C4.5 算法在冠状造影数据处理中的应用

云玉屏,林克正
YUN Yu-ping,LIN Ke-zheng

哈尔滨理工大学 计算机科学与技术学院,哈尔滨 150080
College of Computer Science and Technology,Harbin University of Science and Technology,Harbin 150080,China
E-mail:yunyuping@gmail.com

YUN Yu-ping,LIN Ke-zheng.Application of C4.5 algorithm in data processing of CAG.Computer Engineering and Applications,2008,44(10):244-245.

Abstract: Firstly,we process the original data through data cleaning,data transform and data protocol.And generate the decision tree by using algorithm of C4.5 under weka platform.Then adjust the ratio factor variables to raise the veracity of judgment of the mothod.The results of experiment demonstrate the algorithm of C4.5 can reduce the risk of coronary angiography,and put forward a new mothod for coronary heart disease analysis and diagnoses.

Key words: coronary angiography;data mining;algorithm of C4.5

摘 要: 首先采用数据清理、数据变换、数据规约等预处理技术处理原始数据,并借助 weka 平台,通过 C4.5 算法生成决策树;然后针对决策树对正确率判断不够理想的地方,调整比例因子变量再进行测试提高判断的正确率。由两种方法的比较,以及与医学认识相对照,可以得出,文中所得决策分类树的构成特点同目前已知的高危因素趋于一致。通过 C4.5 算法建立判定决策树,降低了冠状动脉造影(CAG)的危险,为冠心病的分析预测提出了一种新的方法。

关键词: 冠状动脉造影;数据挖掘;C4.5 算法
文章编号: 1002-8331(2008)10-0244-02 **文献标识码:** A **中图分类号:** TP39

1 引言

自从 1961 年在 FRAMINGHAM 随访 6 年的报告中首次提出“危险因素”这个概念以来,有关 CAD 的众多危险因素被不断地发现和证实,并成为防治 CAD 的关键干预点。2004~2005 年组 CAD 患者年龄构成中,小于 60 岁所占的比例明显增加,大于 70 岁所占比例明显减少,年龄构成上呈低龄化趋势。对高血压、糖尿病、脂代谢紊乱、肥胖等危险因素予以早发现、早控制、早治疗,从而降低 CAD 的发病率和死亡率,这是摆在我国医疗卫生工作者面前的严峻课题[1]。

冠状动脉造影是诊断冠心病的一种有效方法,但由于属创伤性检查,如处理不当会引起严重并发症,甚至死亡。针对这一问题,试图用数据挖掘的方法来研究这个问题。可以将 C4.5 算法引入冠心病发病规律研究中,它能从巨大的冠心病数据中挖掘出需要的数据和规则。由于数据挖掘可以从大规模数据中自动进行规则的提取,将数据挖掘应用于冠心病发病规律的研究,一方面可以对大批量的冠心病数据进行处理,找出其发病规则,另一方面也可以仅从实际的冠心病数据中获取实用规则。

2 C4.5 算法

C4.5 算法[2,3]是构造决策树分类器的一种算法。这种算法利用比较各个描述性属性的信息增益值(Information Gain)的大小,来选择 Gain 值最大的属性进行分类。如果存在连续型的描述性属性,首先要把这些连续型属性的值分成不同的区间,即“离散化”。把连续型属性值“离散化”的方法是:

(1)寻找该连续型属性的最小值,并把它赋值给 MIN,寻找该连续型属性的最大值,并把它赋值给 MAX;

(2)设置区间$[MIN,MAX]$中的 N 个等分断点 A_i,它们分别是 $A_i=MIN+\frac{MAX-MIN}{N}\times i$,其中,$i=1,2,\cdots,N$;

(3)分别计算把$[MIN,A_i]$和$[A_i,MAX]$($i=1,2,\cdots,N$)作为区间值时的 Gain 值,并进行比较;

(4)选取 Gain 值最大的 A_k 作为该连续型属性的断点,把属性值设置为$[MIN,A_k]$和(A_k,MAX)两个区间值。

C4.5 算法使用信息增益的概念来构造决策树,其中每个分类的决定都与所选择的目标分类有关,不确定性的最佳评估方法是平均信息量,即信息熵(Entropy):

$$S=-\sum_I (P_i^*\log(P_i)) \qquad (1)$$

本文所述信息增益是指信息熵的有效减少量,根据它就能够确定在什么样的层次上选择什么样的变量来分类。假设存在两个类 P 和 N,并且记录集 S 中包括 x 个属于类 P 的记录和 y 个属于类 N 的记录。那么,用于确定记录集 S 中某个记录属于

作者简介: 云玉屏(1982-),女,硕士研究生,主要研究方向:数据挖掘算法在医疗领域的应用;林克正(1962-),男,博士研究生,教授,主要研究方向:图像处理与机器视觉、计算机保密与编码理论、数字水印技术等。

收稿日期: 2007-06-28 **修回日期:** 2007-09-17

哪个类的所有信息量为:

$$Info(S)=Info(S_p,S_n)=-(\frac{x}{x+y}\cdot\log\frac{x}{x+y}+\frac{y}{x+y}\cdot\log\frac{y}{x+y}) \quad (2)$$

假设使用变量 D 作为决策树的根节点,把记录集 S 分为子类$\{S_1,S_2,\cdots,S_k\}$,其中每个 $S_i(i=1,2,\cdots,k)$中包括 x_i 个属于类 P 的记录和 y_i 个属于类 N 的记录。那么,用于在所有的子类中分类的信息量为:

$$Info(D,S)=\sum_{i=1}^{k}\frac{x_i+y_i}{x+y}\cdot Info(S_{ip},S_{in}) \quad (3)$$

