文章编号：2096-1618(2018)06-0624-08

基于改进决策树的故障诊断方法研究

赵锦阳[1]， 卢会国[1,2]， 蒋娟萍[1,2]， 罗扬燚[1]

(1. 成都信息工程大学电子工程学院，四川 成都 610225；2. 中国气象局大气探测重点开放实验室，四川 成都 610225)

摘要：为解决设备故障无法定位和无法及时预测的问题，提出一种基于决策树的故障诊断方法。该方法采用基尼系数进行分类树的无偏节点分裂，按照最小代价复杂度剪枝法对生成的决策树进行剪枝，并采用袋装技术建立分类回归树的组合预测模型。最后对空调智能远程控制器故障数据进行分类研究，结果证明了该方法的有效性与可行性。

关 键 词：信号与信息处理；数据挖掘；决策树；故障诊断；基尼系数；剪枝

中图分类号：TP311.13 **文献标志码**：A

doi：10.16836/j.cnki.jcuit.2018.06.005

0 引言

随着科技不断进步，民航新科技不断在大型枢纽机场成功应用。为保证飞机安全进近和着陆，民航近年来投入大量资金建设机场盲降设备系统。当然，再精密的设备也有寿命周期，尤其像机场类似场所对设备性能要求都很高，这就使工作人员要对设备的运行状态做到心中有数，做到提前发现故障设备，从而使故障设备带来的危害降到最低。然而，对众多的设备进行盲目检查，无疑是一项耗时、费力的工作，因此实现机场盲降设备故障诊断是亟待解决的问题。为了对所有的设备进行集中监控和管理，能够及时发现和排除故障设备，提高工作人员对设备的检查效率，机场加大了对盲降设备监控系统的开发[1]。

机场工作人员为了对大量的盲降设备进行集中监控和管理，采用机房监控管理，其监控对象包括动力系统、环境系统、消防系统、网络系统等各个子系统。机场盲降台设备属于精密设备，其设备的正常运行对恶劣天气条件下的飞机着陆起着至关重要的作用。而设备的运行状态除了内部机械性能，还取决于设备所处的工作环境；因此机场管理人员对盲降台机房内的温度和湿度也格外关注。空调设备监控在盲降设备监控中扮演着十分重要的角色，可实时监控机房内的环境变化，使得设备工作在规定环境下。空调控制器出现故障对整个监测系统会造成重大影响，如何对其故障进行及时准确的诊断，具有较高实用价值。以四川开潭科技有限公司在重庆机场使用的空调控制器为例，分析机场采集控制器的state、warning、control return参数的值，这些值成为判断控制器故障类型的重要数据来源[2]。

分类是故障诊断中一个重要的手段，科学界研究出了许多分类算法，有决策树、支持向量机、神经网络等。近几年，分类器在设备故障诊断领域也慢慢发展起来，但是单棵树往往会出现诊断精度不高，易出现过度拟合等问题。

提出采用选择合适的决策树剪枝复杂度参数对控制器进行故障检测，利用采集器得到的特征参数数据分析建立组合预测模型，最后通过投票得出诊断结果。

1 决策树

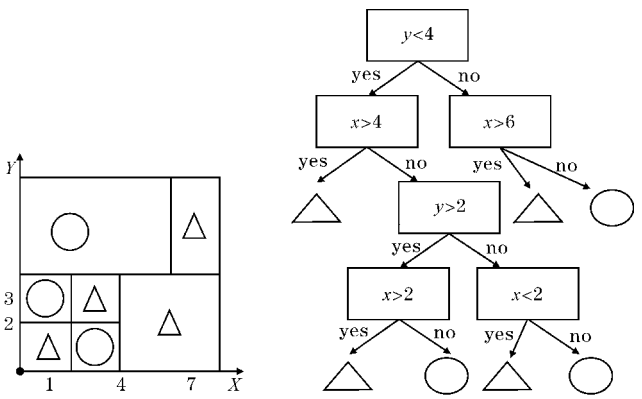
1.1 决策树概述

决策树最早起源于人工智能，其生成最终结果的分析过程如同一棵倒立的树，决策树包含3种重要节点：根节点、中间节点、叶节点。每个节点代表研究事物的属性，而每个分叉路径表示对应的属性值。决策树最上方的节点称为根节点，一棵决策树只有一个根节点。不能产生下级节点的节点称为叶节点，由于决策树是一组合分类器，故其可以有多个叶节点。位于根节点下方且在叶节点上方的节点称为中间节点。若树中每个节点最多只能分裂出两条分支路径，称此决策树为二叉树，若能分裂出不止两条路径，则称其为多叉树[3]。

