Classifying bent radio galaxies from a mixture of point-like/extended images with Machine Learning

David Bastien*, Nadeem Oozeer†‡§ and Radhakrishna Somanah*
*Physics Department, University of Mauritius, Reduit, Mauritius.
†SKA South Africa, The Park, Park Road, Pinelands, Cape Town 7405, South Africa.
email: nadeem@ska.ac.za.
‡African Institute for Mathematical Sciences, 6-8 Melrose Road, Muizenberg 7945, South Africa.
§Centre for Space Research, North-West University, Potchefstroom 2520, South Africa.

***Abstract*—The hypothesis that bent radio sources are supposed to be found in rich, massive galaxy clusters and the avalibility of huge amount of data from radio surveys have fueled our motivation to use Machine Learning (ML) to identify bent radio sources and as such use them as tracers for galaxy clusters. The shapelet analysis allowed us to decompose radio images into 256 features that could be fed into the ML algorithm. Additionally, ideas from the field of neuro-psychology helped us to consider training the machine to identify bent galaxies at different orientations. From our analysis, we found that the Random Forest algorithm was the most effective with an accuracy rate of 92% for a classification of point and extended sources as well as an accuracy of 80% for bent and unbent classification.**

I. Introduction

Galaxies can be observed at different wavelengths, each of these wavelengths reveals different morphology and characteristics of galaxies. Galaxies hosting an active galactic nuclei (AGN) are usually powerful radio sources that produce jets and extended radio emitting regions known as lobes. These sources also show various radio morphologies, from single point like source to extended ones. The lobes are huge and can range from 100pc up to few Mpc in linear extent and can be bent due to different mechanisms.

The first step that radio astronomers use to detect radio sources is to do a source finding. This will basically fit Gaussian at various local maxima and extract various fitting parameters such as the beam major and beam minor axis, and also the beam position angle. Then sources will be classified as point/un-resolved or resolved/extended depending by how much the fitted parameters to the sources deviate from the synthesized beam parameters of the telescope. In order to check whether sources are connected or not, contours are usually drawn overlay onto Optical/Infra Red (IR) images. The core of the radio source normally align with the optical/IR counterpart. With the upcoming deep surveys, the volume of data/images that will be produced will be huge. Manual classification will be a tedious if not impossible. A Machine Learning (ML) approach can be a vital tool for classification analysis.

The goal of the work presented here shows that ML techniques are efficient and robust. ML can be put to use to analyse large amount of data that will result from future telescopes like the Square Kilometre Array (SKA) and its pathfinders such as MeerKAT and ASKAP. This paper is divided as follows: §II describes Machine Learning techniques and feature selection process adopted in in this work. The results and analysis are discussed in §III and we finally conclude the paper in §IV.

II. Machine Learning application

ML algorithms can be broken down into two main categories: (i) Supervised and (ii) Unsupervised learning. Each one of them can be further broken down into classification and regression for supervised learning as well as clustering and dimensionality reduction for unsupervised learning.

In order to analyse the radio images a sample was made from various already classified radio sources. All the images were downloaded into small postage images. The first step was to carry out the point-like/extended classification and second step was the straight/bent classification. Before applying any ML algorithm, features characterising the radio sources were extracted. For example [1] made use of the parameters of the fitted Gaussians. In our case coefficients obtained from the shapelet analysis were used.

Shapelets can be defined as a linear decomposition of an object into a series of localised basis functions with different shapes, which we call Shapelets [2]. The localised basis function will be orthonormal with each other and are a set of weighted polynomials. The first 256 shapelet coefficients were extracted for each images and arrange as vectors. Each radio source is a vector in this vector space. The direction of the vector gives the shape and the orientation while its amplitude is a function of the total flux.

Once normalised only the information about the shape is stored as a vector. The comparison of the sources can be made by finding the distances between the various vectors. If two sources have the exact same shape and orientation with different peak amplitude once the shapelet vector coefficients normalised, their projected vector will be equal. This technique is useful since it enables us to compare faint objects with very bright objects with similar shapes.

Before inputting our coefficients into the ML algorithm, a visualisation is carried out in order to give us an idea of how well features are and to what extent our point and extended sources get separated in the feature space. A dimensionality

TABLE I: Results Machine Learning for point-like and extended sources classification. The confusion matrix shows the number of sources and how they were classified by the algorithms. For example, in the first row 65 sources were classified correctly as point while 15 point sources were classified as extended.

