Charles Leung 400049422
leungt11

DATASCI 3ML3 PROJECT REPORT

INTRODUCTION

Galaxies are one of the largest cosmic structures and are vast collections of stars, gas, and dust, bound together by gravity, and are the building blocks of our universe. Studying them can help us understand the formation and evolution of the universe and answer unsolved questions in physics (e.g. the nature of dark matter). One way to study galaxies is through their images, which are captured by telescopes and other instruments. These images contain a wealth of information about the properties of galaxies, such as their size, shape, color, and brightness.

Accurate classification of galaxies can provide valuable information about their origin, dynamics, and interaction with other celestial bodies. In recent years, machine learning techniques have emerged as powerful tools for automating the classification of galaxies based on their morphological features. Because machine learning models can be trained to recognize patterns and features in galaxy images and classify them according to their properties, they can help astronomers and astrophysicists automate the process of analyzing galaxy images and extract useful information from them more efficiently. Machine learning can also help to identify new types of galaxies.

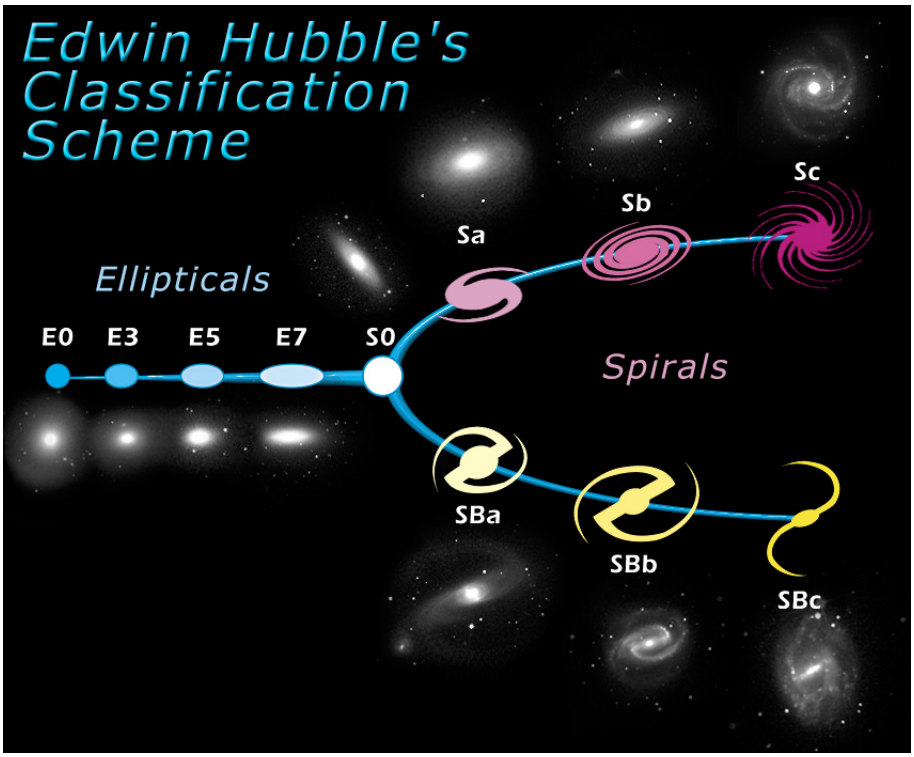

Figure 1: Galaxy types under the Hubble Classification Scheme

This project aims to compare the performance of two machine learning classifiers, namely Decision Tree Classifier and Random Forest Classifier, in classifying galaxies into three categories: ellipticals, spirals, and mergers in order to explore the capabilities of these classifiers in handling complex and high-dimensional astronomical data and to determine the most suitable model for this particular classification task. These are not the only galaxy types that exist; but since they are the most dominant galaxy types, using a ternary classification scheme can greatly simplify the problem at hand.

Decision Tree Classifiers work by recursively splitting the data into subsets based on the features that best separate the classes. On the other hand, Random Forrest Classifiers are an ensemble of Decision Trees that combine multiple trees to improve the accuracy and robustness

of the model. They work by randomly selecting a subset of features and a subset of data points for each tree and aggregating the results. Random forests can help correct for overfitting tendencies in decision trees, and it may be particularly insightful to compare the two in the context of this problem.

The key idea behind Random Forest is that the combination of multiple "weak" models (individual trees) can produce a more accurate and robust "strong" model, analogous to a weighted mean in statistics. The final prediction in a classification problem can be obtained using a majority vote among the individual trees:

$$Prediction \ = \ Majority \ vote(Decision \ tree_1, Decision \ tree_2, \ldots, Decision \ tree_n)$$

where $Decision \ tree_i$ is the prediction of the i^{th} decision tree, and n is the number of trees in the ensemble.

