一种基于梯度提升回归树的系外行星宜居性预测方法

朱维军[1] 王 鑫[1] 钟英辉[2] 樊永文[1] 陈永华[1]
(郑州大学信息工程学院 郑州 450001)[1] (郑州大学物理工程学院 郑州 450001)[2]

摘 要 系外行星的宜居性是近年来探索宇宙的一个热点研究课题,机器学习为系外行星宜居性分类提供了一种可行的手段。然而,现有的宜居性分类效果面临严重不足与局限。为此,给出一种基于梯度提升回归树的系外行星宜居性分类预测方法。首先,使用梯度提升回归树算法对系外潜在宜居行星与非宜居行星的相关物理学与天文学数据集进行训练;然后,利用训练好的模型对相关测试集进行预测。仿真实验结果表明,新方法在测试集上的预测准确率高达 100%。

关键词 梯度提升回归树,系外行星,宜居性,二分类
中图法分类号 TP181,P144 **文献标识码** A

Habitability Prediction of Exoplanets Based on GBRT Algorithm

ZHU Wei-jun[1] WANG Xin[1] ZHONG Ying-hui[2] FAN Yong-wen[1] CHEN Yong-hua[1]
(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)[1]
(School of Physical Engineering, Zhengzhou University, Zhengzhou 450001, China)[2]

Abstract The habitability of exoplanets is a hot research topic in the field of the exploration of the universe in recent years. The Machine Learning(ML) technique provides a viable means for classifying exoplanets according to their habitability. However, the existing ML-based approaches of habitability classification have some serious shortcomings and limitations. To this end, this paper provided a novel method for predicting the habitability of exopla-net based on Gradient Boosted Regression Trees(GBRT). First, the physical and astronomical data on the potentially habitable exoplanets and the inhabitable ones are employed to train by algorithm GBRT. Then, the trained model is used to predict the habitability of the exoplanets in our test set. The simulated experimental results show that the predictive accuracy of the new method is as high as 100%.

Keywords Gradient boosted regression trees, Exoplanet, Habitability, Binary classification

1 引言

探索宇宙与寻找外星生命始终是千百年来人类科学探索的永恒主题与重大课题之一,吸引着无数人的兴趣。从目前的科学认识水平来看,外星生命如果存在,它们最大可能藏身的天体是太阳系之外的行星。在此背景下,研究系外行星是否宜居,成为宇宙探索领域的一个热点与前沿。给定一颗系外行星的物理性质与天文性质,研究人员需要据此分析该星的宜居性。

一方面,近三年被发现的系外行星数量不断增加,截至 2018 年已超过 3000 个,系外行星宜居性分析这一问题需求也随之被提上日程;另一方面,这样的分析需要一种自动化、智能化的辅助方式,近年来火热的人工智能与机器学习技术正好为此提供了一类方法与手段。

2016 年,Bora 等[1]提出使用 K 近邻分类算法分析系外行星宜居性。2018 年初,Saha 等[2]对其进一步改进,使用基于 XGBoost 的提升树算法,根据行星表面温度等若干性质,把行星按照宜居性的不同划分为 5 类。门户人方法对于 Proxima b 这样的系外行星,做到了(计算层面)百分之百的精准分类[2]。2018 年初,Hora[3]做了一项 benchmark 式的研究,其对所有当时已知的潜在宜居型系外行星,辅以十余倍数量的非宜居型系外行星,组成样本集,分别使用 6 种机器学习算法对训练集进行训练,然后预测有关行星是否宜居的二分类结果,效果优异[3]。需要说明的是,本文谈到的"宜居",除非特别说明,否则特指天文学家定义的"潜在宜居",例如,其中一种定义是"质量和从母星得到的辐射与地球接近"[4]。

然而,这些研究本身也存在各自的局限与不足(具体参见第 5 节),为此,我们使用与之不同的一种机器学习方法——梯度提升回归树(Gradient Boosted Regression Trees,GBRT)算法[5]对系外行星的宜居性实施二分类。

