文章编号: 1009-6094(2020)04-1236-05

基于 RF 的森林火灾风险评价模型及其应用研究*

贾 南[1],陈 悦[2],康可霖[2],李俊锋[2]

(1 中国人民警察大学基础部,河北廊坊 065000;
2 中国人民警察大学研究生部,河北廊坊 065000)

摘 要: 为了预测森林火灾发生的可能性,为森林火灾的预防预报提供依据,减少火灾损失,在 Python 平台上应用随机森林算法,以西班牙东北 Tr´as-os-Montes 地区 Montesinho 森林公园 2000 年 1 月至 2003 年 12 月的数据记录,对影响森林火灾的指标变量进行分析和评价。结果表明,随机森林算法对森林火灾预测的准确度约为 80%,表明随机森林算法对森林火灾具有较好的预测能力,可用于对森林火灾的预测预报。

关键词: 安全工程; 森林火灾; 风险评价; 随机森林算法
中图分类号: X43 **文献标志码**: A
DOI: 10.13637/j.issn.1009-6094.2019.0899

0 引 言

森林被誉为地球之肺,在保持水土、调节气候等方面发挥着巨大的作用。调查显示,仅我国的森林面积就达到 2.08 亿 hm^2,巨大的森林面积加上日趋频繁的极端天气使森林火灾的数量也随之增加。据统计,世界森林火灾的年发生次数约为 22 万次,过火面积达 1 000 万 hm^2,仅欧洲就达到 8.1 万次,葡萄牙和西班牙的年发生次数分别为 1.8 万次和 2.3 万次,过火面积为 12.3 万 hm^2 和 14 万 hm^2[1]。在我国,2002—2010 年,森林火灾共发生 96 704 次,受灾面积为 132.34 万 hm^2[2]。森林火灾造成了巨大的经济财产损失和环境损失,威胁动植物乃至人类本身的安全,因此森林火灾的预防问题极为严峻,而对森林火灾的预测更是防止森林火灾发生的一项基础性工作。

森林火灾的危险性与多种指标变量有关,这些指标变量与森林火灾的关系复杂且非线性,使用传统方法对其进行分析不足以揭示复杂过程中的格局和关系,无法得到准确的预测结果[3]。2001 年,Breiman[4] 建立了随机森林(RF)算法模型,该算法既保持了统计学理论的优点,又减少了极端数据的误差,将一系列决策树的预测结果进行组合而得到最优结果。RF 算法的基本思想是利用计算机抽取数据组成分类器,再将多个相对较弱的分类器组合起来成为一个强分类器,弱分类器相互补充从而提高结果的准确性,对于森林火灾这样具有多个指标变量的预测具备很好的效果。

1 随机森林算法

1.1 决策树的生成

基于森林火灾中多指标变量进行预测的 RF 算法的核心是根据各指标变量和数据进行判断,得出相应的风险等级,判断的过程在决策树中进行。此模型的决策树是较简单的二叉型决策树,整棵树由根节点、非叶子节点和叶子节点组成,从根节点到叶子节点代表一组森林火灾数据评价完成,叶子节点即为对应的危险等级。数据集 D_t 的训练过程对应一棵决策树的生长过程,即将 D_t 输入根节点 S_1 中,然后按照相应的标准不断向下进行分割,最终落入相应的风险等级中。图 1 为单棵决策树的生长过程。

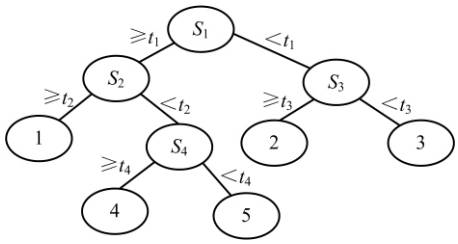

图 1 单棵决策树
Fig. 1 Single decision tree

1) 使用 Bootstrapping 的方法获取训练集。从 N 个样本数据中有放回地随机抽取 n 个数据组成一个训练集,输入根节点 S_1 中。

2) 随机选取各节点的指标变量。设最终的 RF 模型中共有 M 个指标变量,随机从这 M 个指标变量中不放回地选取 m 个作为节点指标,本文按照 Breiman 推荐的取 M 的平方根作为 m[5]。

