一种基于灰色关联度的决策树改进算法

叶明全[1],胡学钢[2]
YE Ming-quan[1],HU Xue-gang[2]

1.皖南医学院 计算机教研室,安徽 芜湖 241000
2.合肥工业大学 计算机与信息学院,合肥 230009
1.Computer Staff Room,Wannan Medical College,Wuhu,Anhui 241000,China
2.Institute of Computer & Information,Hefei University of Technology,Hefei 230009,China
E-mail:ymq@wnmc.edu.cn

YE Ming-quan,HU Xue-gang.Inproved decision tree algorithm based on grey weighted correlated degree.Computer Engineering and Applications,2007,43(32):171-173.

Abstract: In the process of constructing a decision tree,the criteria of selecting partitional attributes will influence the eficiency of classification.The paper analyses the shortcoming of current algorithms for improved ID3 which are inefficient to acertain degree and subjective to measure the attributes which are important or not.Therefore,an efficient and reliable algorithm is proposed by introducing grey weighted correlated degree.The main idea of the approach algorithm is as follows:firstly through grey relational analysis,the relation model of character attributes and classification of the data mining problem is established,then using grey weighted correlated degree to modify information gain of attributes which has many values but not important.Compared with other improved ID3 by an example,the experiment has proved that the improved ID3 algorithm based on grey relational degree is efficient.
Key words: decision tree;classification;ID3 algorithm;grey weighted correlated degree

摘 要:在构造决策树的过程中,分裂属性选择的标准直接影响分类的效果。分析了现有改进的ID3算法不同程度地存在学习效率偏低和对多值属性重要性的主观评测等问题,提出一种高效而且可靠的基于灰色关联度的决策树改进算法。该算法通过灰色关联分析建立各特征属性与类别属性之间的关系,进而利用灰色关联度来修正取值较多但非重要属性的信息增益。通过实验与其它ID3改进算法进行了比较,验证了改进后的算法是有效的。
关键词:决策树;分类;ID3算法;灰色关联度
文章编号:1002-8331(2007)32-0171-02 **文献标识码**:A **中图分类号**:TP301.6

1 引言

决策树算法是机器学习领域的一种重要方法,常用于数据的分类和预测[1]。决策树算法的核心问题是选取在树的每个结点要测试的属性,争取能够选择出最有助于分类实例的属性。为了解决这个问题,ID3算法[2]引入了信息增益的概念,并使用信息增益的多少来决定树的不同结点需要测试的属性。但这种做法存在着对取值情况较多的属性有所偏袒的问题即多值偏向问题。所谓多值偏向,是指决策树算法在选择分裂属性时,倾向于优先选取取值较多的属性。在实际问题中属性取值较多的属性不一定是最优的。多值偏向所带来的问题,是把属性在分类中的重要性跟属性取值不同个数的多少关联起来,最终可能导致从数据集中归纳出不准确的知识。为避免多值偏向问题,出现了许多决策树的优化算法,如基于关联度函数的决策树算法[3]、基于统计估计的决策树算法[4],但改进算法在选择合适的分裂属性时,没有考虑到属性的信息熵;还有学者提出在计算信息熵时引入用户兴趣度[5]和优化参数[6],这二种算法均要求用户反复测试训练集或由决策者根据先验知识及领域知识来确定用户兴趣度和优化参数,造成改进算法效率低、主观性强等缺点。

1982年,由中国学者邓聚龙教授创立的灰色系统理论[7],是一种研究少数据、贫信息不确定性问题的新方法。灰色关联分析是灰色系统理论的一个重要组成部分,其基本思想是根据数列的几何关系或曲线的相似程度来判别因素间的关联程度。

本文在分析文献[5]中决策树改进算法缺陷的基础上,提出了一种新的基于灰色关联度的决策树改进算法。最后本文通过一组训练数据集进行实验,并与其它改进算法进行了比较。结果表明,本文提出的改进算法是切实可行的,算法效率高,且从根本上解决了多值偏向问题。

2 灰色关联度

灰色关联分析是灰色系统分析方法中的一种,是依据各因

基金项目:安徽省自然科学基金(the Natural Science Foundation of Anhui Province of China under Grant No.2005KJ094)。
作者简介:叶明全(1973-),男,硕士,副教授,主要研究方向为数据挖掘、医院信息系统;胡学钢(1961-),男,博士,教授,主要研究方向为知识工程、数据挖掘。

素数列曲线形状的接近程度做发展态势的分析。它以各子因素序列与母因素序列的有关数据为基础计算母子因素的关联度，用关联度来描述母子因素间关系强弱、大小和次序。关联度的几何含义为子因素序列与母因素序列几何曲线的相似与距离程度，如果两序列曲线几何形状相似，距离接近，两者关联度大，反之，两者关联度就小。灰色关联分析吸收了距离空间的量化特性和灰色预测点集拓扑空间的整体比较特点，升华为灰关联空间，建立起整体比较机制，在各领域中应用广泛[8,9]。

设一训练集 S 有 n 个样本，m 个特征属性记为 $X_i(i=1,2,\cdots,m)$，类别属性记为 Y，则 n 个样本的各特征属性值构成一数列：$X_i=\{X_i(1),X_i(2),\cdots,X_i(n)\}$，其中 $i=1,2,\cdots,m$；n 个样本的类别属性值构成一数列：$Y=\{Y(1),Y(2),\cdots,Y(n)\}$。则特征属性数列 X_i 与类别属性数列 Y 在第 k 个点（样本）的灰色关联系数定义为：

$$\xi_i(k)=\frac{\min\limits_i\min\limits_k|Y(k)-X_i(k)|+\rho\max\limits_i\max\limits_k|Y(k)-X_i(k)|}{|Y(k)-X_i(k)|+\rho\max\limits_i\max\limits_k|Y(k)-X_i(k)|} \quad (1)$$

其中 $\min\limits_i\min\limits_k|Y(k)-X_i(k)|$ 为两级最小差，$\max\limits_i\max\limits_k|Y(k)-X_i(k)|$ 为两级最大差，ρ 为分辨系数，在(0,1)中取值，一般取 0.5。综合各特征属性序列点（样本）($k=1,2,\cdots,n$)的关联系数，得到整个各类别属性数列$\{Y(k)\}$和特征属性数列$\{X_i(k)\}$的灰色关联度，即为：

$$RLD(i)=\frac{1}{n}\sum_{k=1}^{n}\xi_i(k)\ (i=1,2,\cdots,m) \quad (2)$$

灰色关联分析对样本量的大小和分布规律没有特殊的要求，且有别于传统分析中常用的因素两两对比的模式，而是将各因素统一置于系统之中进行比较与分析，考虑了各因素间相关性。

3 基于灰色关联度的决策树算法

在决策树建立过程中，关键是如何选择属性进行划分样本数据。Quinlan 于 1986 年提出 ID3 算法是以信息论中的信息熵为基础、信息增益为标准来选取测试属性[2]。ID3 及其衍生 C4.5 算法被人们认为是标准决策树学习算法中最优秀的算法，并且大量的改进性研究也都是基于它们的核心策略即信息熵进行的。ID3 算法首先检测训练集所有的特征属性，选择信息增益最大的属性产生决策树根结点，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。

设 S 是 s 个训练样本数据的集合。假定类别属性具有 m 个同值，定义 m 个不同类别 $C_i(i=1,\cdots,m)$。设 s_i 是类别 C_i 中的样本个数，则对当前训练集 S 进行分类所需的期望信息：

$$Info(S)=Info(s_1,s_2,\cdots,s_m)=-\sum_{k=1}^{n}p_i\log2(p_i) \quad (3)$$

其中 $p_i=s_i/s$，是训练集 S 中任意一个样本对象属于类别 C_i 的概率。

设一个属性 A 具有 k 个不同值$\{a_1,a_2,\cdots,a_k\}$。属性 A 将当前训练集 S 划分为 k 个子集$\{S_1,S_2,\cdots,S_k\}$；其中 S_j 中的样本在属性 A 上具有相同的值 $a_j(j=1,2,\cdots,k)$。设 s_{ij} 是子集 S_j 中属于类别 C_i 的样本数，则属性 A 划分当前训练集 S 的所需要的信息熵：

$$E(A,S)=\sum_{j=1}^{k}\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s}Info(s_j) \quad (4)$$

其中，$\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s}$ 被称为第 j 个子集的权值。

$Info(A,S)$的值越小，子集划分纯度越高。对于给定的子集 S_j 进行分类所需的期望信息：

$$Info(s_j)=Info(s_{1j},s_{2j},\cdots,s_{mj})=-\sum_{i=1}^{m}p_{ij}\log_2(p_{ij}) \quad (5)$$

其中 $p_{ij}=s_{ij}/s_j$ 是 S_j 中任一样本属于类别 C_i 的概率。

利用属性 A 对当前训练集进行划分将获得的信息增益：

$$Gain(A,S)=Info(S)-E(A,S) \quad (6)$$

显然，式(4)的值越小则式(6)的值越大。研究表明 ID3 算法往往偏向于选择取值较多的属性，然而属性取值较多的属性却不总是最优的属性。按照使信息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：Bratko 研究小组在研究判断病情的各种因素时，用 ID3 确定“病人的年龄（有 9 种值）”为应首先判断的属性，即靠近决策树的根结点，但实际中医学专家却认为这个属性在判断病情时没那么重要。

