Separating Stars from Quasars: Machine Learning Investigation Using Photometric Data

Simran Makhija[a,*], Snehanshu Saha[b], Suryoday Basak[c], Mousumi Das[d]

[a] *PES University South Campus, Hosur Road, Electronic City, Bangalore, Karnataka, India - 560100*
[b] *CAMS and CPR, PES University, Outer Ring Road, Bangalore, Karnataka, India - 560085*
[c] *University of Texas Arlington, Texas, USA - 76019*
[d] *Indian Institute of Astrophysics, Bangalore, Karnataka, India - 560034*

Abstract

A problem that lends itself to the application of machine learning is classifying matched sources in the Galex (*Galaxy Evolution Explorer*) and SDSS (Sloan Digital Sky Survey) catalogs into stars and quasars based on color-color plots. The problem is daunting because stars and quasars are still inextricably mixed elsewhere in the color-color plots and no clear linear/non-linear boundary separates the two entities. Diversity and volume of samples add to the complexity of the problem. We explore the efficacy of neural network based classification techniques in discriminating between stars and quasars using GALEX and SDSS photometric data. Both sources have compact optical morphology but are very different in nature and are at very different distances. We have used those objects with associated spectroscopic information as our training-set and built neural network and ensemble classifiers that appropriately classify photometric samples without associated spectroscopic labels. Catalogs comprising of samples labelled using our classifiers can be further used in studies of photometric sources. The design of a novel Generative Adversarial Network (GAN) based classifier is proposed in the paper to tackle the classification problem. To evaluate the correctness of the classifiers, we report the accuracy and other performance metrics and find reasonably satisfactory range of 91-100 % .

Keywords: Generative Adversarial Networks (GANs), Random Forests, SDSS Catalog, Virtual observatory tools

1. Introduction

One of the major challenges of of large scale photometric surveys is the separation of the different classes of sources, especially stars and quasars. Both types of sources have a compact optical morphology and are hence difficult to separate without spectroscopic data (Fan, 1999). In such cases, other parameters of the sources such as their optical variability or their optical colors (Richards et al., 2002) are necessary to distinguish between stars and quasars. Later studies have shown that including the infrared data or UV data with optical photometry results in a more efficient separation (Wu et al., 2012; Bianchi et al., 2005). There are also several template based optical classification schemes that use the optical fluxes of sources to distinguish between quasars and stars (Bovy et al.,

*Corresponding author
Email address: snehanshusaha@pes.edu (Snehanshu Saha)
Center for AstroInformatics, Modeling and Simulation
Center for Pattern Recognition

2011). In such methods, including a larger wavelength range such as near-infrared (DiPompeo et al., 2015) or UV (Jimenez et al., 2009) have been found to increase the efficiency.

In this paper we present two machine learning classification approaches to distinguish between stars and quasars using only optical photometric data and UV data, observed using SDSS (Gunn et al., 1998) and GALEX (Martin et al., 2005; Morrissey et al., 2007) respectively. In the present era of large telescopes and optical surveys, machine learning is already proving to be an important technique for characterizing sources in huge data sets (Morice-Atkinson et al., 2018; Das and Sanders, 2019). Machine learning involves the use of statistics to make useful predictions and to learn essential features; *classification* is the process of marking separations between categories in data. Supervised machine learning techniques make use of labeled data to make predictions of future unseen data. In the present context, the labels in the data comprise of numeric indications of a source being a star or a quasar.

The problem of separating stars from quasars using machine learning has not been studied in much depth in the literature (Krakowski et al., 2016; Bai et al., 2019), but other methods based on template classification have been done earlier (Preethi et al., 2014). The best method is to include spectroscopic data so that the red-shift of quasars (derived from their optical spectra), can be used to differentiate between stars within our Galaxy and quasars (Viquar et al., 2018). However, this becomes difficult when spectroscopic data is unavailable. Spectroscopic surveys also take an enormous amount of observational time. Another application of using machine learning techniques for separating compact sources like stars and quasars is in identifying gravitationally lensed quasars which are rare (Khramtsov et al., 2019)

Photometric surveys use template based methods to separate different classes of astronomical sources but there are difficulties in using the template based methods, such as lack of adequate data for determining spectral energy distributions (SEDs), large error bars or overlapping SED properties of the different classes. As a result the template based classification methods do do not always provide clear results and hence, it is important to explore other methods such as machine learning algorithms for such problems.

1.1. A Gentle Introduction to Classification Methods

The number of quasars in our sample is much larger than the number of stars as collected by the technique mentioned above. This imbalance between the two classes results in a poor automation of the task and it demands the need to up-sample the star observations in order to match the number of quasars. To increase the number of stars we used the following methods. For every instance in the majority class, a minority sample at random is chosen resulting in clones of the minority class. This is done to enhance the classification task which would have otherwise produced results biased towards the majority class.

Methods used for classification involve:

- **Supervised learning approach**: It is the use of samples which belong to a well-defined category for learning the classification task. In our case, the samples that belong to a particular category, as derived from the spectroscopic mapping, are used for training the classifier. Random forest, which is an ensemble of tree-based classifiers, is found to produce remarkable results. Details are elaborated below and implemented using scikit-learn library (Pedregosa et al., 2011).

- **Semi-supervised learning approach**: It is the use of samples which belong to a well-defined category as well as unlabeled ones for training the classifier. These models try to surpass performance by working well when the data isn't adequate for the supervised approach. In our pursuit of finding the best classifier, generative adversarial neural networks (GANs) produced excellent results. They are a competing pair of neural networks which make use of the labeled data as well as noisy unlabeled samples to model the discriminator to work as a classifier.

We will revisit the methods in greater detail once the problem statement and nature of data are articulated.

2. Data and Problem Statement

The Galaxy Evolution Explorer or GALEX is a space telescope that operated between the years 2003 to 2012 and was developed under the NASA Explorer program. It observed astronomical sources in the far-UV and near-UV (FUV and NUV) wavebands (Gil de Paz et al., 2007). The Sloan Digital Sky Survey or SDSS is an optical survey that observed large portions of the sky in the wave bands *u,g,r,i,z* and obtained the spectra of the sources so that their red-shifts could be determined as well. The survey was conducted using a dedicated 2.5m wide-angle optical telescope operated at Apache Point Observatory in New Mexico. We have used the data released in SDSS DR12 and DR 14. The data we use assumes that the ground truth is based on the cross-matching done by Budavári et al. (2009). The latest matching based on (Budavári et al., 2009) was done using GALEX GR $6+7$ and SDSS DR 12 and 14.

We have taken into consideration the GALEX and SDSS matches from two regions.

1. **North Galactic Region:** Data from the region $> 75^\circ$ of galactic latitude is used. Since UV emission is significantly affected by dust extinction, the reason to consider the data from this region is that the stellar extinction is constant and low (Brosch et al., 1995). The results of our experiments on this data provides us with a premise of methods that is further reinforced by taking extinction into consideration. From this region, we had used 2027 quasars and 912 stars.
2. **Equatorial Region:** The extinction observed around the equator is varied and poses additional challenges to the classification process. We have selected data in the range of -30° dec to 30° dec and have used the extinction values present in the SDSS catalog to account for the extinction. The data from this region had 9182 stars and 20716 quasars.

We have used the photometric data from the cross-matched table 'GALEX GR6PLUS7 x SDSS DR12', which was available on the MAST CasJobs server. Using data from GALEX, cross-matching with labels in SDSS provides us with data from both the visible and the ultraviolet spectra. Each of the photometric samples chosen in either region have an associated spectroscopic label from the SDSS database. We have added the queries written to extract the data used for subsequent classification tasks in Appendix C.

For each of these samples, the cross matching also gives us spectroscopic labels from the SDSS database. In our experiments, we use the spectroscopic labels as the primary class label. In the database, there are no photometric labels that tell us if a source is a quasar or a star – this information can be retrieved only from the spectroscopic data.

