PES UNIVERSITY

COMPUTER SCIENCE AND ENGINEERING

MACHINE LEARNING

UE17CS303

Separating Stars from Quasars: Machine Learning Investigation Using Photo metric Data

https://github.com/navneetraju66/SDSS-Data-Classification

AUTHORS :

LAMYA BHASIN
PES1201701244

SUJEETH V
PES1201700958

NAVNEET RAJU
PES1201701545

Contents

1 ABSTRACT

A problem that lends itself to the application of machine learning is classifying matched sources in the Galex (Galaxy Evolution Explorer) and SDSS (Sloan Digital Sky Survey) catalogs into stars and quasars based on color-color plots. The problem is daunting because stars and quasars are still inextricably mixed elsewhere in the color-color plots and no clear linear/non-linear boundary separates the two entities. Diversity and volume of samples add to the complexity of the problem. We explore the efficacy of neural network based classification techniques in discriminating between stars and quasars using GALEX and SDSS photometric data.Catalogs comprising of samples labelled using our classifiers can be further used in studies of photometric sources.The design of a novel Neural Network classifier is proposed in the paper to tackle the classification problem. To evaluate the correctness of the classifiers, we report the accuracy and other performance metrics and find reasonably satisfactory range of 81-100

2 INTRODUCTION

One of the major challenges of of large scale photometric surveys is the separation of the different classes of sources, especially stars and quasars. Both types of sources have a compact optical morphology and are hence difficult to separate without spectroscopic data (Fan, 1999). In such cases, other parameters of the sources such as their optical variability or their optical colors (Richards et al., 2002) are necessary to distinguish between stars and quasars. Later studies have shown that including the infrared data or UV data with optical photometry results in a more efficient separation. In this paper we present a machine learning classification approaches to distinguish between stars and quasars using only optical photometric data and UV data.

The Sloan Digital Sky Survey or SDSS is an optical survey that observed large portions of the sky in the wave bands u,g,r,i,z and obtained the spectra of the sources so that their red-shifts could be determined as well. The survey was conducted using 2.5m wide-angle optical telescope operated at Apache Point Observatory in New Mexico.The data we have used for classification is derived using GALEX and SDSS photometric data.

The pre-processed data (Makhija et al., 2019) given consisted of 4 catalogs:

1. North Galactic Region Only: Selected only samples that have fuv.Populated the entire feature list (with pairwise differences)Random Forests (RF) without

upsampling to generate predicted labels was run.

2. Equatorial Region Only: Selected only samples that have fuv. Populated the entire feature list (with pairwise differences).RF without upsampling was run to generate predicted labels.

3. North Galactic Region and Equatorial Region Combined: Selected only samples that have fuv. Populated the entire feature list (with pairwise differences).RF without upsampling was run to generate predicted labels.

4. Removed fuv and fuv-related features: Populate the entire feature list (even with samples that don't have fuv, with pairwise differences). Run RF without upsampling to generate predicted labels.

Machine learning involves the use of statistics to make useful predictions and to learn essential features; classification is the process of marking separations between categories in data. Supervised machine learning techniques make use of labeled data to make predictions of future unseen data. In the present context, the labels in the data comprise of numeric indications of a source being a star or a quasar. The problem of separating stars from quasars using machine learning has not been studied in much depth

3 IMPLEMENTATION

The classification model used is Deep Learning. the advantages of deep learning are:

1)Has best-in-class performance on problems that significantly outperforms other solutions in multiple domains. This includes speech, language, vision, playing games like Go etc. This isn't by a little bit, but by a significant amount.

2)Reduces the need for feature engineering, one of the most time-consuming parts of machine learning practice.

3)Feature engineering can be automatically executed inside Deep Learning model.

4)Can solve complex problems.

5)Flexible to be adapted to new challenge in the future (or transfer learning can be easily applied).

6)High automation. Deep learning library (Tensorflow, keras, or MATLAB...) can help users build a deep learning model in seconds (without the need of deep understanding). Most deep learning methods use neural network architectures, which is why deep learning models are often referred to as deep neural networks.

The term “deep” usually refers to the number of hidden layers in the neural network. Traditional neural networks only contain 2-3 hidden layers, while deep networks can have as many as 150.

Deep learning models are trained by using large sets of labeled data and neural network architectures that learn features directly from the data without the need for manual feature extraction.

