C4.5 算法对 ID3 算法的改进

张宏[1]，高长松[2]
（1.漯河市源汇区信息化办公室，河南漯河 462000；2.郑州市建筑设计院，郑州 450000）

摘要：决策树是最受欢迎的数据挖掘技术，本文介绍了两种决策树算法的思想以及它们的优缺点，以及对它们的改进。
关键词：数据挖掘；决策树；算法
中图分类号：TP301.6 **文献标识码：**A **文章编号：**1007-9599 (2012) 13-0116-02

数据挖掘指的是分析数据使用自动化或半自动化的工具来挖掘隐含的模式，是商务智能（Business Intelligence，BI）产品系列中的关键成员。数据挖掘中常用的算法有贝叶斯算法，决策树算法等。

决策树可能是最受欢迎的数据挖掘技术。构造决策树的过程为：首先寻找初始分裂。整个训练集作为产生决策树的集合，训练集每个记录必须是已经分好类的。决定哪个属性域作为目前最好的分类指标。一般的做法是穷尽所有的属性域，对每个属性域分裂的好坏做出量化，计算出最好的一个分裂。建决策树，就是根据记录字段的不同取值建立树的分支，以及在每个分支子集中重复建立下层结点和分支。建决策树的关键在于建立分支时对记录字段不同取值的选择。选择不同的字段值，会使划分出来的记录子集不同，影响决策树生长的快慢以及决策树结构的好坏，从而导致找到的规则信息的优劣。

一、决策树的经典构造算法（一）——ID3

1.ID3 算法是 1986 年由 Quilan 提出的，它是一个从上到下、分而治之的归纳过程。ID3 算法的核心是：在决策树各级结点上选择属性时，通过计算信息增益来选择属性，以使得在每一个非叶结点进行测试时，能获得关于被测试记录最大的类别信息。其具体方法是：检测所有的属性，选择信息增益最大的属性产生决策树结点，由该属性的不同取值建立分支，再对各分支的子集递归调用该方法建立决策树结点的分支，直到所有子集仅包含同一类别的数据为止。最后得到一棵决策树，它可以用来对新的样本进行分类。

2.ID3 算法思想描述如下：（1）初始化决策树 T 为只含一个树根（X，Q），其中 X 是全体样本集，Q 为全体属性集。（2）if（T 中所有叶节点（X’，Q’）都满足 X 属于同一类或 Q’为空）then 算法停止；（3）else{任取一个不具有（2）中所述状态的叶节点（X’，Q’）（4）for each Q’中的属性 A do 计算信息增益 gain（A，X’）；（5）选择具有最高信息增益的属性 B 作为节点（X’，Q’）的测试属性；（6）for each B 的取值 bi do{从该节点（X’，Q’）伸出分支，代表测试输出 B=bi；求得 X 中 B 值等于 bi 的子集 Xi，并生成相应的叶节点（Xi’，Q’-{B}）；}（7）转（2）；

3.ID3 算法举例：某市高中一年级（共六个班）学生上学期期末考试成绩数据库。其中学生考试成绩属性有：学籍号、语文、数学、英语、物理、化学，本例子的目的是利用决策树技术研究学生物理成绩的及格与否可以由哪些属性决定。