为更好地理解分类树，可将样本集中的每一个观测样本看成n维特征上的一个点，用圆圈和三角形表示输出变量的类别。从特征空间角度理解，即当完成

收稿日期：2018-03-08
基金项目：四川省教育厅重点科技计划资助项目(14ZA0170)

对 n 维特征空间的区域划分,也就是生成了一棵决策树,而此时的 n 维空间被划分为若干个小矩形区域。如图 1(a)是一个 2 维特征空间划分示例,图 1(b)则为其采用树形方式展示。



(a) 特征空间划分示例　　(b) 决策树

图 1　特征空间划分和决策树

对上述 2 维特征空间继续分析,可知研究的数据集是图形的形状,每个样本的属性为 X、Y,为了表示方便,对每个样本进行编号表示,可知此数据集的内容如表 1 所示。

表 1　数据集

编号	X	Y	所属类别
1	X1:x>4	Y1:y<4	△
2	X2:x>6	Y2:y>4	△
3	X3:x<6	Y2:y>4	○
4	X4:x>2	Y3:y>2	△
5	X5:x<2	Y3:y>2	○
6	X5:x<2	Y4:y<2	△
7	X4:x>2	Y4:y<2	○

1.2　决策树生长过程中的节点分裂算法

决策树无论经过多少次节点分裂,最终只能有一种结果输出,因此可以用于数据的分类和预测。由根节点,产生左右子树时,需要比较不同属性分裂后的结果的优劣,选择最优的属性分裂产生左右子树,这个比较后分裂的过程称之为节点分裂[4]。不同的比较规则就对应不同的决策树生成算法。主要研究 CART 节点分裂算法,此算法是对 CLS 和 ID3 算法的改进[5]。

CART 算法将分裂属性的取值划为两个子集,然后从当前被分成的两个子集出发,计算由训练集决定的 Gini 指标,然后采用递归循环方式,再把当前训练集分成两个子集,从而产生左右两个分枝子树[6]。其计算过程如下:

(1)基尼指数:

$$Gini(K)=1-\sum_{i=1}^{m}P_i^2 \tag{1}$$

其中 P_i 表示当前类别在样本集 K 中出现的概率。

(2)计算当前划分的 Gini 系数:

如果 K 被分割成两个子集 K_1 与 K_2,则当前划分的 Gini 系数为

$$Gini_{\text{split}}(K)=\frac{|K_1|}{|K|}Gini(K_1)+\frac{|K_2|}{|K|}Gini(K_2) \tag{2}$$

CART 是选出最小的 Gini 系数的变量,根据此变量进行节点分裂,最后通过循环的形式,产生决策树。如表 1 所示,属性有两个,每个属性里的值都不同,在决策树的每一个节点可以根据某属性的某个值进行划分。例如分裂方式选择可以如下所示:

(1)X 为 X1 和非 X1;

(2)X 为 X2 和非 X2;

(3)Y 为 Y1 和非 Y1。

对于属性 X 为 X1 时,样本集 K_1 包含 1 个三角形和 0 个圆形;对于属性 X 为非 X1 时,样本集 K_2 包含 3 个三角形和 3 个圆形,如表 2 所示。

表 2　分裂属性选取 X1 和非 X1

	△	○
X1	1	0
非 X1	3	3

由式(1)知,X 属性为 X1 时,其 Gini 指数为

$$Gini(K_1)=1-\sum_{i=1}^{m}P_i^2=1-[1^2+0^2]=0 \tag{3}$$

当 X 属性为非 X1 时,其 Gini 指数为

$$Gini(K_2)=1-\sum_{i=1}^{m}P_i^2=1-\left[\left(\frac{3}{6}\right)^2+\left(\frac{3}{6}\right)^2\right]=\frac{1}{2} \tag{4}$$

可知当前样本集 K 被分割成 K_1 和 K_2,则当前划分的 Gini 系数为

$$Gini_{\text{split}}(K)=\frac{|K_1|}{|K|}Gini(K_1)+\frac{|K_2|}{|K|}Gini(K_2)$$
$$=\frac{1}{7}\times 0+\frac{6}{7}\times\frac{1}{2}=\frac{3}{7} \tag{5}$$

