Photometric Identification of Compact Galaxies, Stars and Quasars using Machine Learning and Deep Learning Algorithms

Project Report

Presented by :
Nakka Vishnu Vardhan

Project Guide :
Dr.Sumohana S Channappayya

Date:
July 10, 2023

Department of Electrical Engineering
IIT Hyderabad

Declaration

I hereby declare that this project report titled "Photometric Identification of Compact Galaxies, Stars and Quasars using Machine Learning and Deep Learning algorithms" was done with the results that I have obtained during the project and some words were taken from the research papers. I have clearly referenced all sources used in the report.

Nakka Vishnu Vardhan

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Sumohana S Channappayya, for his guidance and support throughout this project. Their expertise and valuable insights have been instrumental in the successful completion of my work in this project.

I would also like to thank the Department of Electrical Engineering at IIT Hyderabad for providing the necessary resources and facilities to conduct this study.

Furthermore, I would be grateful to all the fellow LFOVIA lab members which include PhD Scholars and M.Tech Students who guided and helped me throughout the period of internship. I would specially mention Srinadh sir for guiding me and gave me this opportunity to work in this project.

Nakka Vishnu Vardhan

Abstract

Deep Learning Algorithms are used to classify the compact galaxies, stars and quasars from the Sloan Digital Sky Survey (SDSS) Data Release 16 (DR16). The current dataset consists of 2,40,000 compact objects and 1,50,000 faint objects. The model is trained in a way so that it reduces human error and also helps in reducing the heavy machinery in classifying the astronomical images. The current architecture is taken from MargNet, a deep learning classifier and implemented in the Pytorch framework.

Contents

Declaration **1**

Acknowledgements **2**

Abstract **3**

1 Introduction **6**
1.1 Background . . . 6
1.2 Objective . . . 6
1.3 Significance . . . 6

2 Dataset **8**
2.1 Image Data . . . 8
2.2 Photometric Features . . . 8
2.3 PreProcessing . . . 8
2.3.1 Compactness Parameter . . . 8
2.3.2 Magnitude . . . 8
2.4 Train Validation Test Split . . . 9
2.4.1 Experiment1 . . . 9
2.4.2 Experiment2 . . . 9
2.4.3 Experiment 3 . . . 9

3 TSNE PLOT **10**
3.1 Procedure . . . 10

4 Machine Learning Algorithms **11**
4.1 Support Vector Machines . . . 11
4.2 Naive Bayes . . . 11
4.3 Decision Trees . . . 11
4.4 Random Forests . . . 12

5 Deep Learning Algorithms **13**
5.1 Artificial Neural Network . . . 13
5.2 Convolutional Neural Network . . . 13

6 Results **14**
6.1 tSNE Plots . . . 14
6.2 ML algorithms . . . 14
6.2.1 Support Vector Machines . . . 14
6.2.2 Naive Bayes . . . 14
6.2.3 Decision Tree . . . 15
6.2.4 Random Forests . . . 15

6.3 DL algorithms 15
6.3.1 Artificial Neural Network 15

7 Observations **16**

8 Conclusion and Future Scope **17**

References **18**

Chapter 1

Introduction

1.1 Background

Sloan Digital Sky Survey is a large-scale survey that collects photometric and spectrometric data on millions of astronomical objects such as stars, galaxies, and quasars. More than 15 terabytes of data are collected every night, making manual analysis a significant burden and rendering data classification nearly impossible. Machine Learning has become increasingly popular in this computerized era, reducing human effort in many areas. It refers to any algorithm through which a model is trained on existing data and used to work on unseen data. Deep Learning, a subfield of Machine Learning, builds models with neural networks that mimic the functioning of our brain. This specific domain has revolutionized audio and image processing.

In survey images, stars and quasars are referred to as point sources convolved with the Point Spread Function (PSF). Their distinction can only be made through their spectra, which requires large machinery. Machine learning algorithms typically classify objects based on color, but high redshift quasars and low redshift quasars have different colors while having similar spectra. Distinguishing galaxies from stars based on morphology fails when galaxies are highly compacted. Previous attempts using machine learning algorithms failed when the LSST obtained images that were deeper and fainter. The signal-to-noise ratio plays a crucial role in these newer datasets. MargNet is the first successful deep classifier network that worked on the SDSS DR16 dataset, demonstrating high accuracy results. It consists of ensemble Deep Neural Network and Convolutional Neural Network architectures, implemented using the TensorFlow-Keras data framework.

