【统计应用研究】

基于 R 软件 rpart 包的分类与回归树应用

谢益辉
(中国人民大学 统计学院, 北京 100872)

摘要: 对于许多分类和回归问题, 二叉树(Binary Tree)提供了有趣而又形象化的方式来研究数据, 它主要是按照一定的规则拆分自变量, 而完成对因变量的合理分类, 进一步可以对未知分类进行预测。在主要介绍递归分割(Recursive Partitioning)和回归树(Regression Tree)在 R 软件中应用的同时, 对一前列腺癌数据使用生存分析和分类与回归树相结合的方法做出分析, 并得到了对于疾病诊断和预防较有指导意义的结论。

关键词: 递归分割; 分类与回归树; 生存分析; R 软件
中图分类号: C819 文献标识码: A 文章编号: 1007- 3116(2007)05- 0067- 04

一、分类与回归树简介

“分类与回归树”早期被称为“决策树”, 而决策树的自动构建可以追溯到 Morgan 与 Sonquist (1963)和 Morgan 与 Messenger(1973)的社会科学研究。而统计学领域的开山之作当属 Breiman 等人[1], 大约与此同时, 树的方法也在各个领域被广泛使用, 如机器学习[Quinlan, (1979)、(1983)、(1986)、(1993)]、工程学[Henrichon 与 Fu, (1969); Sethi 与 Sarvarayudu, (1982)]等。近年来, 这些方法的发展越来越集中于机器学习领域, 而统计学方面的进展则相当少。软件方面, Therneau 和 Atkinson (1997)的 rpart 库(在 S- Plus 软件中, 后被 Ripley 引入 R 软件的同名附加包 rpart)提供了实现快速计算以及封装好的一系列 S 函数[2]。

树的构建不妨可以看作是一个变量选择的过程, 所有问题基本可以归结为两点: 选择哪个变量作为拆分(节点); 以及如何拆分, 即拆分的规则。在生成树之后, 还面临着一个剪枝(prune)的问题。关于分类与回归树的深入理论, 可以参考 Breiman 等人的著作, 在下文的应用部分中仅仅简略描述部分理论。

递归分割, 顾名思义也就是对变量进行逐层分隔, 直到分割结果满足某种条件才停下来, 这里“分割的结果”可能是得到一些分类值, 也可能是一些描述统计量或预测值。分类树用于因变量为分类数据的情况, 树的末端为因变量的分类值; 回归树则可以用于因变量为连续变量的情况, 树的末端可以给出相应类别中的因变量描述或预测。在不引起混淆的情况下, 下文中“回归树”是指广义上的树方法, 与“分类树”不加区分。

二、R 软件中的回归树应用

R 是一种开源、免费的优秀统计软件, 官方网站在 http://www.r- project.org, 下载镜像(CRAN)为 http://cran.r- project.org; 诸多统计学前沿方法都能以最快的速度在 R 中得到计算机实现, 其中 rpart 包是官方推荐的一个包, 它的功能就是实现递归分割和回归树。

(一) rpart 包介绍及主要函数说明

rpart 不是 R 默认安装的包, 使用时需要从 CRAN 上下载安装。其中包括的函数比较少, 以 rpart() 和 prune() 两个函数为主, 前者是用来拟合一个树模型, 后者用来根据“成本复杂性”对生成的树进行剪枝。之所以要剪枝, 是因为若不加任何限制, 最后生成的树必然能完全拟合原始数据, 这样的树在实际应用中毫无意义, 因为树的枝节太多, 而不能反映数据内在大规律; 而从另一个极端情况来看,

收稿日期: 2007- 03- 03

作者简介: 谢益辉(1984-), 男, 湖北省宜昌人, 硕士生, 研究方向: 统计软件, 机器学习。

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

若树的枝节太少,那么必然也会带来很大的预测误差。综合看来,要兼顾树的规模和误差的大小,因此通常采用一个叫"成本复杂性"的标准来对树进行限制,最后达到的目的是使误差和数的规模都尽可能小。误差的计算通常基于交叉验证等方法,即用一部分训练样本建立模型,而剩下的样本用来作验证看模型的预测误差大小。下面看 R 的函数实现[3-4]:

1. 生成树: rpart()函数

函数用法:

rpart(formula, data, weights, subset, na.action = na.rpart, method, model= FALSE, x= FALSE, y= TRUE, parms, control, cost, ...)

