

Correcting Mislabeled Quasars in Extragalactic Catalogs

ARJUN SHRIVASTAVA

ABSTRACT

Quasars, a type of active galactic nuclei (AGNs), are some of the brightest objects in the universe. They allow astronomers to accurately observe distant objects and look farther back in time, offering researchers a better understanding of our universe's history. However, the X-ray and extragalactic databases that catalog astronomical objects often mislabel quasars as other objects. Therefore, I seek to improve these classifications by identifying quasars in the Deep Fields component of the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS) database and cross-checking them with the extragalactic catalogs. The typical method of classifying an object as a quasar is via manual visual classification: examining unfolded light curves for star-like objects that exhibit irregular short-term and long-term variations in brightness. Unfortunately, this is time-intensive and prone to human error, so I attempted to accelerate the process and boost accuracy with deep neural networks. I first visually classified data of about 4,000 objects for my training set and included four different filters. After preprocessing, I trained a Long Short-Term Memory (LSTM) neural network with different variations of hyperparameters until I achieved an accuracy of 97.5%. When I ran my model through my data, it identified 992 new quasars, 796 of them being actually quasars while the rest were misclassified, yielding an overall binary accuracy of 80% for the entire CFHTLS Deep Field database. Of the identified quasars, 14.8% of which were new quasars and 83.3% were mislabeled in the NASA/IPAC Extragalactic Database (NED). Many of the mislabeled quasars tended to appear as galaxies as well as unidentified sources of ultraviolet or X-ray radiation. In the future, I seek to improve my model for higher accuracy.

1. INTRODUCTION

Quasars are the brightest objects in our universe and a subclass of active galactic nuclei (AGN). Physically, a quasar is an abnormally bright galaxy, a typical quasar is hundreds of times brighter than our own Milky Way. They are extremely bright are due to a supermassive black hole in the center of the galaxy accreting matter around itself. Because of their vast distances, astronomers are able to peek back billions of years in the history of the universe and understand what the early cosmos was like. Unfortunately, many extragalactic databases often mislabel these quasars as other objects, most often as galaxies and stars and occasionally a visual or X-ray source. One of the underlying problems is that quasars are often confused with RR Lyre variables, a well known type of star that varies in brightness periodically, because they both display similar tendencies of varying brightness. Therefore, it is important to develop methods that correctly identify and correct the mislabeling. One of the most promising places to look for quasars is

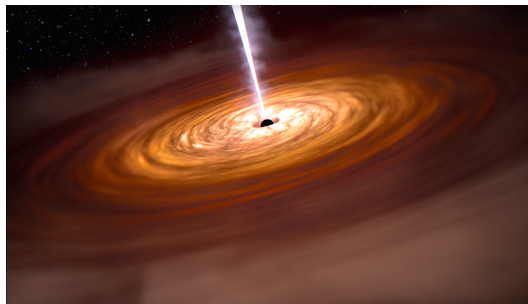


Figure 1. Artist's conception of a quasar

34 in the Deep Fields Component Canada-Hawaii-France Telescope Legacy Survey (CHFTLS), containing about 28000
 35 objects.

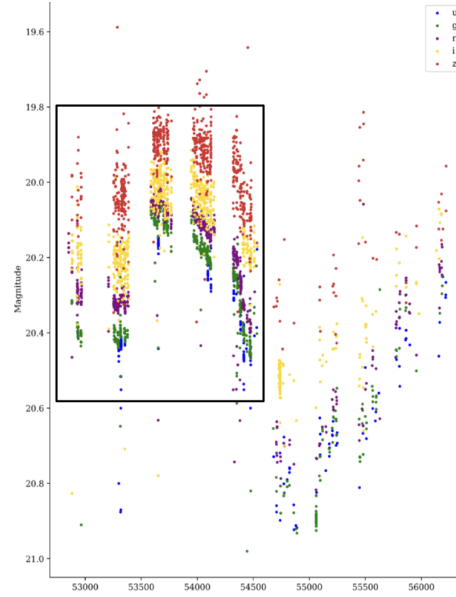


Figure 2. This is the light curve of a quasar (exaggerated). We can see that in all filters (red, green, blue, infrared), the brightness varies irregularly over a long period of time, this can be seen as the different bands shift vertically over time. The y-axis is the apparent magnitude, a measure of how bright an object appears to be. The x-axis is time, modified Julian date.

