Classification of Photometric Data As Stars or Quasars

Sanjay Chari
Department of CSE
PES University
Bangalore, India
sanjaychari2xy2@gmail.com

Aditya Shankaran
Department of CSE
PES University
Bangalore, India
adityashankaran10@gmail.com

Athul Sandosh
Department of CSE
PES University
Bangalore, India
athulsandosh99@gmail.com

***Abstract*—We perform binary classification on a dataset that contains the class labels stars and quasars. The dataset has been obtained from the GALEX and SDSS surveys. Previous methods have reported accuracies in the range 91-100 percent. We employ a decision tree trained using the Classification and Regression Trees approach in order to achieve satisfactory metrics on the dataset. We report performance metrics that are on par with the contemporary approaches on the dataset.**

***Index Terms*—Binary Classification, Decision Tree, CART, GALEX, SDSS**

I. Introduction

The dataset used for the project contains photometric data about stars and quasars. The photometric data was obtained from GALEX and SDSS surveys. The GALEX or Galaxy Evolution Explorer is an orbiting ultraviolet space telescope which was launched on April 28, 2003 and operated until early 2012. The telescope made observations in ultraviolet wavelengths to measure the history of star formation in the universe 80 percent of the way back to the Big Bang. The SDSS or Sloan Digital Sky Survey is a major multi-spectral imaging and spectroscopic redshift survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. We implement a Decision Tree trained using the Classification and Regression Trees approach, with maximum depth hyperparameter set to 8.

II. Problem Statement

We were provided with four catalogues, each with a collection of photometric data related to stars or quasars. We were required to classify each of the data points in the csv file of every catalog, as star or quasar. The csv files contain 39 columns, with column headings such as galexobjid, sdssobjid, and so on. The task at hand is to achieve accuracy in the range 91-100 percent, using any machine learning algorithm of our choice.

III. Methodology

We employ a decision tree for our classification task, as when we compared its performance with other algorithms, we achieved convergence faster with decision tree, and achieved good accuracy too. A decision tree classifier is a binary tree where predictions are made by traversing the tree from root to leaf — at each node, we go left if a feature is less than a threshold, right otherwise. Finally, each leaf is associated with a class, which is the output of the predictor. We make use of a concept called Gini Index in order to quantify the purity of a node. The Gini Index of a given node is given by :

$$G = 1 - \sum_{k=1}^{n} {p_k}^2$$

A node is pure (G = 0) if all its samples belong to the same class, while a node with many samples from many different classes will have a Gini Index closer to 1. The algorithm we employ to train our Decision Tree is a recursive algorithm called CART, or Classification And Regression Trees. Each node is split so that the average of the Gini Index of its children weighted by their size, is minimized.

A. Choosing The Optimal Maximum Depth

The algorithm stops when the maximum depth, a hyperparameter, is reached, or when no split can lead to two children purer than their parent. To choose the optimal value for the maximum depth hyperparameter, we compared the training time and average of validation accuracy, precision, recall and f1 scores for various values of maximum depth across all catalogues.

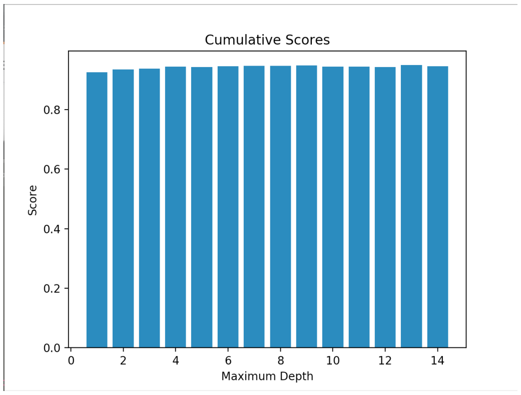

Fig. 1. Relationship between Scores and Depth

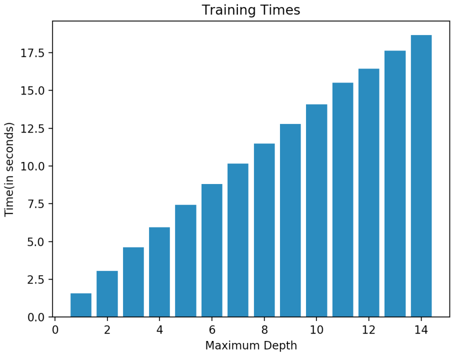

Fig. 2. Relationship between Training Time and Depth

It should be noted, that, in Fig. 1, Score refers to the average of the accuracy, precision, recall, and F1 score metrics, over all 4 catalogues in the dataset. As can be observed in the graphs, the Score plateaus when the value of the maximum depth hyperparameter is 8, while the training time keeps on on increasing. Due to this reason, we decided that setting the value of the maximum depth hyperparameter to 8 was optimal for our classification task.

