算法杂货铺——分类算法之决策树(Decision tree)

2010-09-19 16:30 by T2噬菌体, 93889 阅读, 29 评论, 收藏, 编辑

3.1、摘要

在前面两篇文章中，分别介绍和讨论了朴素贝叶斯分类与贝叶斯网络两种分类算法。这两种算法都以贝叶斯定理为基础，可以对分类及决策问题进行概率推断。在这一篇文章中，将讨论另一种被广泛使用的分类算法——决策树（decision tree）。相比贝叶斯算法，决策树的优势在于构造过程不需要任何领域知识或参数设置，因此在实际应用中，对于探测式的知识发现，决策树更加适用。

3.2、决策树引导

通俗来说，决策树分类的思想类似于找对象。现想象一个女孩的母亲要给这个女孩介绍男朋友，于是有了下面的对话：

女儿：多大年纪了？

母亲：26。

女儿：长的帅不帅？

母亲：挺帅的。

女儿：收入高不？

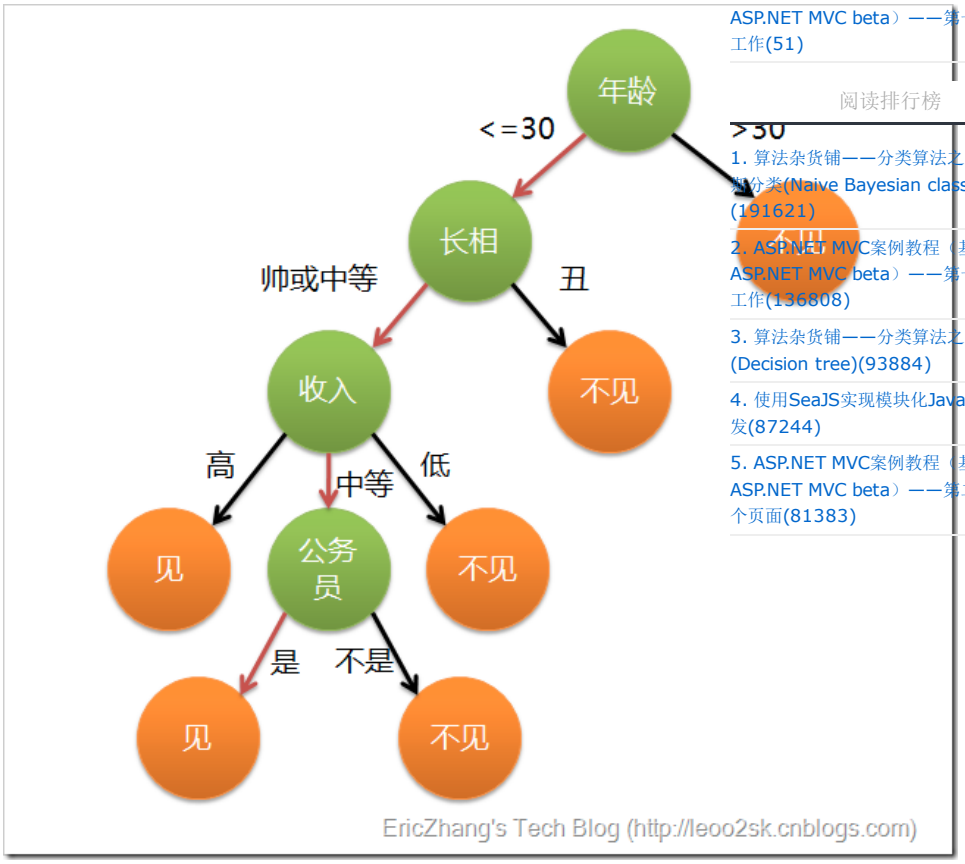
母亲：不算很高，中等情况。

女儿：是公务员不？

母亲：是，在税务局上班呢。

女儿：那好，我去见见。

这个女孩的决策过程就是典型的分类树决策。相当于通过年龄、长相、收入和是否公务员对将男人分为两个类别：见和不见。假设这个女孩对男人的要求是：30岁以下、长相中等以上并且是高收入者或中等以上收入的公务员，那么这个可以用下图表示女孩的决策逻辑（声明：此决策树纯属为了写文章而YY的产物，没有任何根据，也不代表任何女孩的择偶倾向，请各位女同胞莫质问我^_^）：



