分类回归树算法的探讨

梁 茵
(广东技术师范学院,广东 广州 510665)

摘 要:分类回归树是一种优良的决策树算法,有广泛的应用。本文探讨了分类回归树算法及应用,首先回顾了分类回归树的起源及应用,其次分析了分类回归树在均匀成本和非均匀成本下的构造,接着讨论了分类回归树的剪枝和验证过程,最后我们对其进行了总结。

关键字:分类回归树;构造;剪枝
中图分类号:TN 915.16 文献标识码:A 文章编号:1672 - 402X(2008)06 - 0029 - 04

1 引 言

决策树是数据挖掘的一种重要算法,而数据挖掘就是从大量、有噪声、模糊、随机的数据中,提取隐含在其中的、事先不被人们知道、有潜在作用的信息和知识的过程。

分类是数据挖掘应用得最多的一项。分类方法的好坏可从三个方面进行判别:(1)预测准确度;(2)计算复杂度;(3)模式的简洁度。

决策树是以事例(case)为基础的归纳学习算法,利用一系列规则,从一组无次序、无规则的事例中建立树状图用于分类与预测,常用的决策树方法有分类回归树(Classification and Regression Tree, CART)、卡方自动交互检验法(Chi- square Automatic Interaction Detector, CHAID)、等[1]。本文探讨其中的分类回归树方法。

CART 方法是由 Breiman 等人在 1984 年提出的一种决策树分类方法[2]。其采用基于最小距离的基尼指数估计函数,这是因为基尼指数可以单独考虑子数据集中类属性的分布情况,用来决定由该子数据集生成的决策树的拓展形状。CART 创建简单二叉树结构对新事例进行分类,这样可以有效地处理缺失数据,尤其对于分类与预测时更好。并且 CART 方法中有贝叶斯分类的特征,使用者可以提供主观的分类先验概率作为选择分类的权重,则 CART 在获得最终选择树前使用交叉检验来评估候选树的误分类率,这对分析复杂样本数据非常有用。CART 处理离散变量与连续变量同样容易,这是由于它使用了或形状的几乎不依靠无关变量的分支。而且,被 CART 考虑到的分支在任何单调转换下是不变的,如对一个或更多的特征取对数、平方根等都是不变的。

Rousu 等人[3]和 Moisen 与 Frescino 等人[4]分别分析了 CART、神经网络、logistic 回归等决策树方法效率的问题,每个方法都从全部特征值开始预测具有有限个分类的响应变量,产生一个可以定义类规则的较小特征值。Markham 等人[5]分析了使用 CART 和神经网络的实时 Kanban 生产系统,他们发现两个方法在准备性和响应速度方面相差无几,但 CART 具有可解释性和开发速度方面的优势。

2 分类回归树的构造

决策树的构造即决策树的测试属性选择。决策树从根结点开始采用自顶向下的(Top- down)的递归方式在每个结点上对样本集按照给定标准选择分支属性,然后按照相应属性的所有可能取值向下建立分支、划分训练样本,直到一个结点上的所有样本都被划分到同一个类,或者某一结点中的样本数量低于给定值时为止。这一阶段关键的是在树的结点上选择可以将训练样本进行最优划分的分支属性,选择分支属性的标准有信息增益(InformationGain)、熵(Entropy)和基尼指数(GiniIndex)等。

CART (Classification and Regression Tree, CART)二叉树由根结点,中间结点和叶(终)结点组成。每个

收稿日期:2008- 03- 05
作者简介:梁茵(1981-),女,海南人,广东技术师范学院助教,研究方向:计算机应用研究。

根结点和中间结点是具有2个子结点的父结点。每个结点,如,被它包括的子集初始学习样本描述。除终结点外,这个子集被分为两组,为和子结点。每个结点的分支被一个依赖于已选特征规则描述, 令这个特征规则为,并假设为连续型。对于一些常量,分支的形式是或。如果是分类型,分支的形式为或,此时为可能类的某些非空子集。特征在所有可能特征中选择,并且在所有可能的分支中选择,其目的是在两个子结点中减少产生子样本的复杂度。

