ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.14, No.20, July. 2018
E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

决策树几种分类算法的分析比较

徐梦茹,王学明

(宁夏大学,宁夏 银川 750001)

摘要:对数据的处理一直是现代科技一直在力争攻克的难关。现代社会的数据量每天都在急速增长,那么面临的难关也就会越来越多,例如,如何从海量数据中获取有用的数据,进而将有用的数据转化为"知识"。本文将首先对数据挖掘中决策树分类算法中的ID3算法、C4.5算法、CART算法进行详细分析,然后总结出各个算法的优缺点,并提出每种算法应该应用于何种情况之下。

关键词:决策树分类算法;ID3算法;C4.5算法;CART算法

中图分类号:TP31 文献标识码:A 文章编号:1009-3044(2018)20-0193-03

随着现代科技的发展,数据已经成为人们生活中必不可少的元素之一,几乎没有人的生活是可以离开数据的,大到宇宙星体间的关联,小到超市商品的信息,可以说生活处处是数据。每人每天都会产生大量的数据,国际著名的数据公司IDC报告称,2013年全球产生的数据量已达4.4ZB,且将以每两年翻一番的速度增长,到2020年,全球数据量将高达44ZB。[1]由此可知全球每天产生的数据量是极其庞大的,各领域的学者都希望能充分利用这些数据,通过分析,获得大量有用的信息。然而实际上,有用的数据是很少的,要从这些数据中发现有用的信息犹如大海捞针,更不用说再寻找它们之间的联系,因而需要一种能够在海量数据中快速找出有用信息的技术,并对这些信息之间的关系加以分析,从而发掘出对人们生产和生活有用的"知识"。由此产生的一项技术就是数据挖掘技术。

数据挖掘是从大量数据中发现有用知识的过程,是一种专业技术,也是一种分析数据的手段,用于发现海量数据所隐藏的各种规律。数据挖掘技术在对数据进行处理的过程中需要对数据进行分类,这样才能方便之后数据的处理与预测。通常所用到的数据分类技术是决策树分类法。决策树,顾名思义,它是一种树形结构,如图1所示,是一个典型的决策树。决策树包含决策节点、分支和叶节点三部分,其中决策节点代表某个待分类数据集合的某个属性,例如图1的"是否有房"和"是否有车"属性,在该属性上的不同测试结果对应一个分支,每个叶节点表示一种可能的分类结果,例如图1中的"可以"代表可以给贷款人进行贷款,而"不可以"则表示不能给贷款人进行贷款。

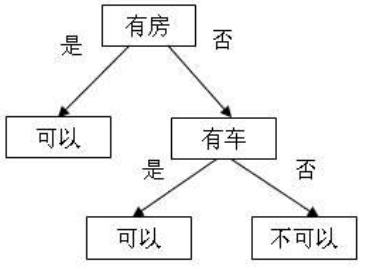

图1 判断是否可以贷款

用决策树进行分类一般有两个步骤:第一步是利用给定的数据集合建立一棵决策树模型;第二步是利用生成的决策树模型对需要分类的样本进行分类。决策树在构建过程中需要重点解决两个问题:

1)如何选择合适的属性作为决策节点去划分数据集合。

2)如何在适当位置停止划分,从而得到大小合适的决策树。

对于何时停止决策树的划分,一般我们认为当属性列表为空,或者数据集中样本都已经分类,此时就可以停止决策树分支的形成及划分,从而得到初始的决策树。而对于第一个问题,不同的决策树算法则给出了不同的解决方法来划分属性,下面依次做出分析。

1 ID3算法分析

1.1 Hunt算法简析

在讨论决策树算法之前,需要先了解一下Hunt算法,此算法是几种经典决策树算法的基础。Hunt算法的基本步骤为:当集合中的数据都属于同一个类,就可以将他们放一起,作为叶子节点;若集合中数据的属性各种各样,就可以先筛选出所有需要的属性,然后从其中选择一个属性,将其分类出来,形成节点,再对剩下的数据进行上述过程,直至全部分离出来。此算法对决策树的递归建立过程如图2所示:

收稿日期:2018-05-07

作者简介:徐梦茹,女,宁夏大学计算机技术在读研究生,主要研究方向:数据库技术;通信作者:王学明(1964—),男,硕士生导师,教授,主要研究方向:数据库技术与应用。