式(5)表示训练集 K 按 X 属性分裂为 X1(K1)和非 X1(K2)两个子集的 Gini 系数为$\frac{3}{7}$,将全部变量进行上述分裂,并根据 Gini 系数最小原则进行比较,得出最佳分裂方式。可见此算法使得决策树的生成过程更加合理,提高了算法的分类精度。

1.3　决策树生长过程中的问题

决策树充分生长后可成长为一棵茂盛的大树,但是由于完整的决策树对训练集样本的描述过于精确,故此棵大树并不是最佳的预测新数据对象树。随着决

策树节点分裂持续进行,势必使后节点的样本量有所减少,从而使得到的决策树很难反映整体数据内部规则。在根节点上进行构建决策树时,处理对象是训练集中的全部训练数据,可知此时样本数最大;当形成第二层分支后,全部数据被分成若干组,而后将根据此分支内各分层的样本产生再下层分枝,此时本分层内的样本数要比第一层根节点的样本数少很多,不断重复此过程生成庞大的一棵树。可见,随着决策树的生长和样本量的不断减少,越下层处的节点所体现的数据特征就越彰显个性化,一般性就越差[7]。

完整的决策树能够准确反映训练样本集中数据的特征,但很可能会具有个性化,无法代表其他众多样本,从而无法用于对新数据的预测,这就是所谓的过度拟合现象。

2 基于改进决策树的故障诊断

分类回归树能很准确的适应学习样本集,但可能会出现过度拟合现象,从而使得模型失去一般代表性。如何提高分类回归树预测的准确性但又不失一般代表性,是需要进一步研究的问题。下面从决策树剪枝和组合预测模型两个方面进行探究。

2.1 决策树的剪枝

决策树修剪技术包括预修剪和后修剪[8]。预修剪技术可以先指定一些参数来控制决策树的生长。这些控制参数通常包括:决策树最大深度、树中节点所包含的最小样本量、树节点 Gini 系数的减少量。当决策树达到控制参数时,达到指定深度或者节点所包含样本数低于指定最小样本量或当前 Gini 系数变化量小于指定值,都会停止生长;后修剪技术则是等到决策树生长到一定程度后再进行统一剪枝[9]。为使生长成的决策树既能包含少的叶节点,又不使预测精度有所降低,可采用二者结合的方法对决策树进行合理的生长控制[10]。

权衡决策树修剪中复杂度和精度之间的关系是得到可靠性模型的关键,通常在得到较低决策树复杂度的同时,又要保持有较高的决策树精度。通常叶节点个数越高,决策树的复杂程度就越大[11]。如果决策树的预测误差看作剪枝带来的代价,用叶节点的个数作为复杂程度的度量,则有决策树 T 的代价复杂度 $R_\alpha(T)$ 定义为

$$R_\alpha(T) = R(T) + \alpha \mid \breve{T} \mid \tag{6}$$

其中:T 表示所包含的所有规则;$R(T)$ 表示 T 在测试集上的预测误差;$\mid \breve{T} \mid$ 表示 T 的叶节点数目;α 为复杂度参数(CP),表示每增加一个叶节点所带来的复杂度。

当 $\alpha = 0$ 时,只有使 $R(T)$ 达到最小,才能使得 $R_\alpha(T)$ 最小,而使得测试集的预测误差最小,只能选择选择叶节点最多的决策树,即决策树最大;当 α 逐渐增大时,$R(T)$ 对 $R_\alpha(T)$ 的影响也随之增加;当 α 足够大时,$R(T)$ 对 $R_\alpha(T)$ 的影响以不占主导作用,此时为了使得生成的决策树不过于庞大,算法选择只有一根节点的决策树。因此应选择恰当的 α,尽量使得误差和复杂度达到最小[12]。