上图完整表达了这个女孩决定是否见一个约会对象的策略，其中绿色节点表示判断条件，橙色节点表示决策结果，箭头表示在一个判断条件在不同情况下的决策路径，图中红色箭头表示了上面例子中女孩的决策过程。

这幅图基本可以算是一颗决策树，说它“基本可以算”是因为图中的判定条件没有量化，如收入高中低等等，还不能算是严格意义上的决策树，如果将所有条件量化，则就变成真正的决策树了。

有了上面直观的认识，我们可以正式定义决策树了：

决策树（**decision tree**）是一个树结构（可以是二叉树或非二叉树）。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。

可以看到，决策树的决策过程非常直观，容易被人理解。目前决策树已经成功运用于医学、制造产业、天文学、分支生物学以及商业等诸多领域。知道了决策树的定义以及其应用方法，下面介绍决策树的构造算法。

3.3、决策树的构造

不同于贝叶斯算法，决策树的构造过程不依赖领域知识，它使用属性选择度量来选择将元组最好地划分成不同的类的属性。所谓决策树的构造就是进行属性选择度量确定各个特征属性之间的拓扑结构。

构造决策树的关键步骤是分裂属性。所谓分裂属性就是在某个节点处按照某一特征属性的不同划分构造不同的分支，其目标是让各个分裂子集尽可能地“纯”。尽可能“纯”就是尽量让一个分裂子集中待分类项属于同一类别。分裂属性分为三种不同的情况：

1、属性是离散值且不要求生成二叉决策树。此时用属性的每一个划分作为一个分支。

2、属性是离散值且要求生成二叉决策树。此时使用属性划分的一个子集进行测试，按照“属于此子集”和“不属于此子集”分成两个分支。

3、属性是连续值。此时确定一个值作为分裂点split_point，按照

ASP.NET MVC beta）——第一篇：准备工作(51)

阅读排行榜

1. 算法杂货铺——分类算法之朴素贝叶斯分类(Naive Bayesian classification)(191621)
2. ASP.NET MVC案例教程（基于ASP.NET MVC beta）——第一篇：准备工作(136808)
3. 算法杂货铺——分类算法之决策树(Decision tree)(93884)
4. 使用SeaJS实现模块化JavaScript开发(87244)
5. ASP.NET MVC案例教程（基于ASP.NET MVC beta）——第二篇：第一个页面(81383)

2008年11月(7)
2008年10月(3)
2008年9月(2)
2008年8月(1)
2008年7月(9)
2008年6月(7)
2008年4月(3)

>split_point和<=split_point生成两个分支。

构造决策树的关键性内容是进行属性选择度量，属性选择度量是一种选择分裂准则，是将给定的类标记的训练集合的数据划分D“最好”地分成个体类的启发式方法，它决定了拓扑结构及分裂点split_point的选择。

属性选择度量算法有很多，一般使用自顶向下递归分治法，并采用不回溯的贪心策略。这里介绍ID3和C4.5两种常用算法。

3.3.1、ID3算法

从信息论知识中我们直到，期望信息越小，信息增益越大，从而纯度越高。所以ID3算法的核心思想就是以信息增益度量属性选择，选择分裂后信息增益最大的属性进行分裂。下面先定义几个要用到的概念。

设D为用类别对训练元组进行的划分，则D的熵（entropy）表示为：

$$info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

其中pi表示第i个类别在整个训练元组中出现的概率，可以用属于此类别元素的数量除以训练元组元素总数量作为估计。熵的实际意义表示是D中元组的类标号所需要的平均信息量。

现在我们假设将训练元组D按属性A进行划分，则A对D划分的期望信息为：

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info(D_j)$$

而信息增益即为两者的差值：

$$gain(A) = info(D) - info_A(D)$$

ID3算法就是在每次需要分裂时，计算每个属性的增益率，然后选择增益率最大的属性进行分裂。下面我们继续用SNS社区中不真实账号检测的例子说明如何使用ID3算法构造决策树。为了简单起见，我们假设训练集合包含10个元素：

