

基于决策树的网上学习资源建设策略研究

毛 布¹, 谢 汶²

(1. 四川自贡广播电视大学, 四川 自贡 643000; 2. 四川大学计算机学院, 成都 610000)

摘 要:现代远程教育中资源建设人员常常根据各类资源的特征进行分析, 以便促进学员网上学习浏览。从数据挖掘技术的角度探讨基于资源特征的分析不仅能促进学员点击率, 又能帮助制定相应的资源建设策略。文章介绍了决策树算法原理, 讨论了网上学习资源的各项具体特征, 应用计算实例说明资源的特征促进点击率决策树分类模型, 利用工具 Analysis Manager 中的决策树方法进行促进点击率规则的数据挖掘。研究表明该应用是切实可行的。

关键词:资源特征; 促进点击率; 数据挖掘; 决策树算法

中图分类号: TP311

文献标识码: A

现在远程教育中利用数据挖掘分析网上资源的特征情况, 促进学员网上学习点击率, 将给远程教育带来一系列新的改革。资源建设策略的建立是一种通过收集网上教学资源特征信息的技术, 并希望通过相关人工智能技术进行知识挖掘, 以支持管理决策^[1]。数据挖掘技术正是数据库营销的技术基础, 其原理大部分都离不开数理统计方法, 该项技术近年的应用集中在一些行为或对象的信息处理上^[2]。例如根据对象(网上教学资源)的特征, 通过收集和处理其关系到的信息来制定资源建设策略。

从实际制定资源建设策略的角度看, 建设者最容易遇到的问题是如何将所面临的对象进行分类。特别是做为资源建设者, 应快速地确定网上的教学资源类型, 提取相关特征, 从而有效的判断其特征点, 做出相应建设规划, 制定建设此类别的资源方案及策略^[3]。从技术的角度看, 决策树是一种不错的分类方法, 简单有效, 应用广泛。

本文讨论资源建设的决策树算法模型, 通过分析资源特征与学员点击行为的关系, 指导制定建设策略, 提高效率。

1 决策树算法原理

1.1 决策树分类模型

决策树模型是最早由 Hunt 提出, 他将概念表示成

“属性—值”形式^[4]。例如, 对网上教学资源的描述有多种属性: 类型、大小、创建时间、资源来源和访问来源等。属性的值域可表示为:

(1) 属性(类型) = { 视频、文本 }。

(2) 属性(大小) = { 大于 10M、小于 10M }。

(3) 属性(创建时间) = { 2008 年、2009 年、2010 年 }。

(4) 属性(资源来源) = { 中央电大资源、省电大资源、地方自建资源 }。

(5) 属性(访问路径) = { 省电大、自贡电大、荣县分校网站、163、百度、搜狗 }。

概念学习系统中的决策树节点就是决策属性, 对应于待分类对象的属性, 由某一节点引出的弧对应于这个属性的可能取值, 叶节点对应于分类的结果。图 1 表示了一棵决策树。显然, 决策树本身就对应着一种分类模式。

要提高搜索树的效率, 首先必须保证树是一棵理想、最优的决策树。为了提高效率, QUINLAN 提出了一种启发式搜索算法, 称为 ID3 算法。它以信息熵和信息增益度为衡量标准, 搜索原则是首先选择熵增益最大的节点。C4.5、C5.0 算法对 ID3 进行了改进^[5]。

1.2 决策树 ID3 算法

ID3 算法的基本步骤是:

(1) 选择属性表 $Attrlist = \{ A_1, A_2, \dots, A_i, \dots, A_n \}$, 检

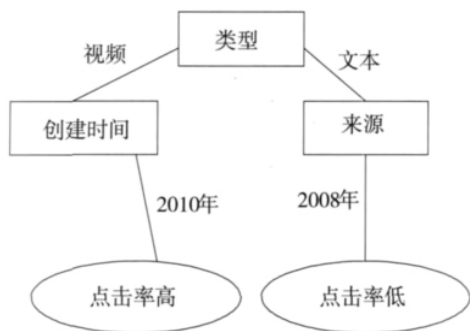


图1 一棵网上教学资源属性构成的点击行为树

测属性设为 A_i 。

(2) A_i 的值域 $Valuetype(A_i) = \{V_1, \dots, V_s\}$ 的 S 个取值把训练实例集 T 分为 S 个子集, 则 $T = \bigcup_{k=1}^s T_k^{(i)}$; 子集 $T_k^{(i)}$ 中的所有实例的属性 A_i 的取值为 V_k 。

(3) T 中实例分类结果组成 $class = \{C_1, C_2, \dots, C_j, \dots, C_m\}$, C_j 的实例数为 e_j , $1 \leq j \leq m$, 且 $\sum_{j=1}^m e_j = |T|$, $|T|$ 表示训练实例集 T 的实例总数, 实例分类结果为 C_j 的概率为 $P_j = e_j / |T|$ 。

