基于随机森林的数学试题难易度分类研究

梁琼芳,莎 仁

(东北师范大学 信息科学与技术学院,吉林 长春 130117)

摘 要:为了实现教育领域的“个性化”,无论是自由组卷的个性化,还是试题推荐的个性化,都首先需要确定试题难易度。研究目标为寻找新的方法解决基于试题难易度的分类问题,提高分类准确率。以高中数学为例,采用2018年多套高考数学试题作为实验数据,对原始数据各个特征进行相关性分析,剔除影响较小的特征,再采用随机森林算法探索试题难易度分类问题,对参数进行改进优化,并与其它分类方法进行对比。实验结果证明,采用随机森林的高中数学试题分类准确率高达90%,而其它3种分类算法准确率分别为72%、74%、74%。因此得出结论,随机森林算法在高中数学试题难易度分类上有较好表现,能够大幅提高分类准确率。

关键词:高中数学;试题难易度;分类算法;决策树;随机森林

DOI:10.11907/rjdk.191358 **开放科学(资源服务)标识码(OSID)**:

中图分类号:TP301 **文献标识码**:A **文章编号**:1672-7800(2020)002-0122-05

Classification of Mathematics Testability Difficulty Based on Random Forest

LIANG Qiong-fang, SHA Ren

(*School of Information Science & Technology, Northeast Normal University, Changchun* 130117, *China*)

Abstract: In order to realize individualization in the field of education, whether it is the individualization of the free test papers or the personalization of the test questions, the difficulty of the test questions must firstly be determined. Therefore, the research goal of this paper is to find new ways to solve the test questions. The classification problem of difficulty is easy, and the accuracy of classification is improved. Taking high school mathematics as an example, in this paper, the mathematics test questions of the college entrance examination in 2018 are used as experimental data, and the correlation analysis of each feature of the original data is carried out to eliminate the features with less influence. Then the random forest algorithm is used to explore the difficulty classification of the test questions, and the parameters are improved and optimized and compared with other classification methods. Experiments show that the accuracy rate of random forests for high school mathematics test classification is as high as 90%, while the accuracy of other classification algorithms is 72% and 74%. Therefore, it is concluded that the random forest algorithm has excellent performance in the classification of high school mathematics questions and can greatly improve the classification accuracy.

Key Words: high school mathematics; test difficulty; classification algorithm; decision tree; random forest

0 引言

近年来,个性化推荐技术正在各个领域迅速兴起,而教育领域作为当今社会必不可少且不容忽视的一部分,越来越需要“个性化”的引入。如今网络试题题库、组卷系统层出不穷,都是为了实现学生的高效练习,而确定试题难易程度是题库构建,以及自由组卷与试题个性化推荐的基础。

在数学试题难易度研究方面,国外Pollitt等[1]在1985年提出难度的3个来源,1996年剑桥考试委员会研究者[2]从权威角度提出影响数学试题难易度的因素,1999年Ahmed等[3]研究了试题认知要求程度对问题难度的影响,直至2006年Leong[4]归纳了影响试卷难度的4个因素,分别为内容、材料、主体因素与命题者决策。在国内,1994年任子朝等[5]提出可从多个客观角度评估试题难度;2002年李红松等[6]提出试题难易度与学生成绩分布有关,并采用主观模糊评价方法结合成绩分布确定试题难易度;2008年,教育部考试中心[7]归纳总结了影响试题难易度的因素,包括知识点个数、运算过程步骤数、推理转折数、设陷数、

收稿日期:2019-03-20

作者简介:梁琼芳(1993-),女,东北师范大学信息科学与技术学院硕士研究生,研究方向为教育支撑软件;莎仁(1993-),女,东北师范大学信息科学与技术学院硕士研究生,研究方向为教育大数据。

创新度、繁琐度、启发度、猜测度等；2016 年，候飞飞[8]根据试题自身特点，结合 C4.5 决策树方法，对物理试题进行难易程度分类研究，验证了决策树分类算法的可行性；2018 年，陈荟慧等[9]进行基于在线测评系统的编程题目难度研究，但仍然依赖于被试的作答通过率；同年曹开奉等[10]总结归纳了我国高考理科试题难度影响因素，为本文研究打下了基础。本文致力于实现高中数学试题的客观难易度分类，以避免通过人为主观判断或过分依赖于被试作答通过率进行难易度分类造成的偏差。