The dataset used in this project is the is a truncated version of the Galaxy Zoo 2 data set, which contains over 300,000 galaxy images labeled by citizen scientists, and consists of various features extracted from images of galaxies. These features include color indices, eccentricity, adaptive moments, and Petrosian radii (the light 'bulge' at the center of the galaxy.). Astronomers have identified these attributes to be instrumental in characterizing the morphological properties of galaxies and therefore play a significant role in determining their classification.

METHODS

A systematic approach will be used for the project. First, dataset will be preprocessed by splitting it into training and testing sets using a 70:30 ratio. The 13 features are then generated from the dataset to be used as input for the classifiers. Decision Tree and Random Forest Classifiers are thereafter trained using the training data.

To evaluate the performance of the models, 10-fold cross-validation is employed, and the classifiers are compared based on the following metrics: accuracy, precision, and recall. Confusion matrices are also generated to provide a visual representation of the classifiers' performance.

The Galaxy-2 dataset used in this project consists of images of galaxies, along with various morphological properties that are critical for classification. The feature extraction process involves generating 13 features for each galaxy, which are used as inputs for the classifiers. These features are the color indices (u-g, g-r, r-i, i-z), eccentricity, adaptive moments ($m4_u, m4_g, m4_r, m4_i, m4_z$), and the ratios of Petrosian radii (petroR50/petroR90 in u, r, and z bands). These attributes help capture essential morphological properties of the galaxies in the dataset.

Color features are calculated using the magnitudes of galaxies in different filters, such as u, g, r, i, and z. For example, the u-g color feature can be calculated as follows:

$$\text{u-g} = \text{mag}_u - \text{mag}_g$$

where mag_u and mag_g are the magnitudes in the u and g filters, respectively. The other color features are calculated similar.

Eccentricity is the measure of how elliptical the galaxy is. It can be calculated by

$$e = \frac{c}{2a}$$

where e is the distance between the foci and a is the length of the major diameter of the galaxy.

This process can be summarized with the following steps:

a. Preprocessing: Split the dataset into training and testing sets using a 70:30 ratio.
b. Feature extraction: Generate the following features from the dataset: color indices, eccentricity, adaptive moments, and concentration indices.
c. Model training: Train the Decision Tree Classifier and the Random Forest Classifier using the training data.
d. Model evaluation: Use 10-fold cross-validation to obtain predictions and evaluate the performance of both classifiers. Compare the models based on accuracy, precision, recall, and confusion matrices.

The project will be implemented with Python as the primary programming language and makes use of the following libraries:

- NumPy for data manipulation and handling

- scikit-learn for implementing Decision Tree and Random Forest Classifiers, as well as for model evaluation
- Matplotlib for plotting confusion matrices

To assess the performance of the classifiers, the following evaluation metrics will be used:

- Accuracy: the ratio of correctly classified instances to the total number of instances. A limitation is that it may not be informative in cases of imbalanced class distribution.
- Precision: the ratio of true positives to the sum of true positives and false positives. It provides an insight into the classifiers' ability to correctly identify positive instances among all instances predicted as positive.
- Recall: the ratio of true positives to the sum of true positives and false negatives. It provides an insight into the classifiers' ability to identify all positive instances in the dataset.

In theory, the Decision Tree Classifier is prone to overfitting, especially when dealing with high-dimensional data, whereas the Random Forest Classifier is more robust but is computationally expensive. Since the problem at hand is a low-dimension ternary classification, the Random Forest Classifier should show slightly improved fit through the aforementioned metrics.

RESULTS

Classifier	Data Set	Accuracy	Precision	Recall
Decision Tree	Training	79.12%	78.92%	79.22%
Decision Tree	Testing	79.91%	79.51%	79.52%
Random Forest	Training	82.78%	82.71%	82.82%
Random Forest	Testing	82.91%	82.39%	82.65%

Table 1: Galaxy Classification Results for Decision Tree and Random Forest

The Decision Tree classifier achieved an accuracy of 79.12%, precision of 78.92%, and recall of 79.22% on the training set. When applied to the testing set, the Decision Tree classifier achieved an accuracy of 79.91%, precision of 79.51%, and recall of 79.52%. The testing set shows similar performance metrics to the training set indicating a good generalization.

