DOI:10.13718/j.cnki.xsxb.2017.01.010

基于决策树模型的计算机课程教学的学生属性影响分析①

成 启 明

河北旅游职业学院 信息技术系,河北 承德 067000

摘要:分析了 ID3 型决策树的挖掘分类过程,并结合距离测度和阈值判断进行了改进. 设计了 5 种学生属性,分别是兴趣属性、课堂学习属性、作业完成属性、预复习属性、课外训练属性. 借助改进的决策树模型,结合 500 个样本数据,分析了学生属性对于计算机课程教学效果的影响. 实验结果表明,兴趣属性、预复习属性、作业完成属性对计算机课程教学效果的提升最为重要.

关 键 词:决策树;学生属性;计算机课程;教学效果

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-5471(2017)01-0060-06

近年来,如何搞好计算机课程的教学工作已经成为教育界普遍关注的焦点问题[1-2]. 人们试图去挖掘影响计算机课程教学效果的主要因素,继而制定有针对性的教学方法以提升教学效果. 计算机课程的教学工作在具体实施过程中受到诸多因素的影响,诸如社会因素、学校因素、教师因素、学生因素等等[3-4]. 为分析各种因素对计算机教学的影响,聚类分析、关联规则挖掘、神经网络关系分析、决策树分析等数据挖掘方法被引入进来[5-6]. 其中,经典的决策树算法包括 ID3 型决策树算法、C4.5 型决策树算法、CRT 型决策树算法等等[7-8]. 现有的决策树算法存在的一些共性问题是:①对于复杂关系的挖掘,决策树的深度过大,导致挖掘过程的执行效率偏低[9];②决策树的构建过分依赖于最优特征,而其他特征则被忽视,导致分类挖掘结果存在片面性[10]. 本文提出一种改进的决策树模型,将其应用于计算机课程的教学分析工作之中,去挖掘学生属性对于计算机课程教学效果的影响.

1 改进的决策树模型

1.1 经典的决策树模型

在各种决策树算法中,ID3 型决策树算法是应用最广泛的一种,它的基本思路是选择最佳的特征属性作为决策树的根节点,然后对根节点的各个取值对应地形成分支;以此类推,在新的分支上再形成新的节点,直到遍历所有特征属性,形成叶子节点. 可见,对于 ID3 型决策树算法而言,最佳特征属性的选择标准是决定决策树优劣的关键所在. ID3 型决策树算法,是采用信息熵和相关增益来设置节点的特征属性,进而做出决策.

假设要处理的数据集合为 T,其中含有 n 个样本. 如果将这个集合映射为 d 个分类,各个分类用 D_i 进

① 收稿日期:2016-06-06

作者简介:成启明(1982-),女,河北承德人,硕士,讲师,主要从事计算机网络安全研究.

行定义，每个分类 D_i 中含有 n_i 个样本，那么可以按照如下的公式计算出信息熵.

$$E(T) = \sum_{i=1}^{d} p_i \log_2(p_i) \tag{1}$$

这里，p_i 表达了集合 T 中的样本可以纳入分类 D_i 中的概率，p_i 的计算为

$$p_i = \frac{n_i}{n} \tag{2}$$

再设定一个属性集合 R_C，它容纳了属性 C 可能出现的所有值. T_u 表达了 T 的一个子集，它对应着属性 C 表现为 u 时 T 中的对应样本，其数学形式如公式(3) 所示.

$$T_u = \{t \in T \mid C(t) = u\} \tag{3}$$

在属性 C 之下的各个分支节点上，T_u 所对应的信息熵用 $E(T_u)$ 来表示. 属性 C 是否被选择，取决于各个节点上 T_u 的加权求和值，这里的权重计算为

$$w_u = \frac{|T_u|}{|T|} \tag{4}$$

属性 C 的期望熵计算，如公式(5) 所示.

$$E(T, C) = \sum_{u \in R_C} w_u E(T_u) \tag{5}$$

进一步可以计算出属性 C 相对于集合 T 的信息增益，如公式(6) 所示.

$$G(T, C) = E(T) - E(T, C) \tag{6}$$

特征属性的信息增益越大，被节点选择的可能性也就越大.

1.2　本文的改进措施

经典的 ID3 决策树方法，对于复杂问题的挖掘时会出现较深的深度，而且除了最佳特征以外其他特征没有被有效利用. 为了在计算机课程教学分析中获得更好的应用效果，本文对经典的 ID3 决策树方法进行改进，结合距离测度和阈值判断完成决策树构建.

数据挖掘的关键在于判断同类数据的相似性，距离测度是一种有效的方法. 令 q 表示数据变量，本文设计的距离测度为

$$\| q - Q \| = (q - Q)^T K^{-1} (q - Q) \tag{7}$$

这里，Q 代表参照数据，K 代表常值矩阵.

