ID3 决策树算法的改进研究

满桂云 林家骏 华东理工大学 200237

摘 要

ID3 决策树算法是数据挖掘中最常用的一种方法，但其存在多值偏向性等问题，文中根据相似性原理，引进属性趋近度概念，以描述属性和决策分类属性的分类样本数的趋近程度确定测试属性，构建决策树，并对 ID3 算法和改进算法 T_ID3 算法的多值偏向性问题和测试预测率进行了理论和实验的验证。

关键词

ID3 算法；多值偏向性；属性趋近度；

引言

Quinlan 的 ID3(Iterative Dichotomizer3)算法是把信息熵作为选择测试属性的标准[1]，而每次信息增益的计算很大程度上将受多值偏向性问题影响，即有优先选取取值较多的属性的倾向。多值偏向所带来的问题是，把属性在分类中的重要性与属性取值数多少关联起来，认为取值较多的属性在分类中具有更重要的作用，而取值较多的属性却不总是最优的属性，这就难以判断得到的测试属性究竟是因为本身比较重要还是由于多值偏向取值较多的缘故而得到的。因此下面将对 ID3 决策树算法进行改进[2-4]。

1.ID3决策树算法的改进——基于属性趋近度的测试属性的选择

定 义 ：

属性趋近度：描述属性的分类样本数趋近决策分类属性分类样本数的程度。

设当前训练样本集：S，有 s 个样本；A 是训练样本集的一个描述属性，它的取值为 $A_1, A_2, \cdots, A_n$；C_j 为决策分类属性，具有 m 个不同值，定义 m 个不同类 C_j(j:1, 2,…,m)，C_j 是 s 当中类 C_j 的样本数，对 $C_1, C_2, \cdots, C_j, \cdots, C_m$ 从大到小进行排序得到新的序列 $C_1', C_2', \cdots, C_j', \cdots, C_m'$，按照 C_j' 的大小顺序定位矩阵纵列顺序，并记 α_{ij} 为 A 的取值为 A_i 时决策为第 j ′类的记录数，得到描述属性和决策分类属性样本取值个数分布情况矩阵：

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{i1} & a_{i2} & \cdots a_{ij} \cdots & a_{im} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{vmatrix} \quad (1)$$

属性趋近度算法具体描述由下面伪代码给出：

```
输入：描述属性和分类属性样本取值个数分布情况
输出：描述属性相对于决策属性的趋近度
Tendency(A1,A2,…,An,C1,C2,…,Cj,…,Cm)
    //A1,A2,…,An为描述属性A的n种取值
    //C1,C2,…,Cj,…,Cm为决策属性C的m种取值
{
    Int count; //count记录属性匹配条数
    While(m>0& n>0)
    {
        t=Max{aij}
            若取得最大值的个数多于1个，则比较相应描述属性取值的样本总数∑ai，选择余数少的描述属性取值样本数目
        count=count+t;
        删去选取的最大值所在的行与列;
        j'=j'-1;
        m=m-1;
        n=n-1;
    }
    Return count/s;
}
```

属性趋近度是判断由单个描述属性取值能得出正确的决策分类属性的可能性大小。把采用属性趋近度作为测试属性的选择标准对各个描述属性进行计算的决策树算法命名为 T_ID3 算法。

2.ID3算法和T_ID3算法的多值偏向性分析

2.1 多值偏向性问题的理论

首先，设 A 是某训练样本集的一个描述属性，它的取值为 A_1，A_2，…A_n，同时为该样本集“创造”另外一个属性 A′，它的取值为 A_1'，A_2'，…，A_n'，A'_{n+1} 并且令 $A_i'=A_i(i=1,2,\cdots,n-1)$。因此 A′等价于把 A 的第 n 个取值 A_n 拆分为 A_n' 和 A'_{n+1} 得到的。显然，拆分属性的某一个取值并不会增加该属性对分类任务的重要性，即 A′不会比 A 更重要。

然后，把决策树算法分别作用在属性 A 和属性 A′上，如果决策树算法的属性选取标准在属性 A′上的取值恒大于在属性 A 上的取值，则说明该算法具有多值偏向问题；如果决策树算法的属性选取标准在属性 A′上的取值与在属性 A 上的取值没有确定的大小关系，则说明该决策树算法不具有多值偏向问题。