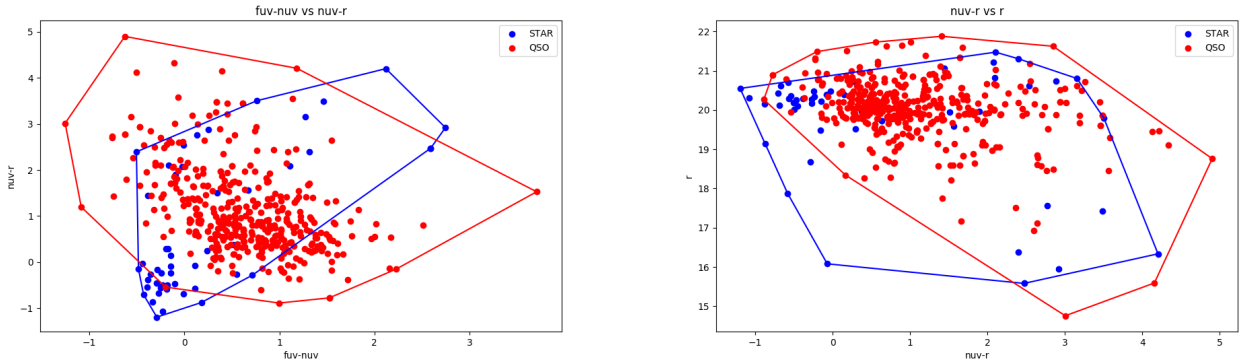

Figure 1: Convex hull indicating the linear inseparability among the classes. Quasars are plotted in red stars in blue, which are seen to be overlapping. In Figure **??**, the features used: fuv-nuv and nuv-r, are amongst the most discriminating features from domain knowledge.

Thus, this is the problem that we're trying to primarily address i.e. to classify photometric data into spectroscopic classes.

Stars and quasars look very similar in their optical images. But the spectral energy distribution for stars and quasars is different and so the optical bands SDSS namely u, g, r, i and z can be used to separate them (note however, that the separation is not obvious). The vast differentiating factor is their UV emissions (Trammell et al., 2007). In the following sections we use both optical and ultraviolet (UV) photometric data with machine learning methods to discriminate between stars and quasars. We first discuss the data and then the machine learning methods in our study.

2.1. Challenges Faced in the Classification Task

The photometric optical data is composed of images of the sources made using the optical filters: u, g, r, i, z in the SDSS survey. The UV data is from the GALEX survey, which is an all-sky in the far-UV (FUV) and near-UV (NUV) wavebands. It uses the two detectors simultaneously to obtain broadband FUV and NUV images. Before using the random forest classifier, the two classes of data were tested for linear separability by finding their convex hulls. The convex hull of a set of points is its subset which forms the smallest convex polygon that contains all the points. Stars and quasars are not linearly separable and a lot of overlap occurs between them as can be seen in Figure 1. The difference between FUV and NUV magnitudes is a measure of dust extinction, which is relatively larger for Quasars, and hence it was used to test for linear separability. (Anjum et al., 2018) This visualization concludes that linear methods cannot model the separation and justifies the exploration of Random forest on our data.

2.2. Verifying Results to Demonstrate the Robustness of the Methods

The quality of classification methods is often brought into question in exploratory applications of machine learning. Generally, statistical measures of correctness such as accuracy, precision, recall, etc. are used. In the present context, however, the attributes in the data inherently allow us to verify the outcome of the classifiers.

We classify the data into the spectroscopic classes of star or quasar (as elaborated in Section 2). Hence, we verify the outcome of the classifiers (after the classification has been done) by considering redshift. The quasars observed by SDSS are generally very far from us and are thought to derive their energy from black holes. Due to their distance from us, they generally have a larger value of observed redshift. Since the stars that are observed by the SDSS are smaller (as stars far away can't be observed easily) and are generally closer to us than quasars, it plays an effect on the redshift we observe of stars and hence their redshift is generally low; and since quasars are far away from us, their observed/estimated redshift is higher. Stars are in our galaxy. Hence we will always consider them to have redshift $z = 0$. Some star clusters and high velocity halo stars have been observed to be distant from us, but they still have $z = 0$ compared to quasars. Quasars are generally at high redshifts and usually have $z >> 0.1$. Since Quasars have evolved earlier in the Universe compared to disk or spiral galaxies, if the spectra of stars and quasars are compared, the redshift of quasars is expected to clearly separate them from stars. However, the separation is not that simple numerically as we discover and discuss, later in section 5.1. Note that we do not use redshift as a feature or attribute in the classification process rather, as a tool for analysis of the values. Redshift of the sources classified can be used to indicate the robustness of the classifiers. However, we didn't use redshift in the training phase of classifier design.

3. Results

3.1. Outcome using random forest

Based on domain knowledge, we began with a parsimony and conservative list of features. We provided a detailed comparison of these results against those resulting from usage of a comprehensive set of features followed by the imputed set in Tables 3 and 4. The results obtained use 10 fold cross-validation, where 10 disjoint test sets are used to test the model, after training on the remaining data, in order to derive a more accurate estimate of model prediction and alleviate concerns of over-fitting. This method of cross-validation also serves the purpose of the model accommodating statistical variance in new(unseen) test data.

The different subset of features used are summarized into cases to determine the best fit based on computation power and size of the dataset:

Case I *Parsimony list*: Using $fuv - nuv$, $nuv - r$ and r: The purpose of using these features is to determine the effectiveness of machine learning on a very small set of features that are thought to be effective in discriminating between stars and quasars (Wu et al., 2012; Bianchi et al., 2005).

Case II *Conservative list*: Using $fuv - r$, $fuv - nuv$, $nuv - r$, r: Extending case I, we include the additional feature of $fuv - r$ in this case.

Case III *Dropped-fuv list*: Using nuv, u, g, r, i, z and pairwise difference between them: One of the reasons to exclude the fuv attribute from the data is that the far-ultraviolet detector was damaged in April 2010 and subsequent observations were done only using the near-ultraviolet detector. Hence, from a data-analysis point of view, this poses an important challenge as a lot of fuv observations are not reliable.

Case IV *Imputed-fuv list*: Using imputed fuv and nuv, u, g, r, i, z and the pairwise differences between them: A step towards tackling the missing values of fuv is to impute them. We did this imputation using linear regression as described in appendix A. Please note, this particular case is not needed once we gathered more samples (Cases 1-8) with FUV values.

Case V *Existing-fuv list*: Using existing fuv, nuv, u, g, r, i, z and pairwise difference between them: Here we select all the data that do not have any missing values of fuv. The purpose of this case is to estimate how well the classifiers can perform when all the features are present.

Case	Precision	Recall	Accuracy
I	1.0	0.943	0.975 ± 0.018
II	1.0	0.93	0.965 ± 0.019
III	0.888	0.948	0.914 ± 0.013
IV	0.895	0.949	0.919 ± 0.013
V	0.983	1.0	0.991 ± 0.010

Table 1: Performance metrics using different set of features based on the cases discussed above (extinction not considered). The accuracy reported is the 95 percent confidence interval of measurement which explains the stability of them when validated on disjoint test-sets.

Case	Precision		Recall		F-score		Accuracy	
	Quasar	Star	Quasar	Star	Quasar	Star	Quasar	Star
I	0.95	1.0	1.0	0.94	0.97	0.97	0.94	1.0
II	0.94	1.0	1.0	0.93	0.97	0.96	0.93	1.0
III	0.94	0.89	0.88	0.95	0.91	0.92	0.94	0.88
IV	0.95	0.90	0.89	0.95	0.92	0.92	0.95	0.90
V	1.0	0.98	0.98	1.0	0.99	0.99	1.0	0.98

Table 2: Class-wise performance metrics using different set of features based on the cases discussed above: extinction not considered in these cases while we considered extinction as a feature in later experiments (see additional results, 3.3)

3.2. Outcome using GAN

Features used for training GAN not only included pairwise subtraction of features, but also the pairwise division. Overall accuracy of 96.78% obtained with 97.46% of quasars and 96.12% of stars being identified correctly. F-score obtained after 50,000 epochs was 96.78%.

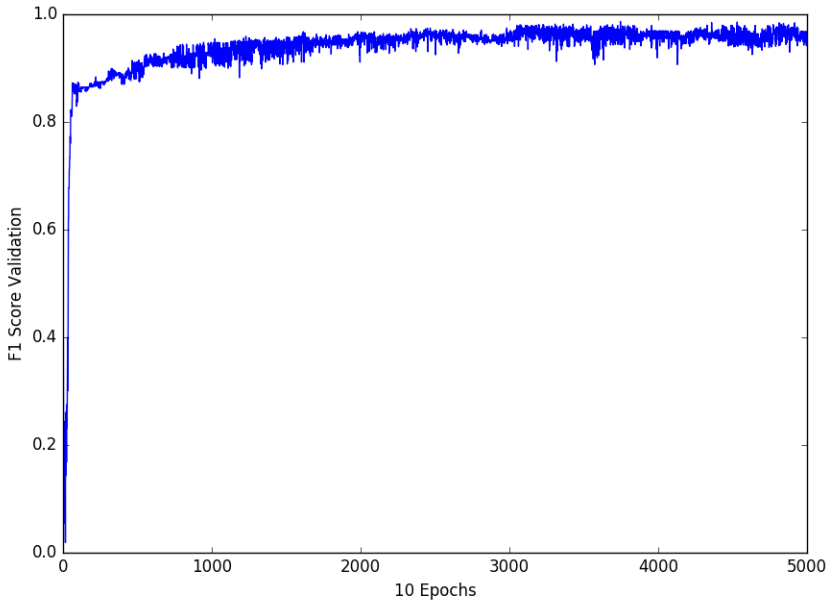

Figure 2: F-score range for 5000 epochs shows remarkable progress made by our competitive pair of networks. The performance peaks up during the first few iterations. It steadily increases until equilibrium.

