Classification of Stars and Quasars

Sanjana Moudgalya
B.Tech, CSE Dept.
PES University
Bangalore, India
PES1201700297

Anusha S Rao
B.Tech, CSE Dept.
PES University
Bangalore, India
PES1201701536

Aninditha Ramesh
B.Tech, CSE Dept.
PES University
Bangalore, India
PES1201700219

I. Introduction

Stars are the most easily recognized objects in the sky. They are self-illuminating bodies that create energy through nuclear fusions. On the other hand, "Quasi-stellar radio source" also known as Quasars are one of the brightest and most distant objects in the universe, which are visible because of the material they pull in. They are powered by black holes billions of times as massive as our sun, and have fascinated astronomers since their discovery.

This project aims on classifying the data into stars and quasars. The photo-metric data had different catalogs.Among the four catalogs present, the data set related to North Galactic pole was used for classification. In the later stages, the model was applied on the other catalogs. Supervised machine learning techniques such as Bayesian classifier and K nearest neighbours(KNN) were used for classification. The accuracies are presented based on the performance metrics. And hence, after comparing the two approaches, it was observed that KNN gave a better accuracy.

II. Problem Statement

The aim was, given photo-metric optical data, classify them as stars and quasars. Stars and Quasars are similar in the optical images. The differentiating factor however was their UV emissions. Thus, a combination of optical and UV photo-metric data was used for classification. The data set was divided into four catalogs based on the region of observation.

- North Galactic Region: Data from the region greater than 75○ of galactic latitude is used.
- Equatorial Region: Data selected in the range of 30○ dec to 30○ dec.
- From both regions: The samples that have fuv(far UV).
- From both regions: The samples that do not have fuv(far UV).

III. Methodology

The data set was already cleaned and hence no further cleaning was required. After analysing the data, columns that represented Index, Galaxy object Id and Sdss Id (Sloan Digital Sky Survey Id) were dropped as they were unique for each data point. Spectrometric redshift is very high for quasars when compared with stars. There is a significant difference and it is a distinguishable factor for classifying stars and quasars. In order to reduce the innate correlation, spectrometric redshift and predicted class labels were also dropped. It was noticed that the data was highly skewed, i.e the number of quasars were very high when compared to the number of stars. Hence the training data was up-sampled in order to balance the classes, after splitting data into train and test to fit the model.

- Bayesian Model
  Bayes classification technique is a supervised learning algorithm. Bayesian machine learning allows us to encode our prior beliefs about what models should look like, independent of what the data depicts. This is especially useful when there is not enough data to confidently learn the model.
  Applying this model to the given data set, the training data was divided class-wise. Mean and standard deviation of each attribute was calculated to give an idea of the probability distribution. Since the attributes were continuous, a Gaussian probability distribution function was used to calculate the probability of attribute value associated with each data point.

  $$\text{PDF} = f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

  Based on the assumption that the attributes are independent, the individual probabilities and assigned to classes. The class with the maximum probability was concluded as the predicted class.
- K nearest neighbours
  K-Nearest Neighbors is one of the most basic classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition and data mining. Here, k is an arbitrary number using which the data is classified after considering the k nearest points, which is preferably an odd number. After a brief analysis, k was chosen to be 5. As each data point was considered to be a n-dimensional vector, five nearest neighbours were calculated using the euclidean distance.

  $$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + \cdots + (q_n-p_n)^2}$$

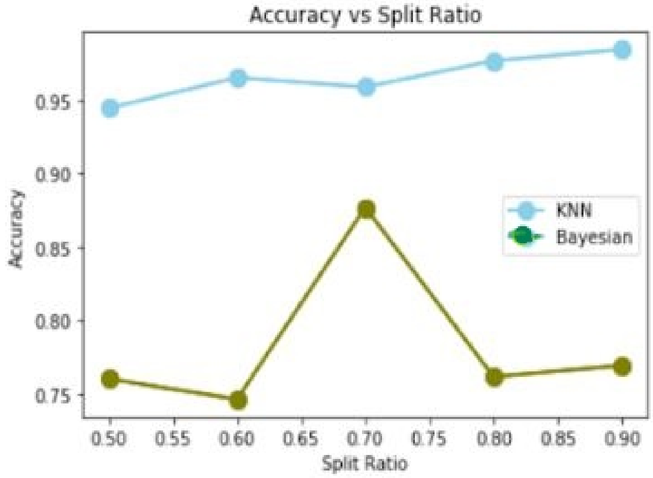
  $$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

The points were then assigned labels based on the majority vote obtained.

IV. Conclusions

Comparing the actual class labels with the predicted labels, a confusion matrix was constructed to find the accuracy measures.

Fig. 1. KNN vs Bayesian



The model was tested for various train-test split ratios. To have a fair comparison, the accuracy was reported class-wise.The average class-wise accuracy of KNN was reported to be 96.61%. On the other hand the accuracy of the Bayesian was reported to be 78.63%.

A general difference between KNN and other models is the large real time computation needed by KNN compared to others. Even though KNN is slower, comparison of the two models showed that KNN gave a slightly better accuracy.Hence KNN was tested on the other two catalogs and the average accuracy was still found to be more than 92%.

$Split\ Ratio$	$Precision$	$Recall$	$Accuracy$
90-10	100	98.33	98.46
80-20	99.12	98.26	97.69
70-30	98.28	97.17	95.89
60-40	98.72	97.48	96.54
50-50	97.25	96.58	94.46

TABLE I
CLASS-QUASARS

$Split\ Ratio$	$Precision$	$Recall$	$Accuracy$
90-10	83.33	100	98.46
80-20	87.5	93.33	97.69
70-30	75	83.33	95.89
60-40	76	86.36	96.54
50-50	70.59	75	94.46

TABLE II
CLASS-STARS

The predicted classes were also cross verified based on spectrometric redshift values. Two thresholds were used to do this-
1. If redshift lesser than 0.004, class-stars
Else class-quasars
2. If redshift lesser than 0.0033, class-stars
Else class-quasars

For an 80-20 split using KNN,
Accuracy = 96.9%

References

[1] Snehanshu Saha, Suryoday Basak, Archana Mathur, Rahul Yedida. ”Machine Learning AstroInformatics.”

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

[3] https://scikit-learn.org/stable/modules/neighbors.html

[4] https://scikit-learn.org/stable/modules/generated/
sklearn.naive$_bayes.GaussianNB.html$