1.2 Objective

The objective of this project is to implement the machine learning and deep learning algorithms especially MargNet architecture on the SDSS DR 16 dataset to train the models so that they can easily classify the images into three categories such as galaxies, stars and quasars without any human intervention. Implementation can be done using PyTorch deep learning framework.

1.3 Significance

Classifying the astronomical data is very crucial for further research works. This task is the basic step for the important studies in different regions of astronomy. Machine Learning models work better when they are trained with a wide variety of data.

We are implementing our models through the Pytorch framework. Pytorch uses dynamic computational graphs which makes it easier for debugging. Its syntax and APIs are similar to python which is highly user friendly. It facilitates easier prototyping. It integrates with the python libraries such as NumPy and SciPy. Many researchers prefer pytorch and is often used in cutting-edge research.

Chapter 2

Dataset

SDSS DR16 Photometric Data collected in five different bands namely u,g,r,i,z is widely used in our project. We have two types of data to build our models.

2.1 Image Data

Image data is extracted from the Flexible Image Transport System in each of the five different filters from the SDSS. It consists of five channels which represent the five bands u,g,r,i,z.

2.2 Photometric Features

The photometric features include deredx, deVRadx, psffwhmx, extinctionx, ug, gr, ri, iz which collectively 24 in number are listed in a csv life.

2.3 PreProcessing

Two different parameters are considered to manage the quality of images.

2.3.1 Compactness Parameter

Here we have used the average ratio of half light ratio to the full width at half the maximum of Point Spread Function in all the five different filters. The less the c value,the more the compactness. Here we have considered the boundary point of c as 0.5. It relates the compactness of the galaxies.

2.3.2 Magnitude

This parameter refers to the faintness of the objects. We have considered its boundary value as 20 because above that value the traditional algorithms fail. It should also be averaged for all the five filters.

```
c < 0.5 |  Compact source dataset | 80k samples from each class
c < 0.5 and mag > 20 | Faint and Compact Source Dataset | 50k samples from each class
```

2.4 Train Validation Test Split

From the research paper, we performed all the algorithms on the two different sets of data in three different experiments.

2.4.1 Experiment1

We have splitted the data in the ratio 6:1:1 for the training, validation and test sets. All the data is from Compact Source Dataset.

2.4.2 Experiment2

We have splitted the data in the ratio 8:1:1 for the training, validation and test sets. All the data is from Faint and Compact Source Dataset.

2.4.3 Experiment 3

We have splitted the sets in the ratio 6:1:1 for the training, validation and test sets but the training and validation sets are from Compact Source Dataset and test data is from Faint and Compact Source Dataset.

Chapter 3

TSNE PLOT

t Distributed Stochastic Neighbor Embedding(tSNE) is a popular dimensionality reduction technique commonly used for visualizing high dimensional data in a lower dimensional space. It is particularly to verify whether the data is linearly seperable or non linearly seperable.tSNE is a powerful visualization tool for revealing hidden patterns and structures in complex datasets.

3.1 Procedure

1. **1.** tSNE models a point being selected as a neighbor of another point in both higher and lower dimensions. It starts by calculating a pairwise similarity between all data points in high dimensional space using a Gaussian Kernel.

2. **2.** The algorithm tries to map higher dimensional data points into lower dimensional space while preserving the pairwise similarities.

3. **3.** It can be achieved by minimizing the divergence between the probability distribution of the original high dimensional and low dimensional. It uses gradient descent to minimize the divergence.

As we are using a large dataset it is not possible to show all the samples in a tSNE plot. Hence we compromised ourselves with 15000 samples from each of the three experiments and plotted the tSNE plot using the scikit learn module.

Chapter 4

Machine Learning Algorithms

Machine Learning is a part of Artificial Intelligence which makes machines learn automatically and improve from experience without being traditionally programmed. Supervised learning is a kind of machine learning in which we train the model with labelled data. Here we have used four most important ml algorithms to classify our images based on the photometrics features listed.

