数据挖掘 ID3 决策树分类算法及其改进算法①

罗雨滋, 付兴宏

(辽阳职业技术学院 信息工程系, 辽阳 111003)

摘 要: 本文通过数据挖掘对传统 ID3 决策树分类算法及性能进行分析研究, 利用高等数学中的微分理论知识, 改进和优化了 ID3 算法中的运算速度和选择测试属性偏向问题, 并进一步给出了改进算法的伪代码.

关键词: ID3 算法; 信息熵; 信息增益; 决策树

Data Mining ID3 Decision Tree Classification Algorithm and its Improved Algorithm

LUO Yu-Zi, FU Xing-Hong

(Liaoyang Vocational Technology Institute Department of Information Engineering, Liaoyang 111003, China)

Abstract: This thesis analyzes the traditional ID3 decision tree classification algorithm and performance by using data mining, using the theory of differential knowledge in higher mathematics, improves and optimizes the operation speed and the test attribute selection bias problem in ID3 algorithm, and further gives the pseudo-code of the improved algorithm.

Key words: ID3; information entropy; information gain; decision tree algorithm

随着计算机技术的快速发展, 超大规模数据库的出现, 海量数据的快速访问以及统计学方法在数据处理领域应用, 激发了人们对海量数据在提取有用信息和知识的数据挖掘技术的开发、应用和研究. 数据挖掘(Data Mining)是指从大量结构化和非结构的数据中提取有用的信息和知识的过程, 是知识发现的有效手段[1].

1 ID3分类算法

ID3 算法是由 Quinlan 首先提出的. 该算法是以信息论为基础, 以信息熵和信息增益为衡量标准, 首先检测所有属性, 选择信息增益值最大的属性产生决策树节点, 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建立决策树节点的分支, 直到所有子集仅包含同一个类别的数据为止, 最后得到一个决策树, 从而实现对新的样本的的归纳分类. [2]

假定 X 为训练集, 是一个离散型随机变量, 其概率发布为: $p(x)=p(X=x), x\in X$; X 的目标属性 $U=\{u_1,u_2,\cdots,u_m\}$, 则 X 的熵 $H(x)=-\sum_{x\in X}p(x)\log_2 p(x)$, U_i 在所有样本中出现的频率为 $P_i(i=1,2,\cdots,m)$, 则该训练集 X 所包含的信息熵(Entropy)定义为:

$$Entropy(X)=Entropy(p_1,p_2,\cdots,p_{m)}=-\sum_{i=1}^{m}p_i\log_2 p_i$$

$$Entropy_i(s)=\sum_{i=1}^{n}\frac{|S_i|}{|S|}Entropy(S_i)$$

其中|Si|(i=1,2,……,n)为样本子集 Si 中包含的样本数, |S|为样本集中包含的样本数. 信息增益是划分前样本数据集的熵和划分后的样本数据集的熵的差值. 即

信息增益:

$$Gain(X,A)=Entropy(x)-Entropy_i(s)$$

信息增益越大, 说明使用的属性 A 划分后的样本子集越纯, 越有利于分类.

2 ID3算法性能分析[3]

通过观察搜索空间和搜索策略, 我们从算法的描述和原理, 将其优、缺点总结如下:

优点: ID3 算法以一种自顶向下的从简单到复杂

① 基金项目:辽宁省教育科学“十二五”规划立项课题(JG12EB052)
收稿时间:2013-04-01;收到修改稿时间:2013-05-02

的爬山策略遍历假设空间, ID3 算法在搜索的每一步用当前的所有训练样本, 以信息增益作为测试属性的技术, 使得在每个非叶子节点进行测试时, 能获得关于被测数据最大的类别信息, 离散化连续属性,降低了对个别训练样例错误的敏感性, 测试次数较少, 分类速度较快. 由于引入了信息熵的概念, ID3 算法能得出结点数最少的决策树. ID3 算法与最基本的决策树算法一样, 非常适合处理离散值样本数据, 算法的计算时间与样本个数, 特征个数, 结点个数三者之积呈线性关系.