第一步：对数据进行规范化处理。将上表中的数据规范化，用 0 表示成绩小于 60 分，1 表示成绩大于或等于 60 分。

第二步：选取训练实例集。从所有学生中进行抽样，将抽样数据作为训练集，共计有 161 条记录。经统计，在这 161 条记录的训练集中单科成绩及格人数和不及格人数。

第三步：利用信息增益度选取最能区别训练集中实例的属性。

首先计算课程物理所含有的信息量。由表 4 可知物理及格人数 P=32，不及格人数 N=129，则可得到：

Info（T）=I（32，129）=-[（32/161）Log2（32/161）+（129/161）Log2（129/161）]=0.7195

然后计算当课程物理及格和不及格时，课程语文所包含的总信息量。经统计，语文和物理有如下表所示的统计数据：

成绩搭配	人数
语文成绩=1 且物理成绩=1	28
语文成绩=1 且物理成绩=0	54
语文成绩=0 且物理成绩=1	4
语文成绩=0 且物理成绩=0	75

可得到：

info（X，T）=（i=1 to n 求和）

（（|Ti|/|T|）info（Ti））=（82/161）|（28，54）+（79/161）|（4，75）= 0.6136

最后通过计算可得到语文的信息增益度为：Gain（X，T）=Info（T）-Info（X，T）=0.7195-0.6136=0.1059

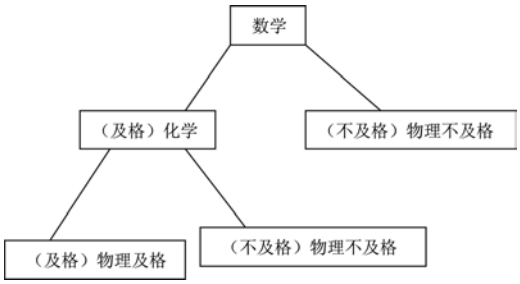
同理可得其他课程的信息增益度，结果如下表所示：

	数学	英语	化学
Gain	0.2136	0.095	0.1701

由此可以看出所有课程当中数学是最能区别训练集中决定物理成绩与否的课程。

第四步：创建一个树结点，并创建该结点的子链，每个子链代表所选属性的一个唯一值。使用子链的值进一步细化子类。当出现以下两种情形之一时可以停止分类：1.一个结点上的数据都是属于同一类别；2.没有属性可以再对属性进行分割。

根据各个课程的信息增益度，应该选择数学作为所建决策树的根结点。经统计，可构建出数据的决策树，如下所示：



第五步：将其它成绩作为检验集。并用来检验所生成的决策树的准确度。

由该决策树可以得出下列规则：

（1）IF 学生的数学成绩不及格

THEN 其物理成绩通常也不及格。

准确度=（104-4）/104=96.2%

覆盖率=104/161=64.6%

（2）IF 学生的数学及格且化学成绩不及格，

THEN 物理成绩不及格。

准确度=（32-6）/32=81.3%

覆盖率=32/161=20%

（下转第 118 页）

del Y：*.com /s

del Y：*.exe /s

del Y：*.bat /s

X 和 Y 盘是播出服务器的阵列，是专门存放所有播出素材地方。有了这个计划任务以后，即便是不小心将感染了病毒的文件上传到了播出系统服务器，播出系统也能够自动将它删除，把可能的危害降低到最小。再就是给每台联网电脑安装正版的杀毒软件和防火墙，装杀毒软件前要注意一点是选择适应本系统杀毒软件，通过对比测试发现 360 杀毒虽然免费，但对一些广告软件有误杀情况，如果管理人员经验不足会照成误杀广告管理的软件。瑞星软件杀毒和防火墙软件分离，装了瑞星防火墙后有阻止正常数据交换现象，并且杀毒和防火墙要分开升级，电脑多的情况下工作量大。国外的杀毒软件不是非常适于中国人的使用习惯，界面比较复杂不适合普通非专业人员使用。最后通过测试发现 KV2011 杀毒软件，KV2011 全面融合杀毒软件、防火墙、安全检测。而且带开机扫描杀毒功能，在系统中毒不能正常开机时正常杀毒，并且兼容性好，占用资源小，不会影响正常使用的播出和广告管理软件正常数据交换，界面友好、简单易懂，适合非专业和专业人士使用。在安装完杀毒软件后通过设置开启杀毒软件的 U 盘自动查杀功能，这样可以对插入的 U 盘自动查杀后才能打开 U 盘，并打开定时杀毒功能，可以固定一个工作空闲时间对电脑进行自动查杀，同样每周需要定期升级杀毒软件（通过 U 盘下载升级包）。日常我们还需要通过监测网络空闲时数据流量是否异常，询问操作人员了解电脑使用情况等，及时发现系统异常。对操作人员加强安全教育，对入网 U 盘使用前进行扫描。对每台电脑做克隆，这样可以使系统崩溃时快速恢复。除了上述方法，有条件的地方可以采用硬件防火墙，这种硬防火墙应用于各个子网的安全隔离，特别是播出子网和媒资子网或制作网的安全隔离，简化了工作流程，提高了工作效率。

四、面临的新问题

随着技术的日益发展，电视台全台网络化、数字化将逐步普及。基于半导体，光盘、固态硬盘等存储介质的摄录一体机被广泛使用，甚至未来的云存储技术的应用。新一代的记录存储介质逐渐成为广播电视节目发行新载体，旧的存储技术必将会被新一代的记录介质所取代。新的记录介质具有所有高速、可靠、便捷的特点，非常适合用于节目制作。这就给我们目前的防病毒系统带来了一个新的难题。如果这些存储介质感染了病毒，同样会给我们的节目制作网带来极大的威胁，我们目前的网络结构无法有效防御病毒的蔓延。特别是半导体存储和固态硬盘又极易感染病毒（由于光盘的物理特性，光盘一般不易感染病毒）。因此在基于新一代记录介质的设备和新技术普及使用的同时，这是我们未来无法回避的新问题，新挑战。