假设选择变量 D 作为分类节点,那么它的信息增量值一定大于其它变量的信息增量值,变量 D 的信息增量为:

$$Gain(D)=Info(S)-Info(A,S) \quad (4)$$

由此可以给出信息增益函数的通用定义:

$$Gain(D,S)=Info(S)-Info(D,S) \quad (5)$$

$$Info(S)=I(P)=I(P_1,P_2,\cdots,P_k)=I\left(\frac{|C_1|}{|S|},\frac{|C_2|}{|S|},\cdots,\frac{|C_k|}{|S|}\right)=-(p_1\cdot\log p_1+p_2\cdot\log p_2+\cdots+p_k\cdot\log p_k) \quad (6)$$

$$Info(D,S)=\sum_{i=1}^{n}(\frac{|S_i|}{|S|})\cdot Info(S_i) \quad (7)$$

3 实验结果与分析

存入数据库中的原始数据首先要检查是否正确地录入,此外表中还有可能存在噪声、空缺和不一致性数据,这些数据的存在将对数据的再处理产生比较大的影响,而且一些冗余数据的存在也会使得影响结果的数据项太多,因此,必须对各表项的数据定义其数据类型,然后转换成计算机能够处理的数据形式再存入数据库,从而形成直接可利用的有效数据库资源。

第一阶段,针对数据的不完整性、噪声和不一致性采用数据清理、数据变换、数据规约等预处理技术处理。初始特征集如表 1 所示。

表 1 病例特征

名称	均值	标准差
性别		
吸烟史/年	2.53	8.53
年龄/年	56.72	8.74
身高/cm	159.74	4.09
体重/kg	61.21	6.26
收缩压/mmHg	144.28	19.58
舒张压/mmHg	89.07	11.93
甘油三脂/(mmol/L)	2.36	1.72
腰围/cm	84.31	11.30
极低密度脂蛋白/(mmol/L)	0.46	0.29
总胆固醇/(mmol/L)	5.02	1.01
空腹血糖/(mmol/L)	5.18	0.87

第二阶段,用 C4.5 算法从中随机抽取 2/3 的数据作为C4.5 算法的训练数据,其余数据作为测试数据。通过对 C4.5 算法的训练,得出如下决策分类树[4,5],如图 1 所示(Y 表示患病类,N 表示未患病类)。利用训练集和测试集的方法进行分类准确率测试,测试结果是对于患病的正确识别率为 95.41%。

第三阶段,通过改进测试属性选择方法进行决策树的简化。首先合并那些对分类贡献小的分枝,并在修正后的测试属性集上选出合适的测试属性。改进后的算法可以有效地降低树的复杂度,提高树的健壮性,如图 2 所示。测试结果是对于患病的正确识别率为 97.18%。

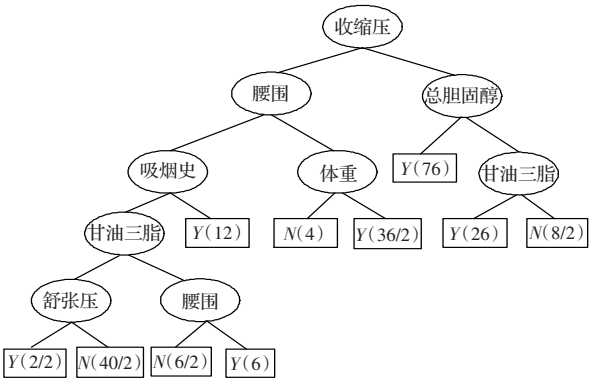

图 1 C4.5 算法得到的决策树

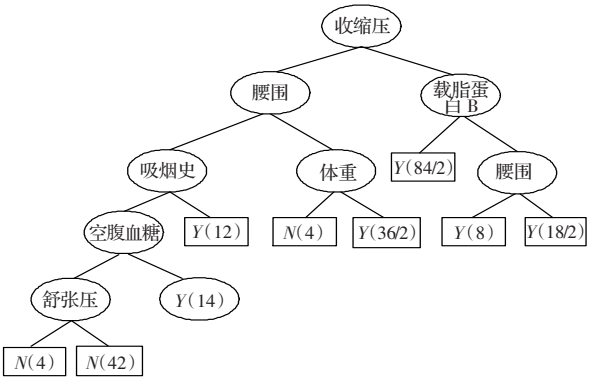

图 2 改进后得到的决策树

出于速度上的考虑,原始数据经过分类器决策的次数越少,决策时间就越快,效率就越高[6]。下面先定义几个概念:

规则 i 在测试集 S 中的错误函数:$E_i(S)$。

集合 S 中类 X 的个数函数:$C(X,S)$。

集合 S 的比例函数 K,即数据集合中分类 N 与分类 Y 的比例:

$$K(S)=\frac{C(N,S)}{C(Y,S)} \quad (8)$$

分类 Y 在测试集合 S 上的正确率:

$$R(Y,S)=\frac{\sum_{i=1}^{n}E_i(S)}{C(Y,S)} \quad (9)$$

其中,n 为分类规则的总个数。

构造正确率随 k 值变化的曲线如图 3。

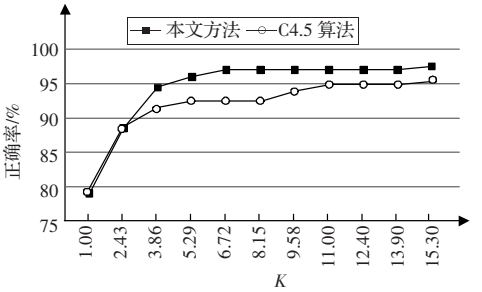

图 3 两种方法的比较图

通过与医学认识相对照,可以得出,本文所得决策分类树的构成特点同目前已知的高危因素趋于一致。这表明 C4.5 算法建立的决策树具有实际物理意义。

(下转 248 页)