Radio Frequency Interference Detection using Machine Learning.

Olorato Mosiane\textsuperscript{1}, Nadeem Oozeer\textsuperscript{1,2,3}, Arun Aniyan\textsuperscript{1,5} and Bruce A. Bassett\textsuperscript{2, 3, 4}

\textsuperscript{1}SKA South Africa, The Park, Park Road, Pinelands, Cape Town 7405, South Africa.
\textsuperscript{2}African Institute for Mathematical Sciences, 6-8 Melrose Road, Muizenberg 7945, South Africa.
\textsuperscript{3}Centre for Space Research, North-West University, Potchefstroom, 2520, South Africa.
\textsuperscript{4}Department of Mathematics and Applied Mathematics, University of Cape Town, 7701, Cape Town, South Africa.
\textsuperscript{5}Department of Physics and Electronics, Rhodes University, 6139, Grahamstown, South Africa

E-mail: nadeem@ska.ac.za

Abstract. Radio frequency interference (RFI) has plagued radio astronomy which potentially might be as bad or worse by the time the Square Kilometre Array (SKA) comes up. RFI can be either internal (generated by instruments) or external that originates from intentional or unintentional radio emission generated by man. With the huge amount of data that will be available with up coming radio telescopes, an automated aproach will be required to detect RFI. In this paper to try automate this process we present the result of applying machine learning techniques to cross match RFI from the Karoo Array Telescope (KAT-7) data. We found that not all the features selected to characterise RFI are always important. We further investigated 3 machine learning techniques and conclude that the Random forest classifier performs with a 98% Area Under Curve and 91% recall in detecting RFI.

1. Introduction

Any radio signal other than the desired astronomical signal is called an unwanted signal, or spurious radiation and classified as Radio Frequency Interference (RFI). \cite{1} & \cite{2} summarised RFI as having a complex nature. The Square Kilometre Array (SKA) is expected to have 100-fold sensitivity and collecting area compared to known radio telescope. Apart from the RFI pick-ups through the primary beam, the peaks of the far sidelobes will tend to be picking up more unwanted signals. The establishment of a consolidated database for the SKA South Africa observatory is very crucial for various science application. RFI data resulting from the many campaigns since 2005 to the present resides currently in an assortment of small databases and test reports. As part of this project, the integration and validation of already measured RF signals in and around the core is essential.

In this paper we investigate the application of Machine learning techniques to detect RFI from data collected using the Karoo Radio Telescope (KAT-7) \cite{3}. The paper is divided as follows: Section(2) describes the methodology adopted in pre-processing the data. Section(3) discusses the Machine Learning techniques and feature selection process together with the implication on
the RFI detection. The analysis and results are discussed in Section(4) and we finally conclude the paper in Section(5).

2. Methodology: Outlier detection

Outliers in observations are data points that do not follow the statistical distribution of the overall data under investigations. There are various ways that outliers are generated, for example, noisy system, human related errors, etc. In order to be able to make sense of the data, pre-processing is required.

The data used in this study were taken from one of the observations done by the KAT-7. The datasets have been flagged by AO Flagger software [4], which is semi-automated and validated by radio astronomers.

Communication signals are intrinsically 100% polarized and [5] describes an example how to use the polarized and unpolarized intensity to detect RFIs. Once the data have been obtained, all the Stokes parameters (I, Q, U, V) are extracted. These are taken as additional layers which represent the original data giving more dimensions from which features can be extracted. From each layer outliers are removed using the modified Z-score, which is given by the equation below:

\[ M_i = \frac{0.6745(x_i - \bar{x})}{MAD} \]  

(1)

Where the Median Absolute Deviation, \(MAD = \text{median}(||x_i - \bar{x}||)\) with \(\bar{x}\) as the median and \(x_i\) are the data points.

Using the modified Z-score anything larger than 3.5 is considered as outlier [6]. The outlier removed data are used to normalize the layer. The normalized layer is then unfolded in such a way that all frequency time series are taken and stitched next to each other to form a larger 1-D pseudo-time series. A similar approach is taken for the spectrum. That is for each time a spectrum is taken and all spectrum are then stitched next to each other to form a larger “spectrum”. With this new perspective of the data one can now visually see periodic RFI as well as wideband RFI features. The normalization of the modified Z-score filtered, layers allows the low level RFI to shoot up.

Downgrading or smoothing the astronomical signal is very sensitive and therefore, filtering seem more appropriate to use in our case. However, filtering always goes at the expense of changing the data structure. We therefore apply the filtering process after the outlier detection procedure. Once the data have been filtered using a Gaussian, statistical analysis is applied to extract features that are fed into a machine learning algorithm.

3. Machine Learning

With the advent of new advanced technologies and the increasingly innovative capacity to gather a massive amount and variety of information, the world is faced with a deluge of data. This is already happening in the radio astronomy community. The MeerKAT data size is expected to be \(\sim 50\) Terabyte for an 1 hour observing run. However, post-processing conventional RFI detection and excision need to be automated. The algorithm presented here serve as a test case to optimize the flagging procedures already in place.

3.1. Data

The data used in this study were taken from one of the observations were done by the Karoo Array Telescope (KAT-7)[3]. The KAT-7 is an array of 7, 12m diameter radio telescopes operating between 1.2 GHz-1.9 GHz. It is a precursor technology to the upcoming MeerKAT. The KAT-7 datasets we investigated were observed at a sampling rate of 0.5 seconds with 644 and 1024 channels. The number of baselines was 15 and 21 respectively. The datasets have been flagged by AO Flagger software [4], which is semi-automated and validated by radio astronomers.
3.2. Feature Extraction
Machine learning algorithms by default are not designed to take raw data in the form of images or signals as learning inputs. Therefore, the data has to be transformed into a different form, which has representative signatures of the input raw data. Feature extraction is the process of transforming the raw data into a different form such that it holds signature properties of the input class.