3) 节点的选择。对于每个节点中指标变量的选择,首先计算节点处的信息熵,假设共有 C 个风险等级,节点 t 中第 c 个等级的频率为 $P(c|t)$,则节点 t 的信息熵为

$$E(t) = -\sum_{c=1}^{C} P(c|t)\ \log_2 P(c|t) \qquad (1)$$

* 收稿日期: 2019-07-19
作者简介: 贾南,讲师,硕士,从事数据挖掘和大数据分析研究,601265674@qq.com。
基金项目: 河北省科技计划项目(16215416);河北省高等学校科学技术研究项目(Z2018020)

根据节点信息熵的定义计算信息熵的下降值,即信息增益。

$$G = E(t_0) - \sum_{k=1}^{K} \frac{nk}{n} E(t_k) \quad (2)$$

式中　K 为决策树中的节点总数,n 为节点 t_0 中的样本总数,n_k 为节点 t_k 中的样本数。G 越大,表示样本的不纯度越小,则选择该分裂方式越优越。

4) 决策树自由生长,不对其进行剪枝处理。

1.2　随机森林算法原理

RF 是由一系列决策树共同组成的集成分类器,森林中的每棵树都会根据自身测试数据的特征,对测试数据进行独立的预测,最后采用投票方法(少数服从多数) 做出最终的预测。从原始数据总集 D 中选取 k 个子训练样本集$\{D_1, D_2, \cdots, D_t, \cdots, D_k\}$,并据此形成 k 棵决策树,最后集成随机森林。图 2 为 RF 的生成步骤。

基于 D_t 可以形成一个基础的决策树模型$h_t(x)$,将测试集输入调试完毕的 RF 各决策树中,最后各叶子节点对应的危险等级即为该决策树的评估结果。通过 Bootstrapping 方法进行有放回地随机抽样获得多份不同的样品数据,使用不同的样本数据对决策树进行训练,可以很好地降低样本的相关性,再加上随机选取节点的指标变量,进一步将相关性降低,基本解决了单棵决策树模型的过拟合问题,使得 RF 模型对噪声具有良好的容忍度,而且任意生长不进行剪枝的决策树能够提高对新数据分类的正确性,最终提高了决策的准确性。

1.3　模型泛化误差与指标重要性评价

随机森林算法是 Bagging 算法的典型代表,总样本数为 N 的原始数据总集 D 中每个样本数据不会被抽到的概率为$(1-1/N)^N$,当 N 趋近于无穷大时,$(1-1/N)^N \to 1/e = 0.368$,即原始数据总集 D 中约有 36.8% 的数据不会出现在抽取出的训练子集 D_t 中[6],这部分数据被称为袋外样本(OOB) 。在使用子训练集中的数据训练完成决策树 t 后,就可以利用 OOB_t 数据进行内部误差估计,计算出该决策树的预测错误率,即袋外误差(E_t),而森林中所有决策树袋外误差的平均值即为本模型的泛化误差。RF 算法使用 OOB 数据计算决策树的预测错误率和各指标变量的重要性,使得袋外误差具有无偏性,比设置测试样本进行交叉检验更加方便有效。

RF 算法的指标重要性评价主要有两种: 一种是计算整个随机森林每个指标变量平均的信息增益,信息增益越大,指标重要性越高; 另一种是人为地向指标变量数据中加入噪声,再计算袋外误差,决策树的袋外误差变化越大,表明该指标重要性越高。本文采用后者来评价随机森林的指标重要性。

