

Scaling Clustering Algorithms for Massive Data Sets using Data Streams

Silvia Nittel
SIE & NCGIA
University of Maine
Orono, ME, USA
nittel@spatial.maine.edu

Kelvin T. Leung
Computer Science
University of California
Los Angeles, CA, USA
kelvin@cs.ucla.edu

Amy Braverman
Earth & Space Science Division
Jet Propulsion Laboratory
Pasadena, CA, USA
amy@jord.jpl.nasa.gov

1 Motivation and Background

Computing clustering techniques on *massive data sets* is still not feasible nor efficient today. For instance, raw satellite imagery data can be replaced with compressed counterparts for many scientific applications. However, to facilitate scientific data analysis the high order correlation between the attributes in the data set as well as their non-parametric distribution must be preserved in the reduced data set. Therefore, practical data reduction can be achieved by partitioning the overall data set via a coarse regular spatial grid, and compressing each grid cell individually by computing multivariate histograms [1] or k-means clustering. Clustering spatial data in high dimensional spaces using k-means is expensive both with regard to computational costs and memory requirements. In a traditional k-means implementation all N data points belonging to a grid cell must be kept in memory to be clustered at a time, which often establishes a bottleneck for scientific data sets. Our objective is to define a clustering algorithm that scales automatically to any number of data points in a single grid cell, and provides high quality clustering results.

2 The Partial/Merge K-Means Algorithm

We introduce the partial/merge k-means which is based on a data stream paradigm to scale up clustering to very large numbers of data points. The first set of operators, consisting of *partial k-means operators*, copes with the memory limitation for highly populated grid cells. Instead of storing all data points of a grid cell C_s in memory, the points of C_s are divided into p partitions P_1, \dots, P_p with the condition that all data points of each partition P_j can be stored into available volatile memory (physical memory, not virtual memory). The size of each data partition can vary based on the available memory. The partial k-means operator selects a set of random k seeds for a partition P_j , and performs a k-means the partition until the convergence criteria is met. This step is repeated for several sets of

random k-seeds, and the representation with the minimal mean square error is selected as cluster representation for P_j . Since the data 'chunk' size can vary, a representative presentation of each partial clustering step with regard to the overall grid cell and other chunks is necessary. Thus, the partial k-means operator computes a set of *weighted centroids* $c_{ij} \in P_j \{(c_{1j}, w_{1j}), (c_{2j}, w_{2j}), \dots, (c_{kj}, w_{kj})\}$, whereby the weight w_{ij} is defined as the number of points that are assigned to the centroid c_{ij} in the converge step for partition P_j . The subsequent *merge k-means* operator computes a second weighted k-means algorithm using the sets of all weighted centroids of all P_j as input for the clustering process. In this step, it is necessary to assure that all centroid presentations of all partitions P_j have the same statistical chance to contribute to the overall cluster result, and do not skew the result. Thus, the merge k-means is computed collectively on all representations instead of incrementally.

3 Experimental Results

In our experimental tests, we compared the performance and clustering quality of a traditional and a data stream-based k-means implementation. We used data sets with nine dimensional data points; the number of data points per grid cell was varied between 2,500 and 75,000. For the partial step, we split grid cells into either 5 or 10 'chunks'. The experimental results show that the partial/merge k-means is highly scalable with regard to the memory bottleneck encountered when clustering large grid cells, and produces clustering presentation that are of significantly better clustering quality than generated by a traditional k-means algorithm.

References

- [1] A. Braverman, E. Fetzer, A. Eldering, S. Nittel, K. Leung, *Semi-Streaming Quantization for Remote-Sensing Data*, Journal of Computational and Graphical Statistics, Special Issue on Massive Data Streams, in print.