可计算中间节点 $\{t\}$ 及其子树 T_t 的代价复杂度来判断一个中间节点 $\{t\}$ 下的子树 T_t 能否被剪掉。

(1)中间节点 $\{t\}$ 的代价复杂度

$$R_\alpha(\{t\}) = R(t) + \alpha \tag{7}$$

其中:$R(t)$ 为中间节点 $\{t\}$ 在测试样本集上的预测误差;中间节点 $\{t\}$ 的代价复杂度可看作减掉其所有子树 T_t 后的代价复杂度。

(2)中间节点 $\{t\}$ 的子树 T_t 的代价复杂度

$$R_\alpha(T_t) = R(T_t) + \alpha \mid \widetilde{T}_t \mid \tag{8}$$

当中间节点 $\{t\}$ 的代价复杂度大于其子树的代价复杂度,即 $R_\alpha(\{t\}) > R_\alpha(T_t)$ 时,应保留子树 T_t。由式(7)、(8)可知,当 $\alpha < \dfrac{R(t) - R(T_t)}{\mid \widetilde{T}_t \mid - 1}$ 时,应该保留子树 T_t,当 $\alpha \geq \dfrac{R(t) - R(T_t)}{\mid \widetilde{T}_t \mid - 1}$ 时,中间节点 $\{t\}$ 的代价复杂度小于等于子树 T_t,应该剪掉。

2.2 建立分类回归树的组合模型

没有剪枝的决策树,对训练样本具有较高的预测精度,但预测结果会因叶子节点样本的较小变化而出现较大波动,稳定性较低。在统计学中较大波动反映为有较大方差。统计学中有

$$D(\bar{x}_1) + D(\bar{x}_2) + \cdots + D(\bar{x}_k) = \frac{\partial^2}{n} \tag{9}$$

其中:x 表示随机变量,k 表示把总体变量分成 k 个不相关的样本;n 表示 x 总体变量以及每个独立样本包含的样本量,∂^2 表示 x 的总体方差。可知

$$\frac{\partial^2}{n} < \partial^2 \tag{10}$$

由(10)式可推知,一棵决策树给出的预测值方差大于多棵决策树给出的平均预测值方差。因此对分类回归树建立组合模型会提高单棵决策树的预测性能。

多个预测模型是建立在多个样本集合上的,获得

多个样本集合和将多个模型组合起来实现“投票表决”,是组合模型预测中的重要手段,涉及统计抽样技术,主要研究有放回的无权重抽样。

无权重抽样也叫 bagging,重抽样自举法[13]。该方法采用可重复的随机抽样,产生每个样本集。Bagging 是将已有的分类算法通过一定方式组合起来,形成一个性能更加强大的分类器。其主要采用 Bootstrap 方法,是一种有放回的抽样方法[14]。

首先,从初始样本集中抽取训练样本,每次从初始样本集中使用 Bootstrap(有放回抽样)方法抽取 m 个训练样本,进行 k 次抽取,得到 k 个相互独立的训练集。由于是有放回抽样,故产生的训练集中可能有重复样本[15]。

其次,根据上述所讨论产生决策树的方法可知,有 k 个训练集可得到 k 个决策树模型。

最后,将上述生成的 k 个决策树模型采用投票的方式,即“少数服从多数”的原则,得到分类结果。过程如图 2 所示。

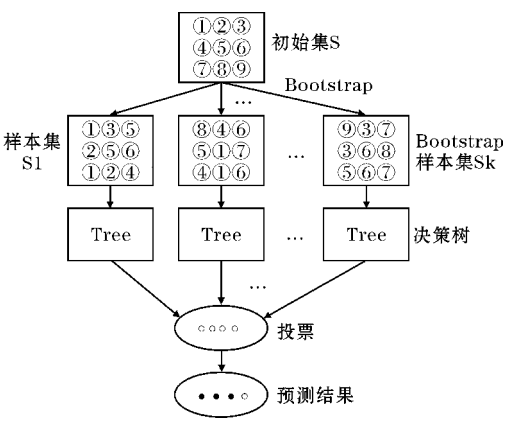


图 2　基于决策树为分类器的 bagging 原理图

3　实验与分析

机场盲降设备系统所使用的空调主要服务对象为盲降台设备,为盲降台机房提供稳定可靠的工作温度、相对湿度,从而使得整个监控系统能正常工作。为了对无人值守的盲降台设备的工作环境进行实时监测和控制,盲降设备机房往往配置空调控制器,以此方便值班室实时掌握房间的温度、湿度;若空调控制器发生故障,会对整个监控系统造成影响。

3.1　实验数据

根据四川开潭科技有限公司对重庆江北机场所做的机场盲降监控系统,汇总出空调控制器故障数据,开潭科技有限公司所使用的空调控制器的故障可以通过其 4 个参数进行诊断:current、state、warning、control return。其中 state 表示空调控制器中所存储的空调状态(为了方便描述空调控制器中所存储的空调状态与实际空调状态的关系,用 0 表示状态不符,1 表示状态吻合)。故障诊断就是利用这些有限的参数所提供的特征信息来确定空调控制器的故障状态的。选取空调控制器故障中的 8 个典型特征作为学习样本输入,如表 3所示。

