

CSDN新首页上线啦，邀请你来立即体验！(http://blog.csdn.net/)

立即体验

CSDN

博客 (http://blog.csdn.net/?ref=toolbar)学院 (http://edu.csdn.net/?ref=toolbar)

下载 (http://download.csdn.net/?ref=toolbar)更多 ▾

登录 (https://passport.csdn.net/account/login?ref=toolbar)

注册 (http://passport.csdn.net/account/mobileregister?ref=toolbar&action=mobileRegister)

3

Opencv2.4.9源码分析——Decision Trees

原创

2016年01月12日 13:42:39

4933

zhaocj (http://blog.csdn....)

+关注

(http://blog.csdn.net/zhaocj)

原创

粉丝

喜欢

未开通

88

1275

0

(https://github.com)

一、原理

决策树是一种非参数的监督学习方法，它主要用于分类和回归。决策树的目的是构造一种模型，使之能够从样本数据的特征属性中，通过学习简单的决策规则——IF THEN规则，从而预测目标变量的值。

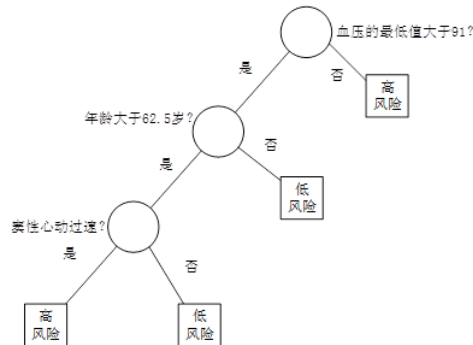


图1 决策树

例如，在某医院内，对因心脏病发作而入院治疗的患者，在住院的前24小时内，观测记录下来他们的19个特征属性——血压、年龄、以及其他17项可以综合判断病人状况的重要指标，用图1所示的决策树判断病人是否属于高危患者。在图1中，圆形为中间节点，也就是树的分支，它代表IF THEN规则的条件；方形为终端节点（叶节点），也就是树的叶，它代表IF THEN规则的结果。我们也把第一个节点称为根节点。

决策树往往采用的是自上而下的设计方法，每迭代循环一次，就会选择一个特征属性进行分叉，直到不能再分叉为止。因此在构建决策树的过程中，选择最佳（既能够快速分类，又能使决策树的深度小）的分叉特征属性是关键所在。这种“最佳性”可以用非纯度（impurity）进行衡量。如果一个数据集中只有一种分类结果，则该集合最纯，即一致性好；反之有许多分类，则不纯，即一致性不好。有许多指标可以定量的度量这种非纯度，最常用的有熵，基尼指数（Gini Index）和分类误差，它们的公式分别为：

$$\text{Entropy} = E(D) = - \sum_{j=1}^J p_j \log_2 p_j \quad (1)$$

$$\text{Gini Index} = \text{Gini}(D) = \sum_{j=1}^J p_j (1 - p_j) = \sum_{j=1}^J p_j - \sum_{j=1}^J p_j^2 = 1 - \sum_{j=1}^J p_j^2 \quad (2)$$

Classification Error = 1 - max{p_j} (3)

上述所有公式中，值越大，表示越不纯，这三个度量之间并不存在显著的差别。式中D表示样本数据的分类集合，并且该集合共有J种分类，p_j表示第j种分类的样本率：

$$p_j = \frac{N_j}{N} \quad (4)$$

式中N和N_j分别表示集合D中样本数据的总数和第j个分类的样本数量。把式4代入式2中，得到：

$$\text{Gini}(D) = 1 - \sum_{j=1}^J \left(\frac{N_j}{N}\right)^2 = 1 - \frac{\sum_{j=1}^J N_j^2}{N^2} \quad (5)$$

目前常用的决策树的算法包括ID3（Iterative Dichotomiser 3，第3代迭戈二叉树）、C4.5和CART（ClassificationAnd Regression Tree，分类和回归树）。前两种算法主要应用的是基于熵的方法，而第三种应用的是基尼指数的方法。下面我们就逐一介绍这些方法。

ID3是由Ross Quinlan首先提出，它是基于所谓“Occam'srazor”（奥卡姆剃刀），即越简单越好，也就是越

他的最新文章

更多文章 (http://blog.csdn.net/zhaocj)

- Opencv2.4.9源码分析——Cascade Classification（三）(http://blog.csdn.net/zhaocj/article/details/54412080)
- Opencv2.4.9源码分析——Cascade Classification（二）(http://blog.csdn.net/zhaocj/article/details/54291762)
- Opencv2.4.9源码分析——Cascade Classification（一）(http://blog.csdn.net/zhaocj/article/details/54015501)

在线课程

腾讯云服务器架构实现介绍

0

腾讯云服务器架构实现介绍 (0)

讲师：董晓杰

Python在58同城的实践

(http://edu.csdn.net/732utm_source=blog9)

Python在58同城的实践 (http://edu.csdn.net/732utm_source=blog9)