IV. Results

TABLE I
Cross Validation Results

Catalogue	Accuracy	Precision	Recall	F1-Score	Training Time(In seconds)
1	0.9505	0.9809	0.9648	0.9725	9.4159
2	0.9391	0.9598	0.9704	0.9650	47.5209
3	0.9478	0.9634	0.9776	0.9704	54.3841
4	0.8504	0.8821	0.8711	0.8764	413.3804

We performed 10-fold cross validation on our model to check our results. The results obtained are summarized in Table I. The value of k=10 for K-fold cross validation was obtained from Kohavi[2].

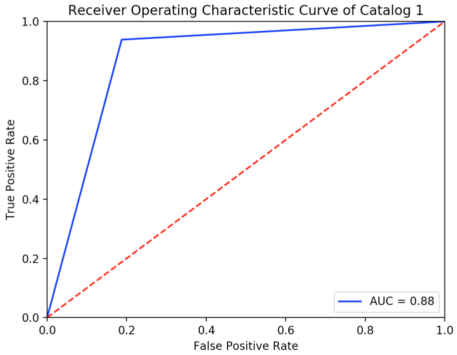

Fig. 3. ROC Curve of Catalog 1

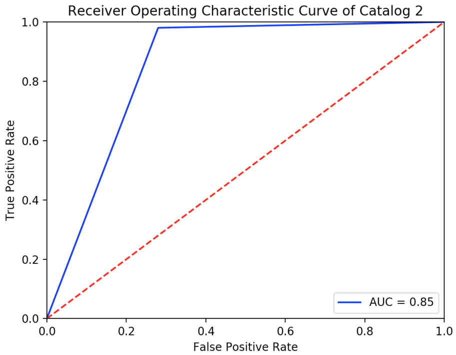

Fig. 4. ROC Curve of Catalog 2

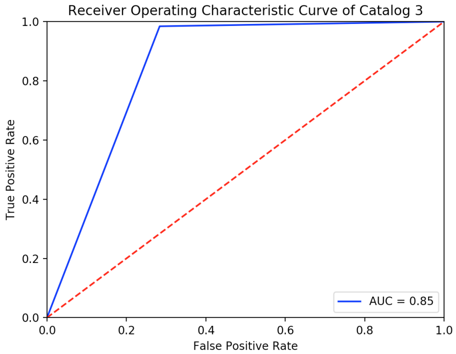

Fig. 5. ROC Curve of Catalog 3

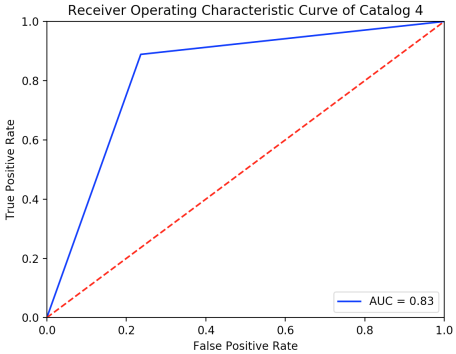

Fig. 6. ROC Curve of Catalog 4

Reciever Operating Characteristic curves for each of catalogues can be observed in Fig. 3-Fig. 6.

V. Conclusion

We were able to make use of a Decision Tree, with the Classification and Regression Trees algorithm for training our model, and achieve accuracy above 93 percent for catalogues 1, 2 and 3. We limited the maximum depth of our tree to 8, as we observed that there was no significant increase in performance metrics beyond this value. Thus, we were able to achieve the specifications mentioned in the problem statement.

References

[1] Simran Makhija, Snehanshu Saha, Suryoday Basak, Mousumi Das, Separating Stars from Quasars: Machine Learning Investigation Using Photometric Data

[2] Ron Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection