机器学习算法在森林火灾风险预测中的有效性比较研究

张朔，潘梦雅*

南京大学信息管理学院，江苏南京 210023

摘要： 森林火灾作为一项全球性的环境问题，对生态系统、经济发展和社会安全构成严重威胁。准确预测森林火灾对于制定有效的预防措施和减少火灾带来的损失至关重要。随着机器学习技术的发展，其在森林火灾风险预测领域中的应用逐渐成为研究热点。鉴于不同的机器学习算法具有不同的数据处理能力和预测精度，比较和选择最适宜的算法对提高预测模型的性能显得尤为关键。本研究通过分析 1990 至 2019 年广西省的气象数据和相应的火灾等级信息，选取了 10 种机器学习算法进行实验。在数据预处理阶段，严格处理了缺失值和异常值，以确保数据质量可靠。随后，利用准确率、精确率、召回率、F1 分数以及 ROC 曲线等多项指标，全面评估各算法的预测性能。为进一步检验模型的稳定性和鲁棒性，并确保所选模型具有更强的泛化能力，研究实施了 5 折交叉验证。结果显示，SVM、贝叶斯、BP 神经网络、逻辑回归、AdaBoost、Gradient Boost 和 XGBoost 在 AUC 上表现较佳，而 KNN 和随机森林在精确度和准确率上具有优势。研究指出，结合多种算法能够进一步提高预测的准确性和可靠性，未来研究可考虑纳入更多影响因素，并探索采用深度学习技术来进一步提升预测性能，为森林火灾的预防和应对提供更加科学、有效的支持。

关键词： 森林火灾；机器学习；数据预处理；预测模型

DOI: 10.57237/j.jaf.2024.02.002

The Effectiveness of Methods Designed to Predict Forest Fire Risk: A Comparison of Machine Learning Algorithms

Zhang Shuo, Pan Mengya*

School of Information Management, Nanjing University, Nanjing 210023, China

Abstract: Forest fires represent a global environmental issue, posing severe threats to ecosystems, economic development, and social security. Accurate prediction of forest fires is crucial for formulating effective preventive measures and minimizing the associated losses. With the advancement of machine learning technologies, their application in forest fire risk prediction has become an emerging research focus. Diverse machine learning algorithms exhibit varying data processing capabilities and predictive accuracies; hence, comparing and selecting the most suitable algorithms is significant for enhancing the performance of predictive models. This study analyzed meteorological data and corresponding fire severity information from Guangxi Province between 1990 and 2019, employing 10 machine

*通信作者：潘梦雅，pmy@smail.nju.edu.cn

收稿日期: 2024-04-01；接受日期: 2024-05-06；在线出版日期: 2024-05-23

learning algorithms in experiments. Initial data preprocessing, including handling of missing and outlier values, ensured data quality. Subsequently, predictive performance across algorithms was assessed using accuracy, precision, recall, F1 score, and the Receiver Operating Characteristic (ROC) curve. To further examine the stability and robustness of the models, a 5-fold cross-validation was implemented. Results indicated that SVM, Bayesian classifiers, BP neural networks, logistic regression, AdaBoost, Gradient Boost, and XGBoost demonstrated superior performance in terms of AUC, while KNN and Random Forest algorithms showed advantages in precision and accuracy. The 5-fold cross-validation confirmed the stability and robustness of the models, revealing that most models maintained stable predictive performance across different datasets. The study suggests that integrating multiple algorithms can improve the accuracy and reliability of predictions and recommends that future research consider additional influencing factors and employ deep learning techniques to further enhance predictive performance.

Keywords: Forest Fires; Machine Learning; Data Preprocessing; Predictive Models

1 引言

森林火灾是指在森林或草原等林地内的燃烧事件。其产生原因包括天气干旱、雷击、人为疏忽等多种因素[1]。森林火灾一旦发生，将对生态环境、经济和社会发展带来灾难性后果，包括林木破坏、土壤质量下降、生物多样性丧失等[2]。这些问题将严重影响森林生态系统的恢复和发展。同时，森林火灾还会对当地经济造成严重损失，如破坏农田、森林资源等。对人类社会而言，森林火灾也会给社会发展带来负面影响，如生命财产损失、环境污染等。

根据森林火灾的统计数据显示，自 1952 年至 2019 年，中国发生了越 77.7 万起森林火灾[3]，如图 1 所示。平均每年发生约 1.2 万起，导致近 346 万公顷森林遭受破坏，经济损失高达约 2 亿美元。中国的森林面积约为 1.75 亿公顷，占据全球森林总面积的 3.9%，位列世界第五，而森林火灾的发生率占全球的 10%。这些数据表明，森林火灾对全球经济构成重大威胁，造成巨大的破坏和损失。

