Experimenting with Distributed Machine Learning Algorithms
Patrick Chase
CPSC 490

Motivation:

So far, there has been a lot of research into developing machine learning and data mining algorithms to solve a variety of problems, such as ranking, prediction, recommendation, and clustering. Many of the algorithms developed assume that all of the data will be held on one machine, however with the large data sets that people now want to study, it is clear that these algorithms must be modified to work in a distributed environment and finish in a reasonable amount of time. Changing the algorithms to work in parallel is a non-trivial problem, and there remains a lot of research and implementation to be done. Internet giants, such as Amazon have done a lot of work in this area for their own applications, like recommending products to millions of users, however there are many applications outside of large tech companies as well. For example, scientists are currently trying to use unsupervised machine learning on the human genome data in the 1000 Genome Project to uncover patterns in human DNA sequences. So far, there have been efforts to open-source packages for large-scale machine learning, such as Apache Mahout, however the number of algorithms in these packages is limited and there remains room for improvement to the existing algorithms.

Project:

For my senior project, I want to experiment with the implementation of a machine learning algorithm on a distributed system, such as Hadoop. Clustering is a standard unsupervised machine learning problem and one that is useful for analyzing a variety of large data sets. I will research the current algorithms for clustering on large data sets and design and implement a clustering algorithm on Hadoop. I will then use the algorithm to perform clustering on a large data set and see if I can gain any insight about the data from observing the clusters.
Data:

I’m still trying to figure out exactly what data I will use for my project. There are a variety of data sets that a clustering algorithm could be applied to. Here are a few potential datasets I have come across in my research and questions that a clustering algorithm could attempt to answer:

- **1000 Genomes Data**
  - Data from the DNA sequences of people
  - Questions
    - Are there DNA sequences that separate people into clusters?
      - Possibly by race or ethnicity?
- **New York Times Articles over last 100 years**
  - Questions
    - Can we separate articles from different time periods or different subjects?
- **Online Shoppers/Consumers**
  - Questions
    - Can we segment users by what they purchased?
- **Sports**
  - Questions
    - How many different “positions” are there in basketball?
      - A startup called Ayasdi did an experiment similar to this for the Miami Heat

Technologies:

- **Hadoop/MapReduce**
  - Amazon EC2
- **Java**
- **MATLAB**
  - For prototyping, modeling, and plotting
- **Spark?**
  - A framework for fast cluster computing that I might experiment with
Deliverables:

- Research current state of distributed machine learning including clustering algorithms
- Design a distributed clustering algorithm
- Implement algorithm on Hadoop
- Test algorithm with one of the data sets described above and see if we can find any patterns
- Report with description of algorithm and findings