山东建筑大学

本科毕业论文

题　　目：基于随机森林的电力能耗预测

院　（部）：计算机科学与技术学院

专　　业：软件工程

班　　级：

姓　　名：

学　　号：

指导教师：

完成日期：2018 年 5 月 25 日

目 录

摘　要

“加快生态文明体制改革、建设美丽中国”是十九大报告的重要议题，而建筑领域绿色低碳发展是实现“美丽中国”目标的重要途径；房屋建筑在全寿命周期中消费了大量资源和能源，对生态文明建设产生巨大影响。

为了在改善住宅建筑的能源性能方面进行高效且有效的城市规划，我们需要对其影响特征有一个清晰的认识。以往我们对影响特征与能耗之间建模的研究仍存在些许空白和限制。建筑环境作为能源消耗的主要载体，通过科学、合理、准确地预测建筑运行能耗是设定合理、明确的节能目标，制定建筑节能政策、法规，以及开展建筑节能工作的重要前提条件，也是实现建筑能耗需求同其他经济领域协调、可持续发展的重要保障。在这种宏观趋势下,加强对建筑能耗的预测和分析,具有重要的理论和实际意义。

本文工作的主要概括如下：了解随机森林算法，针对固定建筑物每日的气象数据进行数据提取和处理，使用随机森林、KNN、支持向量机算法对已收集数据集进行训练，获得预测正确率的模型，之后使用其他多种算法训练出该方法下预测正确率最高的模型；然后针对两种模型分析得出所有所用算法中更适合于解决此类问题的算法；最终得出最佳的预测模型。

关键词：建筑能耗；数据处理；随机森林；KNN；支持向量机

Power consumption prediction based on Random Forest

ABSTRACT

"Accelerating the reform of the ecological civilization system and building a beautiful China" is an important topic in the nineteenth report. House buildings consume a lot of resources and energy in the whole life cycle, which has a great impact on the construction of ecological civilization.

Efficient and effective city planning in improving the energy performance of residential buildings requires a clear understanding of the influential features. Previous studies on modeling the relationships between influential features and the energy consumption have several gaps and limitations, such as the linear modeling methodology and insufficient consideration of particular features. Building environment as the main carrier of energy consumption, through the scientific, reasonable, accurately predict building energy consumption is to set reasonable and clear energy saving goal, building energy conservation policies and regulations, and carry out the work of building energy efficiency and important premise condition, is to realize building energy consumption demand with the rest of the economy to coordinate, the sustainable development important guarantee. Under this macro trend, it is of great theoretical and practical significance to strengthen the prediction and analysis of building energy consumption.

Understand random forest algorithm In this paper, the main work summarized as follows: in view of the fixed building daily meteorological data extraction and data processing, using the random forest algorithm to collect training data set, get the prediction accuracy of the model, then using a variety of other algorithm training the model with highest accuracy under the method; Then, based on the analysis of two models, all the algorithms used are more suitable for solving such problems. Finally, the best prediction model is obtained.

Key words: Building energy consumption;data processing;Random Forests;KNN;SVM

1 前 言

目前，发达国家建筑能耗占国家总能耗 30%--40%，其中建筑运行能耗占 30%以上。我国建筑能耗占社会商品总能耗的 27%左右，暖通空调能耗约又占建筑能耗的 50%～70%，建筑环境设备优化操作和管理可实现 20%-30%的能源节约。有研究指出，建筑消耗了约 40%的全球能源，其温室气体排放量约占全球温室气体排放的 33%。随着全球气候变化、能源短缺、大气污染日益严重，特别是近几年，全国大范围出现的雾霾天气和 PM2．5 值严重超标，给我国能源生产和消费方式提出了严峻挑战。

1.1 研究意义

“加快生态文明体制改革、建设美丽中国”是十九大报告的重要议题，而建筑领域绿色低碳发展是实现“美丽中国”目标的重要途径。

当前，房屋建筑在全寿命周期中消费了大量资源和能源，2015 年全国建筑能耗占全国能源消费总量的 20%，对生态文明建设产生巨大影响。同时，新时代下建筑还承载着人民对更加健康舒适的美好居住空间的需求。因此，加快推进建筑节能工作是实现绿色低碳发展的重要举措，而建筑能耗数据是科学推进建筑节能工作的基础，但当前建筑能耗权威数据缺失，相关研究缺乏系统性。电能能耗影响因素图如图 1.1 所示：

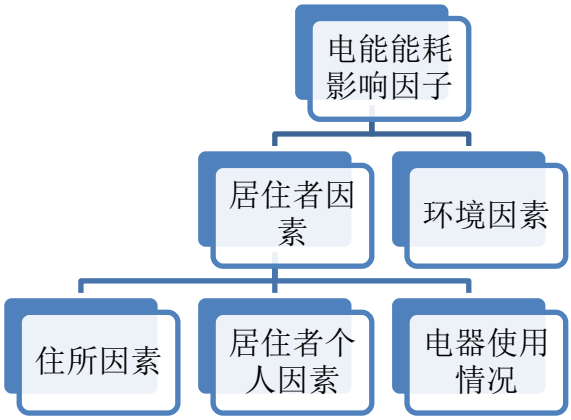

图 1.1

预测的意义在于建筑环境作为能源消耗的主要载体，通过科学、合理、准确地预测建筑运行能耗是设定合理、明确的节能目标，制定建筑节能政策、法规，以及开展建筑节能工作的重要前提条件，也是实现建筑能耗需求同其他经济领域协调、可持续发展的重要保障。得到预测结果之后，我们可以清楚的分析出主要的电能消耗特征和因素，从而做到在这些主要方面对能源消耗进行优化操作，针对像时间此类特殊的特

征因素可以进行效果最好的电能调度时间，从而节省成本、降低能耗，消费者和能源供应商双方都会因此减少支出。

1.2 本文主要内容

本文的主要内容是探究使用随进森林算法对建筑能耗进行预测，目的是为了找出能耗与建筑相关特征之间的联系，方便探究如何进行能源消耗的优化，从而设定合理、明确的节能目标。

第一章，大致介绍了论文题目的含义和在现实中存在的问题，总结了研究的实际意义与用途。

在第二章中，我们对随机森林算法的发展历程、实际应用的情况、其他学者进行的研究进行了分析，总结了随机森林算法的使用场景和方法。

在第三章中，对整个预测流程进行了总结整理，对数据处理的方法进行了细致的总结，对于特征重要性的总结，对基于机器学习模型的预测使用的算法的原理进行研究以及交叉验证和检验函数的使用。

在第四章中，对所要分析的数据集进行预处理，之后使用训练多种不同方法回归算法的预测模型并选择其中最优秀的预测模型。回归模型基于变量和趋势之间的关系的分析，以便做出关于连续变量的预测。

1.3 本文主要工作

本文工作的主要概括如下：针对固定建筑物每日的气象数据进行数据提取和处理，使用随机森林算法对已收集数据集进行训练，获得预测正确率的模型，之后使用其他多种算法训练出该方法下预测正确率最高的模型；然后针对两种模型分析得出所有所用算法中更适合于解决此类问题的算法；最终得出最佳的预测模型，流程图如图 1.2 所示。

具体的工作流程为:

（1）学习并总结随机森林算法使用案例和相关研究。

（2）研究随机森林算法原理。

（3）数据预处理。

（4）使用随机森林算法以及其他算法进行建模，得到回归模型。

（5）对得到的回归模型进行调参，得到准确率最高的模型。

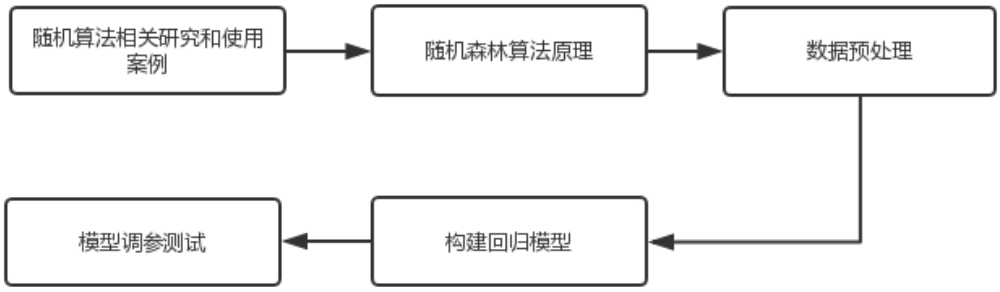

图 1.2

1.4 本章小结

本章主要介绍了当今建筑能耗在国家能源消耗中的重要性，进一步通过对建筑能耗进行预测并且进行分析得出如何进行建筑能耗的节能的最佳方法，这也是这项研究的意义所在。介绍了整篇文章的主要内容和流程，最后对本文中主要的工作进行了概括和总结。

2 相关研究

随机森林是一种比较新的机器学习模型。上世纪八十年代 Breiman 等人发明分类树的算法，通过反复二分数据进行分类或回归，计算量大大降低。2001 年 Breiman 把分类树组合成随机森林，即在变量（列）的使用和数据（行）的使用上进行随机化，生成很多分类树，再汇总分类树的结果。随机森林在运算量没有显著提高的前提下提高了预测精度。随机森林对多元共线性不敏感，结果对缺失数据和非平衡的数据比较稳健，可以很好地预测多达几千个解释变量的作用，被誉为当前最好的算法之一。