日志密度	好友密度	是否使用真实头像	账号是否真实
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

其中s、m和l分别表示小、中和大。

设L、F、H和R表示日志密度、好友密度、是否使用真实头像和账号是否真实，下面计算各属性的信息增益。

$$info(D) = -0.7log_2 0.7 - 0.3log_2 0.3 = 0.7 * 0.51 + 0.3 * 1.74 = 0.879$$

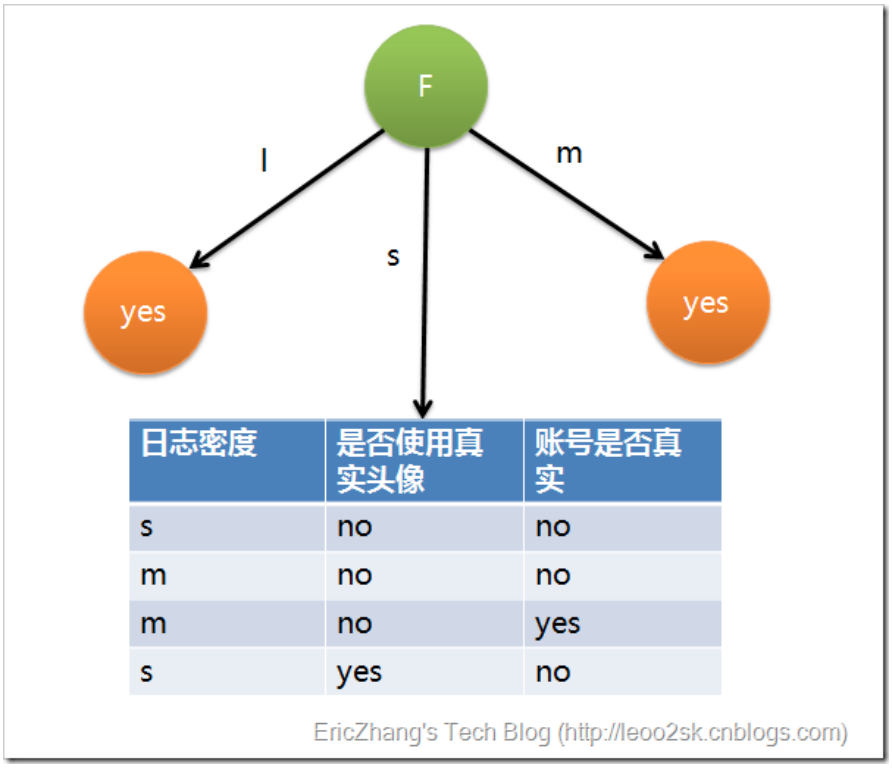
$$info_L(D) = 0.3*(-\frac{0}{3}log_2\frac{0}{3} - \frac{3}{3}log_2\frac{3}{3}) + 0.4*(-\frac{1}{4}log_2\frac{1}{4} - \frac{3}{4}log_2\frac{3}{4}) + 0.3*(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) = 0 + 0.326 + 0.277 = 0.603$$

$$gain(L) = 0.879 - 0.603 = 0.276$$

因此日志密度的信息增益是0.276。

用同样方法得到H和F的信息增益分别为0.033和0.553。

因为F具有最大的信息增益，所以第一次分裂选择F为分裂属性，分裂后的结果如下图表示：



日志密度	是否使用真实头像	账号是否真实
s	no	no
m	no	no
m	no	yes
s	yes	no

在上图的基础上，再递归使用这个方法计算子节点的分裂属性，最终就可以得到整个决策树。

上面为了简便，将特征属性离散化了，其实日志密度和好友密度都是连续的属性。对于特征属性为连续值，可以如此使用ID3算法：

先将D中元素按照特征属性排序，则每两个相邻元素的中间点可以看做潜在分裂点，从第一个潜在分裂点开始，分裂D并计算两个集合的期望信息，具有最小期望信息的点称为这个属性的最佳分裂点，其信息期望作为此属性的信息期望。