CART提供了3种可能的分支方法: 熵(Entropy)、基尼指数(GiniIndex)和"二分法"(Twoing function)方法。每个方法都可能与误分类成本C(i/j)一起使用,C(i/j)就是把事实上属于j类的个体分到i类的成本,$C(i/j)\geq 0$, $C(i/i)=0$。为适应特定的应用,CART使用者需要选择具有高度弹性不同水平的误分类成本C(i/j)。

分支规则来源于复杂度函数(BFOS称为不纯度函数)。令$p(j/t)$, $0\leq p(j/t)\leq 1$, $j=1,2,\cdots,J$, $p(j/t)$为树中结点t处j类个体的比例。J为类的个数。所以,对于每一个结点:

$$\sum_{j=1}^{J} p(j/t)=1 \tag{1}$$

现在我们来介绍结点t处CART使用的三个主要的复杂度函数。我们应该区分两个不同的情况。第一种情况,对于任何(分类)项目的误分类成本,无论它是否为真实的类和无论它被误分进入哪个类,都是均匀的。第二种情况,把一个事实上属于j类的个体误分类进入i类的成本,记为C(i/j),它可能依赖于i和j。

(1)均匀成本下的熵函数如式(2)所示:

$$d_E(t)=-\sum_{j=1}^{J} p(j/t)\log[p(j/t)] \tag{2}$$

非均匀成本下的熵函数如式(3)所示:

$$d_E(t)=-\sum_{j=1}^{J}\sum_{i=1,j\neq 1}^{J} C(i/j)\,p(j/t)\log[p(j/t) \tag{3}$$

这里代表个体被分入的类,代表它的真实类。

(2)均匀成本下的复杂度的基尼指数如式(4)所示:

$$d_G(t)=\sum_{j=1}^{J}\sum_{i=1}^{j-1} p(j/t)\,p(i/t)=\frac{1}{2}\left(1-\sum_{j=1}^{J} p^2(j/t)\right) \tag{4}$$

对于二叉树个体,简化后式(5)所示:

$$d_G(t)=p(1/t)\,p(2/t) \tag{5}$$

非均匀成本下的基尼指数复杂度式(6)所示:

$$d_G(t)=\sum_{j=1}^{J}\sum_{i=1}^{j-1} p(j/t)\,p(i/t)\,[C(j/i)+C(i/j)] \tag{6}$$

(3)具有子结点t_L和t_R,并且t_L和t_R表示个体分别进入结点t_L和t_R的比例的"二元"函数式(7)所示:

$$d_T(t)=\frac{P_L P_R}{4}\sum_{j=1}^{J}\left|p(j/t_L)-p(j/t_R)\right| \tag{7}$$

熵和基尼指数的复杂度函数属于在给定结点处个体的复杂度。因此,作为一个在结点分支个体的工具, 复杂度的改变是从父结点到子结点复杂度总和的变化。

一旦选定了基尼指数或熵复杂度函数, 一个分支规则,即分支值,被用于结点处最大化地降低通过分支获得的复杂度。对一些特征,使用我们刚才用的记法, 我们由获得的通过使用阈值把结点分支成(左)、(右) 两个子结点定义复杂度减少的增益如式(8)所示:

$$\triangle d(s,t)=d(t)-p_L d(t_L)-p_R d(t_R) \tag{8}$$

这里和个体各自进入结点和的比例。这个复杂度降低的增益也被认为是对于结点分支的优良性。分支过程将一直持续到结点处的最佳分支的优良性为正值。在此,我们再次强调这个过程仅用于基尼指数和熵函数。

3 分类回归树的剪枝

构造过程中得到的树并不一定是最简单、最紧凑的决策树。因为许多分支反映的可能是训练样本中的某个结点最优,而不是树全局最优。并且为了防止所建立的树和训练样本的过拟合现象, 需要进行树剪枝[38]。树剪枝过程试图检测和去掉这种分支,以提高对未知样本集进行分类的准确性。通常采用有事前剪枝和事后剪枝两种。