4.1 Support Vector Machines

This algorithm starts with low dimensional data and moves the data into higher dimension and finds the support vector classifier that separates the high dimensional data into groups. It includes the parameters such as kernel which helps in finding the relationships between the samples and also includes in finding the optimum Support Vector Classifier. Another Parameter is Penalty Parameter(C) which controls the tradeoff between smooth decision boundaries. Higher the value of C leads to less misclassification.

4.2 Naive Bayes

Naive Bayes is an algorithm based on Bayes' theorem which assumes that all the features are independent of each other. It is widely used in modern day classification problems. Naive refers that the features are conditionally independent when given a label.It estimates the prior probabilities for each class label and calculates the likelihood probabilities for each feature by conditioning with the given label. For an unseen new data it calculates the posterior probabilities so that the class with high posterior probability will be considered as an output. This algorithm is computationally efficient and can handle high dimensional data effectively.

4.3 Decision Trees

Decision Tree is an innovative and optimal machine learning algorithm which can be used for both classification and regression tasks. It consists of leaves, branches and roots. Based on the impurity level of a feature a leaf will be generated. The impurity can be quantified using Gini Impurity or Information Gain. To overcome the overfitting we can go through stopping techniques like maximum depth, and increase the minimum number of observations in the leaves. They are interpretable and have the ability to handle non linear relationships between features and classes. These are robust to outliers and have the capability to handle the missing values in the input data.

4.4 Random Forests

Decision Trees are easy to build but in practice they are prone to high variance with the unseen data. Random Forest is the solution to this problem which ensembles multiple decision trees which upgrades the robustness of predictions. It results in a vast improvement in accuracy. It starts by creating a boot strapped dataset and builds a decision tree only with a subset of features and this process continues further. This process is called Bagging. This algorithm also handles high dimensional data and has the ability to capture complex non linear relationships. It is less prone to overfitting since it uses a combination result of separately formed decision trees.

Chapter 5

Deep Learning Algorithms

Deep Learning is a subset of Machine Learning which focuses on the development and application of artificial neural networks with multiple layers similar to the neurons in our brain. They are playing a key role in modern day to solve complex problems. Some pretrained models are available to reduce the computational and data requirements.

5.1 Artificial Neural Network

Artificial Neural Network is a popular deep learning modelling technique which is inspired from the functionality of the human brain's neural networks. They consist of neurons which process, transmit information through weighted connections. Activation functions are mathematical functions which are nonlinear in nature.

5.2 Convolutional Neural Network

These types of networks are feed forward networks. CNN has convolutional operations between a set of filters and the inputs. A typical CNN contains three layers namely convolutional layers, pooling layers and fully connected layers. The layers in CNN are sparser than fully connected layers. Kernel is used for convolution of a layer using the same weights across the whole layer. They are significantly used in a wide range of applications like classification, counting.

