IDENTIFY EXOPLANET CANDIDATES WITH DEEP LEARNING MODELS

A THESIS

SUBMITTED TO THE DEPARTMENT OF PHYSICS

OF THE UNIVERSITY OF HONG KONG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Wenchao Wang

August 2021

Abstract of thesis entitled

IDENTIFY EXOPLANET CANDIDATES WITH DEEP LEARNING MODELS

Submitted by

Wenchao Wang

for the degree of Doctor of Philosophy

at the University of Hong Kong

in August 2021

The work is about identifying planet candidates using deep learning models. Finding these objects manually is a very labor intensive task. For example, *The Large Synoptic Survey Telescope (LSST)* is expected to generate about $200,000$ images per year, which is equivalent of more than 10^6 GB of data. Therefore using reliable algorithms to manage the data is necessary. Deep learning can be helpful because it suits well for very large input data. In general, having more data only makes deep learning models perform better.

We use *Kepler Space Telescope* and *Transiting Exoplanet Survey Satellite* (*TESS*) to detect planet candidates by using a convolutional neural network model. We apply the Q1-Q17 (DR24) table as our training and test sets. The model takes two phase-folded light curves and some parameters of each transit-like signal and then outputs whether the signal represents a planet candidate (PC), a non-transiting phenomena (NTP) or a false positive (FP). In the current model, we feed 17 features into a dense

neural network model, such as transit durations and depth of signals. At this stage, the models achieve AUROC and accuracy of about 97.7%, 95.9% respectively for the test set. The accuracy for the training set can be over 99%, which means that the model can easily overfit the data. The most straightforward way to the problem is to use more data to train the model. Therefore, we plan to train it with more simulated data later in order to increase the AUROC and accuracy of predictions.

Identify Exoplanet Candidates with Deep Learning Models

Department of Physics, The University of Hong Kong, Pokfulam Road, Hong Kong

Wang Wenchao

3030053350

Declaration

I hereby declare that this whole dissertation report is my own work, except the parts with due acknowledgment, and that it has not been previously included in a thesis, dissertation or report submitted to this University or to any other institution for a degree, diploma or other qualifications.

Signature: _____

Name: Wenchao Wang

Date: August 2018

Acknowledgments

Lots of people give me much precious help. First of all, I really appreciate my supervisor Dr. Stephen Chi Yung Ng for his great support, patience and kindness. Prof. Takata also provides me much help in term of pulsars' emission mechanisms and numerical simulations and is really charming. In addition, I thank Ms. Ruby Cho Wing Ng for her invaluable guidance and advice. They give me not only knowledge and techniques but also encouragements. Therefore, I would like to express my most sincere thanks to them.

Abstract

The work is about identifying planet candidates using deep learning models. Finding these objects manually is a very labor intensive task. For example, *The Large Synoptic Survey Telescope (LSST)* is expected to generate about $200,000$ images per year, which is equivalent of more than 10^6 GB of data. Therefore using reliable algorithms to manage the data is necessary. Deep learning can be helpful because it suits well for very large input data. In general, having more data only makes deep learning models perform better.

We use *Kepler Space Telescope* and *Transiting Exoplanet Survey Satellite* (*TESS*) to detect planet candidates by using a convolutional neural network model. We apply the Q1-Q17 (DR24) table as our training and test sets. The model takes two phase-folded light curves and some parameters of each transit-like signal and then outputs whether the signal represents a planet candidate (PC), a non-transiting phenomena (NTP) or a false positive (FP). In the current model, we feed 17 features into a dense neural network model, such as transit durations and depth of signals. At this stage, the models achieve AUROC and accuracy of about 97.7%, 95.9% respectively for the test set. The accuracy for the training set can be over 99%, which means that the model can easily overfit the data. The most straightforward way to the problem is to use more data to train the model. Therefore, we plan to train it with more simulated data later in order to increase the AUROC and accuracy of predictions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Transit Photometry

A planet is too faint to be found directly by telescopes, so people study it by observing its host star. Since planets orbit around their host stars, they have some periodical effects on the stars such as gravitational pull and light blocking. When a exoplanet pass in front of its host star, it blocks a very small fraction of the star. By measuring the drops of the starlight, basic stats of the planet can be obtained such as period and size.

Because planets are in general very tiny compared to host stars, the drops of the starlight are very small. Therefore, some statistic methods are used to identify transits. The confirmed transits that passed statistic tests are called TCE (Threshold-Crossing Events). In *Kepler* DR24 data products (the data used in this thesis), the mean and median of TCE are about 12516 ppm and 125 ppm respectively (ppm stands for Parts Per Million). The distribution of TCE depths is shown in the following figures.

