决策树 ID3 算法研究

杜威铭 冉 羽
(西南科技大学计算机科学与技术学院,四川 绵阳 621010)

【摘 要】用于分类的数据挖掘技术的方法有很多,在这些方法中决策树凭借其易理解、效率高等优点而占有重要地位。ID3 算法是决策树构造方法中最为常用的实现方法,它在数据分类和预测领域得到广泛应用。本文重点总结了决策树方法中的 ID3 算法的研究现状,在详细介绍 ID3 算法原理、算法性能的基础上,总结了 ID3 算法以及给出了 ID3 算法的改进算法。

【关键词】数据挖掘;ID3 算法;ID3 优化算法;决策树

中图分类号:TP181 文献标识码:A 文章编号:2095-2457(2018)11-0145-002
DOI:10.19694/j.cnki.issn2095-2457.2018.11.062

Reaserch of ID3 algorithm in decision tree

DU Wei-ming RAN Yu

【**Abstract**】There are many methods used to categorize data mining techniques, in which decision trees play an important role by virtue of their ease of understanding and efficiency. ID3 algorithm is the most commonly used method in decision tree construction method, which is widely used in the field of data classification and prediction. This paper focuses on the research status of ID3 algorithm in decision tree method. Based on the detailed introduction of ID3 algorithm principle, application example and algorithm performance, this paper summarizes the ID3 algorithm and the improved algorithm of ID3 algorithm.

【**Key words**】Data mining; ID3 algorithm; ID3 optimization algorithm; Decision tree

0 绪论

随着软硬件技术的发展,数据库技术也经历了多次演变,在信息数据量剧增的环境下,对于海量的数据以及数据背后的隐藏信息,我们期望通过更高层次的方法,寻找出模型与规则,帮助我们利用数据进行分析与决策。因此,数据挖掘技术应运而生并越发受人重视,高校、研究所与公司在该方面的研究也做了很大的投入。决策树[1]方法作为数据挖掘中的一种重要的方法,也受到了诸多关注。下面将介绍决策树方法中的 ID3[2](Interactive Dichotomic Version 3)算法。

1 ID3 算法研究

1.1 ID3 算法简介

J·Ross Quinlan 等人在 1986 年提出 ID3 算法。其核心是“信息熵”,在创建决策树的过程中,依次查询样本集合中的每个属性,选取出具有最大信息增益值的属性,将该属性作为测试属性与划分标准。通过该标准将原始数据集合划分成多个更纯的子集,并在每个子集中重复这个过程,直到分支子集中的所有样本无法继续分割,即样例属性属于同一类别,此时一棵决策树便创建完成。

1.2 ID3 算法原理

ID3 算法原理包含了信息论[3]中的信息熵和信息增益。信息熵作为属性类别的不纯性度量,熵值越高属性的纯度越低,反之越高。信息增益通过信息熵相减求得,它反映了该属性特征在总体数据集中的重要程度。信息增益和信息熵分别有以下数学定义[4]:

信息增益:已知样本集合 T,决策树的训练属性 B,假设$\{b_1,b_2,\cdots,b_n\}$是属性 B 的 n 个不相同的值,通过这 n 个值可以将集合 T 划分为 n 个子集$\{a_1,a_2,\cdots,a_n\}$。则每个子集对应一个通过节点生成的分支。属性 B 对集合 T 划分的数学期望为:

$$E(B)=\sum_{j=1}^{n}\frac{a_{1j}+\cdots+a_{nj}}{a}+H(a_{1j}+\cdots+a_{nj})$$

可通过属性 B 的样本除 a 求得相应子集的权值$\frac{a_{1j}+\cdots+a_{nj}}{a}$。最终的信息增益为:

$$Gain(B)=H(a_1,a_2,\cdots,a_n)-E(B)$$

信息熵:已知样本集合 T,其样本个数为 a,有 n 个各不相同的值(属性),便可划分 n 个类 $M_i(i=1,2,\cdots,n)$。假设样本 M_i 的个数为 a_i,则样本 Mi 分类所对应的信息熵求值公式为:

$$H(a_1,a_2,\cdots,a_n)=-\sum_{i=1}^{n}p_i\log(p_i)$$

其中,p_i 表示样本集合中 Mi 的概率值,有 $M_i=\frac{a_i}{a}$,log 底数的取值反映了信息熵的单位。

1.3 ID3 算法描述

下面给出 ID3 算法的伪代码描述:

输入:离散型决策属性集合 D 和样本集合 S。

输出:函数 Create_Tree(D , S)返回一棵决策树。

Function Create_Tree(D , S)

作者简介:杜威铭(1994.03.02—),男,汉族,四川广安人,本科在读,西南科技大学计算机科学与技术学院软件工程卓越 1501 班学生,研究方向为数据挖掘,机器学习。

冉羽(1994.01.04—),女,汉族,四川广安人,硕士研究生,西华师范大学教育学院心理健康教育专业,研究方向心理健康教育。

Begin
(1)创建结点 N;
(2)if S 都在同一个类 C then
(3)return N 作为叶子结点,记为类 C;
(4)if D=NULL then
(5)return N 为叶子结点,记为 S 中最普通的类;
(6)选择 D 中拥有最大信息增益的属性 A;
(7)标记结点 N 为 A;
(8)for each A 中的未知值 value
(9)从结点 N 长出一个条件为 A=value 的分枝;
(10)设 Bvalue 是 S 中 A=value 的样本子集;
(11)if Bvalue=NULL then
(12)添加一个叶子结点,记为 S 中最普通的类;
(13)else 添加一个从 Create_Tree (Bvalue,D - {A}) 返回的结点。
End

1.4 ID3 算法应用实例

以表 1 数据为训练样本集,介绍 ID3 算法如何生成一棵决策树。

表 1 训练样本集

NO	条件属性				感冒
	体温	流鼻涕	肌肉疼	头疼	
1	高	是	是	否	是
2	很高	否	否	否	否
3	很高	是	否	是	是
4	正常	是	是	是	是
5	正常	否	否	是	否
6	高	是	否	否	是
7	高	是	否	是	是
8	很高	是	是	否	是
9	高	否	是	是	是
10	正常	是	否	否	否
11	正常	是	否	是	是
12	正常	否	是	是	是
13	高	否	否	否	否
14	很高	否	是	否	是
15	很高	否	是	否	是
16	高	否	否	是	是

(1)信息熵的计算

用 p 表示感冒,n 表示未感冒,初始训练样本感冒人数为 12,未感冒人数为 4,因此可求得分类前训练集的信息熵:

$H(X)=I(p,n)=-(12/16)\log_2(12/16)-(4/16)\log_2(4/16)=0.8113$bits

(2)条件熵的计算

选择属性体温作为划分属性,体温的取值集为{正常,高,很高},其中正常体温人数为 5,高体温人数为 6,很高体温人数为 5,则有:

体温正常:$p_1=3,n_1=2,I(p_1,n_1)=0.9710$bits
体温高:$p_2=3,n_2=2,I(p_2,n_2)=0.6500$bit
体温很高:$p_3=3,n_3=2,I(p_3,n_3)=0.7219$bits

此时可以算出用体温属性划分训练集后熵的期望值为:

E(体温)$=(5/16)I(p_1,n_1)+(6/16)I(p_2,n_2)+(5/16)I(p_3,n_3)=0.7728$bits

(3)信息增益的计算

Gain(体温)$=0.8113-E$(体温)$=0.0385$bits,同理可求得:

Gain(流鼻涕)$=0.5117$bits
Gain(肌肉疼)$=0.0038$bits
Gain(头疼)$=0.0359$bits

选择具有最大信息增益的流鼻涕属性作为根节点进行决策树的创建,引生出流鼻涕和不流鼻涕两个分枝,在流鼻涕分枝,求得新划分的信息增益:

Gain(流鼻涕,体温)$=0.1992$bits
Gain(流鼻涕,肌肉疼)$=0.0924$bits
Gain(流鼻涕,头疼)$=0.1379$bits

选体温作为流鼻涕分枝的结点,在不流鼻涕分枝,求得新划分的信息增益:

Gain(不流鼻涕,体温)$=0.0157$bits
Gain(不流鼻涕,肌肉疼)$=0.0157$bits
Gain(不流鼻涕,头疼)$=0.0032$bits