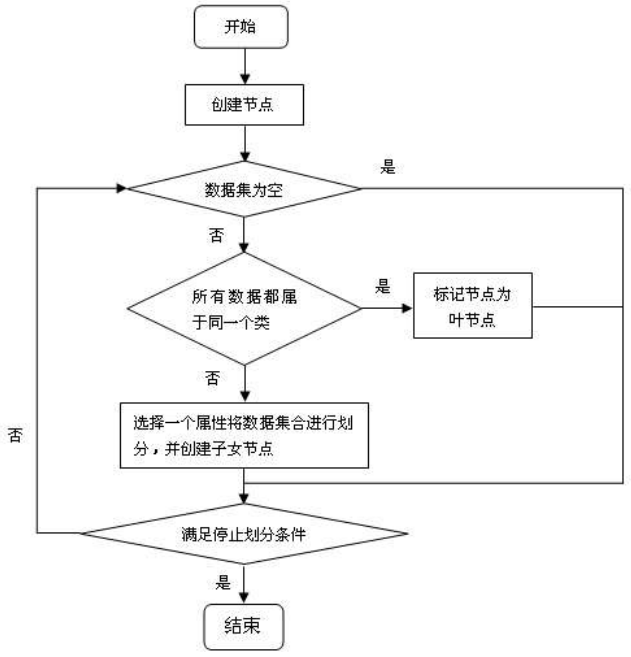

图2 Hunt算法建立过程

决策树的几个经典算法的主要过程就如以上所述，但是根据什么来将属性分离出来，则是值得进一步探讨和改进的问题。

1.2 ID3算法

ID3算法于1986年由Quinlan提出，使用信息增益作为属性选择标准。[2]信息增益是数据划分前后的熵的差值，ID3算法采用使得信息增益最大的特征来划分当前的数据。算法的执行过程如下：首先对所有属性计算其相应的信息增益值，并选择值最大的属性作为决策树的一个节点，之后由该属性的不同取值建立此节点下的分支，再对各分支的子集递归调用以上过程，建立决策树节点和分支，直到剩下的数据集合都属于同一类别为止，最后得到一棵决策树，用来对新的样本进行分类。[3]对应的算法流程图如图3所示：

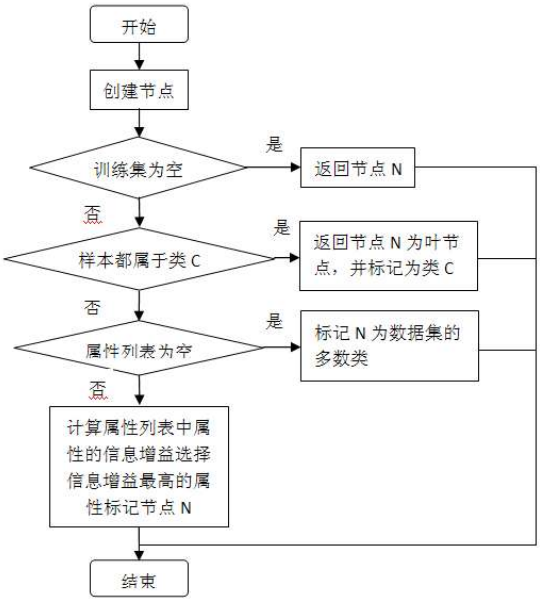

图3　ID3算法流程图

1.2 ID3算法的用途及优缺点

ID3算法是一个典型的决策树分类算法，以一种从简单到复杂的策略遍历空间。ID3算法具有以下的优点：

（1）ID3在建树过程中会包含所有可能的树，建立过程从空的树开始，然后逐步考虑更加复杂的情况。

（2）ID3算法采用自顶向下的搜索策略，分类速度较快。

（3）ID3算法与Hunt算法一样，非常适合处理离散的样本数据。

当然，此算法也有自己的弊端，例如此种算法不能在搜索中进行回溯，因而不能判断有多少其他的决策树也是与现有的训练数据一致的，这样算法只能收敛到局部最优的答案，而不能得到全局最优的答案，且此算法依赖于属性值数目较多的属性，但是属性值较多的属性不一定是分类最优的属性，同时此算法不能处理连续型属性[4]。

2 C4.5算法分析

2.1 C4.5算法

C4.5算法也是用于生成决策树的一种经典算法，是ID3算法的另一种延伸和优化。C4.5算法与ID3算法生成决策树的过程基本相同，但是ID3算法不能处理连续型属性，而C4.5算法可以先离散化连续型属性，然后进行属性的选择分类；属性分类时，ID3算法利用信息增益进行属性分类选择，C4.5算法则用信息增益率进行计算。

在用C4.5算法构造决策树时，信息增益率最大的属性即为当前节点的分裂属性，随着递归计算，被计算的属性的信息增益率就会变得越来越小，到后期则选择相对比较大的信息增益率的条件属性作为分裂属性。C4.5算法的流程图如图4所示：

