Classification of Stars and Galaxies with Machine Learning

Matthew Pooser
Department of Computer Science
University of Georgia
Athens, GA 30602
mpooser@uga.edu

Abstract

This paper addresses the performance of four different machine learning algorithms, Logistic Regression, Gaussian Naïve Bayes, Linear SVM, and Random Forest Classification on an astronomical dataset.

1 Introduction

1.1 Motivation

Excessively large datasets are needed for machine learning algorithms to correctly train discriminative models. In this vein, the size of astronomical data is usually on the order of terabytes (TB) which should work well. In addition, further research for related work shows not many technical papers have attempted to apply machine learning to the field of astronomy and a large amount of astronomical data remain unmined for potential insights. Lastly, astronomy has always fascinated me, and I even was an Astrophysics double major along with Computer Science until I decided to drop Astrophysics. Although weak on the physics side, I still wanted to apply astronomy in some capacity and this project was a well-defined fit.

1.2 Research Question

Which machine learning works best on a specific astronomical survey?

1.3 Solution

Logistic Regression will perform better than Gaussian Naïve Bayes, Linear Support Vector Machine (SVM), and Random Forest Classification.

2 Project's Dataset

The dataset used for this project is a pre-labeled subset of the Sloan Digital Sky Survey Data Release 12 (SDSS DR12) which was published in 2014. Since axial precession of the earth occurs every 25,772 years, this dataset remains relevant for today's use. The CSV file consists of 91,272 data points, 9 columns each (after preprocessing). The target column is class: 0 indicates a galaxy and 1 indicates a star.

[8]:

	ra	dec	u	g	r	i	z	class	redshift
0	300.841762	76.511282	19.19619	17.83329	17.52225	17.40237	17.35182	STAR	-0.000220
1	300.730508	76.551731	21.65541	19.13715	17.92577	17.44741	17.15818	STAR	-0.000008
2	300.871382	76.530570	20.70867	19.20954	18.55966	18.24395	18.10117	STAR	0.000096
3	300.317409	76.374746	22.88806	21.20900	19.90560	19.33555	19.08966	STAR	-0.000247
4	301.252332	76.319520	17.82932	16.11081	15.39808	15.13612	15.00507	STAR	-0.000131
5	301.458518	76.426766	14.23474	14.45194	14.75188	14.99473	15.24296	STAR	-0.000288
6	300.848872	76.320462	21.69935	19.60817	18.18657	17.04788	16.44566	STAR	-0.000267

Figure 1: SDSS DR12 data

2.1 Dataset's Background Information

These are quick pieces of information related to ra, dec, u, g, r, i, z, and redshift.

2.1.1 Right Ascension (ra) & Declination (dec)

- These refer to the coordinates of an object in space.
- Right Ascension: celestial longitude, measured in hours. Each hour represents 15 deg/hr of the night sky. A degree is equivalent to the area that a human finger takes up when the human arm is fully extended towards the night sky.
- Declination: celestial latitude, measured in degrees. An easy way for astronomers to know their declination is to take the latitude of where they are. For instance, since Athens, GA has a latitude of 34º N, people in that area are at a declination of +34º.

2.1.2 Filters

Filters isolate certain intervals of a star's electromagnetic spectrum. U corresponds to ultraviolet light. G and R correspond to green and red respectively and are examples of optical light. The last two filters, I and Z, are in the near-infrared part of the electromagnetic spectrum.

2.1.3 Redshift

As light continues to travel farther away from its source, its spectrum naturally shifts toward the red part of the electromagnetic spectrum. This can be used to determine the distance and radial velocity of stellar phenomena. Figure 2 shows a visual of redshift and blueshift.