主要参数说明:

fomula 回归方程形式:例如 y~ x1+ x2+ x3。

data 数据:包含前面方程中变量的数据框(data frame)。

na.action 缺失数据的处理办法:默认办法是删除因变量缺失的观测而保留自变量缺失的观测。

method 根据树末端的数据类型选择相应变量分割方法,本参数有四种取值:连续型⇒"anova";离散型⇒"class";计数型(泊松过程)⇒"poisson";生存分析型⇒"exp"。程序会根据因变量的类型自动选择方法,但一般情况下最好还是指明本参数,以便让程序清楚做哪一种树模型。

parms 用来设置三个参数:先验概率、损失矩阵、分类纯度的度量方法。

control 控制每个节点上的最小样本量、交叉验证的次数、复杂性参量:即 cp: complexity pamemeter,这个参数意味着对每一步拆分,模型的拟合优度必须提高的程度,等等。

2. 剪枝: prune()函数

函数用法:

prune(tree, ...)　　prune(tree, cp, ...)

主要参数说明:

tree 一个回归树对象,常是 rpart()的结果对象。

cp 复杂性参量,指定剪枝采用的阈值。

(二) 应用示例与解释

这里选取一个相对比较特殊的例子以显示回归树的多种用途(数据来源: http://www.stanford.edu/class/stats202/DATA/stagec.data);例子的研究目的是结合生存分析的方法寻找影响前列腺癌复发的因素,因变量为前列腺癌复发的时间间隔 pgtime,相应的生存分析状态变量为 pgstat(复发 1 或删失 0),自变量有病人的年龄 age、肿瘤等级 grade(取值 1~ 4)、处于 $G2$ 阶段的瘤细胞比例等 $g2$, 关于数据的详细描述参见。

因为这里用的是生存分析的方法,所以 rpart()函数中的因变量要使用生存分析对象(survival object),可以用 survival 包中的 Surv()函数生成,树模型方法参数 method 则应该使用"*exp*"方法(若不指明,程序也会自动挑选它)。首先产生一个树的拟合:

```
> library(rpart)
> library(survival)
> fit < - rpart(Surv(pgtime, pgstat) ~ age+
    eet+ g2+ grade+ gleason+ ploidy, data=
    stagec, method="exp")
```

然后将拟合结果用 print()函数打印在屏幕上,结果如下:

```
> print(fit)
n= 146

node) split n deviance yval
 * denotes terminal node
  1) root 146 195.411 600 1.000 000 0
  2) grade< 2.5 61 45.021 520 0.362 470 1
  4) g2< 11.36 33 9.120 116 0.122 556 2*
  5) g2> = 11.36 28 27.804 100 0.733 529 8
 10) gleason< 5.5 20 14.376 900 0.529 219 0*
 11) gleason> = 5.5 8 11.201 470 1.308 368 0*
  3) grade> = 2.5 85 125.327 400 1.619 062 0
  6) age> = 56.5 75 104.154 700 1.428 731 0
 12) gleason< 7.5 50 66.701 410 1.143 132 0*
 13) gleason> = 7.5 25 33.993 130 2.035 522 0
 26) g2> = 15.29 13 16.555 970 1.349 474 0*
 27) g2< 15.29 12 14.220 260 2.921 048 0*
  7) age< 56.5 10 15.522 810 3.197 743 0*
```

