文章编号：1004-4353(2020)04-0366-05

基于机器学习的高考信息与大学程序设计课程成绩相关性分析研究

金城， 崔荣一， 赵亚慧*
（延边大学 工学院，吉林 延吉 133002）

摘要：为研究学生高考信息与计算机程序设计课程(C 语言)成绩的相关性，提出了一种基于随机森林算法的相关性预测与分析模型. 首先，对 2014—2016 年延边大学计算机科学与技术专业的学生相关数据进行了清洗和筛选，并将 C 语言考试成绩分成 5 类；其次，将学生的高考信息作为特征训练随机森林分类模型；最后，使用 LIME 解释性模型对影响随机森林的主要特征进行了相关性分析. 实验结果表明，影响 C 语言成绩的主要特征为生源、总成绩、民族、数学和语文. 该研究结果可有效识别不同学生学习成绩的主要相关因素，为教师针对不同学生群体设计合理教学模式提供参考依据.

关键词：大学成绩；影响因素；决策树；随机森林
中图分类号：TP399 **文献标识码**：A

Research on correlation analysis between college entrance examination information and college program design course scores based on machine learning

JIN Cheng， CUI Rongyi， ZHAO Yahui*
（*College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: In order to study the correlation between college entrance examination information and computer programming course (C language) scores, a correlation prediction and analysis model based on stochastic forest algorithm is proposed. Firstly, the data related to computer science and Technology majors in Yanbian University from 2014 to 2016 were cleaned and screened, and the C language scores were divided into 5 categories of different levels. Secondly, the college entrance examination information of students is used as a random forest classification model of feature training. Finally, LIME explanatory model is used to analyze the correlation of the most influential characteristics of random forest. The experimental results show that the five characteristics of student origin, total score, nationality, mathematics and Chinese have the greatest influence on the C language score. The results of this study can effectively identify the main factors related to different students academic performance and provide reference for teachers to design reasonable teaching modes for different groups of students.

Keywords: college grades; influential factor; decision tree; random forest

收稿日期：2020-07-30 **基金项目**：吉林省高等教育学会高教科研课题(JGJX2018D347)
* **通信作者**：赵亚慧(1974—)，女，教授，研究方向为自然语言文本处理、教育信息处理.

近年来，一些研究者针对学生的高考成绩与学生进入大学后的学习成绩（尤其是学生入学第一学期的学习成绩）之间的关系进行了研究. 例如：陈小杭[1]对学生高考的数学成绩与学生入学后的大学数学专业课成绩进行了相关性分析，结果表明学生的高考数学成绩与学生入学后的大学数学专业课成绩无显著相关性. 石铁玉等[2]研究表明，学生的高考成绩与学生入学后的考试成绩呈弱相关性. 杜晓燕等[3]对学生的高考成绩和大一单科成绩的关联性进行研究表明，文科类课程的成绩与高考成绩的关联性较大，而理科类的课程的成绩与高考成绩的关联性较弱. 上述文献的研究方法主要是基于统计的相关性分析方法进行的，但该方法对于没有明显统计学规律的多元复杂数据其效果并不理想. 文献[4]研究表明，随机森林算法在处理数据复杂、维度较高的分类任务时可获得较高的准确度. 因此，本文采用基于随机森林算法研究学生的高考信息与大学一年级的程序设计课程成绩之间的相关性，以为教师在程序设计课程教学中设计出更有针对性和有效的模式提供参考.

1　相关技术

机器学习方法可以从一类数据中自动学习规律，特别是对特征种类多、特征数目庞大的复杂数据进行预测时，其效果显著优于基于统计的方法，因此该方法被广泛地应用于回归、拟合和大数据分析等方面. 目前，常使用的机器学习方法包括监督学习、无监督学习和强化学习[5]. 其中：监督学习方法可以从大量没有显著统计规律的数据中学习到有效的模型，因此常用于解决回归、分类的问题；无监督学习可以在较为规律的统计数据中发现潜在的结构，因此常被用于聚类和降维；强化学习则可在基于环境的动态互动中取得最大化的预期利益，因此常被用于控制系统的设计中.

决策树是一种被广泛应用于金融、保险、医疗等领域的树状分类器，但决策树算法在数据复杂时准确率较低. 为此，L.Breiman 结合 Bagging 集成思想[6]与随机子空间方法[7]提出了随机森林算法[8]，该算法具有解释性好、结构简单、计算开销小等优点[9]. 随机森林算法的具体步骤如下：

输入：样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，决策树迭代次数 T

输出：随机森林 $f(x)$

1) for $t=1$ to T：

a) 对训练集进行第 t 次随机采样，共采集 m 次，由此得到包含 m 个样本的采样集 D_t.

b) 用采样集 D_t 训练第 t 个决策树模型 $G_t(x)$. 训练决策树模型的节点时，首先在所有样本特征中随机选择一部分样本特征，然后在选出的样本特征中选取一个最优的特征来划分决策树的左右子树.