2.1 随机森林算法的相关研究

近年来，随机森林模型在界内的关注度与受欢迎程度有着显著的提升，这多半归功于它可以快速地被应用到几乎任何的数据科学问题中去，从而使人们能够高效快捷地获得第一组基准测试结果。在各种各样的问题中，随机森林一次又一次地展示出令人难以置信的强大，而与此同时它又方便实用。需要大家注意的是，随机森林方法有自己的局限性。

谈及随机森林算法的产生与发展，我们必须回溯到 20 世纪 80 年代。可以说，该算法是 Leo Breiman, Adele Cutler, Ho Tin Kam, Dietterich, Amit 和 Geman 这几位大师呕心沥血的共同结晶，他们中的每个人都对随机森林算法的早期发展作出了重要的贡献。Leo Breiman 和 Adele Cutler 最早提出了执行随机森里的关键算法，这一算法也成为了他们的专利之一。Amit, Gemen 和 Ho Tim Kam 各自独立地介绍了特征随即选择的思想，并且运用了 Breiman 的“套袋”思想构建了控制方差的决策树集合。在此之后，Deitterich 在模型中引入了随机节点优化的思想，对随机森林进行了进一步完善。

随机森林是一种多功能的机器学习算法，能够执行回归和分类的任务。同时，它也是一种数据降维手段，用于处理缺失值、异常值以及其他数据探索中的重要步骤，并取得了不错的成效。另外，它还担任了集成学习中的重要方法，在将几个低效模型整合为一个高效模型时大显身手。

在随机森林中会生成很多的决策树，并不像在 CART 模型里一样只生成唯一的树。当在基于某些属性对一个新的对象进行分类判别时，随机森林中的每一棵树都会给出自己的分类选择，并由此进行“投票”，森林整体的输出结果将会是票数最多的分类选项；而在回归问题中，随机森林的输出将会是所有决策树输出的平均值。

2.2 基于随机森林建模与应用相关研究

随着机器学习在技术和商业的发展，随机森林的在实际问题的应用场景已经很多，国内外针对不同问题的研究更是数不胜数。

国内基于随机森林建模与应用的研究：

一、王文寅和郭鹏波在《基于随机森林的股权众筹项目风险评估研究》[1] 中依托于大众融资的互联网金融，正适合目前我国的创业经济发展潮流，为中小微企业提供了一个崭新的融资方式. 但是，正在探索发展的众筹行业也面临诸多困境，我国的众筹体制及相应的法律法规还不完善，众筹平台在快速增长吸引了大量项目的同时，其风险和危机也与日俱增，违约、 非法集资、携款潜逃等问题频频出现在众筹项目中.随着互联网金融的发展，金融交易数据必将激增，信息更新速度将会越来越快，如何有效地利用现有数据改善当前的众筹风险问题至是个值待关注和探讨的问题。因此，对众筹项目进行风险评估与管理，为我国众筹投资者提供投资依据，具有重要的现实意义。

二、刘剑、曹美燕、高治军、许可的《基于随机森林的太阳能辐射预测模型》[2] 描述的是随着环境污染问题的加重，近年来太阳能作为最理想的清洁新能源备受关注。准确的掌握太阳能辐射情况对于太阳能的利用尤为关键。为了高效利用太阳能，准确的预测太阳能辐射情况极其重要。针对太阳能辐射的预测问题，研究了基于随机森林的太阳能辐射预测方法，根据影响太阳能辐射的因子建立了随机森林树型分类器，构建了一种基于随机森林的太阳能辐射量预测模型。准确有效的太阳能辐射预测方法对于合理开发太阳能资源、提高光伏发电的效率具有重要的实际意义。结果表明：采用随机森林模型预测效果较好，减少了均方根误差、提高了模型的预估精度，对复杂环境下的太阳能辐射量预测、光伏发电有效利用具有重要指导意义和应用前景。

三、张雯、刘爱利、齐威、丁浒的《基于随机森林的月貌面向对象分类》[3] 针对月球地貌分类的研究较少且没有相对简单的方法，提出将面向对象和随机森林相结合的方法对月貌进行分类。月貌能直观地反映月表特征及其目前的状态，也记录着月球形成和演变的历史信息。月表地貌学 的研究有利于加深认识和理解月表形貌，对月表年龄的估算、月壤厚度的反演、月球起源及演化历史的探索等具有重要意义，同时可为月球资源的开发利用提供基础。分类结果基本吻合月貌实际情况，总体分类精度达到 84.2%，Kappa 系数 0.71，和面向对象最近邻分类结果相比，总体精度提高了 9.4%，Kappa 系数提高了 0．07。

四、陈苏雨、方宇、胡定玉的《基于随机森林方法的地铁车门故障诊断》[4] 针对

现有地铁车门故障诊断方法存在的诊断速度慢以及大量故障检修数据未得到合理利用等问题，提出一种基于信息增益率的随机森林故障诊断方法。针对现有地铁车门故障诊断方法存在的诊断速度慢以及大量故障检修数据未得到合理利用等问题，提出一种基于信息增益率的随机森林故障诊断方法。该方法将地铁车门历史故障数据集转化成决策表，通过 Bootstrap 重抽样，建立多棵基于信息增益率的决策树，形成随机森林故障诊断模型，实现地铁车门故障的快速诊断。

五、魏金太、高穹的《基于信息增益和随机森林分类器的入侵检测系统研究》[5] 目的是解决互联网以及其他网络上的安全数据通信总是会受到入侵以及滥用等威胁。目前，许多误用检测系统无法检测未知攻击，而异常检测系统虽然能够精确检测未知攻击，但由于入侵检测固有的特性，入侵事件与正常事件类间存在极大的不平衡性，这导致很难利用机器学习的方法高效地进行入侵行为检测。为此提出了一种基于信息增益和随机森林分类器的入侵检测系统.为了解决类之间的不平衡性，对训练数据集应用了合成少数过采样算法。且随着故障数据的增加，其故障诊断模型可以自动更新完善。通过地铁车门实际故障数据，验证了该方法的有效性。同时，通过对随机森林模型中决策树的数目讨论分析，确定了该方法模型的最优设计结构。

六、吕杰、郝宁燕、李崇贵、史晓亮、李宗泽的《利用随机森林和纹理特征的森林类型识别》[6] 结合遥感判读样地、植被指数、纹理信息以及地形因子等多源数据，构建最小距离分类模型、支持向量机分类模型和随机森林分类模型，对黑龙江凉水自然保护区森林优势树种进行分类。结果表明，基于随机森林模型的分类结果总精度和 Kappa 系数分别为 81．0l％和 0．76，较支持向量机分类方法有明显提高。该研究为提高我国高分辨率数据的自给率和森林资源的有效管理提供了一定的参考价值。

七、王利民、刘佳、杨玲波、杨福刚、富长虹的《随机森林方法在玉米-大豆精细识别中的应用》[7] 研究基于遥感影像的作物精确识别技术方法, 对获取作物分布信息具有重要意义。结果表明, MLC、SVM、RFC 的总体分类精度分别为 91.68%、91.49%、94.32%, Kappa 系数分别为 0.87、0.87、0.91, RFC 方法作物识别精度比 MLC 和 SVM 分类显著提升。

通过随机森林算法及其思想，解决了许多实现起来困难的现实问题，这说明随机森林算法针对某些特定问题有着很好的性能，也说明了随机森林适合解决某一类特殊问题。

国外基于随机森林建模与应用的研究：

一、对生长裙带菜的国家的鉴别分析[8] :利用随机森林、分类和回归树对可见光近红外光谱进行初步比较。BSE 病毒爆发后，与蛤蜊、裙带菜相关的问题引起了人们的关注。我们开发了一种科学的方法，利用软独立建模方法来区分生产裙带菜的生产国家。然而，得到的 SIMCA 模型的验证并不是很好。因此，我们尝试了一种非参数分类方法，即基于分类和回归树(CART)的随机森林。我们发现，随机森林可能是一种有效的可见和近红外光谱的分类方法。我们建议使用 CART 对分类条件的基础进行初步解释。

二、基于点或多边形的训练数据对湿地随机森林分类精度的影响[9]。湿地在空间和时间上都是动态的对不同的生态系统提供服务。这种现状使得我们很难保持湿地地图的准确性、效率和一致性。此外，点参考数据可能不代表某一地区的主要土地覆盖类型。在研究中，以三种方式实现随机森林分类器的方法，在不同的生态系统中对两个研究地点进行土地覆盖分类的训练:场和照片解释点;围绕着点的固定窗;和相交点的图像对象。还进行了额外的评估，以确定关键的输入变量。结论是，图像对象区域训练方法是最精确的，最重要的变量包括:复合地形指数、夏季绿色和蓝色波段，以及来自激光雷达点云数据的网格统计数据，特别是与返回高度相关的数据。

三、利用时间序列监测缅甸的水稻农业[10]。由于云层覆盖、光学传感器分辨率的限制、空间和时间的动态变化以及缺乏系统性的雷达，限制了大面积水稻农业的评估和监测。在适度空间分辨率下，密集时间序列的开放数据获取提供了监测农业的新机会。南亚和东南亚的大米对粮食安全至关重要，而且大多是在高云层覆盖的雨季期间生长。在本研究应用中，利用时间序列干涉宽图像绘制了缅甸各地的水稻种植面积、作物日历、洪水和种植强度。使用最新的土地覆盖地图经过整合和分类后使用随机森林算法。然后对密集的数据进行时间序列的物候分析，以评估整个缅甸的水稻信息。分析表明，收获的水稻面积为 6,652,111 公顷，与政府普查统计数据一致(R2 = 0.78)。结果显示，在多云地区，评估和监测水稻产量的能力很强。在缅甸这样人口众多的国家，政府依赖稻米生产，更加健全和透明的监测和评估工具可以帮助政府更好的决策。这些结果表明，系统和开放获取合成孔径雷达(SAR)可以帮助提供食品安全措施和监测、报告和验证程序所需的信息。

