Pattern Discovery and Large-scale Data mining on cosmological datasets

Doris Jung-Lin Lee[1,2], Robert J. Brunner[3]

[1]Department of Computer Science, University of Illinois Urbana-Champaign, [2] Department of Astronomy, University of California Berkeley, [3]Department of Astronomy, University of Illinois Urbana-Champaign

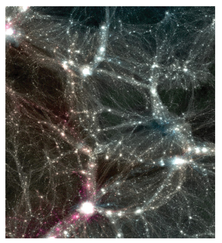
Abstract

With next-generation telescopes capturing tens of TB/night of observational data, the role of scalable and efficient data analysis methods have become central to the knowledge discovery process in astronomy. The diversity and scale of astronomical datasets also presents challenging research problems to the data mining and machine learning community. This poster describes three projects highlighting our recent work in these areas: 1) Current state-of-the-art ML algorithms (SVM, LDA, DNNs) are capable of classifying galaxy morphology at above 90% accuracy, but their results reflect the inherent errors due to human classifiers. We propose a scalable, hybrid technique that integrates active learning in crowdsourcing citizen science platforms for improving the data quality of the training labels. 2) We developed a recursive, source-finding algorithm that automatically corrects for positional inaccuracies in outdated astronomical catalogs. By applying this technique to imaging data from two different sky survey, we recovered all 23,011 sources in a widely used astronomical catalog. 3) Traditional friends-of-friends algorithms and density-estimation methods designed for halo-finding are not only computationally intensive, but especially problematic for detecting substructures within haloes. We explore non-parametric, unsupervised methods for finding haloes in the Dark Sky Simulation, a 34TB N-body simulation containing trillions of particles.

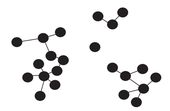
Clustering Dark-Matter Haloes in Large-scale N-body Simulations

Data: Dark Sky Simulation (Skillman et al. 2013) :

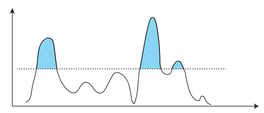
- first-publicly released trillion particle cosmological N-body simulation
- remote access 55TB of data through a custom memory-mapped interface
- result from a purely tree-based adaptive N- body code running 80million CPU hrs on Oak Ridge Titan supercomputer
- high resolution (10240^3) and large volume ($8h^{-1}Gpc$) enable comparison with future observational surveys
- provided halo catalog for grouth-truth comparison
- Science goal:
  Study the structural evolution of the universe by comparing with cosmological parameters served from observational survey.



State-of-the-art algorithms (Knebe et al. 2011) :



- searching in 3D, 4D and 6D phase space (position + velocity)
- OpenMP, MPI for parallel halo-finding
- Friends of Friends :
  - membership determined by whether two particle lie within linking length distance
  - if cluster contain less than a minimum number of particle, then not a halo
- Density-based mechanism:
  1. looks for overdense regions
  2. aggregate particles until density falls below threshold
  3. prune away gravitationally unbound particles

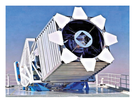


Challenges:

- Hierarchical structure formation makes it hard to find sub-haloes within haloes
- clustered result highly dependent on input parameters such as linking length or halo properties definition

Method:

- Mini-batch KMeans, Mean-shift, hierarchical clustering
- hyperparmeter tuning to minimize MSE comparison with “ground-truth” halo catalog
- divide-and-conquer strategy doesn’t work since local structures don’t capture the full cluster centroids
- Future Work: Scaling up to trillion particles (distributed ML, out-of-core methods)
- Better vizualization technique and characterization methods (mass function, halo density estimation) for comparing against halo catalogs



Source-finding for positional update in the Third Reference Catalogue of Bright Galaxies (RC3)

RC3 Catalog

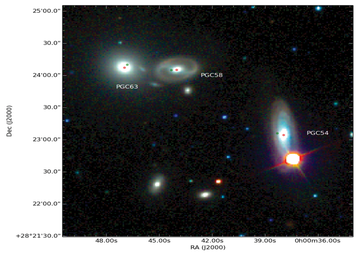
- 23011, large, bright, nearby galaxies
- Popular catalog for both morphological and cosmological studies
  - NYU VAC, NASA-Sloan Atlas, EFIGI Catalogue
- RC3 samples fairly completely and uniformly,
  $\longrightarrow$ but the catalog suffer from large positional uncertainties
- from ~1990s, including subset from RC2 [1976]

Data:

