Classification of stars and quasars

Meishty Pande
Computer Science
PES University Bengaluru, India
meishtypande@gmail.com

Sampreetha V
Computer Science
PES University Bengaluru, India
sampreethav18@gmail.com

Samhitha D
Computer Science
PES University Bengaluru, India
samhithadinesh13@gmail.com

Abstract— Using Machine Learning for the task of Classifying stars and quasars from data obtained from Galex and SDSS photometric data. We have implemented KNN algorithm with variation in number of neighbours and training -testing size ratio.

I. PROBLEM STATEMENT

Classification of Photometric data into stars and quasars.

We have been provided with Galex (GalaxyEvolution Explorer) and SDSS (Sloan Digital Sky Survey) catalogues on which we have to find a suitable method of classifying the data into stars and quasars. This is quite a challenge because there is no clear linear/non-linear boundary separates the two entities. Diversity and volume of samples add to the complexity as well. To evaluate the correctness of the classifiers, we report the accuracy and other performance metrics.

From the base paper we referred to, we know that the infrared data or UV data with optical photometry results in an efficient separation.

In the database, there are no photometric labels that tell us if a source is a quasar or a star – this information can be retrieved only from the spectroscopic data. Thus, the problem that we're trying to primarily address is to classify photometric data into spectroscopic classes.

The photometric optical data is composed of images of the sources made using the optical filters: u,g,r,i,z in the SDSS survey. The UV data is from the GALEX survey, which is an all-sky in the far- UV (FUV) and near-UV(NUV) wavebands. The difference between FUV and NUV magnitudes is a measure of dust extinction, which is relatively larger for Quasars, and hence we use it to test for linear separability.

II. METHOD OF CLASSIFICATION

We are employing KNN(k-nearest neighbours) which is a supervised learning algorithm used for classification. Here is a brief overview of how we implemented it on Catalogue 2- "cat2.csv" file, Catalogue 1- "cat1.csv" file.

1. First, we load the data
2. We initialize K to a certain number of neighbours. We have used k ranging from 1 to 11 to find out the optimal value.
3. For each example in the data we calculate the Euclidean distance between the query example and the current example.

4.Then we add the distance and the index of the example to an ordered collection

5.We sort the ordered collection of distances and indices in ascending order (by the distances)

6.We pick the first K entries from the sorted collection 7.Then we get the labels of the selected K entries 8.Lastly, we return the mode of the K labels

III. RESULT AND ANALYSIS

We print the accuracy list for values starting from k=1 to 11 and hence find the optimal k value citing higher accuracy.

To interpret the performance measures,*Case 1: When the catalogue was 2 and file was cat2.csv*

For the ratio of training and test –

Training Size	Test Size	Optimum number of neighbors	Accuracy
85.0	15.0	8	94.3327239488117%
80.0	20.0	5	96.15384615384616
70.0	30.0	4	95.89743589743590 %
65.0	35.0	4	96.49122807017544 %
60.0	40.0	11	96.15384615384616 %
50.0	50.0	7 or 9	95.38461538461539%
40.0	60.0	9	95.64102564102565 %

Case 2: When the catalogue was 1 and file was cat1.csv

For the ratio of training and test –

Training Size	Test Size	Optimum number of neighbors	Accuracy
85.0	15.0	7	96.93877551020408%
80.0	20.0	9 or 10	96.15384615384616
70.0	30.0	7 or 9	95.89743589743590 %
65.0	35.0	7	96.49122807017544 %
60.0	40.0	8	96.15384615384616 %
50.0	50.0	3	95.38461538461539%
40.0	60.0	3	95.64102564102565 %

IV. CONCLUSIONS

For all even values of k, the method would not be optimal as it is advisable to take odd values for binary classification to avoid the ties i.e. two classes labels achieving the same score. Normally we take a 70:30 ratio for training to test but here 80:20 seems more fitting.

Therefore, after examining the above results, we can conclude that :

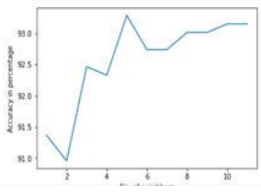
In the cat2.csv ,for the ratio of training and test 80:20, with k-value 5 and accuracy 93.28767123287672% will be a good classification model for our dataset.

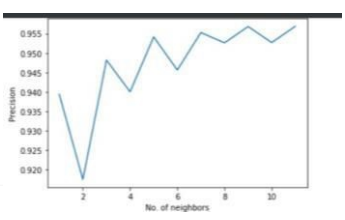
In the cat1.csv ,for the ratio of training and test 85.0 : 15.0, with k- value 7 and accuracy 96.93877551020408% will be a good classification model for our dataset.

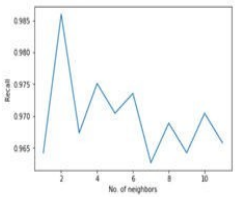
GRAPHS:

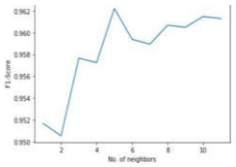
For Catalogue 2, cat2.csv:

A] For ratio of training and test is 80.0 : 20.0, k=5 is the optimal as the accuracy is 93.28767123287672%





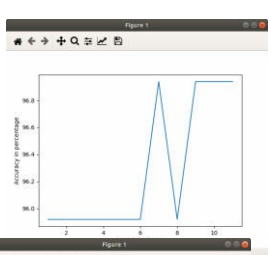


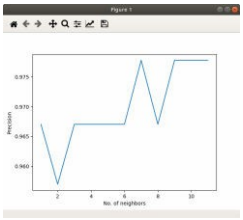


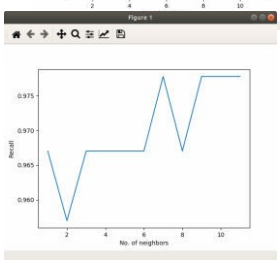
For Catalogue 1, cat1.csv:

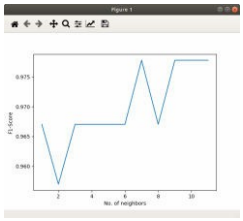
B] For the ratio of training and test 85.0 : 15.0, with kvalue 7 and accuracy

96.93877551020408%









https://github.com/SampreethaV/MLPROJECTKNN

V. ACKNOWLEDGMENTS AND REFERENCES

Dr. Snehanshu Saha, our guide and mentor for this project.

Simran Makhija, Snehanshu Saha, Suryoday Basak,

Mousumi Das. Separating Stars from Quasars: Machine Learning Investigation Using Photometric Data.

Machine Learning-Tom Mitchell Indian Edition

McGrawHill