/series_detail/73?utm_source=blog9)

他的热门文章

- Opencv2.4.9源码分析——MSER (http://blog.csdn.net/zhaocj/article/details/40742191)
- 33102
- Win7下qt5.3.1+opencv2.4.9编译环境的

是小型的决策树越优于大型的决策树。如前所述，我们已经有了熵作为衡量样本集合纯度的标准，熵越大，越不纯，因此我们希望在分类以后能够降低熵的大小，使之变纯一些。这种分类后熵变小的判定标准可以用信息增益（Information Gain）来衡量，它的定义为：

$$G(D,A)=E(D)-\sum\frac{N_i}{N}E(D_i)\quad(6)$$

阅读全文



3



- 搭建 (<http://blog.csdn.net/zhaocj/article/details/38944037>)
🔒 23349
- s3c2440启动文件详细分析 (<http://blog.csdn.net/zhaocj/article/details/5302370>)
🔒 22825
- Opencv2.4.9源码分析——SIFT (<http://blog.csdn.net/zhaocj/article/details/42124473>)
🔒 22688
- Opencv2.4.9源码分析——HoughLinesP (<http://blog.csdn.net/zhaocj/article/details/40047397>)
🔒 21116

相关文章推荐

决策树（八）--随机森林及OpenCV源码分析 (http://blog.csdn.net/App_12062011/article...)

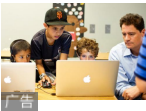
原文：<http://blog.csdn.net/zhaocj/article/details/51580092> 一、原理随机森林（Random Forest）的思想最早是由Ho于1995年首次提出...

App_12062011 (http://blog.csdn.net/App_12062011) 2016年08月08日 11:43 🔒1601

机器学习---决策树CART---opencv源码分析 (<http://blog.csdn.net/lizhengl/article/details/...>)

一、原理 决策树是一种非参数的监督学习方法，它主要用于分类和回归。决策树的目的是构造一种模型，使之能够从样本数据的特征属性中，通过学习简单的决策规则——IF THEN规则，从而预测目标变量的值。...

lizhengl (<http://blog.csdn.net/lizhengl>) 2017年02月10日 12:20 🔒261



都是前端，月薪20K和40k的开发到底差距在哪？
大学毕业后我成为前端开发者，从一开始的小白到现在的“高手”，我把一些感想记录下来...

(http://www.baidu.com/cb.php?c=lgF_pyfqnHmknj0dP1f0lZ0qnfK9ujYzP1nYPH0k0Aw-5Hc3rHnYnHb0TAq15HfLPWRznb0T1YvrH7-uHR4nAFBmyDkPADv0AwY5HDdnHcsnH01nHT0lgF_5y9YIZ0lQzq-uZR8mLPbUB48ugfEIAqspynElvNBnHqdlAdxTvqdThP-5yF_UvTkn0KzujY4rHb0mhYqn0KsTWYs0ZNGujYkPHTYn1mk0AqGujYknWb3rjDY0APGujYLnWm4n1c0ULI85H00TZbqnW0v0APzm1Y1n1cvr0)

Opencv2.4.9源码分析——bilareralFilter (<http://blog.csdn.net/zhaocj/article/details/39520...>)

双边滤波（bilateral filter）是一种非线性滤波技术，它是由Tomasi于1998年提出。它扩展了高斯平滑滤波技术。高斯滤波是一种常见并且有效的滤波方法，简单地说它是以被处理像...

zhaocj (<http://blog.csdn.net/zhaocj>) 2014年09月24日 10:43 🔒8806

Opencv2.4.9源码分析——HoughLinesP (<http://blog.csdn.net/zhaocj/article/details/4004...>)

标准霍夫变换本质上是把图像映射到它的参数空间上，它需要计算所有的M个边缘点，这样它的运算量和所需内存空间都会很大。如果在输入图像中只是处理m（mM）个边缘点，则这m个边缘点的选取是具有一定概率性的，因...

zhaocj (<http://blog.csdn.net/zhaocj>) 2014年10月13日 16:35 🔒21137



Delphi7高级应用开发随书源码 (<http://download.csdn.net/detail/chenxh...>)

(<http://download.csdn.net/detail/chenxh...>) 2003年04月30日 00:00 676KB 下载



程序员跨越式成长指南

完成第一次跨越，你会成为具有一技之长的开发者，月薪可能翻上几番；完成第二次跨越，你将成为拥有局部优势或行业优势的专业人士，获得个人内在价值的有效提升和外在收入的大幅跃迁...

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjfrjD0lZ0qnfK9ujYzP1f4PjnY0Aw-5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1dhrH7-uHuWrrHTYmHFBmH9h0AwY5HDdnHcsnH01nH60lgF_5y9YIZ0lQzqMpgwBUvqoQhP8QvIGIAPCmgfEmvq_lyd8Q1R4uWc4uHf3uAckPHRkPWN9PhcsmW9huWqdlAdxTv5HDkxWFBmhkEusKzujY4rHb0mhYqn0KsTWYs0ZNGujYkPHTYn1mk0AqGujYkn10snjf10APGujYLnWm4n1c0ULI85H00TZbqnW0v0APzm1YYn1bsPf)

Opencv2.4.9源码分析——Gradient Boosted Trees (http://blog.csdn.net/zhaocj/article/de...

