采用数据挖掘技术中 ID3 决策树算法分析学生成绩

中国海洋大学信息科学与工程学院在职研究生 张 媛

1. 引言

当前，职业技术教育随着社会发展和科技进步，其办学软硬件层次正逐步“升级”，办学规模和社会影响力也成倍增长。在学校管理工作中，特别是对学生的成绩管理工作中，普遍存在的问题是学生成绩数据量过于庞大，但对这些数据的处理还停留在初级的数据备份、查询及简单统计阶段，并没有对大量的成绩数据进行深入地分析，加以捕捉有利于教学管理工作的信息，这是对教学信息资源极大的浪费。数据挖掘技术正是解决这个问题的可行而有效的方法。本文使用 ID3 决策树算法生成决策树分析学生成绩优良与哪些因素有关，并利用事后修剪法对决策树进行修剪，最后由决策树产生分类规则。

2. 数据挖掘的方法和技术

数据挖掘方法是由人工智能、机器学习的方法发展而来，结合传统的统计分析方法、模糊数学方法及科学计算可视化技术，以数据库为研究对象，形成了数据挖掘的方法和技术。可分为以下六大类：归纳学习法、仿生物技术、公式发现、统计分析方法、模糊数学方法、可视化技术。

信息论方法（决策树方法）是归纳学习法中的一类。信息论方法是利用信息论的原理建立决策树。在知识工程领域，决策树是一种简单的知识表示方法，它将事例逐步分类成代表不同的类别。由于分类规则是比较直观，易于理解，该类方法的实用效果好，影响较大。由于该方法最后获得知识表示形式是决策树，故一般称它为决策树方法。这种方法一般用于分类任务中。

决策树是通过一系列规则对数据进行分类的过程。它提供一种在什么条件下会得到什么值的类似规则的方法。决策树是以实例为基础的归纳学习算法。从一组无次序、无规则的元组中推理出决策树表示形式的分类规则。它采用自顶向下的递归方式，在决策树的内部节点进行属性值的比较，并根据不同的属性值从该节点向下分支，叶节点是要学习划分的类。从根节点到叶节点的一条路径就对应着一条分类规则，整个决策树就对应着一组析取表达式规则。

信息论方法中较有特色的方法有 ID3,IBLE 方法。目前已形成了多种决策树算法，如 CLS、ID3、CHAID、C ART、FACT、C4.5、GINI、SEE5、SLIQ、SPRINT 等。其中最著名的算法是 Quinlan 提出的 ID3 算法。

3. 决策树的生成

决策树的生成分为学习及测试两个阶段。决策树学习阶段采用自顶向下的递归方式。决策树算法分成两个步骤：一是树的生成，开始时所有数据都在根节点，然后递归地进行数据划分，直至生成叶节点。二是树的修剪，就是去掉一些可能是噪音或者异常的数据。决策树停止分割的条件有：一个节点上的数据都是属于同一个类别；没有属性可以再用于对数据进行分割。

建立一棵决策树可能只要对数据库进行几遍扫描之后就能完成，这也意味着需要的计算资源较少，而且可以很容易地处理包含很多预测变量的情况，因此决策树模型可以建立得很快，并适合应用到大量的数据上。

4. ID3 算法

决策树归纳的基本算法是贪心算法，它以自顶向下递归的方法构造决策树。著名的决策树归纳算法 ID3 算法的基本策略如下：

- 树以代表训练样本的单个节点开始。
- 如果样本都在同一个类中，则这个节点成为树叶节点，并用该类标记。
- 否则，算法使用称为信息增益的基于熵的度量作为启发信息，选择能够最好地将样本分类的属性，该属性成为该节点的“测试”或“判定”属性。在这里，我们假设所有的属性都是分类的，即取离散值。连续值的属性必须离散化。
- 对测试属性的每个已知的值创建一个分支，并据此划分样本。
- 算法使用类似的方法，递归地形成每个划分上的样本决策树。一旦一个属性出现在一个节点上，就不必在该节点的后代上考虑这个属性。
- 整个递归过程在下列条件之一成立时停止：

⑴ 给定节点的所有样本属于同一类。

⑵ 没有剩余属性可以用来进一步划分样本，这时候将该节点作为树叶，并用剩余样本中所出现最多的类型作为叶子节点的类型。

⑶ 某一分枝没有样本，在这种情况下，以训练样本集中占多数的类创建一个树叶。

但是，ID3 算法也存在着如下不足：

⑴ 不能够处理连续值属性，ID3 算法最初定义时是假设所有属性值是离散的，但在现实环境中，很多属性值是连续的。

⑵ 计算信息增益时偏向于选择取值较多的属性，这样不太合理。

⑶ 对噪声较为敏感，所谓噪声是指训练集中属性值或类别给错的数据。

⑷ 在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效。

⑸ 只适合于能够驻留于内存的数据集使用，当训练集大得无法在内存容纳时程序无法运行。

5. 树的剪枝

当决策树创建时，由于数据中的噪声和孤立点，许多分枝反映的是训练中的异常。剪枝方法处理这种过分适应数据问题。通常，这种方法使用统计度量，剪去最不可靠分枝，这可带来较快的分类，提高决策树独立于测试数据正确分类的能力。有两种常用的剪枝方法：

先剪枝方法（prepruning），通过提前停止树的构造而对树剪枝。一旦停止，节点成为树叶。该树叶持有子集样本中出现最频繁的类。在构造树时，如统计意义下的 χ2、信息增益等度量，可以用于评估分裂的优良性。如果在一个节点划分样本将导致低于预定义阈值的分裂，则给定子集的进一步划分将停止。然而，选择一个适当的阈值是困难的。较高的阈值可能导致过分简化的树，而较低的阈值可能使得树的简化太少。

后剪枝方法（postpruning），它由完全生长的树剪去分枝。通过删除节点的分枝，剪掉树节点，代价复杂性剪枝算法是后剪枝算法的一个实例。在该算法中，最下面的未被剪枝的节点成为树叶，并用它先前的分枝中最频繁的类进行标记。对于树中每一个非树叶节点，算法计算该节点上的子树被剪枝后可能出现的期望错误率。然后，使用每个分枝的错误率，结合沿每个分枝观察的权重评估，计算不对该节点剪枝的期望错误率。如果剪去该节点，导致较高的期望错误率，则保留该子树；否则剪去该子树。产生一组逐渐被剪枝的树之后，使用一个独立的测试集评估每棵树的准确率，就能得到具有最小期望错误率的决策树。

也可以交叉使用先剪枝和后剪枝，形成组合式方法。后剪枝所需的计算比先剪枝多，但通常产生更可靠的树。

6. 从决策树提取分类规则

从决策树提取分类规则时，规则使用 if…then 的形式表示出来，对从根到树叶的每一条路径创建一条规则，沿着路径上的每一个属性——值对，形成规则前件（“IF”部分）的一个合取项。叶节点包含类预测，形成规则后件（“THEN”部分）。if…then 规则易于理解，特别是当给定的树很大时，而且便于规则匹配等操作。

7. 结论

数据挖掘虽然还是一门新兴的数据分析技术，但已经具有了强大的生命力，其研究取得了令人瞩目的成就，已经成功地应用到了许多领域。可以说，有数据积累的地方，就有数据挖掘技术的用武之地，这是因为它直接与经济和决策紧密相连。