(4) 求取相对信息熵。定义训练实例集 T 的实例信息量为:

$$I(T) = - \sum_{j=1}^m P_j \log P_j = - \sum_{j=1}^m \frac{e_j}{|T|} \log \frac{e_j}{|T|} = - \left(\sum_{j=1}^m e_j \log \frac{e_j}{|T|} \right) / |T| \quad (1)$$

定义子集 $T_k^{(i)}$ 的实例平均信息量为:

$$I(T_k^{(i)}) = - \left(\sum_{j=1}^m e_j^{(i)} \log \frac{e_j^{(i)}}{|T_k^{(i)}|} \right) / |T_k^{(i)}| \quad (2)$$

子集实例数与实例总数关系为:

$$\sum_{j=1}^m e_j^{(i)} = |T_k^{(i)}| \quad (3)$$

如果选择属性 A_i 作为检测属性, 将训练实例集 T 分为 S 个子集后, 可以由各实例子集的实例总信息量 $|T_k^{(i)}| \cdot I(T_k^{(i)})$ 之和对实例集 T 的实例总数 $|T|$ 的平均值来表示实例集 T 的实例平均信息量, 相对信息熵为:

$$I(T, A_i) = \left(\sum_{k=1}^s |T_k^{(i)}| \cdot I(T_k^{(i)}) \right) / |T| \quad (4)$$

(5) 搜索的启发式称为熵增益原理。

$$GI(T, A_i) = I(T) - I(T, A_i) \quad (5)$$

ID3 选择信息量最大的属性 A_i 作为检测属性划分实例集, 达到分类的目的^[6-7]。

2 资源特征建设策略与决策树模型

现在资源建设人员想通过了解详细的网上资源特征来掌握学员网上学习心理, 为了获得学员在自然状态

下的网上点击资源行为特征、对自己偏好的资源感应规律, 本文通过对网上资源外表特征的分析 and 评价, 在自然状态下跟踪和记录学员对指定资源的感应和行为变化; 再利用以上决策树算法模型建立网上资源建设策略, 并使用决策树挖掘工具进行实例分析^[8]。

2.1 实验方法与步骤

选取自贡电大 2010 年 11 月 27 日 15:35 时网上资源为数据来源。总共收集了 50 门课程资源进行抽样并建立了相应的记录表。并把这些数据输入到数据库中, 然后根据决策树模型由研究人员进行数据处理。此方法称为网上资源特征分析。

2.2 数据预处理

表 1 至表 3 是指定的 50 门课程的网上学员学习行为统计表, 学习时间是 2010-1-13 至 2010-11-27。资源建设者给此课程资源的定位为地方电大自建资源丰富、视频资源众多、文件类型大都大于 10M 和辅导资源齐全的热门课程。

实验中共记录了 50 门课程的特征, 表 2 为网上资源特征分布表, 表 3 为对应的网上资源学习点击情况分布表。50 门课程中点击率超过 50% 的资源课程共 20 门, 其他课程资源学员注册后点击率不足 50%。

表 1 网上资源特征与学员点击学习行为评估指标体系

指标	选择项
资源主要类型(I1)	A. 视频; B. 文本
建立日期(I2)	A. 2010 年; B. 2010 年前
文件大小(I3)	A. 10M 以内; B. 10M 以上
访问来源(I4)	A. 电大网址; B. 百度; C. 其它
访问日期(I5)	A. 4、10 月; B. 1、7 月; C. 其它时间段
资源是否包括自建部分(I6)	A. 是; B. 否
是否有教学大纲(I7)	A. 是; B. 否
是否网上互动(I8)	A. 是; B. 否
是否有教学说明(I9)	A. 是; B. 否
是否有教学辅导(I10)	A. 是; B. 否
是否有习题答案(I11)	A. 是; B. 否
是否有往届考题(I12)	A. 是; B. 否

表 2 各特征各类别分布情况

特征	A	B	C	总计
资源主要类型(I1)	31	19	-	50
建立日期(I2)	10	40	-	50
文件大小(I3)	50	0	-	50
访问来源(I4)	15	10	25	50
访问日期(I5)	29	12	9	50
资源是否包括自建部分(I6)	24	26	-	50
是否有教学大纲(I7)	50	0	-	50
是否网上互动(I8)	50	0	-	50
是否有教学说明(I9)	50	0	-	50
是否有教学辅导(I10)	41	9	-	50
是否有习题答案(I11)	35	15	-	50
是否有往届考题(I12)	12	38	-	50

