华夏35度

Data Mining

分类回归树CART(上)

分类回归树(CART,Classification And Regression Tree)也属于一种决策树，上回文我们介绍了基于ID3算法的决策树。作为上篇，这里只介绍CART是怎样用于分类的。

分类回归树是一棵二叉树，且每个非叶子节点都有两个孩子，所以对于第一棵子树其叶子节点数比非叶子节点数多1。

表1

名称	体温	表面覆盖	胎生	产蛋	能飞	水生	有腿	冬眠	类标记
人	恒温	毛发	是	否	否	否	是	否	哺乳类
巨蟒	冷血	鳞片	否	是	否	否	否	是	爬行类
鲑鱼	冷血	鳞片	否	是	否	是	否	否	鱼类
鲸	恒温	毛发	是	否	否	是	否	否	哺乳类
蛙	冷血	无	否	是	否	有时	是	是	两栖类
巨蜥	冷血	鳞片	否	是	否	否	是	否	爬行类
蝙蝠	恒温	毛发	是	否	是	否	是	否	哺乳类
猫	恒温	皮	是	否	否	否	是	否	哺乳类

导航

博客园 首页 联系 订阅 管理

公告

有粉丝的招聘网站

昵称：Orisun
园龄：6年
粉丝：788
关注：10
+加关注

统计

随笔 - 15 文章 - 272 评论 - 373

搜索

谷歌搜索

随笔分类(14)

生活积累(11)

杂侃天下(3)

文章分类(268)

Algorithms(46)

Android(13)

C/C++(19)

DataBase(7)

Distributed(20)

DM,NLP,AI(45)

Embed(9)

Java(12)

Linux(49)

Python(5)

script(11)

Search Engine(21)

Web(9)

Windows(2)

最新评论

1. Re:拉格朗日乘子法和KKT条件

还行

--萌悦

2. Re:数据挖掘学习清单

豹纹鲨	冷血	鳞片	是	否	否	是	否	否	鱼类
海龟	冷血	鳞片	否	是	否	有时	是	否	爬行类
豪猪	恒温	刚毛	是	否	否	否	是	是	哺乳类
鳗	冷血	鳞片	否	是	否	是	否	否	鱼类
蝾螈	冷血	无	否	是	否	有时	是	是	两栖类

上例是属性有8个，每个属性又有多少离散的值可取。在决策树的每一个节点上我们可以按任一个属性的任一个值进行划分。比如最开始我们按：

1）表面覆盖为毛发和非毛发

2）表面覆盖为鳞片和非鳞片

3）体温为恒温和非恒温

等等产生当前节点的左右两个孩子。按哪种划分最好呢？有3个标准可以用来衡量划分的好坏：GINI指数、双化指数、有序双化指数。下面我们只讲GINI指数。

GINI指数

总体内包含的类别越杂乱，GINI指数就越大（跟熵的概念很相似）。比如体温为恒温时包含哺乳类5个、鸟类2个，则：

$$GINI = 1 - [(\frac{5}{7})^2 + (\frac{2}{7})^2] = \frac{20}{49}$$

体温为非恒温时包含爬行类3个、鱼类3个、两栖类2个,则

$$GINI = 1 - [(\frac{3}{8})^2 + (\frac{3}{8})^2 + (\frac{2}{8})^2] = \frac{42}{64}$$

所以如果按照"体温为恒温和非恒温"进行划分的话，我们得到GINI的增益（类比信息增益）：

$$GINI\_Gain = \frac{7}{15} * \frac{20}{49} + \frac{8}{15} * \frac{42}{64}$$

谢谢！

--蟹粉小笼包

3. Re:LibSVM使用指南

推荐一款最易用的支持向量机软件：Excel+SVM，无需繁琐安装，无需复杂参数设置，一键自动寻优，高精度单因变量、多因变量回归、两分类、多分类。www.plsexcelword.com。

--gystld

4. Re:C4.5决策树

E(subtree)-D(subtree)>E(leaf) 中的减号　但是你在后面案例计算的时候用的是号

--冷冷的那一风

5. Re:FP-Tree算法的实现

因为每一项末尾都是牛奶，可以把牛奶去掉，得到条件模式基（Conditional Pattern Base,CPB），此时的后缀模式是：（牛奶）。复制代码薯片：1，鸡蛋：1薯片：3，鸡蛋：3，面包：3薯.....

--小王子的玫瑰

最好的划分就是使得GINI_Gain最小的划分。

终止条件

一个节点产生左右孩子后，递归地对左右孩子进行划分即可产生分类回归树。这里的终止条件是什么？什么时候节点就可以停止分裂了？直观的情况，当节点包含的数据记录都属于同一个类别时就可以终止分裂了。这只是一个特例，更一般的情况我们计算χ^2值来判断分类条件和类别的相关程度，当χ^2很小时说明分类条件和类别是独立的，即按照该分类条件进行分类是没有道理的，此时节点停止分裂。注意这里的"分类条件"是指按照GINI_Gain最小原则得到的"分类条件"。

假如在构造分类回归树的第一步我们得到的"分类条件"是：体温为恒温和非恒温。此时：

	哺乳类	爬行类	鱼类	鸟类	两栖类
恒温	5	0	0	2	0
非恒温	0	3	3	0	2

我在《独立性检验》中讲述了χ^2的计算方法。当选定置信水平后查表可得"体温"与动物类别是否相互独立。

还有一种方式就是，如果某一分支覆盖的样本的个数如果小于一个阈值，那么也可产生叶子节点，从而终止Tree-Growth。

剪枝

当分类回归树划分得太细时，会对噪声数据产生过拟合作用。因此我们要通过剪枝来解决。剪枝又分为前剪枝和后剪枝：前剪枝是指在构造树的过程中就知道哪些节点可以剪掉，于是干脆不对这些节点进行分裂，在N皇后问题和背包问题中用的都是前剪枝，上面的χ^2方法也可以认为是一种前剪枝；后剪枝是指构造出完整的决策树之后再来考查哪些子树可以剪掉。

