doi：10.3969/j.issn. 1674-2346.2022.01.019

决策树在大学外语等级考试成绩分析中的应用

王渊志

（宁波市教育考试院，浙江 宁波 315000）

摘　要：随着数字化在教育考试领域的不断推进，考试成绩数据分析已成为考试管理领域的一大课题。本文以宁波市大学外语等级考试成绩数据为基础，利用决策树模型挖掘出影响考试成绩的关键因素，为高校进一步提高大学外语教学水平提供科学有效的参考。

关键词：大学外语等级考试；成绩；决策树

中图分类号：G424.74　　**文献标识码：**A　　**文章编号：**1674-2346（2022）01-0096-05

1　大学外语等级考试成绩分析的意义

大学外语等级考试是教育部考试中心负责实施的全国性的教学考试，目的在于对高校学生实际外语应用能力进行客观、准确的测量，这项考试因为题目设计科学合理、考务流程规范，在社会上认可度很高，很多用人单位将该考试成绩作为招录工作人员的重要参考依据之一。对学校而言，考试成绩既直接体现了学生的学习效果，又能评价教师日常教学水平。因此各个高校对于大学外语等级考试的成绩十分关注，如何提高大学外语等级考试成绩，推进外语教学，从而提升学生的实际外语水平成为众多高校追求的目标。

目前，学校使用教育部考试中心研发的大学外语等级考试考务管理系统，主要包括报名信息录入、照片采集、试场编排、准考证打印、缺考违纪数据录入、成绩导入等功能，对系统数据往往停留在查询、增删等基础的应用方面，个别学校通过导出到 EXCEL 功能，进行简单的成绩统计，得出的结果往往比较单一，数据的价值没有完全被开发。如果把数据挖掘技术应用于成绩分析，可以帮助学校深入了解学生各项成绩之间的关联，找出影响成绩的各项因素，对于提高教学质量，提升人才培养水平大有帮助。

2　决策树分类方法介绍

本文采用分类方法中的代表——决策树算法，尝试对影响考试成绩几个关键要素进行分析。分类方法的定义如下：找出同类事物共同性质的特征性知识和不同事物之间的差异性特征知识。基于决策树的分类算法是一种以实例为基础的归纳学习算法，即从一系列无序无规则的元组中推导出分类规则，以树的形式呈现。决策树采用自顶至下的贪婪算法，在其内部结点选择分类效果最优的属性向下分支，直到这棵树能明确地分类训练样本，或所有属性都被使用。决策树中比较著名的是 C4.5 算法。通过这种算法得出的结论很容易把逻辑上的关系以一种非常直观的方法清晰地表达出来。对于判断因素少、逻辑组合

收稿日期：2021 － 09 － 12

作者简介：王渊志，男，助理研究员。研究方向：考试管理，考试信息化

较为简单的项目尤为适合。决策树尤其擅长处理非数值型数据，数据预处理工作量相对较少。

采用决策树技术进行分类包含两个步骤：(1) 使用训练样本构造并优化一棵决策树，搭建模型。从实际应用上看，这个过程就是从样本中获取知识，进行机器学习的过程。(2) 依靠构造完成的决策树对输入数据进行分类。从根结点依次判断输入记录的属性值，直至某个叶结点停止，从而找到该记录对应的类。其中建树与剪枝环节是建立决策树模型的关键步骤。

3 决策树在大学英语四级成绩分析中的应用

大学外语等级考试的开考科目包括英语四级、英语六级、日语四级、日语六级、法语四级等。目前全国每次均有近 1000 万人参加考试，其中宁波市报考人数达到 10 万人，在浙江省内居首位。报考人数最多科目为英语四级，本文主要以 2019 年下半年宁波市英语四级考试成绩作为分析样本。

该样本包括考试成绩记录 41222 条，来自宁波 16 所高校。按学校层次分为重点本科、普通本科、高职专科与成教四大类，按专业类别分为理工类、医药类、人文类、经管类、艺术体育类五大类。

本文借助 Visual Studio SSDT+SQL Server 工具，采用决策树算法，对报考数据中的学校类别、考生专业、入学年级、性别、考生学历等项目进行挖掘分析，找出关联特征，为高校改进教学安排提供参考。主要包括以下几个步骤：(1) 对报考数据进行预数理，即去除无关字段，离散化保留字段；(2) 将报考数据分类为训练集与测试集，并通过 SSDT 中的决策树算法建立挖掘模型；(3) 模型准确率验证。