常用分类算法如下：典型的朴素贝叶斯方法，针对大量数据训练速度较快，并支持增量式训练，对结果的解释便于理解，但在大数据集下才能获得较为准确的分类结果，且忽略了数据各属性值之间的关联性[11]；K-最近邻分类算法比较简单，训练过程迅速，抗噪声能力强，新数据可以直接加入训练集而不必重新进行训练，但在样本不平衡时结果偏差较大，且每次分类都需要重新进行一次全局运算[12]；决策树分类算法易于理解与解释，可进行可视化分析，运行速度较快，可扩展应用于大型数据库中，但容易出现过拟合问题，且易忽略数据属性间的关联性[13]。

自 2000 年以来，深度学习等人工智能技术得到了迅速发展，在很多领域都取得了较好的应用效果。其中随机森林算法在分类方面表现突出，其避免了决策树分类算法中容易出现的过拟合问题，并在运算量未显著增加的前提下，提高了分类准确率[14]。因此，本文旨在利用随机森林算法实现一种更精确、客观的试题难易度分类方法，既能节省人力，又可提升分类准确率与客观性。

1　随机森林

1.1　决策树——随机森林的基分类器

决策树作为随机森林的基分类器，是一种单分类器的分类技术，也是一种无参有监督的机器学习算法[15]。决策树可视为一个树状模型，由节点与有向边组成，其中包括 3 种节点：根节点、中间节点和叶子节点。决策树构建不需要先验知识，并且比诸如神经网络的方法更容易解释。决策树分类思想实际上是一个数据挖掘过程，其通过产生一系列规则，然后基于这些规则进行数据分析。构建决策树的一个关键问题是节点分裂特征选择，由于不同分裂标准对决策树的泛化误差有很大影响，因此根据不同划分标准，学者们提出了大量决策树算法[16]。

其中 Hunt 等[17]提出的 CLS 算法随机选择分裂节点，Quinlan 等[18]提出的 ID3 算法基于信息熵，C4.5 算法基于信息增益率[19]，Breiman 等[20]提出的 CART 算法基于 Gini 指标，然而没有一种算法在各种数据集上都能得到最好结果。决策树采用单一决策方式，因此具有以下缺点：一是包含复杂的分类规则，一般需要决策树事前剪枝或事后剪枝；二是收敛过程中容易出现局部最优解；三是因决策树过于复杂，容易出现过拟合问题。

1.2　随机森林构建

为了克服以上所述决策树算法的不足，结合集成学习思想[21]，研究者们提出了“森林”的概念。森林中的决策树按照一定精度进行分类，最后所有决策树参与投票决定最终分类结果，这是随机森林的核心概念。随机森林构建主要包括以下 3 个步骤：

（1）为 N 棵决策树抽样产生 N 个训练集。每一棵决策树都对应一个训练集，主要采用 Bagging 抽样方法从原始数据集中产生 N 个训练子集。Bagging 抽样方法是无权重的随机有放回抽样，在每次抽取样本时，原数据集大小不变，但在提取的样本集中会有一些重复，以避免随机森林决策树中出现局部最优解问题[22]。

（2）决策树构建。该算法为每个训练子集构造单独的决策树，最终形成 N 棵决策树以形成“森林”。节点分裂原则一般采用 CART 算法或 C4. 5 算法，在随机森林算法中，并非所有属性都参与节点分裂指标计算，而是在所有属性中随机选择某几个属性，选中的属性个数称为随机特征变量。随机特征变量的引入是为了使每棵决策树相互独立，减少彼此之间的关联性，同时提升每棵决策树的分类准确性，从而提高整个森林的性能。

（3）森林形成及算法执行。重复步骤（1）、（2），构建大量决策树，形成随机森林。算法最终输出由多数投票方法实现。将测试集样本输入随机构建的 N 棵决策子树进行分类，总结每棵决策树分类结果，并将具有最大投票数的分类结果作为算法最终输出结果。