3.3. Additional Results

In this section, we discuss the results of both methods[3] attempted by including extinction as a feature in the data sets.

Case 1 *North Galactic Region with Extinction Using Random Forest*: The data from the north galactic region was used with random forests by including extinction as a feature.

Case 2 *North Galactic Region with Extinction Using GANs*: The data from the north galactic region was used with GANs by including extinction as a feature.

Case 3 *Equatorial Region with Extinction Using Random Forest*: The data from the equatorial region was used with random forests by including extinction as a feature.

Case 4 *Equatorial Region with Extinction Using GANs*: The data from the equatorial region was used with GANs by including extinction as a feature.

Case 5 *Combined Data with Extinction using Random Forest*: The data from both the north galactic region as well as the equatorial region was used with random forests by including extinction as a feature.

Case 6 *Combined Data with Extinction using GANs*: The data from both the north galactic region as well as the equatorial region was used with GANs by including extinction as a feature.

Case 7 *Equatorial Region Without Extinction Using Random Forest*: While extinction is an important attribute to consider while comparing various sources, a good accuracy without considering them in regions of varied extinction

[3]The two separate scenarios (Cases I-V and cases 1-8) are considered to stress the fact that the classifiers designed are robust to additional training data and features.

would be something interesting. In this case, we use the data from the equatorial region with random forests without considering extinction as a feature.

Case 8 *Equatorial Region Without Extinction Using GANs*: The data from the equatorial region was used with GANs by excluding extinction as a feature.

Case	Precision		Recall		F-score		Accuracy	
	Quasar	Star	Quasar	Star	Quasar	Star	Quasar	Star
1	1.0	0.99	1.0	0.99	0.99	0.99	1.0	0.99
2	0.95	1.0	1.0	0.57	0.97	0.73	0.94	1.0
3	1.0	0.97	0.97	1.0	0.98	0.98	1.0	0.96
4	0.95	0.68	0.97	0.60	0.96	0.64	0.95	0.68
5	1.0	0.97	0.97	1.0	0.99	0.99	0.99	0.97
6	0.92	0.77	0.98	0.42	0.95	0.54	0.92	0.77
7	1.0	0.97	0.97	1.0	0.98	0.98	0.99	0.97
8	0.91	1.0	1.0	0.11	0.95	0.21	0.91	1.0

Table 3: Class-wise performance metrics using different set of features based on the cases discussed above.

Case	Precision	Recall	Accuracy
1	1.0	0.98	0.99 ± 0.005
2	0.95	1.0	0.95 ± 0.006
3	1.0	0.97	0.98 ± 0.005
4	0.95	0.97	0.93 ± 0.003
5	0.99	0.97	0.98 ± 0.007
6	0.92	0.98	0.91 ± 0.006
7	0.99	0.97	0.98 ± 0.007
8	0.90	1.0	0.91 ± 0.005

Table 4: Performance metrics using different set of features based on the cases discussed above. The accuracy reported is the 95 percent confidence interval of measurement which explains the stability of them when validated on disjoint test-sets. Relaxation on accuracy is insignificant.

4. Detailed Methodology

Having discussed the data and summary of results obtained, we now present deep analysis of methods designed to accomplish the reported outcome.

4.1. Random Forest Classifier

The non-linearity and clutter in the data motivated us to explore tree-based classification (Khaidem et al., 2016). These classifiers are an intuitive mechanism by which we can train a model to learn a bunch of if-else rules to accomplish the separation task. The recursive partitioning of feature space based on the impurity at each node is the driving force of these models. The impurity at each node indicates the mixture of classes at that node and can be calculated using entropy or gini impurity as shown in equation B.1 (see appendix). A decision tree is grown in a top-down manner in which each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The internal nodes represent feature values, each branch represent the criteria of split, the leaf nodes represent the outcome of following that decision path and path represents the classification rules. The root node of a tree contain all the samples and they grow until pure leaf

nodes with samples belonging to one class are obtained. Splitting at intermediate nodes is done by maximizing purity at each level relative to its parent node. These tree-based classifiers are popularly applied in the field of astronomy and produce the best results (Saha et al., 2018a). Random Forest is an ensemble of such decision tree classifiers taking into consideration the aggregate prediction of each of these models. The aggregate prediction of these estimators using random features performs better, especially using continuous features.

4.2. Ensemble Learning

One of the pertinent issues with decision trees is that of over-fitting. Over-fitting refers to the problem wherein a machine learning system generalizes very well to the training data but fails to generalize well to the test data. This can be a serious problem in real-world scenarios since the fundamental idea of machine learning is to make a system generalize well to data that it has not previously seen.

Ensemble learning refers to a methodology wherein multiple *weak learners* are combined to solve a single problem. Generally, ensembles are effective in reducing bias or variance. Random forests are a type of ensemble learner, and the mechanism used in random forests is referred to as bootstrap aggregation or *bagging* (Breiman, 1996). The collective effect of the voting mechanism from multiple trees is a reduction in variance so that the entire classifier does not overfit.

In the context of the current problem, decision trees may overfit particularly in areas of overlap between the two classes. We can see from Figure 1 that the data overlaps in certain ranges. This is further reflected in the results obtained: had the data been really separable, then machine learning classifiers would have performed perfectly well, with an accuracy of 100%, and in today's era of artificial intelligence, classifying linearly separable data is not a challenging task.

4.2.1. Training and growing Random Forest

The specified number of trees are grown with each of them using a random set of features and random samples. Instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features. The selection of random features with replacement in building each tree limits error due to bias, error due to variance and promotes diversity. They don't overfit as they consider a subset of features and a subset of samples. Therefore they are much more robust than a decision tree (Breiman, 2001), (Basak et al., 2019).

4.2.2. Prediction using Random forest

There are multiple samples in the region chosen for exploration which don't have a mapping to its spectroscopic labeled observation due to which we need to automate the task of assigning labels. Classification of these samples is done by tracing the tree top-down from root to a leaf while considering the criteria or rule of split at each subsequent node encountered in the path. Majority voting amongst the tree estimators are used to determine the label in case of conflicting predictions. This results in stability against noisy samples as a noisy point may affect a single tree, but overall prediction of the forest remains the same.

4.3. Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a pair of competing neural networks comprising of a discriminator and a generator. Discriminative models are used to predict the category given features pertaining a sample.

Generative models, on the other hand, attempt to predict probabilities of each feature given a certain label. The former learns the boundary between classes while the latter learns the distribution modeling the data that can be used along with the probabilities associated with features to replicate samples. GANs are used popularly in the tasks like generating images from sketches or text, increasing the resolution of an image etc. They are gaining popularity as an unsupervised and semi-supervised classifier in the recent years where the data is cluttered and not adequate for training. By not considering red-shift as an attribute, we lose the clear demarcation between classes and that renders a complex level of difficulty in the classification task we aim to accomplish. The task of the generator is to use Gaussian noise to produce samples replicating the input data through feedback provided by the discriminator. By detecting fake samples, the discriminator is transformed into a classifier which learns the target function using both labeled input data and unlabeled generated data. The large number of samples being used enhance the learning process and provide for more effective training. The bi-objective function that depicts the working of GANs is as follows:

$$arg\min_{G}\max_{D} V(D,G) = arg\min_{G}\max_{D} E_{x\in p(x)}[logD(x)] + E_{z\in p_G(z)}[log(1-D(G(z)))]. \tag{1}$$

where $arg\min_{G}\max_{D}$ represent the moves made by D and G followed by the payoff function. $x \in p(x)$ and $z \in p_G(z)$ represent the fact that samples come from true data and generated data respectively. The game terminates when the discriminator wins (hopefully) and we achieve the condition $p_G(z) = p(x)$ i.e. despite the best efforts of the generator G, true data can longer be distinguished from generated data. Note, D outputs a scalar 0/1 depending on its ability to classify the generated (fake) data and true data respectively.