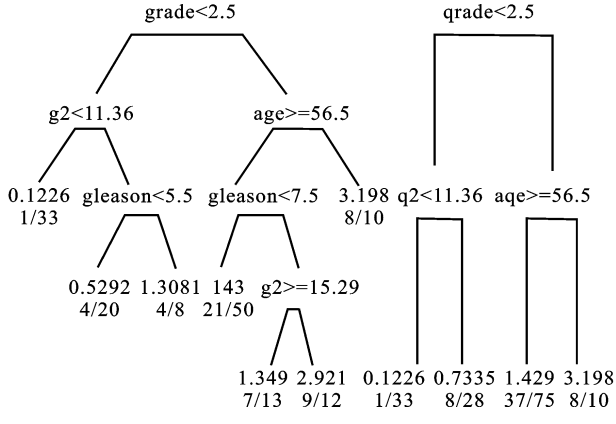

图 1　原始回归树(左)和剪枝后的回归树(右)图

从上面的文本中可以看出每次分类的节点,以及节点中样本量的大小、残差平方和等信息,当然图

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

形是更直观的展示方法, 如图 1(左) 。每次分类情况的信息都在树中展示出来了(比如最下端的 a/b 表示前列腺癌复发的比率) 。

上文提到过, 太大的树会导致种种实际应用中的问题, 所以为了进一步的剪枝, 可以先看看这棵树的复杂性参量表, 它可以通过 printcp() 函数从拟合对象 fit 中获取, 也可以提取 fit 中的对象 cptable, 如以下代码所示:

```
> fit$ cptable
CP      nsplit      rel      error      xerror      xstd
1    0.128 3    0    1.000 0    1.015 6    0.074 1
2    0.041 4    1    0.871 7    0.936 7    0.078 0
3    0.028 9    2    0.830 3    0.967 4    0.083 8
4    0.017 7    3    0.801 4    0.980 4    0.087 2
5    0.016 5    4    0.783 7    1.019 8    0.092 4
6    0.011 4    5    0.767 2    1.028 6    0.095 6
7    0.010 0    6    0.755 8    1.050 5    0.096 5
> plot( fit, uniform = T, branch= 0.4, compress= T)
> text(fit, use. n = T)
> fit2 < - prune( fit, cp = 0.017)
> plot( fit2)
> text(fit2, use. n = T)
```

在剪枝理论中, 比较著名的规则就是 1- SE(1 标准差) 规则, 其意思是: 首先要保证预测误差(通过交叉验证获得, 在程序中表示为 xerror) 尽量小, 但不一定要取最小值, 而是允许它在“最小的误差±一个相应标准差”的范围内, 然后在此范围内选取尽量小的复杂性参量, 进而以它为依据进行剪枝。这个规则体现了兼顾树的规模(复杂性) 和误差大小的思想, 因为一般说来, 随着拆分的增多, 复杂性参量会单调下降(纯度越来越高) , 但是预测误差则会先降后升, 这样, 就无法使复杂性和误差同时降到最低, 因此允许误差可以在一个标准差内波动。

从上面的复杂性参量表中不难发现: 最小的误差是第 2 行, xerror= 0.936 7, 相应标准差为 xstd= 0.078 0, 相加得 1.015, 从而树的最大节点数只能为 3(参见第 4、5 两行的 xerror) , 相应复杂性参量 *cp* 必须大于 0.016 462(本例中选为 0.017) 。剪枝后的树参见图 1(下) , 相应程序代码如下:

```
> newgrp < - fit2$ where
> plot( survfit( Surv( pgtime, pgstat) ~ newgrp,
  data = stagec) , mark. time = F, lty = 1: 4)
> title( xlab = “Time to Progression”,
  ylab = “Prob Progression”)
> legend( 0.2, 0.2, legend = paste(“node”, c
  ( 4, 5, 6, 7) ) , lty= 1: 4, bty= “n”)
```

最后, 为了查看四个节点上的人群前列腺癌复发的生存函数图, 可以从 $fit2$ 中获取原始数据中每个个体所处的节点位置, 也就是根据节点对原始数据产生了一个分组的变量。通过对不同的组进行生存函数拟合, 然后作图(见图 2) 。图 2 中的四个节点分别与图 1 中从左至右的四个叶节点对应, 该图表明: 本例中的四组人群前列腺癌的存活概率和存活时间(此处的“存活”是生存分析中的一般用语, 在此指的是前列腺癌没有复发的情况) 是有差异的, 其中节点 7 的情况最不理想, 生存概率小、时间短, 即前列腺癌最容易复发, 而这一组人群的特征就是肿瘤等级高 grade ⩾2.5 且年龄相对小 age< 56.5。这对于疾病诊断和预防是很有指导意义的结论。

