ENGINEERING AT ILLINOIS

parallel COMPUTING INSTITUTE

Passionate on Parallel REU 2015

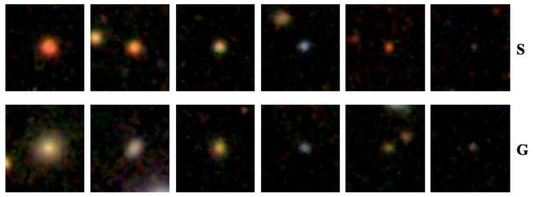
Machine Learning for Astrophysics with Apache Spark

Blake Jackson (ryan.jackson@coloradocollege.edu) | Beatrice Liang (bsliang@colby.edu) | Edward J Kim, Dr. Robert Brunner

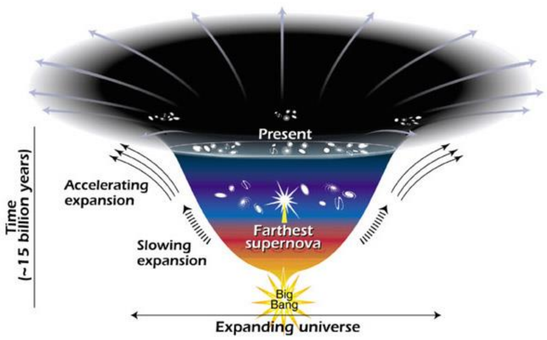
Star-Galaxy Classification

Modern astronomical surveys collect a large amount of photometric data, because it is faster to collect than spectroscopic data. Unfortunately, it is also less accurate.

Machine learning offers an opportunity to train a model on spectroscopic datasets, and then use it to label photometric data. In this case, the model differentiates between stars and galaxies.



Once the galaxies are separated from the stars, researchers can estimate their redshifts, and use these estimates to examine the properties of dark energy that cause the accelerating expansion of the universe.



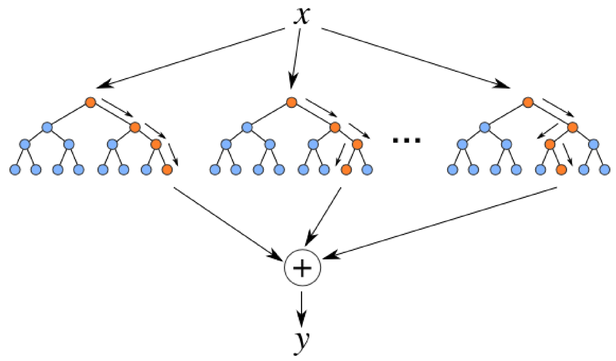
Our Goal

One of the greatest challenges in processing astronomical datasets is their size. For example, the Large Synoptic Survey Telescope is expected to produce 60 PB of raw data and 30 PB for the catalog database. Distributed computing on commodity hardware presents a highly scalable means of storing and analyzing these data.

We aim to deploy Star-Galaxy classification at scale using Apache Spark, a relatively new cluster computing framework that is typically faster than Hadoop's MapReduce. Our cluster is comprised of virtual machines on Microsoft Azure.

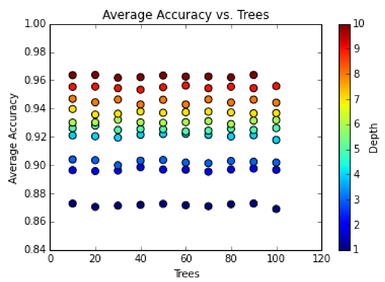
Distributed Random Forest Model

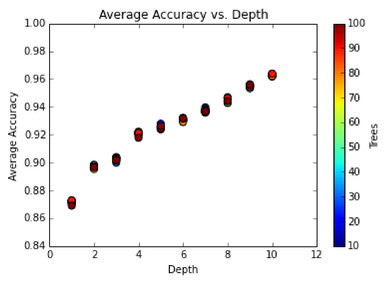
A random forest is an ensemble of decision trees that vote to generate the forest's classification. Random forests are more robust to overfitting than lone decision trees. Because the trees are trained separately, the training process is easily parallelizable. Classification takes advantage of Spark's RDD data structure to distribute the workload among all available nodes.



We trained our classifier on the colors and magnitudes of 66388 objects from the Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS). We validated the model with k fold cross-validation.

Tuning Model Parameters

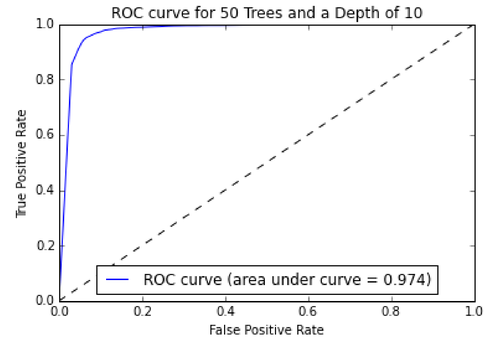


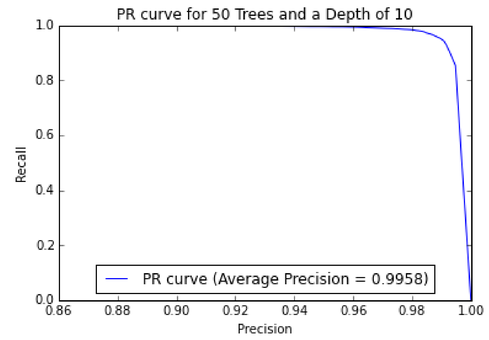


Increasing tree depth results in large performance boosts up to the number of features. However, computational time increases exponentially with depth.

Increasing the number of trees yields comparatively insignificant changes to performance. Computational time scales linearly with the number of trees.

Results





- ❑ Deploying the random forest algorithm on Spark on Hadoop is a viable method of classifying large datasets.
- ❑ This implementation is highly scalable, limited only by the size of the cluster.
- ❑ The performance of the random forest is comparable to that of other prevalent classification techniques on CFHTLenS data.

Future Work

- ❑ Classify larger datasets like the Dark Energy Survey
- ❑ Use classifications to estimate galactic redshifts
- ❑ Include Spark implementations of other machine learning algorithms (e.g. hierarchical Bayesian template fitting) in an ensemble with the Random Forest
- ❑ Extend the model to classify objects that are neither stars nor galaxies (e.g. quasars)

We would like to thank: program coordinators Jill Peckham, Dr. Craig Zilles and Dr. Matthew West; project mentors Edward J. Kim and Dr. Robert Brunner; The University of Illinois; and Microsoft Azure
This REU is co-funded by the ASSURE program of the Department of Defense in partnership with the National Science Foundation REU Site Program under Award No. 1263145.
Image References: http://imgsrc.hubblesite.org/hu/db/images/hs-2001-09-g-full_jpg.jpg
http://cdn-ak.f.st-hatena.com/images/fotolife/k/kazoo04/20131204/20131204173330.png
http://lcdm.astro.illinois.edu/static/images/SGmag.png

https://github.com/beatriceliang/POPREU

CSL: COORDINATED SCIENCE LAB