4.3.1. GANs for classification

Standard GANs turn the discriminator into a classifier that is only useful for determining whether or not a given data point x belongs to the input data distribution. For classification, we aim to train a discriminator that classifies the data into K categories by assigning a label y to each input sample x. The discriminator should not only model data distribution but also learn the feature representation for separating into classes. Using it as a categorical classifier requires a change in objective:

- Instead of the discriminator returning a probability of a sample being real or fake, it should provide class assignments to all the encountered samples, remaining uncertain of class assignments made to synthetic generated data.
- Instead of generating samples replicating the input dataset, the generator should produce samples belonging to precisely one class with a high certainty and represent each class equally. (Springenberg, 2015)

The discriminator $D(x)$ is considered to be a function returning the logits of a sample x for K classes. This can be mapped to a probability using softmax assignment based on discriminator output:

$$p(y = k|x, D) = \frac{e^{D_{k(x)}}}{\sum_{k=1}^{K} e^{D_{k(x)}}} \tag{2}$$

4.3.2. GANs and Nash Equilibrium

GANs begin with the discriminator separating input samples and gaussian noise initiated at the generator into star, quasars and fake samples. We turn the network's sigmoid output into a softmax with 3 class outputs and set up the losses in a manner that can tune the generator (unsupervised loss) as well as classify the samples (supervised loss). The discriminator indirectly provides a feedback to the generator which refines the synthetic samples to mimic the authentic samples. This leads to greater number of samples for the discriminator to be trained on every epoch increasing classification performance. The discriminator grows capable to learn an inferred function which can map a new data point to its desirable outcome. The generator is an additional source of information and can be discarded once the discriminator has been trained. GANs are defined as a zero-sum minimax game that works with on a bi-objective function as described earlier. The non-cooperative pair of networks constantly attempt to undermine the other with the generative forging original samples and the discriminator detecting these replicated samples. An equilibrium between them is achieved when one network will not change it's action irrespective of what the other does and is referred to as the Nash equilibrium. Consider a game where A controls parameter x and B controls parameter y. The goal of A is to maximize xy, while that of B is to minimize it. Nash equilibrium is attained when $x = y = 0$.

At Nash Equilibrium, the discriminative function is learnt and at this stage, it can be used on the samples which aren't labeled in the region of exploration. In the separation of stars and quasars, a Nash equilibrium is achieved somewhere after 3000 epochs and the visualization of it can be observed in the figure 3.

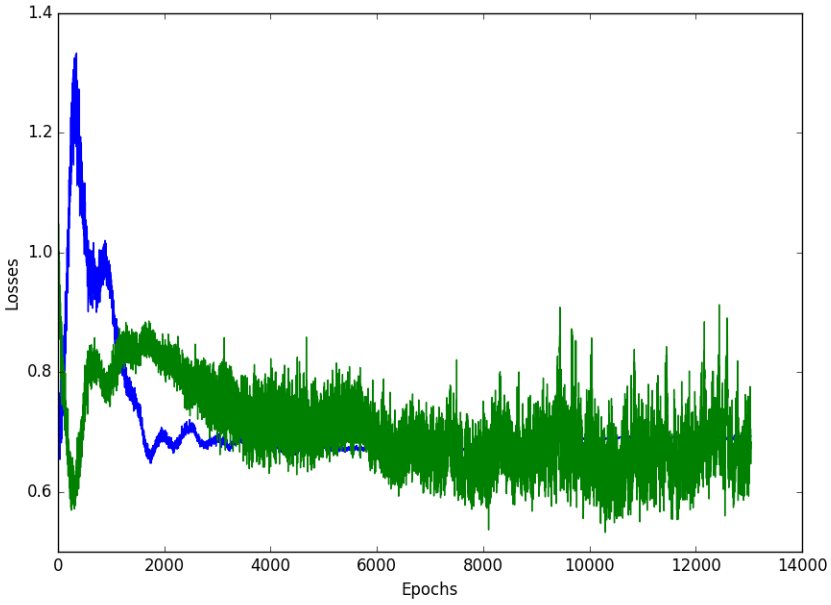

Figure 3: The generator and discriminator loss is plotted against epochs. They are indicative of the non-cooperative nature of the model to begin with but converging eventually which provides for a stable classifier.

It is interesting to note that GAN is based on the minimax principle. This implies non-existence of global optima, rather saddle points are found in the process. The bi-objective optimization problem admits of non-cooperative

structure. The competing goals of the two objective functions are balanced out in the Pareto front and the Nash equilibrium contains all solutions in the Pareto front. It is a compelling optimization problem where we seek the sub-optimal solution in the most effective manner. Since classical optimization methods such as re-scaling may not work effectively, we should seek efficient metaheuristic based optimization technique to hasten convergence of solutions in the Pareto front without compromising the quality metrics such as hypervolume, purity etc. This, in turn, may reduce the training cost. Any positive change in validation accuracy would be a plus! We shall investigate this in the context of the star quasar separation problem, in future.

4.3.3. Optimization trick in GAN classification

The goal of the discriminator, D is to maximize $V(G, D)$ given G and is to enable itself discriminate between the two classes by the use of the optimal discriminator, D^*_G. Goal of G is to minimize D^*_G when $D_G = D^*_G$ i.e. G^* is the optimal G that accomplishes this objective i.e. $G^* = argmin_G V(G, D^*_G)$. It can be shown the optimization problem has a unique solution G^* satisfying $p_G = p_{\text{data}}$. Note, p_G, p_{data} are the probability of single class token generation by G and the probability of class assignment to samples made by D respectively. Please see appendix C for proof.

We use the Jensen-Shanon divergence which is a distance metric and therefore symmetric to establish $p_G = p_{\text{data}}$ which follows from $J(p_{\text{data}}||p_G) = 0$. Note, $J(D) = -\frac{1}{2}E_{x\in p_{data}(x)} log D(x) - \frac{1}{2}E_{z\in p_G(z)} log(1 - D(z)); J(G) = -\frac{1}{2}E_{z\in p_G(z)} exp(\sigma^{-1}(D(G(z))))$. The Maximum likelihood cost for GAN is obtained by applying a cost function to every sample from the generator. The determination of the activation function from $J(G) = E_x \in p_G f(x)$ is non-trivial and important for the classifier to perform well.

By Radon-Nokodym theorem, (cite)$\partial/\partial\theta J(G) = \partial/\partial\theta \int p_G(x) f(x) dx = \int f(x) \partial/\partial\theta (p_G(x)) dx$. By a simple sampling trick, we obtain $f(x) = -\frac{p_data(x)}{p_G(x)}$; We also know that $D(x) = \sigma(a(x)) = \frac{p_{data}(x)}{p_{data}(x)+p_G(x)}$ implying $f(x) = -exp(a(x))$.

4.3.4. Choice of Activation functions in GAN classifier

Let's consider $f(x) = \frac{1}{1+kx^\alpha(1-x)^{1-\alpha}}$ (Saha et al., 2018b) and fix $k = 1, \alpha = 0$. The function assumes the form: $f(x) = \frac{1}{1+1-x} = \frac{1}{1+exp(-x)}$, by first order Taylor series approximation. This helps us choose the activation for G, in this case, either $f(x) = \frac{1}{1+kx^\alpha(1-x)^{1-\alpha}}$ (set parameters k, α in the vicinity of approximation i.e. $k = 0.91, \alpha = 0.5$) or the sigmoid activation. GAN classifier is expected to converge rapidly for the choice of parameters mentioned for the activation, $f(x) = \frac{1}{1+kx^\alpha(1-x)^{1-\alpha}}$. However, when $k = 1, \alpha = 1$; SBAF becomes $\frac{1}{1+x}$ which upon binomial expansion (restricting to first order expansion assuming $0 < x < 1$) yields $y = 1 - x = 1 -$ ReLU. Let us consider ReLU approximate in the positive half and consider the function $f(x) = kx^n$. We know that the ReLU activation function is $y = \max(0, x)$. We need to approximate the values n and k such that $f(x)$ approximates to the ReLU activation f unction over a fixed interval by minimizing relative error. The approximation problem is equivalent to the following optimization problem:
$\min \left\| kx^{n-1} \right\|$ subject to the constraints $k > 0, n > 1, -10 < x < 10$ and results in the following bounds on k and n:

$$0 < k < 1; 1 < n < 2$$

Therefore, we obtain the following continuous approximation of ReLU: $y = kx^n, x > 0$ when $0 < k < 1, 1 < n < 2$, , otherwise $y = 0$. More precisely, the approximation to the order of 10^{-3} yields $k = 0.54, n = 1.3$. Note, unlike ReLU, A-relu is differentiable at $x = 0$. The activation function has no discontinuity at $x = 0$. We choose the value of n between 1 and 2 so that the derivative doesn't explode! *Therefore, the new activation, A-relu can be used instead of E-RELU in the D part of GAN. The efficacy of the new analysis is reflected in better classification accuracy (please refer Table 3*).

