山东农业大学学报(自然科学版),2016,47(6):936-939 VOL.47 NO.6 2016
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2016.06.025

决策树算法在分类预测中的应用与优化

葛 朋[1],彭梦晶[2]
1. 重庆市中医院信息科, 重庆 400010
2. 重庆市中医院设备处, 重庆 400010

摘 要：决策树是一类常见的机器学习方法，具有属性结构和较好的分类预测能力，可以根据既定规则完成基本的决策任务。本文阐述了决策树算法的基本思想，并以某银行信贷问题为例，分析了决策树算法在应用中遇到的一些问题，最后给出了性能调优方案。

关键词：机器学习; 决策树算法; 分类预测
中图法分类号：TP181 **文献标识码**：A **文章编号**：1000-2324(2016)06-0936-04

The Optimization and Application of the Decision Tree Algorithm in the Classification Prediction

GE Peng[1], PENG Meng-jing[2]
1. *Information Center of Chongqing Chinese Medicine Hospital, Chongqing* 400011, *China*
2. *Equipment Department of Chongqing Chinese Medicine Hospital, Chongqing* 400011, *China*

Abstract: Decision tree is a kind of common method in machine learning, it is with an attribute structure and better ability to classify and predict and it can complete basic decision tasks according to the given rules. The paper firstly expounded the basic thoughts and then analyzed some problems in the application of the decision tree algorithm taking a bank credit as case to obtain the performance tuning scheme in the end.

Keywords: Machine learning; decision tree algorithm; classification prediction

随着云计算和大数据的迅速发展，数据挖掘技术得到了广泛的应用。数据挖掘指的就是从大量数据中通过某种算法和工具，挖掘数据背后隐藏价值的过程。而在实际应用中，数据挖掘主要用于分类和预测。数据挖掘需要实现建立数据关系模型，对数据进行分析预测。预测是为了基于历史数据通过机器学习算法分析数据的变化趋势，达到预测的作用。决策树算法是数据挖掘中最常用的方法，其作用于分类阶段，可以直接体现数据特点，分析预测数据，并能方便提取决策规则，达到辅助决策的功效。

1 决策树基本流程

决策树是一种常见的树形结构，一个典型的决策树由一个根节点、若干个内部节点和若干个叶节点组成。叶节点与决策结果相对应。其它每个内部节点表示一个属性测试，每个分支代表一个测试输出，每个叶节点代表一种类别。在机器学习中，决策树学习的目的是针对不同的未知数据产生一颗泛化能力强的决策树。常用的决策树算法是 ID3 和 C4.5 算法，其采用“自上而下、分类治之”的方法，通过一些无序、无规则的事例推测出决策树的分类规则，可以实现对位置数据的分类、预测和数据预处理。决策树一般分为构成和剪枝两个步骤，其工作流程如图 1 所示：

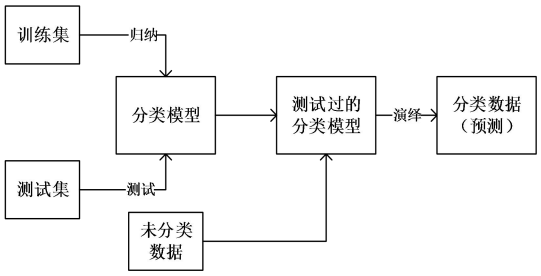

图 1 决策树工作流程图
Fig.1 The running process of the decision tree

收稿日期：2016-03-20 **修回日期**：2016-03-26
作者简介：葛 朋(1975-),男,研究生,工程师,研究领域:数据库技术. E-mail:34170856@qq.com

数字优先出版:2016-12-25 http://www.cnki.net

决策树学习的关键是如何选择最优划分属性。一般而言，随着划分过程不断进行，我们希望决策树的分支节点所包含的样本尽可能属于同一类别，即节点的纯度（Purity）越来越高。因此引入了信息增益的概念。

2　决策树算法分析

2.1　ID3 算法

ID3 算法是决策树的一种，它是基于奥卡姆剃刀原理的，这个算法的基础是越是小型的决策树越优于大的决策树，尽管如此，也不总是生成最小的树型结构，而是一个启发式算法。假定当前样本集合 D 中第 k 类样本所占的比例为 P_k (k=1,2,3...|y|)，那么 D 的信息熵定义为：

$$Ent\,(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$

在信息论中，$Ent(D)$的值越小，那么 D 的纯度就越高。根据上式可以计算属性 a 对样本集 D 进行划分所获得的信息增益（Information gain）。

$$Gain\,(D,a) = Ent\,(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent\,(D^v)$$

