UE17CS303 - MACHINE LEARNING ASSIGNMENT
Separating Stars from Quasars: Machine Learning Investigation Using Photometric Data

1st Skanda VC
PES1201700987
B-Tech, CSE
PES University
Bangalore, India
skandavc18@gmail.com

2nd G R Dheemanth
PES1201700229
B-Tech, CSE
PES University
Bangalore, India
dheemanthgr@gmail.com

3rd Mehul Thakral
PES1201701122
B-Tech, CSE
PES University
Bangalore, India
mehul.thakral@gmail.com

***Abstract*—Machine Leaning in astronomy is a daunting task due to the nature of data-sets available. There is huge diversity in data with no clear linear or non linear separation. In this paper we explore the efficacy of gradient boosted trees in classification of starts and quasars using GALEX and SDSS photo-metric data. We found a reasonably satisfactory range of 88 to 96% accuracy.**

***Index Terms*—gradient boosting,gbt, spectroscopic features, celestial objects**

I. Introduction

With large scale photo-metric surveys, the major challenge is separation of stars and quasars. Since both stars and quasars have compact optical morphology, it is difficult to classify them. So generally spectroscopic data is used for separation. But spectroscopic surveys take enormous amounts of time. Hence, we have proposed Gradient Boosted Tree model to predict whether the galactic object is a star or a quasar without using spectroscopic data.

A. Dataset

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 wavebands. 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.The data-set consists of observations from GALEX and SDSS surveys. 2 regions were considered in these surveys.

- North Galactic Region: Data from the region $> 75°$ of galactic latitude was used. From this region there were 2027 quasars and 912 stars.
- Equatorial Region: Data in the range of $30°$ to $30°$ was used. The data from this region had 9182 stars and 20716 quasars.

All this data is in 4 catalogues.

- Catalogue 1 : It contains observations from North Galactic Region Only. Selected samples had fuv.
- Catalogue 2 : It contains observations from Equatorial Region Only. Selected samples had fuv.
- Catalogue 3 : It contains observations from both North Galactic Region and Equatorial Region. Selected samples had fuv.
- Catalogue 4 : It contains observations from both North Galactic Region and Equatorial Region. Samples without fuv were also present.

Overall, the number of quasars are more than stars in the data-set which is evident from figure 2.

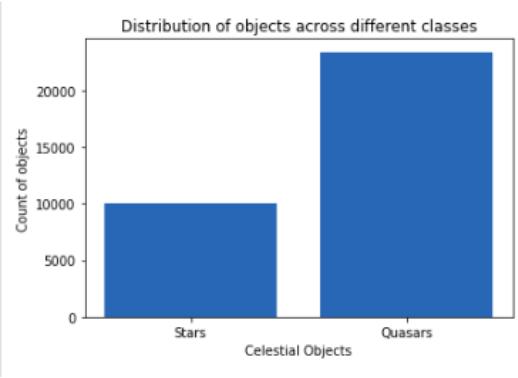

Fig. 1. Distribution of data points across different classes.

II. Ensemble Techniques

When we fit a model to predict a variable, the causes for differences are

- noise
- variance
- bias

Ensemble techniques can be used to reduce these factors and give better accuracy. An ensemble is a set of predictions