(1)先剪枝:该方法在产生完全拟合整个训练集的完全增长的决策树前就停止决策树的生长。一旦停止剪枝,当前结点就成为一个叶结点,其可能包含多个不同类别的训练样本, 判定标准有重要性检验和信息增益。此外,若在一个结点上划分样本集时,会导致结点中样本数少于指定的阈值, 也会停止分支。确定一个合理的阈值常常比较困难。阈值过大会导致决策树过于简单化, 阈值过小又会导致多余树

枝无法修剪。

(2) 后剪枝: 它允许决策树充分生长然后修剪掉多余的树枝。被修剪的结点就成为一个叶结点, 并将其标记为它所包含样本中类别个数最多的类别。常用的事后剪枝方法有最小期望误判成本 (ECM)[6]和最小描述长度(Minimum DescriptionLength, DML)[7]。

与先剪枝相比, 后剪枝倾向于产生更好的结果, 因为其根据完全增长的决策树做出剪枝决策, 先剪枝则可能过早终止决策树的生长。但是, 对于后剪枝, 当子树被剪掉后, 生长完全决策树的额外开销就浪费了。

CART 是后剪枝过程。假设树是由个终结点组成, 那么我们定义 CART 树的复杂度如式(9)所示:

$$D(T)=\sum_{t\in T^t} d(s,t) \tag{9}$$

就像 BFOS 指出的那样, 虽然通过选择最佳分支特征和在每一个结点处特征的最佳分支, 我们选择了树, 但是, 结果树仍然未必是最小化了复杂度的树。

当误分类成本不是均匀时, 期望的误分类成本定义为式(10)所示:

$$R(t)=-\sum_{j=1}^{J}\sum_{i=1,i\neq j}^{J} C(i/j)\,Q(j/t)\,\pi(j) \tag{10}$$

$Q(i/j)$ 表示 j 类的个体被误分类进入 i 类的比例。$\pi(j)$ 是一个体属于 j 类的先验概率。

当然这些估计的误分类比例是非常低估的, 因为它们依赖于开始产生分类规则的数据。在 CART 中, 估计的误分类成本的两个非常有用的方法——交叉验证(cross- validation)[7,8]和检验样本(test- sample)方法。它是一种替代随机二次抽样的方法。在该方法中, 每个记录用于训练的次数相同, 并且用于检验恰好一次。数据集被分成 K 份(K 通常为 10)大小相等的子样本, 使用 K-1 部分数据构建树, 用剩余数据检验它, 重复这一过程 K 次, 总误差就是所有 K 次运行的误差之和, 所获得的 K 次误差总和后再平均就可以得到交叉验证误分类率 $Q^{CV}(i/j)$, 然后把它们加入式(10)中, 获得总的交叉验证误分类率 $R^{CV}(i/j)$, 它考虑了先验概率和非均匀误分类成本。当数据集非常小或非常大时 K 是可以改变的。

交叉验证方法的优点是使用尽可能多的训练记录, 此外, 检验集之间是互斥的, 并且有效地覆盖了整个数据集; 缺点是整个过程重复 N 次, 计算得开销很大, 此外, 因为每个检验集只有一个记录, 性能估计量的方差偏高。

当数据集足够大时, 我们不必求助于交叉验证产生不是非常严重的下偏的误分类率估计量。在那种情况下, 我们简单地从学习样本中抽出随机检验样本, 得到了不包括在检验样本中个体的误分类率作为 $Q(i/j)$ 的估计量。获得的总的误分类率估计量记为 $R^{TS}(i/j)$。

BFOS 获得了 $R^{CV}(i/j)$ 和 $R^{TS}(i/j)$ 标准差(SE)的估计量。这里的标准差就是由检验样本和交叉验证个体子样本随机选择产生的 $R^{CV}(i/j)$ 和 $R^{TS}(i/j)$ 的分布。这些标准差估计量的目的就是用于最大树的剪枝。最大树就是由分支结点最初产生的树直到它们感觉上是纯的(purity), 这种纯度是指每个终结点仅包含属于一个单个类的个体或者它们的复杂度不能通过进一步剪枝而减少。