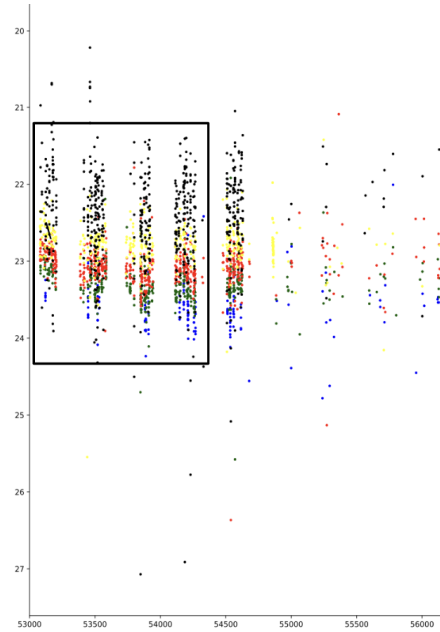


Figure 3. This the light curve of an RR Lyrae Variable Star. We can variation vertically in each band, this indicates variation in brightness over shorter period of time, but over longer periods of time the bands stay parallel to each other.

36
 37 Quasars can be identified by analyzing a graph of their brightness over time. While the graph might show some
 38 similarity when compared to that of a RR Lyare variable star, a quasar will have much greater variation over a longer
 39 period of time. In general, the light curve of an RR Lyre variable will be periodic and nearly sinusoidal, and its period

will be around 1–2 days. On the other hand, quasars tend to vary on a scale of years and exhibit highly irregular variations in brightness. If we plotted the light curves of both of these side by side as shown in figures 2 and 3, we can see the difference more clearly.

A common method is to visually look at the graph and mark each object as a quasar or non-quasar and catalog it. However, this method can be very ineffective and time-intensive for humans to do for thousands of objects, so there have been many statistical methods developed to simplify this process. One method is to take the standard deviation (or root mean squared) of the brightness for a given filter for each object and plot it against the mean magnitude. Then we draw a regression line.

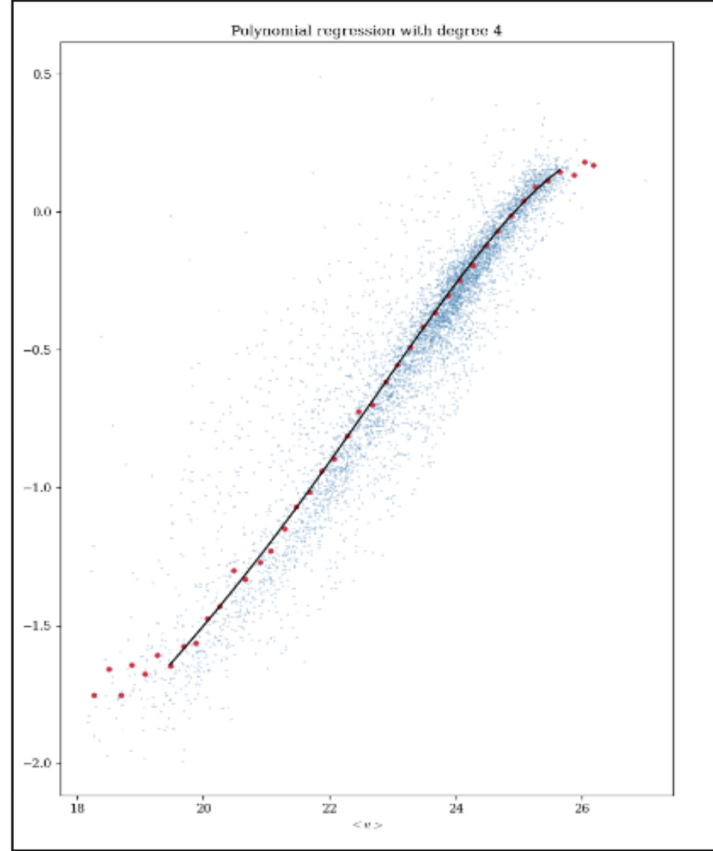


Figure 4. Logarithmic plot of observed RMS vs. median blue magnitude for all 28000 objects. Quasars are expected to lie above the ridgeline (points above the line are too faint to be seen) and can be narrowed down using certain thresholds.