2 基础知识

2.1 梯度提升回归树算法与 Graph Lab

机器学习的一个核心目标是对输入数据进行分类。其

本文受国家自然科学基金(U1204608)资助。

朱维军(1976—),男,博士后,副教授,CCF 高级会员,主要研究方向为人工智能及其多学科应用,E-mail:zhuweijun@zzu.edu.cn;王 鑫(1997—),男,主要研究方向为人工智能;钟英辉(1987—),女,博士,副教授,主要研究方向为宇宙辐射、毫米波、天体物理;樊永文(1993—),男,硕士生,主要研究方向为人工智能;陈永华(1962—),男,博士,副教授,主要研究方向为 DNA 计算,E-mail:ieyhchen@zzu.edu.cn(通信作者)。

中，用来分类的方法有很多，如支持向量机、逻辑回归、随机森林、决策树、提升树(Boosted Trees，BT)和深度学习等。其中BT是一类常用的机器学习算法，它具有效果好、对输入要求不敏感、计算复杂度不高等优点，因此被广泛应用于文本分割[6]、人脸识别[7]、手势识别[8]、多视角目标侦测[9]、情绪识别[10]等领域。本文中使用的GBRT即为一种BT算法。

GBRT是预测分析最有效的机器学习算法之一。该算法生成多棵决策树，所有树的结果累加起来形成最终答案。算法的核心在于，每棵树是从之前所有树的残差中来学习的。它在被提出之初就和SVM一起被认为是泛化能力较强的算法。作为一种回归树，GBRT可用于处理回归问题，也可用于处理二分类问题(设定阈值，大于阈值为正例，反之为负例)。GBRT的优势在于[11]：1)自然而然地处理混合类型的数据；2)预测能力强；3)对于异常值的鲁棒性强。GBRT的劣势在于[11]：由于提升的顺序性，不能进行并行处理。

GraphLab是一款开源的Python机器学习包[12]，由美国卡内基梅隆大学开发。该工具集成了包括GBRT在内的多种机器学习算法，极大地简化了模型的训练学习过程，便于用户操作和实现具体的机器学习算法。

2.2 系外行星的宜居性

搜索系外行星生命被认为[1]：寻找与地球条件类似的行星(地球相似度)、生命在该星以我们已知或未知形式存在的可能性(宜居性)。两个常用指标是：地球相似指数(Earth Similarity Index，ESI)和行星宜居性指数(Planetary Habitability Index，PHI)。前者以地球作为宜居性参考，考查行星与地球的物理相似度；后者基于生命对环境的一些基本要求，评估行星的潜在宜居性[1]。

基于不同的指标，不同研究机构给出的潜在宜居行星列表有所区别。

美国波多黎各大学行星宜居性实验室认为潜在宜居的系外行星有53个[13]。他们把这些行星分为两类[13]：保守估计下的潜在宜居系外行星和乐观估计下的潜在宜居系外行星。前者有13个，行星入选条件是[13]：行星位于保守宜居带，且它的半径小于或等于1.5倍地球半径，同时大于0.5倍地球半径，并且它的质量小于或等于5倍地球质量，同时大于0.1倍地球质量。后者有40个，行星入选条件是[13]：行星位于乐观宜居带，且它的半径小于或等于2.5倍地球半径同时大于1.5倍地球半径，且它的质量小于或等于10倍地球质量而大于5倍地球质量。这样的条件是为了确保入选行星具有固态岩石成分，以及表面拥有液态水[13]。

日本京都大学公布的系外行星数据库[14]有3种宜居性定义：SEAU宜居(根据太阳等价天文单位计算宜居带)、Kopparapu宜居，以及NASA宜居(根据美国国家航空航天局的定义计算)。满足后两种宜居的系外行星数量很少，约30余个，满足SEAU宜居的系外行星则达186个[14]。