2) 在形成的 T 个决策树中，利用投票表决结果. 当结果只有一个类时，将票数最多的类别作为最终类别；当结果包含多个类时，将目标类别作为最终类别.

2　基于分类的成绩影响因素分析

2.1　数据收集

本研究以大学一年级的 C 语言程序设计课程为例，收集的数据为 2014—2016 年延边大学计算机科学与技术专业 3 个年级的学生个人信息. 信息包括：高考成绩、学生生源、民族、考生类别和入学第 1 年的 C 语言期末考试成绩. 3 个年级的学生人数分别为 115 人、157 人和 145 人.

3 个年级学生的高考特征属性及其分布如表 1 所示. 由表 1 可知：在性别方面，男生略高于女生；在民族结构方面，考生以朝鲜族和汉族学生为主，其中朝鲜族学生占总考生的 34.2%；在考生类别方面，城市考生占总考生的 58.7%；在生源方面，考生主要来自吉林省，占总考生的 49.9%.

表 1　考生特征属性分布

特征属性	特征属性类别	比例/%
性别	男	57.8
	女	42.2
民族	朝鲜族	34.2
	汉族	57.2
	其他民族	8.6
考生类别	城市	58.7
	农村	41.3
生源	吉林省	49.9
	其他省份	50.1

2.2 特征选择

由于本数据集中的学生主要为汉族与朝鲜族的考生(占总人数的91.7%),且朝鲜族和非朝鲜族考生的录取政策不同(非朝鲜族的其他少数民族和汉族采用同一录取标准),因此本文将民族特征分为朝鲜族和非朝鲜族进行分析.同时去除不使用全国I卷和全国II卷省份的学生信息.将考生进入大学后的C语言成绩按分数段分为5类:100～90(第1类),89～80(第2类),79～70(第3类),69～60(第4类),59～0(第5类).各年级C语言成绩的分布情况如图1所示.

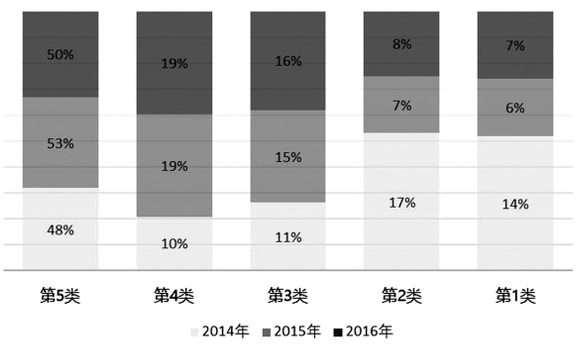

图1 各年级学生的C语言成绩分布

2.3 基于随机森林算法的C语言课程成绩预测模型

构建C语言成绩预测模型的方法如下:

1)将处理后的数据按9∶1分为训练集和测试集;

2)利用各年级的训练集数据训练随机森林模型,并通过调整随机森林的参数得到最优的预测模型;

3)利用Bootstrap方法从训练集中随机抽取多个训练样本子集,并对每个子集分别进行随机森林建模;

4)利用测试集对各随机森林进行测试,并综合多棵随机森林的测试结果以通过投票的方式得出最终的C语言课程成绩预测模型;

5)使用可解释性模型LIME(local interpretable model-agnostic explanations)计算对随机森林模型贡献度最大的特征.

上述步骤中利用LIME计算调整贡献度的方法为:① 在原始样本中随机替换掉若干特征,以此得到含有噪声的数据z'.② 计算随机森林模型对z'预测的值.③ 求出原样本与生成样本之间的距离,并将其作为权重.④ 利用生成样本、预测值和权重训练一个简单的线性模型g.⑤ 按式(1)计算模型g拟合样本的结果与随机森林模型预测样本的结果之间的差值,然后根据差值对随机森林模型进行解释(差值越小贡献度越大).

$$L(f,g,w^y)=\sum_{i=1}^{N}w^y(z_i)(f(z_i)-g(z_i'))^2. \tag{1}$$

其中,f为原模型,w为权重,z为原样本,z'为加入噪声后的样本.基于随机森林算法构建C语言成绩预测模型的流程如图2所示.