我们发现存在相同的信息增益,则选择分枝少的属性作为不流鼻涕分枝的结点,即肌肉疼属性。之后重复上诉步骤,完成下图 1 决策树的创建。

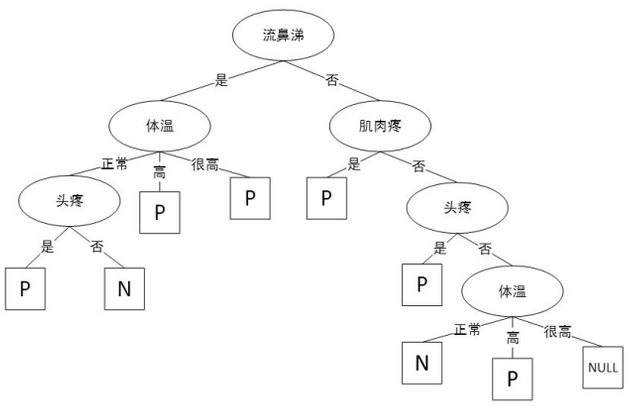

图 1 ID3 决策树

1.5 ID3 算法优缺点

通过 ID3 算法的伪代码描述与实际使用,我们可以发现 ID3 算法是一种采用自顶向下、贪婪策略的算法。其优势主要有以下 3 点:①自顶向下的搜索方式降低了搜索次数,提升了分类速度。②ID3 算法原理清晰,算法思路简单易懂,易于实现。③由于决策树在创建的过程中都使用目前的训练样本,而不是根据独立的训练样本递增的做出判断,在很大程度上降低了对个别训练样本错误的敏感性[5]。ID3 算法不足主要有以下四点:①ID3 算法对噪声数据相对敏感。②ID3 算法循环调用过程中会产生大量的对数运算,随着样本集合、属性以及属性取值个数的增加,对数运算次数将会大大增加,从而降低了 ID3 算法的运算效率,产生了极大的时间开销。③ID3 算法在建树过程中不进行回溯导致生成的决策树节点只是局部最优的,相对于全局,往往不是我们所期待的结果,即如多值偏向所得结果并不总是最优结果。④ID3 只能分类离散型数据,对于非离散型数据需要经过预处理才能使用。

2 ID3 改进算法

(下转第 103 页)

$$F(x)=n\times\frac{1-e^{-\alpha x}}{1+e^{-\alpha x}} \quad (式\ 1)$$

利用如式 1 所示的非线性量化因子替代基本控制器中的线性量化因子，等同于对模糊规则或隶属度函数进行调整，相当于在控制的不同误差阶段，采用不同的论域范围或不同分辨率的隶属度函数，误差较小时，提高隶属度函数的分辨率，从而有效消除静态误差，提高控制精度。二次网温度控制系统框图如图 1 所示。

2.2.2 二次网供、回水压力控制

为保证有足够的末端压差，对回水压力和供水压力进行控制。回水压力采用补水定压的控制方式，若回水压力低于给定值，则开启补水泵补压，压力达到要求时停止。供水压力通过改变循环泵转速进行 PID 调节，当热用户开大热量阀门，用热量增加时，必然造成管网压力降低，压力传感器采集到的压力值低于设定值，则控制器根据偏差的大小，向变频器发出信号，调整提高循环泵的转速，从而增加供水压力，保证末端压差，以保证末端用户正常采暖。

供、回水压差的设定值依据供、回水温差指标设定。压差设定过大，流量过大，温差减小，回水温度升高，造成热资源浪费。反之，压差设定过小，热流量不足，造成回水温度过低。

3 监控系统整体架构[3]

集中供热监控系统实现换热站数据的采集、无线传输和远程监控。利用可编程控制器 PLC 对换热站数据进行采集及控制。PLC 通过串口和无线数传模块 GPRS-DTE 相连，在 DTU 模块中提前设置服务器地址，则由 DTU 模块通过无线网络将串口采集的数据透传至服务器，不受地域分布影响，兼具实时性和准确性。监控中心服务器接收换热站数据，并对数据进行可靠规范的存储、记录、显示、报警等处理，亦可向 PLC 发送指令、传送数据，进行控制系统等参数设置等。监控中心服务器接入互联网，在网络覆盖区域，客户端可随时通过互联网登录监控系统了解换热站运行情况，根据操作权限实施相应操作。监控系统整体架构图如图 3 所示。

