Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(1), 143-152
Published Online February 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.131016

Hans汉斯

基于随机森林算法的机器学习分类研究综述

向进勇[1,2]，王振华[1,2]，邓芸芸[1,2]

[1]伊犁师范大学网络安全与信息技术学院，新疆 伊宁
[2]伊犁师范大学伊犁河谷智能计算研究与应用重点实验室，新疆 伊宁

收稿日期：2023年4月18日；录用日期：2024年2月23日；发布日期：2024年2月29日

摘　要

机器学习是实现人工智能的重要技术,随机森林算法是机器学习的代表算法之一。随机森林算法以简单、有效而闻名工业界和学术界，它是基于决策树的分类器，通过投票选择最优的分类树。随机森林算法有可变重要性度量、包外误差、近似度等优秀特性，因此随机森林被广泛的应用到分类算法中。目前，不仅在医学、农业、自然语言处理等领域被广泛提及，而且在垃圾信息分类、入侵检测、内容信息过滤、情感分析等方面都有广泛的应用。本文主要介绍了随机森林的构建过程以及随机森林的研究现状，主要从分类性能、应用领域以及分类效果加以介绍，分析随机森林算法优缺点以及研究人员对随机森林算法的改进，希望通过分析能够让初学随机森林算法的研究人员掌握随机森林的理论基础。

关键词

决策树，随机森林，机器学习

A Review of Machine Learning Classification Based on Random Forest Algorithm

Jinyong Xiang[1,2], Zhenhua Wang[1,2], Yunyun Deng[1,2]

[1]School of Cyber Security and Information Technology, Yili Normal University, Yining Xinjiang
[2]Key Laboratory of Intelligent Computing Research and Application, Yili Normal University, Yining Xinjiang

Received: Apr. 18th, 2023; accepted: Feb. 23rd, 2024; published: Feb. 29th, 2024

Abstract

Machine learning is an important technology to realize artificial intelligence, and random forest algorithm is one of the representative algorithms of machine learning. The random forest algorithm is well-known in industry and academia for its simplicity and effectiveness. It is a decision tree-based classifier that selects the optimal classification tree through voting. Random forest al-

文章引用: 向进勇, 王振华, 邓芸芸. 基于随机森林算法的机器学习分类研究综述[J]. 人工智能与机器人研究, 2024, 13(1): 143-152. DOI: 10.12677/airr.2024.131016

gorithm is widely used in classification algorithms because of its excellent characteristics such as variable importance measure, out-of-envelope error and approximation. At present, it is not only widely mentioned in medicine, agriculture, natural language processing and other fields, but also widely used in junk information classification, intrusion detection, content information filtering, sentiment analysis and other aspects. This paper mainly introduces the construction process of random forest and the research status of random forest, mainly from the classification performance, application field and classification effect, analyzes the advantages and disadvantages of random forest algorithm and the improvement of random forest algorithm by researchers, hoping that through analysis, researchers who have just learned random forest algorithm can master the theoretical basis of random forest.

Keywords

Decision Trees, Random Forests, Machine Learning

1. 引言

机器学习(ML)作为人工智能的一个分支，使用计算机通过计算对客观事物进行分析，机器学习根据语料标注情况可以分为有监督学习、无监督学习以及半监督学习，(Abdulqader 等人[1]，2020 年；Adeen 等人[2]，2020 年)通过无监督学习对目标进行了预测，主要通过经验进行学习，在结果不明的情况下做出正确预测或选择。有监督学习是机器学习算法中的一大类，有监督学习的目标是根据以前的数据(样本数据)预测新实例的类别标签(离散的、无序的值)。(Zeebaree 等人[3]，2019b；Sadiq 等[4]，2020)进行了有监督学习分类，主要根据样本数据(“训练数据”)创建模型。决策树是一种有监督学习分类算法，它以树的形式构造模型。(Abdulqader 等人[5]，2020；Zeebaree 等人[3]，2019a)通过特征将信息分割成相应的子集，逐步构建成决策树。(Zebari 等[6]，2020a)论文中提出的决策树是一种图表结构，是一种评估过程，用来预测结果的概率。决策树的一个分支对应未来的选择、结果或响应，最终的结果由树最远的节点表示也就是叶子节点。(Sadeeq & Abdulazeez [7], 2018; Najat & Abdulazeez [8], 2017)等认为决策树有以下优点对应用领域来说它的要求低，易于算法设计人员理解，但是对于无监督学习的效果还有待提高，还有决策树更加适合分类而不是预测，使用决策树对类标签中的数据集进行预测可能会导致预测结果的不确定、预测结果的不准确。相对于其他的研究人员(Zebari 等人[6]，2019a)并没有使用单一分类器，而是使用了多个分类器，这些分类器通过决策树进行集成。利用决策树对大量数据进行分类时，在各自的类别标签数据中会产生估计不准确的情况，因此(Mienye 等人[9]，2019)使用多个决策树集成为一个分类器，并不仅仅使用单一的分类器，这种方法就是随机森林。