缺点: ID3 算法每当在树的某一层选择了一个属性进行测试后, 由于不进行回溯, 算法是局部最优的, 而不是全局最优的答案. ID3 算法的计算方法依赖于属性值数目较多的属性, 但他不一定是最优的属性. 而且 ID3 算法不能增量训练样例, 每增加一次实例都要从新构造新的决策树, 此算法不适合增量式学习任务. 每选择一个分裂结点, ID3 算法利用信息熵的值来判断数据集中的分裂属性, 都要进行对数计算, 数据量大, 而且选择偏向取值较多的属性, 生成效率就会受到影响, 增大了决策树的计算成本.

总的来说, ID3 算法以从简单到复杂的爬山策略遍历假设空间, 从空树开始, 逐步考虑更复杂的假设. 对于处理大规模的数据集, ID3 算法不失为一种数据挖掘和机器学习中获取知识的有用工具.

3 改进算法的理论基础

高等数学中的泰勒公式和麦克劳林公式改进简化信息熵的计算复杂度, 能够减少算法生成的决策树时间.

根据高等数学中的微分理论知识(当$|x-x_0|$很小时)

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+o(x-x_0)$$

麦克劳林公式

$$f(x)=f(0)+f'(0)x+\frac{f''(0)}{2!}x^2+\cdots+\frac{f^{(n)}(x)}{2!}x^n+R(x)$$

(其中 $R(x)=\frac{f^{(n+1)}(x)}{(n+1)!}(x-x_0)^{n+1}$)

麦克劳林公式的近似公式:

$$f(x)\approx f(0)+f'(0)x+\frac{f''(0)}{2!}x^2+\cdots+\frac{f^{(n)}(0)}{n!}x^n$$

假设 $f(x)=\ln(1+x)$ 时, 且当 $x\to 0$ 时, $\ln(1+x)\approx x$

4 改进后属性信息熵的运算公式推导

由等式

$$I(p_i,n_i)=-\frac{p_i}{p_i+n_i}\log_2\frac{p_i}{p_i+n_i}-\frac{n_i}{p_i+n_i}\log_2\frac{n_i}{p_i+n_i}$$

$$E(A)=\sum_i^n\frac{p_i+n_i}{p+n}I(p_i,n_i)$$

整理得

$$E(A)=\sum_i^n\frac{1}{(p+n)\ln 2}(-p_i\ln\frac{p_i}{p_i+n_i}-n_i\ln\frac{n_i}{p_i+n_i})$$

因为 $\frac{1}{(p+n)\ln 2}$ 是常量,

设 $G(A)=\sum_i^n(-p_i\ln\frac{p_i}{p_i+n_i}-n_i\ln\frac{n_i}{p_i+n_i})$

因为 $\ln(1+x)\approx x$,

$$\ln\frac{p_i}{p_i+n_i}=\ln(1-\frac{n_i}{p_i+n_i})\approx-\frac{n_i}{p_i+n_i}$$

同理, $\ln\frac{n_i}{p_i+n_i}\approx-\frac{p_i}{p_i+n_i}$

则 $G(A)=\sum_i^n(p_i\frac{n_i}{p_i+n_i}+n_i\frac{p_i}{p_i+n_i})=\sum_i^n\frac{2p_in_i}{p_i+n_i}$

改进后的属性信息熵公式为:

$$H(A)=(\sum_i^n\frac{2p_in_i}{p_i+n_i})N$$

其中 N 是属性值的个数.

$H(A)$ 因只有加法、乘法和除法运算, 比 $E(A)$ 中多次进行的对数运算相比, 节省了运算时间. 通过对改进属性信息熵 $H(A)$ 函数计算, 选出最大信息增益作为分裂节点, 对每个属性的信息熵引入权值 N, 使得计算出的信息熵依赖于属性的取值个数 N, 于是克服了 ID3 算法选择测试属性的偏向问题.

5 改进的ID3算法与原ID3算法的有效性比较分析[4]

5.1 时间复杂度分析

ID3 算法信息熵的计算公式为:

$$E(A)=\sum_i^n\frac{p_i+n_i}{p+n}I(p_i,n_i)$$

其中

$$I(p_i,n_i)=-\frac{p_i}{p_i+n_i}\log_2\frac{p_i}{p_i+n_i}-\frac{n_i}{p_i+n_i}\log_2\frac{n_i}{p_i+n_i}$$

ID3 算法的信息熵的时间复杂度为 $O(n^2 \log_2 n)$.