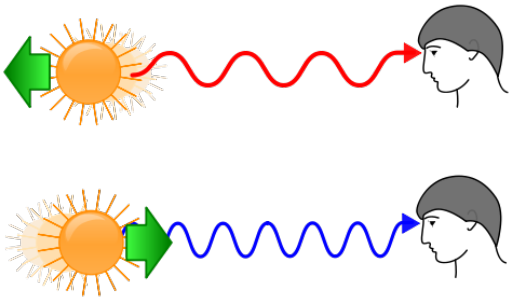

Figure 2: Example of Redshift and Blueshift

2.2 Disproportionate Number of Stars to Galaxies

The dataset has many more galaxies, ~73%, than stars, ~27%, which would heavily impact training the models. Techniques to balance the dataset are further discussed in Section 3.2.

```
GALAXY    66847
STAR      24425
Name: class, dtype: int64
```

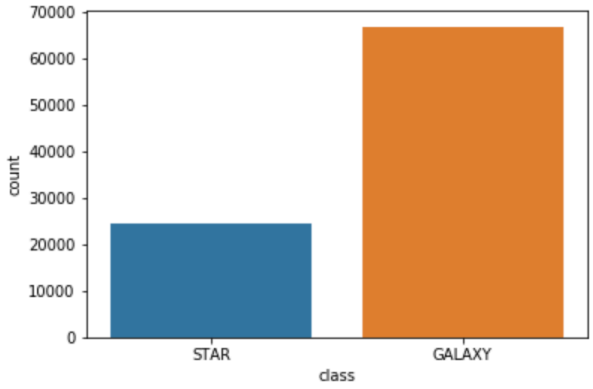

Figure 3: Class Membership

3 Project Methodology

This section details the tools, the preprocessing techniques, and the process to collect the results.

3.1 Tools

- Scikit-learn
- Imbalanced
- Jupyter
- NumPy
- Pandas
- Seaborn & Matplotlib

3.2 Preprocessing

Identifying characteristics such as objid, specobjid, and fibered were removed to prevent the classifier from associating those ids with a certain class. Other original features such as run, rerun, camcol, field, plate, and mjd were removed as they pertained to how the pictures were taken and not so much about the intrinsic properties that I wanted the classifier to learn. The third class, QSO for quasar objects, were removed as it was outside the project's scope. Fortunately, there were no NaN values or missing values in the original dataset.

The pre-processed data was split into a testing set (25%) and training set (75%). SMOTE (Synthetic Minority-oversampling Technique) was used to generate synthetic data points to balance the amount of star and galaxy class representation in the training data, but not the testing data. The testing data was untouched in order to prevent information loss.

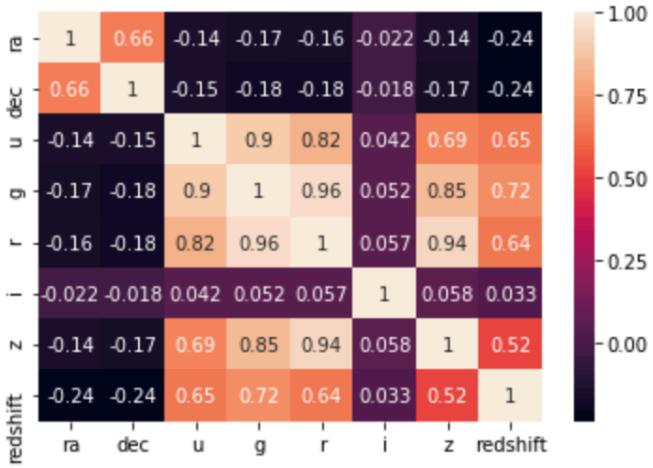

Figure 4: Heatmap of the Remaining Features

3.3 Procedures

1. The dataset was loaded into Jupyter Notebook
2. The unneeded features were removed
3. Data was normalized and target values were given numerical values
4. Training and testing were performed with performance metrics printed for the test data

4 Results

The rest of Section 4 outline the accuracy, F1-score, recall, AUROC, confusion matrices, and class-based performances for each of the four tested classifiers.