表 3 各特征各类别的点击分布情况

特 征	A	B	C	总计
资源主要类型(I1)	13	7	-	20
建立日期(I2)	10	10	-	20
文件大小(I3)	16	4	-	20
访问来源(I4)	10	5	5	20
访问日期(I5)	8	7	5	20
资源是否包括自建部分(I6)	20	0	-	20
是否有教学大纲(I7)	20	0	-	20
是否网上互动(I8)	20	0	-	20
是否有教学说明(I9)	19	1	-	20
是否有教学辅导(I10)	18	2	-	20
是否有习题答案(I11)	10	10	-	20
是否有往届考题(I12)	12	8	-	20

2.3 决策树模型与分类实例

经过对数据预处理后,可以利用信息熵分析网上建设资源定位是否准确,也可以通过最后规则的建立为资源建设人员提供推销策略,以集中力量建设重点资源。下面分别给出面向“资源主要类型、是否包含自建资源和是否有往届考题”的信息熵增益情形。可以看出,资源类型的增益是最大的,这说明,资源建设者首先必须按“资源主要类型”进行分类,其次是“是否包含自建资源”,然后是“是否有往届考题”。这说明建设者的分类基本上是正确的。这样可得到一条建设规则:如果该课程资源主要类型为视频,课程资源中又包含有往届考题,且包含自建资源,则被学员点击学习的可能性会偏大。这说明,通常资源建设人员会根据上面的分类方法进行分,但还需要加入更多的元素,包括资源大小、是否有教学大纲和是否教学辅导等,从而可锁定对象,重点应对,提高学员点击率。

分类属性信息量与增益 “学习行为”信息总量 $I_0 = 1.3689$ 如下。

(1) 基于“资源主要类型”的分类 “资源主要类型”A 的信息量 $I_A = 1.0541$; “资源主要类型”B 的信息量 $I_B = 0.89563$; “资源主要类型”平均信息量 $I_{\bar{A}} = 0.5682$; “资源主要类型”信息增益 $GI = I_0 - I_{\bar{A}} = 0.8007$ 。

(2) 基于“是否包含自建资源”的分类 “是否包含自建资源”A 的信息量 $I_A = 1.4685$; “是否包含自建资源”B 的信息量 $I_B = 0.9586$; “是否包含自建资源”平均信息量 $I_{\bar{A}} = 1.3125$; “是否包含自建资源”信息增益 $GI = I_0 - I_{\bar{A}} = 0.0564$ 。

(3) 基于“是否有往届考题”的分类 “是否有往届考题”A 的信息量 $I_A = 1.3597$; “是否有往届考题”B 的信息量 $I_B = 0$; “是否有往届考题”平均信息量 $I_{\bar{A}} = 1.3445$; “是否有往届考题”信息增益 $GI = I_0 - I_{\bar{A}} = 0.0244$ 。

3 网上学习资源的特征印象决策树挖掘实例

前面的分类过程计算繁琐,在大数量情况下,必须借助计算机技术。决策树的程序化实现也比较简单,目前各大数据库提供商如微软提供的 Analysis Manager (数据分析与联机分析器)里就有决策树工具^[9]。本研究利用此工具进行网上资源印象特征的点击行为挖掘,并给出了分析结果。

分析结果显示了 29 条规则,包括只看或观察(a)、尝试点击(b)、网上学习(c)等三种行为的分类规则,每一类规则包含若干子规则。a 类行为规则包含 14 个子规则, b 类行为规则包含 9 个子规则, c 类行为规则包含 8 个子规则。

根据前面 ID3 算法,下面给出 a 类行为规则的 14 个子规则。a 类行为是“非网上学习”行为,只看或观察,资源建设者无需对这部分人群分出注意力。

基于学员的“非网上学习”规则集如下:

- Rule1: if I1 = B ∩ I5 = C ∩ I11 = B then action = a
- Rule2: if I1 = B ∩ I10 = B then action = a
- Rule3: if I2 = B ∩ I5 = C ∩ I8 = B then action = a
- Rule4: if I2 = B ∩ I5 = C ∩ I7 = B then action = a
- Rule5: if I2 = B ∩ I5 = C ∩ I9 = B then action = a
- ...

c 类行为是“点击”行为,这部分来到网上基本上有点点击学习意向,取决于学习资源对其吸引力。

基于学员的“网上学习”规则集如下:

- Rule1: if I1 = A ∩ I2 = A ∩ I7 = A then action = c
- Rule2: if I1 = B ∩ I5 = A ∩ I8 = A then action = c
- Rule3: if I2 = A ∩ I10 = A ∩ I12 = A then action = c
- Rule4: if I2 = A ∩ I9 = A ∩ I11 = A then action = c

