ГРАВИТАЦИЯ, КОСМОЛОГИЯ И ФУНДАМЕНТАЛЬНЫЕ ПОЛЯ

***

УДК 524.3, 004.8

© Бикмурзин Р. В., Федченко Д. П., 2020

МЕТОД ПОСТРОЕНИЯ КЛАССИФИКАТОРА ЗВЕЗД ИЗ КАТАЛОГА GAIA DR2 НА НАЛИЧИЕ ЭКЗОПЛАНЕТ

Бикмурзин Р. В.[a,1], Федченко Д. П.[a,2]

[a] Сибирский федеральный университет, г. Красноярск, 660041, Россия.

В данной работе предложена методика построения классификатора звезд по наличию экзопланет (ЭП) на основе информации из звездного каталога GAIA DR2 и данных, полученных в ходе наблюдений с космического телескопа «Кеплер». Три типа классификаторов были разработаны с помощью методов машинного обучения таких как: логистическая регрессия, случайный лес и искусственная нейронная сеть. Сравнительная оценка результатов классификаторов проведена с помощью матрицы ошибок.

Ключевые слова: экзопланеты, классификация звезд, нейросеть, случайный лес, звездный каталог GAIA.

METHOD FOR CONSTRUCTING A CLASSIFIER OF STARS FROM THE GAIA DR2 CATALOG FOR THE PRESENCE OF EXOPLANETS

Bikmurzin R. V.[a,1], Fedchenko D. P.[a,2]

[a] Siberian Federal University, Krasnoyarsk, 660041, Russia.

In this paper, we propose a method for constructing a classifier of stars by the presence of exoplanets (EPs) based on information from the GAIA DR2 star catalog and data obtained during observations from the Kepler space telescope. Three types of classifiers have been developed using machine learning methods such as logistic regression, random forest and artificial neural network. Comparative evaluation of the results of the classifiers was carried out using the error matrix.

Keywords: exoplanet, stellar classification, neural network, random forest, Gaia catalogue.

DOI: 10.17238/issn2226-8812.2020.4.4-9

Введение

Космический телескоп «Кеплер» за 9 лет службы на гелиоцентрической орбите с 7 марта 2009 года по 15 ноября 2018 открыл более двух тысяч экзопланет, еще около двух тысяч кандидатов в экзопланеты ждут своего подтверждения другими научными группами. Другой до сих пор действующий телескоп «Гайя» (Gaia) был отправлен во вторую точку Лагранжа системы Земля – Солнце европейским космическим агентством 19 декабря 2013 года. Перед данным проектом было поставлено две научных задачи: прояснить возникновение и развитие нашей Галактики (измеряя с высокой точностью параллаксы, собственную и радиальную скорости более миллиарда звезд), а

[1] E-mail: bikmurzin.sfe@gmail.com
[2] E-mail: fdp@bk.ru

также обнаружение ЭП. В качестве одного из промежуточного результатов работы космической программы в 2018 году был выпущен второй релиз звездного каталога Gaia DR2, содержащий более чем 1,7 млрд. космических объектов.

Один из методов открытия у звезд ЭП — это транзитный. Он основан на обнаружении изменения светимости звезды в момент прохождения планеты на ее фоне. Чтобы увеличить вероятность фиксации данного явления, необходимо вести постоянные, периодические наблюдения за большим количеством звезд. В нашей Галактике по разным оценкам находится от 200 до 400 млрд. звезд, а значит и потенциальных объектов для наблюдений. Такой огромный выбор ставит перед «охотниками» за ЭП важный вопрос: «За какими звездами лучше вести наблюдения, чтобы результат их работы был более плодотворным?» Выбор объекта наблюдений может играть одну из ключевых ролей для успешного поиска ЭП.

В данной работе предлагается метод построения классификатора, который поможет выбрать звезды с более вероятным обнаружением планет на их орбите, на основе информации содержащейся в каталоге Gaia DR2 и уже подтвержденных ЭП, открытых с помощью телескопа «Кеплер».

1. Постановка задачи классификации

Пусть X — множество звезд, $Y = \{0, 1\}$ — множество меток классов (класс определяет наличие планеты на орбите звезды, 0 — нет, 1 — да), $\hat{y} : X \to Y$ — целевая функция, известная только на объектах обучающей выборки $X^l = (x_i, y_i)_{i=1}^l$, $y_i = \hat{y}(x_i)$. Требуется построить алгоритм классификации $a : X \to Y$, аппроксимирующий $\hat{y}$ на всем множестве X.

