改进的多叉决策树在大学专业选择系统的应用①

郑　陶　柴瑞敏 (辽宁工程技术大学 电子与信息工程学院 辽宁 葫芦岛 125105)

摘　要： 考生在高考后对大学专业的选择是其职业生涯的起点。大学专业选择系统可以测定考生的人格特性，根据以往大量数据中提取的模型帮助他们选择适合自身的专业。借助平面向量参考系改进的决策树算法，可以挖掘出训练集中隐含的规则，构建高精度的评判模型，从而避免了以往利用专家系统的方法构建时过分依赖先验知识所产生的误差，更客观的为考生匹配适合的专业。实验证明，算法的挖掘精度和效率都达到了令人满意的效果。

关键字： 多叉决策树；着色叶子节点；平面向量参考系；高考专业选择

Application of Improved Multi-Forks Decision Tree to Major Selection System

ZHENG Tao, CHAI Rui-Min

(School Electronic and Information Enginee, Liaoning Technical University, Huludao 125105, China)

Abstract: It's a starting point of career for the students who have passed the college entrance examination to choose a major. Major Selection System can check the real personality trait and then help select the matching major based on the data obtained. This study improves data mining algorithms by putting the plane vector reference system into it and digs out the implied relevance in the sample set to build the measure system. In this way, the inaccuracy of previous expert systems can almost be avoided so that the major can match the students more objectively. Experiments show that the precision and efficiency of the algorithms are almost satisfactory.

Keywords: multi-forks decision tree; staining leaves; plane vector reference system; major selection

1　引言

高考后的考生对大学专业的选择对其发展至关重要，根据考生自身特性选择适合他们的专业是促进其日后职业发展的前提。目前，我国对大学专业选择领域的研究和应用还处于起步阶段，没有形成系统化、规范化的测评体制。调查发现，有 70%以上的在校大学生当年在“懵懂”的状态下填报了大学专业。学生在专业选择时带有很大的盲目性。而在美国，100%的中学生至少接受过一次由教育考试服务社(ETS)开发的 SIGI 和美国大学考试中心(ACT)开发的 Discover 这样的职业测评，这将有效的帮助学生了解自己的个性特质和职业倾向，从而在选择专业的时候目标明确。在日本，仅仅由专业人才服务公司 RECRUIT 开发的职业生涯评估系统一年就接待了 20 万以上的中学生的测试[1]。

国内外现有的测评系统都依赖专家的先验知识，致使系统存在着一定程度的主观随意性。本文构建的专业选择辅助系统是以数据为基础，把平面向量参考系引入传统决策树，把类别的有序性考虑进来，变类别的离散值为连续域，进行更精细的类别区分，从而使基于数据的思想得以实现，因而避免了使用经验数据所产生的误差。

2　引入力学中有关向量的概念

决策树是用二叉树形图来表示处理逻辑的一种工具。可以直观、清晰地表达加工的逻辑要求。但很少有学者用它去实现与人有关的决策体系，主要有以下两个问题：1)对人的测评会出现很多的模糊性数据。2)训练样本集中会出现很多属性完全相同但结论不相

① 收稿时间:2009-06-07

同的样本，不能随意的把某些样本作为噪声去掉。样本自身所包含的客观的矛盾性致使“开发基于数据的算法”的设想一度陷入困境。

本文所构建的专业选择系统就可以很好的解决以上的问题，把力学中有关向量的概念引入决策树中，为更进一步的类别划分建立参考系，实现对样本的综合处理以及对决策的模糊逼近。力学有关向量的概念如图 1 所示。

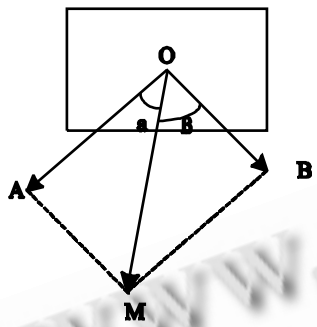

图 1　力学中向量的应用

点 O 为物体的重心，$\vec{OA},\vec{OB}$分别表示两个方向上的分力，$\vec{OM}$ 的方向是合力的方向，是合力的大小。$\vec{OA},\vec{OB},\vec{OM}$满足平行四边形法则[2]。α，β 分别是合力与两个分力的夹角，若 $|\vec{OA}|>|\vec{OB}|$ 则 $\alpha<\beta$，说明 $\vec{OM}$ 更靠近模比较大的向量。

3　大学专业选择系统的总体思路

学生进入系统进行答题，测得考生的兴趣类型 C_1、性格类型 C_2、(测评量表分别是基于霍兰德职业兴趣理论和爱森克人格理论)[3]。性别 C_3、学习成绩 C_4、表达能力 C_5、这三个属性由考生根据自己实际情况填选。计算机将五个属性预处理后作为算法的输入，而后通过计算生成该生在 200 个专业上适合度的排序。

第一步，条件属性的确定

确定训练集的条件属性将直接决定系统的客观性。但是，随着属性个数的增加，我们的计算量也会加大，给我们的测评工作带来一定的困难。因此要适当的选取属性，使属性既能反映出测评的特性，又便于我们的计算处理[4]。属性的选择可以通过计算可能的若干指标的信息增益后进行筛选，本文最终确定了五个条件属性。