对于资源建设者来说,最重要也最需要花时间精力应对的就是 b 类用户,这部分用户中有部分人有网上点击学习倾向,因此,资源建设者必须采用适当的手段,包括:新颖的资源类型、更新及时的资源库、方便友好的界面及其他策略。这些方法的应用必须根据具体的学员特征来实施。图 2 是对尝试点击学习行为规则的展开,显示了 3 条规则。例如:规则 1 表示如果学习资源的主要类型为视频,且是最近建立资源,文件大小适中,学员就愿意点击,反之学员则对该学习资源不感兴趣。

从数据挖掘的结果看,对于该资源的分类应该从资源类型、资源大小和资源更新程度着手,然后是是否为自建资源、是否有为辅导资料等考虑,而做为资源建设者还需要进一步细分才能提高资源应用点击率。

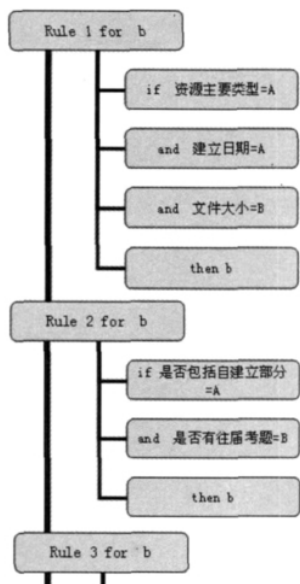


图2 尝试点击行为规则集

4 结束语

建立这种网上学习行为与资源外表特征印象关联模型的好处是为资源建设人员提供一些经验规则,以指导资源建设者在资源建设过程中处理好资源建设的特点^[10]。从另一个角度说,资源开发者可根据学员的喜好特征、群体密度来开发有针对性的产品,从而帮助实现针对性的建设个性化资源^[11]。决策树算法的数据挖掘技术,计算速度快,实现起来比较容易。未来的工作是将资源按类型、大小和更新时间等分类,然后进行实际网上跟踪和观察,集中进行规则挖掘,构成资源建设者

的外表特征印象与消费行为关联规则库,指导资源建设者的工作。

参考文献:

- [1] 丁荣涛. 基于决策树的高职学生网络学习分类模型构建[J]. 远程教育杂志, 2010(5): 24-26.
- [2] Berry M J A, Linoff G S. 数据挖掘技术[M]. 别荣芳, 尹静, 邓六爱, 译. 北京: 机械工业出版社, 2006.
- [3] 毛布. 基于动态聚类的网上学员细分实证研究[J]. 四川理工学院学报: 自然科学版, 2010, 23(6): 534-537.
- [4] Li S T, Shue L Y, Lee S F. Business intelligence approach to supporting strategy-making of ISP service management[J]. Expert System with Application, 2008, 35(3): 739-754.
- [5] 李强. 创建决策树算法的比较研究—ID3, C4.5, C5.0 算法的比较[J]. 甘肃科学学报, 2006, 18(4): 84.
- [6] Rygielski C, Wang J-C, Yen D C. Data mining techniques for customer relationship management[J]. Technology in Society, 2002, 24(24): 483-502.
- [7] Wen W. A knowledge-based intelligent electronic commerce system for selling agricultural products[J]. Computers and Electronics in Agriculture, 2007, 57(1): 33-46.
- [8] 许多顶. 网络数据库营销[J]. 商业研究, 2002(18): 121-123.
- [9] 唐晓宇. 个性化消费需求下的网络数据库营销的竞争优势[J]. 商业研究, 2002(4): 94-95.
- [10] 王慧敏, 陈泽宇, 王敏娟, 等. 移动学习情境中教育智能应用探究[J]. 中国远程教育, 2010(1): 68-71.
- [11] 钟志贤, 黄林凯. 对教学信息系统开发与应用的几点反思[J]. 中国远程教育, 2010(1): 62-67.

Research on Construction Strategies for Online Learning Resources Based on Decision Tree

MAO Bu¹, XIE Wen²

(1. Sichuan Zigong Broadcasting and TV University, Zigong 643000, China;

2. College of Computer Science, Sichuan University, Chengdu 610000, China)

Abstract: Resources construction personnel of modern distance education often base their analysis on the characteristics of different types of resources so as to promote a student-oriented online learning and browsing. A DM-based approach to the analysis of resource characteristics can not only increase learning hits but help formulate related strategies on resources construction. This essay introduces basic theories of decision tree calculation and analyzes various specific features of online learning resources. Examples of calculation are also employed in this essay to demonstrate that resource characteristics are conducive to the construction of decision tree hitting module. The ending part of the essay features DM that promotes hitting rates by way of decision tree in Analysis Manager. Results have justified the feasibility of this application.

Key words: resource characteristics; hitting rate promotion; data mining; decision tree calculation