3.1 数据预处理

为了获得数据挖掘所需的净化数据，必须对海量数据进行预处理，包括数据集成、数据选择和数据清理，本文使用 SQLSEVER2014 软件实现。

(1) 去除不相关字段。由于数据直接从系统中导出，数据整齐，数据噪声情况不存在。但数据集合中共有 35 个字段，这些字段给挖掘提供了海量的信息，但是如果使用过多的字段作为输入值，反而会使挖掘结果可读性下降，影响到最终结果的获取和分析，有必要去掉数据集中与数据挖掘关系不大的字段，如班级、班级名称、校区、编排座位等信息，保留了其中专业名称、年级、性别、准考证号（标志数据的主键）、总分、缺考、报名学校、学历名称等字段供挖掘使用。

(2) 所属学校归类。16 所高校按照学校层次可分为重点本科、普通本科、高职专科与成教四大类，将报名学校列替换为学校类别。

(3) 专业归类。由于考生就读专业较多，不利于数据挖掘，根据专业性质归为理工类、医药类、人文类、经管类、艺体类 5 种。

(4) 总分离散化处理。由于总分为连续数值，不适合决策树算法。新增“是否通过”与“是否优秀”字段。总分大于等于 425 分，“是否通过”为真。总分大于等于 550，“是否优秀”为真。

3.2 创建挖掘项目

使用 VS2017 新建 Analysis Service 多维数据和数据挖掘项目，在关联数据源后，选取 70%的记录作为训练集，指定挖掘结构为决策树，采用“准考证”为主键，选择“学校类别”、“专业类别”、“入学年级”、“性别”、“考生学历”作为输入列，选择“是否缺考”、“是否通过”、“是否优秀”作为可预测列，生成通过率、优秀率与实考率 3 个挖掘模型，从而发现通过率、优秀率、实考率与输入字段之间的规律。

3.3 验证模型准确性

为了保证模型具有较好的精确度和健壮性，将剩余的 30%的数据视为测试集，用来测试和验证模型是否准确。经验证，通过率、优秀率、实考率的测试结果预测概率超过 80%，说明模型结果真实可靠。

4 决策树分析

由于生成的决策树模型对应的规则较多，且树型较大。本文以通过率、优秀率、实考率为例，从模型中抽取出一些强关联型规则加以分析。

4.1 通过率决策树分析

部分强关联规则：

IF 年级=“19 级” then 通过率在 65%左右

IF 年级=“19 级”and 学校类别=“重点本科”then 通过率接近 90%

IF 年级=“19 级”and 学校类别=“重点本科” and 专业类别=“经管类”then 通过率超过 97%

IF 年级=“19 级”and 学校类别<>“重点本科”then 通过率在 60%以上

IF 年级 ◇“19 级” then 通过率不到 25%

IF 年级 ◇“19 级”，学校类别<>“重点本科” then 通过率仅有 12%

可以发现，决定大学英语四级能否通过的首要因素是考生的年级。根据现行政策，考生第一学年允许报考英语四级，因此多数考生都不会放弃第一学年考试的机会，而且由于刚入学，学习热情较高。反观 19 级以前的考生，这些考生大多是重考生，未能在首次考试中一次性通过，一般而言英语基础不够扎实，而英语学科需要长期积累，基础不实的考生往往再次考试通过率也比较低。

对于 19 级考生，学校类别决定了通过率。重点本科的通过率明显高于其他类别的考生。显然，重点本科的生源素质确实是高于其他层次考生，生源素质直接影响了四级的通过率，这与日常经验得出的判断是一致的。对于普通学校考生，性别因素对通过率起了关键的作用，女生的通过率比男生高出 15 个百分点。对于重点本科学生，就读于经管与人文专业的考生的通过率要高于其他学科。

4.2　优秀率决策树分析

部分强关联规则：

IF 学校类别=“重点本科”then 优秀率在 30%以上

IF 学校类别=“重点本科”专业类别=“人文”then 优秀率在 40%以上

IF 学校类别=“重点本科”专业类别=“经管”then 优秀率在 45%以上

决定大学英语四级优秀率的首要因素是学校的类别。重点本科高校聚集了大批最优秀的考生，这类考生参加四级考试优秀率自然要远远高于其他类别学校的考生。其中重点本科高校的考生，修读人文与经管类学科的优秀率要高于其他学科，艺体类的考生优秀率最低。高职专科考生，受制于生源素质，优秀率很低，只有极个别的人文学科考生达到了优秀，而非人文专业的无一优秀。

