# SPSS实用教程：决策树预测分类模型

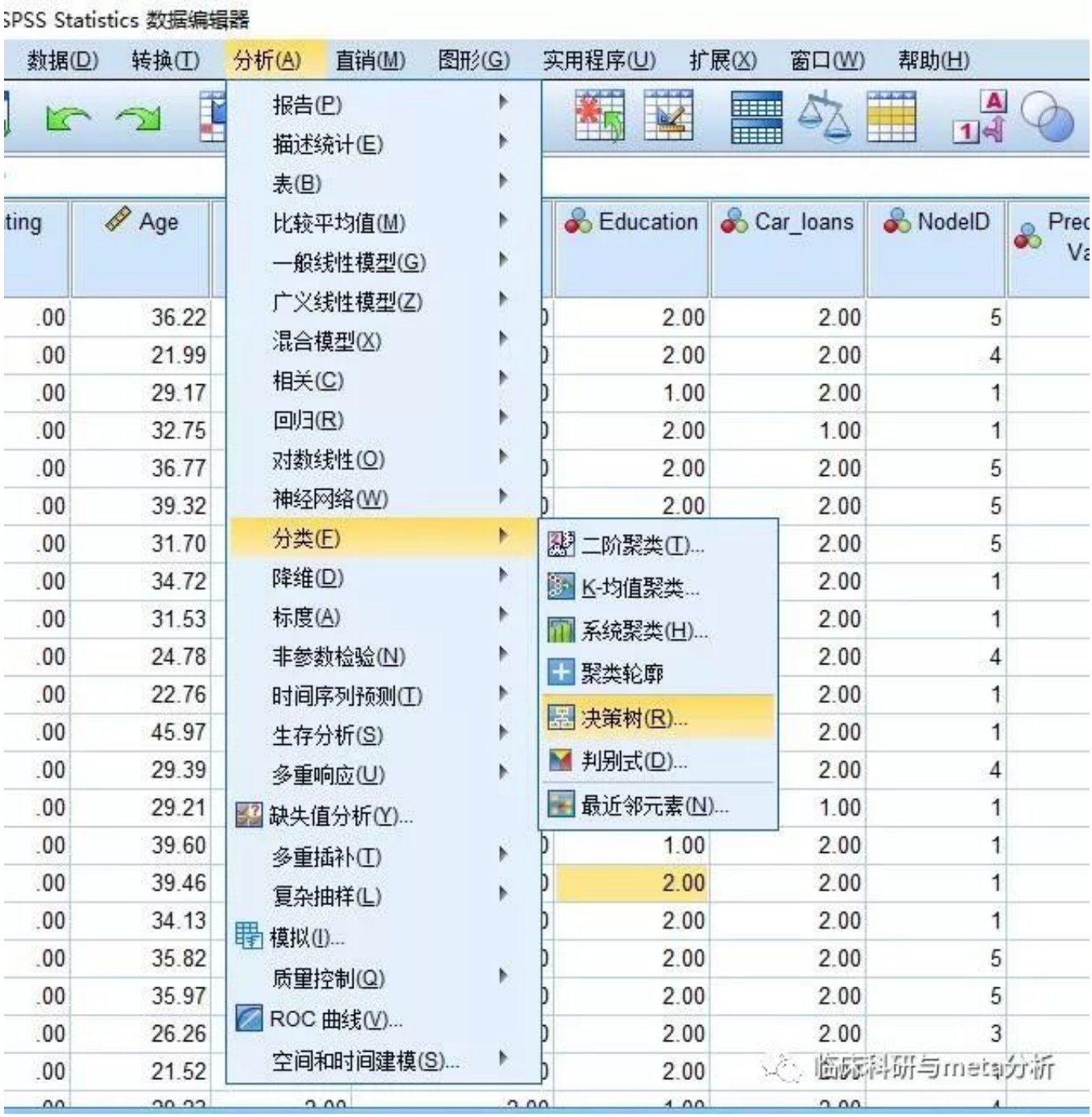
作者：Doctor Jiang 来源：临床科研与meta分析

本文原地址：https://www.iikx.com/news/statistics/573.html

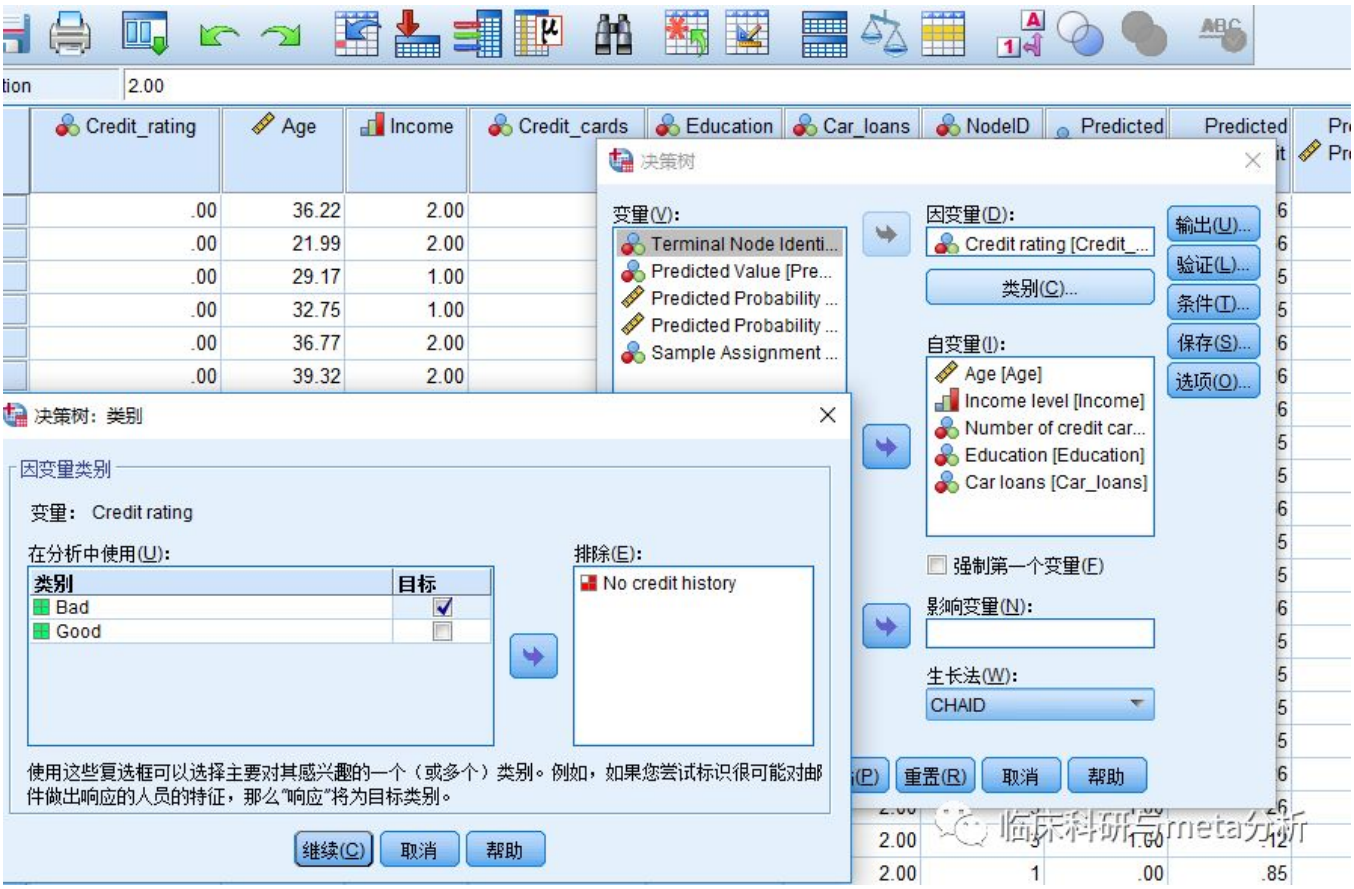


本次使用的数据为银行的信用好坏情况数据。自变量包括了收入水平、信用卡数量、教育水平、贷款次数，年龄。

点击分类，决策树

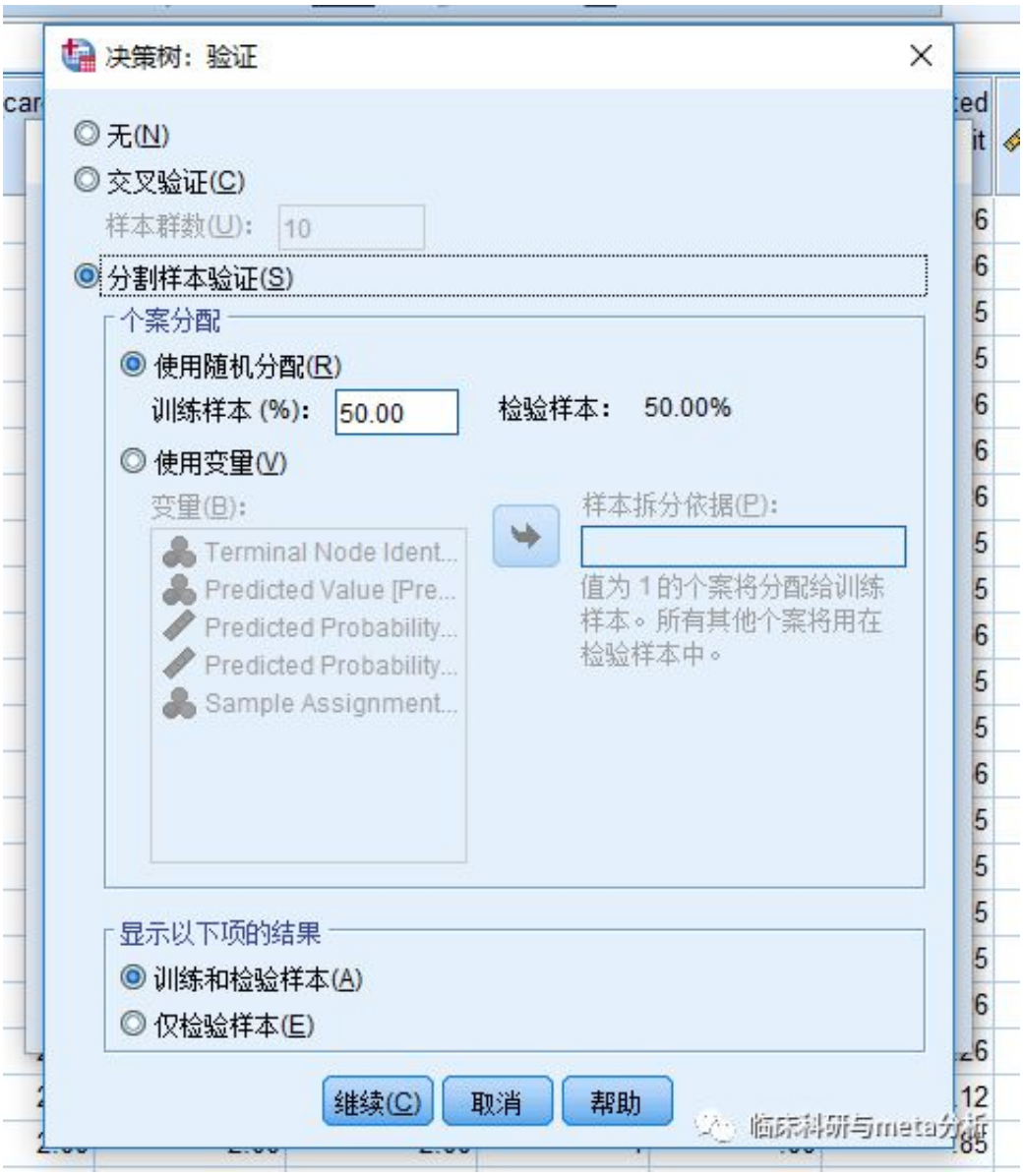



将相应变量选入应变量以及自变量。点击自变量的类别，进行勾选bad，因为我们只对信用差的感兴趣。


Credit_rating
Age
Income
Credit_cards
Education
Car_loans
NodeID
Predicted
决策树
变量(V):
Terminal Node Identi...
Predicted Value [Pre...
Predicted Probability ...
Predicted Probability ...
Sample Assignment ...
因变量(D):
Credit rating [Credit_...
类别(C)...
自变量(I):
Age [Age]
Income level [Income]
Number of credit car...
Education [Education]
Car loans [Car_loans]
强制第一个变量(F)
影响变量(N):
生长法(W):
CHAID
输出(U)...
验证(L)...
条件(T)...
保存(S)...
选项(O)...
重置(R)
取消
帮助
决策树：类别
因变量类别
变量： Credit rating
在分析中使用(U):
类别
目标
Bad
Good
排除(E):
No credit history
使用这些复选框可以选择主要对其感兴趣的一个（或多个）类别。例如，如果您尝试标识很可能对邮件做出响应的人员的特征，那么"响应"将为目标类别。
继续(C)
取消
帮助
临床科研与meta分析

