从决策树到随机森林算法综述

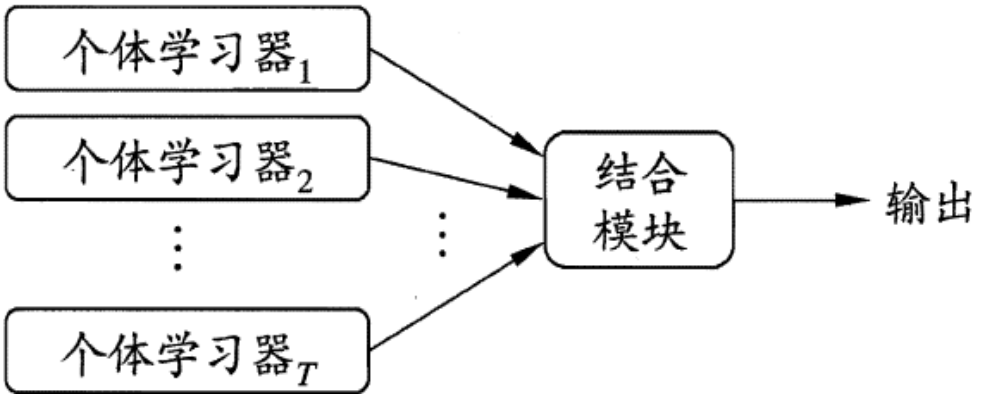
作者：王奕力　　学号：1900011608　　指导老师：陈斌

[摘要]随机森林算法是一种基于决策树的集成学习算法，具有很高的预测准确性，不容易过拟合，并且易于实现特征重要性评估，在许多领域有着广泛的应用。同时，随机森林算法也有不足和改进的空间。为此，本文首先简单介绍了从决策树到随机森林算法的原理，并对该算法做出一些评价，然后综述了近几年来随机森林算法的一些改进研究，最后对该算法的发展方向进行总结和预测。

关键词：随机森林 机器学习 决策树

引言

解决分类问题的算法有很多，例如朴素贝叶斯算法、支持向量机算法、决策树算法等。这些都是单个的分类器，很容易出现过拟合问题，因此集成算法（如图[①]）应运而生。集成算法主要有Boosting（提升法）和Bagging（套袋法）两种，Bagging是通过结合几个模型降低泛化误差的技术，其中又以2001年由Leo Breiman提出的随机森林算法最具代表性。



一、原理简介

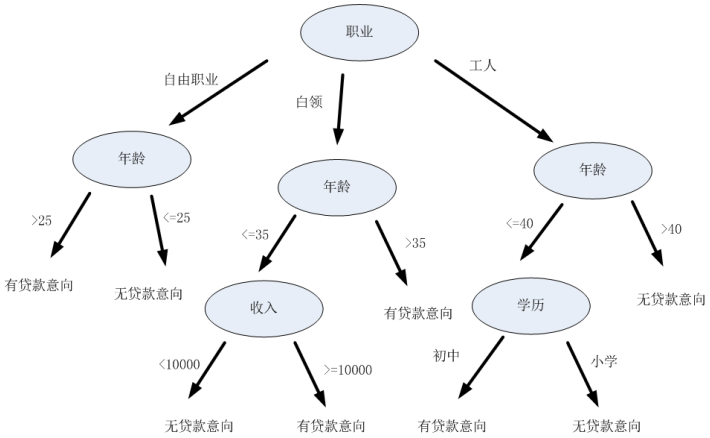
（一）随机森林的基本构成要素——决策树

决策树可以看作是通过对一系列关于数据是/否问题的判定，最终得出一个预判类型（或回归情况下的连续值）的模型，代表对象属性和对象值之间的一种映射关系。可以借助下图[②]进行通俗直观的理解。其中我们把职业、年龄等称作特征，有/无贷款意向是需要预判的类（标签），每一个人则是一个样本。

---

① 图片来自 CSDN 随机森林算法梳理

② 图片来自百度图片搜索



其中，节点的分裂处的问题（如：年龄>25？），并非是人为预先设定好的，而是在对决策树进行训练时生成的。训练的过程中，我们为模型提供特征和标签，帮助它学习如何根据特征对样本进行分类。在CART算法中，所有可能的问题中，得到应答时会导致Gini不纯度减少量最大的问题，也即分类效果最好的问题，会参与决策树的分叉。