随机森林是一种可扩展的、易于使用的机器学习算法，在多数情况下，即使不进行参数调优，也能获得很好的分类结果。由于随机森林算法的灵活性和多样性，它成了最常用的机器学习算法之一，经常被用来解决分类问题以及回归问题。在集成的分类方法中，研究人员(Das 等[10].，2007)根据特定的函数构建多个分类器。然后将多个分类器组合起来创建出一个新的分类器；随机森林其实是一种比单个决策树更强大且性能更加稳定的建模方法。许多决策树组合起来还可以限制过拟合以及偏见造成的误差等问

题，从而产生更高的准确率，随机森林算法是一种强大的机器学习方法，已在各个领域取得了广泛的成功。下面将讨论随机森林算法的优点和缺点。优点：(1) 高准确性：随机森林在许多情况下都能够实现高度准确的预测，这是由于它是一个集成学习方法，通过组合多个决策树的结果来减小单个模型的误差。(2) 抗过拟合：随机森林通过随机抽样数据和特征来构建多个决策树，减少了过拟合的风险。每棵树都是在不同的子集数据上训练的，从而提高了模型的泛化能力。(3) 处理大规模数据：随机森林能够有效处理大规模数据集，因为它可以并行训练多个决策树，从而提高了训练速度。(4) 对于不平衡数据集具有良好的表现：随机森林可以处理不平衡的分类问题，因为它可以通过调整类别权重来平衡不同类别的重要性。(5) 特征重要性评估：随机森林可以提供每个特征的重要性评分，帮助识别哪些特征对于模型的预测最为关键。缺点：(1) 复杂性：随机森林通常包含多个决策树，因此模型的结构相对复杂，不易可视化和解释。这可能会使模型在某些应用中显得笨重。(2) 计算资源需求：由于随机森林涉及多个决策树的构建和集成，因此需要更多的计算资源和内存，特别是在大型数据集上。(3) 可能不适合高维数据：对于高维数据，随机森林可能表现不佳，因为在高维空间中决策树的分割可能变得困难，容易导致过拟合。(4) 预测速度较慢：与一些线性模型相比，随机森林的预测速度较慢，因为需要对多个决策树的结果进行集成。(5) 不适用于序列数据：随机森林主要用于静态数据集的分类和回归问题，不适用于处理时间序列或序列数据。本文其余部分组织如下：第 2 节背景理论，第 3 节相关工作，第 4 节比较和讨论，第 5 节结论。

2. 理论背景

大数据时代背景下，由于数据量巨大，存在许多分类困难的问题，(Zebari 等人[6]，2020b)许多传统的分类算法在某些情况下不能得到理想的结果，随机森林分类算法在某些分类问题上表现出理想的分类效果，随机森林本质上由一组决策数构成，将决策树的结果合并成最终的结果。研究人员(Schonlau & Zou 等人[11]，2020)证明随机森林可以限制机器学习中过度拟合现象并且不会因为很小的偏差而造成很大的误差，这就是随机森林最大的优点。(Han 等人[12]，2019 年；Zhou 等人[13]，2020 年)利用随机森林中最小化方差对多个数据样本进行训练。

2.1. 决策树

决策树根据属性(特征)将一个结点划分成两个或多个子节点，(Kumar 等人[14]，2016)证明制作子节点的方式可以扩大后续子节点的同质性。(Li 等人[15]，2019)证明决策树可以在所有的属性上划分节点，然后选择最同质的子节点进行分裂。随机森林本质上是由多个决策树组成，决策树是构成随机森林的基本分类器。

常用的决策树分类算法有 ID3、C4.5、CART (Classification and Regression Tree)。ID3 算法选择信息增益最大的属性和节点分裂进行构建决策树，(Singh & Giri 等人[16]，2014)认为 ID3 只能支持离散数据处理，而且训练模型容易出现过拟合现象。C4.5 算法是 ID3 算法的增强，引入了剪枝步骤从而防止 ID3 算法出现的过拟合现象，执行过程是指定一个阈值；样本数小于给定的阈值。集合可以直接看作是一个叶子节点，这样就可以减少过拟合现象，但是阈值的选择依赖经验，因此缺乏必要的理论支持。CART 算法(Band 等人[17]，2020)主要根据杂质最小准则对样本数据进行双向递归分割，将当前的样本集合划分成为两个子样本集合，因此生成的决策树的每个非叶子节点只有两个分支(Sarker, *et al* [18], 2020)，也就是说 CART 树是二叉树，而 ID3 和 C4.5 可以是多叉树。决策树训练流程图如图 1 所示。