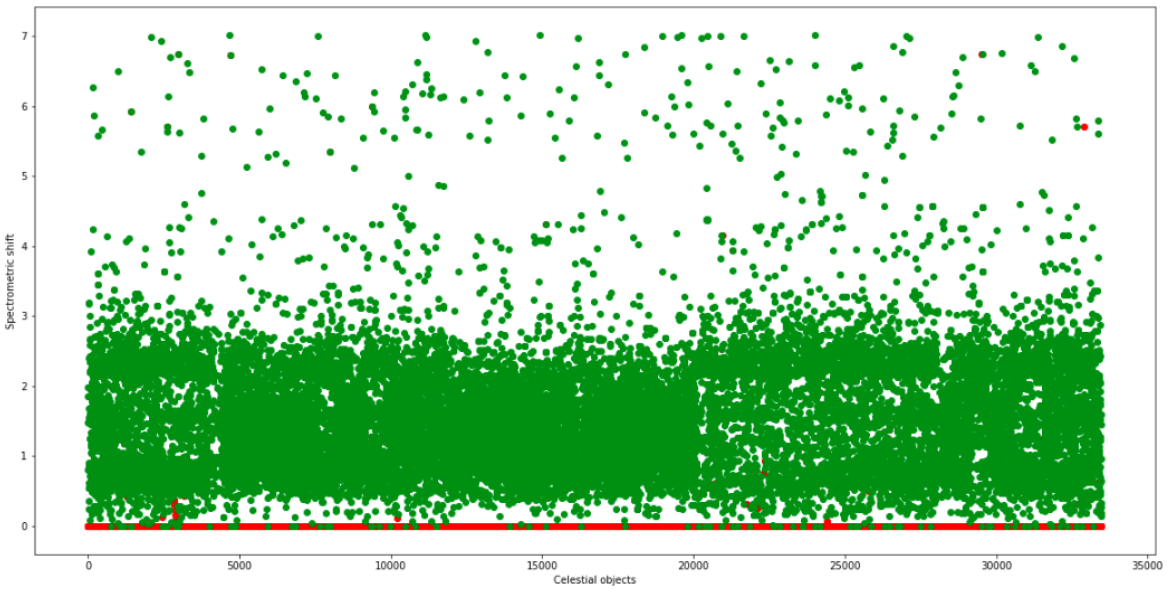

Fig. 2. Scatter Plot of distribution of stars and quasars with respect to spectrometric shift

combined to give a final prediction. There are 2 main ensemble techniques

1. Bagging : Here, we build many learners and combine them using model averaging techniques. Random forests come under this category.This method was proven to be very effective according to Dr Saha's paper [1].

2. Boosting : It is an ensemble technique in which the predictors are not made independently but sequentially. Gradient Boosting comes under this technique.

This paper talks about Gradient Boosted Trees.

III. Gradient Boosted Trees

Gradient Boosted Tree is a technique which produces an ensemble of weak decision tree models.

It involves 3 parts :

- The Loss function
- A weak learner
- An additive model to add weak learners

A. Loss Function

The loss function is a differentiable function which has to be minimized. The loss function defined in our case was L2 loss.

Loss = $\sum(y - \hat{y})^2$

This function was minimized using gradient descent. Gradient Descent is an iterative optimization algorithm for finding the minimum of a function. The predicted values are updated in the direction towards the minimum. They are updated by a size which is called as the learning rate. The learning rate used was 0.22. But unlike the traditional gradient descent where the values are minimized, here we have weak learners/decision trees.

B. Weak Learner

In gradient boosted tree, decision trees are used as weak learners. Trees are constructed in a greedy manner, choosing the best split points based on purity scores.Generally the weak learners are constrained in specific ways such as maximum depth of tree, minimum split gain, etc. We chose a min split gain of 0.1 and max depth of 5.

C. Additive Model

In case of gradient boosted trees, the trees are added one at a time. The existing trees in the model are not changed. After calculating the loss, to perform gradient descent, we must add a tree to the model such that it reduces the loss. This is done by parameterizing the tree. The parameters of the tree are modified to minimize the loss. The output for the new tree is then added to the output of the existing sequence of trees to improve the final output.

IV. Challenges Faced in the Classification Task

The main challenge was that Stars and quasars were not linearly separable and a lot of overlap occurred between them. So we had to go with ensemble techniques. Random forests was not chosen because the data-set was imbalanced. Usually when we fit a model like logistic regression or random forest on such a data-set, there are high chances that the model is biased. Where as in Gradient Boosting, it is a sequential process and thus every time it makes an incorrect prediction, it focuses more on that incorrectly predicted data point. Therefore there is less bias.

V. Results

The accuracy's obtained after training the model separately on catalogue 1,2,3,4 are summarized in the table.

