决策树 ID3 算法与 C4.5 算法研究

席经纬

上海交通大学　电子信息与电气工程学院

摘　要　用于分类的数据挖掘技术方法有很多, 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 算法和 C4.5 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 算法和 C4.5 算法的研究现状, 在详细介绍 ID3 和 C4.5 算法原理、分类依据的基础上, 比较两者的优劣。并且通过实验仿真来具体探究算法构建决策树的过程及结果。

关键词　决策树，ID3 算法，C4.5 算法

数据挖掘是从大量数据中发现有用知识的过程，是一种专业技术，也是一种分析数据的手段，用于发现海量数据所隐藏的各种规律。数据挖掘技术在对数据进行处理的过程中需要对数据进行分类，这样才能方便之后数据的处理与预测。通常所用到的数据分类技术是决策树分类法。决策树，顾名思义，它是一种树形结构，包含决策节点、分支和叶节点三部分，其中决策节点代表某个待分类数据集合的某个属性，在该属性上的不同测试结果对应一个分支，每个叶节点表示一种可能的分类结果。

传统的决策树分类方法有 ID3 和 C4.5，他们都是以信息熵作为分类依据，是单颗决策树。

1　ID3 算法

1.1　算法原理

在 ID3 算法中，选取信息增益最大的属性作为分类依据，然后根据该属性的属性值构造分支。具体构造步骤如下:

step1. 构造一个节点，如果所有的样本都在该节点上，那么停止算法，将该节点改成叶子节点，并用该类标记。

step2. 如果样本都不在一个节点上，求出每一个属性的信息熵，根据公式求出每一个属性的信息增益，然后选择信息增益最大的一个属性作为分类依据。

step3. 对属性中的每一个值创建一个分支，然后划分样本。

step4. 重复上述 step1-step3 步骤，直到属性全部使用完。[2]

1.2　数据划分

ID3 算法以信息熵和信息增益作为衡量标准来对样本进行分类。

1.2.1 信息熵

熵的概念主要是指信息的混乱程度，变量的不确定性越大，熵的值也就越大，熵的公式可以表示为：

$$Entropy(S) = -\sum_{i=1}^{m} p(u_i) log_2 p(u_i)$$

其中 m 为样本类别数，$p(u_i)$ 为类别在样本中出现的概率

$$p(u_i) = \frac{|u_i|}{|S|}$$

1.2.2 信息增益

信息增益指的是划分前后熵的变化，可以用下面的公式表示：

$$InfoGain(S, A) = Entropy(S) - \sum_{V \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

其中，A 表示样本的属性，Value(A) 是属性所有的取值集合。V 是 A 的其中一个属性值，S_v 是 S 中 A 的值为 V 的样例集合。[4]

2 C4.5 算法

2.1 算法原理

C4.5 算法也是用于生成决策树的一种经典算法，是 ID3 算法的另一种延伸和优化。C4.5 算法与 ID3 算法生成决策树的过程基本相同，但是 ID3 算法不能处理连续型属性，而 C4.5 算法可以先离散化连续型属性，然后进行属性的选择分类；属性分类时，ID3 算法利用信息增益进行属性分类选择，C4.5 算法则用信息增益率进行计算。

相比于 ID3 算法中，C4.5 算法选取信息增益率最大的属性作为分类依据，然后根据该属性的属性值构造分支。具体构造步骤如下：

step1. 对数据预处理形成决策树训练集，如果数据是连续型属性变量，首先就应该考虑对其进行离散化，若本身是离散值，此步省略。

step2. 构造一个节点，如果所有的样本都在该节点上，那么停止算法，将该节点改成叶子节点，并用该类标记。

step3. 如果样本都不在一个节点上，求出每一个属性的信息增益，根据公式求出每一个属性的信息增益率，然后选择信息增益率最大的一个属性作为分类依据。

step4. 对属性中的每一个值创建一个分支，然后划分样本。

step5. 重复上述 step2-step5 步骤，直到属性全部使用完。

2.2 数据划分

对于离散特征, C4.5 算法不直接使用信息增益, 而是使用“信息增益率”(gain ratio)[7]来选择最优的分支标准，增益率的定义如下:

$$GainRatio(S,A) = \frac{InfoGain(S,A)}{IV(A)}$$

其中 IV(A) 称作分支标准 A 的“固有值”(intrinstic value）, 属性 A 的可能性数目越多，则 IV(A) 的值通常会越大。

$$IV(A) = -\sum_{V \in Value(A)} \frac{|S_v|}{|S|} * log_2(\frac{|S_v|}{|S|})$$

3 比较分析

通过 ID3 算法的原理及公式描述, 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: a. 自顶向下的搜索方式降低了搜索次数, 提升了分类速度。b. ID3 算法原理清晰, 算法思路简单易懂, 易于实现。c. 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。[1]ID3 算法不足主要有以下四点:a. ID3 算法对噪声数据相对敏感, 算法依赖于属性值数目较多的属性，但是属性值较多的属性不一定是分类最优的属性。b. ID3 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。c. ID3 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。d. ID3 只能分类离散型数据。[6]