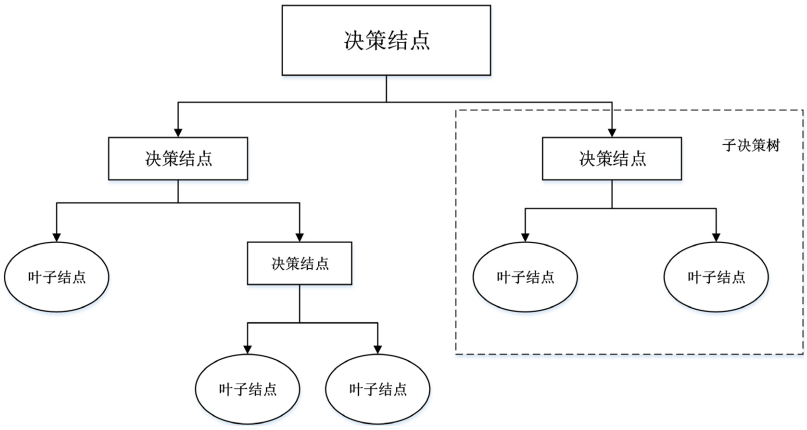

Figure 1. Decision tree training flow chart
图 1. 决策树训练流程图

2.2. 随机森林

随机森林是由独立同分布的随机变量构成的一种树形分类器，每棵树根据输入 X 进行投票。(Ozgode Yigin 等人[19]，2020)生成一个独立于先前相同分布的随机向量，并训练测试生成一棵树，提取随机森林的上限以获得两个参数作为泛化误差和个体的相互依赖性。随机森林流程图如图 2 所示。

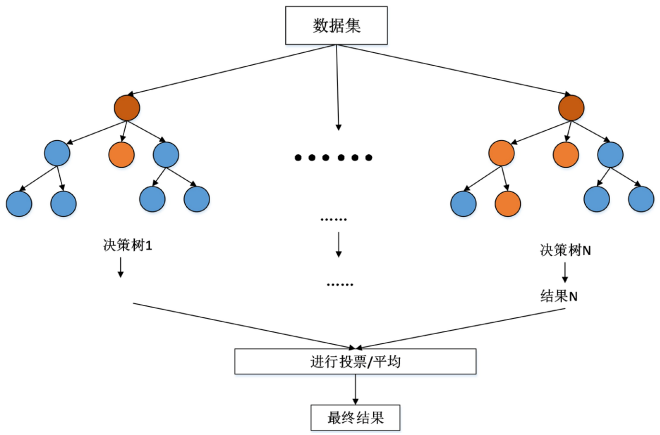

Figure 2. Flow chart of random forest
图 2. 随机森林流程图

随机森林算法的优缺点如表 1 所示。

Table 1. Advantages and disadvantages of random forest algorithm
表 1. 随机森林算法的优缺点

优点	缺点
准确率更高、能够有效地处理大型数据库，特征很多的数据表现良好。它可以快速有效地管理数千个输入变量。提供关于重要且不在“分类”中的变量的信息。提供估计不完整数据的技术。处理丢失的细节而不失准确性。可以判断出不同特征之间的相互影响，训练速度快容易做成并行方法。对于不平衡数据集，可以平衡误差。	发现的主要问题之一是过度拟合单个数据集，特别是在回归任务中。随机森林在多维度处理多值和多值属性方面存在困难。他们更喜欢多层次分类变量，对于有不同取值的属性的数据，取值划分较多的属性会对随机森林产生更大的影响。

2.3. 随机森林算法

2001 年 Breiman 提出了一种随机森林的分类算法，它是以决策树作为基本分类器的集成学习模型。研究人员(Denisko & Hoffman 等人[20]，2018)为了使随机森林适用于多个样本子集，使用 bootstrap 方法，它是利用每个样本子集创建决策树，并最终将多个决策树组合成一个随机森林，该算法被广泛应用到诸如生物信息领域对基因序列的分类和回归。研究人员(Utkin 等人[21]，2020)通过投票决定最终的分类结果。一般来说，(Demidova & Ivkina 等人[22]，2019)学者们选择关联度低的分类器目的就是提高分类的精度。随机森林算法在分类的过程中，每个基本的分类器都是决策树，因此分类效果上有相同的误差分布，因此(Abdulazeez 等人[5]，2020)完成分类效果的归纳。研究人员(Kolhe 等人[23]，2020 年；Gajowniczek 等人[24]，2020 年)取测试集的测试特征并使用随机生成的决策树来预测结果，将预测结果进行存储并对每个预测结果进行投票，将预测结果得票最高的结果作为随机森林算法的最终结果，图 3 说明了随机森林的训练过程。

