Separating Stars from Quasars: Machine Learning Investigation Using Photometric Data

November 2019

NISHA NEMASING RATHOD PES1201701672
BRITNEY D MUNGRA PES1201701801
MANASA H K PES1201701886

1 Abstract

We explore the effectivess of knn 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 knn-classifier that appropriately classify photometric samples without associated spectroscopic labels.

2 Introduction

The data set consits of observed astronomical sources in the far-UV and near-UV (FUV and NUV) wavebands,wave bands of u,g,r,i,z and obtained the spectra of the sources,redshift,class and the pred values.
Also the dataset is divided into various catalogs:
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.
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.
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.

3 Method used for classification

KNN- k Nearest Neighbours is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.
Steps Followed were to load the dataset first ,remove the unwanted columns like galex-objid ,sdss-objid ,spectrometric-redshift.
Then we split the dataset into train and test using folds. then we give different values for k and number of folds to get optimal output.
Then we find the euclidean distance between each of the data points in that row to the corresponding values from training set,we then sort the distances in ascending order.
Then we take the first k entries and classify based on the maximum number of columns having the same classification.
Then the accuracy is obtained using accuracy function

4 Results

We are considering the effectiveness of knn based on the accuracy's obtained

- The accuracy obtained for catalog 1 with 30 folds and 5 nearest neighbours is giving the highest accuracy of 96.508
- All the catalogs are giving us accuracy above 90