表 3　空调控制器典型故障

序号	故障名称	故障现象	故障分析
1	空调供电线路故障	控制器上电后电源指示灯不亮,控制器失效	没有电源输入或者输入电压不足
2	工作状态出现故障	空调出现工作异常	空调工作状态的配置不符合要求
3	空调制冷故障	空调控制器显示 F0	回风口与出风口温度传感器安装错误
4	空调制热故障	空调控制器显示 F1	空调故障不能制热
5	电流传感器故障	空调控制器显示 F3	空调控制器的硬件损坏
6	回风温度传感器故障	空调控制器显示 F4	回风口温度传感器可能脱落或者损坏
7	出风温度传感器故障	空调控制器显示 F5	出风口温度传感器可能脱落或者损坏
8	空调红外发射头故障	空调控制器无法控制空调	红外发射头正负极可能接反或者损坏

用 80 组训练样本和 40 组测试样本进行学习和预测,其分布如表 4 所示。

表 4　实验数据样本集

故障类别	训练样本	测试样本
1	10	5
2	10	5
3	10	5
4	10	5
5	10	5
6	10	5
7	10	5
8	10	5

3.2　故障诊断结果

3.2.1　剪枝的决策树预测结果

对 80 组训练样本创建分类树进行训练学习,其中决策树的异质性指标使用 Gini 系数。为了得到茂盛的决策树,首先让决策树的修剪复杂度最小($CP=0$)。其模型图如图 3 所示。

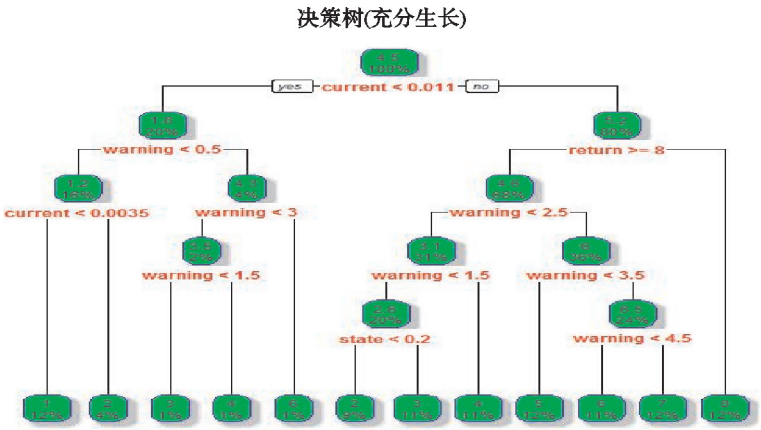


图3 剪枝前决策树

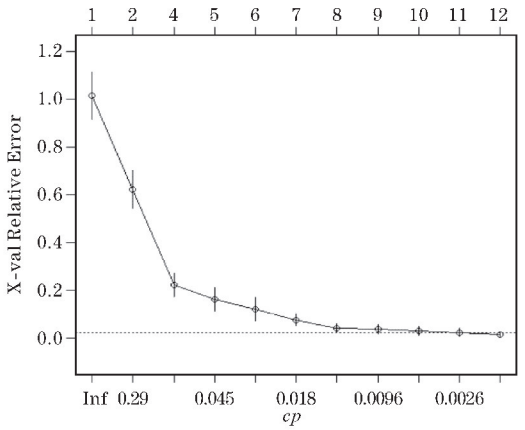


图4 决策树复杂度及误差折线图

由上述理论分析知,未经剪枝的决策树会充分生长,其产生依赖于训练样本集;但为了使产生的决策树具有代表性,对训练样本集不过分依赖(过拟合),进行相应的剪枝必不可少,从而使得到的预测模型对任何测试样本集都有较高正确率的预测。由以上可知,决策树剪枝的参数称为复杂度参数(CP),可得出决策树的预测误差估计值与复杂度参数、叶节点关系如图4所示;并得到相应复杂度下的决策树,用不同复杂度的决策树对40组测试样本进行分类预测,其具体预测结果如图5所示。