Chapter 6

Results

6.1 tSNE Plots

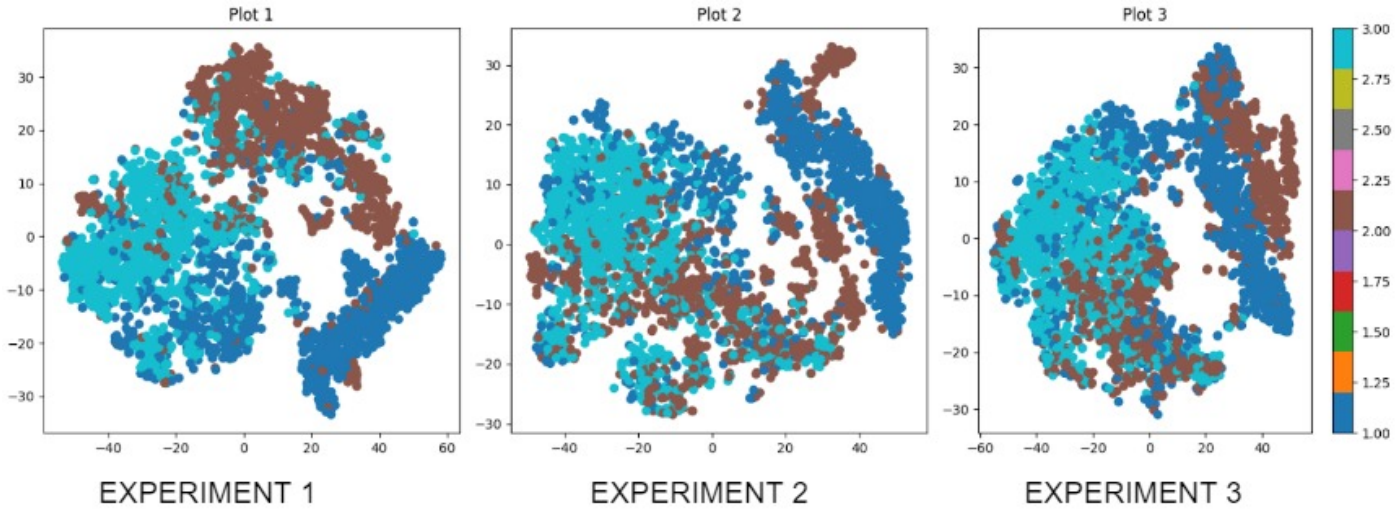

Figure 6.1: tSNE Plots

These plots make clear that the features contain non linear relationships between themselves and hence these are tough to linear separation. It also insists us to use complex deep learning architectures which ensure the non linearity between features.

6.2 ML algorithms

6.2.1 Support Vector Machines

	ACCURACY	PRECISION	RECALL
EXPERIMENT 1	85.05	85.53	85.05
EXPERIMENT 2	73.13	75.98	73.13
EXPERIMENT 3	70.03	75.17	70.03

6.2.2 Naive Bayes

	ACCURACY	PRECISION	RECALL
EXPERIMENT 1	64.83	71.59	64.38
EXPERIMENT 2	60.56	65.69	60.57
EXPERIMENT 3	64.98	70.50	64.98

6.2.3 Decision Tree

	ACCURACY	PRECISION	RECALL
EXPERIMENT 1	88.26	88.25	88.26
EXPERIMENT 2	78.11	78.11	78.11
EXPERIMENT 3	76.11	76.14	76.11

6.2.4 Random Forests

	ACCURACY	PRECISION	RECALL
EXPERIMENT 1	90.86	90.84	90.86
EXPERIMENT 2	83.44	83.37	83.44
EXPERIMENT 3	80.74	80.88	80.74

6.3 DL algorithms

6.3.1 Artificial Neural Network

Galaxy - Star - QSO classification

	ACCURACY	PRECISION	RECALL	F1 SCORE
EXPERIMENT 1	92.77	92.81	92.76	92.78
EXPERIMENT 2	73.13	75.98	73.13	86.2

Galaxy - Star classification

	ACCURACY	PRECISION	RECALL	F1 SCORE
EXPERIMENT 1	97.73	97.7	97.73	97.75
EXPERIMENT 2	96.14	96.2	96.14	96.14

Chapter 7

Observations

The SVM algorithm works well on the compact galaxy dataset and classifies most of the galaxies from stars and quasars.But it shows poor performance as the dataset is huge and also the number of samples is very much larger than the number of features or dimensions. This algorithm takes more time as the dataset needs to be taken into higher dimensions and has to select the suitable hyperplane.

Gaussian Naive Bayes algorithm works well when every feature in the dataset is available in normal distribution.Moreover it fails whenever it doesn't happen.It also assumes that every feature is independent of each other. If it doesn't happen, gaussian NB may not capture the dependencies of features which results in low accuracy.Moreover this algorithm has high bias and low variance.

The Decision Tree works well in the classification of compact galaxies from stars and quasars. Since it is one of the optimal algorithms, it has a higher accuracy than others. But it has high variance to the test data. It undoubtedly suits the training data but failed in predicting the unseen test samples.

Random Forests did a better job compared to other ML algorithms. The accuracy rose above 90 percent in classification of compact galaxies and a decent performance in classification of compact and faint galaxies from stars and quasars.

MargNet ANN architecture did a good performance in classifying the galaxies from stars and quasars. As the neurons are completely connected from each other, it turns into a Dense Neural Network. It worked well on the classification of galaxies from stars as the morphology only distinguished the both.