3.3.2、C4.5算法

ID3算法存在一个问题，就是偏向于多值属性，例如，如果存在唯一标识属性ID，则ID3会选择它作为分裂属性，这样虽然使得划分充分纯净，但这种划分对分类几乎毫无用处。ID3的后继算法C4.5使用增益率（gain ratio）的信息增益扩充，试图克服这个偏倚。

C4.5算法首先定义了“分裂信息”，其定义可以表示成：

$$split\_info_A(D) = -\sum_{j=1}^{v}\frac{|D_j|}{|D|}log_2(\frac{|D_j|}{|D|})$$

其中各符号意义与ID3算法相同，然后，增益率被定义为：

$$gain\_ratio(A) = \frac{gain(A)}{split\_info(A)}$$

C4.5选择具有最大增益率的属性作为分裂属性，其具体应用与ID3类似，不再赘述。

3.4、关于决策树的几点补充说明

3.4.1、如果属性用完了怎么办

在决策树构造过程中可能会出现这种情况：所有属性都作为分裂属性用光了，但有的子集还不是纯净集，即集合内的元素不属于同一类别。在这种情况下，由于没有更多信息可以使用了，一般对这些子集进行“多数表决”，即使用此子集中出现次数最多的类别作为此节点类别，然后将此节点作为叶子节点。

3.4.2、关于剪枝

在实际构造决策树时，通常要进行剪枝，这时为了处理由于数据中的噪声和离群点导致的过分拟合问题。剪枝有两种：

先剪枝——在构造过程中，当某个节点满足剪枝条件，则直接停止此分支的构造。

后剪枝——先构造完成完整的决策树，再通过某些条件遍历树进行剪枝。

关于剪枝的具体算法这里不再详述，有兴趣的可以参考相关文献。

T2噬菌体
关注 - 14
粉丝 - 2527
21 0
荣誉：推荐博客
+加关注

« 上一篇：算法杂货铺——分类算法之贝叶斯网络(Bayesian networks)
» 下一篇：算法杂货铺——k均值聚类(K-means)

分类: [11]数据挖掘
标签: 算法, 数据挖掘, 分类, 决策树

ADD YOUR COMMENT

#1楼 funskiller
2010-09-19 16:58
哈。虽然不太懂。但也燃起了我对算法的兴趣。
支持(0) 反对(0)

#2楼 Jun.lu
2010-09-19 17:09
压力很大！
支持(0) 反对(0)

#3楼[楼主] T2噬菌体
2010-09-19 17:12
@ jelle
@funskiller
其实算法的一大特点就是容易吓唬人，纸老虎一个，又是公式又是图标的。其实静下心来读，都是可以理解的。当然我是指应用，一般算法应用都不困难，不过证明和研究算法理论的可就真是大牛了。我的智商也就能写写应用方面的文章，所以不难懂。
支持(0) 反对(0)

#4楼 Jun.lu
2010-09-19 17:17

@ EricZhang(T2噬菌体)
你已经超级牛了。努力向你学习。

支持(0) 反对(0)

#5楼 指针为空
2010-09-19 17:22

一般这种问题我都是写一堆if判断
然后谁维护这堆if谁晕

支持(0) 反对(0)

#6楼 Tanky Woo
2010-09-19 18:18

同是算法爱好者的路过。。。

支持(0) 反对(0)

#7楼 funskiller
2010-09-19 18:41

@ EricZhang(T2噬菌体)
这点我确实比较赞同。学习嘛。就是研究其规律。最开始学编程的。对那些代码都一头雾水。熟知编程的特点后。学习新语言很快上手。
算法是立足于解决问题。只要知道了那玩意能够解决什么问题。我想学习的兴致还是很足的。比如你这个关于SNS账号检测。让我感觉有趣。也会尝试思考相关。另外一点。我对公式。符号啊。都不太了解。自学的，没系统上课过。如果能够加上个链接（比如说链接到中文维基百科）。那就更给力了。

支持(0) 反对(0)

#8楼[楼主] T2噬菌体
2010-09-19 18:46

@ funskiller
好提议，我马上加。。。

支持(0) 反对(0)

