

【量化专题】：机器学习模型理论—决策树的剪枝

专题报告

摘要

- 决策树生成算法递归地生成决策树。这样生成的决策树对训练数据的分类很准确，但对未知的测试数据的分类不那么准确，容易发生“过拟合现象”。所以需要决策树的剪枝策略优化过拟合问题。
- 本文将针对决策树的预剪枝和后剪枝对决策树的原理进行介绍。

作者姓名：姜慧丽

邮箱：jianghuili@csc.com.cn

电话：023-81157278

期货从业资格号：F3081375

期货投资咨询从业证书号：Z0018496

发布日期：2024年3月28日

风险提示：本报告仅对模型作客观呈现，不具备任何投资建议。历史业绩不代表未来业绩，回测业绩不代表实盘业绩，期市有风险，入市需谨慎。

目录

【量化专题】：机器学习模型理论—决策树的剪枝	1
摘要	1
一、 决策树剪枝的思想	1
二、 决策树损失函数	1
三、 决策树的预剪枝	2
四、 决策树的后剪枝	2
五、 总结	4

一、 决策树剪枝的思想

在决策树学习的过程中，为了尽可能将训练样本正确分类，结点划分过程会不断重复，训练集自身的某些特点可能被当做所有数据具备的一般性质从而导致“过拟合”。决策树的分支越多、层数越多、叶结点越多，越容易“过拟合”，从而导致模型泛化能力差。

为了增强模型的泛化能力，应减少决策树的复杂度、对已生成的决策树进行简化，也就是剪枝。剪枝（pruning）算法的基本思路为剪去决策树模型中的一些子树或者叶结点，并将其上层的根结点作为新的叶结点，从而减少了叶结点甚至减少了层数，降低了决策树复杂度。从基本策略上讲，决策树的剪枝分为预剪枝和后剪枝，下边将分别介绍这两种剪枝策略。

二、 决策树损失函数

决策树的剪枝往往通过最小化决策树整体的损失函数实现，本文首先介绍损失函数。设树 T 的叶结点个数为 $|T|$ ， t 是树 T 的一个叶结点，该叶结点有 N_t 个样本点，其中类别 k 的样本点有 N_{tk} 个， $k=1,2,\dots,K$ ， $H_t(T)$ 为叶结点 t 上的经验熵， $\alpha \geq 0$ 为参数，所以决策树的损失函数可以定义为：

$$C_a(T) = \sum_{t=1}^{|T|} N_t H_t(T) + \alpha |T|$$

其中经验熵公式：

$$H_t(T) = - \sum_k \frac{N_{tk}}{N_t} \log \frac{N_{tk}}{N_t}$$

记

$$C(T) = \sum_{t=1}^{|T|} N_t H_t(T) = - \sum_{t=1}^{|T|} \sum_{k=1}^K N_{tk} \log \frac{N_{tk}}{N_t}$$

则损失函数可以写为：

$$C_a(T) = C(T) + \alpha|T|$$

$C(T)$ 是模型对训练数据的预测误差，表示模型和训练集之间的拟合程度。 $\alpha|T|$ 为惩罚项，相当于对损失函数做了约束， $|T|$ 表示树的叶节点的个数，即表示树的复杂度，参数 $\alpha \geq 0$ 控制二者之间的影响，相当于 α 越大，叶节点的个数对损失函数的影响越大，剪枝之后的决策树更易选择复杂度较小的树， α 越小，表示叶节点的个数对损失函数影响越小， $\alpha=0$ 意味着只考虑模型与训练集的拟合程度，不考虑模型的复杂度。所以 α 的大小控制了预测误差与树的复杂度对剪枝的影响。

剪枝，就是当 α 确定时，选择损失函数最小的模型，即损失函数最小的子树。

三、 决策树的预剪枝

预剪枝是在构建决策树的时候同时进行剪枝工作，当发现分类有偏差时就及早停止。比如决定在某一个节点不再分裂，则一旦停止，该节点就成为叶子节点。

预剪枝的方法有很多，如：

- 1、提前设定决策树的高度，当达到这个高度时，就停止构建决策树；
- 2、当达到某节点的实例具有相同的特征向量，也可以停止树的生长；
- 3、设定某个阈值，当达到某个节点的样例个数小于该阈值的时候便可以停止树的生长，但这种方法的缺点是对数据量的要求较大，无法处理数据量较小的训练样例；
- 4、设定某个阈值，每次扩展决策树后都计算其对系统性能的增益，若小于该阈值，则让它停止生长。

预剪枝的显著缺点是无法预知下一步可能会发生的情况。假设当前决策树不满足最开始的构建要求，进行了剪枝，但实际上若进行进一步构建后、决策树又满足了要求，这种情况下，预剪枝会过早停止决策树的生长。

四、 决策树的后剪枝

后剪枝是人们普遍关注的决策树剪枝策略，与预剪枝恰好相反，后剪枝的执行步骤是先构造完成完整的决策树，再通过某些条件遍历树进行剪枝，其主要思路是通过删除节点的分支并用叶节点替换，剪去完全成长的树的子树。

目前主要应用的后剪枝方法有四种：悲观错误剪枝（Pessimistic Error Pruning, PEP），最小错误剪枝（Minimum Error Pruning, MEP），代价复杂度剪枝（Cost-Complexity Pruning, CCP），错误率降低剪枝（Reduce Error Pruning, REP）。