对于决策树挖掘分析的过程，本文配合设计一个阈值集合 $H = \{h_i\}$，其中 $1 \leqslant i \leqslant d$. 再设定属性集合 $C = \{c_i\}$，其中 $1 \leqslant i \leqslant n$. 对应设置一个包含于 C 的子集 $F = \{f \mid f \subseteq C\}$. 再设置 2 个阈值：① 错误分类阈值 H_e，② 交叉错误分类阈值 H_c.

如果存在一个集合 $D \subseteq C$，并且 D 容纳了不同类别的样本，那么对应于 $f \in F$，D 在这个 f 上的数据取值为 $q = (q_1, \cdots, q_n)$，并且存在 $q_i \in f$ 和某个类的距离测度(计算如公式(7)) 最小，就把这个样本 q_i 纳入这个类别 D.

同时，根据错误分类阈值 H_e 的约束，最大限度地实现正确分类；根据交叉错误分类阈值 H_c 的约束，尽可能地避免交叉分类.

2　学生属性影响计算机课程教学的变量设置

2.1　变量设计

本文的研究目的在于借助改进的决策树模型，对学生属性和计算机课程教学效果之间的关系展开研究，并确定出对计算机课程教学最重要的学生属性.

为了便于构建计算机课程教学效果和学生属性之间关系的决策树，本文设计了这样一些变量：

变量 1：成绩总评，用学生计算机课程成绩来反映，60～100 之间为“通过”，0～59 之间为“未通过”.

变量2：学生性别，学生的静态属性.

变量3：学生学院，学生的静态属性.

变量4：学生专业，学生的静态属性.

变量5：兴趣属性，学生的动态行为属性，反映了学生对计算机课程的喜爱程度，设置了兴趣浓厚、兴趣一般2种情况.

变量6：课堂学习属性，学生的动态行为属性，反映了学生在课堂上的学习效果，分为认真听讲及课堂知识掌握扎实、不认真听讲2种情况.

变量7：作业完成属性，学生的动态行为属性，反映了学生对待作业的态度，分为及时独立完成、完成作业不够及时认真2种情况.

变量8：预复习属性，学生的动态行为属性，反映了学生是否具有预习和复习的主动性，分为进行预复习、没有预复习2种情况.

变量9：课外训练属性，学生的动态行为属性，反映了学生是否会选择其他课外训练方式提升计算机水平的行为，分为参加课外训练、未参加课外训练2种情况.

2.2 数据遴选

为了获取本文模型分析工作的原始数据，本文对全校学生的计算机教学情况进行了调查，结合教师评价和学生自评，形成了对每一位学生的属性评价并纳入一个统一的数据库中. 本次实验中，在数据库中抽取了500个样本，其中3个样本的情况如表1所示.

表1 学生属性影响计算机课程教学效果的3个样本数据

序号	学号	姓名	性别	学院	专业
1	15030201	纪梦	女	信息工程学院	智能信息处理
兴趣属性	课堂属性	作业属性	课外属性	预复习属性	成绩总评
兴趣浓厚	认真听讲	及时独立完成	未参加课外培训	进行预复习	95(通过)
序号	学号	姓名	性别	学院	专业
2	15030202	周远洋	男	信息工程学院	智能信息处理
兴趣属性	课堂属性	作业属性	课外属性	预复习属性	成绩总评
兴趣一般	认真听讲	完成不及时	参加课外培训	没有预复习	82(通过)
序号	学号	姓名	性别	学院	专业
3	15030204	王怀明	男	信息工程学院	智能信息处理
兴趣属性	课堂属性	作业属性	课外属性	预复习属性	成绩总评
兴趣一般	不认真听讲	完成不及时	未参加课外培训	没有预复习	43(未通过)
……	……	……	……	……	……

3 学生属性影响计算机课程教学的决策树分析结果

在实验过程中，借助本文提出的改进决策树模型，以“成绩总评”为根节点，以“兴趣属性”、“课堂属性”、“作业属性”、“课外属性”、“预复习属性”为中间节点和叶子节点，进行决策树的优化迭代构建.

决策树遍历可能出现的结构，结合距离测度和阈值判断，最终形成的决策树结构如图1所示.

对应于图1的决策树结构，500个样本数据在各个节点的比例分配如表2所示.

根据图1中的决策树结构和表2中的数据配置，可知第8号节点实现了最佳的教学效果，符合此节点配置的39名同学全部通过了计算机课程考试，通过率为100%. 满足第8号节点的这39名同学，都具有这样的共同属性：兴趣浓厚、进行预复习、及时独立完成作业，这说明上述3种学生属性对计算机课程的教学效果影响最大.

进一步分析6个叶子节点的增益情况及这些节点对决策树模型整体的贡献度，结果如表3所示.

