决策树分类算法及其应用

曹颖超

(西安科技大学计算机科学与技术学院,陕西 西安 710054)

摘 要:本文详细介绍机器学习分类算法中的决策树算法,并详解如何构造,表示,保存决策树,以及如何使用决策树进行分类问题。

关键词:分类算法;决策树; ID3

1 概述

决策树分类方法是一种自上而下，在分支节点进行属性值的比较得到分裂点属性,根据不同的属性值判断构造向下的分支,最终在叶子节点得到分类结果。

传统的决策树分类方法有 ID3 和 C4.5，他们都是以信息熵作为分类依据,是单颗决策树。然而,由于单棵决策树的局限性,在训练数据过程中由于属性值的过多容易出现过拟合(Over-Fitting)现象。为了弥补单棵决策树的缺陷,本研究使用多颗决策树和 Boosting 算法结合在一起的 GDBT 分类方法。

决策树分类方法是一种自上而下，在分支节点进行属性值的比较得到分裂点属性,根据不同的属性值判断构造向下的分支,最终在叶子节点得到分类结果。

传统的决策树分类方法有 ID3 和 C4.5，他们都是以信息熵作为分类依据,是单颗决策树。

2 传统决策树分类算法

2.1 算法原理

在 ID3 算法中,选取信息增益最大的属性作为分类依据,然后根据该属性的属性值构造分支。具体构造步骤如下:

step1.构造一个节点,如果所有的样本都在该节点上,那么停止算法,将该节点改成叶子节点,并用该类标记。

step2.如果样本都不在一个节点上,求出每一个属性的信息熵,根据公式求出每一个属性的信息增益,然后选择信息增益最大的一个属性作为分类依据。

step3.对属性中的每一个值创建一个分支,然后划分样本。

step4.重复上述 step1-step3 步骤,直到属性全部使用完。

ID3 算法有以下计算公式:

加入一个训练数据集有 N 个分类，用 D 来标记这个训练集,它有 M 的属性,定义 m 个不同的类 Ci(I=1,2,…,m),Ci,D 是 Ci 类的元组的集合。和分别表示 D 和 Ci,D 中元组的个数。对 D 中的元组分类所需的期望信息由(1)公式给出:

$$Info(D)=-\sum_{j=1}^{m} pi\log_2(pi) \tag{1}$$

假设属性 A 它的属性值有 V 的不同的离散值,可以根据属性 A 把数据集 D 划分成 v 个子集{D1, D2,…Dv}。设子集 Dj 中全部的记录数在 A 上具有相同的值 aj。基于按 A 划分对 D 的元组分类所需要的期望信息由(2)公式给出:

$$Info_A(D)=-\sum_{j=1}^{v}\frac{|Dj|}{|D|}\times Info(Dj) \tag{2}$$

信息增益定义为原来的信息需求(基于类比例)与新的信息需求(对 A 划分之后得到的)之间的差,如公式(3):

$$Gain(A)=Info(D)-Info_A(D) \tag{3}$$

2.2 算法示例

有一个训练数据集,有一些与天气相关的属性,分类是在某种天气组合下出不出门,具体数据如表 1 所示。

表 1 样本数据集

属性	Outlook	Temperature	Humidity	Windy	类
1	Overcast	Hot	High	Not	No
2	Overcast	Hot	High	Very	No
3	Overcast	Hot	High	Medium	No
4	Sunny	Hot	High	Not	Yes
5	Sunny	Hot	High	Medium	Yes
6	Rain	Mild	High	Not	No
7	Rain	Mild	High	Medium	No
8	Rain	Hot	Normal	Not	Yes
9	Rain	Cool	Normal	Medium	No
10	Rain	Hot	Normal	Very	No
11	Sunny	Cool	Normal	Very	Yes
12	Sunny	Cool	Normal	Medium	Yes
13	Overcast	Mild	High	Not	No
14	Overcast	Mild	High	Medium	No
15	Overcast	Cool	Normal	Not	Yes
16	Overcast	Cool	Normal	Medium	Yes
17	Rain	Mild	Normal	Not	No
18	Rain	Mild	Normal	Medium	No
19	Overcast	Mild	Normal	Medium	Yes
20	Overcast	Mild	Normal	Very	Yes
21	Sunny	Mild	High	Very	Yes
22	Sunny	Mild	High	Medium	Yes
23	Sunny	Hot	Normal	Not	Yes
24	Rain	Mild	High	Very	No

初始时刻的熵值为:

$$H(X)=-\frac{12}{24}lb\frac{12}{24}-\frac{12}{24}lb\frac{12}{24}=1$$

选取 outlook 属性作为测试属性,此时条件熵为:

$$H(X|Outlook)=\frac{9}{24}(-\frac{4}{9}lb\frac{4}{9}-\frac{5}{9}lb\frac{5}{9})+\frac{8}{24}(-\frac{1}{8}lb\frac{1}{8}-\frac{7}{8}lb\frac{7}{8})+\frac{7}{24}(-\frac{7}{7}lb\frac{7}{7}-0)=0.4643$$

选取 Temperature 属性作为测试属性,则有:

$$H(X|Teamp)=\frac{8}{24}(-\frac{4}{8}lb\frac{4}{8}-\frac{4}{8}lb\frac{4}{8})+\frac{11}{24}(-\frac{4}{11}lb\frac{4}{11}-\frac{7}{11}lb\frac{7}{11})+\frac{5}{24}(-\frac{4}{5}lb\frac{4}{5}-\frac{1}{5}lb\frac{1}{5})=0.6739$$