事实上，SEAU宜居的入选条件较为宽松，而NASA宜居的要求相对严格，这是造成二者数量差异较大的主要原因。

3 新方法

如第1节所述，我们把系外行星宜居性问题规约为机器学习二分类问题。所提方法的核心原理如图1所示。

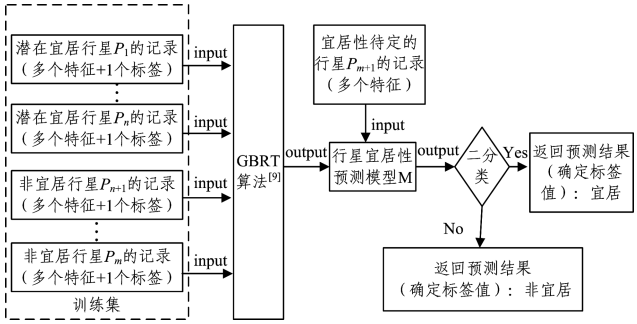


图1 基于GBRT的系外行星宜居性预测方法

新方法的步骤如下。

首先，构建包含n个潜在宜居行星(P_1，…，P_n)和$m-n$个非宜居行星(P_{n+1}，…，P_m)的多项特征值与单项标签值的m条记录作为训练集。

然后，使用GBRT算法对训练集实施训练，获得可预测行星宜居性的机器学习模型M。

最后，把宜居性待定的给定行星P_{m+1}的多项特征值输入M，M的输出即指明了P_{m+1}的宜居性结果(是否宜居)。

4 仿真实验

4.1 实验目的

检验新方法的预测准确率。

4.2 实验平台

(1)CPU：Intel(R) Core(TM) i7-4790 CPU @3.60GHz。

(2)内存：8.0GB RAM。

(3)操作系统：Windows 10。

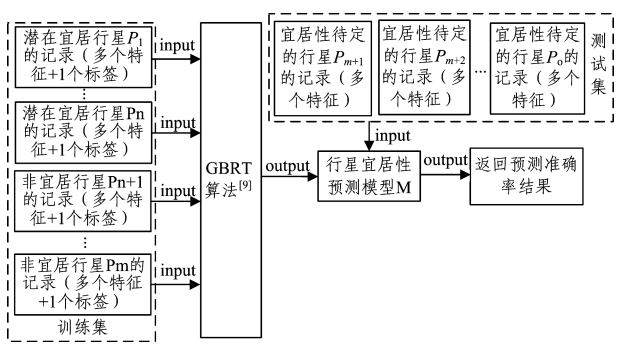
(4)GraphLab：用于本研究实现GBRT算法的机器学习实验。

(5)数据集：日本京都大学系外行星数据库[14]。

本文中为了确保数据分析的效果，使用宽松标准[14]作为系外行星宜居的定义，以避免正例过少带来的数据不均衡问题。使用系外行星数据库的全部正例样本(186条记录)和部分负例样本(214条记录)构成本实验的样本集(400条记录)。在数据集中，使用了绝对磁场强度、表面磁场强度、质量、轨道距离、轨道周期、半径、行星尺寸类别(大小类似太阳系中的哪个行星)、恒星辐射、恒星质量、恒星半径、温度等11个物理、天文性质作为数据特征。

4.3 实验步骤

实验步骤与流程如图2所示。



注：其中GBRT算法实现在Graph Lab上实施

图2 实验流程与步骤

(1)在Graph Lab上把400条记录分为训练集与测试集。

(2)输入训练集，调参训练获得机器学习模型M。

(3)使用测试集测试 M 的预测准确率。

(4)若步骤(3)获得的准确率低于给定阈值，则继续调参获得更优模型；否则，M 即为获得的可对行星宜居性预测的目标模型，返回 M 的预测准确率。

4.4　实验结果与分析讨论

当参数设定为表 1 所列的取值时，可获得最优实验结果，如图 3 所示。图中可见：38 个正例样本均被正确预测，34 个负例样本也均被正确预测，假阴率和假阳率均为 0%；所有 72 个测试样本均被正确预测，预测准确率高达 100%；单条记录的平均预测时间只有约 0.002 s。

表 1　获得图 3 所示实验结果的参数与取值

参数名	参数含义	参数取值
fraction	训练集与测试集的记录数量之比	0.8
seed	其他需要用户调整参数的取值的组合编号	3

```
In [105]: result = model.evaluate(test)
          result['accuracy']
Out[105]: 1.0

In [106]: print ( 'The total running time for %d records is %.18f second ' %( len(test),totaltime ) )
          The total running time for 72 records is 0.165552884505757447 second

In [107]: print ('The average running time for %d record is %.18f second ' %(1,averagetime) )
          The average running time for 1 record is 0.002299345618135520 second

In [108]: result
Out[108]: {'accuracy': 1.0, 'auc': 1.0, 'confusion_matrix': Columns:
                  target_label    int
                  predicted_label int
                  count    int

           Rows: 2

           Data:
           +--------------+-----------------+-------+
           | target_label | predicted_label | count |
           +--------------+-----------------+-------+
           |      1       |        1        |   38  |
           |      0       |        0        |   34  |
           +--------------+-----------------+-------+
```

