C4.5 决策树分类算法性能分析

◆ 刘瑞玲

摘要： 大数据时代，人们日益关注如何获取海量数据背后的重要信息，探寻其存在的关系和规则，帮助决策者做出准确预测。C4.5决策树分类算法就是对海量数据集进行分类处理的经典算法，利用训练集构造决策树模型，从中提取有意义、有价值的分类规则。论文在研究如何构造决策树的基础上，对C4.5算法的性能进行深入分析。

关键词： 决策树构造；C4.5算法性能分析

一、如何构造决策树

构造决策树关键一步是“分裂属性”。在构造过程中，就某个节点所处的位置按照某一特征属性的不同进行划分、构造不同的分支，其最终目标是让每个因分裂而划分出的子集尽可能的“纯净”，尽量做到分裂子集中等待进一步分类的特征属性最大可能地归属于同一类别。在内容方面，构造决策树的关键性一步就是对“属性选择的度量”。它是选择如何分裂的准则，其决定了分裂点的选定和树的拓扑结构。在属性选择度量算法选取方面，虽然此类算法有很多，但是大多采用自顶向下，逐层分解的层层递归分治法，并综合考虑融合使用不回溯的贪心算法思想来解决问题。

总之，决策树的构造就是通过属性选择度量确定各个特征属性间拓扑结构的过程。该过程只是利用属性选择度量进行选择，而不需要专业领域知识，最好做到将训练集尽可能地划分成不同类别的属性，并且要划分的分支数据量越来越小，决策树的分支划分最理想的停止条件是决策树的叶子结点都有同类标记。因为此刻所有类别的属性都已经被清晰的分开了。倘若构造过程中有噪声或某些数据没有被恰当的表述，又或者决策树生成过程中产生重复子树等，都将造成产生过于复杂的决策树。因此，化繁为简以寻找最优决策树是不可缺少的环节。主要解决以下三个问题：①生成的决策树各叶子结点深度最小；②生成的决策树叶子结点数目最少；③生成的决策树不但叶子节点最少而且每个叶子节点的深度也最小。

决策树构造过程中的输入是一组带有类别标记的例子，构造过程的输出结果不外乎以下两种结果：不是一棵二叉树就是一棵多叉树。二叉树的内部节点一般表示为一个形如 if-then 的逻辑判断，树的两边就是逻辑判断的结果分支。多叉树的内部节点一般表示为属性，边是该属性的所有取值，属性值与边一一映射关系，树的叶子结点都是类别标记。

二、C4.5 分类算法构造决策树的步骤

（一）数据预处理形成决策树训练集。

如果数据是连续型属性变量，首先就应该考虑对其进行离散化，若本身是离散值，此步省略。从数据集中找到该连续型属性值的最小取值、最大取值；插入 K 个数值在区间 [最小值，最大值] 内，以将其等分成 K+1 个小区间；再以插入的各个数值作为分段点，进一步将区间划分成各更小的子区间。

（二）计算每个属性的信息增益 Gain（S）和信息增益率 Gain—Ratio（S）。

信息增益率 Gain—Ratio（S）= Gain（S）/I（S）对于取值连续的属性而言，分别计算每个属性分类的分割点所对应分类的信息增益率，选择信息增益率的最大值对应的 K，作为该属性分类的分裂点，信息增益率最大的属性，被选为当前的属性结点，得到决策树的根结点。

（三）根结点每一个可能的取值对应一个子集或一个划分，对训练样本子集递归地执行前面第 2 步的过程，直到每个划分（子集）中的观测数据在分类属性上都取相同的值，最终生成决策树。

（四）根据训练集构造的决策树提取分类规则，再根据分类规则对海量数据集进行分类处理。

三、C4.5 分类算法的性能分析

C4.5 分类算法是经典的决策树算法之一，从大量训练事例中提取分类规则，自上而下形成决策树。通过某种方式的计算，选出被赋予“最适合”属性的起始点作为根节点，该属性的每个取值对应生成新的分支，每个分支对应生成新的结点。该算法除具备 ID3 的优点之外，还做出以下改进：