ID3 算法的实质是利用信息增益来度量属性的选择，在选择分裂属性时，往往偏向于选择取值较多的属性，但是现实应用中取值较多的属性不一定是最重要的属性，因此很有可能会使得生成的决策树的预测结果与现实情况误差较大，为了解决这个问题，引入分支信息熵和属性优先的概念。

2.2　C4.5

C4.5 决策树算法是在 ID3 算法的基础上发展起来的，通过信息熵方法递归形成决策树，应用广泛，效率更高。对比两种算法的不同之处，一方面表现在 C4.5 将信息增益作为测试属性，而 ID3 算法采用基于信息增益的方法选择测试属性。另一方面表现是 C4.5 算法不需要独立测试样本集，提高效率，可以直接处理连续属性和属性空缺的样本，这样的产生决策树分枝减少，而 ID3 算法的连续属性处理是离散化的。

C4.5 算法用信息增益率来选择属性，解决了用信息增益选择属性时偏向选择取值多的属性的缺陷问题，此外，它还能处理非离散的数据和不完整的数据。但是 C4.5 的算法效率比较低。

3　决策树算法的应用

为了验证并测试决策树算法在分类预测中的性能以及调优效果，以某银行客户信贷风险预测为例，详细阐述决策树在算法在分类预测中的应用。现在拥有一个某公司近 2 年内的申办信用卡客户数据集，包含 40700 个客户，每个客户有 40 个属性域。其中有 700 个客户申请了贷款项目（target_flag=1），而另外的 40000 个则是没有申请的（target_flag=0）。同时还提供了一个预测数据集（8000 个样本），其中的 TARGET_FLAG【是否贷款】属性是本次实验重点关注的属性，利用训练集建立模型，然后利用此模型对此属性进行预测，最终得出每个客户申请贷款项目的可能性。

在数据预处理阶段主要完成的是数据清洗工作，主要包括数据格式转换，空缺值处理，离散化和不相关属性的清理等工作，目的是提高数据的质量。为了提高模型的可信度，我们手动从原始数据集中随机选出 4000 个实例作为训练集，21343 个实例作为测试集，用来测试分类预测模型的性能。

在这个数据集中，显然|y|=2，在决策树学习开始时，根节点包含 D 中所有的样例，其中正例占 p_1=700/21343，p_2=700/21343,因此根节点的信息熵为：

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k = 0.998$$

然后我们计算出当前属性集合中每个属性的信息增益，最后选择信息增益最大的属性作为划分属性，选择部分决策树如图 2 所示。

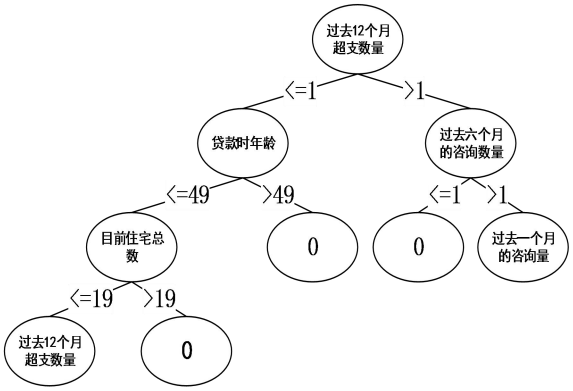

图 2 部分决策树模型

Fig.2 The model of some decision trees

在 Weka 平台中分别选择 ID3 算法和 C4.5 算法对模型进行训练，并将模型应用到集中，分别得到模型的性能结果，如表 1 所示。

表 1 ID3 和 C4.5 性能比较

Table 1 Comparison of ID3 and C4.5 performances

算法 Algorithms	正确率 Correct percentage	均方根误差 Root mean square error
ID3	89.63%	0.1518
C4.5	92.45%	0.1439

4　决策树分类算法的优化

通过分析决策树的两种算法 ID3 算法与 C4.5 算法的不同，总结出来决策树算法的三大问题，即计算效率低，多值偏向和对空缺值敏感。计算信息熵是一个高度复杂的过程，导致其计算开销比较大。除此之外原始数据集由于缺乏属性约减导致额外的计算开销。基于信息熵的属性选择都会偏向于选择取值较多的属性，但是有时候下取值较多的属性不一定就是最重要的属性。决策树算法在遇到缺失的数据时产生错误的概率比较大，这样会对后续的预测和决策产生较大的影响。

针对以上三大问题，对数据预处理、属性的合理选择和连续属性的离散化等方面进行优化。

4.1　预处理

预处理主要是解决决策树空缺值敏感问题。现实任务中经常遇到不完整的样本，即样本缺失，例如由于诊测成本隐私保护等因素，患者的医疗数据在某些属性上的取值未知，尤其是属性较多的数据集上，往往会有大量样本出现缺失值。假如给定数据集 D 和属性 a，$\tilde{D}$表示 D 中在属性 a 上的缺失值的样本集，那么信息增益公式可以拓展为

$$Gain(D,a)=\rho\times Gain(\tilde{D},a)=\rho\times(Ent(\tilde{D})-\sum_{v=1}^{V}\tilde{r}_v Ent(\tilde{D}^v))$$