因此，要解决多值偏向问题，需要解决二个关键问题：(1) 如何判定一个取值较多的属性是否最优的；(2) 对一个取值较多的非最优属性，如何修正其信息增益，以避免被选取为分裂属性。

文献[5]提出引入用户兴趣度的优化法对 ID3 算法进行改进。即对(4)式修改为：

$$EA_1(A,S)=\sum_{j=1}^{k}\left(\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s}+k\right)Info(s_j) \quad (7)$$

相应地式(4)改为：

$$Gain_1(A,S)=Info(S)-EA_1(A,S) \quad (8)$$

该算法虽然解决了多值偏向问题，但存在以下几个主要缺点：(1)算法中提出由决策者根据先验知识或领域知识来判断一个取值较多的属性是否重要，具有一定的主观性；(2)式(7)中 k 的取值是一个模糊的概念，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需要反复测试训练集。由于需要人为确定参数 k，同样具有一定的主观性，可能由于人为因素而影响决策效果；(3)训练集要反复测试确定k，算法执行效率低；(4)没有充分考虑条件属性与分类属性、条件属性间的相互依赖关系。

针对上述问题，考虑到系统中特征属性序列曲线与类别属性序列曲线的紧密程度可用灰色关联度的大小来描述，即灰色关联度最大的特征属性对系统类别属性的影响也最大。因此，取值较多但灰色关联度较低的特征属性对分类结果影响不大，显然不是最优属性。为此，本文引入灰色关联度取代用户兴趣度等对 ID3 算法进行改进。具体方法是首先利用灰色关联分析建立各特征属性与类别属性之间的关系，计算各特征属性与类别属性之间的灰色关联度，并将它们排序；其次对取值较多的属性通过灰色关联度来判断是非最优，从而确定是否通过其灰色关联度来降低它的信息增益，最后对灰色关联度较低且取值较多的属性，在计算其信息增益时利用式(7)时将 k 设其灰色关联度，其它属性计算信息增益时设 k 为 0。

4 实验分析

为便于比较，本文采用文献[5]中提供的一组实验数据来进

行实验和对比分析。

天气舒适度的特征属性表{穿衣指数,温度,湿度,风力}[5]。各属性的值域为:D(穿衣指数)={正常,很多,较多},量化为{0,1,2};D(温度)={适中,很高},量化为{0,1};D(湿度)={正常,很大},量化为{0,1};D(风力)={没有,中等,很大},量化为{0,1,2}。分类结果 D(天气舒适度)={不舒适,舒适},量化为{0,1}。根据训练集样本数据,依次计算各穿衣指数,温度,湿度,风力等特征属性的灰色关联度,结果依次是0.48,0.8,0.68,0.68。由于多值属性穿衣指数的灰色关联度最低,与类别属性的关联最弱,是非最优属性。在计算穿衣指数信息增益时设定 k=0.48,计算其它属性信息增益时设定 k=0(结果见表 1)。表 1 显示,用 ID3 算法确定决策树根结点分裂属性时,穿衣指数的信息增益最大,而用改进算法确定决策树根结点分裂属性时,湿度的信息增益最大,避免了多值非最优属性穿衣指数成为分裂属性。

表 1　改进算法对天气舒适度各属性信息增益的影响

特征属性	ID3 算法	文献[5]中算法		本文中算法	
	gain 值	k	gain 值	k	gain 值
穿衣指数	0.483 5	0.33	0.012 8	0.48	−0.214 7
温度	0.026 7	0.00	0.026 7	0.00	0.026 7
湿度	0.060 0	0.00	0.060 0	0.00	0.060 0
风力	0.005 2	0.00	0.005 2	0.00	0.005 2

通过本文提出的决策树改进算法所生成的决策树结构与文献[5]中一样。但在建树过程中,灰色关联度是根据已知知识(训练集)客观计算的,避免了用户参与,提高了算法效率,保持了决策树算法的一个最大优点即不需要用户具有任何背景知识,是一种高效而且可靠的改进算法。

5　小结

本文提出的基于灰色关联度的决策树改进算法,在解决领域问题中大数据量覆盖小数据量的重要性方面具有一定的优势,能够根据训练集客观的判定多值属性是否最优,并通过其灰色关联度来修正属性的信息增益,解决了 ID3 算法多值偏向问题。该算法具有构思新颖独特,解决了文献[5]算法的缺陷,具有较为重要的理论意义和实用价值。(收稿日期:2007 年 5 月)

参考文献:

[1] 刘奕群,张敏,马少平.基于改进决策树算法的网络关键资源页面判定[J].软件学报,2005,16(11):1958-1966.

