Classification of Stars and Quasars using Weighted KNN

Kumar Abhishek
PES1201700251
CSE, PES University
kumar98abhishek@gmail.com

Abhishek M
PES1201701563
CSE, PES University
abhishek574abhi@gmail.com

Manvith J
PES1201701774
CSE, PES University
manvithj99@gmail.com

github link for code :
https://github.com/Abhi-63/ML_KNN_project

***Abstract*— Quasars and stars are extremely luminous bodies found in several galaxies. But, how does one really distinguish between the two? Well, there are multiple features that do set them apart. Reporting these features with accuracy is quite a task. Reliability of the dataset is questionable. After much cleaning and imputations, a more reliable set has been put forth to classify these objects. This paper will show how using a Weighted KNN model the classification can be done.**

Keywords—classification, weighted knn, stars, quasars

I. Introduction

Stars and quasars are celestial objects which have very little visual difference when observed from our planet.A star consists of a luminous spheroid of plasma held together by its own gravity.A quasar is an extremely luminous active galactic nucleus in which a supermassive black hole is surrounded by a gaseous accretion disk. They seem to be indistinguishable when merely observed however they differ in many aspects, especially their photometric features.

Stars and quasars look very similar in their optical images but the spectral energy distribution for stars and quasars is different and so the optical bands from SDSS namely u, g, r, i and z can be used to separate them. The vast differentiating factor is their UV emissions.

In this project both optical and ultraviolet (UV) photometric data is used with machine learning methods (KNN) to discriminate between stars and quasars. The spectroscopic labels are used as the primary class label. Both stars and quasars have a compact optical morphology and are hence difficult to separate without spectroscopic data. In such cases, other parameters of the sources such as their optical variability or their optical colors are necessary to distinguish between them. Further research has shown that including the infrared data or UV data with optical photometry results in a more efficient separation.

Using data from GALEX, cross-matching with labels in SDSS, provided data from both the visible and ultraviolet spectra. Each of the photometric samples chosen in either region (north galactic or equatorial) have an associated spectroscopic label from the SDSS database.

II. Dataset and Problem Description

Considering the problem statement, to classify photometric data collected from the Galaxy Evolution Explorer (GALEX) and the Sloan Digital Sky Survey (SDSS) over the North Galactic region and Equatorial region in to spectroscopic classes of Stars and Quasars. K Nearest Neighbours Machine Learning algorithm is used to successfully distinguish between Stars and Quasars.Inferences about the dataset and the results of the KNN model have been elucidated.

III. Methodology

Machine learning uses algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. Machine learning techniques are broadly classified into 2 types - Supervised and Unsupervised

1. Supervised: All data is labeled and the algorithms learn to predict the output from the input data.

2. Unsupervised: All data is unlabeled and the algorithms learn to inherent structure from the input data

K Nearest NeighbourAlgorithm

Euclidean distance: If x_i and x_j are two instances and $a_r(x)$ denotes the value of the r^{th} attribute of instance x, then

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

Advantages

1. The algorithm is simple and easy to implement.

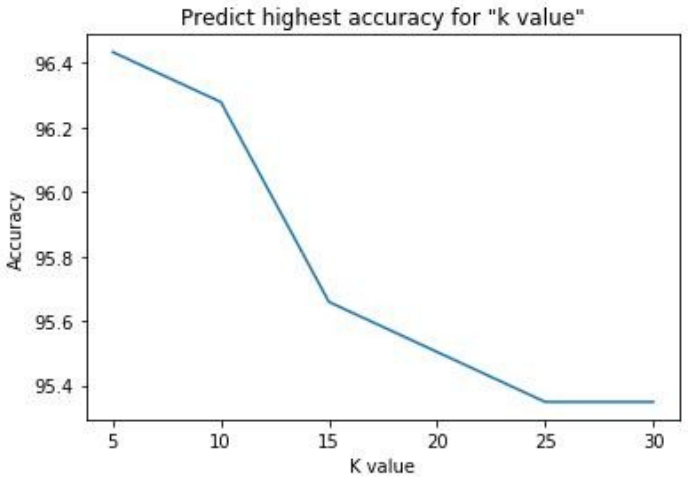
2. There's no need to build a model, tune several parameters, or make additional assumptions.

3. The algorithm is versatile. It can be used for classification, regression, and search (as we will see in the next section).

