Source-Artefact Classification in Inteferometric Images using Machine Learning

Arun Aniyan, Lerato Sebokolodi, Sphesihle Makhathini & Oleg Smirnov © SKA SA & Rhodes University

Objective

Distinguishing real sources from artefacts in radio interferometric images in order to make reliable sources catalogs

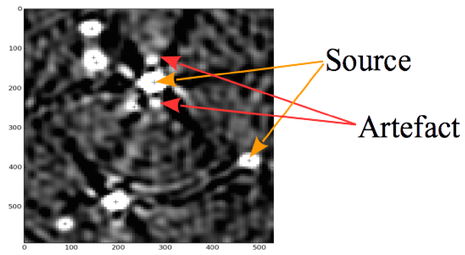

Figure 1: The artefacts around a source makes it challenging to distinguish between them.

Introduction

One of the most important challenges in radio imaging is the discrimination of real astronomical sources from spurious artefacts. At present trained radio astronomers manually identify them and rerun the calibration steps to obtain cleaner images.

We present the application of machine learning to classify real sources and calibration artefacts in radio images with high accuracy. In this study we also identify the useful features that can be used to detect radio artefacts.

The presented method will be beneficial in creating more reliable radio source catalogs.

Data set

In this study we generated simulated skies for the JVLA in C-configuration using the L-band. The sources positions and fluxes were generated randomly to follow normal and power-law distribution respectively.

Direction-dependent effects were induced via polarization beams and pointing errors. Images were generated for different periods of observations ranging from 1 hour to 25 hours in steps of 1 hour. Each set of observations contained 20 set of images in order to provide large training sets.

Ran `PyBDSM` on the resulting images to get catalogs of sources (both real source and artefacts). From known sky models and `PyBDSM` catalogs positions were cross-matched in order to classify detections into sources and artefacts.

Overall we generated a total of 29603 objects.

Results

We did the training/testing with Decision trees, K-Nearest Neighbors, Random forest and Naive Bayes.

The performance of a given classifier in terms of its accuracy was measured against the data set using a 10-fold Cross-Validation.

Classifier	Accuracy	Recall
Decision Tree Classifier	88.6749	92.0579
KNeighbors Classifier	95.9241	99.9492
Random Forest Classifier	95.2267	99.1879
Naive Bayes	82.0414	84.7550

The training parameters for each classifier were optimized using Grid search algorithm.

It can be seen from the table that KNeighbors classifier has the best test accuracy.

Feature Extraction and Analysis

Feature Extraction

A total of 28 features for each detected blob were extracted. The extracted features can be divided into the following categories.

1. Flux features
2. Axis-Angle features
3. Nearest bright source features

The flux features included descriptors like total flux and peak flux of the blobs which basically describe the brightness of each source or artefact. Along with the basic features we derived new quantities like ratios of peak and total flux, peak flux and error in peak flux, total flux and error in total flux etc.

Measures like FWHM of the major/minor axis of the source, position angle of major axis, FWHM of the deconvolved major axis of source were axis-angle related features.

We also extracted features that are related to the nearest bright source for a given object. This is from the intuitive understanding that artefacts will always be associated with a bright real source. The peak flux, distance and angle to the nearest bright object were the features extracted in this category.

Feature Analysis

Having a large feature set may not always be beneficial for better classification accuracy. Often it may be necessary to reduce the dimensionality of the data set and find features that best contribute to separation of the classes.

To find the features that are rich in discriminatory power we used *Boosting* technique. Boosting converts weak classifiers to strong ones and also gives a measure of relative importance of different features in the data set.

We found that the features related to nearest bright source and the features derived from flux descriptors where useful for more accurate classification. This also proved the fact that it is easier to identify artefacts by looking at its position with respect to a nearby bright source.

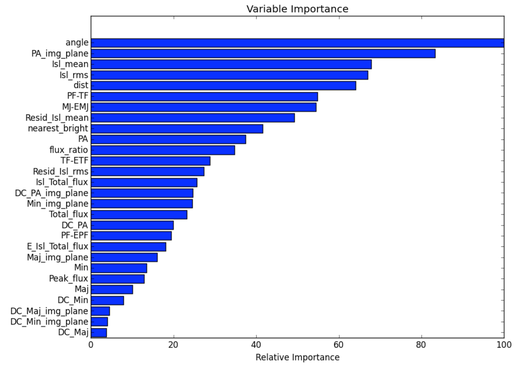

Figure 2: Feature importance by boosting

Conclusion

- In this study we have shown machine learning techniques can be used to discriminate calibration artefacts and real sources with high accuracy.
- We have identified the useful features that can be used to discriminate artefacts from real sources.

References

[1] N. Mohan and D. Rafferty. PyBDSM: Python Blob Detection and Source Measurement. Astrophysics Source Code Library, February 2015.

Future Research

Even though this method has good classification accuracy, the feature extraction process will be computationally intensive with larger number of sources and high dynamic range images. We have developed a ConvNet which directly takes images as input and does the classification with high accuracy.

Contact Information

Web www.ska.ac.za
Email arun@ska.ac.za
Phone +27 (0) 725158171