气象因素是导致森林火灾发生的重要因素之一，其主要包括温度、湿度、风速和降雨等[4]。因此，利用气象数据对森林火灾进行预测和监测，能够有效地减少火灾的发生[5]。在过去，人们通常使用统计模型来预测森林火灾危险等级，但是由于统计模型对数据的要求较高，容易受到噪声和数据缺失的影响，导致其预测精度较低[6]。而随着机器学习技术的发展，可以利用机器学习算法对气象数据进行建模，构建出更为准确和可靠的森林火灾预测模型[7]。

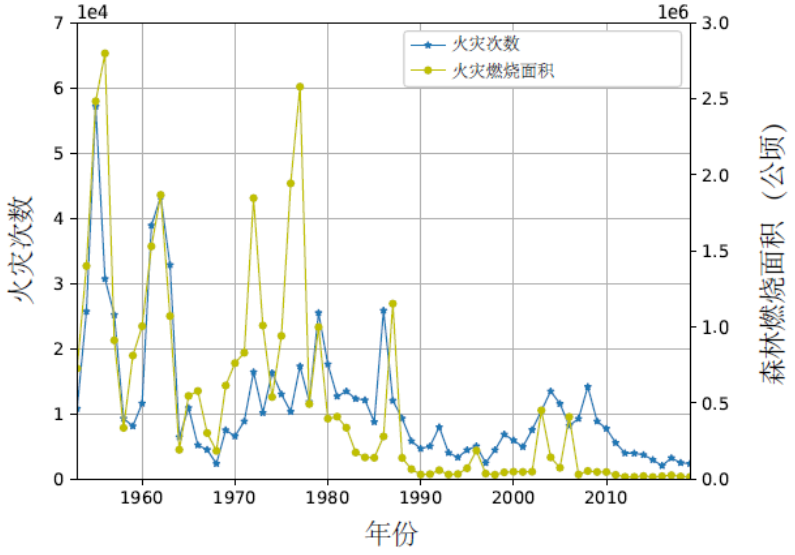

图 1 1952-2019 年中国森林火灾次数和燃烧面积

当前，已经有很多关于森林火灾预测的研究，但是这些研究主要是对单一算法或者方法进行探索和应用，缺乏对不同算法的综合比较和评估。因此，有必要对目前广泛应用的机器学习算法在森林火灾预测中的表现进行比较研究，以便为实际应用提供更为准确和可靠的预测模型。本次研究基于气象数据构建森林火灾危险等级预测模型，并对比不同的机器学习算法，以确定最优的预测模型。

依据上述问题，本研究首先通过对比实验，从多个角度评估不同机器学习算法在森林火灾预测方面的性能，以便更全面地了解不同算法的优劣，为相关领域的研究提供参考价值。其次，本研究在实验中使用了真实的气象数据，提高了模型的实际应用价值。此外，通过分析不同算法的预测结果，可以更好地理解气象因素对森林火灾的影响，并为相关政策制定提供参考依据。最后，本研究还对模型性能的稳定性进行了探究。这些研究成果对于提高森林火灾预警和防范能力具有重要意义。

2 相关研究

森林火灾是一种具有毁灭性的自然灾害，仅造成生态环境的破坏和生命财产的损失，还会对气候环境和社会稳定产生重大影响。因此，研究森林火灾的预测方法具有重要的现实意义。目前，森林火灾预测方法主要分为经验模型和基于物理模型的方法两种[8]。

经验模型是利用历史数据和经验公式进行预测。早期，经验模型主要是根据观测数据和历史火灾数据，采用统计分析[9]和时间序列分析[10]等方法建立模型，如灰色模型[11]、回归模型[12]等。这些模型通常只考虑历史数据和统计分析，无法将更多的影响因素和非线性关系考虑在内，预测精度有限。后来，随着计算机技术和遥感技术的发展，研究者开始运用机器学习方法来建立经验模型[13-16]，这些方法的优势在于能够更全面、更准确地处理多维数据，提高预测精度。

基于物理模型的方法是利用数学模型来描述森林火灾的发生和发展过程。这种方法通常需要考虑各种气象、地理、植被等因素对森林火灾的影响，通过建立物理模型对火灾进行模拟和预测[17]。例如，利用地形地貌、植被信息、风向风速等因素建立 3D 物理模型，对火场蔓延进行模拟，实现对森林火灾危险等级的精确预测[18]，这是目前研究热点之一。此外，基于遥感数据的森林火灾预测方法也受到了广泛关注。遥感数据能够提供丰富的信息，如植被覆盖率、土地利用类型、地形高度等，这些信息可以帮助预测火灾的发生和蔓延[19]。

在气象数据驱动的森林火灾风险等级预测领域，研究者们已经开展了大量工作。在机器学习算法的选择上，当前研究主要聚焦于两类方法：传统的非集成学习算法和集成学习算法。非集成学习算法包括 KNN[20]、决策树[21]、支持向量机[22]、贝叶斯[23]、k-means[24]和 BP 神经网络[25]、逻辑回归[26]等。这些方法基于各自的理论基础和优化目标，展现出不同的特性及适用范围。另一方面，集成学习算法通过结合多个单一学习器，创建一个更为强大的综合模型，以提高预测的准确性和模型的泛化能力。集成学习通过合并多个模型来减少单一模型的偏差和方差，从而在预测任务中实现更优的性能。总的来说，无论是非集成学习算法还是集成学习算法，它们都为森林火灾风险预测提供了多样的工具和方法，研究者可以根据具体问题和数据特点选择合适的算法来进行预测。

