Data Mining: Learning from Large Data Sets - Fall Semester 2015

suyifan@student.ethz.ch
zjia@student.ethz.ch
weixi@student.ethz.ch

December 2, 2015

The goal of this project was to extract the 100 most representative elements from a large image dataset. In other words, what the problem requires is solving K-means on a large dataset, but even using smart initialization techniques (like K-means++) K-means is not fast enough for large datasets. We tried several approaches for solving this problem.

Extracting Representative Elements

The first approach was *K-Means of K-Means*, or in other words, the mapper was running a k-Means algorithm, and the reducer was doing a K-Means based on the output of the mappers. The result was obviously not that much satisfying so we changed our strategy.

We then decided to implement an *online k-Means* method using adaptive coresets based on slides as in fig. 1. In the mapper, our algorithm sent the selected points in the space (the coresets) to the reducer. The next thing is to figure out how many points were necessary to sample in order to get a good coreset

$B \leftarrow \emptyset \quad D' \leftarrow D$

while $D' \neq \emptyset$

$S \leftarrow$ uniformly sample $10dk \ln(\frac{1}{\epsilon})$ points from D'

Remove $\frac{|D'|}{2}$ points nearest to S from D'

$B \leftarrow B \cup S$

Partition D into Voronoi cells D_b centered at $b \in B$

$q(x) \propto \lceil \frac{5}{|D_b|} + \frac{\mathrm{dist}(x,B)^2}{\sum_{x'} \mathrm{dist}(x',B)^2} \rceil, \quad \gamma(x) = \frac{1}{|C|q(x)}$

$C \leftarrow$ sample $10\lceil dk \log^2 n \log(\frac{1}{\delta})/\epsilon^2 \rceil$ from D via q

Figure 1: Coresets via Adaptive Sampling

(β parameter). After finding β and with a clever choice of randomly initializing the cluster centers(like K-means++) during reducer phrase we successfully beat the baseline hard with a final score of: *8.85*.

Team member contribution

- Yifan SU worked for the parameter selection and report writing;
- Xing WEI worked for online k-Means implementation.
- Zheng JIA worked for the parameter selection;