4.4. Training GANs: Minibatch discrimination

The discriminator is fed with authentic input data as well as the synthetic generated data in separate minibatches. The synthetic data begins as random noise, which is refined at every pass to mimic the actual samples. The discriminator learns from a minibatch of real and fake samples alternatively, rather than individual samples, which results in a more stable training method. We compute statistics when discriminator trains on the real minibatch, followed by the ones obtained by training on the fake minibatch. The discriminator loss is then the mean of the loss obtained on training on both. Training minibatch-wise mitigates problem of mode collapse by looking at multiple examples in combination, rather than in isolation. We approximate the closeness between every pair of observations in one minibatch, $c(x_i, x_j)$, and get the overall summary of each data point by summing up how close it is to other points in the same batch which is then explicitly added to the input of the model. (Salimans et al., 2016)

5. Discussion and Conclusion

The experiments have been divided in to two phases. Initially, we tested our algorithms on a smaller data set covering a small patch in the sky. The confidence of the classification accuracy on the smaller data set is important to scale the algorithms on a larger data set. This was done when additional experiments were conducted (refer to section 3.3). Using the Random Forest, the imputation of fuv increases the number of samples used for training, but it lowers the performance measures. It even leads to miss-classification of samples with existing fuv which were classified correctly earlier. This is due to the over-fitting of individual estimators that occur and the fact that very few samples have fuv, which makes it difficult to impute for the remaining and leads to the miss-classifications. The performance results are still pretty encouraging to begin assigning categories to samples with no spectroscopic mapping in the area of exploration, which include all objects with a galactic latitude$>$ 75. Using GANs with just few samples results in excellent performance measures due to the generation of additional samples through the generator. GANs have an unbiased performance despite the imbalance in the data and hence doesn't need up-sampling. We wanted to demonstrate that using imputed samples with GANs resulted in inclusion of more synthetic data which increased the variance of each class, decreasing the classification performance and therefore such measure is uncalled for. Comparatively worse performance using Imputation is no longer a significant issue as we have gathered a lot more samples and obtained better performance using the same.

By deploying these classifiers for prediction over all the samples that do not have an associated spectroscopic label, we can automatically assign them to a category. This will assist in building up a catalog of many objects with spectroscopic classifications and such a catalog can be extremely useful for further physical studies of the properties

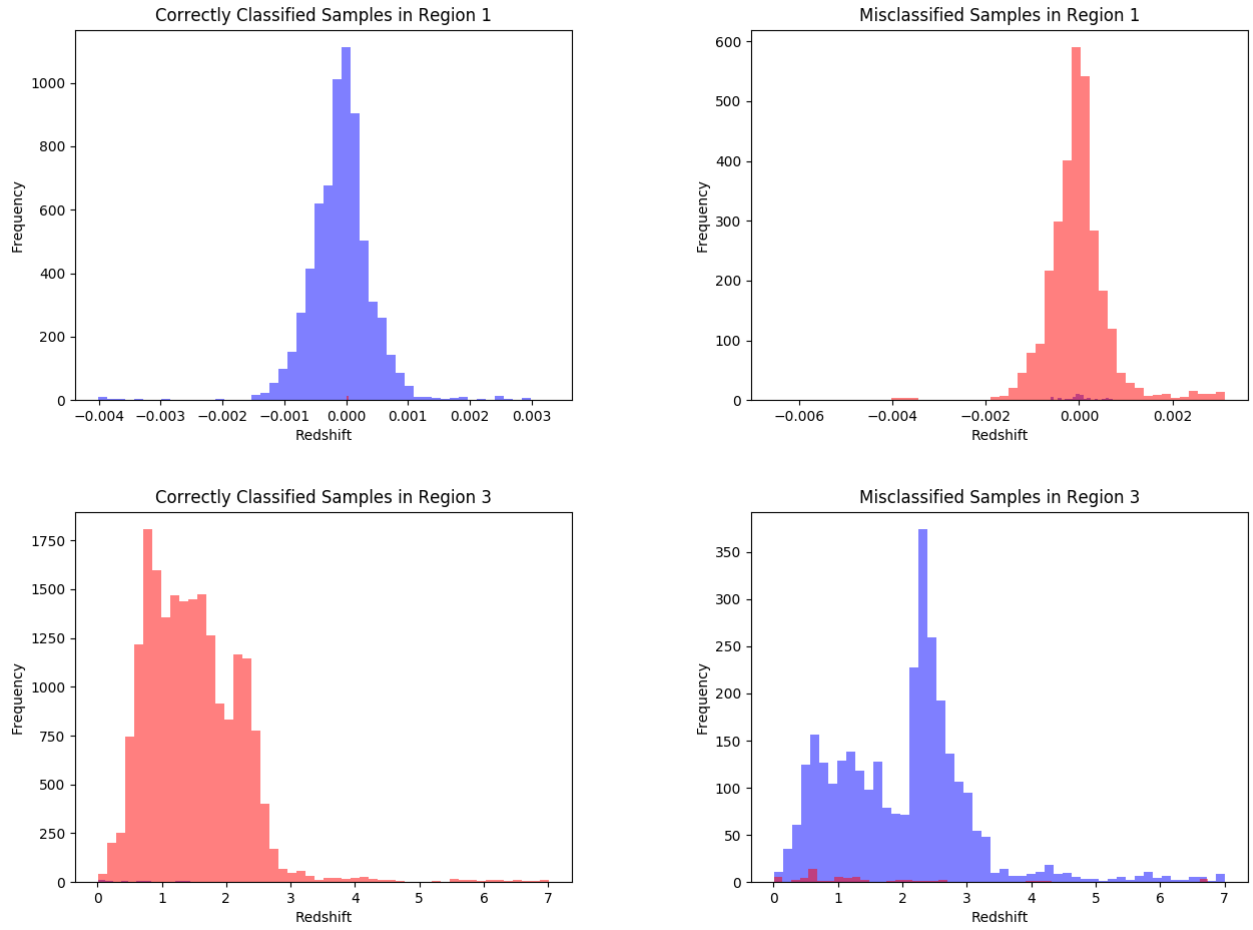

Figure 4: The histograms in blue represent the distribution of redshift of instances in the dataset that are classified as stars and the histogram in red represents the distribution of redshift of instances in the dataset classified as quasars. The smaller, purple regions represent an overlap between the classes of stars and quasars.

of stellar objects. For tagging observations without spectroscopic mapping and with fuv, our system can tag with accuracies as high as 99% whereas for those without fuv, we obtain almost 92% accuracies.

5.1. Verification of Separability by Cross-Checking with Spectrometric Redshift

As discussed in Section 1.1, we use spectrometric redshift to additionally try and verify the results of the classifiers. The distribution of the redshift attributes of classified instances are visualized in Figure 4. For the visualization in Figure 4, we consider the entire dataset comprising of instances from the north galactic region as well as the equatorial region, along with their extinction values used in the RF classifier. In order to be able to use a larger number of samples, we remove *fuv* and other features derived from *fuv* (such as pair-wise differences).

Traditionally, for the purpose of category detection, templates for stars are fitted in the redshift range from 0.004 to 0.004, and templates for quasars are fitted in the redshift range from 0.0033 to 7.00 (Pâris et al., 2017). There is a slight overlap in these ranges; while this can't be used to verify the results of classification completely accurately, a relation of the predicted labels to the redshift values of the samples can nontheless serve to verify the quality of classification. We analyse the distribution of the redshift of the samples after the classification to gain an estimate of the quality of classification. We do this for two cases: (1) considering only samples that have no

missing values of *fuv*; and (2) by dropping *fuv* and considering all the samples. Since the ranges of redshift values corresponding to stars and quasars are well investigated, we try to verify our results in a two-fold manner. First, we divide the samples into the following different catalogs, meant to serve as cases for analysis:

Catalog 1: *North Galactic Pole Only*: We select only samples that have *fuv*; then we populate the entire feature list and un RF on master catalog to label the samples, find confusion matrices, etc.

Catalog 2: *Equatorial Region Only* We select only samples that have *fuv*; then we populate the entire feature list and un RF on master catalog to label the samples, find confusion matrices, etc.

Catalog 3: From both regions, we select only samples that have *fuv*; then we populate the entire feature list and un RF on master catalog to label the samples, find confusion matrices, etc.