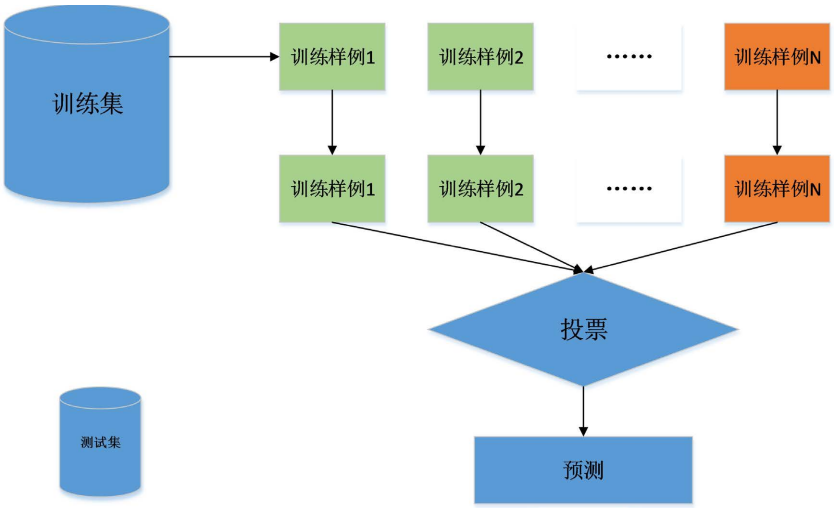

Figure 3. Process of random forest training
图 3. 随机森林训练过程

随机森林的算法的基本步骤如下：

随机森林作为机器学习中主要的分类器之一，它是由许多的独立同分布的决策树构成决策树主要研究样本的规律。(Bingzhen 等人[25]，2020)具体步骤如下，在随机森林算法中，主要有以下两个步骤，一个是随机森林的形成，另一个是对结果进行投票。在这里，首先公开随机森林构建的伪代码(Computer Science & Engineering & GZSCCET Bhatinda, Punjab, India [26], 2017)：

1. 从完整的“m”个特征中随机选择“K”个特征，其中 k << m。
2. 使用最佳分割点计算“K”个特征中的节点“d”。
3. 用最佳划分将数据划分为子节点。
4. 重复执行 1 到 3，直到节点数达到“n”。
5. 重复步骤 1 到 4“n”次创建“n”个树从而构建一个森林。

根据生成的随机森林分类器，我们对数据进行预测。用于随机森林预测的伪代码如下所示：获取测试特征使用每个随机生成的决策树来预测结果并存储预期结果(目标)。对每个预测目标进行投票考虑票数最多的预测目标作为随机森林算法的最终预测结果。

决策公式[(Das 等人[10]，2007)]使用公式 1 所示。

$$H(x)=\arg\max_{y}\sum_{i=1}^{k}I\left(h_i(x)=Y\right) \tag{1}$$

公式中 x 为测试样本数据，h_i 为单个决策树。Y 为输出变量(比如类标签)，I 为指标函数，H 为随机森林模型。

即对测试样本的每个测试树的分类结果进行总结，最终的分类结果是票数最大的类。此外，还出现了几种随机森林改进算法，它们的配对见表 2 和随机森林算法(Lmaizumi 等人[27]，2020)。

Table 2. Improved random forest algorithm
表 2. 改进随机森林算法

算法名称	与随机森林算法的不同
极端随机树(Darbanian 等人[28])	随机森林和额外树(通常称为极端随机森林)之间的关键区别在于，随机森林考虑的是每个特征，选择一个随机值用于分割，而不是确定局部最优的特征/分割组合(对于随机森林) (对于额外树)。
孤立随机森林(IForest) (Chaudhary, n.d., 2020, [29])在信用卡诈骗检测取得了不错的效果	孤立森林的功能比随机森林好一点。因此，它会生成很多决策树，然后确定从树中提取观测值所需的路径长度。
随机生存森林(Random Survival Forest) (Chen 等人[30]，2019)	构建树的决策与 RF 相同。RSF 中的每个决策树都是一个两类生存树，用于处理生存数据。对于高维生存的证据，它优于其他方法的生存分析。