这种理论分析方法的优点：首先，在此分析方法中的属性 A 和属性 A′实际上是同一个属性，即它们对于分类任务的重要性是相同的，因此在此分析方法中不需要使用领域的专家知识来判断属性之间的相对重要性。其次，属性 A′是通过拆分属性 A 得到的，即属性 A′的取值多于属性 A，这就为评估决策树算法的多值偏向提供的方便。

2.2 ID3算法的多值偏向分析

在生成决策树时，ID3 算法采用信息增量作为属性选取的标准，属性的信息增益可表示如下：

信息增益 $gain(A)=I-E(A)$ (2)

其 中 ：

信息熵 $I=-\sum_{j=1}^{m} p(C_j)\log_2(p(C_j))$ (3)

条件熵 $E(A)=\sum_{i=1}^{n} p(A_i)I(A_i)=$

$-\sum_{i=1}^{n} p(A_i)\sum_{j=1}^{m}[p(C_j/A_i)\log_2(p(C_j/A_i))]$ (4)

$P(C_i)$ 是任意样本属于 C_j 的概率，即 $P(C_j)=\frac{c_j}{s}$

$P(C_j/A_i)$ 是 S 中描述属性为 A_i 决策分类属性为 C_j 的样本的概率

把 ID3 算法分别作用在 A′和 A 上得

$$gain(A')-gain(A)=$$
$$(I-E(A'))-(I-E(A))=E(A)-E(A') \quad (5)$$

因为 $A_i=A_i'(i=1,2,\cdots,n-1)$，所以有如下结果：

$$gain(A')-gain(A)=p(A_n)I(A_n)-\sum_{i=n}^{n+1} p(A_i')I(A_i')$$

(5′)

把 $I(A_n)$、$I(A_n')$ 和 $I(A'_{n+1})$ 的表达式代入，两边同时除以 $P(A_n)$ 得：

$$\frac{gain(A')-gain(A)}{p(A_n)}=-\sum_{j=1}^{m} p(C_j/A_n)\log_2 p(C_j/$$

$$A_n)] + \frac{p(A_n^{'})}{p(A_n)} \sum_{j=1}^{m} p(C_j/A_n^{'}) \log_2 p(C_j/A_n^{'})$$

$$+ \frac{p(A_{n+1}^{'})}{p(A_n)} \sum_{j=1}^{m} p(C_j/A_{n+1}^{'}) \log_2 p(C_j/A_{n+1}^{'})] \quad (6)$$

在此，我们只考虑两种情况，故取n=2，为了表达的方便，设定如下参数：

$$r = \frac{p(A_n^{'})}{p(A_n)} \quad x = p(C_1/A_n) \quad p = p(C_1/A_n^{'})$$

$$q = p(C_1/A_{n+1}^{'})$$

故有下面的结果

$$\frac{gain(A^{'}) - gain(A)}{p(A_n)} = r(p\log_2 p + (1-p)\log_2(1-p))$$

$$+ (1-r)(q\log_2 q + (1-q)\log_2(1-q))$$

$$-(x\log_2 x + (1-x)\log_2(1-x)) \quad (6')$$

令 $f(x) = (x\log_2 x + (1-x)\log_2(1-x))$ (7)

则：

$$\frac{gain(A^{'}) - gain(A)}{p(A_n)} = -f(x) + rf(p) + (1-r)f(q)$$

(6'')

由于 $x = rp + (1-r)q$

故：$f(x) = f(rp + (1-r)q)$ (7')

通过分析可知f(x)为凹函数，故有：

$$f(x) = f(rp + (1-r)q) \le rf(p) + (1-r)f(q)$$

(8)

因此，可得：

$$\frac{gain(A^{'}) - gain(A)}{p(A_n)} = -f(x) + rf(p) + (1-r)f(q) \ge 0$$

(9)

故：$gain(A^{'}) - gain(A) \ge 0$ 恒成立，即ID3算法存在多值偏向问题。

2.3 T_ID3算法的多值偏向分析

T_ID3算法在测试属性的选择上，主要依据的是属性趋近度。经过纵列调整过的描述属性A′和决策分类属性样本取值个数分布情况矩阵为：

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{i1} & a_{i2} & \cdots & a_{im} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \\ a_{(n+1)1} & a_{(n+1)2} & \cdots & a_{(n+1)m} \end{vmatrix} \quad (10)$$