图 3　关于预测系外行星宜居性的最优实验结果

以上结果提示：新方法在预测能力、预测效果与预测效率方面表现优异。本实验的预测效率高是因为预测行星宜居性被规约为一个强可学习问题，该近似计算问题可在多项式时间内完成，兼具高预测准确率。此外，本实验的预测准确率高是因为：1)GBRT 算法适用于小样本数据集，在小样本上的预测能力强；2)采样合理，正例与负例的数量比值接近；3)选取比较宽松的 SEAU 宜居性标准，保证了正例样本总数均具有一定的数量，使得合理采样成为可能，进一步避免了数据不均衡问题；4)确保一定数量的特征值被选取，尽可能充分覆盖对宜居性有影响的物理、天文因素；5)如上所述，预测行星宜居性被规约为一个强可学习问题。

5　相关工作比较

目前已有一些工作直接使用机器学习技术对系外行星宜居性进行预测。

Hora 在研究[3]中使用分类回归树(Classification And Regression Trees，CART)、支持向量机(Support Vector Machines，SVM)、随机森林(Random Forest，RF)、逻辑回归(Logistic Regression，LR)、前向神经网络(Forward Neural Networks，FNN)和朴素贝叶斯分类(Naïve Bayes，NB)等 6 种机器学习方法分别对训练集进行训练，然后给出预测有关行星是否宜居的二分类结果[3]。其结果表明[3]：CART 算法表现最优，其预测准确率高达 99.8%；表现最差的 NB 算法的预测准确率也可达 95.5%。然而，由于这项研究采用严格的宜居性标准[13]，导致数据集中的正例数量只有 35 个，测试集中的正例数量更是低至 9 个。测试集中正例数量如此之少，导致其关于假阴率与真阳率的实验结果的价值与意义均被大打折扣。与之对照，本文的实验中，测试集正例数量被提升至 38 个，超过文献[3]中测试集正例数量的 4 倍，而且还有进一步大幅提升的空间。

Saha 等人指出，在他们基于 XGBoost 算法的实验中，分类准确率高达 100%[2]，这固然再一次证实了机器学习在行星宜居性分类中表现优异，然而，文献[2]也指出模型不确定、测量不确定仍是其所述方法的不足[2]。这就导致该方法对具体每一个行星做宜居性分类尚且不能确定结果，所谓“百分之百分类准确率”仅特指用于对 Proxima b、TRAPPIST-1 d、TRAPPIST-1 e 这 3 个行星的分类，对行星 TRAPPIST-1 g 的分类准确率只有 82.3%，对行星 Proxima Cen b 的分类准确率更低至 73%[2]。与之对照，本文实验中，对所有 72 个系外行星的预测(分类)结果均稳定地达到了 100%的准确率。

此外，温度、轨道周期、轨道半径等因素都对行星宜居性产生影响，这些是文献[1-2]中方法“在未来的研究中拟考虑的”[2]。作为对照，新方法已经把上述 3 种物理、天文条件作为样本集的特征数据，使其参与训练与预测。来自母星(行星所围绕公转的恒星)的电磁辐射对行星的宜居性也有重要影响，已有的文献[1-2]并未考虑这些因素，而本研究则把辐射值作为机器学习的一个特征值进行训练与预测。

结束语　本工作提出一种使用梯度提升回归树的系外行星宜居性预测方法，从已有的文献来看，这是第一种可全面地在计算层面达到精准预测系外行星宜居性的核心方法(同时兼顾天体物理学层面的若干必要约束条件)。这是本项研究的贡献。

宜居性的定义松紧事关机器学习算法的预测效果，同时也事关预测结果的天文学价值高低。由于正例数据的稀疏，因此这是一个鱼与熊掌式的问题。若宜居性定义过于严格，则对给定行星的阳性预测结果相对接近于该星真实宜居，但数据量稀少难以训练；若宜居性定义过于宽松，则数据量较大，方便训练与预测，但对给定行星的阳性预测结果，即使逻辑正确、计算过程正确，也未必表明物理宜居。进一步地说，由于天体物理学与天体生物学的极端复杂性，液态水存在只是真实宜居的一个公认的必要非充分条件[4]，而潜在宜居的

(下转第 79 页)