent years, with the advent of big data and its 4 Vs, streaming machine learning has played an increasingly important role in multiple industrial and research applications due to its exponential data generation rate and high streaming data throughput.

The instance-incremental streaming machine learning methods, such as Hoeffding Tree, can achieve similar accuracy as their equivalent batch-incremental learning approaches with better resource efficiency. However, such methods were designed for single thread execution and fail to cope with high throughput streaming data.

This work introduces a new algorithm called Streaming-KNORA, for streaming data analysis in a distributed environment, and a new multi-tier architecture based on Spark, Spark Streaming and AWS. Performance is evaluated by testing various Streaming-KNORA parameters on different datasets, measuring accuracy and throughput, and monitoring resource utilization.

The project contributions are four-fold:

1. Design and development of Streaming-KNORA, that from the literature review, appears to be the first attempt at a Streaming ML Enabled model.
2. Streaming-KNORA is able to analyse streaming data, learn concepts on disjoint streaming data, and achieve higher accuracy than a single streaming learning model on all considered datasets.
3. Streaming-KNORA analyses streaming data in a batch fashion and performs batch execution on Spark.
4. The Streaming-KNORA pipeline's throughput, which runs with a large batch size on Spark, achieves similar or higher throughput to the pipeline running on a single thread.

On all datasets, Streaming-KNORA achieves higher accuracy than single streaming learning model. It also achieves, on average, 2.61 times speedup running with 32 cores on Spark compared with running it on a single executor core. The speedup is up to 4.16 times on small batch datasets. Finally, the Streaming-KNORA pipeline's throughput is up to 1.85 times faster than the pipeline's throughput running on a single thread with datasets processed with large batch sizes.