维普资讯 http://www.cqvip.com

一种新的决策树分裂属性选择方法

刘星毅
(钦州学院,广西 钦州 535000)

摘 要:分类问题是数据挖掘和机器学习中的一个核心问题。为了得到最大程度的分类准确率,决策树分类过程中,非常关键的是结点分裂属性的选择。常见的分裂结点属性选择方法可以分为信息熵方法、GINI系数方法等。分析了目前常见的选择分裂属性方法——基于信息熵方法的优、缺点,提出了基于卡方检验的决策树分裂属性的选择方法,用真实例子和设置模拟实验说明了文中算法的优越性。实验结果显示文中算法在分类错误率方面好于以信息熵为基础的方法。

关键词:决策树;分裂属性;卡方检验;信息熵
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)05-0070-03

A New Splitting Criterion of Decision Trees

LIU Xing-yi
(Qinzhou University, Qinzhou 535000, China)

Abstract: Classification is an important issue on data mining and machine learning. Selecting splitting attributes is the key process during constructing decision tree for receiving the maxinized classification accuracy. Existing methods for classification usually can be the method based on entropy, GINI index, and so on. Analyses the disadvantages and the advantages of the method which is utilized to select splitting attributes based on information gain theory, and proposes a statistical method which employs chi-squared test to get the relation between the condition attributes and the class label. Demonstrate experimental this algorithm and the results show this method is significantly well than the methods based on information theory.

Key words: decision trees; splitting attributes; Chi-squared test; information entropy

1 决策树介绍

在数据挖掘和机器学习中,决策树分类算法以其抽取规则简便、规则易于理解等优点得到了广泛的应用[1]。1986年Quinlan[2]首先提出ID3算法,并在1993年提出了C4.5和C5算法,它们马上就成为了广泛流行的决策树算法。决策树算法一般采用自顶向下的贪婪算法,在每个内结点选择分类效果最好的属性进行下一步的分类,直到这棵树能准确地分类训练样本,或所有的属性都已被使用过。影响决策树分类算法分类效果的主要问题是,在每个内结点如何选取要测试的属性以及剪枝技术。文中主要研究在内结点如何选取要分裂属性的问题。

传统的属性选择标准中应用最为广泛的有信息增益(Information Gain)[2]、信息增益率(Gain Ratio)[3]等。文中讨论基于熵理论的主要分类标准的特性,并针对这类标准的缺点提出了一种基于卡方检验的选择分裂属性的方法,依此建立的决策树,避免了传统决策树区分能力不强的缺点,并且能有效地降低决策树分类错误率。

2 分裂属性的选择

决策树的核心问题是如何选取在树的每个结点要测试的属性,称为测试属性或者分裂属性。在建立决策树时,减少测试后产生的新子节点的凌乱度(Disorder)是选择分裂属性的基本精神。换个方式说,就是希望能够使节点测试的动作尽量少,尽快使每个树叶节点内的每个例子种类都相同,这样建立起来的决策树的深度会比较浅,相同地,决策树也会变得比较小。选择分类属性的方法可以分为两大类:一是直觉上的方法。就是要找到个属性,使测试后的每个例子的子集合之间的差异最大。就是说,想办法使测试的例子,尽量归属于已经不用继续再测试的子集合。但是,一旦训练集合内的例子变多,就有可能发生无论使用哪个属性测试,都无法产生任何一个不需再测试的子集

收稿日期:2007-08-25
基金项目:广西自然科学基金(桂科0640069)
作者简介:刘星毅(1972-),男,广西钦州人,硕士,中国计算机学会会员,研究方向为计算机网络、数据库技术。

维普资讯 http://www.cqvip.com

合的情况。所以直觉上的方法只适用于训练集合很小的时候。第二种方法,即专家们提出的用信息论方法的方法代替直觉方法。"信息论(Information Theorem)"是由Shannon于1949年提出来的,最早用来处理一些与通讯上有关的问题。之后,Quinlan于1979年提出ID3决策树归纳算法,首次使用信息论来当作选择测试属性时的依据,造成了革命性的突破。在ID3中,使用信息增益(Information Gain)方法来帮助确定生成每个节点时所应采用的合适属性。这样就可以选择具有最高信息增益(熵减少的程度最大)的属性作为当前结点的测试属性。该属性使得对结果划分中的样本分类所需的信息量最小,利用该属性进行当前〔节点所含)样本集的划分,使得所产生的各样本子集中的"不同类别混合程度"降为最低。因而采用这样一种信息论方法将帮助有效减少对象分类所需要的次数,从而确保所产生的决策树最为简单,尽管不一定是最简单的,因为这种总体最优的问题通常是NP问题。Quinlan的信息增益定义为:

$$\mathrm{Gain}(A) = I(s_1, s_2, \cdots, s_m) - E(A) \tag{1}$$

其中一个给定的样本数据对象进行分类所需的信息量为:

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2^{(p_i)} \tag{2}$$

这里,$p_{ij} = \frac{s_{ij}}{|S_j|}$,是$S_j$中的样本属于$C_i$的概率。

根据测试属性A划分子集的熵(entropy)或期望信息为:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{S} I(s_{1j}, \cdots, s_{mj}) \tag{3}$$

项$\frac{s_{1j} + \cdots + s_{mj}}{S}$充当第$j$个子集的权,并且等于子集(即,$A$值为$a_j$)中的样本个数除以$S$中的样本总数。熵值越小,子集划分的纯度越高。

文献[1]认为信息增益度量存在一个内在偏置,它偏袒具有较多值的属性,他们提出选择替换的度量标准是增益比率(gain ratio)[2]。但是,此方法的分裂信息项阻碍选择值为均匀分布的属性。即使用增益比率代替增益来选择分裂属性会产生的一个问题:当某个S_i接近$S(|S_i| \approx |S|)$时,分母可能为0或非常小。并且如果某个属性对于S的所有样例有几乎同样的值,这时要么导致增益比率未定义,要么是增益比率非常大。为了避免上述信息熵方法的缺陷,文中利用统计理论中的卡方检验来选择分裂属性。

3　卡方检验方法

卡方检验[4]是对样本的频数分布所来自的总体分布是否服从某种理论分布或某种假设分布所作的假设检验。它属于自由分布的非参数检验。它可以处理一个因素分为多种类别,或多种因素各有多种类别的资料。所以,凡是可以应用比率进行检验的资料,都可以用卡方检验。并且文中的方法可以同时处理离散值的属性和连续值属性,而经典的信息理论方法只能处理离散值属性,遇上连续属性的时候,这些方法通常用离散化方法把连续值转化成离散值,这种转化过程通常造成信息丢失。

文中之所以提出用卡方检验的方法代替信息熵的方法有以下几个原因:

第一,卡方是一种非参数检验,在现实数据中,用户通常对所要处理的数据集没有任何先验知识,这样不可能采用统计上通常采用的参数方法,而此时,非参数方法是最好的替换,它能避免参数方法因为错误的参数估计而导致的不正确估计,能有效地捕捉数据间的关系;

第二,使用卡方统计有着成熟和完备的理论保障;

第三,实验证明,卡方检验方法能取得比著名的信息熵更好的分类结果。

设计的卡方检验选择分裂属性方法可以归纳成以下几个步骤:

(1)把各个属性之间的数据进行规范化。由于在处理多个属性的过程中,容易出现计算结果偏向于数量级较大的属性,从而就会出现决策树构建过程中常见的偏置(bias)问题,所以在进行计算的过程中通常把各种属性值规范化到1～10之间,这样可以有效地避免bias问题。

(2)根据式(4)计算测试集中每个属性的卡方值。

$$\chi^2 = \sum_i \sum_j \frac{(x_{ij} - E_{ij})^2}{E_{ij}} \tag{4}$$

其中x_{ij}代表第i个属性的第j个属性值,E_{ij}为第i个属性的所有属性值的均值。

(3)根据得到的卡方值进行分析,卡方值越大说明该属性与分类属性的相关性越强烈,在选择分裂属性的时候,就选择卡方值最大的为当前分类结点的分类属性。

下面根据一个例子来解释文中的算法。表1是某个城市一个住房调查表,有5个条件属性分别为调查日期、被调查人的住房地区,被调查人的住房房型、被调查人的收入和前一个被调查人对调查结果的应答,决策属性是当前调查人的调查应答。