尽管基于气象数据的森林火灾预测方法已经取得了显著的研究成果，然而仍存在一些亟待解决的问题和有待改进的方向。其中，最为突出的问题在于数据采集的不完善和不精确，这一问题直接影响了模型的预测精度，使其难以达到理想水平。此外，为了应对更为复杂和多变的环境，模型的鲁棒性也亟待提升，以确保其在各种条件下都能保持稳定的预测性能[27]。

此外，在将森林火灾预测方法应用于实际场景时，还需要考虑数据采集、预处理和分析等诸多环节。首先，需要采集高质量的气象数据和其他相关数据，并对数据进行预处理和清洗，确保数据的准确性和完整性[28]。然后，需要选择合适的算法和模型，进行特征提取和数据建模，进行模型训练和评估。最后，需要对模型的预测结果进行可视化和解释，以便将复杂的预测结果以直观的方式呈现给决策者，为他们制定行动计划提供有力的依据。

因此，未来的研究可以从多个维度展开深入探索。一方面，可以致力于进一步完善气象数据的采集和处理方法，提高数据的精确性和可靠性[29]，为预测模型提供更为坚实的基础。另一方面，可以尝试引入更多元化的因素和数据，如遥感数据、人类活动等，从而构建更为综合和全面的预测模型。此外，还可以积极探索物理模型与机器学习方法相结合的新路径，开展

实现对森林火灾危险等级更为精准、细致的预测研究，为森林防火工作提供更为有力的科学支撑。其中，物理模型可以为机器学习算法提供更为丰富和准确的特征数据，从而提高算法的预测能力；而机器学习方法则可以帮助物理模型进行更加快速、高效和准确的参数优化和预测分析，加速研究进程。

总之，基于气象数据的森林火灾危险等级预测研究是一个涉及多学科的复杂课题，在理论研究和实践应用方面都具有重要的价值和意义。未来还需要进一步深入研究和探索，结合多种方法和技术，提高预测精度和效率，为保护森林资源和人类生命财产安全做出更大的贡献。

3 方法

3.1 数据收集与处理

3.1.1 数据来源

据统计，2020 年全国共发生森林火灾 1153 次，火场总面积为 25081 公顷，受害的森林面积为 8526 公顷，共造成人员伤亡 41 人，其他损失折款为 10077.7 万元。各省市中，广西壮族自治区是森林火灾发生次数最多的地区，其在 2020 年共发生了 206 次森林火灾，火场总面积为 2103 公顷，受害的森林面积为 786 公顷，这些火灾不仅造成了两人的人员伤亡，还带来了损失折款高达 247.1 万元的其他损失。这些统计数据凸显了加强森林防火工作的重要性和紧迫性。因此，本研究选取广西省为研究对象，从国家气象科学数据中心（http://data.cma.cn/）搜集了 1990 年至 2019 年间详尽的气象数据，包括日平均温度、日平均湿度、日平均风速、日降水量等数据。同时，还系统整理了广西省内同一时期内的森林火灾等级信息，确保数据的完整性和准确性。本次研究将气象数据和森林火灾等级信息整合在一起，作为模型的输入和输出数据，以期实现对森林火灾风险更为精准、科学的评估和预测。

3.1.2 数据预处理

为满足算法分析的要求，必须在预处理阶段对采集到的数据进行检查与合并。因为任何基于不可靠数据做出的决策，其精度都会大打折扣。数据预处理技术，如数据清洗、数据插值、数据提取等，不仅能显著提高数据挖掘模式的准确性，还能有效减少实际数据挖掘所需的时间。但该阶段面临以下问题：

(1) 数据异常：地理天气参数应在合理范围内，以中国南京为例，温度应保持在[-20 ℃, 40 ℃]范围内，湿度则应在 0%到 100%范围内。一旦天气参数超出范围，便被视为数据异常。

(2) 数据缺失：基于无线传感器网络技术，每天至少采集一次天气参数。但传感器一旦出现故障或无法工作，相应的数据参数就会丢失，这就是所谓的数据缺失。

在这种情况下，采集的气象数据可能是不可靠的，因此必须执行清洗技术。数据清洗无疑是数据分析前不可或缺的关键步骤。当数据出现异常时，我们会将异常的数据值替换为字符串'NAN'；当数据缺失时，同样可以使用字符串'NAN'来标记这些缺失的数据位置。