一、原理 梯度提升树（GBT，Gradient Boosted Trees，或称为梯度提升决策树）算法是由Friedman于1999年首次完整的提出，该算法可以实现回归、分类和排序。GBT的优点是...

zhaocj (http://blog.csdn.net/zhaocj) 2016年05月20日 15:28 5905

OpenCv中决策树源代码解读(一) (http://blog.csdn.net/maweifei/article/details/72763020)

*****...
maweifei (http://blog.csdn.net/maweifei) 2017年05月26日 08:40 509

【opencv】goodFeaturesToTrack源码分析-1 (http://blog.csdn.net/jaych/article/details/5...

本系列文章为goodFeaturesToTrack源码分析，包括：【opencv】goodFeaturesToTrack源码分析-1【opencv】goodFeaturesToTrack源码分...

jaych (http://blog.csdn.net/jaych) 2016年04月20日 20:59 1725



Delphi7高级应用开发随书源码 (http://download.csdn.net/detail/chenhx...

(http://download... 2003年04月30日 00:00 676KB 下载 (

opencv reduce函数 (http://blog.csdn.net/jacke121/article/details/60589168)

opencv reduce函数
jacke121 (http://blog.csdn.net/jacke121) 2017年03月06日 19:16 1178

Opencv2.4.9源码分析——Extremely randomized trees (http://blog.csdn.net/zhaocj/articl...

一、原理 ET或Extra-Trees（Extremely randomized trees，极端随机树）是由PierreGeurts等人于2006年提出。该算法与随机森林算法十分相似，都是由...

zhaocj (http://blog.csdn.net/zhaocj) 2016年06月12日 21:04 4783



Opencv2.4.9源码分析——Support Vector Machines (http://download.cs...

(http://download... 2016年05月02日 17:49 689KB 下载 (





Opencv2.4.9源码分析——FAST (http://download.csdn.net/detail/sinat_3...

(http://download... 2017年10月29日 17:23 953KB 下载 (

Opencv2.4.9源码分析——MSCR (<http://blog.csdn.net/zhaocj/article/details/43191829>)

前面我们介绍了MSER方法，但该方法不适用于对彩色图像的区域检测。为此，Forssen于2007年提出了针对彩色图像的最大稳定极值区域的检测方法——MSCR (Maximally Sta...

 zhaocj (<http://blog.csdn.net/zhaocj>) 2015年01月27日 10:51  4985



3



opencv2.4.9源码分析——SIFT (<http://download.csdn.net/detail/zhaocj/8...>)

 (<http://download.csdn.net/detail/zhaocj/8...>) 2014年12月24日 14:42 1.34MB [下载 \(](#)





opencv2.4.9源码分析——SURF (<http://download.csdn.net/detail/zhaocj/...>)

 (<http://download.csdn.net/detail/zhaocj/...>) 2015年01月10日 16:43 1.32MB [下载 \(](#)



Opencv2.4.9源码分析——SIFT (<http://blog.csdn.net/zhaocj/article/details/42124473>)

SIFT (尺度不变特征变换, Scale-Invariant Feature Transform) 是在计算机视觉领域中检测和描述图像中局部特征的算法, 该算法于1999年被David Lowe提出, 并于2...

 zhaocj (<http://blog.csdn.net/zhaocj>) 2014年12月24日 15:34  22720



Opencv2.4.9源码分析——HoughLinesP (http://blog.csdn.net/Dopamy_BusyMonkey/arti...)

Opencv2.4.9源码分析——HoughLinesP

 Dopamy_BusyMonkey (http://blog.csdn.net/Dopamy_BusyMonkey) 2015年10月10日 15:37  549

Opencv2.4.9源码分析——Cascade Classification (一) (<http://blog.csdn.net/zhaocj/arti...>)

我把级联分类器分为三部分内容介绍, 第一部分内容是原理。物体识别, 尤其是人脸识别, 是近二、三十年里计算机视觉领域一个热门的课题。它的应用范围极广, 目前成熟的算法也较多。OpenCV也集成...

 zhaocj (<http://blog.csdn.net/zhaocj>) 2017年01月04日 09:16  2196

Opencv2.4.9源码分析——HoughCircles (<http://blog.csdn.net/zhazhiqiang2010/article/de...>)

【原文: <http://blog.csdn.net/zhaocj/article/details/50454847>】图形可以用一些参数进行表示, 标准霍夫变换的原理就是把图像空间转换成参数空间...

 zhazhiqiang2010 (<http://blog.csdn.net/zhazhiqiang2010>) 2016年04月08日 15:44  1258