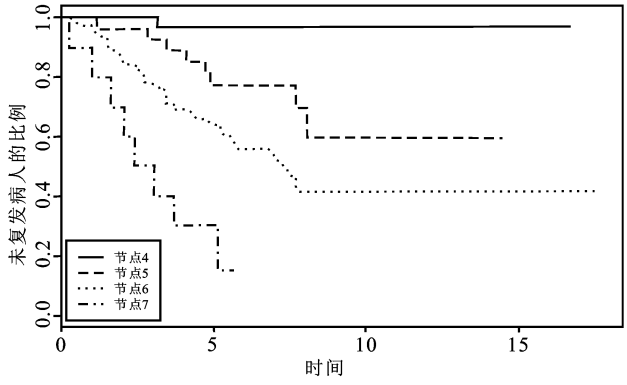

图 2　不同节点的生存函数图

三、总　结

无论从管理决策还是统计方法的角度来看, 分类与回归树都有很多自身的优点, 一个比较显而易见的优势就是它特别直观明了, 决策者根据树形图的分枝走向很容易预测未知因变量的取值, 这一点与神经网络(这种方法的权重以及预测机制往往让人觉得很不透明) 等预测方法相比, 逻辑和操作都更易理解; 另一方面, 它对缺失值的处理机制能提高数据的利用充分程度, 对于缺失数据, 常用的处理办法就是删除含缺失的观测, 但是分类与回归树一般只是删除因变量缺失的观测, 而保留有自变量缺失的观测, 而且它还具有较好的稳健性, 因为划分的依据是观测值的顺序而不是数值的具体大小, 这样即使数据中出现异常值, 对于结果也不会造成太大的影响; 还有一个好的性质就是, 对原始变量作任何单调变换也都不会影响结果。与单纯的回归相比, 分类与回归树也省去了选择变量的麻烦, 因为在构建树的过程中它会自动选择最优划分变量。

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

当然分类与回归树也并非最完美的统计方法,但是,随着时间推移,各种方法的交叉运用也使它的缺陷在逐渐改善,例如预测的准确度问题,可以通过Bagging(Breiman,1996)等方法来提高。

在软件方面,除了本文讲到的 rpart 包之外,R 中还有关于递归分割更为详细的包 party,它包含了 Bagging 方法,可以产生条件推断树(conditional inference tree)等,而 Breiman 的 randomForest 包则实现了分类与回归树的随机森林(random forest)算法。

参考文献:

[1] BREIMA NL, FRIEDMANJ, OLSHENR, STONEC. Classification and Regression Trees[M]. Wadsworth,1984.

[2] VENABLES W N, RIPLEY B D. Modern Applied Statistics with S[J]. Springer, edition, 2002(4):251-269, URL http://www.stats.ox.ac.uk/pub/MASS4/, ISBN 0387-95457-0.

[3] BRIAN S Everitt, TORSTEN Hothorn. A Handbook of Statistical Analyses Using R[S]. Chapman & Hall/CRC, 2006:131-142.

[4] THERNEAU T M, ATKINSON E J. An introduction to recursive partitioning using the rpart routine[J]. Technical Report 61, Mayo Clinic. Section of Statistics, 1997.

(责任编辑:郭诗梦)

The Application of the Classification and Regression Tree Based on the Package rpart in R-Lamguage

XIE Yi-hui

(School of Statistics, Renmin University of China, Beijing 100872, China)

Abstract: For many problems concerning classification and regression, the method of "binary tree" has provided an interesting visualization approach in research. The basic idea of such methods is to split the response variable according to certain rules, thus we can get a reasonable classification of the response variable, and make prediction to the unknown classifications based on new samples. This paper mainly introduces the implementation of recursive partitioning and regression tree in the package rpart of R language, then makes an analysis to a medical data using classification and regression tree and survival analysis, and finally gets some useful instructions on the diagnosis and prevention of illness.

Key words: recursive partitioning; classification and regression tree; survival analysis; R language

©1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net