4.1 Logistic Regression

- Accuracy: 97.1%
- F1-Score: 94.9%
- Recall: 90.5%
- AUROC: 98%

16081	16
637	6084

Table 1: Confusion Matrix for Logistic Regression

	precision	recall	f1-score	support
0	0.9990	0.9619	0.9801	16718
1	0.9052	0.9974	0.9491	6100
micro avg	0.9714	0.9714	0.9714	22818
macro avg	0.9521	0.9796	0.9646	22818
weighted avg	0.9739	0.9714	0.9718	22818

Figure 5: Classification Report for Logistic Regression

4.2 Gaussian Naïve Bayes

- Accuracy: 98.8%
- F1-Score: 97.7%
- Recall: 98.4%
- AUROC: 98.2%

16,621	185
97	5,915

Table 2: Confusion Matrix for Gaussian Naïve Bayes

	precision	recall	f1-score	support
0	0.9890	0.9942	0.9916	16718
1	0.9839	0.9697	0.9767	6100
micro avg	0.9876	0.9876	0.9876	22818
macro avg	0.9864	0.9819	0.9842	22818
weighted avg	0.9876	0.9876	0.9876	22818

Figure 6: Classification Report for Gaussian Naïve Bayes

4.3 Linear SVM

- Accuracy: 98.7%
- F1-Score: 97.7%
- Recall: 95.5%
- AUROC: 99.1%

16,432	0
286	6,100

Table 3: Confusion Matrix for Linear SVM

	precision	recall	f1-score	support
0	1.0000	0.9829	0.9914	16718
1	0.9552	1.0000	0.9771	6100
micro avg	0.9875	0.9875	0.9875	22818
macro avg	0.9776	0.9914	0.9842	22818
weighted avg	0.9880	0.9875	0.9876	22818

Figure 7: Classification Report for Linear SVM

4.4 Random Forest Classification

- Generated 5 decision trees with max_features = 4. No pruning was used.
- Accuracy: 99.4%
- F1-Score: 98.9%
- Recall: 99.9%
- AUROC: 99.8%

16,588	130
13	6,087

Table 4: Confusion Matrix for Random Forest Classification

	precision	recall	f1-score	support
0	0.9997	0.9971	0.9984	16718
1	0.9922	0.9992	0.9957	6100
micro avg	0.9977	0.9977	0.9977	22818
macro avg	0.9959	0.9982	0.9970	22818
weighted avg	0.9977	0.9977	0.9977	22818

Figure 8: Classification Report for Random Forest Classification

5 Conclusions

All four algorithms performed well on the test data, but Random Forest Classification had the highest accuracy, F1-score, recall, and AUROC. All four models may be overfitting, but cross validation gave roughly the same percentages for the above performance metrics. Pruning was attempted for the Random Forest, but accuracy remained the same. See `tree-01.png` (or 2, 3, 4, 5) for what the decision tree looked like.

It seems as though I made a classifier to predict if an object is either a galaxy or not a galaxy based off the classification reports for each model. At the suggestion of the panel during the presentation of the research, plugging in arbitrary points to see its predicted output would be insightful. Training these models on a mixture of astronomical data from a variety of different sources will be the next step forward.

6 Future Work

Future improvements are outlined below to increase the usefulness of the trained models and what else machine learning can be used for in terms of astronomy.

6.1 Future Improvements to the Project

Since the dataset represented such an extremely small portion of the data, the models should be trained on bigger datasets (on the order of TB) in order to verify the models' accuracy.

Additionally, changing the hyperparameters for the Random Forest may reduce potential overfitting and overall size of the decision trees.

6.2 Further Topics for Study

Several topics that could be the next step for machine learning with astronomy involve the detection of gravitational waves, detection of exoplanets, and predicting the presence of dark matter / dark energy.

Acknowledgments

Dr. Sheng Li, Assistant Professor in Department of Computer Science, University of Georgia

References

https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html

https://towardsdatascience.com/formatting-tips-for-correlation-heatmaps-in-seaborn-4478ef15d87f

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html