2.3 基于随机森林的建筑能耗预测相关研究

在建筑能耗分析领域，建立建筑能耗模型是深入分析其能耗特点的重要手段，而

且能够为建筑节能提供优化节能措施。因此，建立可靠准确的建筑能耗模型是建筑节能研究中的重要任务。机器学习方法近年来在建筑能耗分析领域的应用越来越广泛。机器学习方法由于其先进的数据分析能力，可以用于分析多变量之间的复杂模式以及交互作用，在调整完机器模型中算法参数后，能够达到计算速度快的特点，非常适合应用于建筑能耗模型的建立和分析。所建立的机器学习能耗模型，可进一步用来进行建筑能耗的不确定性和敏感性分析、贝叶斯分析和最优化计算等方面。

基于机器学习对建筑能耗进行分析的研究国内外学者已经对基于机器学习方法的建筑能耗模型进行了很多研究：利用多元自适应回归样条法得到了英国伦敦中学建筑的能耗模型；利用多元线性回归和分类回归树模型探讨了意大利北部 80 所学校的能耗特点；根据支持向量机算法，研究了建筑中遮阳控制的相关计算；基于高斯过程和主成分回归等机器学习算法，分析了美国佐治亚理工学院校园建筑的能耗特点。

2.4 本章小结

在机器学习中，随机森林算法的使用场景非常广泛，技术本身是一种多功能的机器学习算法，能够执行回归和分类的任务。可以快速地被应用到几乎任何的数据科学问题中去，从而使人们能够高效快捷地获得第一组基准测试结果。在实际问题的解决上，完美或者很好的完成了诸如基于随机森林的股权众筹项目风险评估研究、基于随机森林的太阳能辐射预测模型、基于随机森林的月貌面向对象分类等困难问题，这也展示出了随机森林算法的优势所在。在本章的最后总结了国内外基于机器学习对建筑能耗进行分析的研究。

3 基于机器学习的预测

基于机器学习进行预测的流程分别是：数据预处理、特征重要性分析、机器学习模型、评价函数和交叉验证。这其中的每一步都对结果的优劣非常重要，所以在第三章详细的介绍了五大步骤中每个步骤的方法和所用模型的原理及优缺点。

3.1 数据预处理

现实世界中数据大体上都是不完整，不一致的脏数据，无法直接进行数据挖掘，或挖掘结果差强人意。为了提高数据挖掘的质量产生了数据预处理技术。数据预处理有多种方法：数据清理，数据集成，数据变换，数据归约等。这些数据处理技术在数据挖掘之前使用，大大提高了数据挖掘模式的质量，降低实际挖掘所需要的时间。

一、数据清洗

1、缺失值处理

处理缺失值分为三类：删除记录、数据补差和不处理。

数据补插方法：

1. 补插均值/中位数/众数 2. 使用固定值 3. 最近邻补插 4. 回归方法 5. 插值法：拉格朗日插值法、牛顿插值法、Hermit 插值法、分段插值、样条插值。

2、异常值处理

1、删除有异常值的记录：删除含有异常值的数据行。

2、视为缺失值。

3、平均值修正 ：使用平均值替换异常值。

4、不处理 ：保持原数。

要仔细分析异常值的原因，再决定取舍。

二、数据集成

将多个数据源放在一个统一的数据仓库中。

1、实体识别、同名异义、异名同义、单位不统一。

2、冗余属性识、同一属性多次出现、同一属性命名不一致。

三、数据变换

对数据进行规范化处理

1、简单函数变换

原始数据进行数学函数变换，平方、开方、取对数、差分运算。用来将不

具有正太分布的数据变换成具有正太性的数据。

时间序列分析中，对数变换或者差分运算可以将非平稳序列转换为平稳序列。

2、规范化

消除指标间规范化

（1） 最小-最大规范化：又称离差标准化，可以对原始数据进行线性变换。

（2） 零-均值规范化：属性 A 的值是基于 A 的平均值与标准差规范化。

（3） 小数指定标规范化：通过移动属性值的小数点位置进行规范化，通俗的说就是将属性值除以 10 的 j 次幂。

3、连续属性性离散化

将连续属性变为分类属性，即连续属性离散化。数据离散化本质上通过断点集合将连续的属性空间划分为若干区，最后用不同的符号或者整数值代表落在每个子区间中的数据。离散化涉及两个子任务：确定分类以及如何将连续属性值映射到这些分类值。

四、数据规约

降低无效的、错误的数据对建模的影响，提高建模的准确性。少量且代表性的数据将大幅缩减数据挖掘所需时间。降低存储数据成本。

1、属性规约

合并属性、逐步向前选择、逐步向后删除、决策树归纳、主成分分析

2、数值规约

通过选择替代的、较小的数据来减少数据量，包含有参数方法和无参数方法两类；有参数方法使用模型评估数据，不需要存放真实数据，只需要存放参数，例如回归、对数线性模型。无参数需要数据，例如直方图、聚类、抽样。

3.2 特征重要性

特征重要性，特征选择(排序)对于数据科学家、机器学习从业者来说非常重要。好的特征选择能够提升模型的性能，更能帮助我们理解数据的特点、底层结构，这对进一步改善模型、算法都有着重要作用。

特征选择主要有两个功能：

1、减少特征数量、降维，使模型泛化能力更强，减少过拟合。

2、增强对特征和特征值之间的理解。。

特征选择的方法：

1、去掉取值变化小的特征：假设某特征的特征值只有 0 和 1，并且在所有输入样本中，95%的实例的该特征取值都是 1，那就可以认为这个特征作用不大。如果 100%都是 1，那这个特征就没意义了。

2、单变量特征选择：单变量特征选择能够对每一个特征进行测试，衡量该特征和响应变量之间的关系，根据得分扔掉不好的特征。对于回归和分类问题可以采用卡方检验等方式对特征进行测试。

主要方法有：Pearson 相关系数、互信息和最大信息系数、距离相关系数、基于学习模型的特征排序。

3、先行模型和正则化：正则化模型、Lasso、Ridge 回归。

4、随机森林：

平均不纯度减少：随机森林由多个决策树构成。决策树中的每一个节点都是关于某个特征的条件，为的是将数据集按照不同的响应变量一分为二。利用不纯度可以确定节点（最优条件），对于分类问题，通常采用基尼不纯度或者信息增益，对于回归问题，通常采用的是方差或者最小二乘拟合。当训练决策树的时候，可以计算出每个特征减少了多少树的不纯度。对于一个决策树森林来说，可以算出每个特征平均减少了多少不纯度，并把它平均减少的不纯度作为特征选择的值。

平均精确率减少：主要思路是打乱每个特征的特征值顺序，并且度量顺序变动对模型的精确率的影响。很明显，对于不重要的变量来说，打乱顺序对模型的精确率影响不会太大，但是对于重要的变量来说，打乱顺序就会降低模型的精确率。

3.3 机器学习模型

机器学习中拥有很多模型，本小节主要介绍随机森林模型、K 临近算法模型。

3.3.1 随机森林模型

随机森林模型的基础是决策树，关于决策树的分类如图 3.1 所示。

一、 决策树

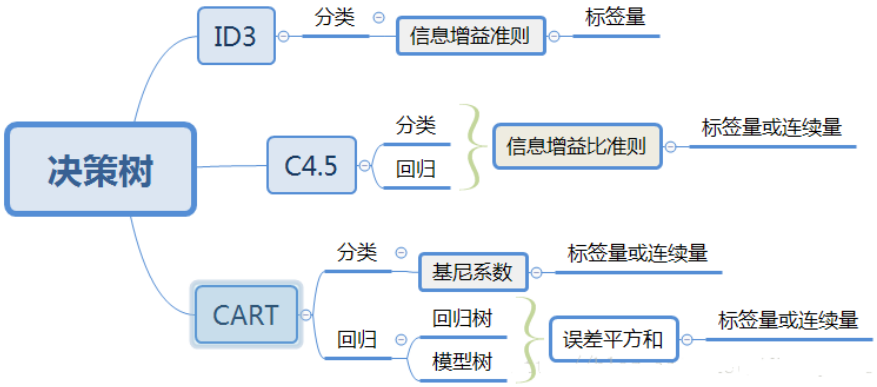

图 3.1

1、定义

决策树是一种监督学习算法。它适用于类别和连续输入（特征）和输出（预测）变量。基于树的方法把特征空间划分成一系列矩形，然后给每一个矩形安置一个简单的模型（像一个常数）。从概念上来讲，它们是简单且有效的。分类决策树模型是一种描述对实例进行分类的树形结构。决策树由结点和有向边组成。结点有两种类型：内部节点和叶节点，内部节点表示一个特征或属性，叶节点表示一个类。

分类的时候，从根节点开始，对实例的某一个特征进行测试，根据测试结果，将实例分配到其子结点；此时，每一个子结点对应着该特征的一个取值。如此递归向下移动，直至达到叶结点，最后将实例分配到叶结点的类中。

2、决策树与 if-then 规则

现在我们可以更抽象一些。决策树可以看成一个 if-then 规则的集合：由决策树的根结点到叶结点的每一条路径构建一条规则；路径上的内部结点的特征对应着规则的条件，而叶结点对应着分类的结论。决策树的路径和其对应的 if-then 规则集合是等效的，它们具有一个重要的性质：互斥并且完备。这里的意思是说：每一个实例都被一条路径或一条规则所覆盖，而且只被一条规则所覆盖。