随机森林算法原理如图 1 所示。

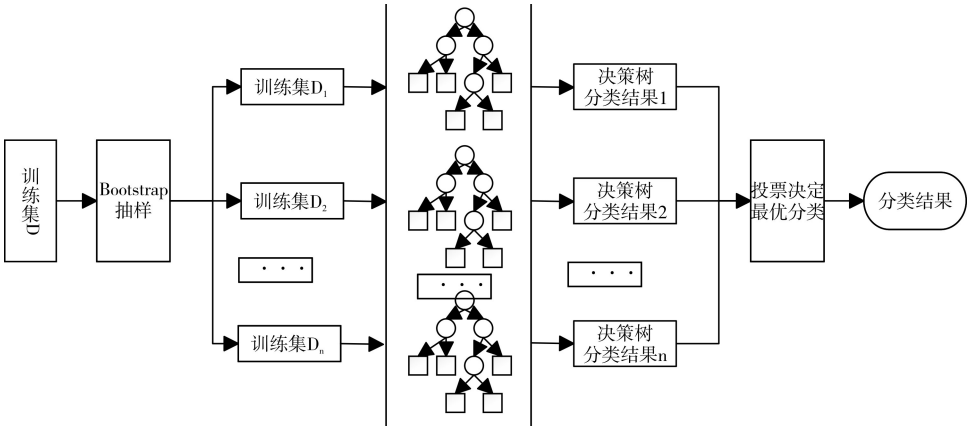

图 1　随机森林算法原理

2 基于随机森林的试题难易度分类模型构建及优化

2.1 数据特征分析与选择

本文采用的试题数据为 2018 年全国各省高考数学试题，部分试题特征来源于组卷网，但其涵盖的试题特征不够全面，故其它影响难易度的试题特征可通过对答案的解析加以确定，并自主进行数据标记，主要字段说明见表 1。

（1）无关数据剔除。表 1 中序 1、2、3、5 特征对试题难易度分类没有价值，不作为训练特性，故删除该字段。

（2）对连续性变量，采用 Pearson（皮尔森）相关系数方法验证与试题难易度值相关关系是否显著[23]，属性中连续变量有 textLength 和 guessMeasure，其与难度值的 Pearson 相关性系数分别为 0.325 031 和-0.095 424，故保留 textLength，删除 guessMeasure。

表 1 试题数据特征及说明

序	字段名（中）	字段名（英）	解释说明	变量类型	取值范围	趋势难度预测（易→难）
1	题目 id	questionId	试题唯一标识	连续变量	1≤n≤300	无
2	试题年份	questionYear	试题属于的高考年份	文本	无	无
3	试题地区	questionPos	试题属于的高考地区	文本	无	无
4	题型	type	试题类型	分类变量	n=0，1	0…1
5	涵盖知识点名称	knowledgeName	解答应用的知识点	文本	无	无
6	知识点数量	knowledgeNum	解答应用的知识点数量	等级变量	1≤n≤6	1…6
7	难度等级	difficultyLevel	客观难度等级、分类标准	等级变量	1≤n≤3	分类标准
8	题目文本长度	textLength	试题题目文本长度	连续变量	2≤n≤329	2…239
9	背景水平	backgroundLevel	试题是否包含参数	等级变量	0≤n≤5	0…5
10	解题步骤	solveStep	解得答案需要多少步骤	等级变量	1≤n≤12	1…12
11	物理运算复杂水平	physicalLevel	解题所需推理与转换次数	等级变量	1≤n≤12	1…12
12	数学运算复杂水平	mathLevel	数学计算难度水平	等级变量	1≤n≤3	1…3
13	知识点位置	knowledgePos	运用的知识点是否冷门	分类变量	n=0，1	1…0
14	内容模块涉及数	moduleNum	题干包含知识点的分散程度	等级变量	1≤n≤6	1…6
15	条件满足性	conditionSatisfact	已知条件与解答条件的关系	分类变量	n=0，1	1…0
16	内容表达方式	expressionWay	文字描述、图形或表格等	分类变量	n=0，1	1…0
17	思维方式	thinkingWay	顺向、逆向或双向思维	等级变量	0≤n≤2	0…2
18	猜测度	guessMeasure	根据题型给定的猜测度	连续变量	0≤n≤1	1…0
19	启发度	inspireMeasure	题干陈述对解题是否有启发	分类变量	n=0，1	1…0
20	内容情境新颖度	novelMeasure	是否是常见题	等级变量	1≤n≤3	1…3