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

[3] 韩松来,张辉,周华平.基于关联度函数的决策树分类算法[J].计算机应用,2005,25(11):2655-2657.

[4] 何劲松,王煦法.参数估计决策树算法[J].模式识别与人工智能,2002,15(3):330-333.

[5] 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,2003,39(25):104-107.

[6] 王静红,王熙照,邵艳华,等.决策树算法的研究及优化[J].微机发展,2004,14(9):30-32.

[7] Deng Julong.Introduction to grey system theory[J].Journal of Grey System,1989(1):1-24.

[8] 马苗,樊养余,谢松云,等.基于灰色系统理论的图象边缘检测新算法[J].中国图象图形学报,2003,8A(10):1136-1139.

[9] Meng Xianlin,Shen Jin,Sun Lixin.Application of grey weighted related degree to the ambient air quality assessment[J].Journal of Harbin Institute of Technolog,2006,13(4):395-397.

(上接 145 页)

由 AODV 路由发现的机制并参照图 2 的拓扑结构可知,仿真开始时,AODV 首先进入的是路由发现阶段。当 n_0 节点发现不存在到达 n_3 节点的路由时,就发送广播请求到下一个节点,如果不是目的节点就继续向下一个节点发请求,如此反复直到返回确认包为止,从而建立一条 n_0 到 n_3 的路由。然后才开始转发数据包。而 GPSR 直接根据邻居节点的位置信息转发数据包,省去 AODV 复杂的路由发现过程。从 AODV 仿真结果图上看,在仿真的前 82 s 内包序列号、吞吐量、传输时延参数表明网络性能有些异常,而 82 s 之后就变得比较正常。这正印证了理论中的 AODV 路由过程即前 82 s 仿真了 AODV 的路由发现过程,之后进入正常的数据转发阶段。从 GPSR 仿真图观察,发现在整个 100 s 的仿真中路由协议很快进入了数据转发阶段且从性能参数看,网络性能良好,这也印证了 GPSR 路由过程。只是在全网电量方面,GPSR 耗能比 AODV 稍多一些,这对于那些对能源优先的 MANET(比如无线传感器网络)来说非常重要。路由协议理论和路由协议仿真结果一致,这表明 J-Sim 很好的完成了 MANET 路由协议仿真的任务,可以作为 MANET 研究的可信工具。

4　结束语

本文引入一种组件化、易扩展的网络仿真器 J-Sim。之后在其无线模块基础上仿真实现了 MANET 中的两种路由协议,实验结果表明在该仿真平台下可以很好的仿真 MANET 的运行。这为以后对 MANET 的路由协议研究提供了新的仿真工具。在以后的工作中,将主要针对现有的路由协议进行改进,提出新的路由算法,并通过 J-Sim 仿真验证设计算法的优劣。(收稿日期:2007 年 4 月)

参考文献:

[1] 英春,史美林.自组网体系结构研究[J].通讯学报,1999,20(9):48-49.

[2] Hung-ying Tyan,Yuan Gao,Jennifer Hou,et al.Tutorial:working with J-Sim [EB/OL].[2003-12-30].http://www.j-sim.org/tutorial/jsim_tutorial.html.

[3] Tyan Hung-ying,Design realization and evaluation of a component-based compositional software architecture for network simulation[D].USA:The Ohio State University,2002.

[4] Sobeih A,Chen W-P,Hou J C,et al.J-Sim:a simulation environment for wireless sensor networks[C]//Proc of the 38th IEEE Annual Simulation Symposium(ANSS'05),2005.

[5] Sobeih A,Hou J C,Tyan H Y.J-Sim:towards composable and extensible network simulation[C]//Proc of the 19th IEEE International Parallel and Distributed Proccessing Symposium(IPDPS'05),2005.

[6] Royer E M,Toh Chai-Keong.A review of current routing protocols for MANET mobile wireless networks[C]//Proc IEEE Personal Communications,April 1999.

[7] Karp B,Kung H T.GPSR:Greedy Perimeter Stateless Routing for wireless networks[C]//Proc of the sixth Annual ACM/IEEE International Conference on Mobile Computing and Networking(MobiComm'00),2000:243-254.

[8] Chen Wei-peng,Ge Ye,Hu Chunyu,et al.J-Sim wireless extension tutorial[EB/OL].http://www.j-sim.org/v1.3/wireless/wireless_tutorial.htm.