Из поставленной задачи следует очевидный вопрос: «Где взять множество меток классов определяющих наличие ЭП?».

Для ответа на данный вопрос, можно воспользоваться виртуальным планетарием Stellarium. Если подсветить в этой программе открытые ЭП, то можно заметить, что на небольшом участке неба, между созвездиями Лебедя и Лиры (рис. 1), космическим телескопом «Кеплер» обнаружены экзопланетарные системы с очень высокой плотностью относительно всей поверхности небосвода. Поэтому предположим, что те звезды, которые находились в этом поле зрения «Кеплера» (рис. 2) и на орбитах которых были обнаружены и подтверждены ЭП, принадлежат классу-1, а звезды, у которых на данный момент не было обнаружено (или подтверждено) наличие ЭП на этом участке неба принадлежат классу-0.

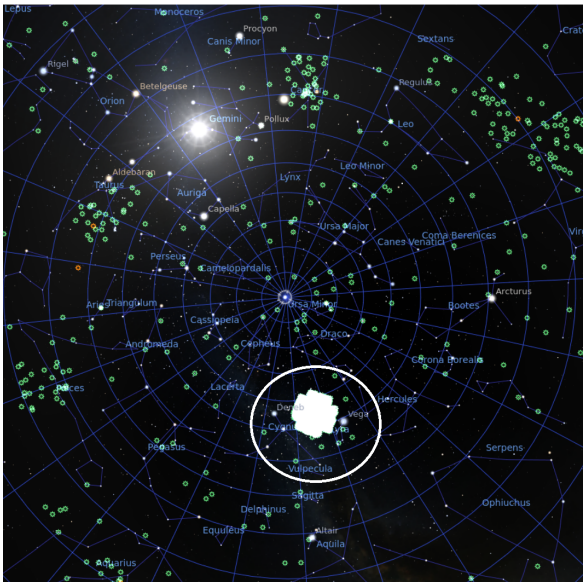

Рис. 1. Местоположения ЭП на небе в виртуальном планетарии Stellarium (отмечены зелеными восьмиугольниками).

2. Предобработка данных

Чтобы получить список всех звезд с их характеристиками на выделенном участке были скачаны файлы с данными и загружены в российскую СУБД ClickHouse из каталога GAIA DR2 (http://www.esa.int/science/gaia). Но в этом звездном каталоге информация о наличии ЭП отсутствует, поэтому с сайта NASA (https://exoplanetarchive.ipac.caltech.edu) был скачан список всех подтвержденных на данный момент планет, что позволило добавить к звездному каталогу дополнительную информацию о наличии ЭП. Так как в каталоге ЭП не было идентификатора звезд, который позволил бы однозначно сопоставить звезды и каталога GAIA, поэтому пришлось сопоставлять звезды по координатам из обоих каталогов. Так же была учтена информация, что вокруг одной звезды может находиться несколько ЭП.

Для дальнейшего изучения был выбран участок неба между созвездиями Лиры и Лебедя в границах координат:

$$Ra = (283,5^\circ; 298,1^\circ)\,, \qquad Dec = (39,2^\circ; 49,5^\circ) \tag{1}$$

где Ra — прямое восхождение, град; Dec — склонение, град (рис. 2). На выбранном участке согласно информации из GAIA DR2 находится 5 173 467 объектов.

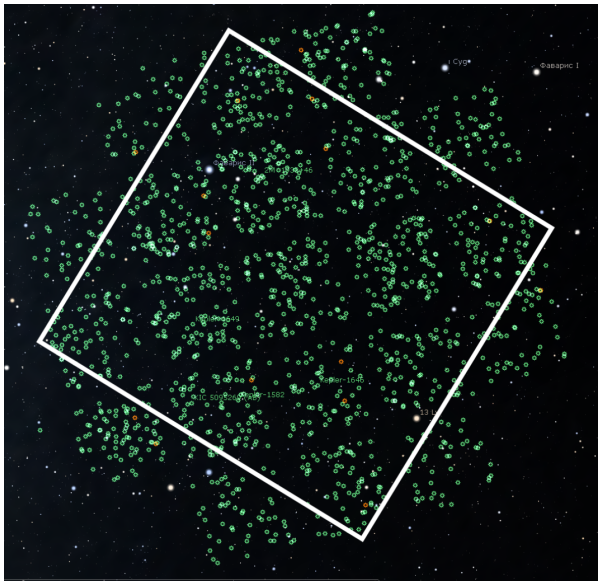

Рис. 2. Местоположения ЭП открытые «Кеплером» между созвездиями Лиры и Лебедя в виртуальном планетарии Stellarium (отмечены зелеными восьмиугольниками).