点击输出


点击验证，我们选择50%的样本用于验证。




我们将收敛限制为父节点最小个案数为400，子节点为200个。


决策树
决策树：标准
增长限制
CHAID
区间
最大树深度
自动(A)
对于 CHAID，最大级别数为 3；对于 CRT 和 QUEST，此级别数为 5。
定制(C)
值(V):
最小个案数
父节点(P):
400
子节点(H):
200
继续(C)
取消
帮助
临床科研与meta分析

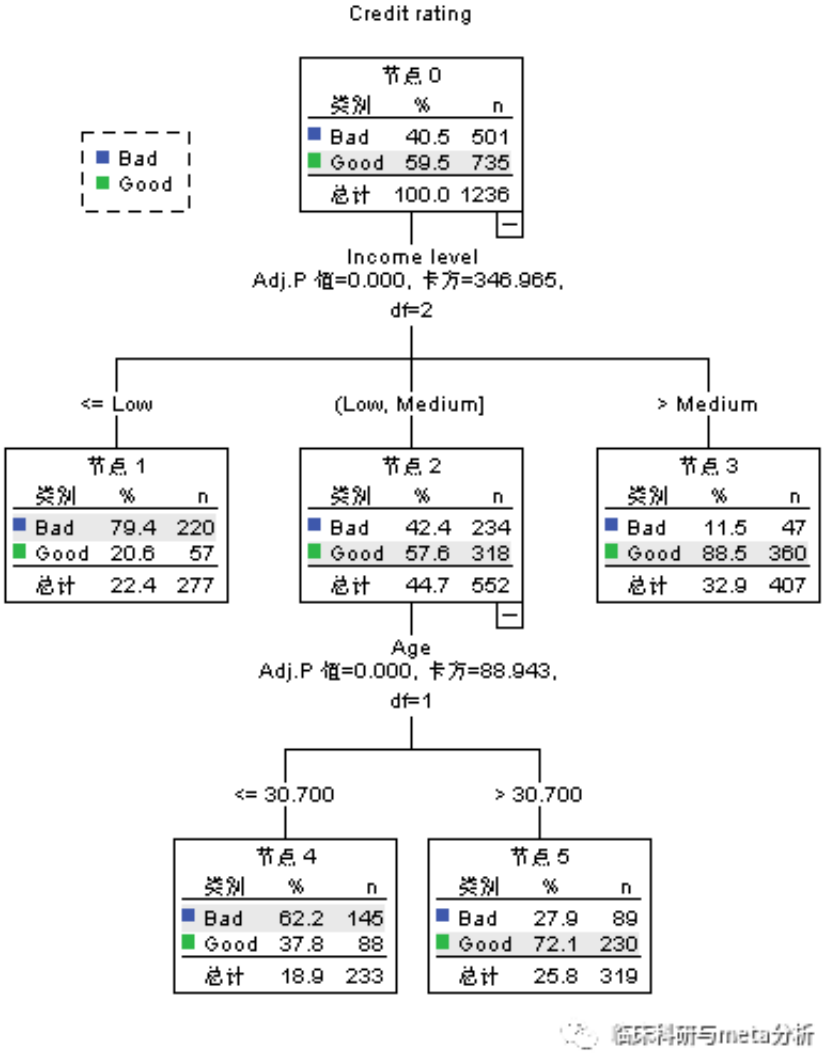
点击保存


变量(V):
因变量(D):
决策树：保存
已保存的变量
终端节点数(T)
预测值(P)
预测概率(R)
样本分配（训练/检验）(S)
将树模型以 XML 格式导出
训练样本(A)
文件(F):
浏览(B)...
检验样本(M)
文件(I):
浏览(O)...
继续(C)
取消
帮助
临床科研与meta分析