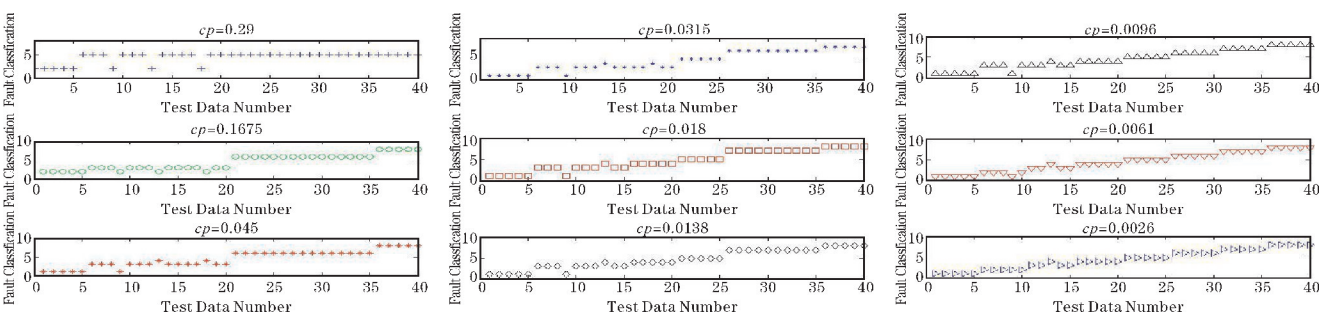


图5 各复杂度下决策树的分类结果

40组测试集是按照故障顺序进行排列添加的,由图5各复杂度下决策树的分类结果可知,当 $CP=0.29$ 时,有34个被误判;当 $CP=0.1675$ 时,有25个被误判;当 $CP=0.045$ 时,有20个被误判;当 $CP=0.0315$ 时,有15个被误判;当 $CP=0.018$ 时,测试样本有11个被误判;当 $CP=0.0138$ 时,测试样本有11个被误判;当 $CP=0.0096$ 时,测试样本有6个被误判,分别为样本6,样本7,样本8,样本9,样本10和样本13;当 $CP=0.0061$ 时,测试样本有2个被误判,分别为样本9和样本13;当 $CP=0.0026$ 时,测试样本有1个被误判,为样本13。故由图5可得相应复杂度下决策树预测准确率如表5所示。

故障类别为顺序加入40组测试样本,为防止生成的决策树过度拟合,一般会对茂盛的决策树进行剪枝,由上图5和表5可知,决策树预测结果的可靠性会随剪枝的复杂度参数的变化而变化,所以选取合适的剪枝复杂度是得到可靠性模型的前提。由以上可知,当复杂度 $CP=0.0026$ 时,决策树测试结果中只有1个样本被误判,此种情况相对于其他情况不仅有最高的预测正确率,而且其修剪复杂度也适当小;此时的决策树如图6所示,与图3剪枝前决策树相比,可以看出其实现了对充分生长的决策树进行合理剪枝,避免其预测结果产生过度拟合现象。

表5 各复杂度下决策树对40组测试集的正确率

复杂度 CP	正确率 Accuracy/%
0.2900	15
0.1675	37.5
0.0450	50
0.0315	62.5
0.0180	72.5
0.0138	72.5
0.0096	85
0.0061	95
0.0026	97.5