3. 相关工作

随机森林算法是一种基于决策树的分类器。它通过投票方式选出最佳分类树作为最终分类器的分类算法(Zebari [11], 2019b)。它可用于分类和回归任务。它通过交叉验证提供更高的准确性(Bargarai 等[31]，2020)。目前，它在新闻分类、入侵检测、内容信息过滤、情感分析等图像处理领域有着广泛的应用。在持续增强和性能指标方面，本文献综述将主要介绍决策树、随机森林的开发方法以及随机森林的研究现状。Iwendi & Jo (2020) [32]提出了一个模型，建议应用随机森林算法，F1 分数为 0.866，通过 AdaBoost 算法在 COVID-19 的患者数据集上进行改进。此外，还指出增强型随机森林算法对不平衡的数据集也是如此。本次分析所回顾的知识表明，武汉本地人的死亡率更高。非本地人则相反。与女性患者相比，男性患者有更高的死亡率风险。受影响的患者中最大的年龄在 20 岁和 70 岁之间。Zhang & Yang (2020) [33]由于大规模的人类迁徙、土地转换和全球环境变化，沿海地区非常紧张。由于它们的地理、异质性和光谱复杂性，绘制城市化沿海地区可能非常困难。在本研究中测试了七种专注于随机森林的可变评级方法。为了选择最佳分类形式，使用 CART 和 CIT 模型实现了特征排除技术。CPVIM 已被证明在根据相关遥感数据提供稳定合理的特征排名方面更加可靠。最佳模型是通过使用 CPVIM 基于 CART 树的 NRFE 过程找到的。它仅使用 10 个特征，即 Green、NIR、SWIR1 和 SWIR2、Greenness、MSAVI、NDII、ED、SVVI 和 DEM，实现了 89.03%的总体准确率。

此外，Saenz-Cogollo & Agelli (2020) [34]提出了从单导联心电图导出的时域特征是由其数据质量严格选择的，并且通过采用(AAMI)和患者间范式原则。分类任务中最具辨别力的特征被认为是相对于 R-R 间隔和 QRS 复合波主波宽度的归一化特征。凭借前六名最具洞察力的特征和一个 40 树 RF 分类器，产生了最好的结果。MIT-BIH 心律失常数据库测量的结果是 NB、SVEB 和 VEB 组的平均精度为 96.14%，个人 F1 评级分别为 97.97%、73.06%和 90.85%。根据在可比条件下测试的最先进方法，结果是迄今为止记录的最佳性能之一。研究结果不仅表明 RF 是一种出色的心跳分类方法，而且还表明实现最先进的效率所需的特征相对较少。

此外，Chai & Zhao (n.d.) [35]提出了一种现代的 OBRF 学习方法由 OBRF-BM 和 OBRF-DIL (具有双增量学习能力的多类倾斜随机森林)组成。计划的系统通过分析测量倾斜的超平面代价来衡量合适的功能和分裂阈值。另外，将决策节点特性投影到一个随机的更高维空间中，该空间将进一步的随机性注入了集合模型，并从提升 OBRF 输出。相比之下，以样本增量和类增量的情况创建渐进方法，以使预定义的模型有效，而无需艰苦的再修订。经验发现表明，OBRF 的出色效率建议。International Conference on Artificial Intelligence and Computer Vision 国际人工智能和计算机视觉会议(2020 年) [36]提到，随机森林是配备数据子样本的决策树的变化，是使用不采样和过度采样的。作者对比了来自评估模型的不同要求的拟合指标，并评估了研究内外的结果。研究结果表明，使用比初始研究小的不平衡子样本的随机森林策略显示出相对于医学数据集使用的随机森林的更高效率和变化。

汤圣君等[37]针对现有三维点云数据分割分类方法存在分类目标内部不一致的问题，提出一种超体素随机森林与 LSTM 神经网络联合优化的室内点云高精度分类方法。根据超体素结构具备内部特征一致性的特点，对原始点云进行超体素划分，并以超体素为基本单元进行多元特征计算，搭建室内点云超体素随机森林分类模型，实现点云数据的粗分类。在公开数据集中对 13 类要素的分类精度可达到 83.2。

徐精诚等[38]提出特征选择技术与随机森林相结合的算法用于 DDoS 攻击检测。这样不仅可以进行样本降维，以降低训练成本和提高训练模型精度同时将特征选择算法嵌入随机森林的单个基学习器，将特征子集搜索范围由全部特征缩小到单个基学习器对应特征，在提高两种算法耦合性的同时提高了模型精度。