Catalog 4: From both the regions, we remove *fuv* and features derived from *fuv* features; then we populate the entire feature list and un RF on master catalog to label the samples, find confusion matrices, etc.

Next, we consider the following non-overlapping ranges, based on the work of Pâris et al. (2017):

Range 1: $z \leq 0.0033$: We expect the types of samples in this range to be predominantly stars.

Range 2: $z \geq 0.004$: We expect the types of samples in this range to be predominantly quasars,.

Range 3: $0.0033 < z < 0.004$: This range of redshifts represents the overlap in the template matching ranges of Pâris et al. (2017).

After the classification is completed, we divide the classified samples for verification into ranges 1, 2, and 3. We separately form the confusion matrices for each of these ranges and investigate the mismatch and the correct matching rates. Note, mismatch rate doesn't indicate the efficacy of our methods in predicting the labels incorrectly based on training data. Rather, it brings out the differences in labels based on redshift separation. This provides a rich insight in to the ground truth as we didn't use redshift in training data while designing the classification methods. The confusion matrices in each of these ranges is given in Table 5 [4]:

[4]The samples representing Figure 4 are presented as catalogs that may be found at `astrirg.org/resources/sqo-catalogs.zip`. The catalogs are organized as `catalog1`, `catalog2`, `catalog1`, `catalog4` where the naming scheme is as follows: `catX_Y_Z` where `X` represents the case, `Y` represents whether the samples are correctly classified or misclassified, and `Z` represents the range of redshift as `r1`, `r2` or `r3` corresponding to ranges 1, 2, or 3.

Catalog	Range	True Class	Predicted Class	
			Star	Quasar
1	1	Star	38	14
		Quasar	0	1
1	2	Star	1	1
		Quasar	0	0
1	3	Star	0	0
		Quasar	19	575
2	1	Star	303	97
		Quasar	5	2
2	2	Star	7	2
		Quasar	0	0
2	3	Star	2	6
		Quasar	120	3102
3	1	Star	352	100
		Quasar	4	4
3	2	Star	9	2
		Quasar	0	0
3	3	Star	3	5
		Quasar	79	3737
4	1	Star	6825	3085
		Quasar	56	27
4	2	Star	29	12
		Quasar	1	0
4	3	Star	62	61
		Quasar	3116	20189

Table 5: Confusion matrices for the post-verification process considering all the samples that do not have missing *fuv* values. The values in each cell represent the classification labels of samples and what the true class is. Thus, from the numbers in the cells, we may infer how many samples that are known to be stars are classified as stars or as quasarw, and how many samples known to be quasars are classified as quasars or stars. If a cell has the number zero in it, it means the corresponding classification in the region does not exist.

From Figure 4, it is evident that in each range, the correctly classified samples are what we primarily expect to exist. However, there also exists a considerable amount of misclassified samples in each of these ranges. The separation based on redshift, thus, promises a correct classification of stars (cross-validated) in for $z \leq 0.0033$ and a correct classification of quasars (cross-validated) in for $z \geq 0.004$. However, the misclassifications of quasars in $z \leq 0.0033$ and stars in $z \geq 0.004$ is high. Overall, there is roughly a 70% correlation between the predicted labels of the classifiers and their redshift ranges. This attempt at cross-validating machine classification results is a logical way to check accuracy and quality of training data, since prediction is based on the quality of training data. The table 5 and percentages reported provide some assurance on the efficacy of the model developed to classify Stars and Quasars. We release the catalog based on the percentage cross-validation, hoping this is one set of stars and quasars with class labels validated by ground truth.

We observe that Random Forests outperformed GANs marginally on the smaller data set. This is not surprising as we're inclined to believe that the Star-Quasar classes are separated statistically, rather than as a strange decision boundary in a high-dimensional space. Since classical ML models have always relied on statistics, RF performs better as observed. But as the data size grew, GAN quickly deciphered the separation since it had more training data to learn. It will be interesting to explore if using ensembles of smaller NNs produce a stronger model. We conclude with the note that our GAN model consists of a non-saturating equilibrium and therefore could always be

improved by exploiting a more powerful divergence metric (between G and D).

Acknowledgements

The authors would like to thank the Science and Engineering Research Board (SERB)-Department of Science and Technology (DST), Government of of India for supporting this research. The project reference number is: SERB-EMR/ 2016/005687. The authors would also like to thank GALEX and SDSS projects.

GALEX (Galaxy Evolution Explorer) is a NASA Small Explorer, launched in April 2003. We gratefully acknowledge NASA's support for construction, operation, and science analysis for the GALEX mission, developed in cooperation with the Centre National d'Etudes Spatiales (CNES) of France and the Korean Ministry of Science and Technology. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

Appendix A. Motivation and Techniques used for Imputation

The strong correlation between fuv and nuv encouraged us to impute fuv using nuv values. The relation between them can be observed in figure A.5 and it motivated us to impute the missing fuv values using linear regression. The samples with existing fuv observation were used to derive the model that fits and then used to derive the missing values. The Pearson correlation constant, which indicates the linear relationship between two variables, of 0.8 measured between existing fuv and nuv values was found to be the same post imputation as well. The mean squared error obtained on the validation set after imputation using linear regression was 0.2 and hence proved to be a good fit.

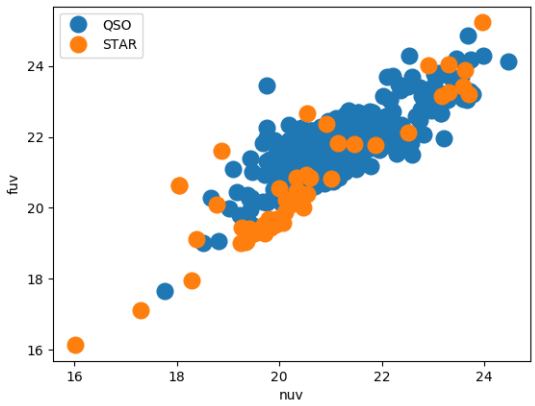

Figure A.5: Linear relationship between fuv and nuv magnitude motivated us to explore imputation of fuv using nuv.

Appendix B. Visualization of Growth of a Tree Estimator

Classification using tree-bases methods results in ease of visualization of the decision path while classifying a sample. Random forest facilitate building of multiple tree estimators and the flow of one such estimator is shown in figure B.6. Deciding on the best feature on which the samples must be split is decided by gini impurity computed as:

$$I = 1 - \sum_{i=1}^{k} (p(c_i|t))^2 \tag{B.1}$$

where I represents the impurity in a node t; $\forall i \in \{1, ..., k\}$ $c_i \in C$, where $C = \{c_1, c_2, ..., c_k\}$ is the set of classes in the learning sample L; $p(c_i|t)$ represents the conditional probability that a learning instance or object in t belongs to class c. The gain at each node is computed as follows:

$P_L = n_L \ / \ n_L + n_R$

$P_R = n_R \ / \ n_R + n_L$

$G = I$ - P_L* I_L - P_R * I_R

At root node or node labeled 0 in figure **??** with 20 samples,3 samples are star and 17 are quasars. The gini impurity at that node is calculates as:

$$1 - \frac{3}{20}^2 - \frac{17}{20}^2 = 0.255 \tag{B.2}$$

Amongst all features considered at that node, 'r' at threshold 0.967 is chosen as the split condition as it reduces the the impurity the most. At node labeled 1, there is 1 sample which is a star. It's pure due to presence of one class only, so that's not split again and it's a leaf node labeled as 'star'. Node labeled 2 has 2 stars and 17 quasars which results in a split in the manner shown above and goes on at each step until every node is pure.

r	nuv-r	fuv-nuv	fuv-r	class
21.2204456329	2.0733261108	-0.0217781067	2.0515480042	STAR
21.3414783478	1.6435031891	0.4095039367	2.0530071258	QSO
21.3414783478	1.6435031891	0.4095039367	2.0530071258	QSO
20.8177127838	2.0895900726	1.1127605439	3.2023506165	STAR
20.6770896912	-0.1303710938	0.8205509186	0.6901798248	QSO
18.4608707428	3.5575504303	1.1403541565	4.6979045868	QSO
21.0197296143	2.0997486115	0.6891441345	2.7888927459	QSO
20.4750003815	2.4873600006	-0.2143669129	2.2729930877	QSO
20.7446346283	2.5522708893	-0.1860618592	2.3662090301	QSO
20.2102870941	2.9362220764	-0.4703674316	2.4658546448	QSO
20.2102870941	2.9362220764	-0.4703674316	2.4658546448	QSO
21.6394004822	0.8932762146	1.745054245	2.6383304596	QSO
21.6001605988	1.6719169617	0.4807434082	2.1526603698	QSO
20.5062675476	1.2284984589	0.8586673736	2.0871658325	QSO
20.6233310699	2.536195755	-0.0056858063	2.5305099488	STAR
19.8593864441	3.4835510254	0.3408527374	3.8244037628	QSO
19.8593864441	3.4835510254	0.3408527374	3.8244037628	QSO
20.5044460297	3.217010498	0.4503669739	3.6673774719	QSO
19.8796005249	2.8876914978	0.3722267151	3.2599182129	QSO
19.6023292542	3.4539089203	0.6114444733	4.0653533935	QSO

Table B.6: Sample of the dataset to demonstrate working of a tree estimator used in Random Forest.