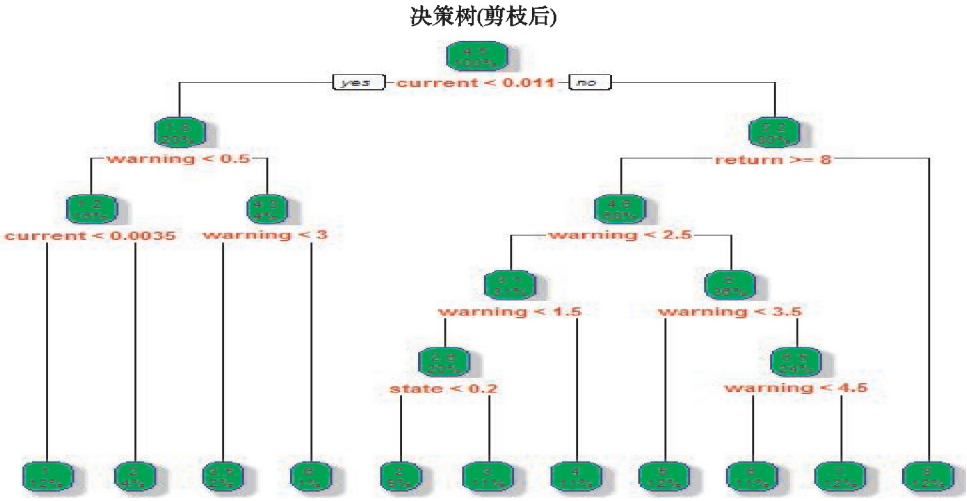


图 6　剪枝后的决策树

3.2.2　组合模型的预测结果

由上述理论分析知,组合模型的预测结果与重抽样自举次数有关,即组合模型不是单一的决策树,而是由许多决策树组合而成,即进行了多少次重抽样自举,就是多少棵决策树相组合。用同样的 80 组训练样本进行学习和 40 组预测样本进行预测,其组合分类树个数和剪枝复杂度相对应的预测结果正确率如图 7 和表 6 所示。其中图 7 组合分类树个数和剪枝复杂度相对应的预测结果正确率的三维图,从三维图的不同视角进行观察可定性的得出最佳组合个数和剪枝复杂度。

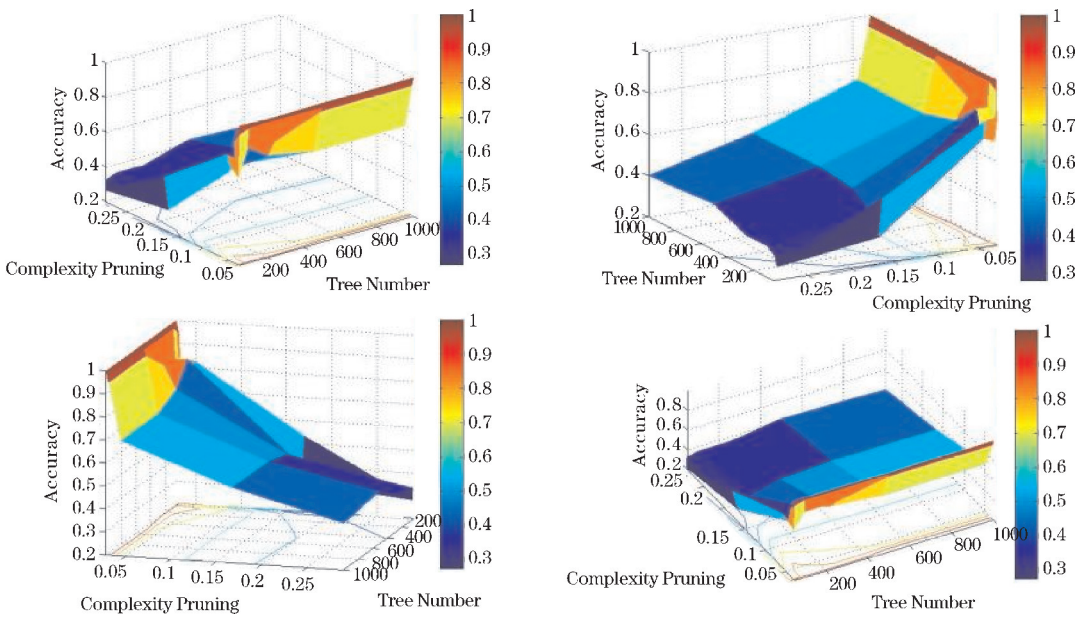


图 7　组合分类树预测结果

从图 7 和表 6 可以看出,组合分类树预测正确率会随着组合数的增加而增加,但此种情况并不会一直持续下去,对于空调控制器故障数据而言,组合数在 250 左右可使分类结果正确率达到最大;与单个决策树对比,组合分类模型有更宽的剪枝复杂度范围,即选择合适的组合数之后,复杂度选择0.029以下就可以有很高的预测正确率,比单个决策树更具有选择灵活性和可靠性。文中选择组合分类树模型生成 250 棵分类树,剪枝复杂度选择0.029,能使预测分类结果完全正确。

实验表明:对单棵决策树进行剪枝,不仅能维持决策树分类准确率而且能避免过度拟合,但过分剪枝会使得决策树准确度有所降低;而组合分类树会弥补过多剪枝带来的准确度问题,从而间接说明决策树的组合预测模型比单个决策树更具有分类优势。但决策树的组合数和剪枝参数都会影响预测结果,如何权衡二者之间的关系,是得到可靠预测模型的前提,所以在建立预测模型前选择合适的参数是成功建立可靠模型的关键。