这里式（1）与式（10）中前（n-1）行完全相等，，即 $\alpha_{ij} = a_{ij}$ $(i = i^{'} \mid i = 1,2,\cdots,n-1)$ 而 $\alpha_{nj} = a_{nj} + a_{(n+1)j}$

根据属性趋近度的计算现在可分两种情况讨论：

（1） $m \le n$

根据算法可知，这时两个矩阵count都会进行m次迭加，将属性A每次得到的t命名为t_j，属性A′每次得到的t命名为$t^{'}_j$，除了能确定$t_1 \ge t^{'}_1$，进行第一次的取值所在分量的行列删除后，其他(m-1)次两个属性的t值都不能确定大小关系。即count(A)与count(A′)的大小关系不能确定，而训练集样本总数s不变，所以两个属性的趋近度大小关系也不能确定。

（2） $m \ge n+1$

这时属性A的矩阵count要迭加n次，属性A′的矩阵count要迭加（n+1）次，同样在每次得到t值的时候，除了能确定$t_1 \ge t^{'}_1$，其他两个属性的t值也都不能确定大小关系，尽管属性A′要比属性A多迭代一次，但因为前n次迭代两个属性的大小已经不能确定大小关系，所以属性A′第（n+1）次也不能决定确定两个count值的关系。

3.两种算法预测率试验比较

通过以上分析，可知T_ID3算法不具有多值偏向问题，有效地避免了弱相关属性因为取值过多而覆盖掉强相关但是取值较少的属性，避免了判定树向多值的弱相关属性倾斜。

为了在更大范围、更多数据集中对比算法的预测率，本文选用了一些某高校学生信息管理系统中的数据集以及加州大学Irvine分校（UCI）维护的一个用于分类算法的测试的机器学习知识库的数据集。关于此知识库的详细介绍以及本文所使用的样本集的详细介绍可以进入网址：http://www.ics.uci.edu/~mlearn/MLSummary.html进行相关查询。

表1提供了本文在实验中所使用的样本集以及相关描述：样本集名称、样本个数、离散属性个数，最后列出了两种算法的预测准确率。

（1）当描述属性数量较多，对于测试属性的选择计算上就需要进行大量的循环计算，这时属性的多值偏向问题对于测试属性的选择上产生的影响就较大，进而对整个决策树的层次构建上产生影响，从实验条目3、4上可以看出在这种情况下，两者的预测率有较大的差别。

（2）样本的数目较小时，构建和剪枝决策树相对规模较小，产生的规则两者相近度较高，预测率差别较小，样本数目较大时，预测率差别相差也相对大些，可参考文献2、5[5-6]。

4.结论

本文针对ID3算法多值偏向性问题引入了一种基于相似性理论的属性趋近度计算方法，选择描述属性和决策分类属性样本数具有最大趋近度的描述属性作为当前测试属性构造决策树。接着对ID3算法和T_ID3的多值偏向性进行了论证分析，然后把T_ID3算法和ID3算法运用于学生信息管理系统和加州大学Irvine分校（UCI）机器学习知识库中的部分数据样本集，可以明显地看出，在预测准确性方面，T_ID3算法也优于ID3算法。

参考文献

[1] Holte RC Very simple classifaction rules Perform well on most commonly used datasets [J], Machine Learning, 1993,11:63—90

[2] I.Kononenko,SJ.Hong,Atribute selection for modeling,Future Generation Computer Systems 13: 18 1-195,1997

[3] 曲开社,成文丽,王俊红.计算机工程与应用.2003,39(25):104-107

[4] 韩松来,张辉,周华平.基于关联度函数的决策树分类算法.计算机应用.2005.25(11):2655-2657

[5] 梁循.数据挖掘算法与应用.北京:北京大学出版社.2006

[6] 邵峰晶，于忠清.数据挖掘-算法和原理.北京:中国水利水电出版社.2004

作者介绍

满桂云(1979—),女,硕士研究生,研究方向:数据处理.

表1 两种算法预测率

	样本集名称	样本个数	决策分类属性个数	离散描述属性个数	ID3算法	T_ID3算法
1	毕业生就业分析评定表（2005）	3341	3	5	81.5.5%	85.7%
2	学生成绩总评评定表（2002-2006）	12863	4	7	68.2%	75.5%
3	学生性格分析表（2006）	3860	8	20	70%	81%
4	Statlog DNA	2000	3	60	61%	75%
5	Contraceptive	1473	3	7	86.2%	88.7%