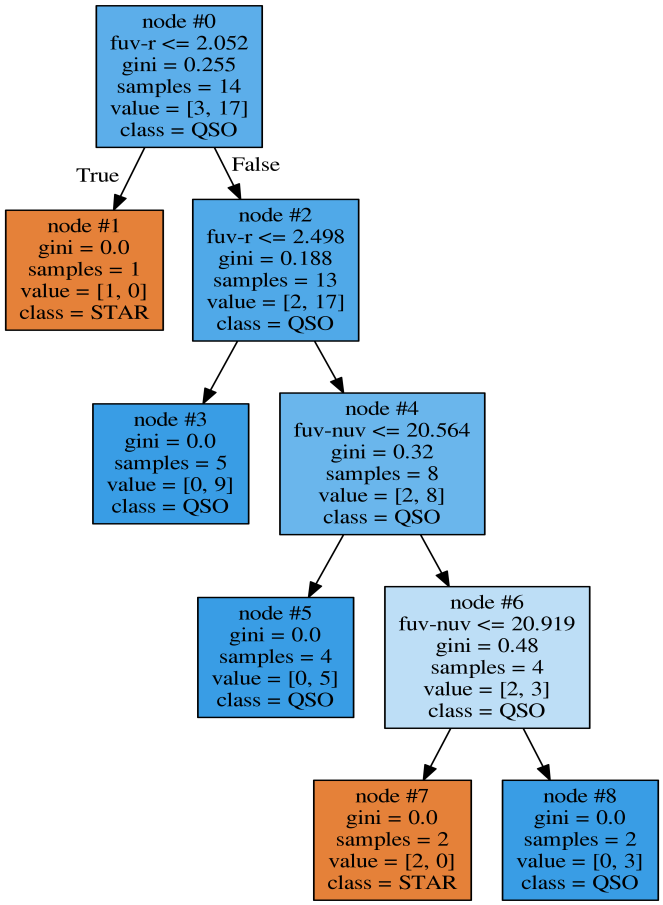

Figure B.6: Growth of a decision tree estimator from samples shown in table.

Appendix C. Queries Used to Fetch Data from MAST CasJobs

In this section, we explain the data acquisition process and the queries used to fetch the data from the MAST CasJobs service. Some amount of pre-filtering has been done in the process of data acquisition such that spectrometric instances with multiple photometric matches are not considered as they could create confusion within the classification algorithms. The data thus acquired has less confusion.

All of the following queries are run sequentially in the MAST CasJobs service. The following queries wil fetch the data that we have used corresponding to the north galactic region.

Query 1: First, a table of matches of IDs is created. The table `xSDSSDR12` had the cross matches, which means that the IDs of the matched SDSS and GALEX sources are available here. We use the additional qualifiers: `distanceRank = 1`, `reverseDistanceRank = 1`, `multipleMatchCount = 1` and `reverseMultipleMatchCount = 1` to ensure that a clean sample is acquired.

```
SELECT objid, SDSSobjid from xSDSSDR12
INTO preMatchesDR12_bestCase
WHERE distanceRank = 1
AND reverseDistanceRank = 1
AND multipleMatchCount = 1
AND reverseMultipleMatchCount = 1;
```

Query 2: After an initial list of IDs have been acquired, we fetch the attributes from other tables in the database, that contain the attributes along with the IDs. The IDs are used to match photometric instances across tables.

```
SELECT g.objid as galex_objid,
s.objid as sdss_objid,
s.ra, s.dec,
s.u, s.g, s.r, s.i, s.z,
g.nuv_mag, g.fuv_mag
FROM preMatchesDR12_bestCase as pm
INTO bestCaseData_northGalactic
INNER JOIN SDSS_DR12.photoobjall as s on pm.SDSSObjid=s.objid
INNER JOIN GALEX_GR6Plus7.photoobjall as g on pm.objid=g.objid
WHERE g.glat > 75;
```

After this query has been implemented, the photometric data is ready. What remains is to match these photometric attributes with spectrometric IDs so that we can extract the appropriate class labels.

Query 3: In this query, we match the photometric data with the appropriate spectrometric data.

```
SELECT p.galex_objid, p.sdss_objid, p.ra, p.dec,
p.u, p.g, p.r, p.i, p.z, p.extinction_u, p.extinction_g,
p.extinction_r, p.extinction_i, p.extinction_z,
```

```
p.nuv_mag, p.fuv_mag, s.class, s.subclass, s.z as spec_redshift
FROM bestCaseData_northGalactic as p
INTO northGalactic_specMatch
JOIN SDSSDR14.specobjall AS s
ON p.sdss_objid = s.bestObjID
WHERE s.bestObjID = s.fluxObjID;
```

The above queries can be used to fetch data corresponding to the equatorial region. For the equatorial region, the range of the coordinates used should be $-30° \leq$ declination $\leq 30°$.

Bibliography

Anjum, A., Das, M., Murthy, J., Gudennavar, S.B., Gopal, R., Bubbly, S.G., 2018. Template-based classification of SDSS- GALEX point sources. Journal of Astrophysics and Astronomy 39, 61. doi:10.1007/s12036-018-9552-3.

Bai, Y., Liu, J., Wang, S., Yang, F., 2019. Machine Learning Applied to StarGalaxyQSO Classification and Stellar Effective Temperature Regression. 157, 9. doi:10.3847/1538-3881/aaf009, arXiv:1811.03740.

Basak, S., Kar, S., Saha, S., Khaidem, L., Dey, S.R., 2019. Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance 47, 552–567. URL: https://doi.org/10.1016/j.najef.2018.06.013, doi:10.1016/j.najef.2018.06.013.

Bianchi, L., Seibert, M., Zheng, W., Thilker, D.A., Friedman, P.G., Wyder, T.K., Donas, J., Barlow, T.A., Byun, Y.I., Forster, K., Heckman, T.M., Jelinsky, P.N., Lee, Y.W., Madore, B.F., Malina, R.F., Martin, D.C., Milliard, B., Morrissey, P., Neff, S.G., Rich, R.M., Schiminovich, D., Siegmund, O.H.W., Small, T., Szalay, A.S., Welsh, B.Y., 2005. Classification and Characterization of Objects from the Galaxy Evolution Explorer Survey and the Sloan Digital Sky Survey. 619, L27–L30. doi:10.1086/423710.

Bovy, J., Hennawi, J.F., Hogg, D.W., Myers, A.D., Kirkpatrick, J.A., Schlegel, D.J., Ross, N.P., Sheldon, E.S., McGreer, I.D., Schneider, D.P., Weaver, B.A., 2011. Think Outside the Color Box: Probabilistic Target Selection and the SDSS-XDQSO Quasar Targeting Catalog. 729, 141. doi:10.1088/0004-637X/729/2/141, arXiv:1011.6392.

Breiman, L., 1996. Bagging predictors. Machine learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. URL: https://doi.org/10.1023/A:1010933404324, doi:10.1023/A:1010933404324.

Brosch, N., Almoznino, E., Leibowitz, E.M., Netzer, H., Sasseen, T.P., Bowyer, S., Lampton, M., Wu, X., 1995. A study of ultraviolet objects near the north galactic pole with FAUST. The Astrophysical Journal 450, 137. URL: https://doi.org/10.1086/176125, doi:10.1086/176125.

Budavári, T., Heinis, S., Szalay, A.S., Nieto-Santisteban, M., Gupchup, J., Shiao, B., Smith, M., Chang, R., Kauffmann, G., Morrissey, P., Schiminovich, D., Milliard, B., Wyder, T.K., Martin, D.C., Barlow, T.A., Seibert,

M., Forster, K., Bianchi, L., Donas, J., Friedman, P.G., Heckman, T.M., Lee, Y.W., Madore, B.F., Neff, S.G., Rich, R.M., Welsh, B.Y., 2009. GALEX-SDSS CATALOGS FOR STATISTICAL STUDIES. The Astrophysical Journal 694, 1281–1292. URL: https://doi.org/10.1088/0004-637x/694/2/1281, doi:10.1088/0004-637x/694/2/1281.