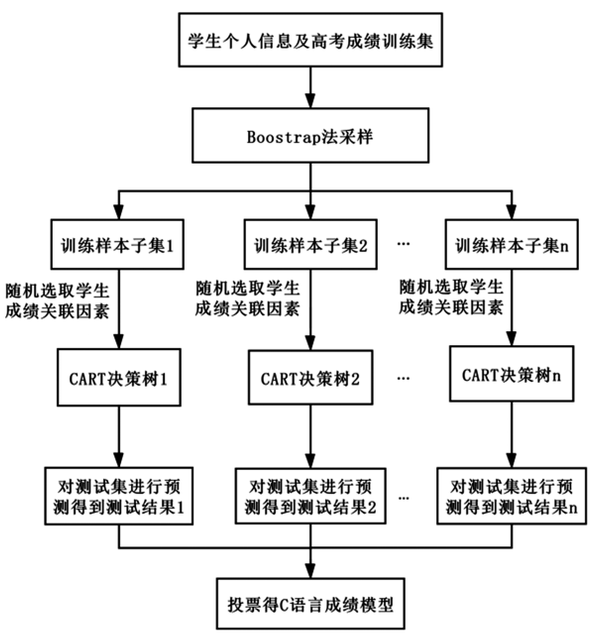

图2 基于随机森林算法的C语言成绩预测模型

3 实验结果与分析

利用随机森林算法对数据进行训练,结果如表2所示.

表2 训练集和测试集的准确度

年级	训练集准确度	测试集准确度
2014	0.921	0.600
2015	0.830	0.538
2016	0.884	0.615

为获得最佳的分类效果,本文利用实验对模型的参数进行了选定,结果如表3所示.

利用 LIME 模型计算每个特征对随机森林模型的贡献度，结果如表 4 所示.

根据表 4 中的贡献度结果，本文将各年级中排序为前 2 名的特征作为最大的相关性特征. 这些特征包括生源、民族、总成绩、数学和语文 5 个特征. 在所有特征中任取 5 种特征，并按不重复原则组合方案进行排列组合，共得到 126 种组合方式. 为验证本文选择的特征方案为最佳方案，对 126 种不同的特征组合使用随机森林进行了训练和测试，其中部分特征组合方案测试集的平均准确率的结果如图 3 所示.

表 3　最优模型参数

参数	参数值
ccp_alpha	0.0
class_weight	None
criterion	‘gini’
max_depth	5
max_features	auto
max_leaf_nodes	None
min_impurity_decrease	0.0
min_impurity_split	None
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0.0

表 4　各特征对模型的贡献度

年级	贡献度	特征	年级	贡献度	特征	年级	贡献度	特征
	0.160 0±0.098 0	生源		0.092 3±0.061 5	总成绩		0.169 2±0.115 1	总成绩
	0.100 0±0.126 5	民族		0.076 9±0.000 0	数学		0.092 3±0.115 1	语文
	0.080 0±0.080 0	综合		0.061 5±0.061 5	综合		0.061 5±0.150 7	生源
	0±0.000 0	总成绩		0.061 5±0.061 5	语文		0.061 5±0.061 5	性别
2014	0±0.000 0	英语	2015	0.061 5±0.061 5	生源	2016	0.030 8±0.075 4	数学
	0±0.000 0	数学		0.046 2±0.075 4	民族		0±0.000 0	考生类别
	0±0.000 0	语文		0.030 8±0.075 4	英语		−0.015 4±0.061 5	英语
	0±0.000 0	考生类别		0.030 0±0.075 4	性别		−0.015 4±0.061 5	民族
	0±0.000 0	性别		0±0.000 0	考生类别		−0.061 5±0.115 1	综合

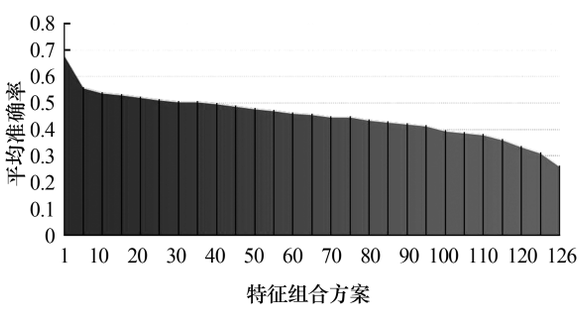

图 3　不同特征组合方案对预测精度的影响程度

由图 3 可知，在所有的特征组合方案中，本文提出的特征选择方案的准确率最高(68%)，故本文提出的特征组合方案为最优组合方案. 在本文提出的特征组合方案中，5 种特征与 C 语言成绩相关度最大的原因是：

1)学生的学习能力与地区的经济和教育发展水平存在一定相关性，因此来自不同地区的学生其学习能力存在一定的差异.

2)高考总成绩是反映一个学生学习能力的重要指标，因此 C 语言成绩与高考总成绩呈一定的相关性.

