Streaming and parallel coreset construction for low rank matrix approximation

Wei Ma
weima@cmu.edu
October 2, 2016

• Background
Recent years have witnessed an explosion in the amount and dimension of data from various sources. Due to the huge “volume” and “velocity”\(^1\) of the data being produced, learning from these datasets requires infeasible amount of computational power. Different approaches are proposed to tackle this difficulty, pruning redundant data before learning algorithm is one of the promising methods to reduce computational time. A coreset of a dataset is defined as a sketch of the dataset which yields \((1 + \varepsilon)\) approximation to the original dataset. Surprisingly, the dimension of coreset is proved to be independent of the number of data (Feldman et al. 2013).

Another exciting feature of coreset is that it is closed with “union” operation, which suggests that coreset can be computed separately and then combined together. This feature allows us to implement streaming and parallel algorithms to calculate coreset of a dataset.

• Detail
In this project, we plan to implement the whole streaming and parallel computation framework for coreset construction described in Feldman et al. (2013). In the paper, authors only proved the feasibility of the computational framework, there’s no open source library available. So we will implement the whole framework and try to optimize the efficiency.

Then we will focus on the low rank matrix approximation using coreset construction. We will implement other streaming low rank matrix approximation methods and compare them with the coreset approach.

If time allows, we would like to discover how to conduct the matrix completion process through coreset, and discuss its advantage compared to conventional matrix completion methods.

• Expected results
After the project, we should have a well organized package for streaming and parallel computation of coreset for large data set. The package can be used for our future researches and can also be open sourced on-line.

We will also have a clear understanding about how data matrix can be approximated by low rank matrix by different methods, especially the coreset construction. We are able to wisely select different low rank approximation algorithms according to situations.

\(^1\)http://www.gartner.com/newsroom/id/1731916
The whole project is actually a preparation for analysing the time-varying traffic speed data all over U.S. in my own research. I plan to conduct the low ran approximation to find the hidden relationship between speed data and use matrix completion to perform the real-time congestion/incident detection.

- **Dataset**

  The implemented framework will first be test on the fake dataset generated by random number generators (Achlioptas & McSherry 2007). Then the Yale face database\(^2\) will be used to conduct the low rank approximation. We’ll also test it on a term-document matrix generated from Bag of Words Data Set in UCI ML database\(^3\).

- **Teammates**

  I’m actively searching for a teammate to work with me on the framework implementation and method comparison.

---

\(^2\) [http://vision.ucsd.edu/datasetsAll](http://vision.ucsd.edu/datasetsAll)

\(^3\) [http://archive.ics.uci.edu/ml/datasets/Bag+of+Words](http://archive.ics.uci.edu/ml/datasets/Bag+of+Words)
References
