Software-Engineered Library Development to Support a High Performance Machine Learning Visualization System

Introduction

The Northeastern Interactive Clustering Engine (NICE):
- Interactive open source data analysis tool for researchers
- GPU acceleration for quick results

Current Application:
- Study of link between environmental factors and preterm births
- Data sets generated studying a range of environmental factors

Machine Learning

- Clustering provides meaningful insight into previously unknown relationships between the data’s attributes
- Two machine learning algorithms to cluster the users’ data:
  - K-Means
    - Clustering is based on points’ distance from the cluster centers
  - Spectral Clustering/Alternative Views
    - Clusters by density; distance from point to point
    - Alternative clusterings made by dissimilarity

How It Works

- User uploads their data to the NICE server and selects model
- Using the CPU and GPU operations, the server clusters the data in the way that the user selected
- Front end uses clustering data to create a visual environment for the client

Future Work

- Further accelerating algorithms with GPU implementations
- More machine learning algorithms based on beta testing
- Implement a recommendation engine based on users’ interests in the different forms of clustering
  - May use a utility matrix that holds the views for each model
  - Data will be normalized to find similarity between the users

Acknowledgements

This research has been generously supported by the following grants:
- Grant Award ACI - 15529294 from the National Science Foundation
- Grant Award Number P42ES017198 from the National Institute of Environmental Health Sciences
- NSF IIS-1546420 BIGDATA IA: Exploring Analysis of Environment and Health Through Multiple Alternative Clustering