在分类回归树中可以使用的后剪枝方法有多种，比如：代价复杂性剪枝、最小误差

剪枝、悲观误差剪枝等等。这里我们只介绍代价复杂性剪枝法。

对于分类回归树中的每一个非叶子节点计算它的表面误差率增益值α。

$$\alpha = \frac{R(t) - R(T_t)}{|N_{T_t}| - 1}$$

$|N_{T_t}|$是子树中包含的叶子节点个数;

$R(t)$是节点t的误差代价，如果该节点被剪枝;

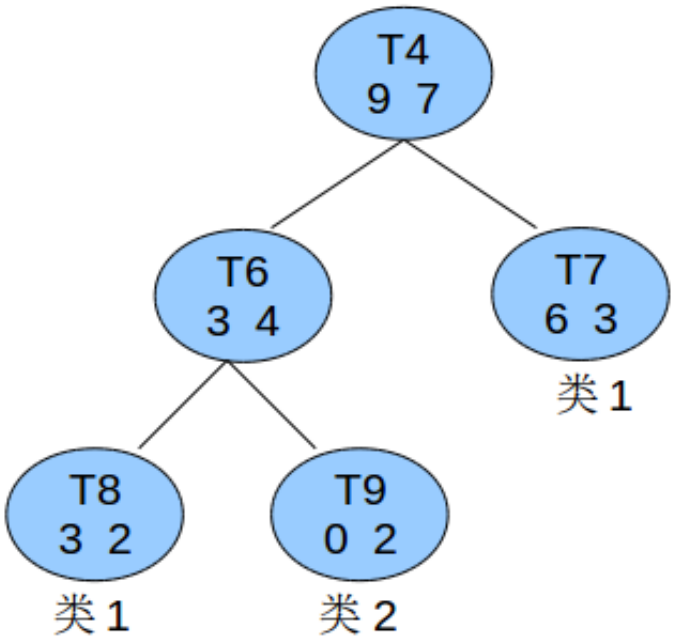
$$R(t) = r(t) * p(t)$$

r(t)是节点t的误差率;

p(t)是节点t上的数据占所有数据的比例。

$R(T_t)$是子树T_t的误差代价，如果该节点不被剪枝。它等于子树T_t上所有叶子节点的误差代价之和。

比如有个非叶子节点t4如图所示：



已知所有的数据总共有60条，则节点t4的节点误差代价为：

$$R(t) = r(t) * p(t) = \frac{7}{16} * \frac{16}{60} = \frac{7}{60}$$

子树误差代价为：

$$R(T_t) = \sum R(i) = (\frac{2}{5} * \frac{5}{60}) + (\frac{0}{2} * \frac{2}{60}) + (\frac{3}{9} * \frac{9}{60}) = \frac{5}{60}$$

以t4为根节点的子树上叶子节点有3个，最终：

$$\alpha = \frac{7/60 - 5/60}{3-1} = \frac{1}{6}$$

找到α值最小的非叶子节点，令其左右孩子为NULL。当多个非叶子节点的α值同时达到最小时，取$|N_{T_t}|$最大的进行剪枝。

源代码。拿表1作为训练数据，得到剪枝前和剪枝后的两棵分类回归树，再对表1中的数据进行分类测试。

```
+ View Code
```

C4.5克服了ID3的2个缺点：

1.用信息增益选择属性时偏向于选择分枝比较多的属性值，即取值多的属性

2.不能处理连贯属性

详细可参考这篇博客。

原文来自:博客园（华夏35度）
http://www.cnblogs.com/zhangchaoyang 作者:Orisun

分类: DM,NLP,AI

Orisun
关注 - 10
粉丝 - 788
+加关注

« 上一篇：不要一个人吃饭
» 下一篇：数据挖掘学习清单

posted on 2012-10-01 21:41 Orisun 阅读(15393) 评论(6) 编辑 收藏

评论

#1楼 2013-09-19 20:38 Ja °

分类回归树CART(下)呢？

支持(0) 反对(0)

#2楼 2013-10-25 14:23 cherrywq

剪枝到什么时候停止 ？？

支持(0) 反对(0)

#3楼 2013-11-13 19:15 ramboww

CART这个 Regression 在哪里看出来的？连续值不也是预处理成离散型吗？

支持(0) 反对(0)

#4楼 2014-01-16 11:10 ywgl

@ramboww

> 引用
> CART这个 Regression 在哪里看出来的？连续值不也是预处理成离散型吗？

回归啥时候也讲讲啊，很想看看啊，博主啊

支持(0) 反对(0)

#5楼 2014-05-14 19:17 zhangxiaoer

博主 animal这个文件的数剧是什么呢？ 能不能讲讲啊

支持(0) 反对(0)

#6楼 2015-11-17 16:08 xconming

博主，表1中的没有找到两个"鸟类"的数据呀？？

支持(1) 反对(0)

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

最新**IT**新闻：
· 作为员工，如何识别初创企业健康状况
· 如何招到靠谱的产品经理？
· 创业跟风者的15项特征
· 在网上没人知道你是一条狗的时候，你会怎么做？
· 看完豆瓣读书这份年度榜单，才知道今年错过了多少好书
» 更多新闻...

最新知识库文章：
· Git协作流程
· 企业计算的终结
· 软件开发的核心
· Linux概念架构的理解
· 从涂鸦到发布——理解API的设计

过程

» 更多知识库文章…