MargNet CNN architecture has a good architecture composed of 5 different patches with every layer activated with a ReLU function. This big architecture tends to capture different features. As we are using different kernels with dimensions matching the dimensions of input 5 channel images. It needs more time compared to other algorithms as it needs to pass the input through certain layers. It is the most efficient of all the other algorithms.

Chapter 8

Conclusion and Future Scope

The current project mainly focused on basic Machine Learning algorithms and a part of Deep Learning Network MargNet. MargNet implementation should be carried further by making the ensembler which ensembles the work of ANN and CNN together. It must be implemented using pytorch data framework on which the work has been started. Machine Learning algorithms are not that much fit for the current project as they need feature engineering whereas deep learning algorithms automatically detect the features of images. It can also be further investigated with the modern networks like VGGNet, GoogleNet, AlexNet etc...

References

1. Photometric identification of compact galaxies, stars and quasars using multiple neural networks

   Siddharth Chaini,1 Atharva Bagul,1† Anish Deshpande,2 Rishi Gondkar,3 Kaushal Sharma,4 M. Vivek,5 Ajit Kembhavi6

2. Star-galaxy classification in the Dark Energy Survey Y1 dataset *I. Sevilla-Noarbe1 , B. Hoyle2,3 , M.J. March˜a4 , M.T. Soumagnac5 , K. Bechtol6 , A. Drlica-Wagner7 , F. Abdalla4,8 , J. Aleksi´c9 , C. Avestruz10, E. Balbinot11 , M. Banerji12,13, E. Bertin14,15, C. Bonnett9 , R. Brunner16, M. Carrasco-Kind17 , A. Choi18, T. Giannantonio12,13,2 , E. Kim17, O. Lahav4 , B. Moraes4 , B. Nord7 , A.J. Ross18, E.S. Rykoff19,20, B. Santiago21,22, E. Sheldon23, K. Wei10,24 , W. Wester7 , B. Yanny7 , T. Abbott25, S. Allam7 , D. Brooks4 , A. CarneroRosell22,26, J. Carretero9 , C. Cunha19, L. da Costa22,26, C. Davis19, J. de Vicente1 , S. Desai27, P. Doel4 , E. Fernandez9 , B. Flaugher7 , J. Frieman7,10 , J. Garcia-Bellido28, E. Gaztanaga29,30, D. Gruen19,20, R. Gruendl16,17 , J. Gschwend22,26, G. Gutierrez7 , D.L. Hollowood31, K. Honscheid18,32, D. James33 , T. Jeltema31, D. Kirk4 , E. Krause34,35, K. Kuehn36, T. S. Li7,10, M. Lima37,22 , M. A. G. Maia22,26, M. March38, R. G. McMahon4,8 , F. Menanteau16,17 , R. Miquel39,9 , R. L. C. Ogando22,26, A. A. Plazas35, E. Sanchez1 , V. Scarpine7 , R. Schindler20, M. Schubnell40, M. Smith41, R. C. Smith25, M. Soares-Santos42 , F. Sobreira43,22, E. Suchyta44, M. E. C. Swanson17, G. Tarle40, D. Thomas45 , D. L. Tucker7 , A. R. Walker25*

3. Star-galaxy Classification Using Deep Convolutional Neural Networks *Edward J. Kim1? and Robert J. Brunner1,2,3,4*

4. Decision Tree Classifiers for Star/Galaxy Separation *E. C. Vasconcellos1, R. R. de Carvalho2, R. R. Gal3, F. L. LaBarbera4, H. V. Capelato2, H. F.Campos Velho5, M. Trevisan6 and R. S. R. Ruiz1*

5. MACHINE LEARNING APPLIED TO STAR-GALAXY-QSO CLASSIFICATION AND STELLAR EFFECTIVE TEMPERATURE REGRESSION *Yu Bai1, JiFeng Liu1,2, Song Wang1, and Fan Yang2*

6. Machine-learning identification of galaxies in the WISE × SuperCOSMOS all-sky catalogue *T. Krakowski1, K. Małek1, 2, M. Bilicki3, 2, A. Pollo1, 4, 2, A. Kurcz4, 2, M. Krupa4, 2*