采用 RF 算法对森林火灾风险的指标重要性进行评价时,首先计算每颗决策树 t 的袋外误差 E_t,然后在其他指标变量不变的情况下,随机改变袋外样本中某一特定指标变量 X_i 的顺序,重新计算改变顺序后的袋外误差 E',通过比较袋外误差的变化来评价指标重要性,指标变量 X_i 的重要性得分为

$$\mathrm{VI}(X_i) = \frac{1}{n_{\mathrm{tree}}} \sum_t (E' - E_t) \quad (3)$$

式中　n_{tree} 为森林中决策树的数量。$\mathrm{VI}(X_i)$ 越大,指标重要性越高。

2　指标变量的选取

2.1　指标变量的选取原则

目前还未建立统一的森林火灾指标变量体系,我国对森林火灾的研究更是处于起步状态。本文在参考美国指标的基础上,结合实际情况确定森林火灾指标变量体系。在建立指标变量体系时应遵循如下一些原则[7]。

1) 可行性原则: 即能够在实际工作中进行量化收集,每项指标的数据都是可查询的,如果所选用的指标没有统计部门进行收集,会给评估工作带来极

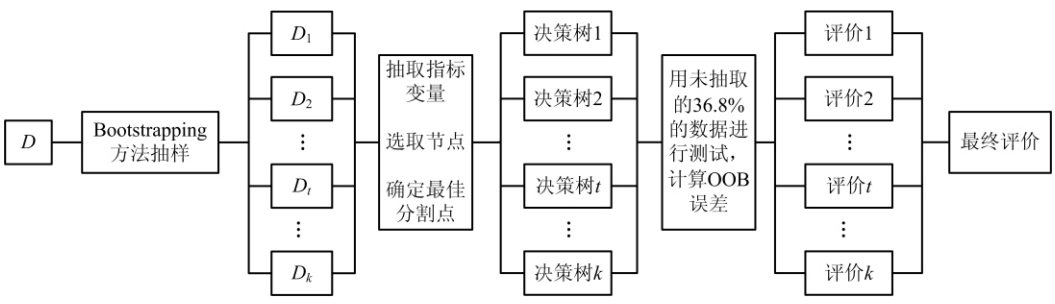

图 2　RF 生成步骤

Fig. 2　RF generation steps

大的困难。

2) 代表性原则: 在森林火灾中,应选取对森林火灾的严重程度有重大影响的突出因素,这些指标的改变能够较大地影响森林火灾发生的几率。

3) 全面性原则: 森林火灾风险评价的指标变量必须能够较为全面地覆盖引起火灾的各方面属性,指标准确、科学、全面,包含森林火灾的全部内容。

4) 独立性原则: 要保证选取的各指标变量之间的相对独立性,相关性不能太大,避免指标体系过于冗余、复杂,降低评价的准确性,使指标体系整体优化。

5) 可比性原则: 森林火灾评价体系中的指标变量应该能够以一定的方式进行量化,按照灾情属性的内在关系进行组合,使森林火灾的量值能够进行横向或纵向的比较。

2.2 指标变量的选取

根据以上原则,剔除了一些数据缺少严重的指标(气压、辐射等),初步筛选了 12 个因子作为模型的指标变量: 空间位置(X、Y)、月份(M)、星期(W)、细小可燃物湿度码(FFMC)、腐殖质湿度码(DMC)、干旱码(DC)、初始蔓延速度(ISI)、温度(T)、相对湿度(RH)、风速(S)、降雨量(R)和地区(A)。

FFMC、DMC、DC 的计算模型来源于加拿大的森林火灾天气因素系统[8-9]。FFMC 体现森林中干重为 0.25 kg/m^2、厚度为 1.2 m 的枯枝落叶等较为细小的可燃物的含水率,FFMC 越大,则可燃物的湿度越大,越不容易被引燃,其与降水量、日照强度、相对湿度、风速等天气因素的影响有关,范围为 0 ~ 101。DMC 体现浅层半分解腐烂的可燃物的湿润程度,其与温度、降水量和相对湿度有关,最小值为 0,无上限,但一般不会超过 150。DC 表示较长时期的干旱对森林中可燃物的影响力,主要观测深层的落叶层和大型段木,最小值为 0,无上限,但一般不会超过 1 000。ISI 主要与风速有关,一般来说,风速每增加 13 km/h,ISI 会变为原来的两倍。