Since quasars display higher variation they are expected to be further above the ridgeline, so we set numerical boundaries and pick objects beyond them. Running this algorithm on the CHFTLS Deep Field Data of 28000 objects shortened this list to about 4000 objects. However, not all of these objects were quasars, manual classification had to be done to find these quasars. Only 730 of them were actually quasars, implying a success rate of about 18%. The low success rate is because variable stars are also likely to be above the ridgeline and get mixed in with the quasars. I seek to improve upon the way quasars are identified by using deep learning and correct the mislabeling of quasars in the NASA Extragalactic Database (NED). Specifically, I seek to use Long-Short Term Memory (LSTM's) neural networks, a type of neural network that can identify short and long term patterns in sequences of data. While neural networks are not new to solving such problems, using them to classify objects based on their time-series is rare.

2. METHODS

In order to train a deep learning neural network, I needed much training data as possible. The data is all files containing information for 28000 objects in the CHFTLS Deep Field, each object has one file. A subset of about 1000 objects was already classified by the previous statistical method and I visually classified 3000 more objects, yielding a total of 4000 objects in our training set. The objects were labelled (1) for non-quasar and (2) for quasar. We used data from 2 of the 6 filters, red and green, as they had the least error in measurement and less noise. I parsed the sequences of magnitudes (a number quantifying brightness) for the red and green filters for all 4000 objects and stored them in a data frame in python. However, after I plotted some of the graphs for a few objects, I realized the data was too noisy and had a very high variance. In order to reduce noise, I took a rolling median of 19 steps, reducing the number of data points while still capturing the overall trend. It is important to use the median rather than mean because the median is not affected by any outliers or errors in measurement. After standardizing the rolling median, the data was ready to be trained. After data preprocessing, I constructed a deep LSTM model using python libraries TensorFlow and Keras. When creating the model, I tested the different hyperparameters, such as the number of epochs, output shape, and learning rate until I was satisfied with the accuracy and loss. My model was a sequential an LSTM layer with an output shape of 128, a dense layer of output shape 1, a learning rate of 0.001, and an activation of sigmoid for output. I used 200 epochs with patience of 50.

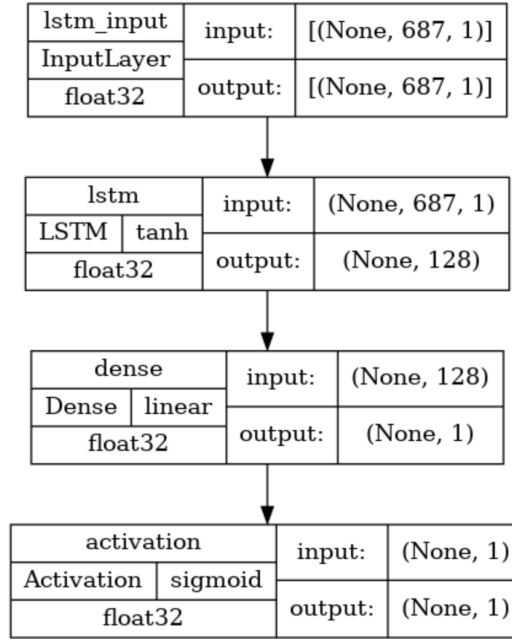


Figure 5. Diagram of the deep neural network model

3. RESULTS

Overall, the total binary accuracy was 97.5% with a loss of 0.0796. Overall, this means the model is expected to correctly differentiate between quasars and non-quasars with a 97.5% accuracy. Compared to the traditional method, which has an accuracy of 18% and requires manual intervention, this is an extremely good result. The model is also significantly faster than the previous statistical method used due to its simplicity, it was able to classify all the objects in the database in 2–3 minutes. Overall, the model predicted 992 quasars in the database. I decided to check the model's accuracy on the database to check if the model was performing as expected. It turns out that 796 out of the 992 quasars were actually quasars, the rest were misclassified. This yields an overall accuracy of 80%. The pie chart summarizes the distribution of the 796 correctly classified quasars.