3、决策树与条件概率分布

决策树还是给定特征条件下类的条件概率分布的一种表示。该条件分布定义在特征空间的划分（partition）上，特征空间被划分为互不相交的单元（cell），每个单元定义一个类的概率分布就构成了一个条件概率分布。决策树的一条

路径对应于划分中的一个单元。决策树所表示的条件概率分布由各个单元给定条件下类的条件概率分布组成。给定实例的特征 X，一定落入某个划分，决策树选取该划分里最大概率的类作为结果输出，如图 3.2 所示：

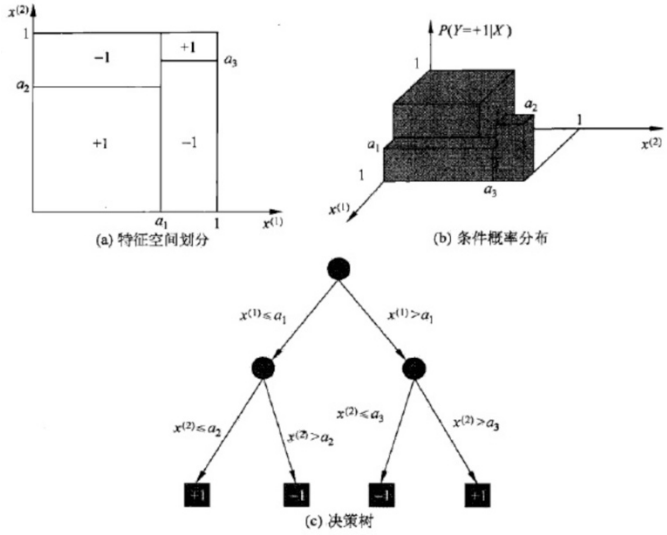

图 3.2

4、决策树的学习

决策树学习算法包含特征选择、决策树的生成与剪枝过程。决策树的学习算法通常是递归地选择最优特征，并用最优特征对数据集进行分割。开始时，构建根结点，选择最优特征，该特征有几种值就分割为几个子集，每个子集分别递归调用此方法，返回结点，返回的结点就是上一层的子结点。直到所有特征都已经用完，或者数据集只有一维特征为止。

二、 决策树的生成

此处主要介绍两种决策树学习的算法：ID3 和 C4.5。

1、ID3 算法由 Ross Quinlan 发明，建立在“奥卡姆剃刀”的基础上：越是小型的决策树越优于大的决策树（be simple 简单理论）。ID3 算法中根据信息增益评估和选择特征，每次选择信息增益最大的特征作为判断模块建立子结点。ID3 算法可用于划分标称型数据集，没有剪枝的过程，为了去除过度数据匹配的问题，可通过裁剪合并相邻的无法产生大量信息增益的叶子节点（例如设置信息增益阀值）。使用信息增益的话其实是有一个缺点，那就是它偏向于具有大量值的属性。就是说在训练集中，某个属性所取的不同值的个数越多，那么越有可能拿它来作为分裂属性，而这样做有时候是

没有意义的，另外 ID3 不能处理连续分布的数据特征，于是就有了 C4.5 算法。

2、C4.5 算法用信息增益率来选择属性，继承了 ID3 算法的优点。并在以下几方面对 ID3 算法进行了改进：

（1）克服了用信息增益选择属性时偏向选择取值多的属性的不足

（2）在树构造过程中进行剪枝

（3）能够完成对连续属性的离散化处理

（4）能够对不完整数据进行处理

C4.5 算法产生的分类规则易于理解、准确率较高；但效率低，因树构造过程中，需要对数据集进行多次的顺序扫描和排序。也是因为必须多次数据集扫描，C4.5 只适合于能够驻留于内存的数据集。在实现过程中，C4.5 算法在结构与递归上与 ID3 完全相同，区别只在于选取决决策特征时的决策依据不同，二者都有贪心性质：即通过局部最优构造全局最优。

三、 CART 算法

CART（Classification And Regression Trees，分类回归树）算法，CART 是一个独立于其他经典决策树算法的算法，所以导致 CART 相对来说较为复杂。因为它不仅仅可以作为分类树，还可以作为回归树。采用的是 Gini 指数（选 Gini 指数最小的特征 s）作为分裂标准,同时它也是包含后剪枝操作。ID3 算法和 C4.5 算法虽然在对训练样本集的学习中可以尽可能多地挖掘信息，但其生成的决策树分支较大，规模较大。为了简化决策树的规模，提高生成决策树的效率，就出现了根据 GINI 系数来选择测试属性的决策树算法 CART。

基尼指数：表示在样本集合中一个随机选中的样本被分错的概率。

基尼指数（基尼不纯度）= 样本被选中的概率 * 样本被分错的概率

$$\text{Gini(p)} = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} {p_k}^2$$

CART 分类生成算法

输入：训练数据集 D，停止计算的条件；

输出：CART 决策树；

根据训练数据集，从根结点开始，递归地对每个结点进行以下操作，构建二叉决策树：

1、设结点的训练数据集为 D，计算现有特征对该数据集的基尼指数。此时，对每一个特征 A，对其可能取的每个值 a，根据样本点对 A=a 的测试为“是”或“否”将 D 分割成 D1 和 D2 两部分，之后计算基尼指数。

2、在所有可能的特征 A 以及它们所有可能的切分点 a 中，选择基尼指数最小的特征及其对应的切分点作为最有特征与最优切分点。依照最优特征与最优切分点，从现有结点生成两个子节点，将训练数据集按照特征分配到两个子节点中去。

3、对两个子结点递归地调用两个子结点，将训练数据集按特征分配到两个子节点中去。

4、生成 CART 决策树。

四、 剪枝

在决策树学习中将已生成的树进行简化的过程称为剪枝。决策树的剪枝往往通过极小化决策树的损失函数或代价函数来实现。实际上剪枝的过程就是一个动态规划的过程：从叶结点开始，自底向上地对内部结点计算预测误差以及剪枝后的预测误差，如果两者的预测误差是相等或者剪枝后预测误差更小，当然是剪掉的好。但是如果剪枝后的预测误差更大，那就不要剪了。剪枝后，原内部结点会变成新的叶结点，其决策类别由多数表决法决定。不断重复这个过程往上剪枝，直到预测误差最小为止。

五、 信息熵

1、信息增益（ID3）

信息增益：以某特征划分数据集前后的熵的差值。

在熵的理解那部分提到了，熵可以表示样本集合的不确定性，熵越大，样本的不确定性就越大。因此可以使用划分前后集合熵的差值来衡量使用当前特征对于样本集合 D 划分效果的好坏。

使用某个特征 A 划分数据集 D，计算划分后的数据子集的熵 entroy(后)。

信息增益 = entroy(前) - entroy(后)

$$g(D,A) = H(D) - H(D|A)$$

2、信息增益比（C4.5 算法）

信息增益比 = 惩罚参数 * 信息增益

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)}$$

信息增益比本质是在信息增益的基础之上乘上一个惩罚参数。特征个数较多时，惩罚参数较小；特征个数较少时，惩罚参数较大。

六、随机森林算法

1、概念

随机森林是一种多功能的机器学习算法，能够执行回归和分类的任务。同时，它也是一种数据降维手段，用于处理缺失值、异常值以及其他数据探索中的重要步骤，并取得了不错的成效。另外，它还担任了集成学习中的重要方法，在将几个低效模型整合为一个高效模型时大显身手。

在随机森林中，我们将生成很多的决策树，并不像在 CART 模型里一样只生成唯一的树。当在基于某些属性对一个新的对象进行分类判别时，随机森林中的每一棵树都会给出自己的分类选择，并由此进行“投票”，森林整体的输出结果将会是票数最多的分类选项；而在回归问题中，随机森林的输出将会是所有决策树输出的平均值。

在随机森林算法中，用到了许多个决策树。决策树是一种监督学习算法，它适用于类别和连续输入（特征）和输出（预测）变量。基于树的方法把特征空间划分成一系列矩形，然后给每一个矩形安置一个简单的模型（像一个常数）。从概念上来讲，它们是简单且有效的。首先我们通过一个例子来理解决策树。然后用一种正规分析方法来分析创建决策树的过程。分类和回归树（简称 CART）是 Leo Breiman 引入的术语，指用来解决分类或回归预测建模问题的决策树算法。

2、袋装（Bootstrap Aggregating——Bagginng）

在统计学中，Bootstrap 是依靠替换随机采样的任意试验或度量。我们从上文可以看见，决策树会受到高方差的困扰。这意味着如果我们把训练数据随机分成两部分，并且给二者都安置一个决策树，我们得到的结果可能就会相当不同。Bootstrap 聚集，或者叫做袋装，是减少统计学习方法的方差的通用过程。

这里有一个问题，即我们不能获取多个训练数据集。相反，我们可以通过从（单一）训练数据集提取重复样本进行自助法（bootstrap）操作。在这种方法中，我们生成了 B 个不同的自助训练数据集。我们随后在第 b 个自助训练数据集得到了一个预测结果$\hat{f}^{*b}(x)$，从而获得一个聚集预测（aggregate prediction）。

$$\hat{f}_{bag} = \begin{cases} \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) & \text{for Regression Problems} \\ \\ arg\max_{b=1\ldots B} \hat{f}^{*b}(x) & \text{for Classification Problems} \end{cases}$$

