Machine Leaning Project - UE17CS303 Separating Stars from Quasars: Machine Learning Investigation Using Photometric Data

TEAM:

Keerthi Priya P	PES1201701643
Archana C	PES1201701384
Thanusha Sai Melavoy	PES1201701457

1 ABSTRACT

The problem statement of our project is to classify matched sources in the Galex (Galaxy Evolution Explorer) and SDSS (Sloan Digital Sky Survey) spectroscopic objects over the North Galactic Region and Equatorial Region into stars and quasars based on color-color plots.The challenge here is there is no clear linear/non-linear boundary that separates the two entities. Hence after exploring all the classification techniques in machine learning,we chose to implement K nearest neighbours algorithm to classify the spectroscopic objects into Stars and Quasars and the efficiency along with the results are interpreted.

2 INTRODUCTION

Quasars are star like point sources with immense energy and much greater red shifts, whereas the stars are self illuminating celestial objects with energy outputs much less than that of quasars.Both of them have compact optical morphology which is a major challenge in distinguishing them.Hence using machine learning approach we have classified the spectroscopic objects into stars and quasars using optical photometric data and UV data observed using SDSS and Galex.

2.1 Brief Abstract on Machine Learning

Machine Learning is the field of study that gives systems the capability to learn from its experiences without being explicitly programmed.Machine Learning techniques are classified into three namely:

- **Supervised learning:** Labeled data is used to train the model.Examples of supervised learning are Regression,Support vector machines,Decision trees,K-nearest neighbours,etc

- **Unsupervised learning:** Unlabeled data is used to train the model.Examples of unsupervised learning K-means clustering,Apriori algorithm

In this project we have used supervised learning algorithm namely K-nearest neighbours which is appropriate for binary classification problems when there is little or no prior knowledge about the distribution data.

2.2 K Nearest Neighbours Approach

2.2.1 Introduction

KNN is a non-parametric and a lazy learning algorithm and is based on feature similarity approach.In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor.The number of neighbors(K) in KNN is the contolling parameter that decides the classification of the given data point.In case k is small,the noise will have a higher influence on the result, and if k is large it may become computationally expensive.Hence choosing k value is a major challenge for any given dataset. KNN is said to be a lazy algorithm because it does not use the training data points to do any generalization. In other words, there is no explicit training phase or it is very minimal.Hence, the training phase is pretty fast.The testing phase of K-nearest neighbor classification is slower and costlier in terms of time and memory.KNN requires scaling of data and hence choosing an appropriate distance method also plays a major role.And an another disadvantage is the features with high magnitudes will weight more than features with low magnitudes.And KNN is not suitable for large dimensional data.
Considering the pros and cons of Knn,we chose Knn because it is more appropriate for the given dataset as it consists many instances and few dimensions.

2.2.2 KNN Algorithm

Steps to perform KNN Algorithm are.
1)Classify the dataset into training set and test set.
2)Determine the parameter k=number of nearest neighbours.
3)Choose the distance method.
4)Calculate the distance between the query instance and training data.
5) Sort the distance and determine the nearest neighbours based on the K-th minimum distance method.
6)Generate the category of the nearest neighbours.
7)Use the majority of the category of the nearest neighbours and assign the category as the prediction value to the test data instance.

3 IMPLEMENTATION

3.1 Data and problem statement

The dataset provided consists of optical and photometric data features.We have been provided with four different catalogs.Each catalog has the following features namely

optical bands from SDSS namely u, g, r, i and z and far-UV and near-UV (FUV and NUV) wavebands' magnitudes.Each catalog has different specifications namely.

- **Catalog 1:** North Galactic Region Only
  The dataset contains samples that have fuv.
- **Catalog 2:** Equatorial Region Only
  The dataset contains samples that have fuv
- **Catalog 3:** North Galactic Region and Equatorial Region Combined
  This dataset also contains samples that have fuv
- **Catalog 4:** Removed fuv and fuv-related features
  This dataset contains all the samples even with the samples that don't have fuv.

We have applied K-NN Algorithm to all the above catalogs.We chose K=3,7,10 as the number of nearest neighbours so as to find the optimal k value with better accuracy measures.We have split 70% of the data as the training data for catalog 1, 80% for catalog 2 and 3,90% for catalog 4.We have used Euclidean distance method as the scaling factor to classify the test instances.Since spectrometric redshift can categorize the data well,we dropped the feature and performed the classification.We performed cross validation on the dataset so that classifying instances would be more accurate.And the results are elucidated in further sections.

3.2 Results

After performing KNN on four different catalogs.The following are the results:

- We have tabulated Accuracy,Precision and F1 Score for 3,7,10 nearest neighbours on normal validation for all the catalogs.

	Accuracy		Precision		F1 Score	
K	3	7	3	7	3	7
Catalog 1	99.43%	96.9%	100%	100%	99.45%	96.88%
Catalog 2	91.55%	82.7%	98.4%	93.6%	90.9%	79.8%
Catalog 3	98.7%	97.7%	99.9%	99.2%	98.8%	97.8%
Catalog 4	99.4%	99.06%	99.8%	99.68%	99.4%	99.06%

- We have tabulated Accuracy,Precision and F1 Score for 3,7,10 nearest neighbours where we have used catalog 3 as training data and catalog1,catalog 2 and catalog 4 as test data.

Catalog 3	Accuracy		Precision		F1 Score	
K	3	5	3	5	3	5
Catalog 1	99.7%	99.5%	100%	100%	99.7%	99.5%
Catalog 2	99.675%	99.3%	100%	99.9%	99.6%	99.3%
Catalog 4	88.3%	89.5%	83.1%	85.7%	89.2%	90.1%

• We have performed 5 fold cross validation on catalog 3 and recorded the following observation.

Catalog 3	Accuracy		Precision		F1 Score	
K	3	7	3	7	3	7
Catalog 1	79.3%	84%	71%	76.4%	82.7%	86.1%

4 CONCLUSION

· After performing KNN on all the four catalogs,we have observed how the number of nearest neighbours effects the accuracy.

· We have seen the computational costs and the efficiency of implementing KNN for the given dataset.We have observed that with decrease in k,the noise influences the test instance.

· We have observed that there are many discrepancies with the data that is many instances were classified differently in different catalog.

· We observed that the optimal values for k are 3, 7 and 5 as there is a good accuracy when measured with differing k values.But having a very high accuracy would overfit the model.