（3）对于二分类变量，采用点二列相关系数方法验证与试题难易度值相关关系是否显著[24]，特征中二分变量与难易度的点二列相关系数分别为 type0.295 424、knowledgePos-0.149 294、conditionSatisfact-0.442 642、expressionWay-0.011 241 和 inspireMeasure0.011 241，故只保留 type 与 conditionSatisfact 特征，删除其它特征。

（4）对于等级变量，采用 Spearman（斯皮尔曼）等级相关系数方法验证与试题难易度值相关关系是否显著[25]，特征中等级变量与难易度 Spearman 相关系数分别为 knowledgeNum0.460 722、backgroundLevel0.266 939、solveStep 0.580 002、physicalLevel0.587 000、mathLevel0.514 686、moduleNum0.406 973、thinkingWay0.066 568 和 novelMeasure0.130 309，删除 thinkingWay 与 novelMeasure 特征，保留其它特征。

综上，最终选择影响试题难易度的 9 个特征。采用随机森林算法作特征选择，可以很好地解决过拟合问题，同时也能过滤掉重要性很低的特征，提高模型分类准确率。

2.2 模型构建与优化

采用 CART 算法作为随机森林构建决策树的方法，采用 Gini 系数最小准则进行节点分裂。CART 算法在训练过程中需要计算每个属性的 Gini 指标，并选择一个具有最小 Gini 指标的变量对当前节点进行分裂，通过递归形式构建决策树，直至达到停止条件。Gini 系数计算公式如下：

$$Gini(m)=1-\sum_{k=1}^{K}{p_{mk}}^2 \quad (1)$$

式（1）中 K 表示有 K 个类别，p_{mk} 表示节点 m 中类别 k 所占比例，当 Gini 取最小值 0 时，此时数据类别最纯；当 Gini 取最大值 1 时，则表示当前节点的数据类别不同。根据式（1）计算特征的 Gini 系数，将 Gini 值最小的点作为该层分裂节点，递归地构建决策树。重复上述步骤，形成随机森林。构建过程中各特征重要性见表 2。

对随机森林的 minimal node size 与 mtry 进行参数寻优，最终确定构建的最优随机森林 node size 为 33，mtry 为 4。其中 minimal node size 寻优过程中测试集分类准确率

变化见图 2。

表 2　随机森林中特征重要性分值

特征因子名称	重要性
solveStep	100.000
physicalLevel	87.811
mathLevel	69.851
textLength	63.551
knowledgeNum	33.252
moduleNum	11.492
conditionSatisfact	6.986
backgroundLevel	2.355
type	0.193

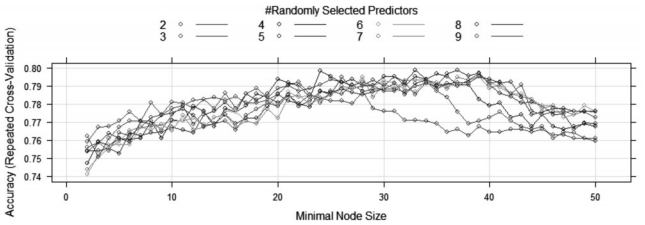

图 2　参数寻优过程中分类准确率变化

3　实验与分析

3.1　实验设计

实验分为两个阶段：模型训练阶段与测试阶段。将数据集按 7∶3 的比例划分为训练集和测试集，分别利用朴素贝叶斯分类、KNN 分类、决策树分类以及本文构建的随机森林方法进行分类预测实验，并将不同算法的混淆矩阵指标及准确率 Accuracy 进行对比[26]。