（一）C4.5 分类算法克服 ID3 用“信息增益”选择属性的不足，改用“信息增益率”进行属性选择。信息增益率和分裂信息公式如下：

$$\text{GainRatio(S,A)} = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (1)$$

$$\text{GainRatio(S,A)} = \sum_{i=1}^{c} \frac{|s_i|}{|s|} \log_2 \frac{|s_i|}{|s|} \quad (2)$$

（二）C4.5 算法在处理连续性属性变量时有优势。假定对于某一连续性属性变量 D，某节点上数据集的样本总量为 T，将该节点上全部数据样本的具体数值，按升序排列后得到属性值的取值序列 $\{D_1, D_2, \cdots\cdots D_T\}$，从在其上生成的 T-1 个分割点中选择最佳分割点。以第 k（0<k<T）分割点的取值 $M_k = (D_k+D_{k+1})/2$ 为分割点，将该节点上的数据集划分为两个子集。C4.5 优势在于以计算每个属性的信息增益比的方式从计算结果中选择信息增益比最大的分割点来划分数据集，而 ID3 算法不足在于通过计算信息增益的方式，偏向于选择拥有多个属性值的属性作为分裂属性。

（三）C4.5 算法对于缺失数值的处理也比较好。有时我们会遇到等待测试的数据可能缺少某些属性值。处理此种问题的常用

方法有二：一是简单粗暴地赋给节点所对应的训练实例中该属性的最常见值；二是温柔地对节点的每个可能取值赋予一个概率，C4.5 算法采用的此种策略稍显复杂。如给定一个布尔属性值 D，如果节点 N 中包含 3 个 D=1 和 7 个 D=0 的实例，那么 D（k）=1 的概率是 0.3，而 D（k）=0 的概率是 0.7。于是，实例 K 的 30% 被分配到 D=1 的分支，70% 被分配到 D=0 的分支。如果还有其他缺失数据值的属性必须被测试，这些片段样例就可以在树的后续分支中进一步被细分。

（四）C4.5 算法中采用后剪枝技术。为避免数据过度拟合造成无法实现对新样本数据的合理分析及决策树树高无限度的增长，C4.5 算法常利用公式（3）和（4）做以下处理：

$$Pr\left[\frac{f-q}{\sqrt{q(1-q)/N}} > z\right] = c \quad (3)$$

$$e = \frac{f + \frac{z^2}{2N} + Z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \quad (4)$$

计算真实误差率 q 的置信区间上限，z 为对应于置信度 c 的标准差，f=E/N 为观察到的误差率，以真实误差率 q 的置信区间上限为该节点误差率 e 做一个悲观估计，通过判断剪枝前后误差率 e 的大小，进而判定决策树是否要真正实施剪枝，此种技术在实际应用中速度快，精度高。

总之，C4.5 算法的优点是构造树的准确率较高，产生的分类规则容易理解，以信息增益率选择分裂属性克服 ID3 的缺点，并能把连续型的属性变量离散化，构造决策树过程中，能够对缺失属性值的进行处理，采用后剪枝技术等。缺点是虽然能做到处理连续属性变量的训练样本数据集，但是需对训练集反复进行排序及按一定规则顺序扫描，导致算法效率低下。同时，在选择分裂属性时没有充分考虑到条件属性间的相互关联性，单纯计算数据集中每一条件属性与决策属性间的期望信息，有可能影响到属性的正确选择。另外，C4.5 算法会出现训练集过大导致算法程序无法运行的情况，只适合于能够完全驻留于内存的数据集。

参考文献

[1] 范明，范宏建等译 .Introduction to Data Mining 数据挖掘导论 [M]. 北京：人民邮电出版社 ,2011.