Survey	SDSS DR10	POSS-II
Bands	u,g,r,i,z	R,B,I
Coverage	35.28 % (N)	78.27%
Resolution	0.396”/pix	1.0”/pix
Technique	CCD	Photographic Plates
Data Size	11.55 TB	1.1 TB

data access through SkyServer and NASA IRSA query interface

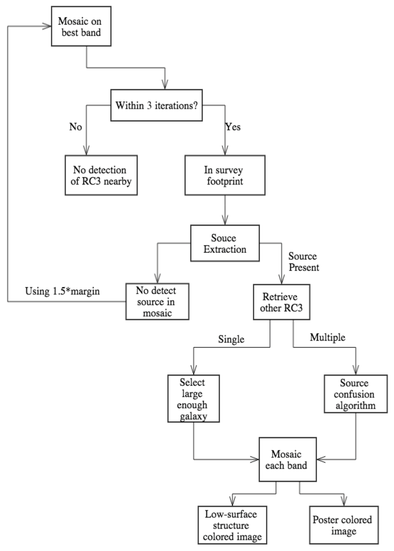
Results:



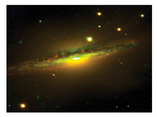
- Demonstrate both source confusion and positional update algorithm

Algorithm:

- Source-finding by recursively generating mosaic images from neighboring fields
- Generate Mosaics from Input Fields
- Updating Astrometry by finding the sources



Resolving Source Confusion



Problem :

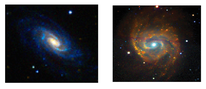
- Large galaxies often reside near other galaxies, or in clusters
- Can not rely on the existing astrometry from catalog (large uncertainties)

Data Products:

SDSS:
- 4283 RC3s improved astrometry > 1’
- 41% catalog coverage

POSS-II:
- 3431 RC3s improved astrometry > 1’
- Full catalog coverage

Web Database and Gallery:



lcdm.astro.illinois.edu/data/rc3/search.html

Potential Applications:
- Target selection
- Masking bright extended sources
- Model-fit light distribution of galaxy
- Spectroscopic fiber placement

Pipeline:

Survey-independent:

- Wavelength independent
- Increase sky coverage

Catalog-Independent

Open-Source: [ascl:1411.006]

Doris Jung-Lin Lee, & Robert J. Brunner (2015). Creating updated, scientifically-calibrated mosaic images for the RC3 catalogue, 1–11. http://arxiv.org/abs/1512.01204

Combining Human and Machine Intelligence for Classifying Galaxy Morphology

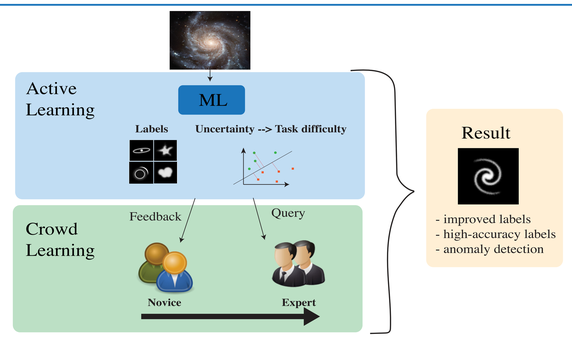
Science Goal: Population studies of galaxies and characterize the relationship between galaxy morphology and their stellar composition

Crowdsourcing labels: Galaxy Zoo

- popular online citizen science project where users help classify telescope images of galaxies according to their morphology
- grew into the popular citizen science platform Zooniverse, with over a million registered users.
- initial sample consist of 900,000 galaxies in SDSS but has since covered other surveys (HST, CANDELS, DECaLS)

Machine learning labels :

- Galaxy morphology classification has been done with SVM, DNN, PCA, random forest at accuracies comparable to the GZ classiciations.
- Problem: supervised methods are trained on GZ classifications and use GZ results for “ground-truth” comparison
  $\longrightarrow$ the algorithm reproduce these human errors
- Ferrari et al. (2015) uses LDA with the original and modified versions of the CASGM coefficients (features) to classify spirals and ellipticals at above 90% accuracy.
- This dataset provides an opportunity to integrate active learning within citizen science crowdsourcing applications.
- The distance to hyperplane from LDA can be used as an uncertainty measure for determining task difficulty in active learning .
- Provide training and feedback to novice users for low-uncertainty images and query high-uncertainty images on expert users.



Doris Jung-Lin Lee, Joanne Lo, Moonhyok Kim, Eric Paulos (2016). Crowdclass: Designing classification-based citizen science learning modules.

Acknowledgements

The RC3 work was supported by the Google Summer of Code Program. RJB would also like to acknowledge support from the National Science Foundation Grant No. AST-1313415. We thank Matthew Turk for useful inputs that helped the halo-finding project.