ID3 算法及其改进*

徐 雯 张 扬
（中国地质大学计算机学院 武汉 430074）

摘 要 文章对 ID3 算法的基本概念和原理进行了相应的详细阐述以及解释说明，并针对 ID3 算法倾向于取值较多的属性的缺点，引进信息增益率对 ID3 算法作了改进，并通过实验对改进前后的算法进行了比较，实验表明，改进后的算法行之有效。

关键词 决策树 ID3 算法 信息增益 增益率
中图分类号 TP301.6

ID3 Algorithm and Its Improvement

Xu Wen Zhang Yang

（Department of Computer Science College, China University of Geosciences, Wuhan 430074）

Abstract The article largely describes and explains the basic concept and the principle of decision tree ID3 algorithm, however, decision tree ID3 algorithm tends to choose attributes that are sampled more often, so focusing on this deficiency. The paper introduce gain ratio to improve the ID3 algorithm, and then compare the original algorithm with the modified algorithm by experiment, whose result proves the improved algorithm more efficient than original one.

Key words decision tree, ID3 algorithm, information gain, gain ratio
Class Number TP301.6

1 引言

数据挖掘是在大量的数据中发现潜在的、有价值的模式和数据间关系的过程[1]。而在处理实际的各类问题中，决策树 ID3 算法会偏向于选择取值较多的属性，本文针对这个不足之处，对 ID3 算法作了改进，并通过实验验证改进后的算法是有效的。

2 ID3 算法原理

设 S 为一个包含 s 个数据样本的集合，其类别属性可以取 m 个不同的值，对应于 m 个不同的类别 C_i，$i\in\{1, 2, \cdots, m\}$。假设 S_i 为类别 C_i 上的样本个数，那么要对一个给定数据对象进行分类所需要的信息量为：

$$I(S_1, S_2, \cdots, S_m)=-\sum_{i=1}^{m} p_i\log_2(p_i) \qquad (1)$$

再设一个属性 A 取 v 个不同的值$\{a_1, a_2, \cdots, a_v\}$。利用属性 A 可以将集合 S 划分为 v 个子集$\{S_1, S_2, \cdots, S_v\}$，其中 S_j 包含了 S 集合中属性 A 取 a_j 值的数据样本。若属性 A 被选为测试属性，设 S_{ij} 为子集 S_j 中属于 C_i 类别的样本数。那么利用属性 A 划分当前样本集合所需要的信息熵为：

$$E(A)=\sum_{j=1}^{v}\frac{S_{1j}+S_{2j}+\cdots+S_{mj}}{s}I(S_{1j}, \cdots, S_{mj}) \qquad (2)$$

其中$\frac{S_{1j}+S_{2j}+\cdots+S_{mj}}{s}$是由所有子集中属性 A 取 a_j 值的样本数之和除以 S 集合中的样本总数。$E(A)$计算结果越小，就表示其子集划分结果越“纯”。而对于一个给定子集 S_j，它的信息为：

$$I(S_{1j}, \cdots, S_{mj})=-\sum_{i=1}^{m} p_{ij}\log(p_{ij}) \qquad (3)$$

利用属性 A 对当前分支节点进行相应样本集合划分所获得的信息增益就是：

* 收稿日期：2009 年 5 月 27 日，修回日期：2009 年 6 月 28 日
作者简介：徐雯，女，硕士研究生，研究方向：数据挖掘与智能计算。张扬，女，硕士研究生，研究方向：计算机技术。

$$Gain(A)=I(S_1,S_2,\cdots,S_m)-E(A) \quad (4)$$

显然, $E(A)$的值越小则信息增益 $Gain(A)$的值越大, 说明选择测试属性 A 对于分类提供的信息越大, 选择 A 之后对分类的不确定程度越小, ID3 算法即采取以 A 作为测试属性的选取标准分割训练实例集最终生成决策树。在开始时, 所有的数据都在根节点, 属性都是种类字段, 所有记录用所选属性递归的进行分割属性的选择是基于信息熵, 每次分类都是选择信息增益最大的那个属性进行分裂, 停止分割的条件是一个节点上的数据都是属于同一个类别, 没有属性可以再用于对数据进行分割, 这时再重复上述分裂直到所有的叶节点都是纯的。