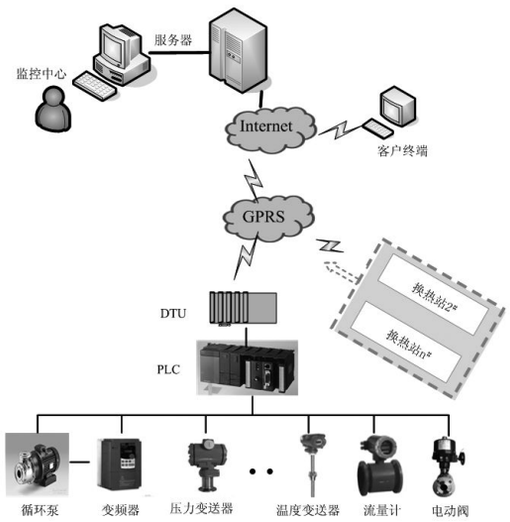

图 3 监控系统整体架构图

4 结论

采用模糊控制策略的换热站变流量控制方案，可以在热用户用热量不断调整变化的情况下高效运行，保证供需平衡，避免能源浪费。在此基础上搭建的城市集中供热监控系统，通过互联网实现对换热站及热用户实时连续监控，可在不影响供热质量的前提下提高供热企业的生产效率，避免人力资源浪费。

【参考文献】

[1]李凯.变流量城市集中供热分布式控制系统[D].河北科技大学,2008.

[2]H.F.WANG,X.M.LIANG,Z.Y.LIU. Simulation Research of Fuzzy Controller with Sigmoid Scaling Factor, Advanced materials research.2013,9:264-267.

[3]余保付.基于物联网的智能供热系统控制技术研究[D].天津工业大学,2016.

(上接第 146 页)

由于 ID3 算法的不足与局限性，J·Ross Quinlan 于 1993 年对原算法进行了改进并提出了 C4.5 算法。该算法将信息增益率作为划分标准，解决了 ID3 算法无法处理连续特征属性的问题，同时降低了计算的复杂度，提升了分类效率。研究者还提出了如下改进算法：基于分类矩阵的 ID3 算法改进、基于粗糙集的 ID3 算法改进、基于粒计算的 ID3 算法改进等、基于相关系数的决策树优化算法、基于神经网络的分类改进算法、基于朴素贝叶斯与 ID3 算法的决策树分类、粗糙模糊决策树归纳算法等[6]。

3 总结与展望

随着决策树分类法再次受到人们重视，并被广泛的研究和使用。作为决策树中经典算法，ID3 算法使用信息增益作为分割标准，凭借其分类速度快、实现方式简单等优点，成为了具有适用与研究价值的示例学习算法与知识获取的有效工具。目前，ID3 应用领域广，如医学中的病症分类预测和基因与高分子序列分析、商业活动中的市场分析和人力资源管理、教育行业中的成绩分析、高校管理等。同时，研究者们也在不断对 ID3 算法进行优化与改进，提升了分类效率，获得了更好的分类结果。在当前大数据技术背景下，会有更多 ID3 改进算法被提出，ID3 算法也会在更多的领域得到应用。

【参考文献】

[1]Jiawei Han,Micheline Kamber. Datamining Concepts and Techniques 范明,孟小峰,译.数据挖掘概念与技术,机械工业出版社,2001.

[2]Quinlan J R. Induction of decision trees" Machine Learning[J]. in Data:Goals and General Description of the IN L.EN System." in, 1986:257--264.

[3]陈文伟.数据仓库与数据挖掘教程.清华大学出版社, 2006-8.

[4]杨洋.决策树 ID3 算法及其改进[J].软件导刊,2016,15(08):46-48.

[5]李华.基于决策树 ID3 算法的改进研究[D].电子科技大学,2009.

[6]杨霖,周军,梅红岩,杜晶鑫.ID3 改进算法研究[J].软件导刊,2017,16(08):21-24.