Classification Between Galaxies and Stars with Multi-layer Perceptron and Random Forest

Wooseok Lee* Department of Energy Resources Engineering,
Seoul National University, Seoul, Korea
Email: *andylws@snu.ac.kr

***Abstract*—In this paper, I propose a quick and concise method to distinguish between galaxies and stars. Thanks to two-layer perceptron, classification between galaxies and stars could be done quickly in the early stages of astronomical research.**

I. Introduction

If you are at somewhere far from artificial lights, you can see a lot of bright spots in the night sky. You might think that all of those far tiny spots would be stars. That is correct, but to be more precisely, if you take a bright point in the sky, it could be a single star but also it could be a cluster of a lot of stars which is called galaxy.

Distinguishing between galaxies and stars is a very important task in starting astronomical research, especially in observational astronomy.

However, those unknown celestial bodies and structures are literally astronomical distances away. Therefore in so many cases, we cannot figure out the physical features of each target with only one photograph. That leads us to take multiple pictures of a target which are taken at a time interval and there, time delay follows. This also applies to the distinction between galaxies and stars.

What if a classifier built by machine learning method can distinguish them only with a single picture?

II. Problem Definition

The goal of this research is to construct a two-layer perceptron and random forest which can successfully recognize whether the object of the given image is a galaxy or a star. Also, it aims to maximize an accuracy of classification. For two-layer perceptron, I checked if the dimension size of hidden layer or the number of training epoch affects significantly in accuracy. For random forest, I checked how the number of trees affects accuracy.

III. Solving Approach

First of all, preprocessing of image data was needed. I converted 10,000 galaxies and stars picture data which are given as PNG files of 512×512 pixel size into PNG files with 28×28 pixels(Figure 1 and Figure 2). This is to reduce time and memory required in calculating. 4,000 pictures for each galaxy and star are used for training and the other 1,000 pictures for each are used for testing.

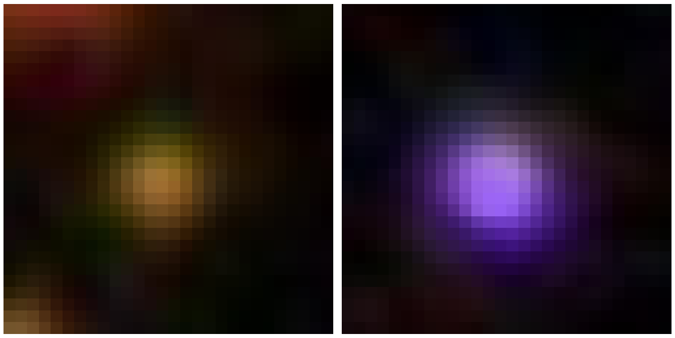

Fig. 1. The sample images of galaxies with 28×28 pixels.

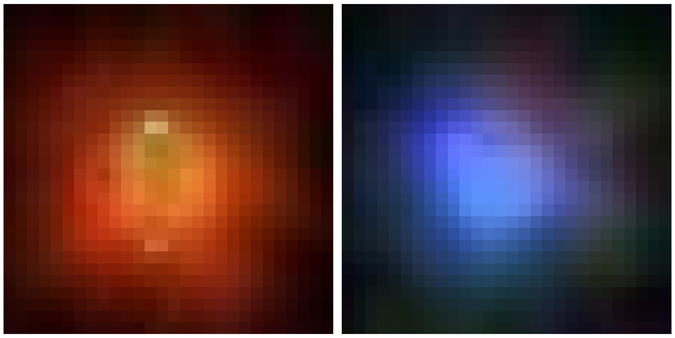

Fig. 2. The sample images of stars with 28×28 pixels.

Then, converting PNG files into CSV files follows. In front of each data, I added target numbers which indicate the real classification of the object; 0s for galaxies and 1s for stars.

Two-layer perceptrons with sigmoid function and a hidden layer with 100 and 200 dimensions are constructed. Each of them are trained with 10 or 20 training epoch. This makes four perceptron structures in total. These perceptrons are compared with each other in terms of accuracy to find out presence or absence of influence of the pixel size and the number of training epoch on accuracy.

Random forest structures are made with various numbers of trees. Each structure has 1, 5, 10, 50, 100, 500 trees. These six random forest structures are compared with each other in terms of accuracy on testing with training dataset and new dataset which was prepared for test to find out the influence of the number of trees on accuracy.

Additionally, to check if the pixel size of image files affects on accuracy, I made PNG files with 56×56 pixels and checked an accuracy with random forest structures.

IV. Evaluation

First, two-layer perceptrons are constructed successfully and four accuracies are calculated. See the Table I.

Training epoch	100 hidden dim.	200 hidden dim.
10	93.45%	93.80%
20	93.65%	94.00%

TABLE I
ACCURACIES FROM EACH TWO-LAYER PERCEPTRON.

Four accuracies from two-layer perceptrons are almost same and are about 93.5% $\sim$ 94%. Accuracies changed slightly as the number of training epoch and the dimension of hidden layer increased, but it is hard to say this amount of difference is noticeable and significant.

Next, random forest structures are constructed successfully and accuracies are calculated as Table II.

The number of trees	Training dataset	Test dataset
1	95.73%	89.34%
5	99.15%	92.80%
10	99.39%	93.60%
50	99.99%	94.50%
100	99.99%	94.50%
500	100%	94.50%

TABLE II
ACCURACIES FROM RANDOM FOREST STRUCTURES TRAINED WITH 28×28 PIXEL IMAGES.

For random forest which used image files with 56×56 pixels, accuracies are calculated as Table III.

The number of trees	Training dataset	Test dataset
1	95.46%	88.74%
5	98.91%	92.80%
10	99.55%	93.55%
50	99.99%	94.40%
100	100%	94.45%
500	100%	94.45%

TABLE III
ACCURACIES FROM RANDOM FOREST STRUCTURES TRAINED WITH 56×56 PIXEL IMAGES.

Accuracies from random forest structures increased as the number of trees increased in the both cases, tested with training dataset and test dataset. In the case of testing with training dataset, accuracies increased to 100%. And in the case of testing with test dataset, accuracies increased to 94.5% and didn't increased anymore. In the cases used the image files with 56×56 pixels, accuracies slightly decreased from the cases used the image files with 28×28 pixels.

V. Conclusion

I constructed concise galaxy-star classification algorithm using two-layer perceptron and random forest. It is encouraging that the best accuracies from each method are higher than 90%. This accuracy is high enough to be used in the light project or the early stages of astronomical research for quick and concise classification of galaxy and star. I hope that this perceptron structure and method will be used by astronomy researchers and save their precious time.

However, it is not recommended to use this in research requiring precision. Because the accuracy is still low, the drastic reduction in error rate is certainly needed. Other kinds of neural network structures and deep learning methods might be the way for it.

Additionally, even though the amount of decrease of accuracy in the change of image pixel size is not significant, the cause of this decrease could be studied.