很显然是用不同的复杂度测量、不同的误分类成本结构、交叉验证和检验样本方法、不同水平的标准差规则(0 或 1), 通常获得各种各样的分类树。所以选择 '最佳 '树的标准是必须的。一个标准就是树的成本—复杂性。

树的成本—复杂性定义为式(11)所示:

$$R_\alpha(T)=R(T)+\alpha|T^t| \tag{11}$$

这里 α是复杂性系数, $0<\alpha$, T^t 是树得终结点的个数。因为当树得终结点的个数增多时估计的误分类率趋向于上升, 被采用的成本—复杂性处罚了树终结点个数的繁殖; 复杂性参数 α可能被认为是每个结点复杂性。然后这个成本—复杂性被用于与通过上述谨慎已选的方法获得的少量的树进行比较。

4 总　结

CART 有良好的优越性, 但是, 并不是说在任何情况下 CART 方法都好。对于许多数据集, CART 方法产生的树并不稳定。训练样本集的一点轻微改变都可能完全改变树的结构, 这些特点存在于具有显著相关特征的数据集中。在 CART 中, 问题就转换为在单个结点处存在几个分支, 而这几个分支在减少子结点的所有复杂度方面几乎是等价的。从而一个特定的分支选择是比较随意的, 但是它将导致更多可能不同的树。这种不稳定性意味着使用者必须十

分清楚由CART产生的树中特定特征的充分解释。另一方面,这一特点暗含着具有相似判别能力的不同树的有用性,它允许通过树的使用改变特征的选择。

那么,CART方法在何时是首选的呢? 对于小的数据集,与logistic回归相比,CART趋向于提供低准确的分类。但是,对大部分使用者,特别是如风险/成本分类方面的应用, 在透明度和易用性是极为重要的情况下,稍微低的准确性不是决定性的。

CART涉及到数据库、统计学、人工智能与机器学习等多个领域,在计算机、医学、市场营销等领域都得到了充分的应用。

参考文献:

[1] 王煜.基于决策树和K最近邻算法的文本分类研究[D].天津大学,管理科学与工程,2007.

[2] Breiman L, Friedman J, Olshen R, al et. Classification and Regression Trees [M]. New York: Chapman & Hall, 1984.

[3] Rousu J, L.Flander, M.Suutarinen, al et. Novel computational Tools in Bakery Process Data Analysis: A Comparative Study [J]. Journal of Food Engineering, 2003(57):45-56.

[4] Moisen G.G, T.S. Frescino. Comparing Five Modeling Techniques for Predicting Forest Characteristics [J].Ecological Modeling, 2003(30):209-225.

[5] Markham, I, B.G. Mathien, B. Wray. Kanban Setting Through Artificial Intelligence: A Comparative Study of Artificial Neural Networks and Decision Trees [J]. Integrated Manufacturing Systems: The International Journal of Manufacturing Technology Management, 2000(11):239-246.

[6] 赵玮,温小霓. 应用统计学教程(下册)[M].西安:西安电子科技大学出版社,2003:12-14.

[7] Richard O.Duda, Peter E. Hart, David G. Stork. 模式分类[M].北京:机械工业出版社,2004.

[8] 范明,范宏建等译.数据挖掘导论[M].北京:人民邮电出版社, 2006.

A Discussion of Classification and Regression Tree

LIANG Yin

(Guangdong Polytechnic Normal Univ., Guangdong Guangzhou 510665, China)

Abstract: Classification and regression tree, a wide range of applications, is a good decision tree algorithm. This article discussion it. In first we review the origin and application of the classification and regression tree, in the second we analyze the construction of the classification and regression tree in the cost of uniform and non-uniform costs, again discuss the pruning method and validation, finally we summarized.

Key words: classification and regression tree; structure; pruning