Dataset Category	accuracy
cat1	96.9%
cat2	95.06%
cat3	94.88%
cat4	88.9%

TABLE I
CATALOGUE WISE ACCURACY

In the original data available from both GALEX and SDSS, the class labels were not available. To know whether the given galactic object is a star or a quasar, the spectrometric red-shift data had to be used. Since it was not used for training and the ground truth (class labels) was not 100% accurate, we had to verify the obtained results with spectrometric red-shift values.

First the spectrometric red-shift values were divided into 3 ranges based on the work of P^aris et al(2017):

- Range 1 : z <= 0.0033: They are predominantly stars.
- Range 2 : z >= 0.004: They are predominantly quasars.
- Range 3 : 0.0033 < z < 0.004: This range of red-shifts represents the overlap in the template matching ranges.

These range values were correlated against the predicted class labels. Spearman correlation was found between the red-shift ranges and the predicted class labels. Only the values corresponding to range 1 and range 2 were used. The results are as follows

Dataset Category	Correlation
cat1	0.88
cat2	0.76
cat3	0.73
cat4	0.74

TABLE II
CATALOGUE WISE CORRELATION BETWEEN RED-SHIFT AND CLASS LABELS

Bias Variance decomposition was done to know the bias of the model. The results have been summarized in the table given below.

Dataset Category	Bias	Variance
cat1	0.004	0.003
cat2	0.0365	0.014
cat3	0.0248	0.012
cat4	0.105	0.045

TABLE III
CATALOGUE WISE BIAS VARIANCE DECOMPOSITION

Also analysis was also done to find out features of importance which influenced the output of the gradient boosted tree which can be seen in figure 1.

Also, the loss falls drastically within a few epochs with gradient boosted trees and the overall training time is a fraction of what it is with neural nets, random forests,etc.

VI. CONCLUSIONS

Seeing the good accuracy of our model the task of classification of celestial objects can be automated using such gradient boosting frameworks which not only scale with large amount of data but also are memory efficient. Also it shows that machine learning approach comes to rescue in such gigantic and illegible problems where even ground truth suffers from errors.

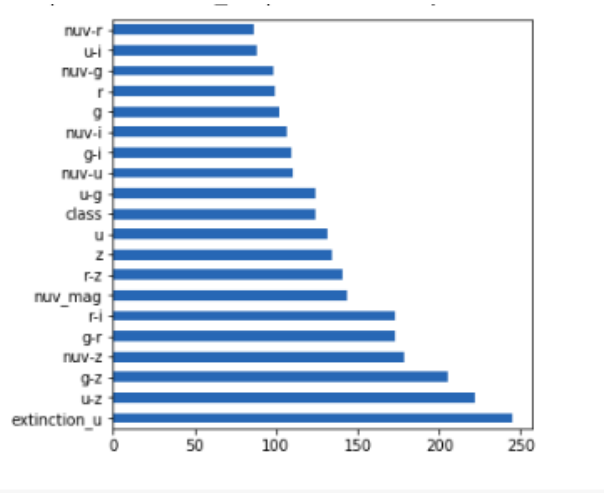

Fig. 3. Most important 20 features of the boosting tree

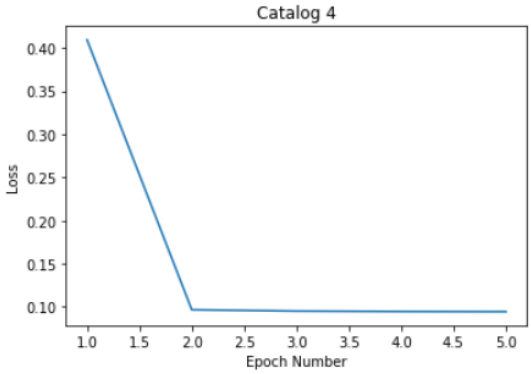

Fig. 4. category 4 : loss vs epochs

REFERENCES

[1] Makhija, S., Saha, S., Basak, S. and Das, M. (2019). Separating stars from quasars: Machine learning investigation using photo-metric data. Astronomy and Computing, 29, p.100313.