3 实例应用

3.1 数据来源

实际案例为西班牙东北部 Tr´as-os-Montes 地区的 Montesinho 森林公园[10],数据来源于两个方面: 一是负责该公园森林火灾事件的检查员在森林火灾发生时收集的时间、位置、植被类型、烧毁面积等信息; 二是由布拉甘茨理工学院收集的位于 Montesinho 森林公园中心的一个气象站的天气信息,这些信息经两位专家整理后,可以在 http://www.dsi.uminho.pt/~pcortez/forestfires/中查询。数据集属性见表 1,空间位置属性见图 3。

表 1 预处理的数据集属性

Table 1 Properties of the preprocessed data sets

指标变量	基本属性描述
X	X 轴坐标(1 ~ 9)
Y	Y 轴坐标(1 ~ 9)
M	一年中的月份(1 ~ 12)
W	星期(1 ~ 7)
FFMC	FFMC 指数(18.70 ~ 96.20)
DMC	DMC 指数(1.1 ~ 291.3)
DC	DC 指数(7.9 ~ 860.6)
ISI	ISI 指数(0 ~ 56.10)
T	地面环境温度(2.20 ~ 33.30 ℃)
HR	相对湿度(15.0% ~ 100%)
V	平均风速(0.40 ~ 9.40 km/h)
R	日平均降雨量(0 ~ 6.4 mm/m^2)
S	烧毁面积(0 ~ 1 090.84 hm^2)

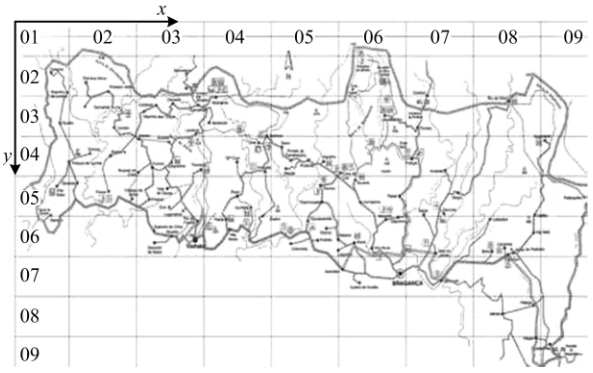

图 3 Montesinho 森林公园地图

Fig. 3 Map of Montesinho forest park

3.2 模型实现与参数设定

在开源软件 Python 平台上对模型进行实现,从 Montesinho 自然公园 2000 年 1 月至 2003 年 12 月的 517 条数据记录中抽取 500 个作为样本数据集 D,随机森林算法的实现过程如下。

1) 数据集分组。使用五折交叉验证法: 将样本数据集 D 随机等分成 5 个子集 D_1、D_2、D_3、D_4、D_5,其中 D_i 为测试集,剩下的 4 个子集为训练集,从而构成第 i 个训练测试组。

2) 参数设定。在创建 RF 模型时,需要设定相应的参数,经过多次调试得到最佳的模型参数,本模型中 RF 算法的决策树设置为 100 棵,节点分叉变量数量为 2。

3) 准确度评价。用分好的 5 个子集进行训练和测试,当准确度符合要求时直接进行下一步的计算,当准确度不符合要求时需要调整参数以增加准确度或重新取样。

4) 指标变量重要性计算。将整个样本数据集 D 全部输入 RF 模型中,计算指标变量的重要性。

5) 森林火灾危险性评价。将数据 area 项(着火面积)中"0"评为无危险,其他为有危险,根据模型分析结果,判断该地区的森林火灾危险性。

3.3　结果分析

3.3.1　评价准确度分析

对 RF 模型的训练和测试进行评价,由表 2 可知,RF 模型的准确度约为 80%,准确度较高,能够为森林火灾的预防预测提供有效的参考。经过分析,模型出现误差的主要原因有两个: 一是信息量不足,影响森林火灾的因素复杂多样,而由于成本和技术等原因,能够收集到的影响因素数据较少; 二是分析的样本数据较少,样本数据不足会导致随机森林算法形成的决策树准确性不足。