После предварительного анализа данных, содержащихся в каталоге GAIA, из 94 параметров были выбраны пять, определяющие наблюдательные и теоретические характеристики звезды ($parallax$, $phot_g_n_obs$, $phot_g_mean_mag$, $teff_val$, lum_val, описывающие значения параллакса, количество наблюдений и среднюю звездную величину в соответствующей полосе пропускания, эффективную температуру и светимость звезд соответственно). Была проведена проверка данных на пропущенные значения в ячейках, если звезды содержали пропуски по какому-либо признаку или имели отрицательный параллакс, то они удалялись из выборки. В итоге, после предварительной чистки информации, на исследуемом участке неба осталось 458 760 объектов, удовлетворяющих всем нашим критериям. Из них 892 звезды с ЭП и 457 868 без них (рис. 3). Соотношение количества звезд с ЭП к звездам без ЭП составило: $892/457868 = 0,001948$.

Проблема при моделировании классификаторов на данных с большим дисбалансом классов (один класс составляет доли процента от другого) заключается в том, что модель может отнести все объекты к численно превосходящему классу, показывая высокую точность, но допуская при

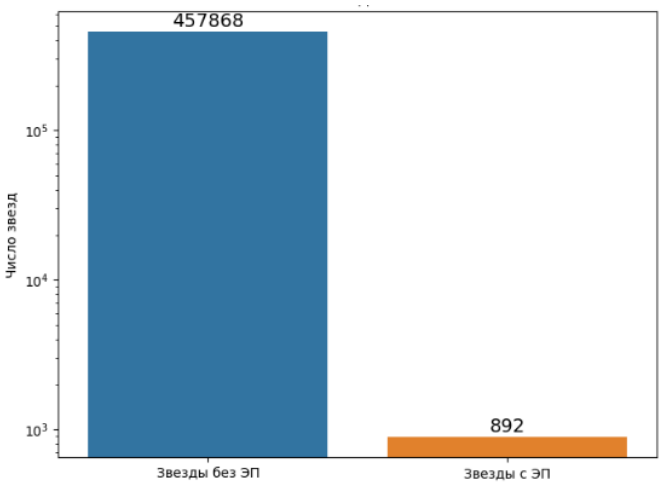

Рис. 3. Дисбаланс классов звезд с ЭП $(99,81\%)$ и без ЭП $(0,19\%)$.

этом ошибки только на малочисленном классе. Таким образом, смысл в работе данного классификатора теряется, т.к. в таких случаях важнее идентифицировать объекты, которые относятся к малочисленному классу, допуская при этом как можно меньше ложноположительных ошибок. В нашем случае задача выявления звезд с ЭП сводится к поиску модели, которая обнаруживает наибольшее количество звезд с ЭП при минимальном количестве звезд, где ЭП не обнаружены, но были ошибочно отнесены к классу 1 (ложноположительная ошибка). Следует отметить, что классификаторы с большим дисбалансом классов разрабатываются и уже применяются в разных областях, так например, в медицине при разделении пациентов на здоровых и больных какой-нибудь редкой болезнью, а в банковской сфере, когда пытаются детектировать мошеннические трансакции среди большинства легитимных. Существует несколько наиболее популярных методов решения такой проблемы с балансировкой данных: случайное удаление примеров мажоритарного класса, дублирование примеров миноритарного класса, алгоритм создания синтетических объектов SMOTE, немного отличающийся алгоритм ASMO, поиск связей Томека и другие [1, 2].

Для оценки предсказывающей способности классификатора в качестве тестовой выборки были случайно выбраны 30% объектов из очищенного набора данных. Таким образом в тренировочной части осталось 321 132 объекта (из которых с ЭП 624 звезды), а в тестовой выборке 137 628 звезд (268 с ЭП). Разделение на тренировочную и тестовую выборки было проведено с сохранением отношения числа звезд с ЭП и без.

3. Модели классификаторов

С помощью предложенной идеи были построены 3 вида классификаторов, на основе логистической регрессии, случайного леса и нейронной сети. Лучшие результаты продемонстрировали модели случайного леса и нейронной сети. Для оценки качества классификаторов использовалась матрица ошибок (рис. 4), которая представляет собой сводную таблицу действительных значений меток классов и значений меток и предсказанных моделью классификатора.

