Distributed Hierarchical Clustering for Speech Applications

Chaikesh Chouragade,Karan Malhotra,Shubham Bansal

1. Problem Statement

Clustering is often an essential first step in various speech related tasks like Speaker recognition, Hierarchical Sampling techniques for ASR, Phoneme clustering etc. Hierarchical clustering, a widely used clustering technique, can offer a richer representation by suggesting the potential group structures but its higher computation cost limits its use on large datasets.

Moreover, parallelization of such an algorithm is challenging as it exhibits inherent data dependency during the hierarchical tree construction. In this project, We would like to parallelize hierarichal clustering using two different frameworks - mapreduce and spark and compare their performances. Further, we would like to leverage these clusters to obtain a more variable dataset which would be further labelled for training an ASR. Finally, we would be comparing ASR performance on test dataset when trained with samples obtained after clustering and randomly sampled.

2. Prior Work

Due to the advance of modern computer architectures and large-scale system, a lot of efforts have been taken to parallelize. There are various implementations using different platforms, (1) was based on milti-core and (2; 3)were based on map-reduce framework. SHRINK (1), is an parallel single-linkage hierarchical clustering algorithm .

SHRINK exhibits good scaling and communication behavior, and only keeps space complexity in O(n) with n being the number of data points. The algorithm trades duplicated computation for the independence of the subproblem, and leads to good speedup.

Further, Chen Jin et al. (4) had presented DiSC, a distributed algorithm for single-linkage hierarchical clustering on Mapreduce framework. In 2015 Chen Jin et al. (5)had presented SHAS, a new parallel algorithm for single-linkage hierarchical clustering and demonstrated how the algorithm scaled using Spark in contrast with MapReduce .

The above implementations were on their own synthesized datasets. For speech data corpus Lerato (6) had done clustering of Acoustic Segments Using Multi-Stage Agglomerative Hierarchical Clustering technique.

3. Proposed Solution

Hierarchical clustering generally falls into two types: agglomerative and divisive. In the first type, each data point starts in its own singleton cluster, two closest clusters are merged at each iteration until all the data points belong to the same cluster. In our approach, we will be implementing agglomerative clustering. As a typical example of agglomerative approach, single-linkage hierarchical clustering (SHC) merges the two clusters with the shortest distance, i.e. the link between the closest data pair (one in each cluster) at each step.

Calculating the SHC dendrogram of a dataset is equivalent to finding the Minimum Spanning Tree (MST) of a complete weighted graph, where the vertices are the data points and the edge weights are the distances between any two points. Now, our effective task would be to find the solution to the following problem: ”Given a complete weighted graph G(D) induced by the distances between points in D, design a parallel algorithm to find the MST in the complete weighted graph G(D)”.

The above problem can be divided into different disjoint subproblems which can be solved and the solutions for these sub-problems would be merged to form a final solution .

4. Identified Datasets

LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech, prepared by Vassil Panayotov with the assistance of Daniel Povey. The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned.

5. Automated Metrics

Runtime: The time in seconds to implement the clustering methods both in conventional ways and proposed parallelized way.

Word Error Rate: Its the shortest edit distance between predictied and true utterance. We will be using this metric to evaluate our ASR performance.

6. Baseline

Our baseline will be work of (4), where they had presented DiSC, a distributed algorithm for single-linkage hierarchical clustering on Mapreduce framework. Although they had done on their own synthesized dataset but we will try to replicate their proposed algorithm for above mentioned speech corpus.

7. Workplan

- Replicating baseline on their synthesized dataset on mapreduce framework.
- Extending baseline mapreduce framework on mentioned corpus.
- Extending to spark framework and comparing the performances of both frameworks.
- Finally, leveraging clustering for our ASR task and reporting the results.

References

[1] W. Hendrix and et al. *Parallel hierarchical clustering on shared memory platforms* . HiPC, 2012.

[2] V. Rastogi and et al. *Connected components on mapreduce in logarithmic rounds* . CoRR, 2013.

[3] S. Wang and H. Dutta. *A parallel random-partition based hierarchical clustering algorithm for the mapreduce framework.* . CCLS-11-04, 2011.

[4] Chen Jin and et al. *DiSC: A Distributed Single-Linkage Hierarchical Clustering Algorithm using MapReduce* .

[5] Chen Jin and et al. *A Scalable Hierarchical Clustering Algorithm Using Spark* . 2015 IEEE First International Conference on Big Data Computing Service and Applications

[6] Lerato Lerato and Thomas Nieslar. *Clustering Acoustic Segments Using Multi-Stage Agglomerative Hierarchical Clustering* .