4. 比较和讨论

近期研究的回顾展示了随机森林算法在生命科学不同领域的评估。该算法作为强大的机器学习技术之一的成功点已在基于分类和回归的问题中得到认可。Iwendi & Jo (2020) [32]使用新的改进版 RF 算法成功预测了 COVID-19 患者数据集的分类及其状态。(Saenz Cogollo & Agelli, 2020) [34]报告了 RF 在心跳分类中的出色表现，该分类基于从不同患者获得的心脏数据集。(Li H, Lin J & Lei X, 2022) [39]将随机森林算法应用到玄武岩纤维混凝土抗压强度预测并取得不错的效果。(Guo Q & Zhang J, 2022)将随机森林算法应用到城市树木检测并且取得不错的分类效果。同样，RF 工具在处理医学不平衡数据集方面具有巨大潜力，并且在输出预测方面具有良好的性能(国际人工智能和计算机视觉会议，International Conference on Artificial Intelligence and Computer Vision, 2020) [36]。有的研究人员强调了基于变量选择过程的 RF 框架的计算效率，以预测最佳域区域以改进城市化沿海地区的土地覆盖分类。在这个框架中，RF 与基于相似性的方法相结合。结果清楚地表明，所开发模型的准确度(71.79%)比单独或单独应用两种方法 RF (66.67%)和基于相似性的方法(58.97%)更高。此外，RF 方法本身仍然比基于相似性的方法具有更好的精度。汤圣君等将随机森林与深度学习算法相结合取得了不错的效果。徐精诚将特征选择算法嵌入随机森林的单个基学习器应用到分布式拒绝服务攻击(DDoS)攻击检测领域降低了训练成本和提高了训练精度。随机森林算法是一种流行的机器学习技术，它通过结合多个决策树来提高预测的准确性和健壮性。Guo Q [40]将随机森林算法应用到树木分类中并且取得了不错的效果。未来随机森林算法的发展方向为算法优化和效率提升：研究者可能会探索新的方法来优化随机森林的训练过程，比如更有效的树构建算法、并行计算技术的利用，以及减少内存使用的策略。特征选择和处理的改进：进一步研究如何更有效地选择和处理特征，以提高随机森林在各种数据集上的表现。深度学习的整合：随着深度学习的发展，将深度学习技术与随机森林相结合，比如使用深度神经网络来提取特征，然后利用随机森林进行分类或回归。应用领域的拓展：随机森林在许多领域已被成功应用，如生物信息学、金融分析等。未来可能会有更多新兴领域，如物联网、自动驾驶车辆的数据分析等领域，开始使用随机森林算法。解释性和透明度的增强：提高模

型的可解释性，使得用户不仅能获得准确的预测结果，还能理解模型是如何做出这些预测的。处理大数据和流数据：改进算法以更有效地处理大规模数据集和实时流数据。与其他算法的结合：将随机森林与其他机器学习算法结合，形成更复杂的混合模型，以解决更加复杂的问题。

5. 结论

本文概述了随机森林及其在分类模型中的性能。随机森林是一个集成分类器，它包括多个分类器，用过去的数据集预测类标签值。随机森林构建速度快，预测速度更快。它们不需要任何交叉验证或完全可并行化。随机森林算法通常比单个分类器更准确。它可以在没有预处理的情况下处理数据，这意味着数据不需要重新缩放或转换。然而，作为一种广泛使用的算法，在提高分类精度方面值得进一步研究。

基金项目

校级项目资助(2022YSYB007)国家自然科学基金资助项目(62266046)。

参考文献

[1] Abdel-Hamid, O., Mohamed, A., Jiang, H. and Penn, G. (2012) Applying Convolutional Neural Networks Concepts to Hybrid NN-HMM Model for Speech Recognition. 2012 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Kyoto, 25-30 March 2012, 4277-4280. https://doi.org/10.1109/ICASSP.2012.6288864

[2] Adeen, I.M.N., Abdulazeez, A.M. and Zeebaree, D.Q. (2020) Systematic Review of Unsupervised Genomic Clustering Algorithms Techniques for High Dimensional Datasets. *Technology Reports of Kansai University*, **62**, 355-374.

[3] Zeebaree, D.Q., Haron, H., Abdulazeez, A.M. and Zebari, D.A. (2019) Machine Learning and Region Growing for Breast Cancer Segmentation. 2019 *International Conference on Advanced Science and Engineering* (*ICOASE*), Zakho-Duhok, 2-4 April 2019, 88-93. https://doi.org/10.1109/ICOASE.2019.8723832

[4] Sadiq, S.S., Abdulazeez, A.M. and Haron, H. (2020) Solving Multi-Objective Master Production Schedule Problem Using Memetic Algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, **18**, 938-945. https://doi.org/10.11591/ijeecs.v18.i2.pp938-945

[5] Abdulqader, D.M., Abdulazeez, A.M. and Zeebaree, D.Q. (2020) Machine Learning Supervised Algorithms of Gene Selection: A Review. *Technology Reports of Kansai University*, **62**, 233-243.