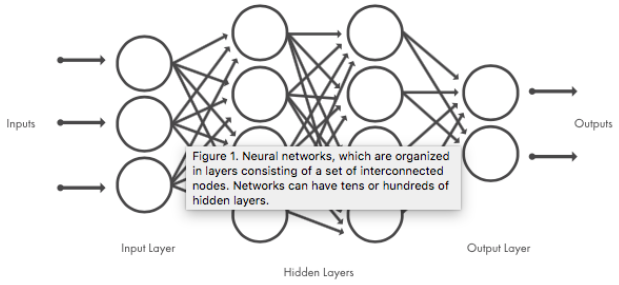
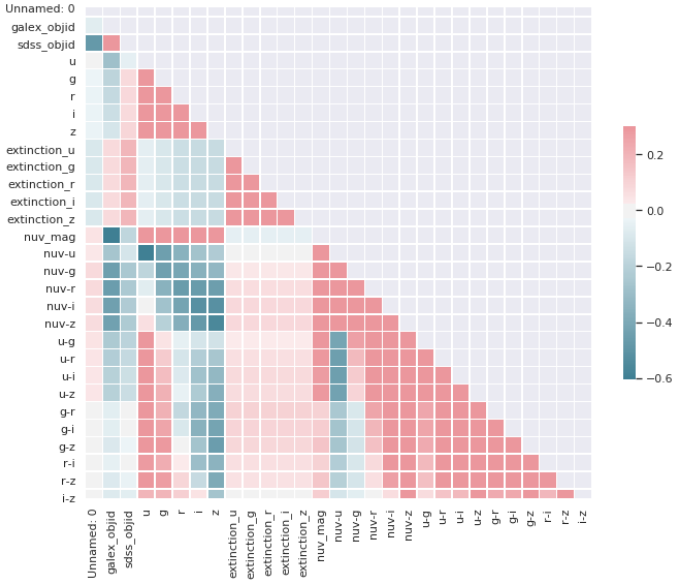


Figure 1. Neural networks, which are organized in layers consisting of a set of interconnected nodes. Networks can have tens or hundreds of hidden layers.

Before jumping into the classification problem and developing the model,it is important to understand the data that we are working with and the kind of feature vectors that would be useful in our classification problem. One of the most common and important techniques to understand the data is to see and understand if there is any kind of co relation between the features. Identifying co relation is important as it could help in dimensionality reduction techniques such as principal component analysis. Below is the co relation matrix, and as we can see although there is correlation it is not too strong and hence PCA or any dimensionality reduction wouldn’t be of much help.



Over-sampling using SMOTE(Synthetic Minority Oversampling Technique) With our training data created, up-sample the no-subscription using the SMOTE algorithm. At a high level, SMOTE:Works by creating synthetic samples from the minor class (no-subscription) instead of creating copies.Randomly choosing one of the k-nearest-neighbors and using it to create a similar, but randomly tweaked new observations. Over-sampling was done on the data set as there was a large imbalance in the classes(especially in the catalog 3 where there are only a few 500 data points for a class).

The deep neural net that we have used consists of an input vector of (16,) and 2 hidden layers.The diagrammatic representation of the deep neural network is shown in the figure.For the first 2 hidden layers the ReLU activation function is used for its linear property (f(x)=max(0,x)).For the activation of the single output neuron we have used a sigmoid function which is sort of a thresholding function and also gives a sort of "probabilistic" output for the binary classification. The loss function which was used for the back propagation of the model was the binary cross-entropy function which was found to be more suitable for this problem of binary classification of stars and quasars.The gradient descent optimization algorithm was used for the back propagation of the neural net and subsequent updation of the weights and biases in the neural network.