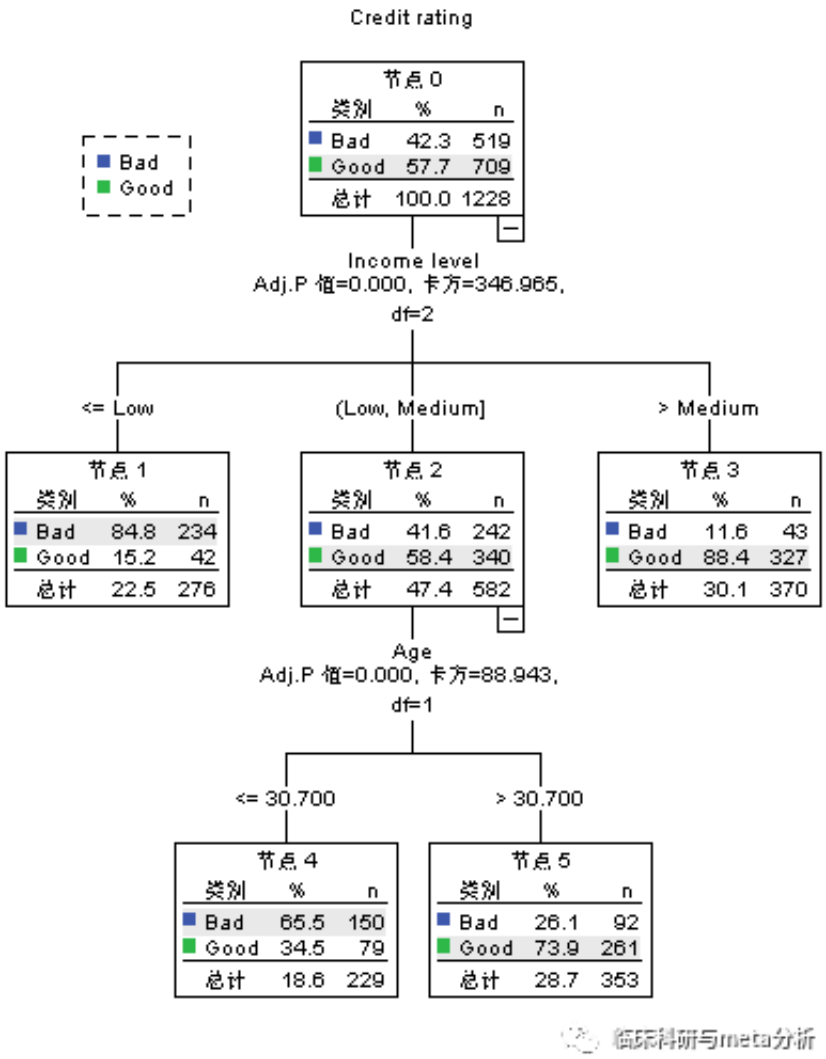
结果

模型摘要对总体模型进行描述

训练样本决策树




验证样本决策树


Credit rating
Bad
Good
节点 0
类别 % n
Bad 42.3 519
Good 57.7 709
总计 100.0 1228
Income level
Adj.P 值=0.000, 卡方=346.965, df=2
<= Low
(Low, Medium]
> Medium
节点 1
类别 % n
Bad 84.8 234
Good 15.2 42
总计 22.5 276
节点 2
类别 % n
Bad 41.6 242
Good 58.4 340
总计 47.4 582
节点 3
类别 % n
Bad 11.6 43
Good 88.4 327
总计 30.1 370
Age
Adj.P 值=0.000, 卡方=88.943, df=1
<= 30.700
> 30.700
节点 4
类别 % n
Bad 65.5 150
Good 34.5 79
总计 18.6 229
节点 5
类别 % n
Bad 26.1 92
Good 73.9 261
总计 28.7 353
临床科研与meta分析

节点增益

## 目标类别：Bad

**节点的增益**

| 样本 | 节点 | 节点 个案数 | 节点 百分比 | 增益 个案数 | 增益 百分比 | 响应 | 指数 |
|---|---|---|---|---|---|---|---|
| 训练 | 1 | 276 | 22.5% | 234 | 45.1% | 84.8% | 200.6% |
| | 4 | 229 | 18.6% | 150 | 28.9% | 65.5% | 155.0% |
| | 5 | 353 | 28.7% | 92 | 17.7% | 26.1% | 61.7% |
| | 3 | 370 | 30.1% | 43 | 8.3% | 11.6% | 27.5% |
| 检验 | 1 | 277 | 22.4% | 220 | 43.9% | 79.4% | 195.9% |
| | 4 | 233 | 18.9% | 145 | 28.9% | 62.2% | 153.5% |
| | 5 | 319 | 25.8% | 89 | 17.8% | 27.9% | 68.8% |
| | 3 | 407 | 32.9% | 47 | 9.4% | 11.5% | 28.5% |