3)因朝鲜族考生的录取分数普遍低于汉族考生，且入学初期存在一定的汉语表达障碍[10](因朝鲜族考生在高考前主要接受的是朝鲜语教学)，因此朝鲜族学生在大一初期的学习成绩普遍偏低.

4)学好计算机程序设计课程需要学生具有较好的逻辑思维能力，而数学成绩在一定程度上能体现一个学生的逻辑思维能力，因此其与 C 语言成绩具有较大的相关性.

5)语文成绩能够体现学生的表达能力和理解能力，其对学习和理解知识至关重要，因此语文成绩和 C 语言成绩也具有较大的相关性.

为进一步说明基于随机森林分析方法的有效性，本文基于相同的数据集，计算了 2014—2016 年级的不同特征与 C 语言成绩间的 Pearson 相关系数、Spearman 相关系数、Kendall 相关系数，结

果(平均值)如表5所示.由表5可知,不同的特征和C语言成绩之间的相关系数均较低(低于0.36),表明其相关性较弱.

利用随机森林模型对各相关系数排名前5的特征进行训练,得到的模型准确率如图4所示.由图4可以看出,采用本文提出的随机森林分析法得出的模型准确率均高于采用3个相关系数分析法所得的准确率,因此表明采用本文提出的基于随机森林的方法分析高考信息和C语言成绩之间的相关性更为准确.

表5 不同特征与C语言成绩间的相关系数

特征	Pearson 相关系数	Kendall 相关系数	Spearman 相关系数
性别	0.064 337	0.064 336 815	0.064 336 815
民族	0.345 888	0.345 888 159	0.345 888 159
生源	0.205 697	0.138 249 956	0.211 637 290
考生类别	0.114 971	0.114 970 949	0.114 970 949
语文	0.085 457	0.085 456 932	0.085 456 932
数学	0.162 210	0.162 209 821	0.162 209 821
英语	0.108 745	0.108 745 335	0.108 745 335
综合	0.307 351	0.307 351 014	0.307 351 014
总成绩	0.353 254	0.353 253 833	0.353 253 833

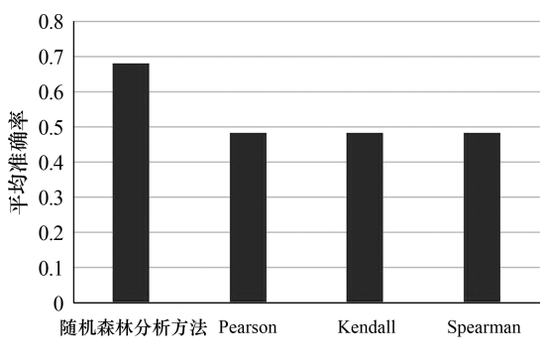

图4 不同特征组合方案的模型准确率

4 结论

本文利用基于随机森林算法的预测和分析方法对C语言成绩的影响因素进行了分析,结果表明生源、总成绩、民族、数学、语文5种特征与C语言成绩的相关性最高.本文的研究结果有助于教师根据新生的实际情况设计出具有针对性的教学模式,以提高程序设计课程的教学质量.本文在研究中所使用的数据量相对较少,因此在今后的研究中我们将进一步增加实验数据量以提高模型的拟合能力,使实验结果更具有普适性.

参考文献:

[1] 陈小杭.高考数学成绩与大学数学专业课学习能力相关性分析[J].长春教育学院学报,2019,35(2):8-10.

[2] 石铁玉,王维维,袁帅.工科学生高考成绩对大学阶段学习成绩的影响分析[J].中国电力教育,2014(8):239-240.

[3] 杜晓燕,丁厚成,林晓飞,等.大一成绩与高考成绩的相关性研究[J].安徽工业大学学报(社会科学版),2016,33(4):56-57.

[4] BREIMAN L. Statistical modeling: the two cultures (with comments and a rejoinder by the author)[J]. Statistical Science, 2001,16(3):199-231.

[5] 李航.统计学习方法[M].2版.北京:清华大学出版社,2019.

[6] BREIMAN L. Random forests[J]. Machine Learning, 2001,45(1):5-32.

[7] ZEITOUNI K, CHELGHOUM N. Spatial decision tree-application to traffic risk analysis[C]//ACS/IEEE International Conference on Computer Systems and Applications. Beirut: IEEE, 2001:203-207.

[8] 吕红燕,冯倩.随机森林算法研究综述[J].河北省科学院学报,2019,36(3):37-41.

[9] 王奕森,夏树涛.集成学习之随机森林算法综述[J].信息通信技术,2018(1):49-55.

[10] 申英美.中国朝鲜族教育问题研究[D].北京:中央民族大学,2006.