在异常或缺失的数据从原始数据中被移出后，将进入数据插值阶段，这一阶段的主要任务是通过只保留一个时间戳来规范化数据。为了处理包含字符串'NAN'的数据，可以在这一阶段中使用几个库和数据插值方法。数据插值策略主要有以下四种：

(1) 前一个数据

带有字符串'NAN'的数据将被以前的数据替换或填充，例如，如果$X_i = \{date, temperature = 'NAN'\}$，其以前的数据是$X_{i-1}$={$date$, $temperature$=20 ℃}。则经过数据插值后，

$X_i(temperature)$=$X_{i-1}(temperature)$。

(2) 后一个数据

带有字符串'NAN'的数据将被其后期相邻的有效数据替换或数填充。

(3) 年平均数

带有字符串'NAN'的数据将被该数据所在年的平均数据所取代或填充。

(4) 前一个和后一个数据的平均数据

带有字符串'NAN'的数据将被前一个和后一个数据的平均值所取代或填充。对于天气参数，它可能只在相邻的时间内略有变化。例如，如果连续三天的天气数据是X_{i-1}={$date$, $temperature$=20 ℃}，X_i={$date$, $temperature = 'NAN'$}，X_{i+1}={$date$, $temperature$ = 21 ℃}，此时，可以假设X_i={$date$, $temperature$=22 ℃}。因此，一般情况下，数据插值应用的方法是用前一个和后一个数据的平均值来代替字符串'NAN'的数据。这种方法之所以广泛应用，不仅在于其普遍性，易于理解和实现，同时处理速度也相当迅速。更值得一提的是，它可以用多种编程语言轻松实现，为数据预处理

提供了极大的便利性和灵活性。

3.2 模型选择和设计

本研究选取了 10 种机器学习算法，包括 6 种非集成算法和 4 种集成算法，旨在构建并深入对比森林火灾危险等级预测模型的性能。这些算法分别是：

(1) KNN：KNN 算法的核心思想在于通过考察样本在特征空间中的 K 个最近邻的类别归属，来预测该样本的类别。选择合适的距离度量（例如欧氏距离）和确定最优的 K 值是 KNN 算法的关键所在。KNN 算法直观易懂，实现简便，但计算成本会随数据集规模的增长而显著增加。

(2) 决策树：决策树算法通过递归地选择最佳特征并对数据进行分割，构建层次分明的树形结构模型。每个内部节点代表一个特征上的测试，每个分支代表测试的结果，而每个叶节点代表最终的类别。决策树模型直观易懂，易于解释，但在处理噪声数据时容易发生过拟合现象。

(3) SVM：SVM 算法的目标是找到一个能够最大化不同类别间间隔的超平面。对于线性不可分的数据，SVM 通过核技巧将数据映射到高维空间，从而实现线性分割。SVM 在处理高维数据和小样本集时表现出色，但对大规模数据集的训练可能较为缓慢。

(4) 贝叶斯：贝叶斯分类器基于贝叶斯定理进行概率预测，通过计算每个类别的后验概率来进行分类决策。朴素贝叶斯假设特征之间相互独立，这一假设简化了计算过程，但也可能忽略了特征间的潜在关联。然而，贝叶斯算法在处理大规模数据集时表现优异，计算效率极高。

(5) BP 神经网络：BP 神经网络是一种多层前馈网络，通过反向传播算法进行训练，不断调整网络中的权重和偏置来最小化预测误差。BP 神经网络能够逼近任何连续函数，但训练过程可能陷入局部最优的风险，并且对初始权重和学习率的选择较为敏感。

(6) 逻辑回归：逻辑回归虽然名字中有“回归”，但实际上是一种分类算法。它通过 sigmoid 函数将线性模型的输出映射到[0,1]区间，用于表示某类别的概率。逻辑回归在处理二分类问题时尤其流行，且模型参数可以通过最大似然估计进行高效求解。

(7) 随机森林：随机森林通过构建多个决策树并进行投票来提高预测准确性。每棵树在训练过程中使用随机选择的特征和样本子集，这既增加了模型的多样性，也有效降低了过拟合的风险。尽管，随机森林在许多实际问题中都表现出色，但由于其内部机制的复杂性，模型的解释性相对较差。

(8) AdaBoost：AdaBoost 是一种自适应的集成学习方法，它通过迭代地调整样本权重和弱分类器的权重来增强模型性能。在每一轮迭代中，AdaBoost 会根据分类器的表现调整样本权重，使得错误分类的样本在后续迭代中获得更多关注。尽管 AdaBoost 对异常值较为敏感，但在众多分类问题中，它都展现出了高准确率的特点。

(9) Gradient Boost 算法：Gradient Boost 通过逐步构建一系列弱学习器（通常是决策树），每个学习器都试图纠正前一个学习器的错误。通过累积这些学习器的预测结果，Gradient Boost 能够实现高精度的预测。然而，为充分发挥其性能，需要仔细调整学习率和其他超参数，以确保模型的稳定性和泛化能力。