#9楼 funskiller
2010-09-19 18:50

@ EricZhang(T2噬菌体)
楼主给我提供点信息吧。就是关于算法由初级到专业，从理论到应用相关的书籍（Y文不佳，所以只有找中文作品）。另外，我也相当讨厌成功学。

支持(0) 反对(0)

#10楼[楼主] T2噬菌体
2010-09-19 19:06

@ funskiller
初期学习算法的话，《算法导论》绝对是本好书，另外就是建议读一读运筹学，因为很多算法问题其实是运筹学讲得更好，推荐图灵数学系列的《运筹学导论》，有两册。数据挖掘初期我推荐机械工业出版社的《数据挖掘 概念与技术》。
如果是研究某个具体算法，建议Google相关论文，一般书籍不会对某个算法讲得很深。
如果有从数学角度学习算法的决心，推荐《计算机程序设计艺术》。

支持(0) 反对(0)

#11楼 Allen Zhang
2010-09-19 20:16

@ 指针为空
用算法，一般也一样看不懂。

支持(0) 反对(0)

#12楼 chenkai
2010-09-19 22:58

@ EricZhang(T2噬菌体)
hehe <<算法导论>> 这让我想起Google GDJ编程大赛中一个选手
他在编程大赛中失败而归 于是乎就拿起了这本<<算法导论>> 重现读了一遍 终于从这本这本开山之作中收益颇多.....

支持(0) 反对(0)

#13楼 wilby
2010-09-27 10:02

“熵的实际意义表示是D中元组的类标号所需要的平均信息量。”，这句话能解释一下么？

支持(0) 反对(0)

#14楼 Everlonely
2010-10-26 17:04

我看到wiki上有一种 Gini impurity，似乎也是一种选择分裂属性的算法，但看不懂其含义。。。
http://en.wikipedia.org/wiki/Decision_tree_learning

Used by the CART algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it were randomly labelled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

如有兴趣可以去看看并解释一下。。。谢了

支持(0) 反对(0)

#15楼 S*S*S
2010-11-24 12:13

本科AI就学这个

#16楼 牛皮糖NewPtone
2011-07-24 22:14

@ wilby
如果一个元组中的元素数目趋于相等，那么熵越大，趋向于1，表明信息的多样化，信息量越大；相反，若某个元素的数目过大或过小，熵越小，趋向于0。
楼主好文，学习了。

支持(0) 反对(0)

#17楼 学习share
2011-09-12 23:34

学习学习

#18楼 苹果苹果大苹果
2011-09-22 10:04

看了一些聚类的文章，都晕晕乎乎，看了楼主的一系列算法杂货铺，可算明白大概是怎么回事了。
非常感谢。

支持(0) 反对(0)

#19楼 Chris Lee
2012-09-29 15:54

文章不错，熵还有更多可讲的。

支持(0) 反对(0)

#20楼 看向星空
2012-10-15 11:14

请问图是用什么画的？

支持(0) 反对(0)

#21楼 jackliu8722
2012-11-14 10:59

我来学习了，

支持(0) 反对(0)

#22楼 gxiaob
2013-03-21 20:11

博主，正在看您的ID3代码，请问输入是怎么控制的？

支持(0) 反对(0)

#23楼 cococo点点
2013-09-25 17:24

感谢博主分享 记得学信息论的时候其实就学了信息熵的概念~

支持(0) 反对(0)

#24楼 涛哥99
2014-06-11 21:21

感谢楼主。时至今日，我只能说，活到老，学到老。算法研究任重而道远！

支持(0) 反对(0)

#25楼 mavarick
2014-06-24 18:13
顶一下，希望自己实现下！
支持(0) 反对(0)

#26楼 cz大侠
2014-07-11 11:33
感谢楼主，非常喜欢你的算法系列
支持(0) 反对(0)

#27楼 chizi
2015-07-22 17:23
讲得真好！一个对算法很抵触的人看了都很感兴趣，谢谢！
支持(0) 反对(0)

#28楼 kmlxk
2015-09-16 09:30
赞！学习使用统计学R的rpart.plot，但是教科书上没有这部分原理，博主v5
支持(0) 反对(0)

#29楼 迈克儿
2016-01-12 22:36
@ 看向星空
显然是excel啊
支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

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。