[2] 冯少荣 . 决策树算法的研究与改进 [J]. 厦门大学学报（自然科学版）,2007,17(5).

（作者单位：河南轻工职业学院）

（上接第 152 页）

断该数据块是否为“热”，如果是热数据块，则 Label=H，接着判断是否为读倾向的数据块，若为读倾向的数据块，则 Label=R 并将其迁入 NVM。其余情况均根据该数据块参数满足的判断条件将 Label 修改为对应的字母表示，并把数据块保留在磁盘中。

根据算法的基本思想画出该算法的流程图，如图 1 所示。其中 XR 表示该数据块的读操作次数，XW 表示该数据块的写操作次数，T 表示该数据块在系统中的生产期，X0 表示冷热标签区分的阈值，Y0 表示读倾向和无明显倾向标签区分的阈值，T0 表示周期性读写数据的访问周期。

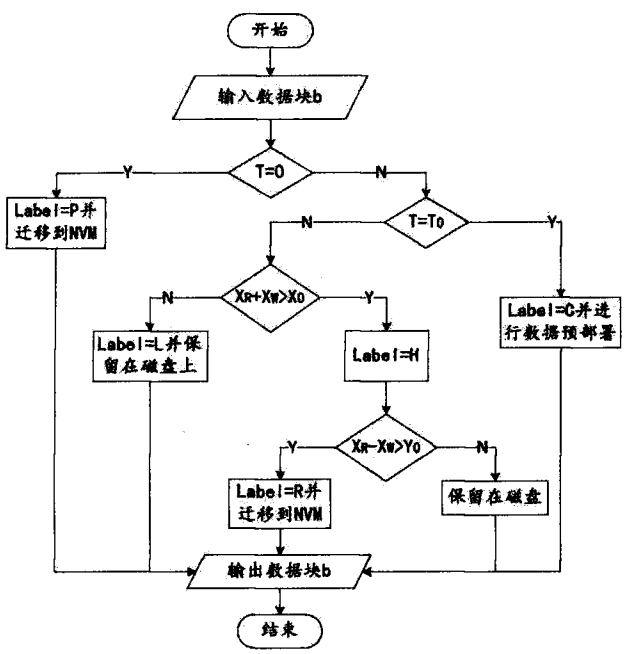

图 1 标签化数据调度算法流程图

四、结语

本文首先介绍了混合存储架构下数据调度机制的研究背景，描述了大数据分析在传统存储架构下存在的瓶颈、新型非易失性存储器的优点以及混合存储架构的优势，最终在混合存储架构下提出了一种基于数据标签化的数据调度策略。

参考文献

[1] 张引，陈敏，廖小飞 . 大数据应用的现状与展望 [J]. 北京：计算机研究与发展 ,2013,50(S2):216-233.

[2]Ramos L, Gorbatov E, Bianchini R. Page Placement in Hybrid Memory System[C].Proc of the Int Confon Super Computing.NewYork:ACM,2011:85-95.

[3]Chen K, Jin P, Yue L. A Novel Page Replacement Algorithm for the Hybrid Memory Architecture Involving PCM and DRAM[M].Network and Parallel Computing,2014:108-119.

[4] 刘巍 . 基于空间局部性的 PCM 和 DRAM 混合内存页面调度算法 [J]. 中国科技论文 ,2014,9(1).

[5] 金培权，郝行军，岳丽华 . 面向新型存储的大数据存储架构与核心算法综述 [J]. 计算机工程与科学 ,2013,35(10).

[6] 徐德 . 嵌入式 Linux 操作系统调度算法改进 [J]. 电脑知识与技术：学术交流 .2011,07(7):1572-1574.

（作者单位：许道强，国网江苏省电力有限公司；夏冬，国网江苏省电力有限公司徐州供电分公司；宋剑枫，国网江苏省电力有限公司南京分公司；葛崇慧，江苏方天电力技术有限公司）