(10) XGBoost 算法：XGBoost 是 Gradient Boost 算法的一种高效实现，它通过引入正则化项和并行计算技术来优化目标函数，提高了训练速度和模型性能。XGBoost 提供了丰富的参数来控制模型的复杂度和训练过程，这使得它在许多机器学习竞赛中都取得了优异的成绩。

通过对这些算法的深入描述，我们可以更好地理解它们在森林火灾危险等级预测中的潜在优势和局限性。在实际应用中，选择合适的算法需要考虑数据特性、问题复杂度以及计算资源等因素。

3.3 性能评估指标

性能评估指标是评价模型预测性能的关键指标。本研究采用以下几个指标来评估模型的性能：

(1) 准确率（Accuracy）：是模型预测正确的样本数占总样本数的比例。准确率越高，模型的预测性能越好。

(2) 精确率（Precision）：是模型预测为正例且预测正确的样本数占所有预测为正例的样本数的比例。精确率越高，模型预测为正例的准确

性越高。

(3) 召回率（Recall）：是模型预测为正例且预测正确的样本数占所有真实正例的样本数的比例。召回率越高，模型识别正例的能力越强。

(4) F1 值（F1-score）：是精确率和召回率的调和平均值，是综合评价指标。F1 值越高，模型的预测性能越好。

(5) ROC 曲线和 AUC 值：ROC 曲线是用于衡量二分类模型性能的重要指标，它以假阳性率为横轴，真阳性率为纵轴，描述了模型在不同阈值下的性能表现。AUC 值是 ROC 曲线下的面积，反映了模型分类能力的好坏。AUC 值越大，模型的预测性能越好。

通过这些评估指标的全面考量，能够深入对比不同机器学习算法在基于气象数据的森林火灾危险等级预测中的性能差异，进而筛选出最优算法。这将为森林火灾的预防和控制工作提供有力参考，有助于提升森林防火工作的准确性和效率。

4 模型对比试验结果

4.1 实验设计和数据集划分

本研究的实验数据来源于广西省 1990-2019 年间的气象数据和森林火灾危险等级信息，其中气象数据包括空气温度、相对湿度、风速和降雨量。为确保研究的客观性与科学性，我们随机打乱并拆分了这近三十年的每日气象数据，最终得到 10957 条数据样本。其中，70%的数据用于模型的训练，而剩余的 30%则用于测试模型的预测性能。本研究选择了气象因素作为模型的输入特征，输出为森林火灾危险等级，共分为 5 个等级：低、较低、中、较高和高。

在研究方法上，本次研究共选择了 10 种机器学习算法进行性能比较。整个实验流程包括数据预处理、数据分割、模型训练、性能测试以及结果比较分析等关键步骤。

Step 1：在预处理阶段，本研究对原始数据进行了缺失值处理、异常值处理和数据标准化等操作，以保证数据的准确性和可靠性。

Step 2：为了训练模型并测试模型性能，本研究将数据集分为训练集和测试集。训练集用于模型的构建与学习，而测试集则用于客观评估模型的预测能力。

Step 3：在训练模型阶段，本研究分别采用 10 种不同的机器学习算法对训练集进行训练，构建了 10 个预测模型。这些模型通过学习数据集的模式和规律，能够有效识别气象因素与森林火灾危险等级之间的复杂关系，从而实现森林火灾危险等级的预测。

Step 4：在模型性能测试环节，本次研究将训练好的模型应用于测试集，并计算了包括准确率、召回率和 F1 值等在内的多项性能指标。这些指标全面反映了模型在预测森林火灾危险等级方面的表现，为评估模型优劣提供有力依据。

Step 5：在比较分析阶段，本研究对不同模型的性能指标进行比较和分析，旨在选出性能最优的模型。本研究综合考虑了模型的预测精度、稳定性以及鲁棒性等多个方面，以确保所选模型在实际应用中能够发挥最佳效果。

4.2 不同模型的实验结果比较和分析

在本研究中，选择了 10 种常用的机器学习算法进行比较。首先，对这些算法在训练集上进行了训练，并在测试集上进行了预测。为全面、客观地评价各算法的预测性能，本次实验计算了精确度、准确率、召回率、F1 值和 AUC 值这 5 个性能指标，如表 1 所示。为了更直观地展示各算法在分类性能上的优劣，本研究还绘制了 ROC 曲线图，如图 2 所示。

表 1 林火预测模型的性能指标

模型	精确度	准确率	召回率	F1	AUC
KNN	0.797	0.662	0.797	0.711	0.637
决策树	0.641	0.661	0.641	0.651	0.625
SVM	0.834	0.628	0.834	0.714	0.657
贝叶斯	0.830	0.653	0.830	0.715	0.651
BP	0.834	0.628	0.834	0.714	0.667
逻辑回归	0.834	0.628	0.834	0.714	0.663
随机森林	0.798	0.668	0.798	0.713	0.651
AdaBoost	0.833	0.628	0.833	0.714	0.691
Gradient Boost	0.832	0.739	0.832	0.716	0.668
XGBoost	0.818	0.675	0.818	0.717	0.669