这就叫做袋装（bagging）。注意，聚集（aggregating）在回归和分类问题中可能有不同的均值。当平均预测值在回归问题中的效果很好时，我们将会需要使用多数票决（majority vote）：由于分类问题中的聚集机制，整体预测就是在 B 个预测值中最常出现的那个主要类别。

3、Out-of-Bag（OOB）误差

Bagging 方法最大的优势是我们可以不通过交叉验证而求得测试误差。回想一下，Bagging 方法的精髓是多棵树可以重复地拟合观察样本的自助子集。平均而言，每一个袋装树可以利用 2/3 的观察样本。而剩下的 1/3 观察样本就可以称为 out-of-bag (OOB) 观察样本，它们并不会拟合一一棵给定袋装树。我们可以使用每一棵树的 OOB 观察样本而计算第 i 个观察样本的预测值，这将会导致大约有 B/3 的预测值可以预测第 i 个观察样本。现在我们可以使用和 Bagging（平均回归和大多数投票分类）类似的聚集技术，我们能获得第 i 个观察样本的单一预测值。我们可以用这种方式获得 n 个观察样本的 OOB 预测，因此总体的 OOB MSE（回归问题）和分类误差率（分类问题）就能计算出来。OOB 误差结果是 Bagging 模型测试误差的有效估计，因为每一个样本的预测值都是仅仅使用不会进行拟合训练模型的样本。

4、特征重要性度量

通过使用单一树，Bagging 通常会提升预测的精确度。但是，解释最终的模型可能很困难。当我们袋装大量的树时，就不再可能使用单一的树表征最终的统计学习流程，因此，Bagging 是以牺牲阐释性能力为代价来提升预测精确度的。有趣的是，一个人可使用 RSS（用于 bagging 回归树）或者基尼指数（用于 bagging 分类树）得到每一个预测器的整体总结。在 bagging 回归树的情况中，我们可以记录由于所有的 B 树上平均的给定预测分子分裂而造成的 RSS 减少的所有数量。一个大的值表示一个重要的预测器。相似地，在 bagging 分类树的情况下，我们可以添加由于所有的 B 树上平均的给定预测分子分裂而造成的基尼系数降低的所有数量。一旦训练完成，sklearn 模块的不同袋装树（bagged tree）学习方法可直接访问特征的重要性数据作为属性。

5、随机森林算法运行流程

在随机森林中，每一个决策树“种植”和“生长”的规则如图 3.3 所示：

1、假设我们设定训练集中的样本个数为 N，然后通过有重置的重复多次抽样来获得这 N 个样本，这样的抽样结果将作为我们生成决策树的训练集；

图 3.3

2、如果有 M 个输入变量，每个节点都将随机选择 m(m<M)个特定的变量，然后运用这 m 个变量来确定最佳的分裂点。在决策树的生成过程中，m 的值是保持不变的； 3.每棵决策树都最大可能地进行生长而不进行剪枝； 4.通过对所有的决策树进行加总来预测新的数据（在分类时采用多数投票，在回归时采用平均）。

随机森林的优点与缺点

优点：

1、正如上文所述，随机森林算法能解决分类与回归两种类型的问题，并在这两个方面都有相当好的估计表现。

2、随机森林对于高维数据集的处理能力令人兴奋，它可以处理成千上万的输入变量，并确定最重要的变量，因此被认为是一个不错的降维方法。此外，该模型能够输出变量的重要性程度，这是一个非常便利的功能。

3、在对缺失数据进行估计时，随机森林是一个十分有效的方法。就算存在大量的数据缺失，随机森林也能较好地保持精确性。

4、当存在分类不平衡的情况时，随机森林能够提供平衡数据集误差的有效方法

5、模型的上述性能可以被扩展运用到未标记的数据集中，用于引导无监督聚类、数据透视和异常检测。

6、随机森林算法中包含了对输入数据的重复自抽样过程，即所谓的 bootstrap 抽样。这样一来，数据集中大约三分之一将没有用于模型的训练而是用于测试，这样的数据被称为 out of bag samples，通过这些样本估计的误差被称为 out of bag error。研究表明，这种 out of bag 方法的与测试集规模同训练集一致的估计方法有着相同的精确程度，因此在随机森林中我们无需再对测试集进行另外的设置。

缺点：

1、随机森林在解决回归问题时并没有像它在分类中表现的那么好，这是因为它并不能给出一个连续型的输出。当进行回归时，随机森林不能够作出超越训练集数据范围的预测，这可能导致在对某些还有特定噪声的数据进行建模时出现过度拟合。

2、对于许多统计建模者来说，随机森林给人的感觉像是一个黑盒子——你几乎无法控制模型内部的运行，只能在不同的参数和随机种子之间进行尝试。

3.3.2KNN 模型

原理：KNN（K-Nearest Neighbor）：存在一个样本数据集合，也称为训练样本集，并且样本集中每个数据都存在标签，即我们知道样本集中每一数据与所属分类对应的关系。输入没有标签的数据后，将新数据中的每个特征与样本集中数据对应的特征进行比较，提取出样本集中特征最相似数据（最近邻）的分类标签。一般来说，我们只选择样本数据集中前 k 个最相似的数据，这就是 k 近邻算法中 k 的出处，通常 k 是不大于 20 的整数。最后选择 k 个最相似数据中出现次数最多的分类作为新数据的分类。

说明：KNN 没有显示的训练过程，它是“懒惰学习”的代表，它在训练阶段只是把数据保存下来，训练时间开销为 0，等收到测试样本后进行处理。

在 KNN 中，通过计算对象间距离来作为各个对象之间的非相似性指标，避免了对象之间的匹配问题，在这里距离一般使用欧氏距离或曼哈顿距离：

$$\text{欧式距离：}\ d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}\ ,\qquad \text{曼哈顿距离：}\ d(x,y)=\sqrt{\sum_{k=1}^{n}|x_k-y_k|}$$

同时，KNN 通过依据 k 个对象中占优的类别进行决策，而不是单一的对象类别决策。这两点就是 KNN 算法的优势。

接下来对 KNN 算法的思想总结一下：就是在训练集中数据和标签已知的情况下，

输入测试数据，将测试数据的特征与训练集中对应的特征进行相互比较，找到训练集中与之最为相似的前 K 个数据，则该测试数据对应的类别就是 K 个数据中出现次数最多的那个分类，其算法的描述为：

1、计算测试数据与各个训练数据之间的距离；

2、按照距离的递增关系进行排序；

3、选取距离最小的 K 个点；

4、确定前 K 个点所在类别的出现频率；

5、返回前 K 个点中出现频率最高的类别作为测试数据的预测分类。

KNN 算法的优点：

1、简单、有效。

2、重新训练的代价较低（类别体系的变化和训练集的变化，在 Web 环境和电子商务应用中是很常见的）。

3、计算时间和空间线性于训练集的规模（在一些场合不算太大）。

4、由于 KNN 方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN 方法较其他方法更为适合。

5、该算法比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。

KNN 算法缺点：

1、KNN 算法是懒散学习方法（lazy learning,基本上不学习），一些积极学习的算法要快很多。

2、类别评分不是规格化的（不像概率评分）。

3、输出的可解释性不强，例如决策树的可解释性较强。

4、该算法在分类时有个主要的不足是，当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的 K 个邻居中大容量类的样本占多数。该算法只计算“最近的”邻居样本，某一类的样本数量很大，那么或者这类样本并不接近目标样本，或者这类样本很靠近目标样本。无论怎样，数量并不能影响运行结果。可以采用权值的方法（和该样本距离小的邻居权值大）来改进。

5、计算量较大。目前常用的解决方法是事先对已知样本点进行剪辑，事先去除对分类

作用不大的样本。

3.3.3SVM 模型

1、定义

在机器学习中，支持向量机（SVM，Support Vector Machine）是与相关的学习算法有关的监督学习模型，可以分析数据，识别模式，用于分类和回归分析。给定一组训练样本，每个标记为属于两类，一个 SVM 训练算法建立了一个模型，分配新的实例为一类或其他类，使其成为非概率二元线性分类。一个 SVM 模型的例子，如在空间中的点，映射，使得所述不同的类别的例子是由一个明显的差距是尽可能宽划分的表示，如图 3.4 所示。新的实施例则映射到相同的空间中，并预测基于它们落在所述间隙侧上属于一个类别。

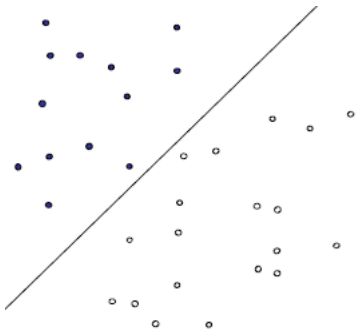

图 3.4

除了进行线性分类，支持向量机可以使用所谓的核技巧，它们的输入隐含映射成高维特征空间中有效地进行非线性分类。

支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面，分隔超平面使两个平行超平面的距离最大化。假定平行超平面间的距离或差距越大，分类器的总误差越小。

2、线性分类

这里我们考虑的是一个两类的分类问题，数据点用 x 来表示，这是一个 n 维向量，w^T中的 T 代表转置，而类别用 y 来表示，可以取 1 或者-1，分别代表两个不同的类。