3.2　实验结果

KNN 分类算法中，neighbors 值变化与最终分类准确率关系变化见图 3，故最终选用 5-nearest neighbor model 模型。

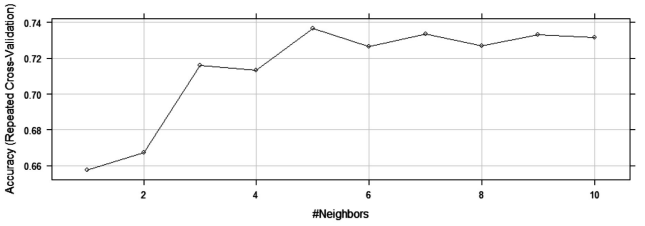

图 3　KNN 参数寻优过程

4 种分类算法实验结果见表 3、表 4。

表 3　4 种分类算法实验结果(一)

算法	朴素贝叶斯(nativeb)			KNN		
三分类	Class:X1	Class:X2	Class:X3	Class:X1	Class:X2	Class:X3
Sensitivity	1.000 0	0.631 6	1.000 0	0.285 7	0.921 1	0.000 0
Specificity	0.744 2	1.000 0	0.933 3	0.953 5	0.166 7	0.977 8
Accuracy		0.720 0			0.740 0	

表 4　4 种分类算法实验结果(二)

算法	决策树(dt)			随机森林(rf)		
三分类	Class:X1	Class:X2	Class:X3	Class:X1	Class:X2	Class:X3
Sensitivity	0.571 4	0.736 8	1.000 0	0.857 1	0.921 1	0.800 0
Specificity	0.883 7	0.750 0	0.888 9	0.975 0	0.833 3	0.953 5
Accuracy		0.740 0			0.900 0	

3.3　结果对比

将朴素贝叶斯、KNN、决策树和随机森林分类算法的实验结果召回率 Sensitivity、特异度 Secificity 与准确率 Accuracy 进行对比，结果如图 4-图 6 所示。

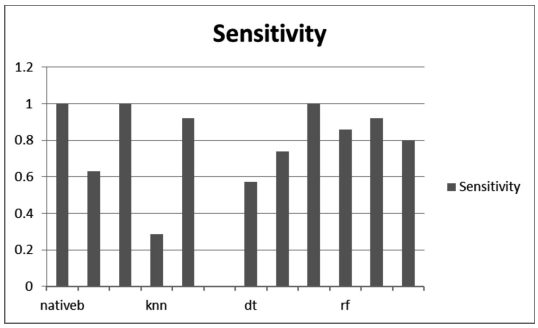

图 4　召回率对比

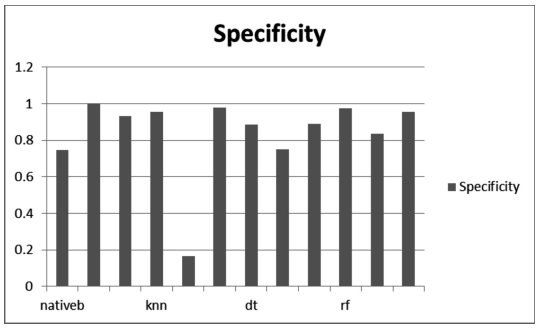

图 5　特异度对比

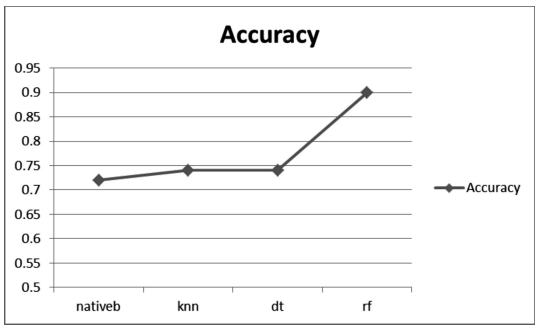

图 6　准确率对比

由上图可以看出，随机森林的召回率和特异度优于其它 3 种分类算法，且分类准确率明显高于其它 3 种分类算法，故验证了本文方法的正确性及有效性。

4　结语

本文将随机森林分类方法应用于高考数学试题客观

难易度分类,大幅提高了分类准确率,为试题个性化推荐与自由组卷系统奠定了基础。但由于网上开源的教育数据较少,故应用的实验数据集较小,使用大数据集应能进一步提高分类准确率,但有待后续进一步验证。另外,本文只分析了影响数学学科试题难易度的因素,对于英语、语文、生物等学科试题,其难易度影响因素还有待进一步分析与探索,这也将是未来的研究方向。

参考文献:

[1] ALASTAIR P, CAROLYN M, et al. Language, contextual and cultural constraints on examination performance[C]. Jerusalem:the International Association for Educational Assessment,2000.