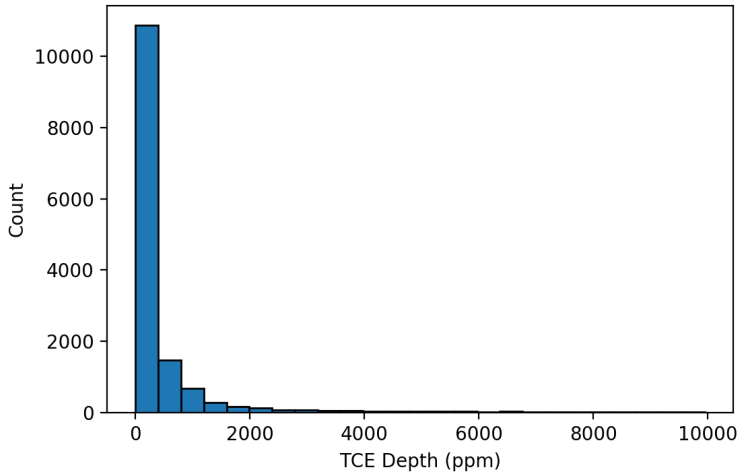

Figure 1.1: The distribution of TCE depths of *Kepler* DR24 data. There are 20367 TCE in the data and only 1831 are larger than 10000 ppm (less than 10%). The histogram shows that most of TCE depths are smaller than 500 ppm.

1.2 *Kepler Space Telescope*

Kepler Space Telescope is a space telescope for finding Earth-like terrestrial exoplanets. It was launched on March 7th, 2009 and was retired in late 2018. During its service, *Kepler Space Telescope* observes more than 500 thousand stars and finds about 2600 exoplanets.

The *Kepler* pipeline is a set of programming tools which can generate calibrated light curves which can then be fed into the algorithms to discriminate between planet candidates (PC) and other types of light curves. Threshold Crossing Events (TCE) are generated after Transiting Planet Search (TPS) which is part of the *Kepler* pipeline. And then we can identify planet candidates by analyzing the generated TCE.

Kepler team developed a tool called *Kepler Robovetter* [1] to identify PCs and false positives and it achieves a high accuracy (over 97%). It is a traditional algorithm

[1] https://github.com/nasa/kepler-robovetter

which is purely written in C++ and is very fast. However, it is difficult to understand and one can view its source code on the github. Therefore, try different method like machine learning may be a feasible choice.

1.3 Machine Learning

Machine learning is a subset of artificial intelligence and has been successfully used in many areas for a variety of tasks such as self-driving cars and optical character recognition (ORC) which can extract words and their meanings from images. Machine learning methods can even generate art works such as pictures which are nearly indistinguishable from art works created by artists. Therefore, it is natural to think that machine learning methods can also be applied to astronomy.

Machine learning can be classified differently based on different criteria. For example, if we predict a product's price which is a concrete value, we are doing a regression task. Otherwise, if we need to classify pictures as dogs or cats, then it is a classification task. Meanwhile, an algorithm is called supervised learning if we train the machine learning model by giving corresponding labels. Otherwise, it is a non-supervised learning algorithm. In this thesis, we input the TCE with corresponding category labels generated by *Kepler* pipeline and output a number 0 or 1 to indicate PC or non-PC. Thus, we use a supervised learning method to do a classification task.

There are a large quantity of machine learning algorithms for classification tasks such as logistic regression, decision trees and support vector machines. However, in this thesis, even though we are going to try these machine learning methods, we focus more on deep learning methods, especially convolution neural networks (CNN).

1.3.1 Deep Learning

more

Chapter 2

Data Preparation

As previously mentioned, we use *Kepler* DR24 data [1] with labels to train out model. It is a CSV file while other types are also supported. Note that the data contains training labels generated by autovetter which is basically a random forest technique. Random forest is also a supervised machine learning algorithm, thus it also needs labels to tell the algorithm which one is PC. The labels fed into random forest algorithm are classified by humans. The reason why use deep learning method to redo the classification task is because deep learning performs very well when trained with large amount of data. Therefore, we try to use the human-made and normal machine learning method generated labels to train a deep learning model in order to get better accuracy.