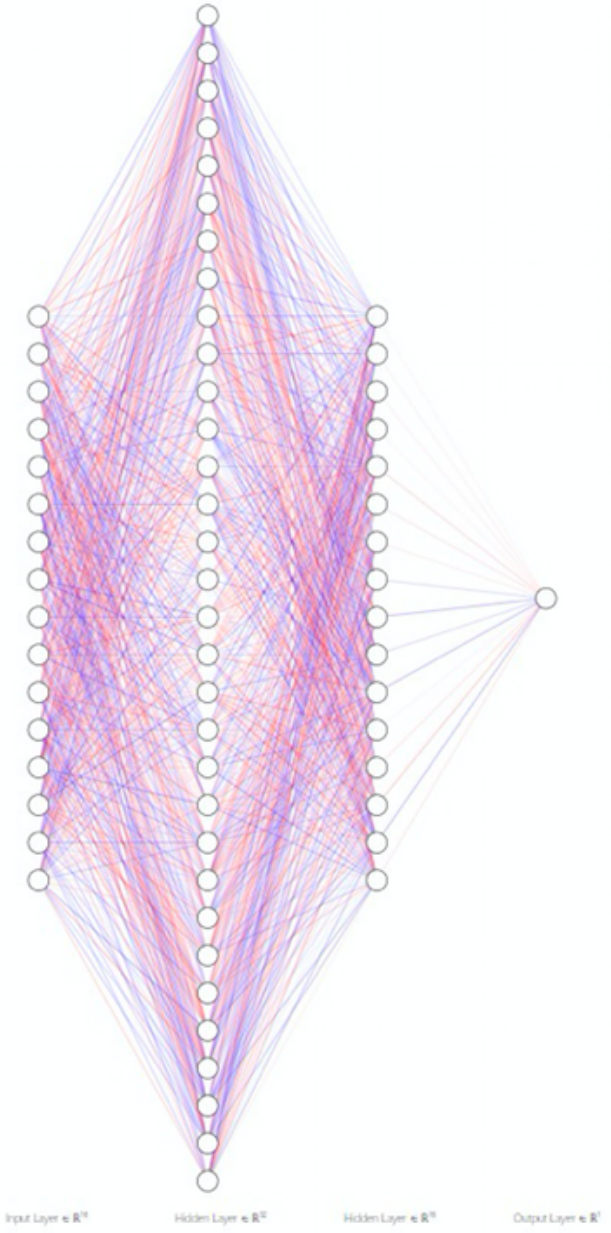
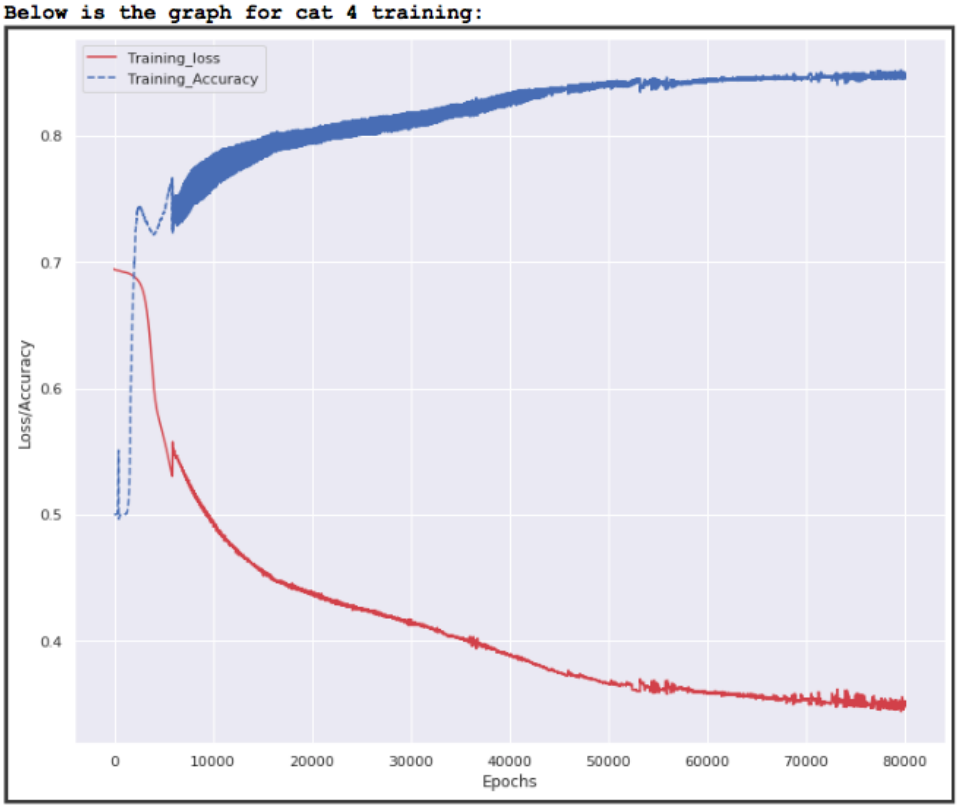


fig. neural network

4 RESULTS

Train accuracy on cat 4 $:84.9:
Test accuracy on test cat 4 data:
0.8434710153553149
Below is the graph for cat 4 training:



	precision	recall	f1-score	support
0	0.79	0.94	0.86	4201
1	0.92	0.75	0.83	4200
accuracy			0.84	8401
macro avg	0.86	0.84	0.84	8401
weighted avg	0.86	0.84	0.84	8401

