Machine Learning Techniques in Radio-Frequency Interference Classification

MULAN XIA | BERKELEY CS 2019 | MENTORS: STEVE CROFT, PRAGAASH PONNUSAMY | BERKELEY SETI RESEARCH CENTER

Abstract

The **Green Bank Radio Telescope** (GBT) is currently being used in one of the world's deepest searches for artificial signals. Using an ABACAD pattern of viewing ON and OFF targets, the telescope searches for potential signals of interest (SOI) by detection of signals which appear repeatedly in ON sources, but are absent in their OFF counterparts. A signal present in both would likely be radio frequency interference (RFI).

This project explores the potential of machine learning in developing a fast and accurate classifier to identify, extract, and classify signals of interest. This involved filtering through data, classifying obvious RFI candidates into 1 of 9 categories, and reducing the dataset to a more manageable size.

Several machine learning techniques, in conjunction with varying feature sets, were evaluated:

- Principle component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) with unlabelled data
- Support vector machines (SVM), logistic regression, linear distriminant analysis (LDA) and quadratic discriminant analysis (QDA) with labelled data

Methods

GBT data was first run through a pipeline which detected hits above a certain threshold. It was then run through three variations of the histogram of oriented gradients (HOG). These comprised the feature vectors on which classifiers were run. Additionally, several correlation coefficients between ON and OFF pairs were computed and used in plotting the PCA projection of unlabelled data.

fig 1. Original data

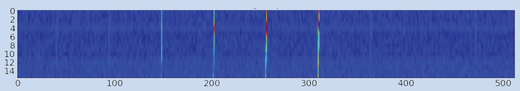


fig 2. HOG data

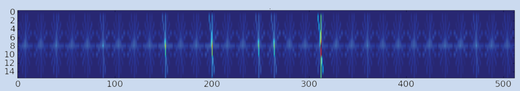


fig 3. Log normalized data

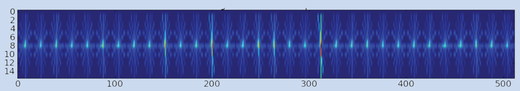
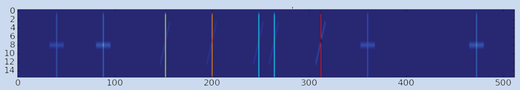


fig 4. Binarized HOG data



Data Analysis

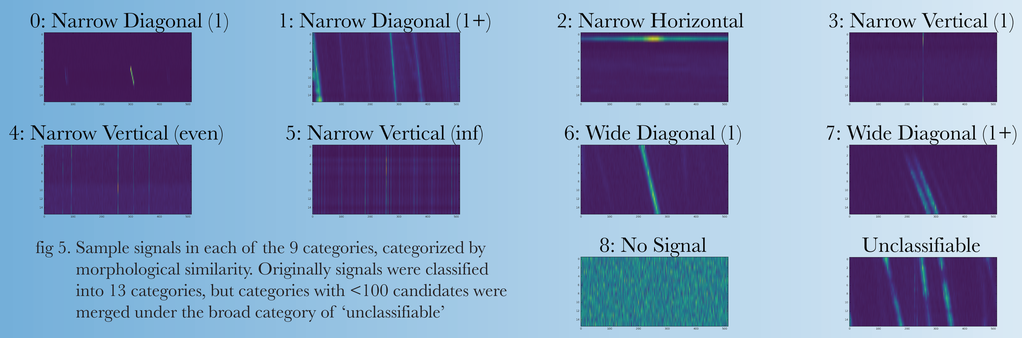


fig 5. Sample signals in each of the 9 categories, categorized by morphological similarity. Originally signals were classified into 13 categories, but categories with <100 candidates were merged under the broad category of 'unclassifiable'