The labels contain four unique values: PC (Planet Candidates) , AFP (Astronomical False Positive), NTP (Non-transiting Phenomenon) and UNK (Unknown). There are 27 columns in the data including the label column named "av_training_set". Each other column represents a property of one TCE and some of the columns are more important for our classification task, such as "tce_period", "tce_duration", "tce_prad", "tce_depth", etc. Meanings of the some column names are listed in the

[1] https://exoplanetarchive.ipac.caltech.edu/docs/Kepler_TCE_docs.html

following table. [2]

Column Name	Definition
tce_period	time interval between two consecutive transits in days
tce_depth	starlight drops in ppm
tce_prad	the planet radius in Earth Radii

Table 2.1: Descriptions of column names. The listed properties are more important for identifying PC.

The CSV table needs to be preprocessed. First of all, we need to drop all data labeled with 'UNK'. Then since deep learning algorithms only deal with numerical labels, we need to label PC and non-PC as 1 and 0 respectively. Non-PC includes both AFP and NTP because we do binary classification to find out planet candidates. After processing the basic properties for each TCE, we also need to download their light curves.

2.1 Download Light Curves

Light curves are downloaded based on the column name "kepid" in the *Kepler* DR24 data table previously discussed. All the light curves can be found on the website *https://archive.stsci.edu/pub/kepler/lightcurves/0020/* . However, there are 20367 TCE in the table containing more than 330,000 light curve files. Downloading these files can be very time consuming (about 2 weeks by estimation). Therefore, I scrape the webpage in parallel to filter the light curves data and download them.

Take kepid 000757137 as an example. The URL of the light curves for this TCE is *http://archive.stsci.edu/pub/kepler/lightcurves/0007/000757137* . There are 17 light curve files as the following screen shot shown.

[2] https://exoplanetarchive.ipac.caltech.edu/docs/API_tce_columns.html

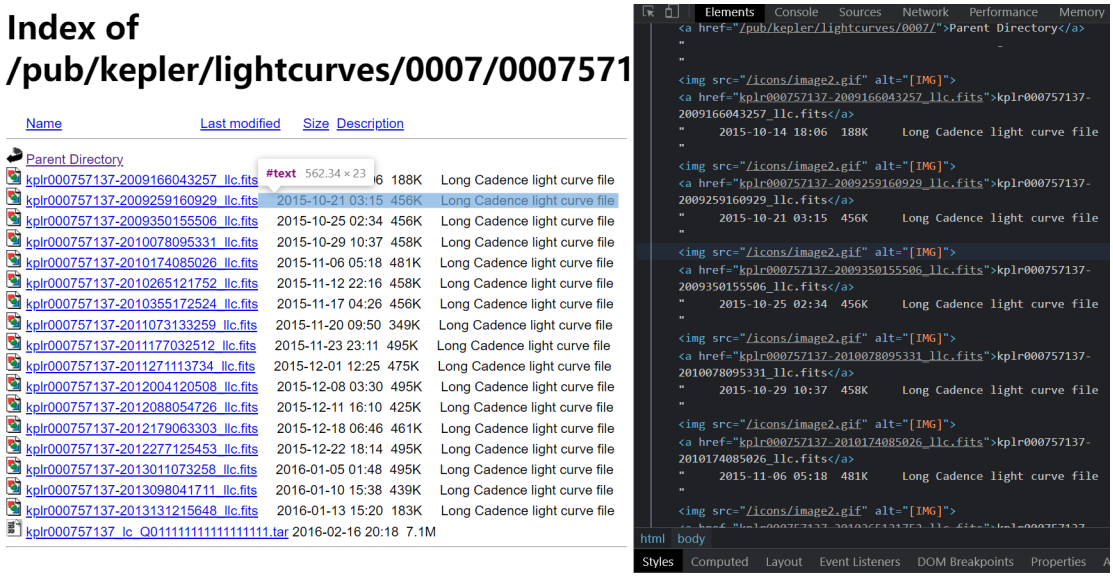

Figure 2.1: The screen shot of webpage containing light curves of kepid 000757137

We can find the hyper link from the anchor tag and navigate to the URL of the desired data. The code is written in Golang, which is very good at concurrency. All code can be found on *github* . In short, the basic procedure is

1. Get all the hyper links from the root URL *http://archive.stsci.edu/pub/kepler/lightcurves/* . Join the current URL with the parsed hyperlinks to generate the new URL.

2. Launch a new goroutine to search from that webpage.

3. Get the hyper links from the child URL. If the hyper links are fits files, then download them to a local directory. Otherwise, like the procedure 1, get the new URL and launch a new goroutine to do the search recursively.

Bibliography