一个线性分类器的学习目标就是要在 n 维的数据空间中找到一个分类超平面，其方程可以表示为：w^Tx+b=0。

3、函数间隔与几何间隔

在分离超平面固定为w^Tx+b=0 的时候，|w^Tx+b|表示点 x 到超平面的距离。通过观察w^Tx+b 和 y 是否同号，我们判断分类是否正确，这些知识我们在感知机模型里都有讲到。这里我们引入函数间隔的概念，定义函数间隔γ′为：

$$\gamma' = y(w^T x+b)$$

可以看到，它就是感知机模型里面的误分类点到超平面距离的分子。对于训练集中 m 个样本点对应的 m 个函数间隔的最小值，就是整个训练集的函数间隔。

函数间隔并不能正常反应点到超平面的距离，在感知机模型里我们也提到，当分子成比例的增长时，分母也是成倍增长。

几何间隔才是点到超平面的真正距离，感知机模型里用到的距离就是几何距离。

3.4 评价函数

1、均方根误差（RMSE）

RMSE 是预测值与真实值的误差平方根的均值。

均方根误差 RMSE(root-mean-square error)， 均方根误差亦称标准误差,它是观测值与真值偏差的平方与观测次数比值的平方根。均方根误差是用来衡量观测值同真值之间的偏差。标准误差对一组测量中的特大或特小误差反映非常敏感，所以，标准误差能够很好地反映出测量的精密度。可用标准误差作为评定这一测量过程精度的标准。计算公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{n}^{i=1}(Y_i - \hat{Y}_i)^2}{n}}$$

优点：标准化平均方差对均方差进行了标准化改进，通过计算拟评估模型与以均值为基础的模型之间准确性的比率，标准化平均方差取值范围通常为 0～1，比率越小，说明模型越优于以均值进行预测的策略，NMSE 的值大于 1，意味着模型预测还不如简单地把所有观测值的平均值作为预测值，

缺点：但是通过这个指标很难估计预测值和观测值的差距，因为它的单位也和原变量不一样了，综合各个指标的优缺点，我们使用三个指标对模型进行评估。

2、R 方（R^2）

R^2 方法是将预测值跟只使用均值的情况下相比，看能好多少。其区间通常在（0,1）之间。0 表示还不如什么都不预测，直接取均值的情况，而 1 表示所有预测跟真实结果完美匹配的情况。

计算公式如下：

$$R^2 = 1 - \frac{\sum_n^{i=1}(Y_i - \hat{Y}_i)^2}{\sum_n^{i=1}(Y_i - \bar{Y})^2}$$

3、平均绝对误差（MAPE）平均绝对误差（MAE）

相对百分误差绝对值的平均值 MAPE(mean absolute percentage error):可以用来衡量一个模型预测结果的好坏。

MAE 平均绝对误差=｜原值-估计值｜/n

$$\mathrm{MAE} = \frac{\sum_n^{i=1}\left|Y_i - \hat{Y}_i\right|}{n}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_i - \hat{Y}_i\right|}{Y_i}$$

3.5 交叉验证

1、定义

交叉验证（Cross Validation），有的时候也称作循环估计（Rotation Estimation），是一种统计学上将数据样本切割成较小子集的实用方法，该理论是由 Seymour Geisser 提出的。在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。把每个样本的预报误差平方加和，称为 PRESS(predicted Error Sum of Squares)。

2、基本思想

交叉验证的基本思想是把在某种意义下将原始数据(dataset)进行分组,一部分做为训练集(train set),另一部分做为验证集(validation set or test set),首先用训练集对分类器进行训练,再利用验证集来测试训练得到的模型(model),以此来做为评价分类器的性能指标。

3、交叉验证方法

（1）K 折交叉验证

初始采样分割成 K 个子样本，一个单独的子样本被保留作为验证模型的数据，其他 K-1 个样本用来训练。交叉验证重复 K 次，每个子样本验证一次，平均 K 次的结果或者使用其它结合方式，最终得到一个单一估测。这个方法的优势在于，同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次，10 折交叉验证是最常用的。

（2）Holdout 验证

将原始数据随机分为两组,一组做为训练集,一组做为验证集,利用训练集训练分类器,然后利用验证集验证模型,记录最后的分类准确率为此 Hold-OutMethod 下分类器的性能指标.。Hold-OutMethod 相对于 K-fold Cross Validation 又称 Double cross-validation ，或相对 K-CV 称 2-fold cross-validation(2-CV)。

（3）留一验证

留一验证只使用原本样本中的一个样本来当做验证集， 而剩余的则留下来当做训练资料。 这个步骤一直持续到每个样本都被当做一次验证资料。 事实上，这等同于 K-fold 交叉验证是一样的，其中 K 为原本样本个数。 在某些情况下是存在有效率的演算法，如使用 kernel regression 和 Tikhonov regularization。

4、用途

交叉验证用于评估模型的预测性能，尤其是训练好的模型在新数据上的表现，可以在一定程度上减小过拟合，还可以从有限的数据中获取尽可能多的有效信息。

3.6 本章小结

本章主要内容详细的介绍了数据集预处理的方法和随机森林算法原理，包括随机森林的基础决策树、决策树的构建原理、随机森林算法以及四种评价函数；还有 K 临近算法的基本原理以及优缺点，还有支持向量机模型的原理以及优点。

4 基于 RandomForest 的模型设计与实现

本章主要使用三种模型对建筑能耗进行预测，分别为随机森林模型、KNN 模型和支持向量机模型。严格按照机器学习预测方法步骤进行实验，包括环境配置、数据集介绍、特征重要性计算、模型训练、评价函数以及交叉验证，最后经过对比分析得出结果。

4.1 环境配置

操作系统：macOS High Sierra 10.13.4

IDE: PyCharm CE

语言：Python3.6

4.2 数据集介绍

数据集命名：energydata.csv。

表 4.1 列出数据集中特征名及其含义。

表 4.1 特征名与实际意义

特征	释义	特征	释义
Date-time	日期时间	T6	建筑物北侧温度
Appliances	电能使用	RH_6	建筑物北侧湿度
Lights	电灯使用	T7	熨衣室温度
T1	厨房温度	RH_7	熨衣室湿度
RH_1	厨房湿度	T8	青少年房间温度
T2	客厅温度	RH_8	青少年房间湿度
RH_2	客厅湿度	T9	父母房间温度
T3	洗衣房温度	RH_9	父母房间湿度
RH_3	洗衣房湿度	Temperature outside	室外温度
T4	办公室温度	Pressure	室外压力

RH_4	办公室湿度	RH_out	室外湿度
T5	浴室温度	Wind speed	室外风速
RH_5	浴室湿度	Visibility	能见度
		Tdewpoint	露点

数据集包含每十分钟提取一次的信息，共约 4.5 个月，采用 ZigBee 无线传感器网络对室内温湿度条件进行监测。最近的气象站(比利时 Chievres 机场)的天气信息可以从公共数据集网站下载，并与使用日期和时间列的实验数据集合并在一起。

数据集共 19735 个实例，无空缺值。

4.3 数据预处理

数据集使用的是源于蒙斯大学的电量能耗数据 energydata.csv

4.3.1 数据集的合成

我将原数据集划分为两个数据集：

1、除预测结果外的特征值(data.csv)

2、电能使用数值（target.csv）

之后使用 load_csv 方法导入上述 csv 文件

```
def load_csv(filename):  #导入csv文件
    dataset = list()
    with open(filename, 'r') as file:
        csv_reader = reader(file)
        for row in csv_reader:
            if not row:
                continue
            dataset.append(row)
    return dataset
```

4.3.2 数据集标准化

时间特征的处理，共使用了 4 种方法：

1、将每天的 144 个特征转化为数值格式 1-144：

```
for i in range(len(X)):
    chuli = X[i][0].split(':')
    X[i][0] = int(chuli[0]) * 6 + (int(chuli[1]) / 10) + 1
```

2、将每天的每个特征转换为秒数：

```
for i in range(len(X)):
    # print(X[i][0])
    chuli = X[i][0].split(':')
    X[i][0] = int(chuli[0]) * 3600 + (int(chuli[1])*60)
```

3、将每天的每个特征转换为分钟数：

```
for i in range(len(X)):
    chuli = X[i][0].split(':')
    X[i][0] = int(chuli[0]) * 60 + (int(chuli[1]))
```

4、将每天的每个特征归一化（0-1）：

```
for i in range(len(X)):
    chuli = X[i][0].split(':')
    X[i][0] = (int(chuli[0]) * 6 + (int(chuli[1]) / 10) + 1)/len(X)
```

数据集中能耗与时间的关系图如图 4.1 所示：

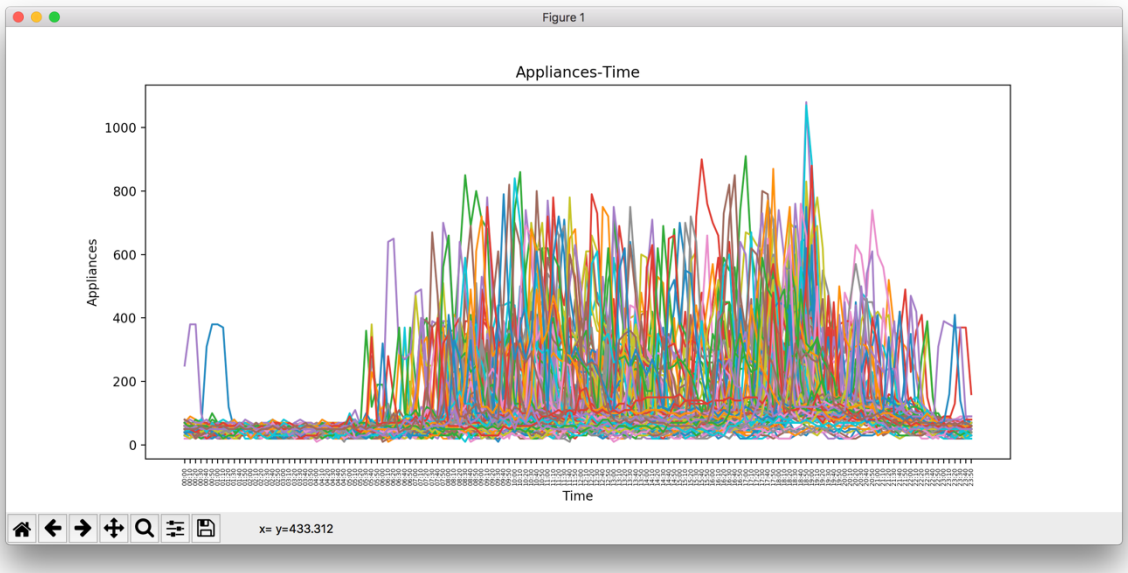

图 4.1 时间-能耗折线图

4.3.3 数据集切分

使用 2/3 数据集用来训练，1/3 数据集用来测试

训练数据-特征：

```
X = train_data[:int(len(train_data)*2/3)]
# print(len(X))# 13056
print('----------data.csv导入成功----------')
```