首先基于收入属性和决策属性两个属性建立它们的关系表(见表2),因为此例子的各个属性值在设想的范围内,不用规范化各个属性值。

维普资讯 http://www.cqvip.com

表 1 某城市住房调查表

日期	住宅区域	房型	收入	前结果	现结果
3/10/06	近郊	分离	高	无应答	无应答
14/9/06	近郊	分离	高	已应答	无应答
2/4/05	农村	分离	高	无应答	已应答
18/1/06	城市	半分离	高	无应答	已应答
3/4/06	城市	半分离	低	无应答	已应答
15/10/05	城市	半分离	低	已应答	无应答
15/10/05	农村	半分离	低	已应答	已应答
2/3/04	近郊	综合	高	无应答	无应答
4/5/06	近郊	半分离	低	无应答	已应答
2/1/06	城市	综合	低	无应答	已应答
3/10/06	近郊	综合	低	已应答	已应答
3/10/06	农村	综合	高	已应答	已应答
8/4/06	农村	分离	低	无应答	已应答
6/5/05	城市	综合	高	已应答	无应答

表 2 收入和决策属性关系表

	已应答	无应答	总计
高	3	4	7
低	6	1	7
总计	9	5	14

接着根据表 2 计算收入属性的卡方值：

$$\chi^2_{income}=\frac{(3-3.5)^2}{3.5}+\frac{(4-3.5)^2}{3.5}+\frac{(6-3.5)^2}{3.5}+\frac{(1-3.5)^2}{3.5}=3.7$$

同理可以得到其他属性的卡方值，从表 3 中可以知道，日期属性的卡方值最大，所以就选择日期属性为决策树当前结点的分裂属性。

表 3 3 种方法分裂属性选择结果

	信息熵	增益比率	卡方值
日期	0.600	0.180	9.333
住宅区域	0.246	0.156	4.400
房型	0.049	0.031	2
收入	0.151	0.151	3.7
前结果	0.048	0.048	2

为了好比较，在表 3 中列出了实验过程中其他两种方法(即信息熵方法和增益比率方法)的结果。从以上三种方法得到的结果来看，虽然“日期”属性在三种方法中都被选为当前结点的分裂属性，但是明显的卡方检验方法区别能力更强(各种属性的卡方值分别是 9.333、4.4、2、3.7 和 2)，而信息比率区别能力最差(各种属性的卡方值分别是 0.180、0.156、0.031、0.151 和 0.048)。

4 实验分析

为了显示文中方法的优越性，从 UCI[5]数据库下载了 8 个数据库，把这 8 个数据库用上述三种方法在软件 C5(可以在 Quinlan 的主页上免费下载到)进行分类比较，然后比较这三种方法的分类错误率，实验结果显示在表 4 中。

表 4 8 个 UCI 数据集的实验结果

	Ecoli (%)	Breast (%)	Heart (%)	Thyroid (%)	Australia (%)	Mushroom (%)	Voting (%)	Cars (%)
信息熵	26.33	27.63	21.35	19.63	26.54	24.58	20.25	18.52
增益比率	25.31	28.53	19.56	22.81	25.84	25.34	21.63	16.25
卡方检验	20.23	22.89	18.23	17.65	25.39	24.59	19.30	15.62

5 结 语

基于统计学理论，提出了利用卡方检验的方法在决策树中进行分裂属性的选取。通过实验证明，文中的方法是非常有效的，并且卡方检验方法简单实用，应用起来非常方便，一些计算器和一些常用的软件例如 EXCEL 上都有计算卡方值的功能。

参考文献：

[1] Mitchell T M. 机器学习[M]. 曾华军，张银奎译. 北京：机械工业出版社，2003.

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

[3] Quinlan J R. C4.5: program for machine learning[M]. New York, US: Morgan Kaufmann, 1993.

[4] Hunt E B, Marin J, Stone P T. Experiments in Induction [M]. New York, US: Academic Press, 1966.

[5] Blade C L, Merz C J. UCI repository of machine learning databases(website)[D]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

(上接第 69 页)

参考文献：

[1] Myron D, Ganeshram R. The Truth About CRM Success & Failure[J/OL]. CRM Magazine, 2002. http://www.destinationcrm.com/articles/default.asp? ArticleID=2370.

[2] Kellen, Vince. CRM Measurement Frameworks[EB/OL]. 2002. http://www.kellen.net/crmmeas.htm.

[3] Dipak J, Siddhartha S. Customer Lifetime Value Research In Marketing: A Review and Future Direction[J]. Journal of Interactive Marketing, 2002, 16(2):34－46.

[4] 沈兆阳. SQL Server 2000 OLAP 解决方案——数据仓库与 Analysis Services[M]. 北京：清华大学出版社，2001.

[5] 巴斯蒂安 M. 数据仓库与数据挖掘[M]. 北京：冶金工业出版社，2003:175－177.