还可以用一个全局常量，属性的均值来填补缺失值或根据不同元祖分类数据，选择同一个属性的平均值替换空缺值。

4.2　属性选择

虽然从理论角度来讲，选择的属性越多，信息量越大，然而当数据量达到某种水平时，性能便会急剧下滑。由于在实际的应用场景中，样本不可能是无限的，因此，在设计分类器时进行属性消减是很重要的。

削减属性可以通过属性提取和属性选择。属性提取其实就是一种映射。若 X 是原始的测量空间，X'是属性空间，则 $X\rightarrow X'$的映射就叫作属性提取器。属性选择是通过选择有效属性来达到空间降维目的的过程。属性选择是属性提取的一种特殊情况。提取变量往往会降低结果的可解释性。尤其对于离散变量而言，进行属性提取是没有意义的。所以主要研究属性选择方法。删除与目标属性无关的属性例如 ID 编号等可以提高模型的应用效率。

4.3 连续属性离散化

在本次数据集中，客户的收入是连续属性，但是在建模过程中需要将其离散化，便于分类处理。离散化是在面对分类问题时处理连续数据常用的方法，像 ID3 算法只能够处理离散属性，而 C4.5 不仅能够处理连续数据，同时还能有效处理离散数据。离散化方法可以分为两类：全局离散和局部离散。全局离散需要考虑到属性之间的相互作用，局部离散方法限制一次只能对一个属性进行离散。局部离散相对于全部离散要简单。

全部离散要比局部离散的计算开销更大。

经测试，优化后的算法在准确率和最小均方误差等参数均有所提升，正确率提高至 96.71%，均方根误差为 0.1279。

5 总 结

随着大数据和云计算时代的到来，数据挖掘分类问题和算法研究成为热点，本文对数据挖掘中决策树算法的应用进行了优化研究。通过分析决策树常见的两种算法以及决策树算法中存在的问题，针对实际问题，主要提出数据预处理、从属性选择、连续属性离散化等几方面改进决策树分类算法，提高决策树算法的效率和性能。

参考文献

[1] 王 珊,萨师煊.数据库系统概论[M].4 版.北京:高等教育出版社,2006
[2] 张云涛,龚 玲.数据挖掘原理与技术[M].北京:电子工业出版社,2004
[3] 赵 基.基于数据挖掘的银行客户分析管理关键技术研究[D].杭州:浙江大学,2005
[4] 石振华.银行卡用户数据挖掘系统的设计与实现[D].西安:西安电子科技大学,2010
[5] 郑英姿.基于数据挖掘的商业银行客户关系管理研究[D].西安:西安科技大学,2010
[6] 李明江,唐 颖,周力军.数据挖掘技术及应用[J].中国新通信,2012(22):66-67
[7] 薛 薇.数据挖掘中的决策树技术及其应用[J].统计与信息论坛,2002,17(2):4-10
[8] 杨学兵,张 俊.决策树算法及其核心技术[J].计算机技术与发展,2007,17(1):43-45
[9] 栾丽华,吉根林.决策树分类技术研究[J].计算机工程,2004,30(9):94-96,105
[10] 陈 文,史金成.决策树分类技术研究[J].福建电脑,2005(8):5-6
[11] 郭彦伟.电信行业客户流失分析的决策树技术[J].科技和产业,2005,5(11):7-9
[12] 李晓卉.决策树技术在客户信用分析中的应用[J].武汉科技大学学报:社会科学版,2008,10(2):26-28,32
[13] 陈沛玲.决策树分类算法优化研究[D].长沙:中南大学,2007
[14] 何 清,李 宁,罗文娟,等.大数据下的机器学习算法综述[J].模式识别与人工智能,2014,27(4):327-336
[15] 李 运.机器学习算法在数据挖掘中的应用[D].北京:北京邮电大学,2015
[16] 姜百宁.机器学习中的特征选择算法研究[D].青岛:中国海洋大学,2009
[17] 孙 亮.若干机器学习算法的研究与应用[D].长春:吉林大学,2012
[18] 胡 蓉.增量机器学习算法研究[D].南京:南京理工大学,2013
[19] 王淑珍.机器学习算法的 Weka 嵌入[D].广州:华南理工大学,2013
[20] 王 静.基于机器学习的文本分类算法研究与应用[D].成都:电子科技大学,2015