表6 组合模型预测正确率

分类树组合/棵	剪枝复杂度	正确率/%	分类树组合/棵	剪枝复杂度	正确率/%
25	0.2900	27.5	250	0.2900	35
	0.1675	52.5		0.1675	50
	0.0450	82.5		0.0450	75
	0.0315	82.5		0.0315	85
	0.0300	92.5		0.0300	97.5
	0.0290	92.5		0.0290	100
50	0.2900	32.5	500	0.2900	42.5
	0.1675	32.5		0.1675	52.5
	0.0450	72.5		0.0450	70
	0.0315	70		0.0315	95
	0.0300	95		0.0300	97.5
	0.0290	100		0.0290	100
100	0.2900	35	1000	0.2900	40
	0.1675	42.5		0.1675	52.5
	0.0450	85		0.0450	70
	0.0315	85		0.0315	95
	0.0300	97.5		0.0300	97.5
	0.0290	100		0.0290	100

4 结束语

提出用改进决策树的方法利用空调控制器中各种特征参数的数据建立预测模型,对空调控制器故障进行诊断,实验证明预测模型准确度比单个决策树分类器有所提高,利用此方法构建的组合分类器能对机场盲降台空调控制器故障做出合理判断,提高机场维修人员对设备检查的效率,在解决机房空调控制器故障诊断领域具有广阔的应用前景。

参考文献:

[1] 陈增杰,余世明,雷霞. 机场导航台设备监控系统的改造[J]. 信息技术,2010(5).

[2] 朱文发,吴浩,郑树彬,等. 一种地铁车辆空调温度控制器故障检测平台的设计[J]. 城市轨道交通研究,2014,17(6).

[3] Breiman L I, Friedman J H, Olshen R A, et al. Classification and Regression Trees(CART)[J]. Biometrics,1984,40(3).

[4] 钟龙申. 随机森林算法处理不平衡数据的改进及其并行化[D]. 广州:广东工业大学,2016.

[5] 曹正凤. 随机森林算法优化研究[D]. 北京:首都经济贸易大学,2014.

[6] Mehta M, Agrawal R, Rissancn J. SLIQ: A Fast Scalable Classifier for Data Mining[R]. IBM Almaden Research Center, San Jose, California, 1995.

[7] 薛薇. R 语言数据挖掘[M]. 北京:中国人民大学出版社,2016.

[8] Breslow L A, Aha D W. Simplifying decision trees: a survey [J]. Knowledge Engineering Review, 1997,12(1):1-40.

[9] 项婧,任劼. 决策树分类器在分析基因微阵列数据中的应用[J]. 计算机工程与设计,2006,27(15).

[10] 苑擎飏. 基于决策树中文文本分类技术的研究与实现[D]. 沈阳:东北大学,2008.

[11] 王鑫. 基于 FRMI 的有序决策树算法及其比较研究[D]. 保定:河北大学,2014.

[12] 吴喜之,马景义,吕晓玲,等. 数据挖掘前沿问题[M]. 北京:中国统计出版社,2009:41-71.

[13] L Breiman. Bagging predictors[J]. Mach. Learn, 1996,24(2):123-140.

[14] Dietterich T. An Experimental Comparison of Three Methods for Constructing Ensembles of De-

cision Trees: Bagging, Boosting and Randomization[J]. Machine Learning,2000,40(2).

[15] 曹博宇. 一种基于密度的改进决策树算法[D]. 大连:大连理工大学,2016.

Research on Fault Diagnosis Method based on Improved Decision Tree

ZHAO Jin-yang[1], LU Hui-guo[1,2], JIANG Juan-ping[1,2], LUO Yang-yi

(1. College of Electronic Engineering, Chengdu University of Information Technology, Chengdu 610225, China; 2. Key Laboratory of Atmospheric Sounding of CMA, Chengdu 610225, China)

Abstract: In order to solve the problems that equipment failure can't be located and the fault of equipment can't be predicted in time, a fault diagnosis method based on decision tree is proposed. The Gini coefficient is used to classify the unbiased node of the classification tree. The decision tree is pruned according to the minimum cost complexity pruning method, and the combination forecasting model of classification and regression tree is established by bagging technology. At last, the fault data of air conditioning intelligent remote controller are classified and studied. The results prove the effectiveness and feasibility of the method.

Keywords: signal and information processing; data mining; decision tree; fault diagnosis; Gini coefficient; prune