根据表 1 和图 2 的结果，可以清晰地看出不同算法在森林火灾危险等级预测中的性能差异。其中，KNN、SVM、贝叶斯、BP 神经网络、逻辑回归、AdaBoost、Gradient Boost 和 XGBoost 的 AUC 值都在 0.65 以上，说明这些模型的分类效果较好。而决策树和随机森林的 AUC 值略低，说明它们的分类效果相对较弱。从精确度和准确率来看，KNN 和随机森林的表现相对较好，其次是 XGBoost、AdaBoost、BP 神经网络、Gradient Boost 和贝叶斯，而决策树、SVM 和逻辑回归的表现

较为一般。从召回率和 F1 值来看，KNN、SVM、贝叶斯、BP 神经网络、逻辑回归、AdaBoost、Gradient Boost 和 XGBoost 的表现相对较好，其中，KNN 和 SVM 的召回率和 F1 值最高，表明这两种模型对于正样本的识别能力较强。而决策树和随机森林的召回率和 F1 值相对较低，说明这两种模型对于正样本的识别能力较弱。

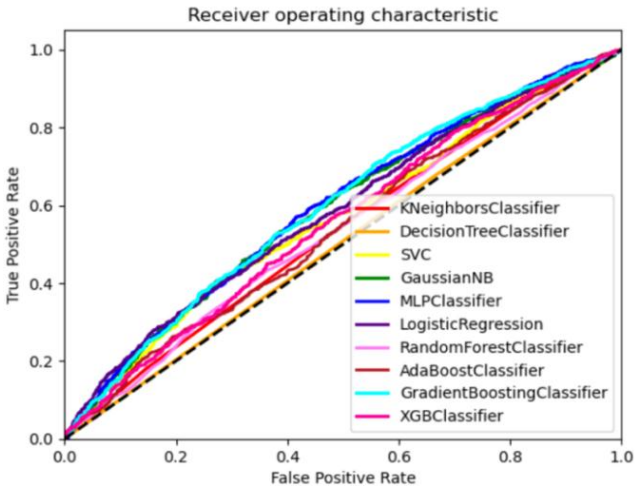

图 2 林火预测模型性能比较 ROC 曲线

综上所述，KNN、SVM、贝叶斯、BP 神经网络、逻辑回归、AdaBoost、Gradient Boost 和 XGBoost 的表现相对较好，其中 KNN 和 SVM 的分类效果最好，而决策树和随机森林的表现相对较差。此外，不同模型的性能表现还与数据集的质量、大小等因素有关，因此在实际应用中需要根据具体情况选择最适合的模型。

4.3 摘要和关键词模型稳定性

除比较各个模型在不同指标下的表现外，还需要考虑模型性能的稳定性和鲁棒性，这些指标反映了模型的泛化能力和可靠性。为全面评估模型的稳定性，本研究使用了 5 折交叉验证技术，即将训练集分为 5 个互斥的子集，轮流使用其中的 4 个子集作为训练集，剩余 1 个子集作为测试集，进行 5 次独立的实验，确保每个子集都有机会作为测试集接受模型的检验。通过 5 折交叉验证技术，本研究获得了更为丰富可靠的性能评估数据，具体如表 2 至表 5 所示。

表 2 林火预测模型 5 折交叉验证 F1 值

模型	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
KNN	0.6954	0.6936	0.70356	0.6973	0.6989
决策树	0.6571	0.6423	0.6388	0.6474	0.6383
SVM	0.6998	0.6998	0.7003	0.7003	0.6997
贝叶斯	0.6987	0.6969	0.6996	0.6996	0.6947
BP	0.7010	0.6990	0.7000	0.7001	0.6998
逻辑回归	0.6998	0.6998	0.7003	0.7003	0.6997
随机森林	0.6938	0.6990	0.6915	0.7032	0.7089
AdaBoost	0.6998	0.6998	0.6998	0.6991	0.6992
Gradient Boost	0.7000	0.7015	0.7023	0.7025	0.7008
XGBoost	0.7009	0.7032	0.7066	0.7057	0.6975

表 3 林火预测模型 5 折交叉验证精确度

模型	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
KNN	0.7886	0.7863	0.7916	0.7861	0.7839
决策树	0.6471	0.6467	0.6337	0.6378	0.6273
SVM	0.8237	0.8237	0.8240	0.8240	0.8236
贝叶斯	0.8201	0.8109	0.8195	0.8172	0.8085
BP	0.8232	0.8237	0.8240	0.8240	0.8236
逻辑回归	0.8237	0.8237	0.8240	0.8240	0.8236
随机森林	0.7909	0.7904	0.7761	0.7857	0.7839
AdaBoost	0.8237	0.8237	0.8231	0.8213	0.8227
Gradient Boost	0.8205	0.8214	0.8227	0.8185	0.8190
XGBoost	0.8109	0.8114	0.8085	0.8053	0.8030

