PINK, a self-organised map for radio astronomy

Tim Galvin, Minh Huynh (CASS, CSIRO)
and Ray Norris (WSU), Rosalind Wang (Data61, CSIRO)
The Challenge

Aus SKA Pathfinder

Square Kilometre Array

MWA
The Challenge

Aus SKA Pathfinder

Evolutionary Map of the Universe (EMU)

- Survey of whole southern sky
- 40 times more sensitive and 5 times resolution of existing all sky survey (VLA, NVSS)
- Expect to detect ~70 million radio sources (c.f. ~2 million currently known!)
The Challenge

- Radio sources can be complex
- So far, best classification and cross-matching is by human eye
- But too many sources in future

- Machine learning is an attractive solution!
  - When trained: ‘smart’, fast, never sleeps or tires
Self-Organised Maps

An unsupervised method of clustering data based on Self Organising Maps (SOM)

SOM ‘simple’ Idea:

- Start off with an empty grid of weights (neurons) equal to shape of training data
- Select a random subject from training data
- Find best matching neuron from the grid
- Reward neuron by making more similar to random subject
- Reward surrounding neurons, although not as much
- Repeat to completion, reducing area of influence throughout
PINK

- Parallelized rotation and flipping invariant Kohonen maps
  - Polsterer et al. 2015
- A SOM works great if you data remains consistent
- Radio galaxies are not:
  - Orientated differently
  - SOM would see these as different objects
    - Pixels get flattened to 1D array
- PINK brute forces problem
  - Produce all ‘realizations’ of single image
  - Compare all ‘realizations’ to map

Polsterer et al. 2015

Figure 1. Both image transformations as they are applied to measure the similarity are shown exemplarily. The flipping (left) is shown on FIRSTJ075843.0+611936 and the rotation (right) is shown on FIRSTJ072529.5+614732.
PINK

Polsterer et al. 2015

- Trained against 200,000 images from Radio Galaxy Zoo objects using FIRST data (Becker et al. 1994)
- Clustering of objects pretty obvious
- Each source in training set should have a representative neuron on the map
- Classify neurons, classify the corresponding images
• Trained against 200,000 images from Radio Galaxy Zoo objects using FIRST data (Becker et al. 1994)
• Clustering of objects pretty obvious
• Each source in training set should have a representative neuron on the map
• Classify neurons, classify the corresponding images
Radio Galaxy Zoo
Radio Galaxy Zoo Dataset Used

- Roughly 12,000 subjects
- Both FIRST radio and WISE infrared images
- Labelled features: peaks and components labels

Example classifications from Wu et al. 2018, submitted
FIRST/WISE Trained Neurons

FIRST images had sigma clipping and were logged. Both FIRST and WISE_W1 normalized to 0 – 1.
Trained SOM and distribution of labels

- Given source find closest neuron
  - PINK uses Euclidean distance for comparison

- Take labels of source and do some book keeping

- Labels correspond to number of components and number of peaks
  - ‘1_1’ = 1 component, 1 peak
  - ‘2_3’ = 2 components, 3 peaks
Assessing Predictive Power

Work in progress

- Divide data in training/validation subsets
- Train SOM, predict!
- Currently taking most likely label from most similar neuron
  - Expanding to weight based on total number of labels in set
  - Incorporate likelihood of a source’s position on map
# PINK Predictive Power: FIRST and WISE

<table>
<thead>
<tr>
<th>Pre-Processing</th>
<th>SOM</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Width</td>
<td>Height</td>
</tr>
<tr>
<td>WISE True</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>False</td>
</tr>
<tr>
<td>WISE True/False</td>
<td>3</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>True</td>
</tr>
</tbody>
</table>
Hard nut to crack

• Pre-processing very important
• Preserve/highlight important features
• PINK is time consuming

LEFT
- FIRST, no pre-processing

RIGHT
- FIRST, sigma-clipped and normalized to 0-1
Radio Galaxy Zoo with RCNN

- Chen Wu et al. 2018, submitted to MNRAS
- Trained using the same RGZ subjects
- Based on Faster Region-Based CNN (Ren et al. 2017)
- Network performs both localization and recognition
- 29 layers, in total 136,777,443 ‘trainable’ parameters

![Image of RCNN diagram](image-url)

![Table of RCNN model parameters](image-url)
Radio Galaxy Zoo with RCNN

Chen Wu, et al. 2018

**Table 5.** Evaluation of 5 data pre-processing methods using AP and mAP. Each row represents APs achieved by all five methods for a given morphology class. The highest AP for each morphology class is highlighted in the bold face. Each column denotes APs achieved by a particular method over all six morphology classes and the overall mAP. Method D4 has achieved the highest mAP, highest APs for morphology 1C_1P and 2C_2P, and second highest AP for 3C_3P.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$F$</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1C_1P</td>
<td>0.8087</td>
<td>0.8539</td>
<td>0.8242</td>
<td>0.8485</td>
<td>0.8784</td>
</tr>
<tr>
<td>1C_2P</td>
<td>0.6376</td>
<td>0.6882</td>
<td>0.6843</td>
<td>0.6746</td>
<td>0.7074</td>
</tr>
<tr>
<td>1C_3P</td>
<td>0.8250</td>
<td>0.8816</td>
<td>0.8561</td>
<td>0.8876</td>
<td>0.8941</td>
</tr>
<tr>
<td>2C_2P</td>
<td>0.7474</td>
<td>0.7014</td>
<td>0.7231</td>
<td>0.7983</td>
<td>0.8200</td>
</tr>
<tr>
<td>2C_3P</td>
<td>0.8087</td>
<td>0.7099</td>
<td>0.6989</td>
<td>0.8047</td>
<td>0.7916</td>
</tr>
<tr>
<td>3C_3P</td>
<td>0.7708</td>
<td>0.8636</td>
<td>0.8561</td>
<td>0.9424</td>
<td>0.9269</td>
</tr>
<tr>
<td>mean AP</td>
<td>78.5%</td>
<td>78.4%</td>
<td>77.4%</td>
<td>82.6%</td>
<td><strong>83.6%</strong></td>
</tr>
</tbody>
</table>
Summary

• PINK used on a cube (i.e. multiwavelength dataset) for first time
• Encouraging start, but needs work
• Need to develop better measure of predictive accuracy
  • Multiple trials of same dataset
  • Based on labels of whole dataset, not just those within neuron
• Improve predictive power
  • Refine preprocessing and gridsize/number of neurons
  • More time to train
  • Larger training set
• If predictive power improves, we use the trained neurons in other problems
  • Likelihood methods for cross cataloging, supply priors to source extraction codes (e.g. LARPY; Weston et al. 2018)
We acknowledge the Wajarri Yamatji people as the traditional owners of the Murchison Radio-astronomy Observatory site

Thank you

CSIRO Astronomy and Space Science
Minh Huynh
Senior Data Scientist / Research Astronomer

t  +61 8 6436 8696
E  Minh.Huynh@csiro.au
w  www.csiro.au