第二步，获得样本数据

集中大三大四学生在计算机上进行答题，得到兴趣(R、I、A、S、E、C)、性格(内向 1、一般 2、外向 3)等属性，而后由若干老师及其本人判定他在此专业的适合度(很适合 1、较适合 2、一般 3、较不适合 4、很不适合 5)最后根据权重汇总成综合意见(即类别)。以计算机专业为例，如表 1 所示：

表 1　训练样本集

编号	属性					类别
	兴趣 C_1	性格 C_2	性别 C_3	学习能力 C_4	表达能力 C_5	
1	I	1	1	2	2	2
2	A	2	1	2	1	4
3	I	1	1	2	2	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
…	I	1	1	2	2	4

第三步：计算信息增益，确定多叉决策树的层次结构

1) 考虑样本数为 s 的训练集 S。假定类标号属性(target attribute)具有 m 个不同值,即有 m 个类别，分别用 $C_i(i=1,\cdots,m)$定义.设 s_i 是 S 中属于类 C_i 的样本数，对一个给定的样本分类所需的期望信息熵由下式给出

$$I(T)=-\sum_{i=1}^{m}P_i\log_2(P_i) \tag{1}$$

其中 P_i 是任意样本属于 C_i 的概率,用 s_i/s 估计。

2) 分别计算每一个属性划分所形成的子集的信息熵。设属性 A 具有 v 个不同值$\{a_1, a_2,\cdots,a_v\}$，那么属性 A 可将 S 划分为 v 个子集$\{S_1,S_2,\cdots,S_v\}$，其中 S_j 是在属性 A 上取值为 a_j 的 S 的子集.设 s_j 是子集 S_j 的样本数，由 A 划分成子集的信息熵的计算公式为

$$I_T(A)=\sum_{j=1}^{v}\frac{S_i}{S}I(S_j) \tag{2}$$

3) 计算各属性划分样本的信息增益。对于属性 A，信息增益的计算公式为

$$Gain(A)=I(T)-I_T(A) \tag{3}$$

属性的信息增益计算结果为：兴趣>性格>性别>

学习能力>表达能力。根据信息增益由强到弱，自上而下为多叉决策树节点分配属性

第四步，为叶子节点着色

若在传统的决策树中，一种叶子节点代表一种类别，五种类别对应五种叶子节点，但在本系统中需要根据叶子节点对 200 个专业进行全排序，显然用传统的方法无法实现。于是本文把传统决策树做了进一步改进，把类别存在序关系的特性考虑进来，把五个类别放在连续域中，为了使问题的表述更形象，把叶子节点用由深到浅不同的灰度值进行“着色”(当然在计算机种存储的是各灰度值，看不到颜色)，且相邻两灰度值的差相等。通过算法得出的最终叶子节点(类别)不止五类，这样有助于对类别进行更进一步的划分。如图 2 所示。它们的灰度值可能是 A 到 E 中的任意值。

图 2　叶子节点的灰度表示

在对人的测评中会出现许多条件属性完全相同而决策属性不同的记录。由于是对人的测评，我们不能草率的把哪个样本作为噪声去掉，而应该把它作为一个决策规则的影响因素去考虑问题。把上述样本的类别值和对应的数量作为输入引入向量参考系进行计算得到此类样本叶子节点的灰度值。如图 3 所示。

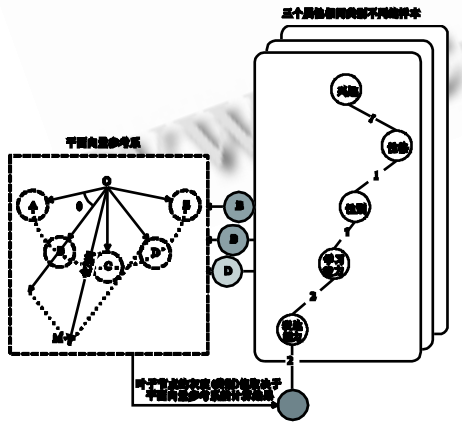

图 3　通过向量计算叶子节点的灰度值

向量是既有大小又有方向的量，本研究利用 5 个由同一点出发，模为 e(表示单位向量)的向量作为参考系，向量间的夹角相等，设为 θ。为了更清晰的描述问题，此处令 $\angle AOE$ 为略小于 180 度的钝角(等于 180 度时，两侧的向量的和可能出现模为零的情况)在这个扇形参考系中只有 θ、e 两个参考量，问题的描述与 θ、e 的取值无关，所以这种描述是比较客观的。

