Using Machine Learning to Quantify the Accuracy of High-Redshift Galaxy Classification Algorithms in the Extended Groth Strip Field

Sarah Bruce

The University of Texas at Austin

May 2022

Abstract

While it is possible to select high-redshift galaxies in a purely automated way, visual inspection by human moderators is usually required to remove contaminants. As the process of automating high-redshift galaxy selection routines for target validation is perfected, we must determine the accuracy of computer-generated classifications using statistical analysis. We visually classified sources in EGS, studying SED's, imaging, IRAC residual mapping, and photometric redshift plots. Human classification is a conglomeration of qualitative elements observed per source; our objective is to define these quantitatively, in order to feed parameters into machine learning algorithms to accurately classify. We focus on the quality of IRAC residual maps; poorly modeled IRAC fluxes result in inaccurate classifications. We have assigned values to the relationship between IRAC residual maps and the likelihood that sources will be falsely identified as high-redshift galaxies and created a supervised machine learning algorithm to predict IRAC classifications with 96.8% accuracy.

1 Introduction

In order to study the early universe, it is necessary to identify high-redshift distant galaxies. Using data from Spitzer/IRAC and Hubble, classification algorithms can automatically give observers a set of suggested high-redshift galaxies; however, spurious sources and other imperfections in data cause misclassifications, leading to the necessity of human oversight. In preparation for the James Webb Space Telescope, it is useful to improve efficiency and accuracy of these classification algorithms.

2 Statistical Analysis

Data collected for this project consists of a set of manually-classified sources accumulated through a collaborative Zooniverse classification program. After classifying, the data must be isolated from the EGS field residual map, sorted into good or bad IRAC, and grouped by certain characteristics depending on the analysis needed.

The next step in the process is to search for patterns in data that could be a common differentiator for good vs bad IRAC data. Using MatLab programming, we observed plots of good vs bad data in terms of individual statistical values, including:

- Variance
- Standard Deviation
- Correlation Coefficient
- Average Brightness

- Maximum and Minimum Brightness

To ensure minimal accuracy lost due to small data values, we used a linear mapping technique to map values to a larger scale.

After observing the statistical data, we determined that while most statistical values are consistent among both good and bad IRAC residuals, there was a trend of low standard deviation and variance among sources determined to have bad IRAC data. We are continuing to analyze for a cause of this trend. As our goal was to determine a consistent delineation line to separate good vs. bad IRAC sources, we initially consider discarding all sources with a certain threshold of low variance or standard deviation. However, choosing to use this as our deciding factor would result in a statistically significant loss in good sources; since our plots show that many good IRAC sources also display low variance and standard deviation, choosing these trends as our delineation line would cause our algorithm to falsely identify a multitude of good sources as unusable, thus we must choose a different technique.

3 Machine Learning

Perceptron is a type of supervised machine learning algorithm made of single-layer binary linear classifiers. Perceptron allows optimal weight coefficients to be automatically learned and outputs results using a step function between two linearly separable classes. As a simple single-node neural network, Perceptron uses the weighted sum of inputs and a bias to create an activation, as well as accepting one data row that will be used to predict a class. In this project, the simple data set used was a training set of sources that had already been classified by hand. Each source was labeled as good or bad IRAC, allowing for Perceptron's binary decision-making.

Using this algorithm and a training set of 2,602 pre-classified sources, we were able to train the algorithm to classify a given IRAC residual as good or bad to an accuracy of 96.8% in 25 epochs. This can be used to identify bad IRAC sources without human supervision. In observing a location plot of all classified sources through the algorithm, we are able to confirm that there is almost no correlation between location and IRAC quality.

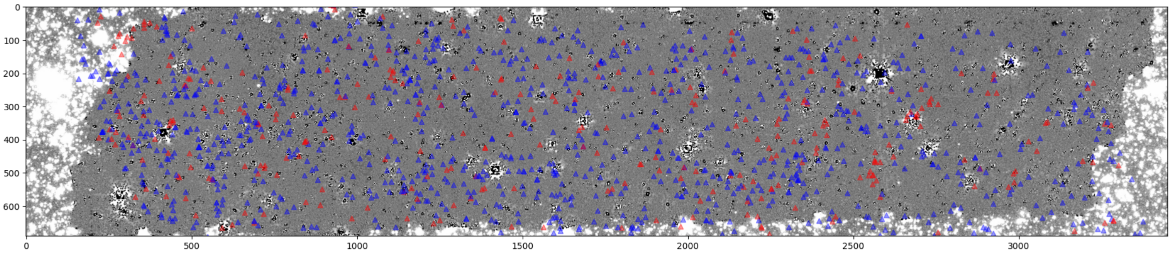

Figure 1: Location plot of all sources classified by the algorithm (good IRAC in blue, bad in red).

While a 96.8% accuracy in 25 epochs is a good start to consistent classification, some work can be done to optimize the algorithm's accuracy. For this project, 25 epochs was an estimation of the ideal amount of training sessions, meant to avoid over-training. If a machine learning algorithm is provided too many training sessions, the algorithm begins to "memorize" test data, rather than learn to separate accurately; this is called overfitting, and appears as extremely high accuracy on the test data (i.e. 99%+), but ultimately low accuracy once it is asked to classify new data. On the other hand, using too few epochs results in underfitting, meaning low classification accuracy altogether. In future work on this project, the number of epochs should be optimized for correct fitting. This can be accomplished by testing for the number of training sessions that has equal accuracy between both the original training set and other pre-classified data sets.

By using the Perceptron algorithm to identify sources with unusable or untrustworthy data from Spitzer/IRAC, the efficiency of galaxy selection can be improved. Sources with bad data can be thrown out of the identification process, creating a smaller and more trustworthy data set for researchers to work with when identifying high-redshift galaxies.

4 Conclusion

In anticipation of the James Webb Space Telescope's soon-to-be-available advanced imaging, this process can be perfected and studied in order to limit algorithm selection errors and improve efficiency in high-redshift galaxy detection. It may be useful to further study an unsupervised clustering machine learning algorithm rather than supervised in an attempt to observe more delineation factors in data, as grouped by the algorithm.