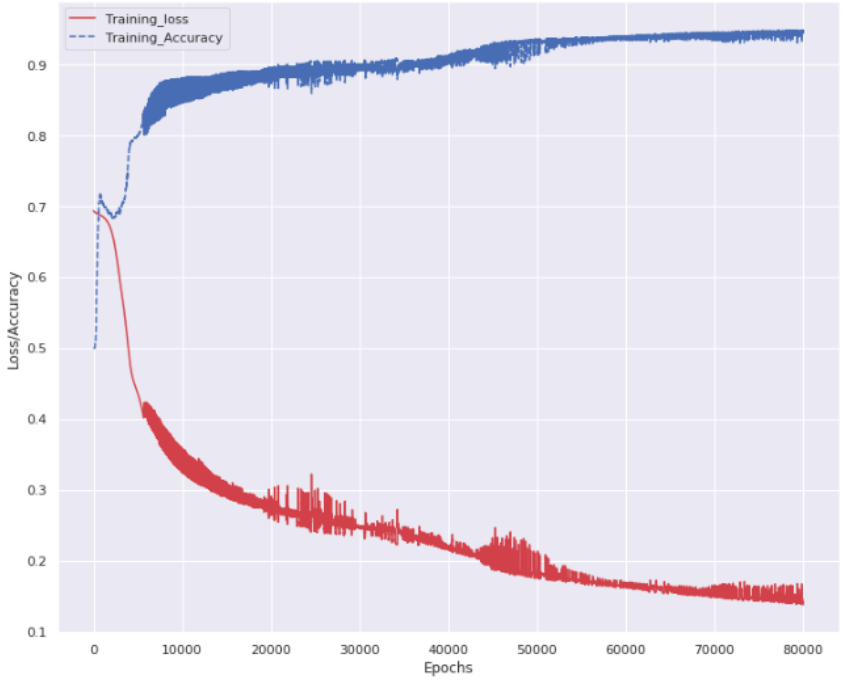
Testing on cat 3 accuracy: 0.86

	precision	recall	f1-score	support
0	0.42	0.88	0.57	466
1	0.98	0.85	0.91	3787
accuracy			0.86	4253
macro avg	0.70	0.87	0.74	4253
weighted avg	0.92	0.86	0.88	4253

Testing on cat2 accuracy:0.8598337950138504

	precision	recall	f1-score	support
0	0.44	0.87	0.59	413
1	0.98	0.86	0.92	3197
accuracy			0.86	3610
macro avg	0.71	0.86	0.75	3610
weighted avg	0.92	0.86	0.88	3610

Trained on CAT3:
Train accuraccy:94.6%



Tested on cat3:

	precision	recall	f1-score	support
0	0.91	0.94	0.93	1034
1	0.94	0.91	0.93	1033
accuracy			0.93	2067
macro avg	0.93	0.93	0.93	2067
weighted avg	0.93	0.93	0.93	2067

Tested on CAT 2:

	precision	recall	f1-score	support
0	0.60	0.95	0.74	413
1	0.99	0.92	0.95	3197
accuracy			0.92	3610
macro avg	0.80	0.93	0.85	3610
weighted avg	0.95	0.92	0.93	3610

Tested on CAT4:

	precision	recall	f1-score	support
0	0.65	0.75	0.69	9973
1	0.88	0.83	0.85	23156
accuracy			0.80	33129
macro avg	0.77	0.79	0.77	33129
weighted avg	0.81	0.80	0.81	33129

5 CONCLUSION

The choice of using a deep neural Network proved to be successful in the classification of stars and quasars. After multiple iterations of trying out different loss functions,optimization algorithms and the neural network hyperparameters we found that a simple gradient descent along with the binary cross entropy loss function provided good results in the binary classification. Over 90 percent accuracy was observed across the catalogs(this value varies across the catalogs as shown in the results section.

6 BIBLIOGRAPHY

Makhija, Simran Saha, Snehanshu Das, Mousumi Basak, Suryoday. (2019). Separating Stars from Quasars: Machine Learning Investigation Using Photometric Data. 10.13140/RG.2.2.24220.74889.