Machine Learning Techniques	AUC	Accuracy	Recall	Precision	F1-Score	Confusion Matrix		
Random Forest (RF)	0.94	0.92	0.67	0.81	0.73		**Point**	**Extended**
						Point	65	32
						Extended	15	460
***k*-Nearest Neighbour (kNN)**	0.94	0.90	0.77	0.67	0.72		**Point**	**Extended**
						Point	75	22
						Extended	37	438
Adaboost (Ada)	0.94	0.91	0.73	0.75	0.74		**Point**	**Extended**
						Point	71	26
						Extended	24	451
Naive Bayes (NB)	0.89	0.85	0.89	0.53	0.67		**Point**	**Extended**
						Point	86	11
						Extended	75	400

reduction is required since we are dealing with 256 dimensions. The Isomap [3] also referred to as Isometric Mapping was used. It is an algorithm that maps manifold on our high dimensional data and projects it on lower dimensions, it is implemented using the ***sci-kit*** [4] learn module in Python. It takes as input the set of vectors that have been described and the vectors are reduced to 3-D and further plotted in 2-D.

Our result for the isomapping (plotted on a x-y grid ranging from [-1,1]) shows that most of the data points lie in the region -0.5$\leq$y$\leq$0 and -0.5$\leq$x$\leq$+0.5. By inspecting the radio images we found that the radio galaxies in this region appears frequently in the data set and they are single Gaussian. A centre point can be drawn at x = 0 and y = -0.25 and data points away from this centre corresponds to images for which more than 1 Gaussian can be fitted. Those centred at x = 0, y = -0.25 are can be fitted with Gaussians that are circular.

III. Results and Discussions

Once convinced of the features selected from the images from [5], we input the whole 256 coefficient into the ML algorithm. Four algorithms have been used; k-Nearest Neighbour (kNN), Adaboost, Random Forest (RF) and Naïve Bayes (NB).

We make use of a stratified cross-validation techniques and with an Receiver Operating Characteristic (ROC) plot we could visualise the result of each techniques. Three of the algorithms had the same ROC Area under curve (AUC) of 0.94. In such a case other performance metrics are used as an efficiency indicator. The RF has the best accuracy and precision but NB had the best Recall and Adaboost had the best F1-Score, Table(I).

The dataset from [6], was chosen as it provides a sample of classified Bent and Unbent sources at high resolution using the Faint Images of the Radio Sky at Twenty-centimeters (FIRST) survey. The Bent-Unbent classification requires and additional rotational training as compared to the point extended classification. This is a new feature extraction process where the machine is trained with sources that have been rotated at different angles. In this new training scheme the source is rotated from 0° to 360° in steps of 10°. At each step, the source is decomposed and the shapelets coefficient are stored.

To ensure that the same source at different orientation is not used in training and testing, the following procedures are adopted to cross-validate the results:

- The number of folds n is a multiple of 369. In that case n=9 is the best fold number.
- An unstratified cross-validation is used, this ensures that the records are not scrambled and hence that the same source does not appear in different folds.

For the case of Bent-Unbent classification from extended sources, a similar approach as explained above and that used by [7] are applied. 5 metrics are used; the accuracy, the recall, the precision, the F1-score and the ROC AUC curve, to define the efficiency the algorithms. The RF was the best classifier for bent and unbent sources, with an AUC of 0.9, 80% Accuracy, 93% Recall and 86 %F1-score. The kNN had the largest precision of 84% in this case.

From our analysis, the Naive Bayes is here the black sheep of the family, the latter has an ROC curve which has an AUC less 0.5. This shows that the Naive Bayes classifier is worst that a random classifier and cannot be used in our analysis.

IV. Conclusion

In order to apply the ML algorithm, we had to extract features from the radio images using shapelet analysis that proved useful. We showed that using shapelets coefficients as features and Random Forest as a classifier yield to an accuracy of 92% in classifiying point-like from extended sources. Through the use of vectors and visualisation tools like Isomap and t-SNE a geometrical approach was brought forward to implement the classification problem. A rotational training technique inspired from a neuro-psychological phenomena known as mental rotation was implemented. The ML algorithms were then trained using sources at different orientations. We made use of a dataset of bent and unbent sources and using Random Forest we achieved an accuracy of 80% in properly classifying them.

References

[1] D. D. Proctor, "Comparing Pattern Recognition Feature Sets for Sorting Triples in the FIRST Database," *apjs*, vol. 165, pp. 95–107, Jul. 2006.

[2] A. Refregier, "Shapelets - I. A method for image analysis," *mnras*, vol. 338, pp. 35–47, Jan. 2003.

[3] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, p. 2319, 2000.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] S. van Velzen, H. Falcke, P. Schellart, N. Nierstenhöfer, and K.-H. Kampert, "Radio galaxies of the local universe. All-sky catalog, luminosity functions, and clustering," *aap*, vol. 544, p. A18, Aug. 2012.

[6] J. D. Wing and E. L. Blanton, "Galaxy Cluster Environments of Radio Sources," *aj*, vol. 141, p. 88, Mar. 2011.

[7] L. du Buisson, N. Sivanandam, B. A. Bassett, and M. Smith, "Machine learning classification of SDSS transient survey images," *mnras*, vol. 454, pp. 2026–2038, Dec. 2015.