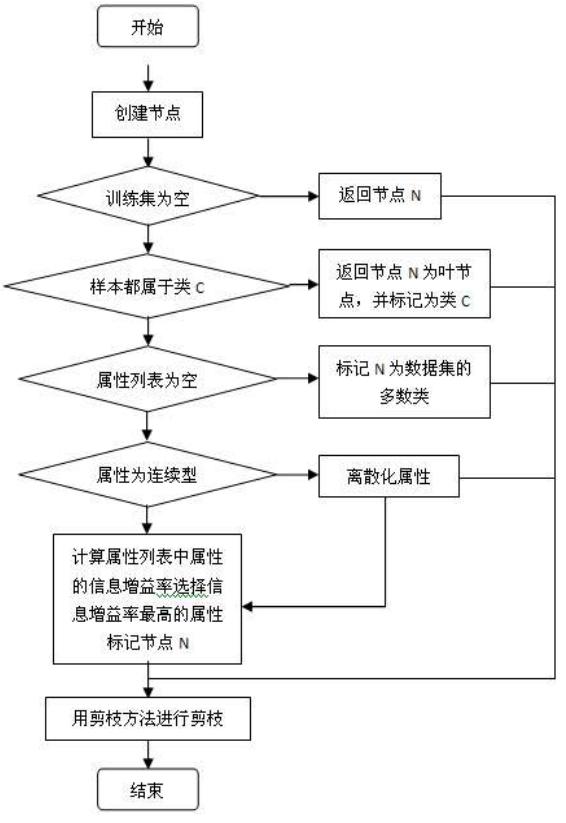

图4 C4.5算法流程图

2.2 C4.5算法的用途及优缺点

C4.5算法的主要优点有：

（1）可以处理数据不完整和连续型属性的数据集；

（2）分类的正确率比较高；

（3）建模速度较快。

C4.5算法的缺点：

（1）在建立决策树的步骤流程中，必须重复地对相应的数据集合进行一次扫描和逐个排序，所以造成了算法的分类效率不高；

（2）C4.5算法的计算公式涉及了大量的对数运算，计算机在进行计算时，会频繁地调用函数，增加了算法的时间开销；

（3）C4.5算法尽管是ID3算法的改进，但是还不能处理很多其他形式的数据集。

3 CART算法分析

3.1 CART算法

CART决策树算法是Breiman于1984年提出的决策树构建算法，采用二元切分法，每次把数据切成两份，分别进入左子树、右子树，并且每个非叶节点都有两个孩子，这样建立起来的树就是二叉树。CART算法采用基尼指数(Gini)来选择要分割的属性，CART每一次迭代都会降低Gini系数，当数据所含的类别越多，此系数就越大，只有当系数越小，那么数据所含不同种类越少，特征就越好，当一个节点中所有样本数据都属于一个类时，Gini系数为0。CART算法的主要过程包含以下三个方面：

（1）二分：在每次判断过程中，都是对观察变量进行二分。算法总是将当前数据集合分割为两个子数据集，使得生成的决策树的每个非叶节点都只有两个分支，因此CART算法生成的决策树是结构简洁的二叉树。算法对于连续特征的处理则与C4.5算法相似。

（2）单变量分割：每次最优划分都是针对单个变量。

（3）剪枝策略：是CART算法的关键点。剪枝过程在最优决策树生成过程中占有特别重要的地位。有研究表明，剪枝过程的重要性要比树的生成过程更为重要，对于不同的划分标准生成的决策树，在剪枝之后都能够保留最重要的属性划分，于是剪枝方法的不同对决策树尤其是最优决策树的生成显得尤为重要。常用的剪枝方法有REP、PEP、CCP等。

3.2 CART算法的用途及优缺点

CART算法虽然可以通过剪枝来避免过拟合的情况，但是此算法效率较低，在构造树的过程中，需要对数据集进行多次的顺序扫描；且此算法只适合于能够驻留于内存的数据集，当训练的数据集大得内存中无法容纳时，程序无法执行。

4 结束语

通过上述对三种算法的分析与比较，可以知道，面对不同的情况可以选择不同的算法来进行决策树的构造，从而来对数据集进行分类，完成数据挖掘过程中的重要的一步。

这三种算法是决策树分类算法中最经典的算法，之后有大量学者对此类算法进行了不同程度的改进，例如有C5.0算法，还有很多基于实际应用的改进算法。通过对分类算法的一步步优化与改进，数据挖掘技术得以更好地发现数据之间的关联，从而可以广泛地应用到实际生活中，对人们的生活和生产产生极为重大的作用。

参考文献：

[1] 米允龙,米春桥,刘文奇.海量数据挖掘过程相关技术研究进展[J].计算机科学与探索,2015,9(06):641-659.

[2] J R Quinlan Induction of Decision Tree [J]. Machine Learning, 1986(1):81-106.

[3] 赵微,苏健民.基于ID3算法决策树的研究与改进[J].科技信息(科学教研),2008(23):383+392.

[4] 马明莉,决策树分类方法及其应用研究[D].河北工业大学,2010,22.

(上接第192页)

完成类内专家集结，得到专家类决策矩阵。

4 结论

通过专家聚类方法，可以有效减少决策者数量，降低决策难度，并且大规模群决策允许的误差内高效地完成决策分析。

参考文献：

[1] 张全.复杂多属性决策研究[M].沈阳：东北大学出版,2008.

[2] 何立华,王栎绮,张连营.基于聚类的多属性群决策专家权重确定方法[J].运筹与管理,2014,23(6): 65-72.