С помощью логистической регрессии мы смогли построить модели классификаторов с показателями TP и TN, которые одновременно превышают значение 0,7. При подборе порога отсечки модели идентифицируют 90% объектов 1-класса (звезд с ЭП), но при этом для 0-класса правильно определяются только 67% звезд. Таким образом на одну обнаруженную звезду приходится от 149 до 190 звезд, ошибочно идентифицированных как 1-класс (ложноположительная ошибка).

Используя случайный лес для построения классификатора, отношение ложнопложительных ошибок (FP) к правильно определенным объектам с ЭП (TP) удалось снизить до 135, но доля верно классифицированных звезд с ЭП (1-класс) снизилась до 72%.

		Classifier Prediction	
		Positive	Negative
Actual Value	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Рис. 4. Матрица ошибок.

Выбрав для пробы несложную архитектуру нейронной сети (рис. 5), классификатор показал почти такие же результаты, что и случайный лес, но коэффициент отношения правильно определенных объектов к ложноположительным был немного улучшен. На одну звезду, правильно классифицированную как звезда с ЭП, приходится 133 звезды без подтвержденных ЭП. Модель нейросети была экспортирована для дальнейшего использования к другим объектам.

тип слоя	размерность слоя
входящий вектор	5
полносвязный слой	10
tanh	
Dropout (0.1)	
полносвязный слой	10
ReLU	
Dropout (0.1)	
полносвязный слой	2
tanh	

Рис. 5. Архитектура нейронной сети.

После применения классификатора на основе нейронной сети к 24 698 432 звездам в радиусе 1000 св. лет от Земли из каталога GAIA DR2 была получена выборка из 1 877 009 звезд, у которых классификатор определил возможное наличие экзопланет.

Заключение

В данной работе была предложена методика построения классификатора звезд по наличию экзопланет на основе объединенной информации содержащейся в звездном каталоге GAIA DR2 и данных, полученных в ходе наблюдений с космического телескопа «Кеплер». Полученные классификаторы позволяют выбирать объекты для наблюдений, которые имеют более высокие шансы для обнаружения ЭП.

Дальнейшее изучение данной темы возможно относительно улучшения эффективности показателей классификаторов, т.к. на тестовой выборке на одну звезду с ЭП классификаторы находят еще дополнительные 133 звезды (ложноположительная ошибка). Это возможно, используя другую архитектуру нейронной сети или изменяя гиперпараметры модели. Следует отметить, что в данной работе вводилось предположение отсутствия экзопланеты только из факта отсутствия данной звезды в списках подтвержденных ЭП (что, строго говоря, не говорит о том, что около данных звезд нет ЭП, а только то, что на данный момент они там не были зафиксированы), то среди ложноположительных результатов также возможно обнаружение ЭП, а значит значение эффективности классификаторов фактически может быть выше полученных.

This work has made use of data from the European Space Agency (ESA) mission Gaia

(https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

Список литературы

1. Шолле Ф. *Глубокое обучение на Python.* Санкт-Петербург: Питер, 2018. 400 с.

2. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 2002, vol. 16, issue 1, pp. 321–357.

References

1. Chollet F. *Deep Learning with Python.* NY: Manning Publications Co., 2017. 384 p.

2. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 2002, vol. 16, issue 1, pp. 321–357.

Авторы

Бикмурзин Рустам Василович, Сибирский федеральный университет, пр. Свободный, д. 79, г. Красноярск, 660041, Россия.
E-mail: bikmurzin.sfe@gmail.com

Федченко Дмитрий Петрович, канд. физ.-мат. наук, Сибирский федеральный университет, пр. Свободный, д. 79, г. Красноярск, 660041, Россия.
E-mail: fdp@bk.ru

Просьба ссылаться на эту статью следующим образом:
Бикмурзин Р. В., Федченко Д. П. Метод построения классификатора звезд из каталога GAIA DR2 на наличие экзопланет. *Пространство, время и фундаментальные взаимодействия.* 2020. № 4. С. 4—9.

Authors

Bikmurzin Rustam Vasilovich, Siberian Federal University, Svobodny pr., 79, Krasnoyarsk, 660041, Russia.
E-mail: bikmurzin.sfe@gmail.com

Fedchenko Dmitriy Petrovich, Ph.D., Siberian Federal University, Svobodny pr., 79, Krasnoyarsk, 660041, Russia.
E-mail: fdp@bk.ru

Please cite this article in English as:
Bikmurzin R. V., Fedchenko D. P. Method for constructing a classifier of stars from the GAIA DR2 catalog for the presence of exoplanets. *Space, Time and Fundamental Interactions*, 2020, no. 4, pp. 4–9.