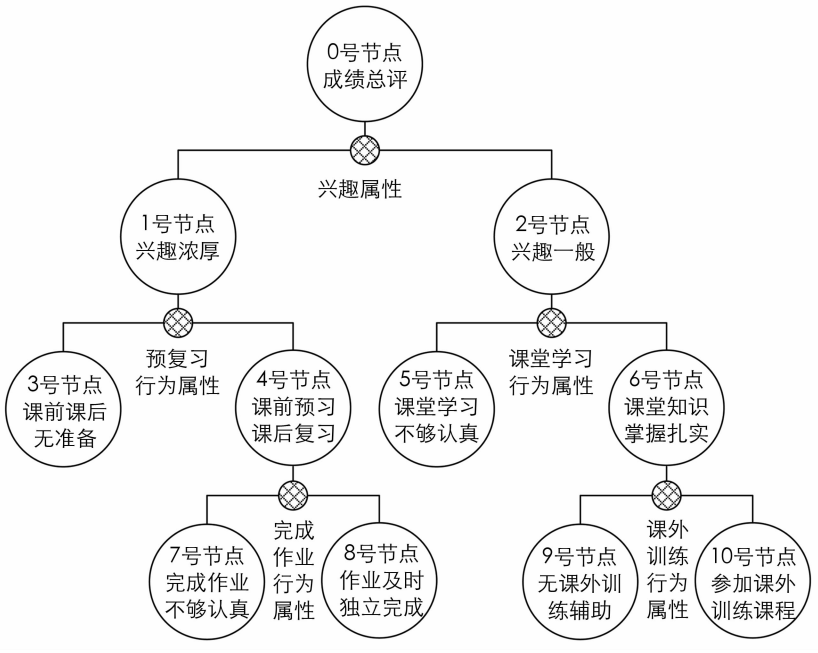

图 1　决策树模型最终的构建结果

表 2　样本数据在各个节点的分配情况

节点	属性	参数	通过	未通过	合计
第 0 号节点	成绩总评	比例/%	82.2	17.8	100.0
		学生数	411	89	500
第 1 号节点	兴趣属性	比例/%	88.8	11.2	100.0
	兴趣浓厚	学生数	294	37	331
第 2 号节点	兴趣属性	比例/%	69.2	30.8	100.0
	兴趣一般	学生数	117	52	169
第 3 号节点	预复习属性	比例/%	82.6	17.4	100.0
	无预习复习	学生数	90	19	109
第 4 号节点	预复习属性	比例/%	91.9	8.1	100.0
	有预习复习	学生数	204	18	222
第 5 号节点	课堂学习	比例/%	55.6	44.4	100.0
	不够认真	学生数	20	16	36
第 6 号节点	课堂学习	比例/%	72.9	27.1	100.0
	掌握扎实	学生数	97	36	133
第 7 号节点	完成作业	比例/%	90.2	9.8	100.0
	不够认真	学生数	165	18	183
第 8 号节点	完成作业	比例/%	100.0	0.0	100.0
	及时独立	学生数	39	0	39
第 9 号节点	课外训练	比例/%	73.2	26.8	100.0
	没有参加	学生数	60	22	82
第 10 号节点	课外训练	比例/%	72.5	27.5	100.0
	有参加	学生数	37	14	51

表3 叶子节点的增益和贡献情况分析结果

节点	属性值		增益值		响应度/%	贡献度/%
	学生数	比例/%	学生数	比例/%		
第8号节点	39	39/500=7.8	39	39/411=9.5	100.0	100/82.2=121.7
第7号节点	183	183/500=36.6	165	165/411=40.1	90.2	90.2/82.2=109.7
第3号节点	109	109/500=21.8	90	90/411=21.9	82.6	82.6/82.2=100.5
第9号节点	82	82/500=16.4	60	60/411=14.6	73.2	73.2/82.2=89.1
第10号节点	51	51/500=10.2	37	37/411=9.0	72.5	72.5/82.2=88.2
第5号节点	36	36/500=7.2	20	20/411=4.9	55.6	55.6/82.2=67.6

表3中属性值的比例一项，用叶子节点中的学生合计总数除以样本总数得出；增益值的比例一项，用叶子节点中的通过学生总数除以样本通过总数得出；响应度一项，用叶子节点中的学生通过总数除以学生合计总数得出；贡献度一项，用叶子节点的响应度除以样本总体通过率得出.

综合表3中的结果可以看出，6个叶子节点对于决策树整体构建的贡献度有所差异. 其中，第8号节点的贡献度最大，为121.7%；第5号节点的贡献度最小，为67.6%. 这一结果再次证实，兴趣浓厚、进行预复习、及时独立完成作业这3项属性对于计算机课程教学效果提升的重要性.