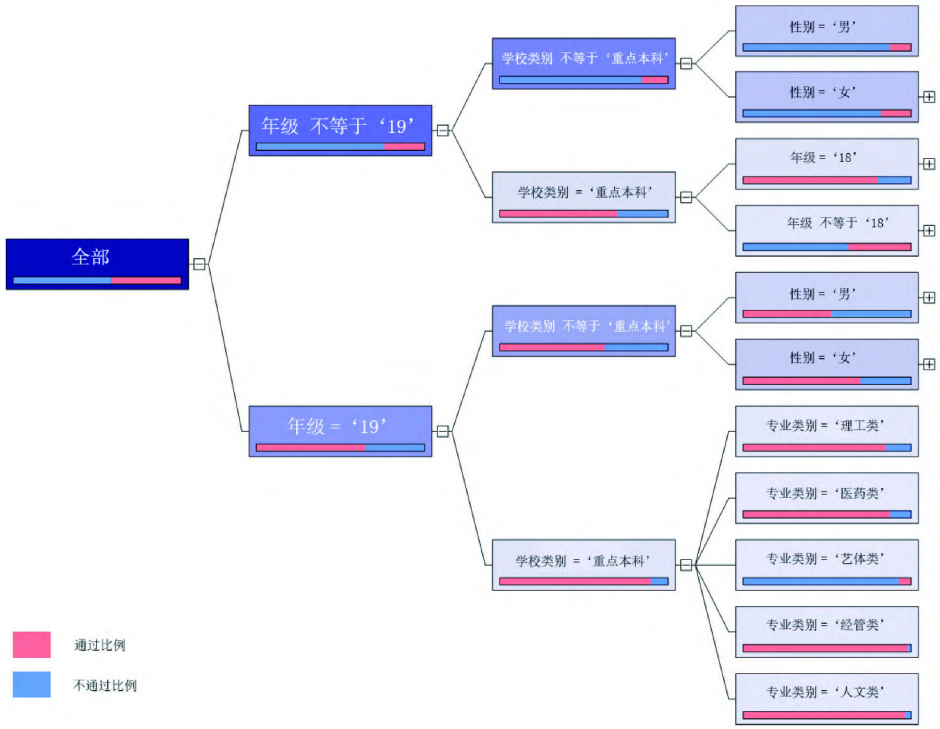

图 1　通过率决策树模型

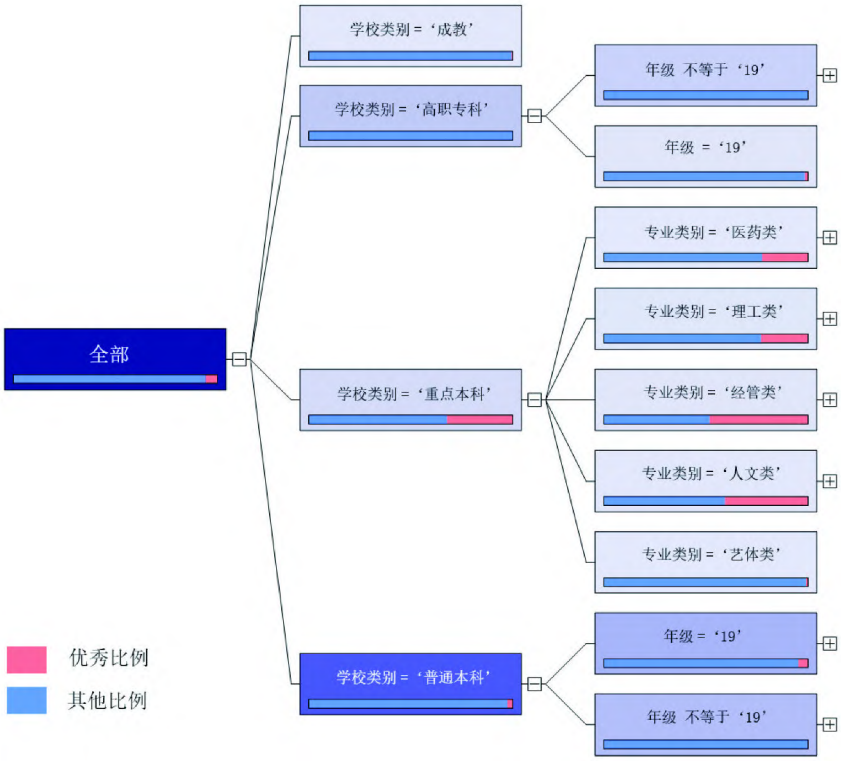

图 2　优秀率决策树模型

4.3　实考率决策树分析

部分强关联规则：

IF 年级=“19”then 实考率在 95%以上

IF 年级=“19” 学校类别 ◇“成教”then 实考率接近 97%

IF 年级 ◇“19”then 实考率在 80%以上

IF 年级＝“18”then 实考率在 85%以上

决定实考率高低的首要因素还是年级，这与通过率的首要因素保持一致。19 级的考生，第一次参加考试，往往比较重视这项考试。而 19 级前的考生，往往是多次参加考试，其对考试的重视程度不如 19 级的考生，因此缺考人数明显增加。对于 19 级的考生而言，成教学生与其他全日制学生产生了明显的差异。成教学生英语基础较差，考生自信心不足，无法认真对待这项考试，因此有一半考生放弃了考试。全日制学生首次报名缺考比较少。对于 19 级的全日制学校的学生而言，性别依然是决定到实考率的关键因素，女生的实考率要比男生高出 3 个百分点。

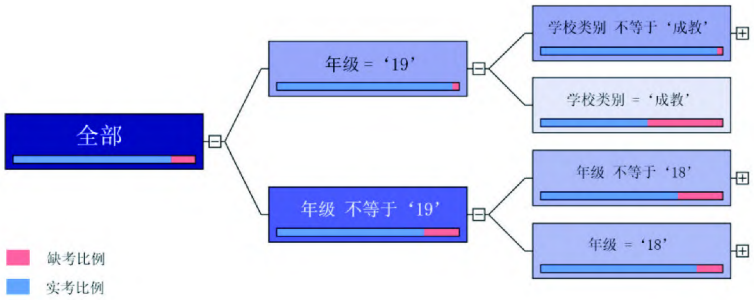

图 3　实考率决策树模型

5　结语

影响通过率首要因素是考生年级，第一学年的考生通过率明显高于其他年级。影响优秀率首要因素是考生学校类别，重点高校的考生优秀率明显高于其他类别。影响实考率首要因素是考生年级，第一学年考生的实考率明显高于其他年级。

实验表明，3 个模型的预测准确率超过 80 %，分析结果也符合现实认知。当然该模型还存在不足，比如对成绩库中相关字段选取过程人为因素较大，未采用更为先进的算法等，都值得进一步研究。

参考文献

[1]袁乐泉，朱亚辉.基于随机森林的大学英语四级通过率预测模型[J].电子测试，2021(4)：54-55.
[2]叶泽俊.基于数据挖掘的大学英语四级通过率预测建模研究[J].长春师范大学学报，2019(12)：8.
[3]栾红波.数据挖掘在大学英语教学和测评中的研究与应用[D].北京：北京邮电大学，2017：22-25.

The Application of Decision Tree in the Analysis of College Foreign Language Test Scores

WANG Yuan-zhi
(Ningbo Education Examinations Authority, Ningbo, Zhejiang 315000, China)

Abstract: With the continuous advancement of digitalization in the field of education examinations, the analysis of examination result data has become a major topic in the field of examination management. Based on the scores data of Ningbo's College Foreign Language Test, this paper uses the Decision Tree model to dig out the key factors that affect test scores so as to provide scientific and effective reference for colleges and universities to further improve college foreign language teaching.

Key words: College Foreign Language Test; score; Decision Tree

(上接第 95 页)

Preliminary Exploration of the Vocal Music Teaching in Secondary Vocational School for Kindergarten Teachers Based on Learning and Application Docking

JIANG Yu
(Ningbo Vocational and Adult Education College, Ningbo, Zhejiang 315041, China)

Abstract: Through the investigation and research of teachers, students, school leaders and kindergarten personnel related to vocal music teaching, this paper analyzes the scene characteristics of vocal music teaching in preschool education: enlightenment, entertainment and activity. Based on the principle of the music scene, the music teaching scene of "behavior, activity and docking" and the teaching realization of the all-element integration of "learning and application docking" are presented.

Key words: all-element integration; music scene; learning and application docking; secondary vocational school for kindergarten teachers