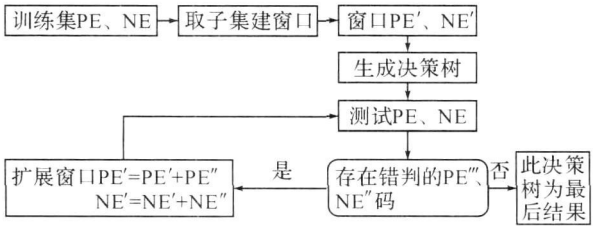

图 1　ID3 算法流程图

3　改进的 ID3 算法

3.1　ID3 算法的不足

ID3 采用自顶向下不回溯的策略搜索全部的属性空间, 它建立决策树的算法简单, 深度小, 分类速度快。传统的ID3 算法选择某个属性 A 作为测试属性的原则是使得信息增益 $Gain(A)$最大。研究表明这种方法存在一个弊端: 算法往往偏向于选择取值较多的属性, 因为加权和的方法使得实例集的分类趋向于抛弃小数据量的数据元组, 然而取值较多的属性却不总是最优的属性, 即按照使熵值最小和信息增益最大的原则被 ID3 算法列为应该首先选取的属性在现实情况中却并不那么重要, 也就是说对这些属性进行测试不会提供太多的信息, 例如: 在股票市场, 个股分析需要对某些少量的元素组有足够大量重视, 而用 ID3 则会忽略个股的重要属性[7~8]。

3.2　改进的思路

3.2.1　引入用户兴趣度

给定 $a(0\leqslant a\leqslant 1)$为用户对不确定知识的兴趣度, a 的大小由决策者根据先验知识或领域知识来确定。它是一个模糊的概念, 通常指关于某一事务的先验知识, 包括领域知识和专家建议, 具体到决策树学习中则是指在决策树训练过程中除了用于生成和修改决策树的实例集之外的所有影响决策树规则生成和选择的因素[2]。

3.2.2　C4.5 也是对 ID3 的改进算法

ID3 算法不能处理数值属性、残缺值等问题, C4.5 算法从这几个方面很好地弥补了 ID3 算法的不足, 并且还加入了剪枝等处理方法, 使得精度有了很大的提高。

3.3　实现的算法

本次设计中实现的是采用增益率(gain ratio)代替信息增益的改进方法: 因为采用信息增益的方法会倾向于选择拥有较多可能属性值的属性, 为了弥补这一缺陷, 所以采用增益率来代替信息增益。增益率是考虑了属性分裂数据集后所产生的子节点的数量和规模, 而忽略任何有关类别的信息, 只计算一个内在的信息值: Sinfo, 也就是分裂信息量。然后用信息增益除以这个分裂信息量得到增益率的值, 即 $Gain(A)/\mathrm{Sinfo}$, 再根据每个属性的增益率的值来选出增益率最大的那个属性作为分类的属性, 再依次方法选出后面分类的属性, 当训练集里包含多个拥有相同属性值, 但是属于不同类别的样本时, 该分类过程不能停止, 只有当所有的叶节点都是纯的, 也就是当叶节点包含的实例拥有相同的类别时停止分类。

4　实验分析

实验所用数据给出了影响夏天天气舒适度的几个相关指标的数据集合, 有 4 个属性: 穿衣指数、温度、湿度、风力, 并有舒适和不舒适两个类别。可以得到决策树如图 2 所示。根据图 2 中从根节点到叶节点的路径及数据集所包含记录的多少, 可以得到表 1。

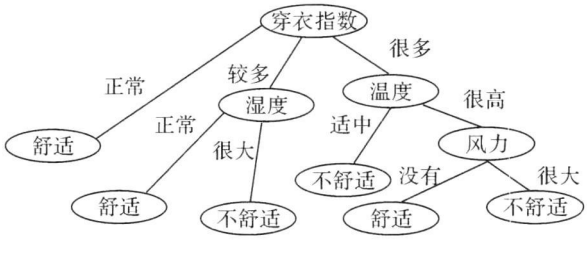

图 2　ID3 算法生成的决策树

当用增益率代替信息增益来用改进后的算法 NEWID3 测试这个数据集时, 通过 Weka 计算出的决策树和分析结果和 ID3 算法是一样的, 也是属性穿衣指数排在首位。表 2 基于穿衣指数的信息增益及增益率的计算结果。

从表 2 中可以看出, 虽然 NEWID3 算法和 ID3 算法分别是用增益率和信息增益来进行属性的分类, 但是在本次测试中属性穿衣指数仍是排在首位,

通过计算可以知道接下来的属性选择依然是湿度、温度、风力,但是从表中可以看出利用增益率来进行属性的分类,穿衣指数和湿度的值更接近一些,这样穿衣指数作为首选项的优势有所降低,用另外的数据集做测试,会有两者更加接近的情况。