表 2　五折交叉验证法评价结果

Table 2　Evaluation results of five-fold cross validation

子集	样本数		RF 评价			
	训练	测试	错误数量		准确度/%	
			训练	测试	训练	测试
D_1	400	100	81	21	79.75	79.00
D_2	400	100	78	19	80.50	81.00
D_3	400	100	80	22	80.00	78.00
D_4	400	100	80	18	80.00	82.00
D_5	400	100	79	19	80.25	81.00
平均准确度					80.10	80.20

3.3.2　指标重要性分析

根据 RF 模型的模拟结果来看,月份、地面环境温度、相对湿度、DMC 为森林火灾的主要驱动因子,其中相对湿度对森林火灾的发生概率影响最大,且相对湿度与 Tr´as-os-Montes 地区的 Montesinho 森林公园中森林火灾的发生概率反向相关。

3.3.3　结果对比

将本文得出的结果与国内专家的结果进行比较: 郭福涛等[11]对大兴安岭地区森林火灾与天气变量之间的关系进行了研究,本文研究结果与其结果基本一致; 郑琼等[12]对伊春地区森林火灾与时空变量之间的关系进行了研究,结论呈现明显的东多西少的特征,但本文预测结果中与空间分布关系不大,可能是由于该公园面积较小、植被种类较为单一,因此空间异质性不大; 秦凯伦等[13]对森林火灾的评价模型选择进行了研究,但由于影响森林火灾的指标变量无显著的线性规律,准确性和便利性相对于随机森林算法较差。将随机森林算法应用于我国森林火灾的预测,可以极大地提高森林火灾预测的准确性和全面性。

4　结　论

本文阐述了 RF 模型算法的原理并将该算法在 Python 平台上进行实现,取得了较好的预测结果。根据以往数据,预测的成功率在 80% 左右,其中月份、地面环境温度、相对湿度、DMC 为森林火灾的主要驱动因子,而相对湿度对森林火灾的发生概率影响最大,且相对湿度与 Tr´as-os-Montes 地区的 Montesinho 森林公园中森林火灾的发生概率反向相关。分析结果表明 RF 模型算法与森林火灾预测有很好的兼容性,可以用于预测森林火灾。

References(参考文献):

[1] CHAS-AMIL M L, PRESTEMON J P, MCCLEAN C J, et al. Human-ignited wildfire patterns and responses to policy shifts [J]. *Applied Geography*, 2015, 56: 164 - 176.

[2] BAI Haifeng(白海峰). *Forest fires prevention and control capabilities in Sanming Region, Fujian Province*(福建三明地区森林火灾综合防控能力研究) [D]. Beijing: Beijing Forestry University, 2012.

[3] PAN Deng(潘登), YU Peiyi(郁培义), WU Qiang(吴强). Application of random forest algorithm on the forest fire prediction based on meteorological factors in the Hilly Area, Central Hunan Province [J]. *Journal of Northwest Forestry University*(西北林学院学报), 2018, 33(3): 169 - 177.

[4] BREIMAN L. Random forests [J]. *Machine Learning*, 2001, 45(1): 5 - 32.

[5] LAI Chengguang(赖成光), CHEN Xiaohong(陈晓宏), ZHAO Shiwei(赵仕威), et al. A flood risk assessment model based on Random Forest and its application [J]. *Journal of Hydraulic Engineering*(水利学报), 2015, 46(1): 58 - 66.