Accuracy of Different Classifiers

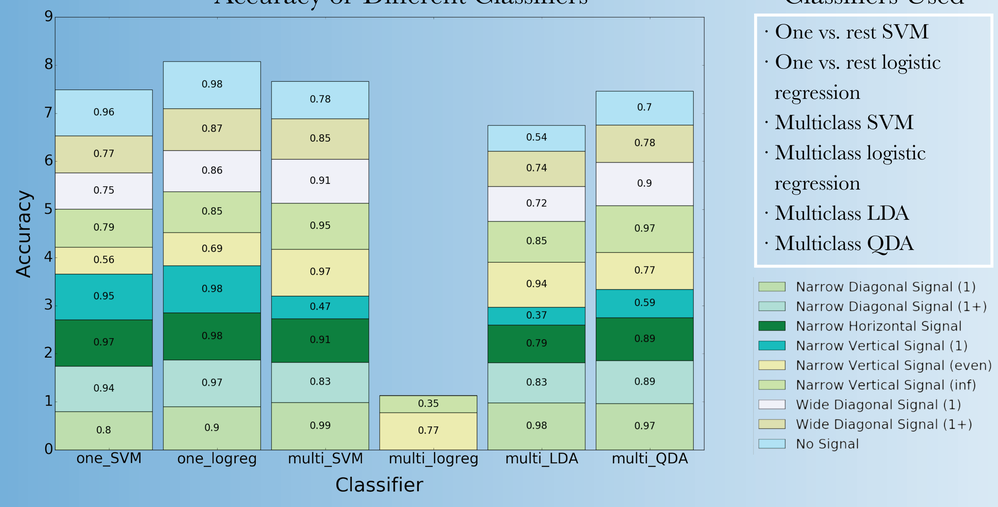


fig 6. Stacked bar graph of classification accuracies for individual signal categories

Classifiers Used

- One vs. rest SVM
- One vs. rest logistic regression
- Multiclass SVM
- Multiclass logistic regression
- Multiclass LDA
- Multiclass QDA

Multiclass SVM Confusion Matrix

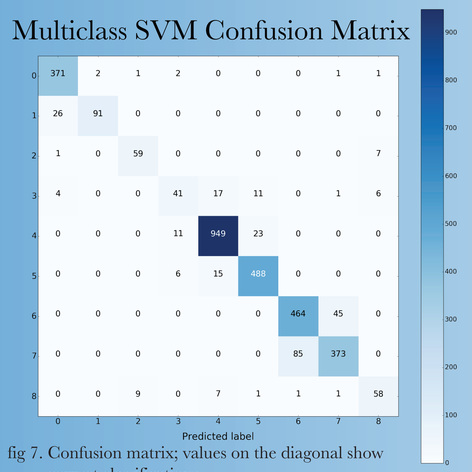


fig 7. Confusion matrix; values on the diagonal show correct classifications

PCA and t-SNE Projection of ON/OFF Pairs

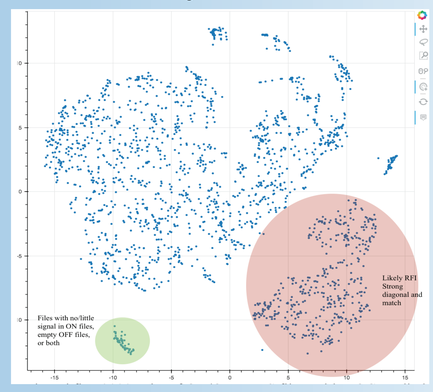


fig 8. PCA projection of HOG processed ON/OFF pairs, clustered by correlation coefficients

Conclusion

Classifier	Accuracy
One vs. rest SVM	0.97
One vs. rest logistic regression	**0.98**
Multiclass SVM	**0.92**
Multiclass logistic regression	0.31
Multiclass LDA	0.84
Multiclass QDA	0.85

fig 10. Overall accuracies of each classifier, averaged across signal categories.

It can be confidently said that well over 99% of GBT data consists not of extraterrestrial signals, but RFI. And while it could be debated whether we should focus on:

a) finding signals which differ significantly from RFI or;

b) searching for RFI–like signals with unexpected features (occurs at an unusual frequency),

this project demonstrates that it is in fact possible to classify certain classes of RFI based solely on morphological characteristics. Thus, machine learning can be well applied to both approaches.

In the first approach, chained classifiers can filter the database of known classes of RFI.

In the second approach, we can first gather a class of classifier-determined RFI, and then apply anomaly detection to extract outliers, or potential SOIs which may otherwise look very similar to RFI to the naked human eye.

Future Work

It should be recognized that the current classifiers are limited in that they do not find outliers (potential SOIs). Each signal must be classified into a predetermined bin. A solution to circumvent this problem would be to use a chain of one class classifiers: remaining, unclassified samples at the end of the chain would ideally comprise a class of SOI. Yet all attempts to implement this had limited success (acc ~0.62) as accuracy significantly drops as the training set size drops. **More labelled data** would be required to make this a viable method.

Additionally, while PCA and t-SNE seem to show that ON and OFF pairs can be grouped based on similarity, further investigation into **better defined features** is necessary, for more accurate groupings.