In this work we took 8 second segment sliding window of the signals and extracted different statistical features. The features are extracted from the spectral and time series information and are as follows: The mean of the entire channel and the kurtosis, skewness, the maximum of the cumulative sum (cumsumx), the variance (var), the percentile 25 and 75 (per25 and per75) and lastly the sum of the 8 second segment (sumvalues).

3.3. Machine Learning Techniques
At first three machine learning algorithms are used to classify the data. The Three algorithms are K-Nearest Neighbour (k-NN), Random Forest Classifier (RFC), Naive Bayesian (NB). In our experiment a random 70/30 % split of the data is done first. That is 30% of the data were withheld for testing the classifier and Cross-validation was performed with 70% of the data to train the algorithm. This split is essential in the true test of comparing classifiers so that the test is not biased [7].

3.4. Receiver Operating Characteristic Curve and Area Under the Curve
The Receiver Operating Characteristic Curve (ROC) is a tool used to evaluate a classifier [8]. The ROC curve is a graphical representation of the true positive rate versus the false positive rate of a binary classifier as the threshold varies between [0,1] in 0.01 increments as in [9].

The ROC curve is a tool used to study the trade-off between the false positives and the false negatives. ROC curve is a plot of True Positive Rate versus the False Positive Rate at different threshold values [9]. From the ROC curve by taking the area under the curve (AUC) one will obtain the probability that a classifier will correctly classify a randomly chosen instance [7].

While investigating all the above ML techniques, we realized that not all the features might be useful and need to be investigated [10]. There are 3 general classes of features selection algorithms: filter methods, wrapper methods and embedded methods. In this paper we will focus mainly on the wrapper methods. These types of methods take the feature importance measure obtained by a classification method and they come at a computational expense, but the result is a view of the importance of each feature in the feature space.

4. Results and Discussions
Our data contain much more non-RFI data than RFI data. This may cause a bias towards the detecting false positives. The training set was hence split up randomly 15 times in order to take into account the above mentioned bias.

The best performing classifier is the RFC and this is true for all the other splits. Since the RFC performs well in all the data splits it suggests that there is a feature or set of features that ranks higher than the other features. Because the RFC is basically a set of small decision trees it means the there is a tree or set of trees that are given more weighting to them than the others, this allows one to calculate which features are important and the features’ importance. The Features that are highly ranked are those that have been extracted along each time series. This suggests that time domain information seems to be more important than spectral information in detecting RFI.
The Top 5 features were taken and the procedure of splitting and evaluation were repeated. The Best performance split is finally summarized in Table 1.

Table 1. Best performing scores of the best data split with top 5 features.

| ML Tech | Type(Data/RFI) | Scores          |
|---------|----------------|-----------------|
|         |                | AUC  | precision | Recall | F1-score |
| NB      | Data           | 0.80 | 0.99      | 0.95   | 0.97     |
| k-NN    | RFI            | 0.07 | 0.32      | 0.12   |          |
| RFC     | Data           | 0.86 | 1.00      | 1.00   | 1.00     |
| RFC     | RFI            | 0.84 | 0.54      | 0.66   |          |
| RFC     | Data           | 0.98 | 1.00      | 1.00   | 1.00     |
| RFC     | RFI            | 0.96 | 0.91      | 0.93   |          |

From Table 1 the RFC and k-NN have an overall improvement whereas the NB has has performed slightly worse but with these 5 features the best performing machine learning technique is the Random Forest Classifier.

5. Conclusion
With the deluge of astronomical data coming from new telescopes such as MeerKAT, conventional RFI detection and excision will need to be re-evaluated. We have shown that machine learning techniques can be applied to detect RFI. Pre-processing and understanding the data is a very time consuming process. However, a lot of care should be taken in pre-processing the data. From our preliminary analysis we have shown that the random forest classifier seems to produce positive results. Further investigation needs to be carried out in order to get a more robust algorithm by using a bigger sample of flagged and un-flagged raw visibility data.

Acknowledgment
The authors would like to acknowledge assistance, support and encouragement from the SKA Africa colleagues and friends who have contributed to this work.

References
[1] Pandey V and Shankar N U 2005 URSI General Assembly Proceedings URL http://www.ursi.org/Proceedings/ProcGA05/pdf/J03b.5(01214).pdf
[2] Fridman P and Baan W 2001 Astronomy & Astrophysics 378 327–344
[3] Foley A, Alberts T, Armstrong R, Barta A, Bauermeister E, Bester H, Blose S, Booth R, Botha D, Buchner S et al. 2016 Monthly Notices of the Royal Astronomical Society 460 1664–1679
[4] Offringa A 2010 Astrophysics Source Code Library 1 10017
[5] Deshpande A A 2005 Radio science 40
[6] Seo S 2006 A review and comparison of methods for detecting outliers in univariate data sets Ph.D. thesis University of Pittsburgh
[7] Buisson L d, Sivanandam N, Bassett B and Smith M 2014
[8] Kim E J, Brunner R J and Kind M C 2015
[9] McCauliff S D, Jenkins J M, Catanzarite J, Burke C J, Coughlin J L, Twicken J D, Tenenbaum P, Seader S, Li J and Cote M 2015 The Astrophysical Journal 806 6
[10] Guyon I and Elisseeff A 2003 The Journal of Machine Learning Research 3 1157–1182