训练数据-结果：

```
train_target = load_csv('target.csv')
print('----------target.csv导入成功----------')
# 处理结果target数组 得到y
a = []
for i in train_target:
    # print(i)
    a.append(float(i[0]))
    # a.append(i[0])

train_target = a
y = train_target[:int(len(train_target)*2/3)]
# print(len(y))# 13056
print('----------target.csv处理完成----------')
```

4.4 特征重要性

特征重要性通过随机森林算法进行计算分析并且可视化，如图 4.2。

使用 Sklearn 工具包中的随机森林林分类器训练出模型后便可以得到生成决策树时使用的特征判断因素，也就是特征重要性程度，核心代码：

```
def feature_importances_(self):
    check_is_fitted(self, 'estimators_')

    all_importances = Parallel(n_jobs=self.n_jobs,
                               backend="threading")(
        delayed(getattr)(tree, 'feature_importances_')
        for tree in self.estimators_)

    return sum(all_importances) / len(self.estimators_)
```

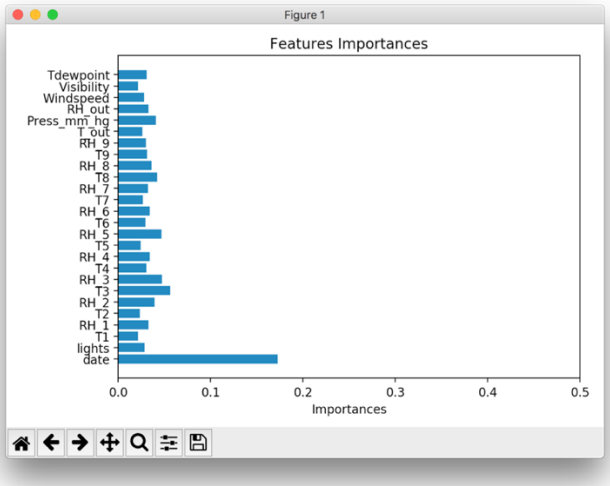

图 4.2 特征重要性柱状图

4.5 机器学习模型

4.5.1 随机森林模型

随机森林模型使用 sklearn 工具包中的 RandomForestRegressor 方法构建。

拟合模型方法：

clf = RandomForestRegressor(n_estimators=100,oob_score = 'true')

参数调优：RandomForestRegressor()常用内置参数

1、max_features：随机森林允许单个决策树使用特征的最大数量，增加 max_features 一般能提高模型的性能，因为在每个节点上，我们有更多的选择可以考虑。然而，这未必完全是对的，因为它降低了单个树的多样性，而这正是随机森林独特的优点。 但是，可以肯定，你通过增加 max_features 会降低算法的速度。因此，你需要适当的平衡和选择最佳 max_features。

2、n_estimators：建立子树的数量，较多的子树可以让模型有更好的性能，但同时让你的代码变慢。你应该选择尽可能高的值，只要你的处理器能够承受的住，因为这使你的预测更好更稳定。

3、min_sample_leaf：最小样本叶片大小。

4、random_state 此参数让结果容易复现。一个确定的随机值将会产生相同的结果，在参数和训练数据不变的情况下。我曾亲自尝试过将不同的随机状态的最优参数模型集成，有时候这种方法比单独的随机状态更好。

5、random_state 是随机数生成器使用的种子;如果是 RandomState 实例，random_state 就是随机数生成器;如果为 None，则随机数生成器是 np.random 使用的 RandomState 实例。

6、oob_score 这是一个随机森林交叉验证方法。它和留一验证方法非常相似，但这快很多。这种方法只是简单的标记在每颗子树中用的观察数据。然后对每一个观察样本找出一个最大投票得分，是由那些没有使用该观察样本进行训练的子树投票得到。

7、criterion string, optional (default=”gini”) 字符串，可选择(默认值为“gini”)。衡量分裂质量的性能（函数）。

8、max_depth（决策）树的最大深度。

9、min_samples_split 分割内部节点所需要的最小样本数量。

10、min_samples_leaf 需要在叶子结点上的最小样本数量。

11、max_leaf_nodes 以最优的方法使用 max_leaf_nodes 来生长树。

12、bootstrap 建立决策树时，是否使用有放回抽样。

13、estimators_决策树分类器的序列。

14、feature_importances_特征的重要性（值越高，特征越重要）。

15、n_jobs 这个参数告诉引擎有多少处理器是它可以使用。“-1”意味着没有限制，而“1”值意味着它只能使用一个处理器。

调优过程：

1、n_estimators 的大小对于正确率的影响不是很大，n_estimators 值与正确率的关系如图 4.3 所示：

```
k_range = range(1, 100)

k_scores = []

for k in k_range:
    random = RandomForestRegressor(n_estimators=k)
    random.fit(X, y)
    predict = random.predict(X)
    a = 0
    wucha = 10
    for i in range(len(y)):
        if abs((predict - y)[i]) < wucha:
            a += 1
    acc = a / len(y)
    k_scores.append(acc)
```

表 4.2 n_estimators 不同参数所用时间及正确率

参数值	所用时间	正确率
10	14 秒	75.5%
100	43 秒	76.3%
1000	598 秒	76.8%

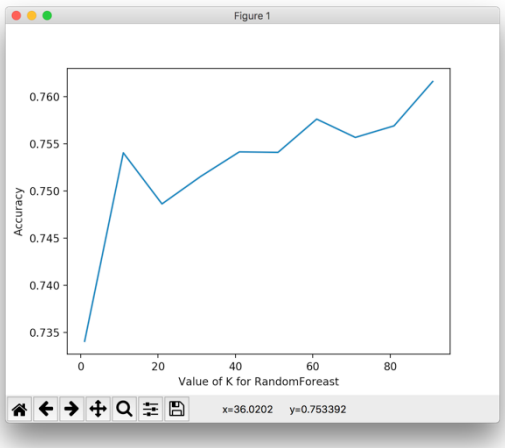

图 4.3 n_estimators 不同参数下正确率

2、oob_score 对于正确率影响不大

表 4.3 oob_score 不同参数下正确率

参数值	正确率
‘true’	75.5%
‘false’	75.5%

3、max_depth 在一定区间内对于正确率影响巨大，表 4.1

表 4.4 max_depth 不同参数下正确率

参数值	正确率
5	30%
10	38%
100	75%
1000	76%

max_depth 值与正确率的关系如图 4.4 所示：

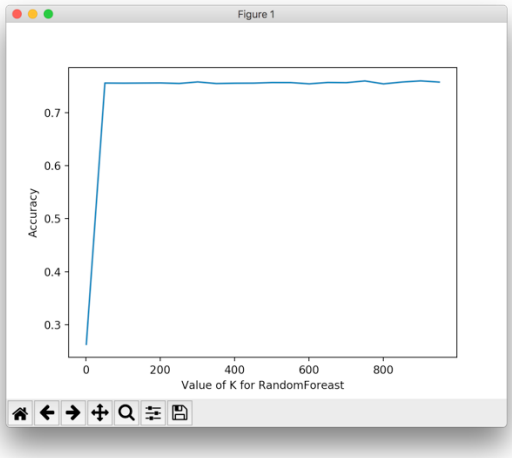

图 4.4 max_depth 不同参数下正确率

计算误差在 N Wh 内正确率，默认为 10Wh：

```
a = 0
wucha = N
for i in range(len(y)):
    if abs((predict - y)[i])<wucha:
        a += 1
acc = a/len(y)
```

4.5.2KNN 模型

K 临近模型使用 sklearn 工具包中的 RandomForestRegressor 方法构建。

拟合模型方法：

clf = KNeighborsRegressor(n_neighbors=10)

