Classification of Stars and Quasars

Kishan K P
B.Tech, CSE Dept.
PES University
Bangalore, India
kishankp9@gmail.com

Madhav Agal
B.Tech, CSE Dept.
PES University
Bangalore, India
madhav.agal1d12@gmail.com

Ashwath Janardhanan
B.Tech, CSE Dept.
PES University
Bangalore, India
ashwath.j99@gmail.com

***Abstract*—The problem at hand is to classify Stars and Quasars as two separate entities. There are many difficulties involved in doing this as both of them seem to exhibit similarities and there is no exact boundary between the two. The paper uses Decision Trees to classify the two entities. The accuracy and other performance metrics seem to be in the range of 85-95%.**

***Index Terms*—Decision Trees, Gini Index, Information Gain**

I. Introduction

The classification of Stars and Quasars as two separate classes is a complicated problem as they exhibit similarities in their optical morphology. With the help of Machine Learning techniques we can classify them into two separate classes given their data. Stars are astronomical objects consisting of a luminous spheroid of plasma held by its own gravity. They are easily recognisable by us. Quasars are extremely large galactic bodies. They derive their energy from black holes.

For the problem, supervised machine learning technique such as Decision Trees was used to classify Stars and Quasars. The data had four catalogs and the model was applied on all four of them. The performance metrics are also displayed. Techniques such as K-Means clustering was also applied. However, Decision Trees was observed to show a better accuracy and this model was chosen.

II. Problem Statement

The problem given, is to classify Stars and Quasars as two separate classes. These two classes are very similar in their optimal images. There is no photometric label which tells us whether it is a star or a quasar. The major differentiating factor is their UV emissions. The ultraviolet and the photometric data were combined to differentiate stars and quasars. The data was obtained from the GALEX telescope and the SDSS. The data was obtained for the folllowing regions.

- North Galactic Region: Data from this region greater than 75° was used.
- Equatorial Region: Data from this region was selected in the range of $-30°$ to $30°$.

III. Techniques

On observing the data given, it was observed that a few columns were unnecessary for training the model. Columns such as galex objid, sdss objid and the predicted values is of no use in training the model and hence they were dropped. Spectrometric redshift is used to test the model based on a certain threshold and is not used for training the data. The dataset consists of mostly contiguous variables. Building a Decision Tree when there are an enormous amount of such variables would take a huge amount of time. To overcome this, equal size bins were constructed. This makes the values discrete and decreases the computational time of building the tree manifold. The number of quasars in the dataset exceeds the number of stars. Training the model without upsampling would result in a very skewed outcome. Hence, we upsample the stars in our training set so that it has an equal number of stars and quasars. The model is then fit on the new training set and is then tested.

- Decision Trees
  Decision Tree learning is one of the most widely used and practical methods for inductive reference. It is a method for approximating discrete-valued functions that is robust to noisy data. It can handle both numerical and categorical data and also take care of multi-output problems. There are various decision tree algorithms namely ID3(Iterative Dichotomiser 3), CART(Classification and Regression Tree), CHAID(Chi-Squared Automatic Interaction Detector) and MARS. The CART model is used to do our classification.
  The model is applied on the training data to train our model. We calculate the information gain of each attribute and also the gini index. The split in the tree is performed at the attribute with the highest information gain. The gini index is used to measure the impurity of the data sample or the set of training tuples. We stop the recursion once the gain becomes zero. All the tuples at this stage would belong to one class.
  Decision Trees also handle noisy data very well. They are widely used in classifying data. Decision Trees are also very beneficial with imbalanced data and they perform frequently well on them as well. They work by learning a hierarchy of if/else questions and can force both classes to be addressed. Decision Trees have performed well and can also be considered as a good model for the problem at hand.
  The Gini Index in a data sample D is calculated as:

$$\mathit{Gini(D)} = 1 - \sum_{i=1}^{m} p_i^2 \qquad (1)$$

where p_i is the nonzero probability that a tuple in D belongs to class C. The Information needed to classify a tuple in a data sample D is given as:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{2}$$

where p_i is the nonzero probability that a tuple in D belongs to class C.
The Gain on an attribute A is measured as:

$$Gain(A) = Info(D) - Info_A(D) \tag{3}$$

IV. Conclusion

The model was tested and the performance metrics was recorded classwise. It was also tested on different split ratios. The AUC-ROC graph was also plotted for all catalogs and the average area was observed to be 0.86. The AUC-ROC curve for Catalog 1 is shown:

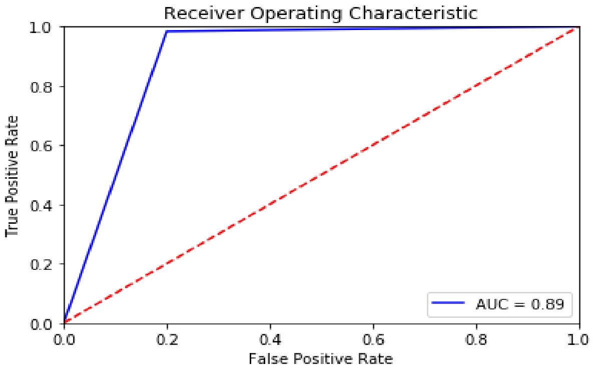

Fig. 1. AUC-ROC Curve for Catalog 1

The accuracy vs split ratio curve for various split ratios is also shown below:

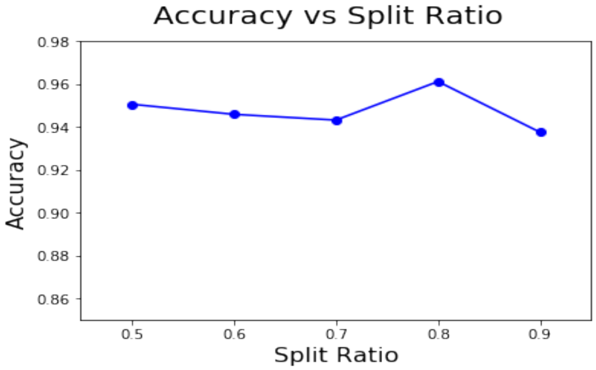

Fig. 2. Accuracy vs Split Ratio for Catalog 1

The graph for the Information Gain vs the number of iterations for the entire tree is shown:

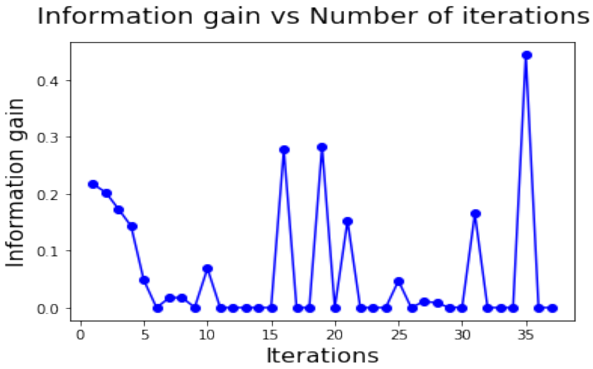

Fig. 3. Information Gain vs Iterations for Catalog 1

It can be seen that, every time the information gain becomes zero, a leaf node is encountered. At points where the information gain is non zero, a decision node is present there, which branches into two child nodes.

On testing the model through for various split ratios, an average accuracy of 95% is observed. This can be verified from the graph. The results were also verified with the spectrometric redshift data. It is known that values which have z less than or equal to 0.0033 are predominantly stars and z greater than or equal to 0.004 are quasars. The objects which fall in between this range represents an overlap in this template. Confusion matrices were also made for these observations.

The classwise performance metrics for stars and quasars are displayed in the tables below:

TABLE I
Class - STARS

Catalog	Precision	Recall	F-Score
1	0.85	0.85	0.85
2	0.79	0.64	0.71
3	0.80	0.755	0.79
4	0.77	0.76	0.76

TABLE II
Class - QUASARS

Catalog	Precision	Recall	F-Score
1	0.96	0.97	0.97
2	0.97	0.96	0.96
3	0.96	0.97	0.96
4	0.89	0.90	0.90

An Average accuracy of 92% is observed using Decision Trees.

V. Github Repository

https://github.com/kishankp9/Classification-of-Stars-and-Quasars

REFERENCES

[1] Simran Makhija, Snehanshu Saha, Suryoday Basak, Mousumi Das ”Separating Stars from Quasars: Machine Learning Investigation using Photometric Data”, April, 2019.

[2] Tom M. Mitchell, ”Machine Learning”, Indian edition, McGrawHill Education.

[3] Jiawei Han, Micheline Kamber, Jian Pei, ”Data Mining - Concepts and Techniques” , 3rd Edition, Morgan Kauffman.