表 4 林火预测模型 5 折交叉验证准确率

模型	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
KNN	0.6433	0.6405	0.6623	0.6464	0.6486
决策树	0.6556	0.6623	0.6563	0.6635	0.6551
SVM	0.6133	0.6133	0.6137	0.6137	0.6131
贝叶斯	0.6134	0.6296	0.6159	0.6331	0.6184
BP	0.6133	0.6130	0.6137	0.6139	0.6131
逻辑回归	0.6133	0.6133	0.6137	0.6137	0.6131
随机森林	0.6475	0.6461	0.6387	0.6672	0.6594
AdaBoost	0.6133	0.6133	0.6135	0.6134	0.6129
Gradient Boost	0.6497	0.6739	0.6804	0.6584	0.6740
XGBoost	0.6486	0.6597	0.6731	0.6648	0.6349

表 5 林火预测模型 5 折交叉验证召回率

模型	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
KNN	0.6638	0.6615	0.6668	0.6613	0.6591
决策树	0.5246	0.5224	0.5020	0.5130	0.4970
SVM	0.6989	0.6989	0.6992	0.6992	0.6987
贝叶斯	0.6953	0.6861	0.6947	0.6924	0.6837
BP	0.6989	0.6975	0.6988	0.6983	0.6988
逻辑回归	0.6989	0.6989	0.6993	0.6992	0.6988
随机森林	0.6674	0.6693	0.6518	0.6572	0.6645
AdaBoost	0.6989	0.6989	0.6983	0.6965	0.6979
Gradient Boost	0.6957	0.6975	0.6974	0.6937	0.6947
XGBoost	0.6861	0.6866	0.6837	0.6805	0.6782

根据表 2 至表 5 中的交叉验证结果，可以全面评估不同模型的稳定性和鲁棒性。通过 5 折交叉验证，每个模型在不同折次下的 F1 值、精确度和准确率得分都有所不同，但各个折次之间的差异不大。这表明，这些模型在测试集上的表现相对稳定，具有较好的泛化能力和可靠性。其中，XGBoost 模型在 F1 值和精确

度上表现最好，但准确率略低；支持向量机模型在精确度和准确率上表现最好，但 F1 值稍低。

5 总结

根据实验结果，不同机器学习算法在基于气象数据的森林火灾危险等级预测中表现不同。从 F1 值、精确度和准确率等指标综合考虑，XGBoost 模型在 F1 值和精确度上表现最好，但准确率略低；支持向量机模型在精确度和准确率上表现最好，但 F1 值稍低。KNN、贝叶斯、逻辑回归、AdaBoost 和 Gradient Boosting 等模型的表现相对较为平均，综合考虑性能和稳定性较好。而决策树模型的表现相对较差，可能是因为该模型容易过拟合，导致在测试集上的表现不稳定。由于本研究所采用的数据仅包括气象数据，未考虑其他因素对森林火灾的影响，因此模型的预测能力仍有限制。此外，气象数据的获取和质量也可能影响模型的预测效果。

未来可以考虑将更多因素纳入预测模型，如地形、植被、人类活动等因素，构建更为全面的预测模型，以提高预测准确性。此外，可以探索更为先进的机器学习算法和深度学习技术，如深度神经网络等，以进一步提高模型的预测能力和稳定性。同时，在实际应用中还需要注意数据质量和采集的完整性，以确保模型的预测结果的可靠性和准确性。

参考文献

[1] Edwards R B, Naylor R L, Higgins M M, et al. Causes of Indonesia’s forest fires [J]. World Development, 2020, 127: 104717.

[2] Pourghasemi H R, Gayen A, Lasaponara R, et al. Application of learning vector quantization and different machine learning techniques to assessing forest fire influence factors and spatial modelling [J]. Environmental research, 2020, 184: 109321.

[3] Ying L, Cheng H, Shen Z, et al. Relative humidity and agricultural activities dominate wildfire ignitions in Yunnan, Southwest China: Patterns, thresholds, and implications [J]. Agricultural and Forest Meteorology, 2021, 307: 108540.

[4] Zheng Z, Gao Y, Yang Q, et al. Predicting Forest fire risk based on mining rules with ant-miner algorithm in cloud-rich areas [J]. Ecological Indicators, 2020, 118: 106772.

[5] Alley R B, Emanuel K A, Zhang F. Advances in weather prediction [J]. Science, 2019, 363(6425): 342-344.

[6] Ma W, Feng Z, Cheng Z, et al. Identifying Forest fire driving factors and related impacts in china using random forest algorithm [J]. Forests, 2020, 11(5): 507.

[7] 朱馨, 李建微, 郭伟, 毕胜, 伍跃飞. 基于机器学习的森林火险预测模型 [J]. 中国安全科学学报, 2022, 32(09): 152-157.