[6] Zebari, D.A., Haron, H., Zeebaree, D.Q. and Zain, A.M. (2019) A Simultaneous Approach for Compression and Encryption Techniques Using Deoxyribonucleic Acid. 2019 13*th International Conference on Software, Knowledge, Information Management and Applications* (*SKIMA*), Island of Ulkulhas, 26-28 August 2019, 1-6. https://doi.org/10.1109/SKIMA47702.2019.8982392

[7] Sadeeq, H. and Abdulazeez, A.M. (2018) Hardware Implementation of Firefly Optimization Algorithm Using FPGAs. 2018 *International Conference on Advanced Science and Engineering* (*ICOASE*), Duhok, 9-11 October 2018, 30-35. https://doi.org/10.1109/ICOASE.2018.8548822

[8] Najat, N. and Abdulazeez, A.M. (2017) Gene Clustering with Partition around Mediods Algorithm Based on Weighted and Normalized Mahalanobis Distance. 2017 *International Conference on Intelligent Informatics and Biomedical Sciences* (*ICIIBMS*), 24-26 November 2017, Okinawa, 140-145.

[9] Mienye, I.D., Sun, Y. and Wang, Z. (2019) Prediction Performance of Improved Decision Tree-Based Algorithms: A Review. *Procedia Manufacturing*, **35**, 698-703. https://doi.org/10.1016/j.promfg.2019.06.011

[10] Das, K., Behera, R.N. and Tech, B. (2007) A Survey on Machine Learning: Concept, Algorithms and Applications. *International Journal of Innovative Research in Computer and Communication Engineering*, **5**, 1301-1309.

[11] Schonlau, M. and Zou, R.Y. (2020) The Random Forest Algorithm for Statistical Learning. *The Stata Journal*: *Promoting Communications on Statistics and Stata*, **20**, 3-29. https://doi.org/10.1177/1536867X20909688

[12] Han, J., Fang, M., Ye, S., Chen, C., Wan, Q. and Qian, X. (2019) Using Decision Tree to Predict Response Rates of Consumer Satisfaction, Attitude, and Loyalty Surveys. *Sustainability*, **11**, Article 2306. https://doi.org/10.3390/su11082306

[13] Zhou, Z., Wang, Y., He, X. and Zhang, X. (2020) Optimization of Random Forests Algorithm Based on ReliefF-SA. *IOP Conference Series*: *Materials Science and Engineering*, **768**, Article ID: 072065. https://doi.org/10.1088/1757-899X/768/7/072065

[14] Kumar, G.K., Viswanath, P. and Rao, A.A. (2016) Ensemble of Randomized Soft Decision Trees for Robust Classifi-

cation. *Sādhanā*, **41**, 273-282. https://doi.org/10.1007/s12046-016-0465-z

[15] Li, Y., Jiang, Z.L., Yao, L., Wang, X., Yiu, S.M. and Huang, Z. (2019) Outsourced Privacy-Preserving C4.5 Decision Tree Algorithm over Horizontally and Vertically Partitioned Dataset among Multiple Parties. *Cluster Computing*, **22**, 1581-1593. https://doi.org/10.1007/s10586-017-1019-9

[16] Singh, S. and Giri, M. (2014) Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: A Survey. *International Journal of Advanced Information Science and Technology*, **3**, 47-52.

[17] Band, S.S., Janizadeh, S., Saha, S., Mukherjee, K., Bozchaloei, S.K., Cerdà, A., Shokri, M. and Mosavi, A. (2020) Evaluating the Efficiency of Different Regression, Decision Tree, and Bayesian Machine Learning Algorithms in Spatial Piping Erosion Susceptibility Using ALOS/PALSAR Data. *Land*, **9**, Article 346. https://doi.org/10.3390/land9100346

[18] Sarker, I.H., Colman, A., Han, J., Khan, A.I., Abushark, Y.B. and Salah, K. (2020) BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model. *Mobile Networks and Applications*, **25**, 1151-1161. https://doi.org/10.1007/s11036-019-01443-z

[19] Ozgode Yigin, B., Algin, O. and Saygili, G. (2020) Comparison of Morphometric Parameters in Prediction of Hydrocephalus Using Random Forests. *Computers in Biology and Medicine*, **116**, Article ID: 103547. https://doi.org/10.1016/j.compbiomed.2019.103547

[20] Denisko, D. and Hoffman, M.M. (2018) Classification and Interaction in Random Forests. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 1690-1692. https://doi.org/10.1073/pnas.1800256115

[21] Utkin, L.V., Kovalev, M.S. and Coolen, F.P.A. (2020) Imprecise Weighted Extensions of Random Forests for Classification and Regression. *Applied Soft Computing*, **92**, Article ID: 106324. https://doi.org/10.1016/j.asoc.2020.106324

[22] Demidova, L. and Ivkina, M. (2019) Defining the Ranges Boundaries of the Optimal Parameters Values for the Random Forest Classifier. 2019 1*st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency* (*SUMMA*), Lipetsk, 20-22 November 2019, 518-522. https://doi.org/10.1109/SUMMA48161.2019.8947569

