Preliminary Design Review

SSE 498

Classification of Radio Astronomy Observations through Machine Learning

Kenneth Carroll

11-28-18

**Table of Contents**

* Introduction .………………………………………………………………………… 3
* Significant Issues ……………………………………………………………………. 3
* Work Breakdown ……………………………………………………………………. 4
* Timeline ……………………………………………………………………………... 4
* Costs Estimates ……………………………………………………………………… 5
* Current Status ………………………………………………………………………... 5
* Risk Areas …………………………………………………………………………… 5

**Introduction**

Machine learning is a rapidly evolving technology that revolves around analyzing data. Algorithms used in machine learning use large quantities of data to calculate a prediction that then gets compared with an expected value. Based on the accuracy of the prediction, the algorithms change the weight of each input to closer match the expected value over time. The more data the machine is trained on, the more accurate the prediction. By applying this principle to data collected over radio observations, the data can be rapidly processed on an optimized algorithm to assist with the classification along with assisting with the isolation of anomalous data. This can serve as a benefit for new data but can also assist with the review of existing data. Minor inconsistencies in data that may be missed or decided as negligible could have a correlation with other minor anomalies.

Machine learning for the classification of astronomical observations was chosen since there are not many applications being developed under machine learning for aerospace applications. As a proof of concept for the utilization of machine learning, a binary classifier for the detection of pulsars will be created. This will take a series of pulsar and non-pulsar data sets and make a prediction. Success will be determined by an accuracy rate equal or greater than 98%. A binary classifier was chosen since it is much easier to implement and serves as a prerequisite to a categorical classifier using similar radio astronomy data. At the required accuracy, results differing from an expected value can point to either an anomaly or a point of further investigation.

**Significant Issues**

Since the majority of pulsar research within supernova remnants has been based off of observations of the Crab Pulsar within SN 1054, some radio astronomers are interested in knowing if the criteria based off of this pulsar are typical or anomalous. Once approach that can be used is to train a form of machine learning, called a neural network, to classify radio observations as either pulsars/pulsar candidates or non-pulsars/non-pulsar candidates. Since a well-trained algorithm can recognize connections which may not be visible to a human and can infer imperially from data, there is the possibility that a good classifier could take inconclusive or emerging data and accurately classify pulsars. This could save time in the identification portion of pulsar research, freeing up more time to study newly found pulsars. Some issues with this approach are the accuracy of the neural network model being used. If the model does not achieve significantly high accuracy, the potential for false results becomes higher. More so, a model needs to minimize the number of false negatives that it can generate since a false positive can be corrected with manual inspection, but a false negative could result in losses of large amount of research data. In addition to that, if the model is set to a specific set of data, a large amount of preprocessing on the data may be required to make observation compatible. If done improperly, the fidelity of the data is compromised, and the accuracy of the model may not match with the expected accuracy since the usage of a neural network assumes that the data does not contain significant errors. This leads into the goals of this project where high accuracy with high fidelity data is paramount. The minimum accuracy required for this project would 98% with an accuracy approaching 99% being a long-term goal and accuracy > 99.5% being ideal albeit unrealistic.

**Work Breakdown Structure**

The work for this project can be thought of in several portions: general research, data collection, data formatting, model design, model development, model training, model testing, and result analysis. First and foremost is research, where knowledge in machine learning and knowledge in radio astronomy, particularly in pulsars, is required. To bridge a lack of understanding in some areas of radio astronomy and make connections with the finer technicalities of machine learning, a good portion of study and research in both areas is required for the success of this project. Next is the data collection, since large volumes of data are required for the proper training of a model. Large being a relative term, it’s important to understand that 20,000 data points is relatively small for machine learning. But having a large amount of data is not sufficient, as formatting between research organizations is not standardized in pulsar research. This requires data to be formatted or preprocessed into a manner that makes all data inputs universal to the model. Since the range of formats that the data can be in varies vastly, this proves to be both a mathematical challenge and a software engineering challenge as well. Once the format of the data is understood, the model design can be thought out, where the structure of the hidden layers of the neural network is the majority of the design; input and output layers are very straightforward. The same can be said for the model development, as modern frameworks allow a model to be developed from a design in less than 100 lines of code. The actual software to train the model is grouped within the model development code but setting up how the model processes data and configuring paths to the data is the bulk of the effort for training; it takes more time to set up GPU training but also yields faster results in the long run. Once setup, model training is an unsupervised, autonomous process. Once trained, testing the model takes a separate portion of code so that the data can be exported in a usable fashion. This data can then be analyzed based off of personal preference which can take days or weeks based off of both the amount of data and the depth of analysis.

**Timeline**

September: Project Research and Development

October: Initial Data Collection and Formatting

November: Model Design, Development, Testing, Training, Analysis, PDR Development, Begin Final Paper

December: Additional Research into Machine Learning and Radio Astronomy, Randomization Algorithm Development, Secondary Model Planning

January: Secondary Model Design, Development, Testing, Training, Analysis

February: Tertiary Model Design, Development, Testing, Training, Analysis

March: Results Analysis, Documentation, CDR Presentation Development, CDR Corrections, Begin Final Presentation

April: CDR, Final Paper, Final Presentation

**Cost Estimates**

****

**Current Status**

By utilizing the HTRU2 Dataset and a mass model generation algorithm, a peak accuracy of 98.2% has been achieved on the first generation of the pulsar classification model. Additional research has now been targeted into more datasets, data conversion between varying formats, and the potential of using the model on existing but inconclusive data. Plans for a better model generation algorithm with layer and neuron randomization has been planned and will be developed in the next month. This algorithm will be used for the secondary model development, and the results of this model set will lead to tweaking of the algorithm before the tertiary model development. Collaboration with Dr. Pannuti is leading to potential datasets from his colleagues and looking into varying types of input for newer models.

**Risk Areas**

The primary risk area for this project is the availability of data. Each data has a limit to how accurately it can be classified, and the introduction of more data can increase this limit. However, the availability of data is proving to be a real challenge and is dependent on third party sources. To counteract this, development with the current dataset will continue to push the accuracy as high as possible. Another potential risk is the probability of false negatives which are significantly more important that false positives. The need to reduce the probability of a false negative going further is going to prove a technical challenge since there is currently a 50% change that in the event of an error it will be a false negative. Since the project is primarily research based, there are no other risk areas that are currently visible.