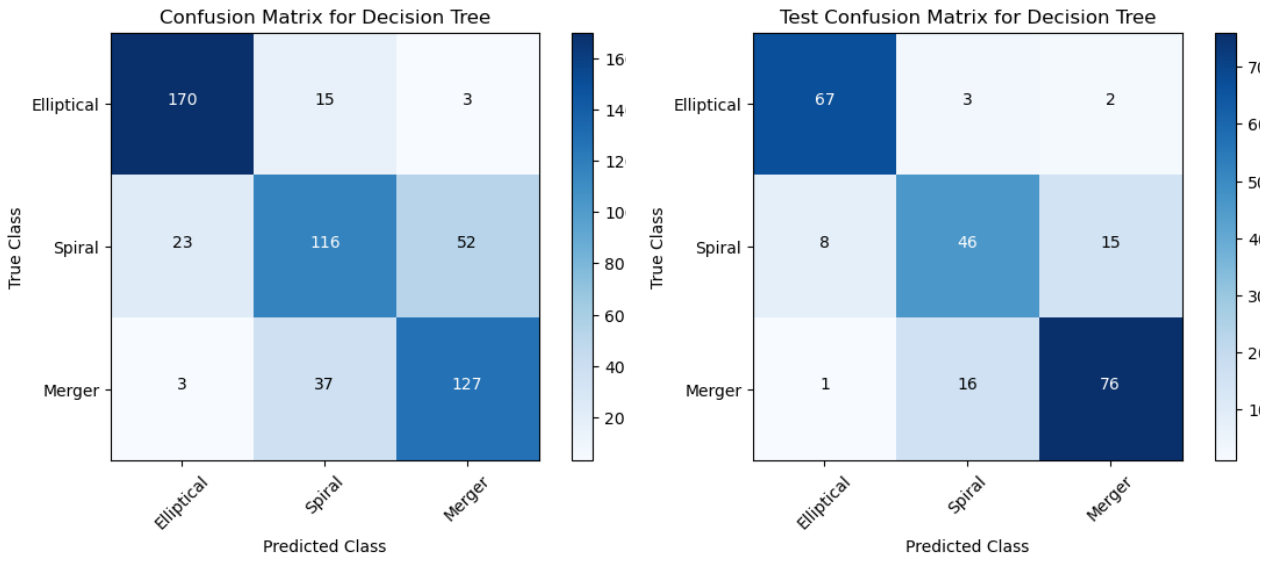

Figure 2: Confusion matrix for Decision Tree, training set (left), testing set (right)

In the confusion matrix, the rows represent the true classes, and the columns represent the predicted classes. In this case, the matrix shows that out of the 182 elliptical galaxies, 166 were correctly classified, 13 were misclassified as spiral, and 3 as merger. For spiral galaxies, 21 were misclassified as elliptical, and 38 as merger. For merger galaxies, 4 were misclassified as elliptical, and 35 as spiral.

The Random Forest classifier, which consists of an ensemble of 50 decision trees, achieved an accuracy of 82.78%, precision of 82.71%, and recall of 82.82% on the training set. When applied to the testing set, the Random Forest classifier achieved an accuracy of 82.91%, precision of 82.39%, and recall of 82.65%. Similarly, performance metrics for the testing set are similar to that of the training set.