表 1　以表格形式表示的分类规则

	穿衣指数	温度	湿度	风力	类别
规则 1	较多		正常		舒适
规则 2	较多		很大		不舒适
规则 3	较多	正常			不舒适
规则 4	较多	很高		很大	不舒适

表 2　穿衣指数的信息增益及增益率的计算

	穿衣指数	温度	湿度	风力
信息增益	0.484	0.027	0.060	0.005
分裂信息量	1.581	0.993	0.993	1.559
增益率	0.306	0.027	0.062	0.003

在本数据集中,为了看出利用信息增益和增益率的不同,可以采用一种极端的方法,就是在数据集中加一个标识码属性,使数据集中标识码这个属性对于每一个实例都存在一个不同的属性值。可以得到决策树如图 3 所示。

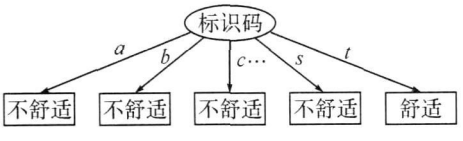

图 3　标识码属性的树桩

在上面几个图中,图 3 是对标识码属性进行分裂而产生的树桩。给定这个属性的值,要说明它的类别所需的信息量是:

info([0, 1])+info([0, 1])+info([0, 1])+…+info([0, 1])+info([0, 1])

由于 20 项中的每一项都是 0,所以信息量是 0。这个结果表明:标识码属性能够区别每个实例,所以它能确定类别,而不会出现任何模棱两可的情况。所以这个属性的信息增益就是在根节点上的信息量,即 info([9, 11])=0.993 位。它比其他任何属性上获得的信息增益要大,毫无疑问标识码将被选为分裂属性。

在图 3 的情形中,每个分支只含 1 个实例,所以分裂后的信息值为:

info(1, 1, 1, …, 1)=−1/20 *log(1/20) *20=4.322 位

则标识码属性的增益率为 info([9, 11])/info([1, 1, 1, …, 1])=0.230

由表 2 可以看出,标识码的增益率小于穿衣指数的增益率,所以穿衣指数仍为分类的首选属性,其次再是标识码、湿度、温度、风力这些属性。

通过这两个算法的对比,虽然在测试原数据集时两个算法的分类是一样的,但是通过加入标识码属性再进行测试时可以发现,增益率的获得是考虑了属性分裂数据集后产生的子节点的数量和规模,所以在实际的开发过程中,当数据集被分裂的子集个数少时,其增益率会有所提高。然而,在某些情况下增益率修正法补偿过度,会造成倾向于选择某个属性的原因,仅仅是因为这个属性的内在信息值比其他属性要小很多。一个标准的弥补方法是选择能够得到最大增益率的属性,而且那个属性的信息增益至少要等于所有属性的信息增益的平均值。

5　结语

随着技术的进步和发展,数据挖掘和机器学习已获得了令人瞠目结舌的进步。统计学、机器学习、信息理论以及计算技术的有机结合,创建了一门具备坚实数学基础的强大工具的完备科学。ID3 算法是基于决策树学习中最重要的一种算法,近年来有大量的学者围绕该算法作了很多研究,并提出了各种各样的改进方法。本文针对 ID3 算法倾向于取值较多的属性的缺点,改进了 ID3 算法,并通过实验证明算法是有效的。

参 考 文 献

[1] 董琳,邱泉,于晓峰,等译. 数据挖掘实用机器学习技术(第二版)[M]. 北京:机械工业出版社,2006

[2] 曲开社,成文丽,王俊红. ID3 算法的一种改进算法[J]. 计算机工程与应用,2003,25:104~107

[3] D. Riano. Learning rules within the framework of environmental sciences[C]. in: BESAI'98, 1998: 151~165

[4] 雷蕾,秦侠,姚小丽. 数据挖掘技术在环境科学中的应用[J]. 环境与可持续发展,2006,(6):8~10

[5] Dzeroski, S., J. Grbovic, W. Walley, et al. Using machine learning techniques in the construction of models. ii. data analysis with rule induction[J]. Ecological Modelling, 1997, 95(1): 95~111

[6] 孙敏. 数据挖掘及在绿地生态评价中的应用研究[D]. 广西大学,2005,6

[7] 王建华,王菲,黄国建. 数据挖掘技术研究的现状及展望[M]. 香港:Global-Link 出版社

[8] Quinlan J R. Induction of decision tree[J]. Machine Learning, 1986,(1):81~106

[9] 郭景峰,等. 决策树算法的并行性研究[J]. 计算机工程,2002,28(8):77~78