[2] HANNAH F H, SARAH H. What makes mathematics exam questions difficult[R]. Research and Evaluation University of Cambridge Local Examinations Syndicate,2006.

[3] AYESHA A, ALASTAIR P. Curriculum demands and question difficulty [C]. Slovenia:IAEA Conference,1999.

[4] CHENG L S. On varying the difficulty of test items[C]. Annual Conference of theInternational Association for Educational Assessment, Singapore, 2006.

[5] 任子朝.高考数学命题研究[J].中学数学教学参考,1994(5):1-4.

[6] 李红松,田益祥.试题难易程度的判断及其集对分析测定方法研究[J].武汉科技大学学报:自然科学版,2002, 25(2):216-217.

[7] 教育部考试中心.2008 年普通高等学校招生全国统一考试大纲:理科[M].北京:高等教育出版社,2008.

[8] 候飞飞.基于 C4.5 决策树的试题难易程度分类研究[D].新乡:河南师范大学,2016.

[9] 陈荟慧,熊杨帆,蒋滔滔,等.基于在线测评系统的编程题目难度研究[J].现代计算机:专业版,2018(13):28-32,36.

[10] 曹开奉,王伟群,刘芳.我国高考理科试题难度影响因素的文献分析[J].考试研究,2018 (3): 40-46.

[11] LEWIS D D. Naive (Bayes) at forty: the independence assumption in information retrieval[C]. European Conference on Machine Learning,1998.

[12] TANG Q Y, ZHANG C X. Data Processing System (DPS) software with experimental design, statistical analysis and data mining developed for use in entomological research [J]. 中国昆虫科学:英文版, 2013, 20(2):254-260.

[13] ROMERO C, VENTURA S. Educational data mining: a survey from 1995 to 2005[J]. Expert Systems with Applications, 2007, 33(1): 135-146.

[14] SVETNIK V, LIAW A, TONG C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling [J]. Journal of Chemical Information & Computer Sciences, 2003, 43(6):1947.

[15] 张琳,陈燕,李桃迎,等.决策树分类算法研究[J].计算机工程,2011,37(13):66-67.

[16] 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,2018,12(1):49-55.

[17] 曹正凤.随机森林算法优化研究[D].北京:首都经济贸易大学,2014.

[18] UTGOFF P E. ID: an incremental ID3[M]. Massachusetts:University of Massachusetts,1987.

[19] QUINLAN J R. C4.5: programs for machine learning [M]. San Mateo:Morgan Kaufmann Publishers Inc,1992.

[20] DEATH G, FABRICIUS K E. Classification and regression trees: a powerful yet simple technique for ecological data analysis [J]. Ecology, 2000, 81(11):3178-3192.

[21] 孔英会.基于混淆矩阵和集成学习的分类方法研究[J].计算机工程与科学, 2012, 34(6):111-117.

[22] 沈学华,周志华,吴建鑫,等. Boosting 和 Bagging 综述[J].计算机工程与应用,2000, 36(12):31-32.

[23] HUBER P J, STRASSEN V. Minimax tests and the neyman-pearson lemma for capacities[J]. Annals of Statistics, 1973 (2):251-263.

[24] 陈冠民,张选群,陈华.多序列相关系数及其估计[J].数理医药学杂志, 1999, 12(2):101-102.

[25] ZAR J H. Significance testing of the Spearman rank correlation coefficient[J]. Publications of the American Statistical Association, 1972, 67(339):578-580.

[26] 宋亚飞,王晓丹,雷蕾.基于混淆矩阵的证据可靠性评估[J].系统工程与电子技术,2015,37(4):974-978.

(责任编辑:黄 健)