为了进一步观察本文提出的改进决策树挖掘方法的性能，选择ID3型决策树方法作为比较. 进行的数据挖掘实验表明，相比于ID3型决策树方法，本文提出的改进决策树挖掘方法性能更优. 不仅如此，本文方法的执行效率也要优于ID3型决策树方法，2种方法的执行效率比对结果，如图2所示.

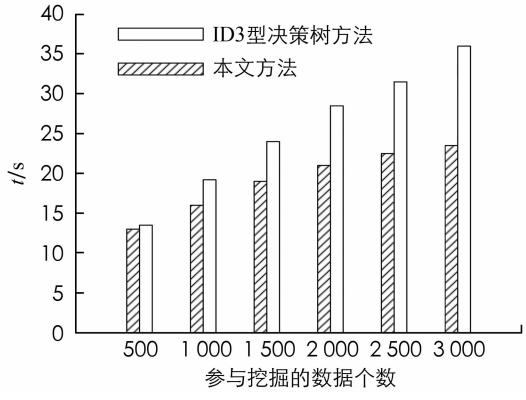

图2 2种方法效率的对比结果

从图2中可以看出，随着被挖掘数据量的逐渐增大，本文方法的挖掘时间都低于ID3型决策树方法. 随着被挖掘数据量的逐渐增大，2种方法的挖掘时间都在增加，但本文方法挖掘时间的增加速度明显低于ID3型决策树方法.

4 结 论

本文在经典ID3型决策树的基础上，提出了一种改进的决策树模型，并用于本文核心问题的分析之中，设计了兴趣属性、课堂学习属性、作业完成属性、预复习属性、课外训练属性来分析学生属性对于计算机课程教学效果的影响.

依托学校数据库中500个样本数据展开实验研究，根据构造出的决策树模型和叶子节点的增益贡献情况，可知兴趣浓厚、进行预复习、及时独立完成作业这3项属性对计算机课程教学效果的影响最大.

因此，要进一步提升计算机课程的教学效果，必须要全力激发学生的学习兴趣，做好课前预习、课后复习工作，并培养学生养成及时独立地完成作业的习惯.

参考文献：

[1] CHEN Q F, CHEN Y P. Discovering Inhibition Pathways for Protein Kinases [J]. IEEE Transactions of Intelligent Systems, 2012, 27(5): 19-26.

[2] 曾 旭, 司马宇. K-Means算法在计算机等级考试成绩分析中的应用 [J]. 软件导刊, 2012, 11(11): 19-21.

[3] 张慧琴, 郭 璐. 数据挖掘(DM)技术在教学档案管理中的应用研究 [J]. 黑龙江教育理论与实践, 2016(Z1): 91-92.

[4] CHEN Q F, CHEN Y P, ZHANG C Q. Detecting Inconsistency in Biological Molecular Databases Using Ontology [J].

Data Mining and Knowledge Discovery，2014，15(2)：275－296.

[5] 周　骏. 计算机组成原理课程教学改革思考 [J]. 西南师范大学学报(自然科学版)，2014，39(6)：161－165.

[6] KRIEGEL H P，BORGWARDT K M，KROGER P，et al. Future Trends in Data Mining [J]. Data Mining and Knowledge Discovery，2016，15(1)：87－97.

[7] 吴春琼，胡国柱，徐　静. 高职院校项目驱动模式下基于数据挖掘决策树分类的教学效果分析 [J]. 吕梁教育学院学报，2015，32(1)：18－21.

[8] 刘光洁，王文永，吴登峰，等. 基于模糊聚类的决策树算法在教学质量评价中的应用 [J]. 东北师范大学学报(自然科学版)，2009，41(3)：36－39.

[9] JANSIRANI P G，BHASKARAN R. Extraction of Dominant Attributes and Guidance Rules for Scholastic Achievement Using Rough Set Theory in Data Mining [J]. International Journal of Computer Science Issues，2015，7(3)：22－28.

[10] 戴慧珺，桂小林，张　成，等. 基于历史大数据决策树分类的 MOOC 教学评估方法研究 [J]. 计算机教育与教学研究，2015(22)：52－55.

On Influence of Computer Course Teaching upon Students′ Attributes Based on Decision Tree Model

CHENG Qi-ming

Hebei Tourism Vocational College，Chengde Hebei 067000，China

Abstract：The classification process of ID3 decision tree has been analyzed，and the combination of distance measure and threshold value judgment been improved. 5 kinds of students′ attributes have been designed，which are interest attributes，classroom learning attributes，job completion attributes，pre-review attributes，and extracurricular training attributes. With the aid of the improved decision tree model，combined with the 500 sample data，the influence of the students′ attributes on the teaching effect of computer course has been analyzed. The experiment results show that the most important is the improvement of the effect of the computer course teaching by the property of interest，the attribute of the pre-review and the completion of the work.

Key words：decision tree；students attribute；computer course；teaching effect

责任编辑　夏　娟