如图所示 C_1=I,C_2=1,C_3=1,C_4=2,C_5=2 的记录有 3 条，其中两条记录的类别为 B，另外的一条为 D，我们分别用 $\vec{OB}$ 和 $\vec{OD}$ 表示，根据平行四边形法则，$\vec{OM}$ 为向量的和，此向量必然夹在两决策向量之间，并且与其中一个决策向量的夹角小于 $\theta/2$，此时容易发现 $\vec{OM}$ 更靠近 $\vec{OC}$，我们称 $\vec{OM}$ 趋近于 $\vec{OC}$，趋近度为

$$\mu_{OM \to OC} = \angle MOC / 0.5\theta \tag{4}$$

而后产生的规则就是：IF C_1=I AND C_2=1 AND C_3=1 AND C_4=2 AND C_5=2

THEN D=C, $\mu = \mu_{OM \to OC}$

由此可见，这条规则结论的模糊性可以通过两个变量精确表示，这种双变量规则可以解决两个问题：1)不需要回避样本集中条件都相同但结论不同的样本。2)如果通过传统的决策树在二百个专业上建立 200 棵树，而最终只有五种适合度，则无法对二百个专业的适合度高低进行全排序。但由向量法构建的模糊决策树就可以更精细的区分适合度，从而解决了这个难题。另外，叶子节点的灰度值可以通过 $\angle AOM$ 与 $\angle AOE$ 的比值计算得到。系统的结构图如图 4 所示：

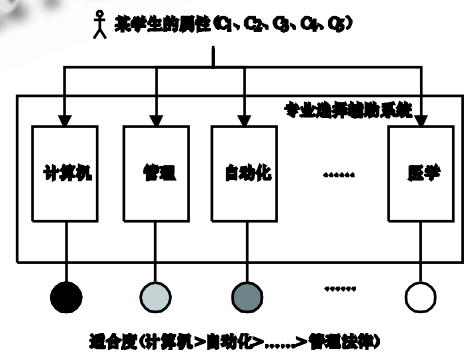

图 4　系统整体功能示意图

另外，还可能出现这样一个问题，几个不同专业上的向量和的方向相同。此时可以根据各条规则的支持度在这个小区域内进行排序。限于篇幅不予详述。

综上所述，传统的决策树算法无法解决全排序问

题，而改进的决策树算法则可以通过对给定样本计算叶子节点的灰度值进行更精确的区分，灰度值相同专业用支持度加以区分，最终得到二百个专业适合度的全排序。

4 基于置信度的剪枝

剪枝过程是让决策树更一般化，以适应新样本的过程，常见的剪枝策略有预剪枝(pre-pruning)技术和后剪枝(post-pruning)技术。预剪枝技术主要是通过某些规则限制决策树的充分生长，后剪枝技术则是待决策树充分生长完毕后再进行剪枝，后剪枝技术应用较多。

不同应用对决策树的要求有不同的侧重，有的注重精确性，有的注重效率，笔者鉴于本研究的特殊性，设计了一种基于平面向量置信度的新剪枝算法。

在对人的预测和测评的情况比较特殊，所以不能使用传统的剪枝算法处理，这里应采用预处理过的数据，注意，预处理过的数据涉及到一个置信度的概念，只有把它考虑进去才能保证预测的精确度。首先我们给出置信度的定义。

定义 1. 在平面向量参考系中，合向量的模与合成此向量的分向量个数的比值称为此合向量的置信度。

显而易见，分向量越分散，置信度就越低，分向量越靠近，置信度越高。这间接反映了一条规则在预处理后的可信程度。通过置信度阈值的设定来对规则进行划分，小于阈值的规则用基于错误的 EBP 作为剪枝算法。可以在保证误报率和错报率的前提下提高效率和系统的预测能力。实验数据来自智友人才测评公司，选取 70000 条记录，其中 50000 条用于训练，20000 条用于测试。机器配置为：AMD Athlon(tm) X2 DualCore Ql-62,内存为 1G，编程工具为 VC#。

表 2 基于置信度的新剪枝算法与 EBP 的性能比较

	误报率/%	错报率/%	时间/s
EBP	7.32	4.97	5.32
新剪枝算法	7.79	4.23	2.93

5 结论与展望

人才测评可以帮助人们更好地了解自己的特点，并可以根据自己的优势进行工作的选择[5]。基于平面向量的模糊多叉决策树是由专业选择这个特定问题而提出的，它适用于决策属性值存在序关系的决策中，它也可以解决另一类问题，假设有两套方案或策略(即两个记录)它们的条件不同，但产生的决策相同，那么就可以根据以上的方法找出哪个方案或策略更好，当然这仍需要以往的大量的数据作为样本集。

另外，此算法也可以向决策属性无序关系的决策问题眼神，对于常见的 Y-N 型和 A-B-C 型，我们可以通过类似平面直角坐标系和空间直角坐标系中通过向量去建立参考系，但对于 n 个决策值(n≥4)的情况，还不能确定是否可以建立有效地向量参考系，有待更多的专家进行更深入的研究。

参考文献

1 崔丽娟.性格气质与大学专业选择——高考生必读.北京:人民军医出版社, 2006.1－2.
2 刘华.平面向量的应用案例与反思.中学数学月刊, 2006,8:25－26.
3 童腮军.高校学生专业选择行为研究[硕士学位论文].南昌:江西师范大学, 2003.
4 徐钰.高校人才测评系统的研究与实现[硕士学位论文].天津:河北工业大学, 2006.
5 王垒.实用人事测量.北京:经济科学出版社, 2002.41－43.