Title: Image Clustering of Scientific Databases

Author(s): Barnes, David C.
Canada, Curtis Vincent

Intended for: Web

Issued: 2015-10-05

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Departmentof Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Title: Image Clustering of Scientific Databases
Author: David C Barnes

The Cinema image database stores a subset of images produced from large-scale scientific simulation. For any simulation, the database contains thousands of images that scientists can use to explore the results and gain meaningful insight. The focus of my project is to cluster the images so that it is possible to see the underlying groups that are present and make the data more viewer-friendly.

It was decided to uses k-means clustering, an unsupervised machine learning (ML) algorithm to perform the image grouping. The project was started using Apache Spark [1]and Amazon EC2[2]. Spark's speed and rich machine learning library along with Amazon EC2's scalable cloud computing capability works well for parallel implementation of the task. However, a serial implementation using scikit-learn[3], python's ML library, and scikit-image[4], python's image processing library was pursued.

The training data is produced from a simulation of MPAS-Ocean. The folder of images is read and converted into a large matrix. This matrix is fed into the clustering algorithm, which returns the clusters, assignment of images and centroids. The images contain millions of features because each one consists of three channels and almost a million pixels. To reduce time and improve results, variants of k-means and preprocessing methods are applied.

Scikit-learn provides feature extraction as a form of data preprocessing method. This allows for a feature to be removed if its variance falls below some set threshold. Another preprocessing method, PCA, reduces the dimensions of the feature matrix to any number of dimensions desired. A variant of k-means clustering, known as mini batch k-means reduces computation time by randomly selecting b samples from the data to form a mini-batch. Scikit-learn also allows for the randomly chosen initial centroids to be spaced far apart from each other in order to increase the chance of a global solution.

Each preprocessing method is combined with both k-means and mini batch k-means then the results of the combinations are compared. One method used to compare the quality of the clusters is the ratio of the inter-cluster distance to the intra-cluster variance. A higher ratio indicates a better quality clustering. The mini batch k-means on unreduced data was chosen to cluster the images as it has the highest ratio.

[1] https://spark.apache.org
[2] https://aws.amazon.com/ec2/
[3] http://scikit-learn.org/stable/
[4] http://scikit-image.org

The code outputs a *.txt* file containing the clusters, images belonging to each cluster and the 3 nearest images to a centroid in a cluster. The results of the clustering algorithm are fed into the web interface for Cinema[5].

[5] created by Uzma Shaikh, Masters student at Purdue University and summer student at CCS-7, Los Alamos National Laboratory.