[6] BREIMAN L. Bagging predictors [J]. *Machine Learning*,

1996, 24(2): 123 – 140.

[7] CHANDLER C. Forest fire behavior and effects [M] // *Fire in forest*. New York: Whey and Sons, 1999: 24 – 28.

[8] WOTTON B M. Interpreting and using outputs from the Canadian forest fire danger rating system in research applications [J]. *Environmental & Ecological Statistics*, 2009, 16(2): 107 – 131.

[9] AGEE J K, SKINNER C N. Basic principles of forest fuel reduction treatments [J]. *Forest Ecology and Management*, 2005, 211(1/2): 83 – 96.

[10] CORTEZ P, MORAIS A. A data mining approach to predict forest fires using meteorological data [C] // *Proceedings of the 13th Portugese Conference on Artificial Intelligence*. Guimaraes, Portugal: Portuguese Association for Artificial Intelligence, 2007.

[11] GUO Futao(郭福涛), SU Zhangwen(苏漳文), MA Xiangqing(马祥庆), et al. Climatic and non-climatic factors driving lightning-induced fire in Tahe, Daxing'an Mountain [J]. *Acta Ecologica Sinica*(生态学报), 2015, 35(19): 6439 – 6448.

[12] ZHENG Qiong(郑琼), DI Xueying(邸雪颖), JIN Sen(金森). Temporal and spatial patterns of forest fires in Yichun Area during 1980 – 2010 and the influence of meteorological factors [J]. *Scientia Silvae Sinicae*(林业科学), 2013, 49(4): 157 – 163.

[13] QIN Kailun(秦凯伦), GUO Futao(郭福涛), DI Xueying(邸雪颖), et al. Selection of advantage prediction model for forest fire occurrence in Tahe, Daxing'an Mountain [J]. *Chinese Journal of Applied Ecology*(应用生态学报), 2014, 25(3): 731 – 737.

Improved forest fire risk assessment model and its application based on the RF algorithm

JIA Nan[1], CHEN Yue[2], KANG Ke-lin[2], LI Jun-feng[2]

(1 Department of Foundation, China People's Police University, Langfang 065000, Hebei, China; 2 Department of the Graduate, China People's Police University, Langfang 065000, Hebei, China)

Abstract: In order to predict the possibility of the forest fire and provide a basis for the forest fire prevention and reduce the fire loss, the present paper intends to apply a random forest algorithm on the Python platform based on the data recorded from January, 2000, to December, 2003, in the Montesinho forest park in the Trás-os-Montes region of the north-east area of Spain. The indicator variables affecting the forest fires have been analyzed and evaluated by using a series of decision-making trees through an integrated classifier of the random forest, in which each tree in the forest can help to predict the testing data independently according to its own feature. At the same time, the Bootstrapping method has been used to register the randomly chosen different samples of data with a place-back sequence. The correlation of the samples can be reduced and properly determined by using the different sampling data to train a decision-making tree plus the random selection of the index variables of nodes. By artificially adding noise to the data of the index variable and calculating the out-of-bag error, a state can be gained that, the bigger the change of the out-of-bag error, the greater the importance of the index would be. And, finally, the eventual prediction can be made by the majority vote and deciding method, i. e. the minority has to be subordinate to the majority. The results of decision show that the accuracy of the forest fire prediction by random forest algorithm can be so accurate as to about 80%, suggesting that the random forest algorithm has better prediction power for the forest fire occurrence and control. Meanwhile, the per-month ground environment temperature, the relative humidity and DMC can also be the chief driving factors of the forest fire, of which relative humidity may have the greatest impact on the probability of the forest fire. For instance, when the relative humidity is below 30%, the probability of forest fire should be the greatest. Therefore, the random forest algorithm can be used to predict the forest fire and provide a certain reference role for the forest fire forecast and prevention.

Key words: safety engineering; forest fires; risk assessment; RF algorithm

CLC number: X43 **Document code:** A

Article ID: 1009-6094(2020)04-1236-05