Advancing Galaxy Classification

Jordan Glesener, Camdin Dirnberger, Micah Glesener

The project explores whether the type of galaxy can be predicted based on cosmological data and investigates the impact of synthetic data on the accuracy of such predictions. The aim is to develop a classification model to categorize galaxies based on their cosmological characteristics. A significant aspect of this investigation is assessing whether incorporating synthetic data, generated by a Generative Adversarial Network (GAN)-trained Generator, enhances the model's accuracy. Classifying galaxies presents a complex challenge, making a predictive model particularly valuable in the field of astronomy. This study also evaluates if the realistic yet synthetic data from the Generator significantly improves the model's performance. This approach, similar to ensemble training but applied to modified datasets, could offer new insights into the utility of synthetic data in advancing classification accuracy.

The goal is to develop a classification model capable of accurately determining the type of galaxy based on light wave, eccentricity, and magnitude data. The target is an initial model accuracy exceeding 75%, with a similar benchmark for a secondary model if its predictions are within 1-2 planets of the actual number of exoplanets. Additionally, realistic synthetic training data regarding light waves and positions of galaxies will be generated to enhance model performance. This synthetic data, designed to simulate real observations, is expected to improve the model's accuracy and provide valuable insights into the efficacy of synthetic data in classification tasks.

A lot of the value of this project is not necessarily readily available, but rather is the same value as most scientific questions. Can we learn more about the secrets of the universe and how the galaxies that we can see change over time? If we can that might help us find something out that will assist our understanding of how our own galaxy is going to change. While many resources have gone into predicting galaxy types in the past, astronomy is often limited by the amount of data available being limited to mostly the late 1990s. Galaxy Zoo provides data about galaxies that have a level of abstraction from an actual photograph to make the computational load more manageable for our lower-powered systems.

Two distinct machine-learning models were employed to enhance galaxy classification analysis. A Random Forest classifier was utilized to make predictions regarding the class of a galaxy based on its features. This model excels at handling complex data with multiple features and provides robust classification results.

Additionally, a Generative Adversarial Network (GAN) was used to generate synthetic data. The GAN created realistic synthetic samples to augment the existing dataset. By generating new, plausible data points, the overall performance and generalization of the galaxy classification model were aimed to be improved. Together, these models were intended to refine predictions and explore the potential of synthetic data in enhancing machine learning outcomes.

For the Random Forest model, the scikit-learn Python package was used. The analysis identified the most important features and determined the optimal number of estimators. The most influential features in predicting the class of a galaxy were:

- Magnitude Differences between the ultraviolet and green bands (indicating how "blue" a galaxy is)
- Eccentricity (describing the shape of the galaxy)
- Petrosian Magnitude (indicating the size of the galaxy)

These features proved to be the most significant in the model's classification of galaxy types. Cross-validation was conducted to find an optimal number of estimators. Random forests were used twice: once on the original data and once on data generated from the GAN model. For training on the GAN model, the original dataset was split in half, with one half fed into the GAN model to produce synthetic data and the other half used for testing.

We created a GAN model that could generate this data on galaxies. The GAN has two neural networks: a generator and a discriminator. Our generator neural network contained a structure similar to most GANs we found in our research. We set it up to take a random noise vector as the input, which it then passes through two layers of nodes, one with 128 nodes and the other with 256 nodes. The Generator then takes the values from the last layer of nodes and creates synthetic star data. The discriminator is also a pretty straightforward neural network. It takes the input vector of the data and passes it through 2 layers of nodes, before using a sigmoid function to output a probability for the data being fake or not.

Once we had our neural networks set up, we created our training loop. For this we used a binary cross entropy loss function to determine the success of each model, which is standard. We used an ADAM optimizer, which again we pulled from our research into existing GAN models as this seems to be a preferred pick. We tried many different learning rates, but then settled on 0.0002 as it seemed to give the best results. It was tricky to get the correct number of Epochs for our run. Our research told us that GANs typically have a few hundred Epochs as they are more complicated neural networks, however, we settled on the lower end and found around 300 to be a reasonably good amount of training iterations, especially with the learning rate that we selected.

Once we had our GAN set up, we began to do our analysis and compare the results of the fake data to the real data. Cross-validation was performed on the model trained with the original data, using a 60/40 training/testing split. The optimal number of estimators was found to be 25. The model achieved an overall accuracy of 78.53%, with accurate classification rates of 94.69% for elliptical galaxies, 68.22% for merger galaxies, and 70.65% for spiral galaxies.

For the GAN-trained model, training was conducted on 10,000 rows of synthetic data created by the GAN, with testing performed on the other half of the original data. Cross-validation revealed that the optimal number of estimators was 20. The results were notably less favorable, with an overall accuracy of 60.90%. The model classified elliptical galaxies at a 62.83% rate, merger galaxies at a 60.75% rate, and spiral galaxies at a 58.70% rate.

Our project into predicting galaxy types using cosmological data and the impact of synthetic data generated by a GAN has revealed that using synthetic data does not enhance the accuracy of the classification model. Given the models above it is evident that the GAN-generated data did not benefit classification accuracy. The failure of the GAN to produce realistic data highlights the limitations of relying on synthetic data in this context. Future work may focus on improving the GAN's capability to generate more accurate data or exploring alternative methods to enhance model performance. We did try changing many of the tuning parameters to get to a dataset that worked, but we were unable to find something that worked well enough. We suspect that one big reason for failure was that the generator was not constructively training, as the discriminator was significantly better at picking fake data than the generator was at making it. If we were to take another stab at the project, we could try dumbing down the discriminator a little, or try to squeeze some more out of the generator so that they balance better. Another fun avenue that we could try to go down would be to use GAN in a more traditional sense of image creation by having it create realistic galaxies for a model to classify.