而改进后算法的信息熵公式为的 $H(A)=(\sum_{i}^{n}\frac{2p_in_i}{p_i+n_i})N$，它的时间复杂度为 $O(n^2)$，改进后的算法公式比原算法公式在时间复杂度上有所提高.

5.2 改进后算法与原算法分类性能分析

为了使实验数据具有一般性，采用每组数据集多次实验，再取平均值的数学统计方法. 在同一个计算机系统下完成改进后算法与原算法在分类准确度和建决策树所用时间消耗上进行比较. 如表 1 所示:

表 1　分类准确度的比较

	记录数量	ID3的分类准确度	改进后ID3的分类准确度
数集1	500	67.20%	69.80%
数集2	600	68.40%	71.20%
数集3	700	71.60%	74.70%
数集4	800	74.80%	77.50%
数集5	900	76.90%	80.10%

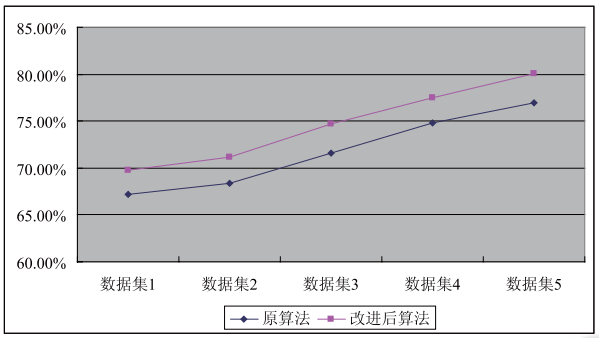

图 1　改进后的 ID3 算法与原算法分类准确度的比较

从图 1，可以看出改进后的 ID3 算法的准确度比原算法的要高，随数据集数量的增大，分类准确率也增大，符合数学中近似线性递增关系. 这种趋势与高等数学中泰勒公式的理论原理吻合，并且呈逐步稳定趋势. 通过多次、大量的数据集测试实践证明了，改进后的算法能够提高和改进了原算法的分类准确度，有一定的应用价值.

从图 2，我们可以看到改进后的算法比原算法在建立决策树的时间上要消耗的少，并且随着数据集数据的增加，时间差呈近似线性递增，数据集数据量越大，改进后的算法越能比原算法消耗的时间差值越明显，说明改进后的算法对大数据集更具有应用价值和优越性.

表 2　构建决策树的时间比较

	记录数量	ID3建决策树所用平均时间（ms）	改进后ID3建决策树所用平均时间（ms）
数集1	500	146.2	118.5
数集2	600	183.4	150.6
数集3	700	209.8	171.1
数集4	800	237.6	190.3
数集5	900	264.9	205.6

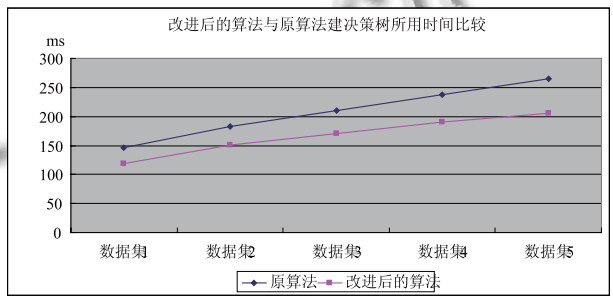

图 2　改进后 ID3 算法与原算法建决策树所用时间比较

6　改进的ID3算法的伪代码

输入：训练样本数据 Smaples，数据属性集合 Attributclist.