节点的Gini不纯度，即分类时，从节点中随机选择的样本被分错的概率，用1减去每个类的样本比例的平方和表示：

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2$$

Gini不纯度越低，说明该集合中单一类的比例越高，即纯度越高。Gini不纯度减少量越大，分类的效果越好。在每个节点，决策树要在所有特征中搜索最适合用于拆分的值，从而可以最大限度地减少树每层的加权总Gini不纯度：

$$\mathrm{I}_{加权} = \frac{n_{left}}{n_{总}} * I_{left} + \frac{n_{right}}{n_{总}} * I_{right}$$

（二）过拟合带来的思考——如何实现方差-偏差权衡

在决策树模型中，若对划分的深度不加限制，它将以贪婪递归的过程重复这种拆分，直到达到最大深度，或者最后一层的加权总Gini不纯度变为0。但这种“完美分类”往往意味着模型可能过拟合，因为所有节点都是使用训练数据构建的，这种“完美”仅仅是对于给出的训练数据，而对于从未见过的新数据无法很好地泛化。

无限灵活的决策树模型通过紧密拟合来“记住”训练数据，但除了训练数据的实际关系，还学习了存在的噪音，导致学到的参数随着训练数据的不同变化很大，具有很高的方差。

但如果限制了最大深度，即规定拆分次数，则又增加了偏差，降低了预测结果的精确性。

为了实现偏差-方差权衡，将许多决策树组合成一个集成模型，即为随机森林。

（三）三个臭皮匠，顶个诸葛亮——随机森林算法

随机森林算法不是决策树的简单平均，它的关键在于在构建树时对训练数据进行随机抽样，以及分割节点时考虑特征的随机子集。

（1）自助抽样法

从原始训练样本集N中有放回地重复随机抽取k 个样本生成新的训练样本集合，根据样本集生成k 棵决策树，并且随机组合得到随机森林，新数据的分类结果按决策树投票多少形成的分数而定；

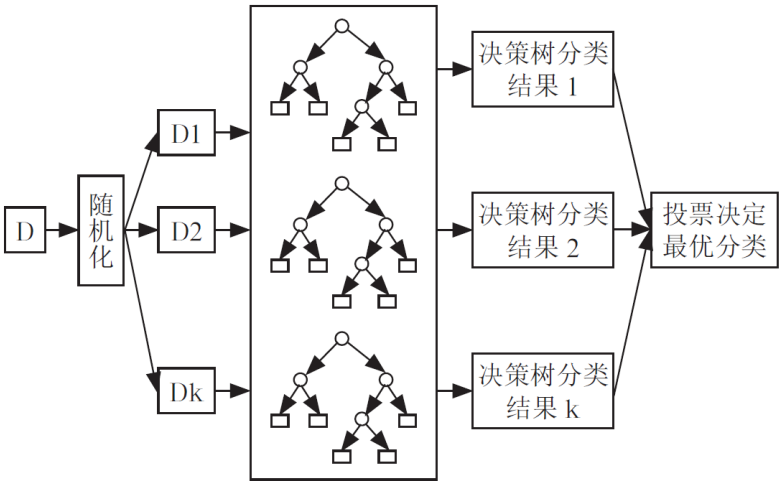


图 1　随机森林示意图

D 是样本集，D1，D2，Dk 分别是每次随机抽样后生成的决策树。随机森林示意图如图1[③] 所示。

采取自主抽样法后，虽然每棵树相对于特定的训练数据具有高方差，但整个森林具有较低的方差，并且不会增加偏差。

（2）特征的随机子集

对于普通的决策树，我们会在节点上所有的m个样本特征中选择一个最优的特征来做决策树的左右子树划分。但是随机森林的每个树，其实是随机选用一部分特征，在这些少量特征中，每次选择一个最优的特征来做决策树的左右子树划分，将随机性的效果扩大，进一步增强了模型的泛化能力。

假设每棵树选取n个特征，n越小，此时模型对于训练集的拟合程度会变差，偏差增加，但是会泛化能力更强，模型方差减小。n越大则相反。在实际使用中，一般会将n的取值作为一个参数（推荐值$n = \log_2 m$[④]），通过开启OOB验证或使用交叉验证，不断调整参数以获取一个合适的n的值。

简单来说，随机森林算法里的每棵决策树都只随机使用了部分的样品，并且根据随机的部分特征进行分类，因此具有较强的普适性，同时许多树组成森林又保证了预测的相对精确。

[③] 图片来自孙明喆, 毕瑶家, 孙驰. 改进随机森林算法综述[J]. 现代信息科技, 2019(20).