[23] Kolhe, M.L., Tiwari, S., Trivedi, M.C. and Mishra, K.K. (2020). Advances in Data and Information Sciences: Proceedings of ICDIS 2019. Springer, Singapore. https://doi.org/10.1007/978-981-15-0694-9

[24] Gajowniczek, K., Grzegorczyk, I., Ząbkowski, T. and Bajaj, C. (2020) Weighted Random Forests to Improve Arrhythmia Classification. *Electronics*, **9**, Article 99. https://doi.org/10.3390/electronics9010099

[25] Zhang, B.Z., Qiao, X.M., Yang, H.M. and Zhou, Z.B. (2020). A Random Forest Classification Model for Transmission Line Image Processing. 2020 15*th International Conference on Computer Science & Education* (*ICCSE*), Delft, 18-22 August 2020, 613-617. https://doi.org/10.1109/ICCSE49874.2020.9201900

[26] Goel, E. and Abhilasha, E. (2017) Random Forest: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, **7**, 251-257. https://doi.org/10.23956/ijarcsse/V7I1/01113

[27] Imaizumi, T., Okada, A., Miyamoto, S., Sakaori, F., Yamamoto, Y. and Vichi, M. (2020) Advanced Studies in Classification and Data Science. Springer, Singapore. https://doi.org/10.1007/978-981-15-3311-2

[28] Darbanian, E., Rahbari, D., Ghanizadeh, R. and Nickray, M. (2020) Improving Response Time of Task Offloading by Random Forest, Extra-Trees and Adaboost Classifiers in Mobile Fog Computing. *Jordanian Journal of Computers and Information Technology*, **6**, 345-360. https://doi.org/10.5455/jjcit.71-1590557276

[29] Chaudhary, A., Kolhe, S. and Kamal, R. (2016) An Improved Random Forest Classifier for Multi-Class Classification. *Information Processing in Agriculture*, **3**, 215-222.

[30] Chen, S., Mulder, V.L., Martin, M.P., Walter, C., Lacoste, M., Richer-De-Forges, A.C., Saby, N.P.A., Loiseau, T., Hu, B. and Arrouays, D. (2019) Probability Mapping of Soil Thickness by Random Survival Forest at a National Scale. *Geoderma*, **344**, 184-194. https://doi.org/10.1016/j.geoderma.2019.03.016

[31] Bargarai, F.A.M., Abdulazeez, A.M., Tiryaki, V.M. and Zeebaree, D.Q. (2020) Management of Wireless Communication Systems Using Artificial Intelligence-Based Software Defined Radio. *International Journal of Interactive Mobile Technologies* (*IJIM*), **14**, 107-133. https://doi.org/10.3991/ijim.v14i13.14211

[32] Iwendi, C. and Jo, O. (2020) COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health*, **8**, Article 357. https://doi.org/10.3389/fpubh.2020.00357

[33] Zhang, F. and Yang, X. (2020) Improving Land Cover Classification in an Urbanized Coastal Area by Random Forests: The Role of Variable Selection. *Remote Sensing of Environment*, **251**, Article ID: 112105. https://doi.org/10.1016/j.rse.2020.112105

[34] Saenz-Cogollo, J.F. and Agelli, M. (2020) Investigating Feature Selection and Random Forests for Inter-Patient Heartbeat Classification. *Algorithms*, **13**, Article 75. https://doi.org/10.3390/a13040075

[35] Chai, Z. and Zhao, C. (2020) Multiclass Oblique Random Forests with Dual-Incremental Learning Capacity. *IEEE*

Transactions on Neural Networks and Learning Systems, **31**, 5192-5203.

[36] Azar, A.T., Gaber, T., Oliva, D., Ṭulbah, M.F. and Hassanien, A.E. (2020) Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020). Springer. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=6144671

[37] 汤圣君, 张韵婕, 李晓明, 等. 超体素随机森林与 LSTM 神经网络联合优化的室内点云高精度分类方法[J]. 武汉大学学报(信息科学版), 2023, 48(4): 525-533.

[38] 徐精诚, 陈学斌, 董燕灵, 等. 融合特征选择的随机森林 DDoS 攻击检测[J]. 计算机应用, 2023, 43(11): 3497-3503.

[39] Li, H., Lin, J., Lei, X. and Wei, T.X. (2022) Compressive Strength Prediction of Basalt Fiber Reinforced Concrete via Random Forest Algorithm. *Materials Today Communications*, **30**, Article ID: 103117. https://doi.org/10.1016/j.mtcomm.2021.103117

[40] Guo, Q., Zhang, J., Guo, S., *et al*. (2022) Urban Tree Classification Based on Object-Oriented Approach and Random Forest Algorithm Using Unmanned Aerial Vehicle (UAV) Multispectral Imagery. *Remote Sensing*, **14**, Article 3885. https://doi.org/10.3390/rs14163885