生长法：CHAID
因变量：Credit rating



风险为误判率，分类为具体的分类情况

下面是对**决策树**归纳算法重要特点的总结：

(1)决策树归纳是一种构建分类模型的非参数方法。换句话说，它不要求任何先验假设，不假定类和其他属性服从一定的概率分布。

(2)找到最佳的决策树是NP完全问题。许多决策树算法都采取启发式的方法指导对假设空间的搜索。

(3)已开发的构建决策树技术不需要昂贵的计算代价，即使训练集非常大，也可以快速建立模型。此外，决策树一旦建立，位置样本分类非常快，最坏情况下的时间复杂度是O(w)，其中w是树的最大深度。

(4)决策树相对容易解释，特别是小型的决策树，在很多简单的数据集上，决策树的准确率也可以与其他分类算法相媲美。

(5)决策树是学习离散值函数的典型代表。然而，它不能很好地推广到某些特定的布尔问题。一个著名的例子是奇偶函数，当奇数(偶数)个布尔属性为真时其值为0(1)。对这样的函数准确建模需要一颗具有2d个结点的满决策树，其中d是布尔属性的个数。

(6)决策树算法对于噪声的干扰具有相当好的鲁棒性，采用避免过分拟合的方法之后尤其如此。

(7)冗余属性不会对决策树的准确率造成不利的影响。一个属性如果在数据中它与另一个属性是强相关的，那么它是冗余的。在两个冗余的属性中，如果已经选择其中一个作为用于划分的属性，则另一个将被忽略。然而，如果数据集中含有很多不相关的属性(即对分类任务没有用的属性)，则某些不相关属性可能在树的构造过程中偶然被选中，导致决策树过大庞大。通过在预处理阶段删除不相关属性，特征选择技术能够版主提高决策树的准确率。

(8)由于大多数的决策树算法都采用自顶向下的递归划分方法，因此沿着树向下，记录会越来越少。在叶结点，记录可能太少，对于叶结点代表的类，不能做出具有统计意义的判决，这就是所谓的数据碎片(data
fragmentation)问题。解决该问题的一种可行的方法是，当样本小于某个特定阈值时停止分裂。

(9)子树可能在决策树中重复多次，如图4-19所示，这使得决策树过于复杂，并且可能更难解释。当决策树的每个内部结点都依赖单个属性测试条件时，就会出现这种情形。由于大多数的决策树算法都采用分治划分策略，因此在属性空间的不同部分可以使用相同的测试条件，从而导致子树重复问题。

(10)迄今为止，本章介绍的测试条件每次都只涉及一个属性。这样，可以将决策树的生长过程看成划分属性空间为不相交的区域的过程，直到每个区域都只包含同一类的记录。两个不同类的相邻区域之间的边界称作决策边界(decision boundary)。由于测试条件只涉及单个属性，因此决策边界是直线，即平行于“ 坐标轴 ”，这就限制了决策树对连续属性之间复杂关系建模的表达能力。

斜决策树(oblique decision tree)可以克服以上的局限，因为它允许测试条件涉及多个属性。

尽管这种技术具有更强的表达能力，并且能够产生更紧凑的决策树，但是为给定的结点找出最佳测试条件的计算可能是相当复杂的。

构造归纳(constructive induction)提供另一种将数据划分成齐次非矩形区域的方法，该方法创建符合属性，代表已有属性的算术或逻辑组合。新属性提供了更好的类区分能力，并在决策树归纳之前就增广到数据集中。与斜决策树不同，构造归纳不需要昂贵的花费，因为在构造决策树之前，它只需要一次性地确定属性的所有相关组合。相比之下，在扩展每个内部结点时，斜决策树都需要动态地确定正确的属性组合。然而，构造归纳会产生冗余的属性，因为新创建的属性是已有属性的组合。

(11)研究表明不纯性度量方法的选择对决策树算法的性能的影响很小，这是因为许多度量方法相互之间都是一致的。