Adaptive Clustering for Big Data: Turning Big Data into Insight

Tarik Adnan Moon, Abdul Wasay, Manos Athanassoulis, Stratos Idreos
Dept. of Computer Science, Harvard University

Motivation
• Clustering widely used in Machine Learning and Computer Vision
• Hard with big dataset (NP Complete)
• Can give insight about the “structure” of very large data sets
• Efficient clustering necessary for fast unsupervised learning
• Algorithms can leverage modern hardware architecture

Methodology
• Leverage the database metadata (Histogram)
• Uniform Random Sampling
• Smart Sampling base on Coreset
• Canopy Clustering as “index” clusters
• Hardware conscious (multicore, cache conscious, NUMA and SIMD aware) design

Approaches
Histogram:
• Analyze the distribution of the data in each dimension.
• Assumes independence between dimensions

Uniform Random Sampling:
• Sample data points uniformly
• Often representative of the data
• Cluster the whole dataset initialized at the clusters centers found from clustering samples

Smart Sampling (Coreset):
• Finding coreset as the representative subset of the whole dataset
• Perform clustering on small corset and use that as a seed for the whole clustering

Hardware Consciousness:
• One pass clustering: Making sure the samples fit into L2 cache
• Reduce cache misses
• Making sure system is NUMA aware
• Leverage SIMD instructions and parallelism to speed up calculations

Results
• Histogram method performs poorly for high dimensional data
• Uniform random sampling outperforms the naïve clustering
• Uniform random sample can perform worse than naïve in some cases, on average faster

Figure 1: Kmeans clustering applied on the image for segmentation

Figure 2: Histogram based analysis for cluster center

Figure 3: Clustering on randomly sampled data

Figure 4: Time taken to cluster 10k image from MNIST dataset

Figure 5: Percentage error (w.r.t. baseline) for different number of clusters