[8] 高超, 林红蕾, 胡海清, 宋红. 基于气象因子的黑龙江黑河林火发生概率预测 [J]. 森林与环境学报, 2022, 42(05): 529-535.

[9] Huesca M, Litago J, Palacios-Orueta A, et al. Assessment of forest fire seasonality using MODIS fire potential: A time series approach [J]. Agricultural and Forest Meteorology, 2009, 149(11): 1946-1955.

[10] 高博, 陈响, 单仔赫, 韩喜越, 单延龙, 尹赛男, 于渤. 基于 Logistic 回归模型的大兴安岭地区林火发生概率预测研究 [J]. 中国安全生产科学技术, 2022, 18(11): 163-168.

[11] 王丹丹, 黄家荣, 刘伟, 孟庆玲, 田鹏飞. 基于人工神经网络的森林火险预测 [J]. 西北林学院学报, 2010, 25(03): 143-146.

[12] Xie L, Zhang R, Zhan J, et al. Wildfire risk assessment in Liangshan Prefecture, China based on an integration machine learning algorithm [J]. Remote Sensing, 2022, 14(18): 4592.

[13] Pastor E, Zárate L, Planas E, et al. Mathematical models and calculation systems for the study of wildland fire behaviour [J]. Progress in Energy and Combustion Science, 2003, 29(2): 139-153.

[14] Pai M L, Varsha K S, Arya R. Application of Artificial Neural Networks and Genetic Algorithm for the Prediction of Forest Fire Danger in Kerala [C] // Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2. Springer International Publishing, 2020: 935-942.

[15] Lin H, Liu X, Wang X, et al. A fuzzy inference and big data analysis algorithm for the prediction of forest fire based on rechargeable wireless sensor networks [J]. Sustainable Computing: Informatics and Systems, 2018, 18: 101-111.

[16] 姚艳霞, 苗双喜, 黄旭, 张波. 耦合 Rothermel 模型与粒子系统的林火蔓延模拟方法 [J]. 测绘通报, 2022(08): 75-80.

[17] 张文文, 王劲, 王秋华, 曹恒茂, 王金波, 左军宏, 王加庆. 森林火灾遥感探测技术研究进展 [J]. 西北林学院学报, 2023(01): 123-130.

[18] Chuvieco E, Aguado I, Salas J, et al. Satellite remote sensing contributions to wildland fire science and management [J]. Current Forestry Reports, 2020, 6: 81-96.

[19] 高超, 林红蕾, 胡海清, 宋红. 我国林火发生预测模型研究进展 [J]. 应用生态学报, 2020, 31(09): 3227-3240.

[20] Karo I M K, Amalia S N, Septiana D. Wildfires Classification Using Feature Selection with K-NN, Naïve Bayes, and ID3 Algorithms [J]. Journal of Software Engineering, Information and Communication Technology (SEICT), 3(1): 15-24.

[21] Jaafari A, Zenner E K, Pham B T. Wildfire spatial pattern analysis in the Zagros Mountains, Iran: A comparative study of decision tree based classifiers [J]. Ecological informatics, 2018, 43: 200-211.

[22] Jaafari A, Pourghasemi H R. Factors influencing regional-scale wildfire probability in Iran: an application of random forest and support vector machine M] // Spatial modeling in GIS and R for Earth and environmental sciences. Elsevier, 2019: 607-619.

[23] 黄琼, 司颖, 王浩宇. 基于加权贝叶斯网络的森林火灾风险评估 [J]. 消防科学与技术, 2021, 40(11): 1671-1676.

[24] Sirat E F, Setiawan B D, Ramdani F. Comparative analysis of K-means and isodata algorithms for clustering of fire point data in Sumatra region[C]//2018 4th International Symposium on Geoinformatics (ISyG). IEEE, 2018: 1-6.

[25] WU C, ZHONG Y, LIU H, et al. A Novel Application of BP Neural Networks to Evaluate the Safety of Power Grid in Wildfire Disasters [J]. International Journal of Simulation: Systems, Science and Technology, 2016, 17(27): 32.1-32.5.

[26] Papagiannaki K, Giannaros T M, Lykoudis S, et al. Weather-related thresholds for wildfire danger in a Mediterranean region: The case of Greece [J]. Agricultural and Forest Meteorology, 2020, 291: 108076.

[27] Demin G, Haifeng L, Anna J, et al. A forest fire prediction system based on rechargeable wireless sensor networks [C] // 2014 4th IEEE International Conference on Network Infrastructure and Digital Content. IEEE, 2014: 405-408.

[28] 高德民, 林海峰, 刘云飞, 吴国新. 基于无线传感网的森林火灾 FWI 系统分析 [J]. 林业科技开发 2015, 29(01): 105-109.

[29] Vikram R, Sinha D, De D, et al. PAFF: predictive analytics on forest fire using compressed sensing based localized Ad Hoc wireless sensor networks [J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12: 1647-1665.