输出：一棵决策树[5]

① 创建结点 N;

② if 样本 Samples 全部属于同一个类别 C then

③ 创建一个叶节点 N，并标记类标号为 C;

④ return;

⑤ else

⑥ for (i=0; i<numputattrirbute;i++)

若 i 大于属性个数时，循环结束

⑦获取属性值子串 numvalues=domain{i}.size()

⑧ for(j=0;j<numvalues;j++)

if alreadyUsedTocompose(node,i,j) then

判断属性值是否为分类属性值，若是，本次循环结束

⑨ else Subset=getSubset(node.data,i,j)

计算属性值的信息熵

Complement=getComplement(node.data,subset)

C1=calculate Entropy(subset)

C2=calculte Entropy(complement)

Entropy=(C1*subset.size()+C2*complement.si

(下转第 187 页)

两个指令转发到相应的控制节点，观察控制节点的 LED 灯 D1 点亮, D2 处于熄灭状态. 几乎同时，电灯处于点亮状态，表明完成了控制执行任务. 这时计算机监控终端控制状态中的风机和天窗控制状态显示为绿色，表示两个设备处于打开状态.

② 修改温度的期望范围为 30ºC-35ºC，由于温室实际温度(25ºC)低于期望范围，这时的温度报警状态同样是红色报警，控制节点的 LED 灯 D2 点亮，D1 熄灭，同时电灯熄灭，表明控制节点关闭了后面的执行设备. 这时计算机监控终端上的控制状态区中的风机和天窗状态同时为红色，表明这 2 个设备已关闭.

③ 手动点击计算机监控终端界面上的“风机开关”按钮，此时下面的风机状态显示由红色(关闭)变成绿色(打开)，汇聚节点显示“WIND_OPEN_CMD”，表明打开风机的控制指令已经传送到风机控制节点，这时控制节点上的 LED 灯 D1 点亮，D2 熄灭. 电灯瞬间点亮，表明控制节点执行了打开开关控制指令.

④ 以同样的方式对其它控制节点进行测试，得到同样的测试结果.

⑤ 以同样的方式对湿度控制效果进行测试，得到相应的控制结果，表明系统的能够按照预定的控制策略完成机电设备的开关任务.

3 结语

本文将无线传感器网络应用于温室测控环境中，发挥其组网快捷、应用方便、覆盖区域广的特长. 通过在温室内部署一定数量的传感器节点，使其自组织构成无线传感器网络. 该系统不仅实现了温湿度信息采集监测功能，同时还完成了控制机电设备调节温室环境变化的工作，使其成为一个完整的闭环测控系统. 解决了以往温室测控系统中的大量布线问题，使温室测控技术向无线化、无人化方向发展迈进了一步.

参考文献

1 高建平,赵龙庆.温室计算机控制与管理技术的发展概况及在我国的应用前景.计算机与农业,2003(2):12–15.
2 李萍萍,毛罕平,王多辉,谢明岗,陈庆芳.智能温室综合环境因子控制的技术效果及合理的环境参数研究.农业工程学报,1998,(3):9–14.
3 张军,吴建锋.基于无线传感器网络的温湿度检测系统.杭州电子科技大学学报,2010,12.
4 高守伟,吴灿阳.ZigBee 技术实践教程—基于 CC2430/31 的无线传感器网络解决方案.北京:北京航空航天大学出版社,2009:57–59.

(上接第 138 页)

ze())/mumdata

⑩ if entropy<bestEntropy then
Selected=true;
BestEtropy=entropy;
selectedAtrribute=i;
selectedValue=j;

7 结束语

本文主要利用高等数学中的泰勒公式和迈克劳林公式，将 ID3 算法中每一次分裂节点都要进行的对数运算简化，从而加快了运算速度，提高了创建决策树的效率. 并将引入的权值 N 乘以信息熵，解决了 ID3 算法选择测试属性的偏向问题，使得信息熵的计算与属性的取值个数相关. 同时对改进算法与原算法进行了有效性分析，最后给出了优化和改进的 ID3 算法的伪代码.

参考文献

1 苏新宁.数据仓库和数据挖掘.北京:清华大学出版社，2006:115–139.
2 蒋盛益.数据挖掘原理与实践.北京:电子工业出版社，2011:51.
3 蒋盛益.数据挖掘原理与实践.北京:电子工业出版社，2011:54.
4 薛薇,陈欢歌.Clementine 数据挖掘方法及应用.北京:清华大学出版社,2007:120–130.
5 朱玉全,杨鹤标.数据挖掘技术.南京:东南大学出版社,2006:90–130.