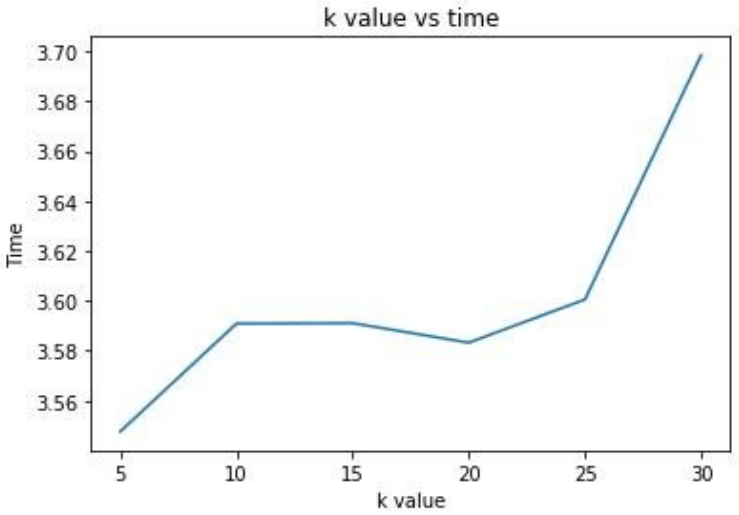
Disadvantages

1. The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase

IV. ANALYSIS

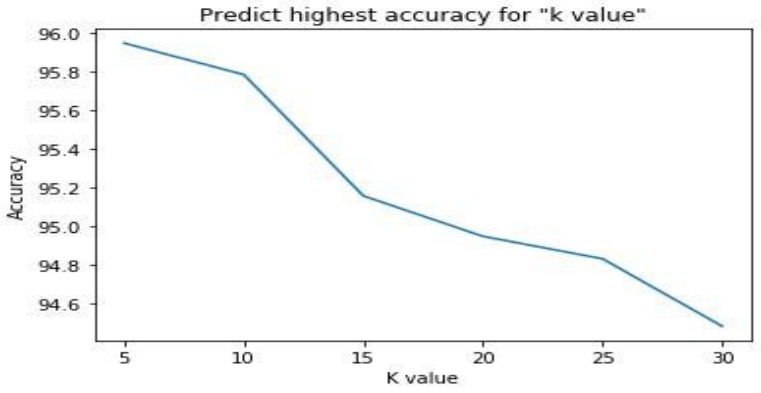


cat-1 (figure-A)

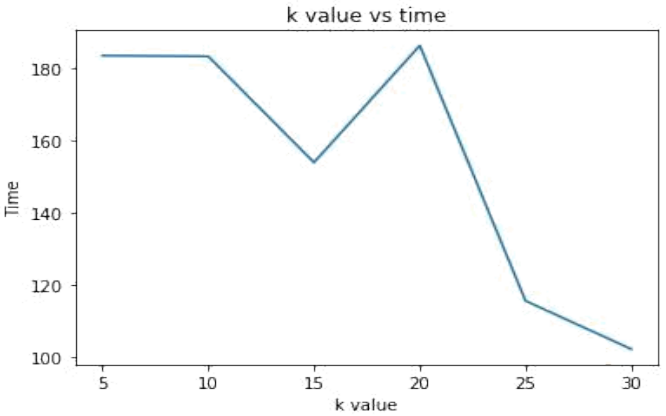


cat-1 (figure-B)

Conclusion from cat-1 figure A and figure B:
-Accuracy vs k value gives the highest highest accuracy w.r.t k value
-If data is large the time taken to calculate the knn will be more for the k-value.

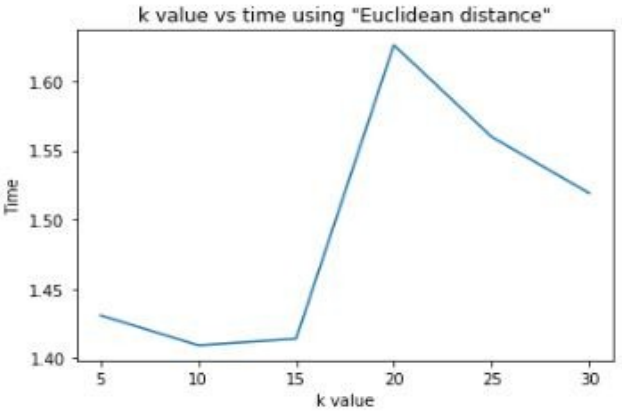


cat-2 (figure-A)

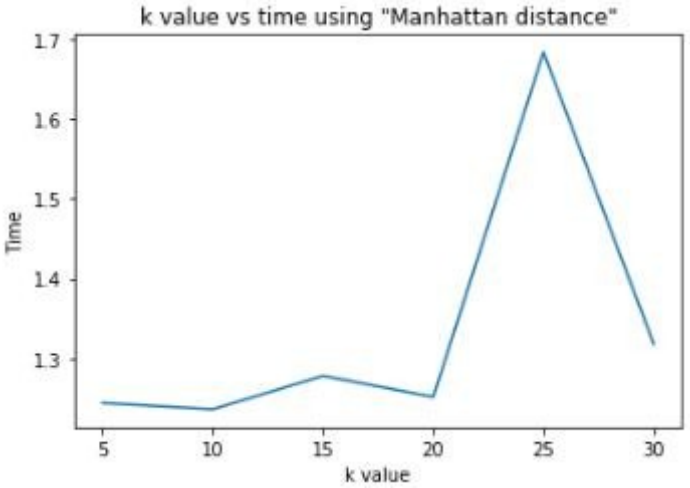


cat-2 (figure-B)

Conclusion from cat-1 figure A and figure B:
-Accuracy vs k value gives the highest highest accuracy w.r.t value
-If data is large the time taken to calculate the knn will be more for the k-value.



cat-3 (figure-A)



cat-3 (figure-B)

Conclusion from cat-3 figure A and figure B:

IV. CONCLUSION

- All the catalogs are giving us accuracy above 90
- Accuracy vs k value gives the highest highest accuracy w.r.t k value
-If data is large the time taken to calculate the knn will be more for the k-value.
- Time taken to calculate accuracy by using the Manhattan Distance runs efficiently for some values of k-value.

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.