C4.5 算法继承了 ID3 算法的优点，并在以下几方面对 ID3 算法进行了改进[5]：a. 用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足。b. 在树构造过程中进行剪枝。c. 能够完成对连续属性的离散化处理。d. 能够对不完整数据进行处理。这样，C4.5 算法构造树的准确率更高，产生的分类规则更容易理解，以信息增益率选择分裂属性克服 ID3 的缺点，并能把连续型的属性变量离散化。同时，在构造决策树过程中，能够对缺失属性值的进行处理，采用后剪枝技术等。[3]

4 实验仿真

选取 UCI 隐形眼镜数据集，通过 ID3 算法与 C4.5 算法构建决策树来根据患者的年龄 (age)、症状 (prescript)、是否散光 (astigmatic)、眼泪数量 (tearRate) 等属性推荐隐形眼镜类型。

4.1 实验数据集

实验采用 UCI 隐形眼镜数据集，它包含很多患者眼部状态的观察属性以及医生推荐的隐形眼镜类型。隐形眼镜类型包括硬材质 (hard)、软材质 (soft) 以及不适合佩戴隐

形眼镜 (no lenses)。

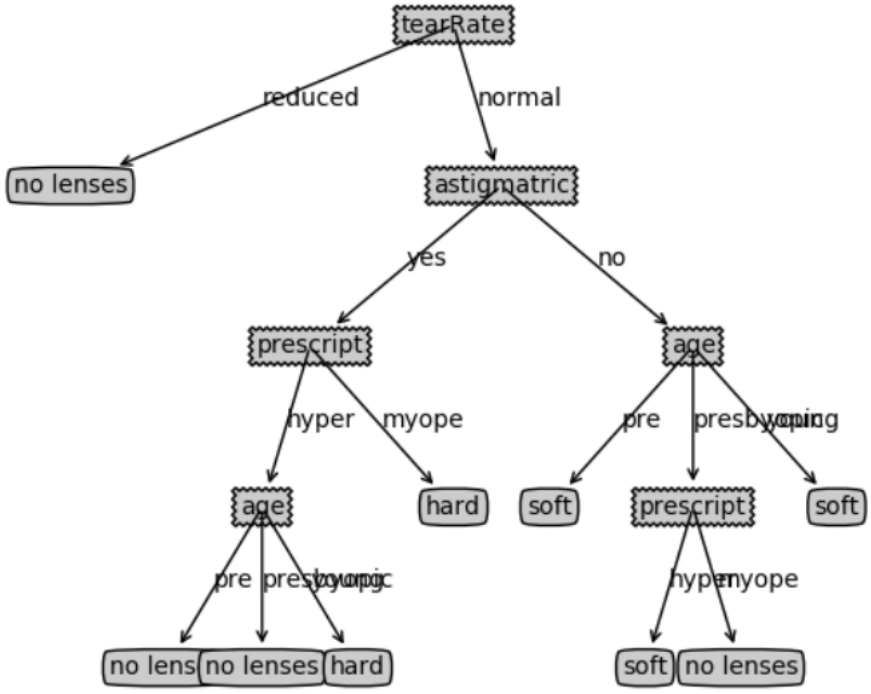
数据集每一行数据的 Labels 依次是序号 (ID)、年龄 (age)、症状 (prescript)、是否散光 (astigmatic)、眼泪数量 (tearRate)、分类 (class)。训练数据集有 100 组数据，测试数据集有 24 组数据。

标签属性：

1. age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic
2. spectacle prescription: (1) myope, (2) hypermetrope
3. astigmatic: (1) no, (2) yes
4. tear production rate: (1) reduced, (2) normal

4.2 实验结果

根据上面描述的算法，我们构建出隐形眼镜决策树：



ID3 算法在训练数据集上可以达到 94% 的准确率，在测试数据集上准确率为 91.6667%，C4.5 算法在训练数据集上可以达到 94% 的准确率，在测试数据集上准确率为 91.6667%

5 总结与展望

作为决策树中经典算法,ID3 算法使用信息增益作为分类标准, 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。C4.5 算法在 ID3 算法的基础上进行改进，提高了准确率。目前，决策树分类算法应用领域十分广泛, 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决策树分类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, 会有更多改进算法被提出，决策树算法也会在更多的领域得到应用。

参考文献

[1] 徐梦茹，王学明. 决策树几种分类算法的分析比较.《电脑知识与技术》, 2018.07

[2] 曹颖超. 决策树分类算法及其应用《信息技术》，2018.01

[3] 马俊宏. 数据挖掘技术决策树分类算法分析、比较与实验. 《北京印刷学院学报》, 2017.11

[4] 杜威铭, 冉羽. 决策树 ID3 算法研究.《科技视界》, 2018.11

[5] 张宏，高长松. C4.5 算法对 ID3 算法的改进. 《计算机光盘软件与应用》，2012.06

[6] 刘瑞玲. C4.5 决策树分类算法性能分析. 《信息系统工程》，2014.05

[7] 王文霞. 数据挖掘中改进的 C4．5 决策树分类算法. 《吉林大学学报》，2017.09