参数调优：KNeighborsRegressor()常用内置参数

1、n_neighbors：默认情况下用于 kneighbors 查询的邻居数量。

2、weights：权重函数用于预测。‘uniform’统一的质量,‘distance’权重点距离的倒数，‘callable’用户定义的函数。

3、algorithm：最小样本叶片大小。‘ball_tree’使用 ball_Tree，‘kd_tree’使用 KDTree，‘brute’使用蛮力搜索。

4、leaf_size：叶子大小传递给 BallTree 或 KDTree。这会影响构建和查询的速度，以及存储树所需的内存。最佳值取决于问题的性质。

5、n_jobs：运行邻居搜索的并行作业数量。如果-1，则作业数量设置为 CPU 内核数量。不影响 fit 方法。

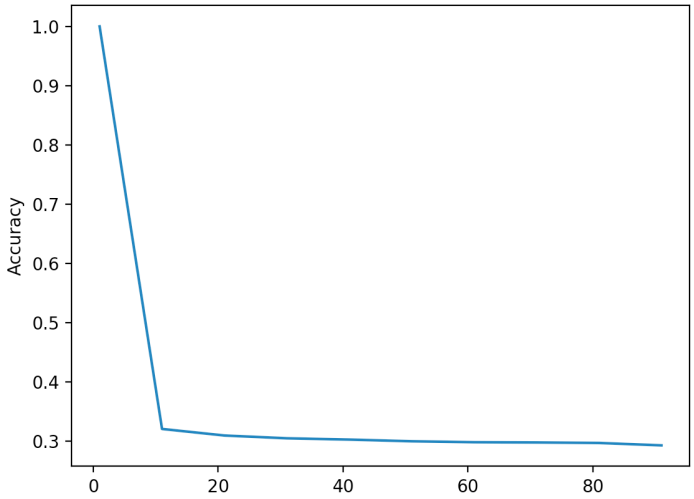

图 4.5 n_neighbors 不同参数下正确率

4.5.3 支持向量机模型

支持向量机模型使用 sklearn 工具包中的 clf = SVC()方法构建。

拟合模型方法：

clf = SVC()

参数调优：clf = SVC()常用内置参数

1、C：错误项的惩罚参数 C。

2、kernel：指定要在算法中使用的内核类型。它必须是'linear'，'poly'，'rbf'，'sigmoid'，'预计算'或可调用的。如果没有提供，将使用'rbf'。如果给出了可调用函数，它将用于从数据矩阵中预先计算内核矩阵；该矩阵应该是一个形状的数组。(n_samples, n_samples)。

3、gamma：'rbf'，'poly'和'sigmoid'的核系数。如果 gamma 是'auto'，那么将会使用 1 / n_features。

4、coef0：内核函数中的独立术语。它只在'poly'和'sigmoid'中很重要。

5、degree：多项式核函数的度数（'poly'）。被所有其他内核忽略。

4.6 评价函数

使用预测得到的数据与正确数据进行计算得到评价函数的值。

1、RMSE

```
print('----------RMSE----------')
rmse = np.sqrt(((predict - y) ** 2).mean())
print(rmse,'\n')
```

2、R 方

```
average = np.sum(y)/len(y) # 平均值
a = []
for i in range(len(y)):
    a.append(average)
r = 1 - (((predict - y)**2).sum()/(((predict - a)**2).sum()))
print('----------R方----------')
print(r,'\n')
```

4、MAE

```
mae = (abs((predict - y)).sum())/len(y)
print('----------平均绝对误差（MAE）----------')
print(mae,'\n')
```

5、MAPE

```
mape = (abs((predict - y))/predict).sum()/len(y)
print('----------平均绝对百分误差（MAPE）----------')
print(mape,'\n')
```

表 4.5 三种模型的评价指数比较

模型	RMSE	R^2	MAE	MAPE
RandomForest	28.06%	90.27%	12.36%	9.72%
KNN	64.05%	33.93%	31.79%	22.79%
SVM	70.24%	29.56%	35.91%	25.47%

随机森林模型的决定系数 RMSE=0.28、R^2=0.90、MAE=0.12、MAPE=0.10，模型拟合效果较其他两个更为理想。

4.7 交叉验证

在研究中我们使用 sklearn 中的 K 折交叉验证模块进行交叉验证。

首先将数据集按照格式分割成训练集和测试集共 4 部分，经过 cross_val_score 和 KFold 方法得到交叉验证分数，结果如图 4.6 所示：

```
# 使用K折交叉验证模块
scores = cross_val_score(knn, X, y, cv=5, scoring='accuracy')
KFold(10, n_folds=2)
# 建立测试参数集
k_range = range(1, 100)
k_scores = []
# 藉由迭代的方式来计算不同参数对模型的影响，并返回交叉验证后的平均准确率
for k in k_range:
    random = RandomForestRegressor(n_estimators=k)
    scores = cross_val_score(random, X, y, cv=10, scoring='accuracy')
    k_scores.append((scores).mean())
```

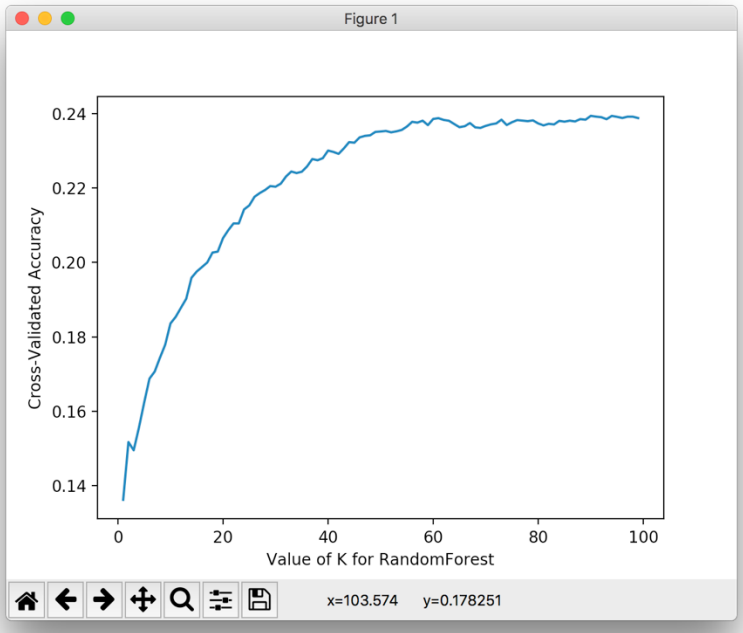

图 4.6

4.8 本章小结

本章主要描述了针对特定数据集的分析以及预测的过程，包括了数据集的处理，特征重要性的分析以及随机森林模型、KNN 模型和支持向量机模型的训练与调优，计算了评价函数并进行了交叉验证；科学的、有参照对比的得到随机森林对于此项目优于其他算法的结论。

5 结　论

本文引入随机森林算法构建决策树，通过随机森林重要性评价功能优化筛选因子，得到变量重要性，构建建筑能耗预测模型，研究并提出了一种基于随机森林算法的建筑能耗预测模型。在相同的条件下，利用KNN算法和SVM算法构建了预测模型，将3种模型进行比较，结果表明，各种特征条件相同的情况下，使用随机森林模型得到的预测结果要更优越另外两种算法的结果，对数据集的适应能力强，能合理有效地预测建筑能耗。本文还有一些因素没有考虑在内，比如地形因素、地理位置因素等，有待于继续深入研究。

谢 辞

衷心感谢我的指导老师。她严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。老师学识渊博，品德高尚，平易近人，在我学习期间不仅仅传授了做学问的秘诀，还传授了做人的准则，这些都将使我终生受益。在我毕业论文的写作过程中，老师始终给予我精心的指导和不懈的支持。她循循善诱的教导和不拘一格的思路给予我无尽的启迪。在此谨向老师致以诚挚的谢意和崇高的敬意。

同时，我也要向身边的同学表示感谢，在整个毕业设计的过程中他们帮我解决了许多棘手的问题，并且论文中某些观点提出和修正与他们的讨论分不开。

路漫漫其修远兮，吾将上下而求索。我愿在未来的学习和研究过程中，以更加丰厚的成果来答谢以前关心、帮忙和支持过我的老师和同学。

参考文献

[1]王文寅,郭鹏波.基于随机森林的股权众筹项目风险评估研究[J].河南科学,2018,(2):283-289.

[2]刘剑,曹美燕,高治军,等.一种基于随机森林的太阳能辐射预测模型[J].控制工程,2017,(12):2472-2477. DOI:10.14107/j.cnki.kzgc.150745.

[3]张雯,刘爱利,齐威,等.基于随机森林的月貌面向对象分类[J].遥感信息,2018,(1):93-98. DOI:10.3969/j.issn.1000-3177.2018.01.014.

[4]陈苏雨,方宇,胡定玉.基于随机森林方法的地铁车门故障诊断[J].测控技术,2018,(2):20-24.

[5]魏金太,高穹.基于信息增益和随机森林分类器的入侵检测系统研究[J].中北大学学报（自然科学版）,2018,(1):74-79,88. DOI:10.3969/j.issn.1673-3193.2018.01.013.

[6]吕杰,郝宁燕,李崇贵,等.利用随机森林和纹理特征的森林类型识别[J].遥感信息,2017,(6):109-114. DOI:10.3969/j.issn.1000-3177.2017.06.018.

[7]王利民,刘佳,杨玲波,等.随机森林方法在玉米-大豆精细识别中的应用[J].作物学报,2018,(4):569-580. DOI:10.3724/SP.J.1006.2018.00569.

[8]Mikio Kaihara,Satoshi Kikuchi.Discriminant analysis of countries growing wakame seaweeds: a preliminary comparison of visible-near infrared spectra using soft independent modelling, Randomforests and classification and regression trees[J].Journal of near infrared spectroscopy,2007,(6):371-377.

[9]Corcoran Jennifer,Knight Joseph,Pelletier Keith, et al.The Effects of Point or Polygon Based Training Data on RandomForest Classification Accuracy of Wetlands[J].Remote Sensing,2015,(4):4002-4025. DOI:10.3390/rs70404002.

[10]Torbick Nathan.Monitoring Rice Agriculture across Myanmar Using Time Series Sentinel-1 Assisted by Landsat-8 and PALSAR-2[J].Remote Sensing,2017,(2). DOI:10.3390/rs9020119.