[④] 该推荐值为 CSDN 搜索获得。

二、对随机森林算法的评价

举一个通俗的例子，假设你想要投资股票，样本是过去一个月几百只股票的相关信息。决策树就是一个人分析所有数据，建立一个精细的模型，但这个模型可能受到大量无关、异常信息干扰，而难以预测未来走势。随机森林就像是一群专家，每个人根据其中几个方面的信息，对一部分股票样本进行分析，给出一个建议，最后根据投票多少决定最终方案。可见，随机森林算法对缺失值、异常值不敏感，模型训练结果准确度更高，有足够多的树，分类器就不会过度拟合。

随机森林非常方便且易于使用，既可以用于回归也可以用于分类任务，并且很容易通过置换法查看模型的输入特征的相对重要性：在包外样本集中随机挑选两个样本，如果要计算某一变量的重要性，则置换这两个样本的这个特征值，统计置换前和置换后的分类准确率。

变量重要性的计算公式为：

$$\mathrm{V} = \frac{n_{before} - n_{after}}{n_{总}}$$

其中n_{before}和n_{after}为置换前后正确分类的样本数，$n_{总}$为样品总数。

但同时，随机森林算法也有其缺点：由于使用大量的树会使算法变得很慢，越准确的预测需要越多的树，这将导致模型越慢，因此无法应对实时性要求很高的情况；另外，因为训练集是随机选取的，加剧了数据集的不平衡性，使随机森林算法在不平衡数据的分类性能明显不如支持向量机。

三、对随机森林的改进方向的总结和预测

通过文献搜集和阅读发现，正因为随机森林算法的不足和潜力，近年来很多研究都提出了随机森林算法的创新性改进。

例如，针对随机森林算法面对特征维度高且不平衡的数据时，算法分类性能严重削弱的问题，有研究者结合权重排序和递归特征筛选，提出了基于特征约减的随机森林改进算法研究；对于部分分类性能差和相似度较高的决策树影响模型整体性能的问题，提出基于聚类约减决策树的改进方法；还有引入梯度提升算法、采用Spark分布式、融合因子分析等改进方法。

根据对近年相关论文的总结，现有的改进方案大致可以分为三类：一是进行数据的预处理，通过投票环节的加权，提高对非平衡数据的敏感度；二是优化随机森林的构建过程，通过对特征重要性的处理，减少决策树数量，加快收敛速度；三是引入新的理论，和随机森林算法融合，集成各家的优势提升性能。

据此推测，这三点也很可能是未来一段时间对随机森林算法的改进方向。

参考文献：

[1] 王诚. 一种基于聚类约简决策树的改进随机森林算法[J]. 南京邮电大学学报：自然科学版, 2019, 39(3):91-97.

[2] 吕红燕, 冯倩. 随机森林算法研究综述[J]. 河北省科学院学报, 2019(3):37-41.

[3] 孙明喆, 毕瑶家, 孙驰. 改进随机森林算法综述[J]. 现代信息科技, 2019(20).

[4] 杨晔民.随机森林的可解释性可视分析方法研究

[5] 袁志聪.人工智能_随机森林技术分析

[6] 王诚.基于特征约减的随机森林改进算法研究

[7] 谢坤.基于数据集成的随机森林算法

[8] 知乎.一文读懂随机森林的解释和实现

[9] CSDN 机器学习之十大经典算法（十） 随机森林算法

[10] CSDN 随机森林算法梳理

[11] CSDN 随机森林算法工作原理

[12] CSDN 随机森林算法总结

[13] CSDN 随机森林算法学习(RandomForest)

[14] Python 机器学习笔记——随机森林算法!