选取 Humidity 属性作为测试属性,则有:

$$H(X|Humidity)=\frac{12}{24}(-\frac{4}{12}lb\frac{4}{12}-\frac{8}{12}lb\frac{8}{12})+\frac{12}{24}(-\frac{4}{12}lb\frac{4}{12}-\frac{8}{12}lb\frac{8}{12})=0.8183$$

选取 Windy 属性作为测试属性,则有:

$$H(X|Windy)=\frac{8}{24}(-\frac{4}{8}lb\frac{4}{8}-\frac{4}{8}lb\frac{4}{8})+\frac{6}{24}(-\frac{3}{6}lb\frac{3}{6}-\frac{3}{6}lb\frac{3}{6})+\frac{10}{24}(-\frac{5}{10}lb\frac{5}{10}-\frac{5}{10}lb\frac{5}{10})=1$$

(转下页)

接入网组网结构优化及关键技术研究

房 毅
（中国联通吉林省分公司，吉林 长春 130000）

摘　要：文章首先对接入网的特征及其结构进行简要分析，在此基础上对接入网组网结构优化技术进行论述。期望通过本文的研究能够对接入网运行稳定性和可靠性的提升有所帮助。

关键词：接入网；组网结构；优化技术

1 接入网的特征及其结构

近年来，随着对通信技术研究力度的不断加大，使得此项技术获得了长足进步与发展，与此同时，人们对电信业务的需求也呈现出多样化的趋势，不仅如此，主干网上的相关技术日渐成熟，如 SDH、PON、ATM、DWDM 等等，这些技术的使用为一线入户提供了可能，从而使数据、语音、图像三线合一的目标得以实现。接入网具体是指骨干网络与用户终端间的全部设备，它的长度从数百米至几千米不等，骨干网的传输介质为光纤，其传输速度较快，正因如此，使得接入网成为整个网络的瓶颈。

1.1 接入网的基本特征

大体上可将接入网的特征归纳为以下几个方面：

1.1.1 对接入的各种业务，接入网能够提供足够的承载能力，由此可以实现业务的透明传送，同时对用户发布的信令也是透明的，除某些特殊用户的信令格式需要进行转换之外，业务处理功能仍在节点范围之内。

1.1.2 对接入网引入的业务和类型基本无任何限制，可借助有限的标准化接口实现业务节点的有效连接。

1.1.3 具有独立于业务节点意外的网管系统，其能够利用标准化接口与 TMN 进行连接，同时，TMN 可以对接入网进行运维和管理。

1.2 接入网的结构

接入网有以下几种结构类型：

1.2.1 总线形。这种结构具体是指将光纤作为公共总线，利用耦合器将各个用户终端与总线进行连接，进而形成网络结构。该接入网结构具有如下特点：所有网络上的用户终端均可以共享主干光纤，由此大幅度节省了前期的线路投资，节点的增设和删减更加方便，终端设备之间的干扰相对较小。其唯一不足之处是损耗积累比较严重，用户终端设备接受信号对主干光纤具有较强的依赖性，一旦主干光纤出现故障，将会对大部分用户造成影响。

（转下页）

可以看出 H(X|Outlook)最小，即 I(X,Outlook) 最大，所以应该选择 Outlook 作为测试属性。

对每个结点执行上面的过程，最后生成下决策树如图 1 所示。

由于每次进行属性划分时都需要遍历所有样本计算信息增益，导致算法的性能较低。同时，为了提升样本训练时的准确率，极易造成模型的过拟合问题。

3 结论

本文演示的是最经典 ID3 决策树，但它在实际应用中存在过度匹配的问题。在以后的研究中将讨论如何对决策树进行裁剪。ID3 决策树算法只能用于标准型数据。对于数值型数据，需要使用 Cart 决策树构造算法。

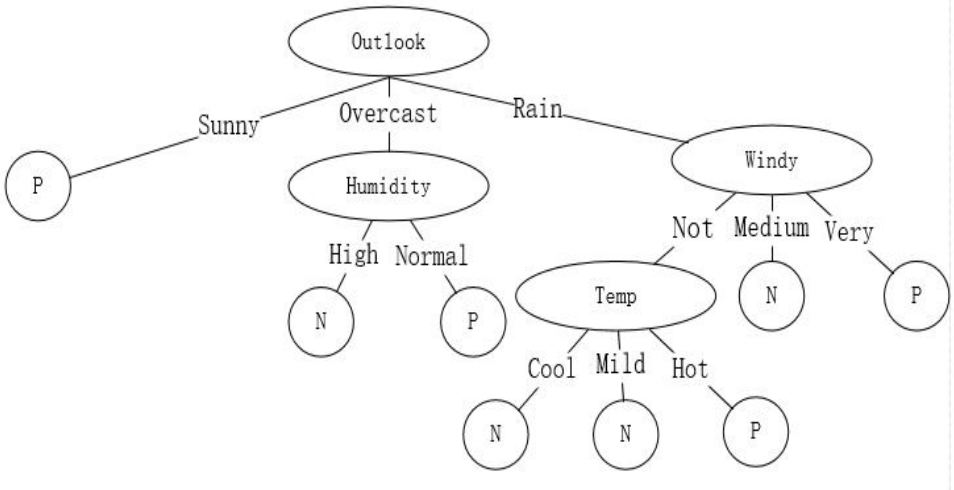

图 1 ID3 算法生成的决策树

参考文献

[1]孟岩，汪云云. 典型半监督分类算法的研究分析[J]. 计算机技术与发展，2017，(09)：1-7.

[2] 龙浩. 用于不平衡分类问题的自适应加权极限学习机研究[D].深圳：深圳大学，2017.

[3]杨志辉. 基于机器学习算法在数据分类中的应用研究[D].太原：中北大学，2017.

[4]沈龙凤，宋万干，葛方振等.最优路径森林分类算法综述[J].计算机应用研究，2018，(1)：1-9.