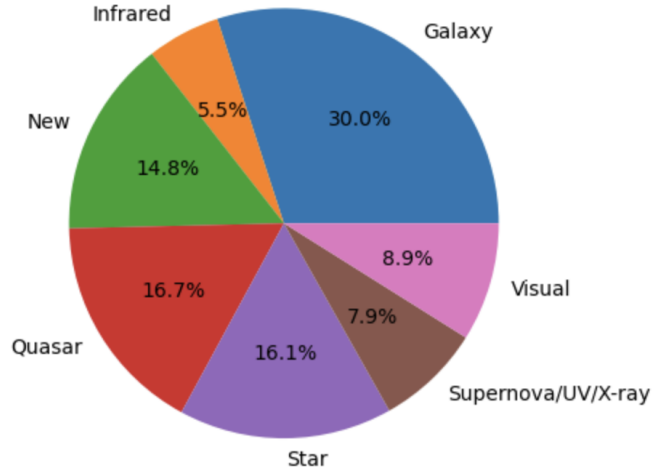


Figure 6. The pie chart shows the distributions of the classifications of the correctly identified quasars in NED.

4. DISCUSSION

While the accuracy during training and testing was 97.5%, the accuracy seemed to decrease to 80% when the model classified. While 80% is still significantly better than the accuracy of the traditional statistical method, 18%, the model can still be improved. One of the main problems is how the data points are different for each object is. While most objects have data that contain at least a 200-300 data points for all the filters combined, some data points contain very few, some as little as 20. Most of the objects the model incorrectly classified were non-variable stars, and they contained significantly fewer data points than normal. The model was mostly trained on data that was less noisy and contained more data points, so it makes sense that it is less accurate in classifying objects with fewer data points. I plan to fix this problem by including these weaker samples.

Another pattern I noticed was that my model had difficulty differentiating objects whose light curves appeared to be RR Lyres and quasars at the same time. This pattern can also be seen in the distribution of incorrectly classified quasars, 92 objects listed as stars in NED were misclassified, yielding an accuracy of 58.20% in the category. This is most likely due to the fact that the objects in the training data were given labels by humans who might classify several such objects as quasars and some as stars, contradicting the patterns that the model identifies and confusing it.

The model had a fairly high accuracy (90.2%) in classifying objects mislabeled as galaxies. 148 objects (14.8%) are new quasars and 662 (83.3%) objects were corrected by the model since they were given labels other than quasars. If we look at the correctly classified quasars, we notice that the highest number of these objects (30%) were listed as just galaxies in the NED database. This misclassification in the database is likely due to the fact that a quasar is basically a galaxy, but much brighter and with an active supermassive black hole in the center. A far away quasar may appear to be a normal galaxy and exhibit similar characteristics.

While using the moving median method to reduce noise and increase accuracy has yielded sufficient results, in the future I plan to use a Fourier transformation on the data before classification to better extract the different patterns of quasars and RR Lyres. The overall accuracy of 80% suggests that LSTM models can be used in differentiating and classifying different celestial objects that exhibit similar seasonality.

5. ACKNOWLEDGEMENTS

I would like to thank the Science Internship Program (SIP), hosted by the University of California, Santa Cruz, for providing access to the CFHTLS database and unfolded light curves. Preliminary research, including visually identifying quasars, examining RMS, and cross-referencing objects with an extragalactic database was conducted during SIP.

I would like to acknowledge Amanda Guo for collaborating with me in preparing the model and the training data. I would also like to acknowledge Puragra GuhaThakurta, Manjima Talukdar, Spencer Jaseph, Joanne Zhao, Christopher Donnelly, Casey Peters, Talise Oh, Aparajito Bhattacharya, Iona Xia, Kareena Gupta-Martinez, Ethan Ubando, Diya Lansberg.

6. REFERENCES

Aas (IPosterSessions - an AMuze! Interactive System), <https://aas241-aas.ipostersessions.com/Default.aspx?s=CE-26-B7-46-52-5C-34-F6-EF-30-E1-7E-50-E4-BF-46>.

“NASA/IPAC Extragalactic Database.” Retrieve Data for Near- Object/Position List, ned.ipac.caltech.edu/forms/nnd.html. Accessed 9 June 2023.