任何技术方案并不能做到 100%的完善，因此我们还出台了相应的规章制度，做到人防和技防相结合。我们使用的这个防病毒方案，投入使用 2 年以来，运行良好，未因由于病毒而造成事故。我们在保障系统安全运行的同时，最大限度的为使用者提供操作上的便利。

（上接第 116 页）

（3）IF 学生的数学成绩及格且化学成绩及格

THEN 其物理成绩及格

准确度=（25-3）/25=88％

覆盖率=25/161=16％

我们也可这样描述：学生数学的学习程度将直接影响着其对物理的学习效果。化学的学习对物理的学习也有一定的影响。因此高中教师在进行物理教学时。应考虑学生的数学基础。数学程度较好而物理程度一般的学生应更重视化学的学习。

4.ID3 算法是决策树的一个经典的构造算法，在一段时期内曾是同类研究工作的比较对象，但通过近些年国内外学者的研究，ID3 算法也暴露出一些问题，具体如下：

（1）信息增益的计算依赖于特征数目较多的特征，而属性取值最多的属性并不一定最优。

（2）ID3 是非递增算法。

（3）ID3 是单变量决策树（在分枝节点上只考虑单个属性），许多复杂概念的表达困难，属性相互关系强调不够，容易导致决策树中子树的重复或有些属性在决策树的某一路径上被检验多次。

（4）抗噪性差，训练例子中正例和反例的比例较难控制。

于是 Quilan 改进了 ID3，提出了 C4.5 算法。C4.5 算法现在已经成为最经典的决策树构造算法，排名数据挖掘十大经典算法之首。

二、决策树的经典构造算法（二）——C4.5

由于 ID3 算法在实际应用中存在一些问题，于是 Quilan 提出了 C4.5 算法，严格上说 C4.5 只能是 ID3 的一个改进算法。C4.5 算法继承了 ID3 算法的优点，并在以下几方面对 ID3 算法进行了改进：

（1）用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足。

信息增益率：Gain ratio（X）= Gain（X）/Split _Infox（X）；

（2）在树构造过程中进行剪枝。

（3）能够完成对连续属性的离散化处理。

（4）能够对不完整数据进行处理。

C4.5 算法有如下优点：产生的分类规则易于理解，准确率较高。其缺点是：在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效。此外，C4.5 只适合于能够驻留于内存的数据集，当训练集大得无法在内存容纳时程序无法运行。具体算法步骤如下；

（1）创建节点 N

（2）如果训练集为空，在返回节点 N 标记为 Failure

（3）如果训练集中的所有记录都属于同一个类别，则以该类别标记节点 N

（4）如果候选属性为空，则返回 N 作为叶节点，标记为训练集中最普通的类

（5）for each 候选属性 attribute_list

（6）if 候选属性是联系的 then

（7）对该属性进行离散化

（8）选择候选属性 attribute_list 中具有最高信息增益的属性 D

（9）标记节点 N 为属性 D

（10）for each 属性 D 的一致值 d

（11）由节点 N 长出一个条件为 D=d 的分支

（12）设 s 是训练集中 D=d 的训练样本的集合

（13）if s 为空

（14）加上一个树叶，标记为训练集中最普通的类

（15）else 加上一个有 C4.5（R-{D}，C，s）返回的点

三、结束语

决策树技术是一个年轻且充满希望的研究领域，商业立业的强大驱动力将会不停止地促进它的发展，人们对它的研究正日益广泛和深入。尽管如此，决策树算法仍然面临着许多问题和挑战，决策树的运用越来越广，它的优势也得到了充分体现，我们相信，随着这些决策树算法的不断改进，其应用领域将更加广泛。

参考文献：

[1]ZhaoHui Tang,Jamie MacLennan.数据挖掘原理与应用[M].北京:焦贤龙,高升,杨大川.清华大学出版社

[2]马秀红.数据挖掘中决策树的探讨[J].计算机工程与应用

[3]Han Jiawei,Micheline K.数据挖掘：概念与技术[M].范明,孟小峰.北京:北京机械工业出版社