4.1 错误率降低剪枝法

该方法将数据集分为训练数据集和测试数据集，训练数据集用来训练生成决策树模型，测试数据集用来预测决策树模型精度。通过对比剪枝前后决策树模型对测试数据集的预测精度，决定是否进行剪枝处理。如果修剪后的决策树预测精度没有降低，则进行剪枝处理，否则不进行剪枝处理。

错误率降低剪枝法（REP）是一个比较简单的决策树剪枝方法，但是，由于使用独立测试集，与原始决策树相比，修改后的决策树可能偏向于过度修剪，这是因为一些在测试数据集中没有出现过的训练数据集所对应的分支很容易被修剪掉。

4.2 悲观错误剪枝法

与 REP 方法相似，悲观错误剪枝法采用对比剪枝前后决策树模型的精度决定是否进行剪枝处理，不过该方法引入了统计学上连续修正的概念弥补 REP 缺陷。

使用 PER 方法进行决策树剪枝，必须满足的条件为：

$$E_{\text{subtree}} \leq E_{\text{leaf}} + S(E_{\text{subtree}})$$

其中， E_{subtree} 表示剪枝前子树各叶子节点误判次数， E_{leaf} 表示剪枝后叶子节点误判次数， $S(E_{\text{subtree}})$ 表示剪枝前子树误判次数的标准差。

假设子树误差的精度满足二项分布，根据二项式性质， E_{subtree} 、 E_{leaf} 和 $S(E_{\text{subtree}})$ 三者的计算公式为：

$$E_{\text{subtree}} = \sum_{i=1}^m (e_i + 0.5)$$

$$E_{\text{leaf}} = e + 0.5$$

$$S(E_{\text{subtree}}) = \sqrt{np(1-p)}$$

其中， e_i 为剪枝前决策树第 i 个叶子节点误判次数， e 为剪枝后叶子节点误判次数， p 为剪枝前子树误判率（ $p = E_{\text{subtree}}/N$ ，其中 N 为子树样本量）。

另外，上述公式中 0.5 为修正因子，由于子节点是父节点进行分裂的结果，从理论上讲，子节点的分

类效果总比父节点好，分类的误差更小，如果单纯通过比较子节点和父节点的误差进行剪枝就完全没有意义。因此对节点的误差计算方法进行修正，修正的方法是给每一个节点都加上误差修正因子 0.5，在计算误差的时候，子节点由于加上了误差修正因子，就无法保证总误差低于父节点。

4.3 代价复杂度剪枝法

一棵完整决策树的非叶子节点为 $\{T_1, T_2, T_3, \dots, T_n\}$ ，计算所有非叶子节点表面误差率增值值（ α 值），该方法通过修剪表面误差增益最小的非叶子节点，完成对决策树的剪枝处理，表面误差增值值的计算公式为：

$$\theta = \frac{R(t) - R(T)}{N(T) - 1}$$

其中， $R(t)$ 为叶子结点误差代价， $R(T)$ 为子树误差代价， $N(T)$ 为子树节点个数， $R(t)$ 和 $R(T)$ 计算公式如下：

$$R(t) = r(t) * p(t)$$

$$R(T) = \sum_i^m r_i(t) * p_i(t)$$

其中， $r(t)$ 为节点的错误率， $p(t)$ 为节点数据量的占比； $r_i(t)$ 为节点 i 的错误率， $p_i(t)$ 为节点 i 的数据量的占比。

五、总结

本文介绍了决策树剪枝的原理。

预剪枝和后剪枝方法比较上：一是后剪枝决策树一般比预剪枝决策树保留更多的分支；二是一般情况下，后剪枝决策树欠拟合风险较小，泛化性能往往优于预剪枝决策树；三是后剪枝决策树训练时间比预剪枝决策树和未剪枝决策树要大很多。

联系我们

全国统一客服电话：400-8877-780

网址：www.cfc108.com

获取更多研报报告、专业客户经理一对一服务、
了解公司更多信息，扫描右方二维码即可获得！



重要声明

本报告内容仅供符合《证券期货投资者适当性管理办法》规定可参与期货交易的投资者参考。在任何情形下都不构成对接收本报告内容投资者的任何投资建议，投资者应充分了解各类投资风险并谨慎考虑本报告发布内容是否符合自身特定状况，自主做出投资决策并自行承担投资风险。中信建投期货不因任何订阅或接收本报告的行为而将订阅人视为中信建投的客户，投资者依据本报告内容作出的任何决策与中信建投期货或作者无关。

本报告发布内容如属于系列解读，则投资者可能会因缺乏对完整内容的了解而对其中假设依据、研究依据、结论等内容产生误解，提请投资者参阅我司已发布的完整系列报告，仔细阅读其所附各项声明、数据来源及风险。

中信建投期货对本报告所载资料的准确性、可靠性、时效性及完整性不作任何明示或暗示的保证，本报告意见仅代表报告发布之时的判断，相关研究观点可能依据我司后续发布的报告在不发布通知的情形下作出更改。

本报告发布内容为中信建投期货所有。未经我司书面许可，任何机构和个人不得以任何形式对本报告进行翻版、复制和刊发，如需引用、转发等，需注明出处为“中信建投期货”，且不得对本报告进行任何增删或修改。亦不得从未经我司书面授权的任何机构、个人或其运营的媒体平台接收、翻版、复制或引用本报告发布的全部或部分内容。版权所有，违者必究。

全国统一客服电话：400-8877-780

网址：www.cfc108.com