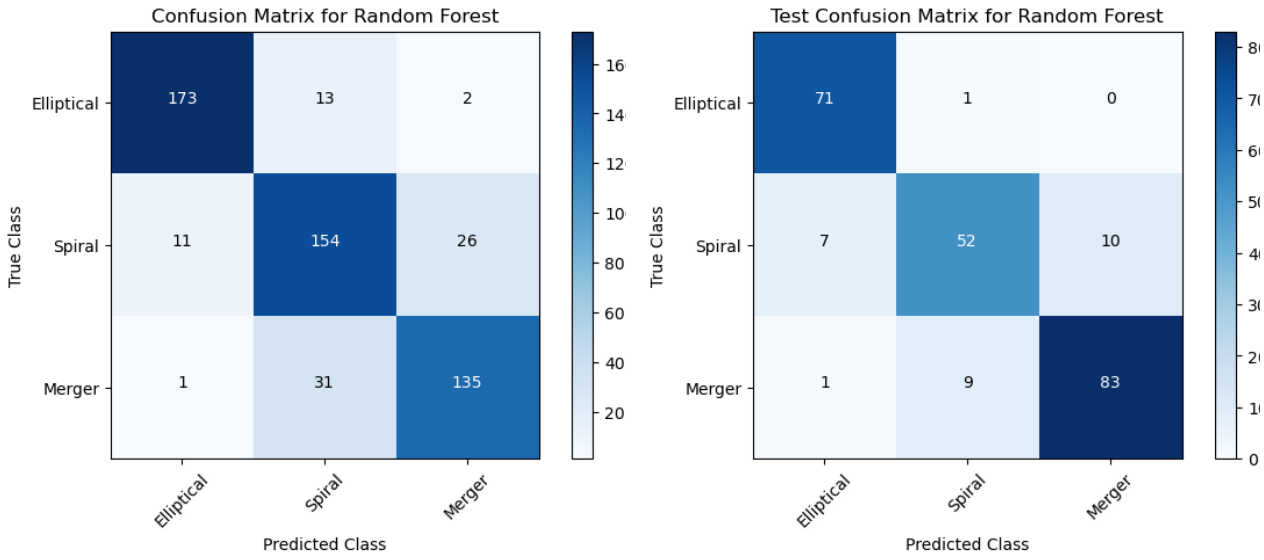

Figure 3: Confusion matrix for Random Forest, training set (left), testing set (right)

Charles Leung 400049422
leungt11

DISCUSSION

The Random Forest classifier outperforms the single Decision Tree classifier in terms of accuracy, precision, and recall for both training and testing sets, although not by a significant margin. This provides credence to the theory that the ensemble nature of the Random Forest classifier should improve the accuracy of the classification results.

The Decision Tree classifier has a good overall performance, with a training accuracy of 0.791 and a testing accuracy of 0.799. The slight difference between the training and testing sets suggests that the model generalizes well to unseen data. The Decision tree is reasonably effective in minimizing both false positives, with a training precision of 0.789 and testing precision of 0.795, and false negatives, with a training recall of 0.792 and testing recall of 0.795. However, the confusion matrices show slightly higher misclassification rates, particularly between the spiral and elliptical classes.

The Random Forest classifier shows slightly better performance compared to the Decision Tree classifier. The model demonstrates excellent generalization to unseen data with a training accuracy of 0.828 and a testing accuracy of 0.829. The Random Forest classifier also outperforms the Decision Tree classifier in minimizing false positives, with a training precision of 0.827 and testing precision of 0.824, and false negatives, with a training recall of 0.828 and testing recall of 0.826.

Both classifiers demonstrate a reasonable performance in classifying the three types of galaxies. The Random Forest classifier has a more balanced classification across the three classes, whereas the Decision Tree classifier shows slightly higher misclassification rates, particularly between the spiral and elliptical classes.

How can the model be further improved? Aside from a larger quantity of training data, better preprocessing and using additional features could be key. Improving the preprocessing steps can help enhance the quality of the input data, which, in turn, can lead to better model performance. For example, outlier detection via regression and handling missing or noisy values via deconvolution could ensure that the model learns from more representative data. Furthermore, feature scaling, such as normalization or standardization, can help prevent the dominance of a single feature and facilitate the convergence of the learning algorithms.

In addition to preprocessing, incorporating more features, such as higher-order moments of the galaxy light distribution or incorporating data from other wavelengths, could potentially enhance the model's ability to capture the complex relationships between galaxy properties and their classifications. Feature selection techniques (e.g. forward selection, backwards elimination) could also be employed to identify the most relevant features and remove redundant or irrelevant ones, improving model efficiency and interpretability.

An aspect that was not explored in this project is the computational time required by each classifier. Decision Trees are generally faster to train and predict than Random Forest classifiers,

as they involve building a single tree compared to an ensemble of trees (in this case, 50) in the latter. However, the increased accuracy and reduced overfitting provided by Random Forest classifiers may justify the additional computational cost for higher-dimensional problems, such as using the entire range of Hubble galaxy classes as well as dealing with esoteric or unique types such as the Sombrero galaxy. Additional study involving comparison of the computational time required by each classifier could be included, taking into account factors such as dataset size, feature complexity, and the number of trees in the ensemble.

It would also be interesting to explore other classification methods, such as Support Vector Machines, k-Nearest Neighbors, or deep learning techniques, to assess their performance on galaxy classification tasks. This would provide a broader perspective on the most suitable methods for classifying galaxies based on their morphological features.

CONCLUSION

In conclusion, this project shows that Decision Tree and Random Forest classifiers are both capable of effectively classifying galaxies based on their morphological features with a roughly 80% accuracy. The Random Forest classifier did show slightly better performance than the Decision Tree classifier as expected. However, as discussed in earlier sections, the choice of the classifier should be guided by the specific requirements of the project, such as computational resources and interpretability, which could become significant when dealing with an expanded list of galaxy types and a wider variety of data types.

Charles Leung 400049422
leungt11

REFERENCES

https://data.galaxyzoo.org/

https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

https://ned.ipac.caltech.edu/level5/Sept11/Buta/Buta15.html

https://skyserver.sdss.org/dr1/en/proj/advanced/color/definition.asp