Das, P., Sanders, J.L., 2019. MADE: a spectroscopic mass, age, and distance estimator for red giant stars with Bayesian machine learning. 484, 294–304. doi:10.1093/mnras/sty2776, arXiv:1804.09596.

DiPompeo, M.A., Bovy, J., Myers, A.D., Lang, D., 2015. Quasar probabilities and redshifts from WISE mid-IR through GALEX UV photometry. 452, 3124–3138. doi:10.1093/mnras/stv1562, arXiv:1507.02884.

Fan, X., 1999. Simulation of Stellar Objects in SDSS Color Space. 117, 2528–2551. doi:10.1086/300848, arXiv:astro-ph/9902063.

Gil de Paz, A., Boissier, S., Madore, B.F., Seibert, M., Joe, Y.H., Boselli, A., Wyder, T.K., Thilker, D., Bianchi, L., Rey, S.C., Rich, R.M., Barlow, T.A., Conrow, T., Forster, K., Friedman, P.G., Martin, D.C., Morrissey, P., Neff, S.G., Schiminovich, D., Small, T., Donas, J., Heckman, T.M., Lee, Y.W., Milliard, B., Szalay, A.S., Yi, S., 2007. The GALEX Ultraviolet Atlas of Nearby Galaxies. 173, 185–255. doi:10.1086/516636, arXiv:astro-ph/0606440.

Gunn, J.E., Carr, M., Rockosi, C., Sekiguchi, M., Berry, K., Elms, B., de Haas, E., Ivezić, Ž., Knapp, G., Lupton, R., Pauls, G., Simcoe, R., Hirsch, R., Sanford, D., Wang, S., York, D., Harris, F., Annis, J., Bartozek, L., Boroski, W., Bakken, J., Haldeman, M., Kent, S., Holm, S., Holmgren, D., Petravick, D., Prosapio, A., Rechenmacher, R., Doi, M., Fukugita, M., Shimasaku, K., Okada, N., Hull, C., Siegmund, W., Mannery, E., Blouke, M., Heidtman, D., Schneider, D., Lucinio, R., Brinkman, J., 1998. The Sloan Digital Sky Survey Photometric Camera. 116, 3040–3081. doi:10.1086/300645, arXiv:astro-ph/9809085.

Jimenez, R., Spergel, D.N., Niemack, M.D., Menanteau, F., Hughes, J.P., Verde, L., Kosowsky, A., 2009. Southern Cosmology Survey. III. QSOs From Combined GALEX and Optical Photometry. 181, 439–443. doi:10.1088/0067-0049/181/2/439, arXiv:0811.4134.

Khaidem, L., Saha, S., Basak, S., Dey, S.R., 2016. Predicting the direction of stock market prices using random forest. URL: "https://www.researchgate.net/publication/301818771_Predicting_the_direction_of_stock_market_prices_using_random_forest".

Khramtsov, V., Sergeyev, A., Spiniello, C., Tortora, C., Napolitano, N.R., Agnello, A., Getman, F., de Jong, J.T.A., Kuijken, K., Radovich, M., 2019. KiDS-SQuaD II: Machine learning selection of bright extragalactic objects to search for new gravitationally lensed quasars. arXiv e-prints , arXiv:1906.01638arXiv:1906.01638.

Krakowski, T., Małek, K., Bilicki, M., Pollo, A., Kurcz, A., Krupa, M., 2016. Machine-learning identification of galaxies in the WISE $\times SuperCOSMOSall - skycatalogue. 596, A39. doi:$, arXiv:1607.01188.

Martin, D.C., Fanson, J., Schiminovich, D., Morrissey, P., Friedman, P.G., Barlow, T.A., Conrow, T., Grange, R., Jelinsky, P.N., Milliard, B., Siegmund, O.H.W., Bianchi, L., Byun, Y.I., Donas, J., Forster, K., Heckman, T.M., Lee,

Y.W., Madore, B.F., Malina, R.F., Neff, S.G., Rich, R.M., Small, T., Surber, F., Szalay, A.S., Welsh, B., Wyder, T.K., 2005. The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission. 619, L1–L6. 10.1086/426387, arXiv:astro-ph/0411302.

Morice-Atkinson, X., Hoyle, B., Bacon, D., 2018. Learning from the machine: interpreting machine learning algorithms for point- and extended-source classification. 481, 4194–4205. 10.1093/mnras/sty2575, arXiv:1712.03970.

Morrissey, P., Conrow, T., Barlow, T.A., Small, T., Seibert, M., Wyder, T.K., Budavári, T., Arnouts, S., Friedman, P.G., Forster, K., Martin, D.C., Neff, S.G., Schiminovich, D., Bianchi, L., Donas, J., Heckman, T.M., Lee, Y.W., Madore, B.F., Milliard, B., Rich, R.M., Szalay, A.S., Welsh, B.Y., Yi, S.K., 2007. The Calibration and Data Products of GALEX. 173, 682–697. 10.1086/520512.

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

Preethi, K., Gudennavar, S.B., Bubbly, S.G., Murthy, J., Brosch, N., 2014. Photometric identification of objects from Galaxy Evolution Explorer Survey and Sloan Digital Sky Survey. 437, 771–776. 10.1093/mnras/stt1935, arXiv:1308.2398.

Pâris, I., Petitjean, P., Aubourg, E., Myers, A.D., Streblyanska, A., Lyke, B.W., Anderson, S.F., Armengaud, E., Bautista, J., Blanton, M.R., Blomqvist, M., Brinkmann, J., Brownstein, J.R., Brandt, W.N., Burtin, E., Dawson, K., de la Torre, S., Georgakakis, A., Gil-Marin, H., Green, P.J., Hall, P.B., Kneib, J.P., LaMassa, S.M., Goff, J.M.L., MacLeod, C., Mariappan, V., McGreer, I.D., Merloni, A., Noterdaeme, P., Delabrouille, N.P., Percival, W.J., Ross, A.J., Rossi, G., Schneider, D.P., Seo, H.J., Tojeiro, R., Weaver, B.A., Weijmans, A.M., Yèche, C., Zarrouk, P., Zhao, G.B., 2017. The sloan digital sky survey quasar catalog: Fourteenth data release 10.1051/0004-6361/201732445, arXiv:arXiv:1712.05029.

Richards, G.T., Fan, X., Newberg, H.J., Strauss, M.A., Vanden Berk, D.E., Schneider, D.P., Yanny, B., Boucher, A., Burles, S., Frieman, J.A., Gunn, J.E., Hall, P.B., Ivezić, Ž., Kent, S., Loveday, J., Lupton, R.H., Rockosi, C.M., Schlegel, D.J., Stoughton, C., SubbaRao, M., York, D.G., 2002. Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample. 123, 2945–2975. 10.1086/340187, arXiv:astro-ph/0202251.

Saha, S., Basak, S., Safonova, M., Bora, K., Agrawal, S., Sarkar, P., Murthy, J., 2018a. Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets. Astronomy and Computing 23, 141–150. https://doi.org/10.1016/j.ascom.2018.03.003, 10.1016/j.ascom.2018.03.003.

Saha, S., Mathur, A., Bora, K., Agrawal, S., Basak, S., 2018b. SBAF: A new activation function for artificial neural net based habitability classification. CoRR abs/1806.01844. http://arxiv.org/abs/1806.01844, arXiv:1806.01844.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, Curran Associates Inc., USA. pp. 2234–2242. http://dl.acm.org/citation.cfm?id=3157096.3157346.

Springenberg, J.T., 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv:arXiv:1511.06390.

Trammell, G.B., Vanden Berk, D.E., Schneider, D.P., Richards, G.T., Hall, P.B., Anderson, S.F., Brinkmann, J., 2007. The UV Properties of SDSS-Selected Quasars. 133, 1780–1794. 10.1086/511817, arXiv:astro-ph/0611549.

Viquar, M., Basak, S., Dasgupta, A., Agrawal, S., Saha, S., 2018. Machine learning in astronomy: A case study in quasar-star classification. arXiv:arXiv:1804.05051.

Wu, X.B., Hao, G., Jia, Z., Zhang, Y., Peng, N., 2012. SDSS Quasars in the WISE Preliminary Data Release and Quasar Candidate Selection with Optical/Infrared Colors. 144, 49. 10.1088/0004-6256/144/2/49, arXiv:1204.6197.